Non-convex Projections for Low-rank Matrix Recovery

Prateek Jain

Microsoft Research, India

Acknowledgements:

a) Matrix Linear Regression: Raghu Meka, Inderjit Dhillon

b) Matrix Completion: Praneeth Netrapalli

c) Robust PCA: Anima Anandkumar, Praneeth Netrapalli, Niranjan U N, Sujay Sanghavi

Overview

- Provable non-convex projections for low-rank matrix recovery

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

- Projected gradient descent:

$$X_{t+1} = P_r(X_t - \eta \nabla f(X_t))$$

- $P_r(Z)$: projection onto set of rank-r matrices
    - Non-convex set

$$P_r(Z) = \arg \min_{\mathrm{X,\, rank(X)} \leq r} ||X - Z||_F^2$$

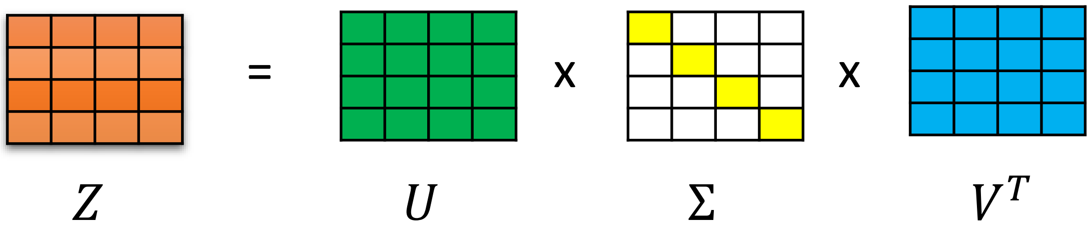
Non-convexity of Low-rank manifold

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

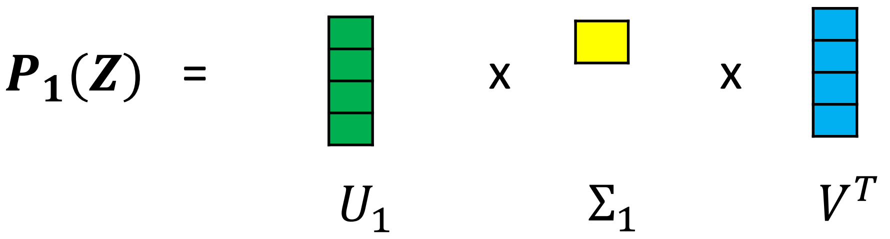
Projection onto set of Low-rank Matrices

- 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!!!

Our Results

- RIP/RSC based Linear Regression

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

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

- Matrix Completion

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

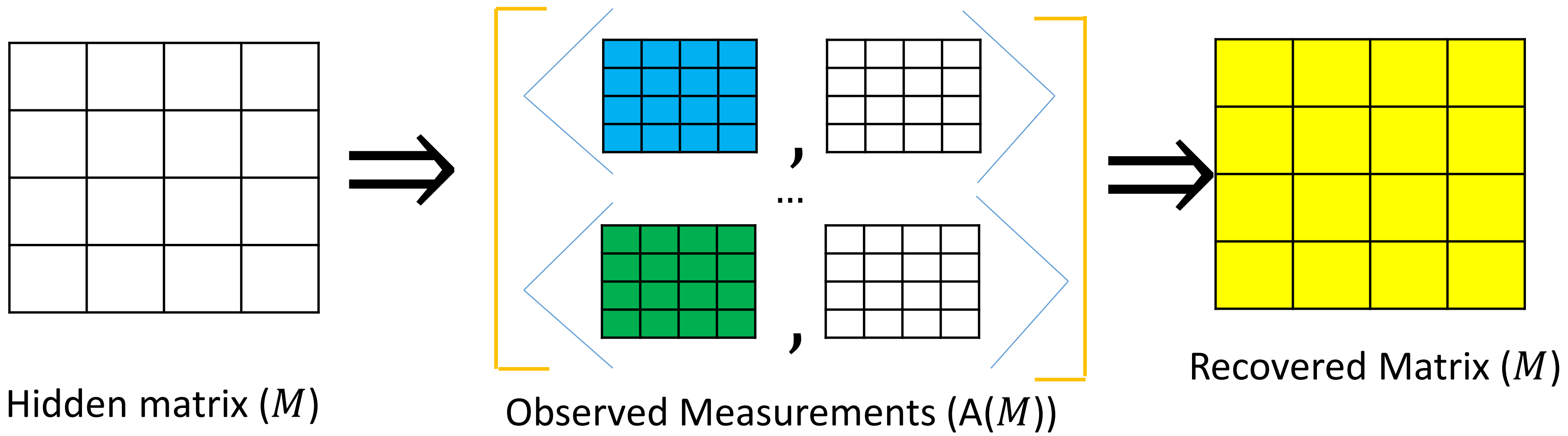
  - Ω: randomly sampled, M: incoherent matrix

- Non-convex Robust PCA

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

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

Low-rank Matrix Sensing



Matrix Linear Regression

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

- $\mathbb{A}: \mathbf{R}^{\boldsymbol{n \times n}} \to \boldsymbol{R^d}$
  - 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:

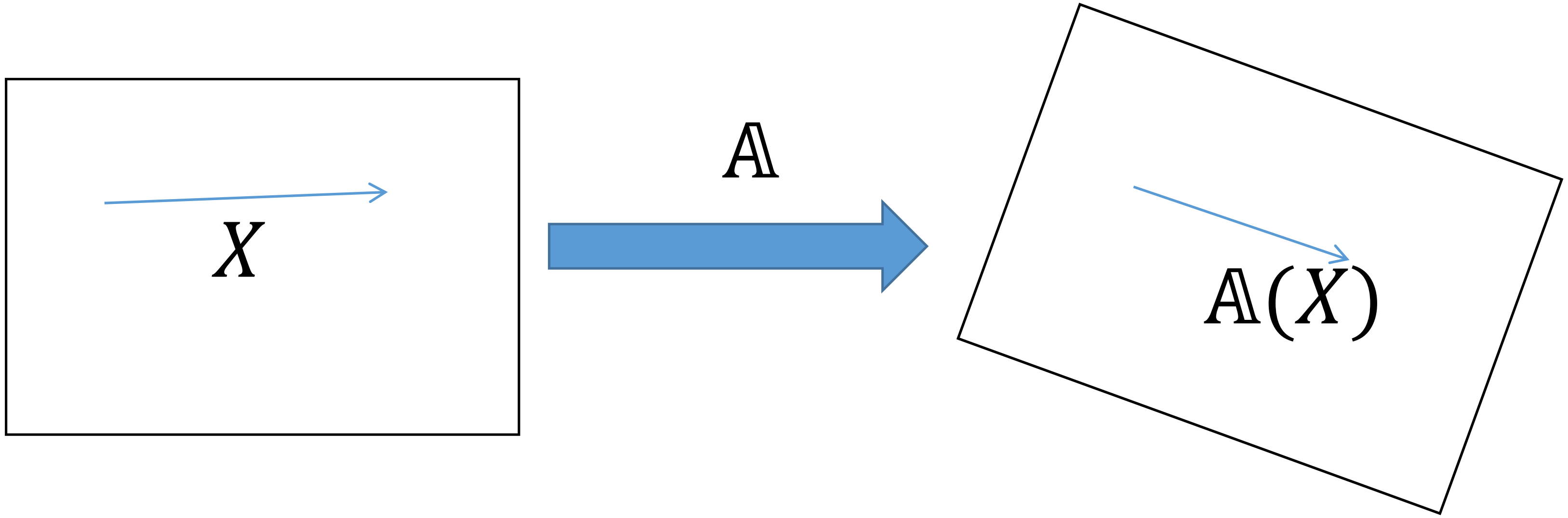
$$\min_X ||\mathbb{A}(X) - b||_2^2$$
$$s.t. \;\; rank(X) \leq r$$

Low-rank Matrix Estimation

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

- 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):

$$(1-\delta_r)||X||_F^2 \leq ||\mathbb{A}(X)||_2^2 \leq (1+\delta_r)||X||_F^2$$

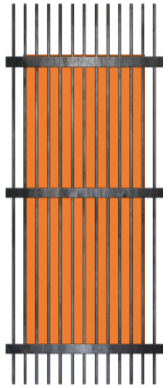
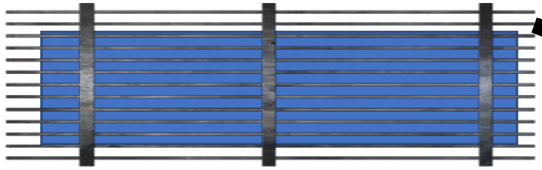
- Examples:
  - $\mathbb{A}$: sampled from multivariate normal distribution
  - $\mathrm{m} = O(\frac{r}{\delta_r^2} n \log n)$

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

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

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

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

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

- 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

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

- 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

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

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

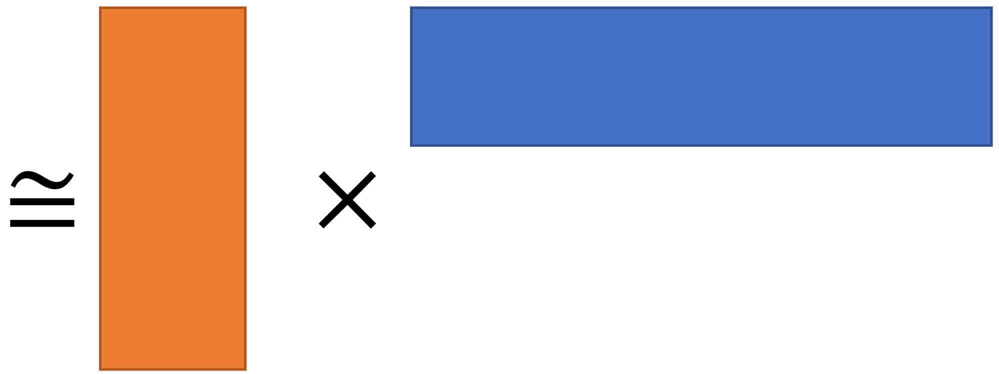
- Task: Complete ratings matrix
- Applications: recommendation systems, PCA with missing entries

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



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

- 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

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

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

- 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]

Approach 2

- Alternating Minimization: $X = UV^T$
- Provably recovers M if:
  - $|\Omega| \geq O(poly(r) n \log n \log\left(\frac{\sigma_1}{\sigma_r}\right) \log(\frac{1}{\epsilon})$
  - σ_i: i-th singular value of M
  - ϵ: accuracy parameter
- Computation time: $O(|\Omega| r^2)$
  - Nearly linearly computation time
- Sample complexity: dependence on $\kappa = \sigma_1 / \sigma_r$
- Refs: [J., Netrapalli, Sanghavi'13], [Hardt, Wooters'14]

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: $X_t \rightarrow M$ 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)

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

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

$$\begin{aligned}&||M - X_t||_\infty \\ &\le ||M - M\,\frac{vv^T}{\lambda}M||_\infty + \sum_{a\ge 0, b\ge 0, a+b\ge 1}^{\infty} \left|\left(\frac{H_t}{\lambda}\right)^a \frac{Mvv^TM^T}{\lambda}\left(\frac{H_t}{\lambda}\right)^b\right|_\infty\end{aligned}$$

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

$$\begin{aligned}||M - M\,\frac{vv^T}{\lambda}M||_\infty &\le \max_{i,j} e_i^T u^* \left(1 - {u^*}^T\frac{vv^T}{\lambda}u^*\right){u^*}^T e_j \\ &\le \max_{i,j} |e_i^T u^*||e_j^T u^*|\,|1 - ({u^*}^T v)^2/\lambda| \\ &\le \frac{\mu^2}{n}4||H_t||_2 \le 8\mu^2\delta||M - X_{t-1}||_\infty\end{aligned}$$

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

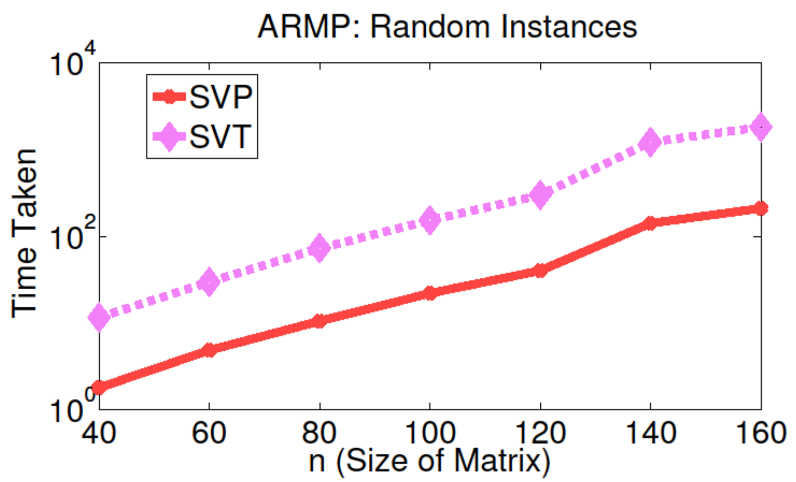
- After t-th step of k-th stage:

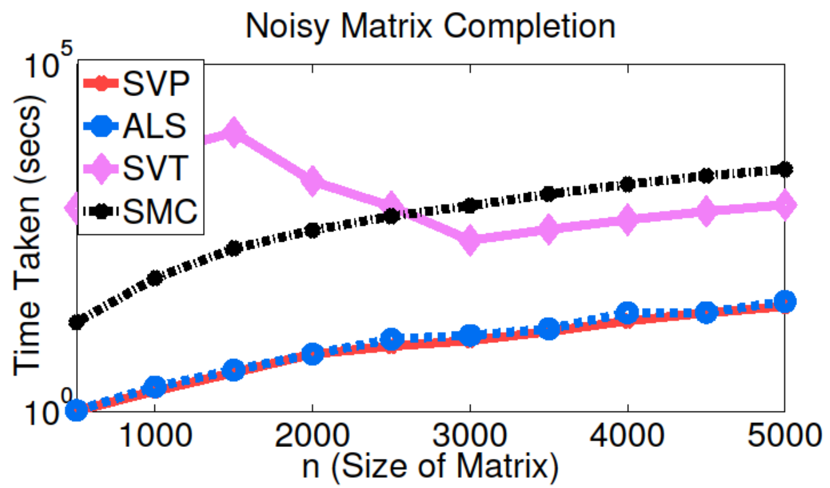
$$||M - X_t||_\infty \leq \frac{2\mu^2 r}{n}\left(\sigma_{k+1} + \left(\frac{1}{2}\right)^t \sigma_k\right)$$

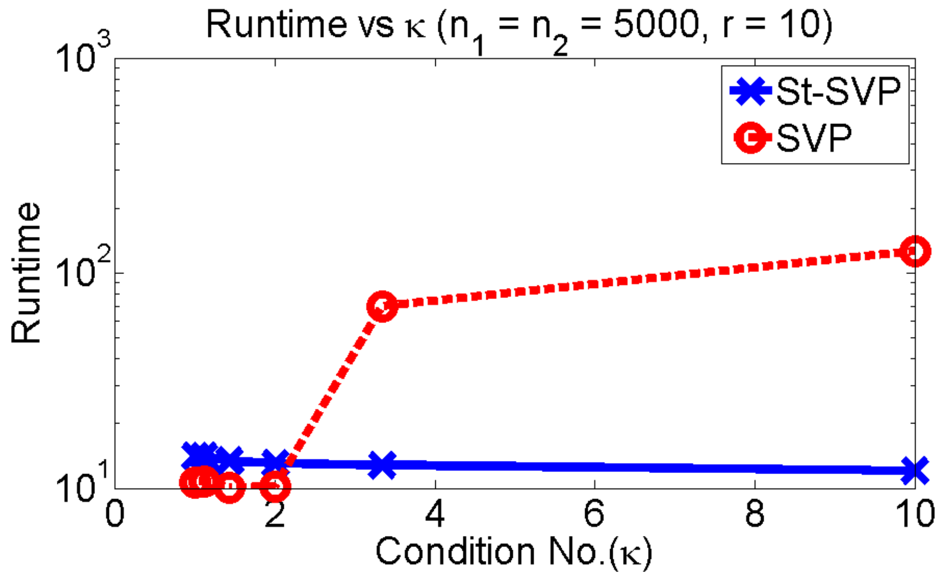
- M: rank-r i.e. $\sigma_{r+1} = 0$
- After $\mathrm{T} = \log(\frac{1}{\epsilon})$ steps of r-th stage: $||M - X_T||_\infty \leq \epsilon$
- Sample complexity: $|\Omega| \geq nr^4\mu^2 \log n \log 1/\epsilon$
- Computation complexity: $O(nr^6\mu^2 \log n \ \log\frac{1}{\epsilon})$
  - Linear in n
  - 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 PCA

Robust PCA

- M=L+E
  - Standard PCA: recover L upto $||E||_2$
  - $|| \hat{L} - L|| \leq ||E||_2, rank(\hat{L}) \leq rank(L) = r$
- Corrupted with arbitrarily large (but sparse) errors

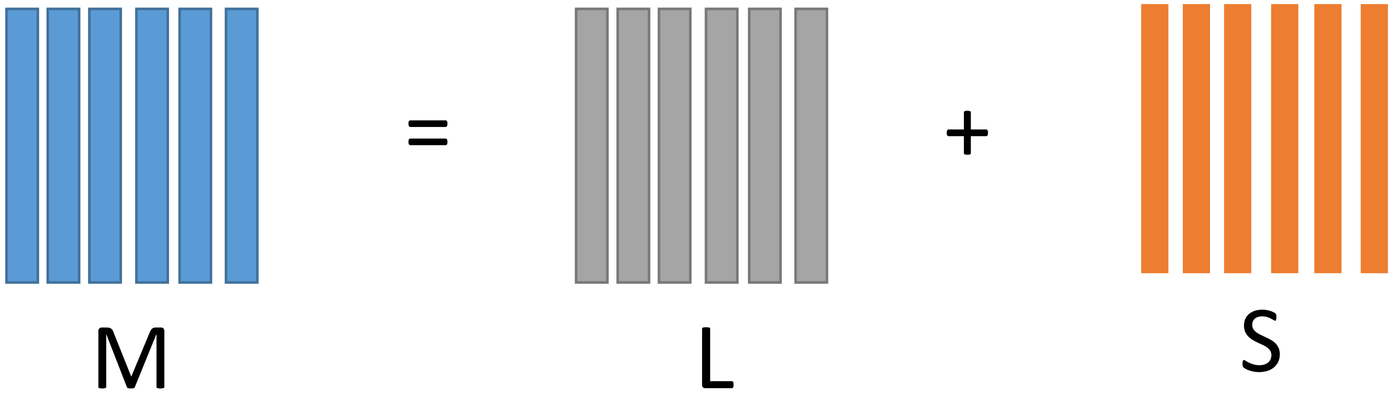
$$M = L + S$$

  - L: low-rank matrix
  - S: sparse matrix
- Goal: Given $M \in R^{n\times n}$, decompose matrix into L, S

Motivation

- Adversarial corruption of a few coordinates per data point
- Foreground-background subtraction

Original Video Background Foreground



Harder Problem than Matrix Completion ?

1		3			5			5		4	
		5	4			4			2	1	3
2	4		1	2		3		4	3	5	
	2	4		5			4			2	
		4	3	4	2					2	5
1		3		3			2			4	

Corrupt ratings - rating between 1 to 5

- But, in MC: known and correct entries are only $O(\log n)$ per row
- In Robust PCA, we can allow $O(n)$ correct elements per row

Identifiability?

- Unique decomposition not achievable in general:
  - $L = e_1 e_1^T,\ S = e_1 e_1^T$
- Assumptions:
  - L: rank-r μ −incoherent matrix
    - $L = U\Sigma U^T$
    - $||U^i||_2 \leq \frac{\mu\sqrt{r}}{\sqrt{n}}$
  - S: d-sparse matrix
    - Each row and column of S has at most d nonzeros

Existing Method

$$\min_{\hat{L},\hat{S}} ||\hat{L}||_* + \lambda||\hat{S}||_1$$
$$s.t.\ \ M = \hat{L} + \hat{S}$$

- Convex program
- Running time: $O(n^3)$
- Assumption: $d \leq \frac{n}{\mu^2 r}$
- Question: PCA time complexity for Robust PCA?
  - $O(n^2 r)$ algorithm?

Our Approach (NcRPCA)

- $M_0 = 0$
- $L_0 = 0$
- For k=1…r
  - For t=1, 2… T
    - $M_t = M_{t-1} - H_\tau(M_{t-1} - L_{t-1})$ //Hard Thresholding
    - $L_t = P_r(M_t)$ //Projection onto low-rank matrices
  - End For
- End For
- Runtime: $O(n^2 r^2)$

Results

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

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

- Assumption: $d \leq \frac{n}{\mu^2 r}$ (same as convex relaxation)
- Running time: $O(n^2 r^2 \log \frac{1}{\epsilon})$

[NUSAJ'14]

Proof Technique

- $M_t = M_{t-1} - H_\tau(M_{t-1} - L_{t-1})$
- $L_t = P_r(M_t)$
- Let $M_t = L + S_t$
- Good properties only if S_t is “sparse”
- Set τ s.t.
  - $\mathrm{supp}(S_t) \subseteq supp(S)$
  - $||S_t||_\infty \leq .5\ ||S_{t-1}||_\infty$
- But for this, we need $||L_t - L||_\infty \leq .1\ ||S_{t-1}||_\infty$
  - Somewhat similar to matrix completion, but different assumptions

Proof setup

- $L_t = P_1(L + S_{t-1}), \quad L_t = \lambda v v^T$

$$(L + S_{t-1})v = \lambda v$$

$$\left(I - \frac{S_{t-1}}{\lambda}\right)v = \frac{Lv}{\lambda}$$

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

$$L - L_t = L - \lambda v v^T$$

$$= L - L\frac{vv^T}{\lambda}L - \sum_{a\geq 0, b\geq 0, a+b\geq 1}^{\infty} \left(\frac{S_{t-1}}{\lambda}\right)^{a} \frac{Lvv^T L^T}{\lambda} \left(\frac{S_{t-1}}{\lambda}\right)^{b}$$

Result

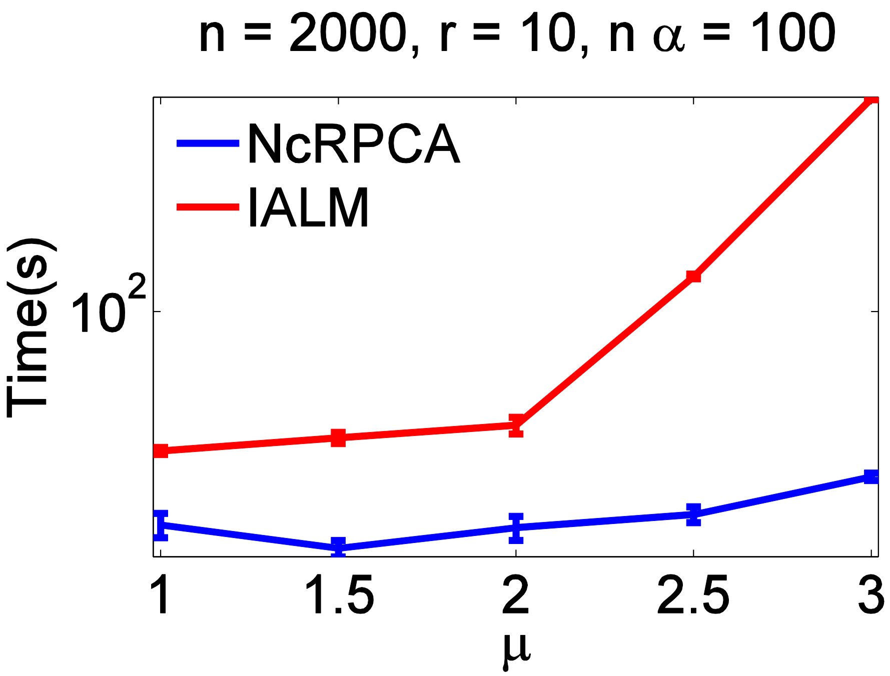
- After t-th step of k-th stage:

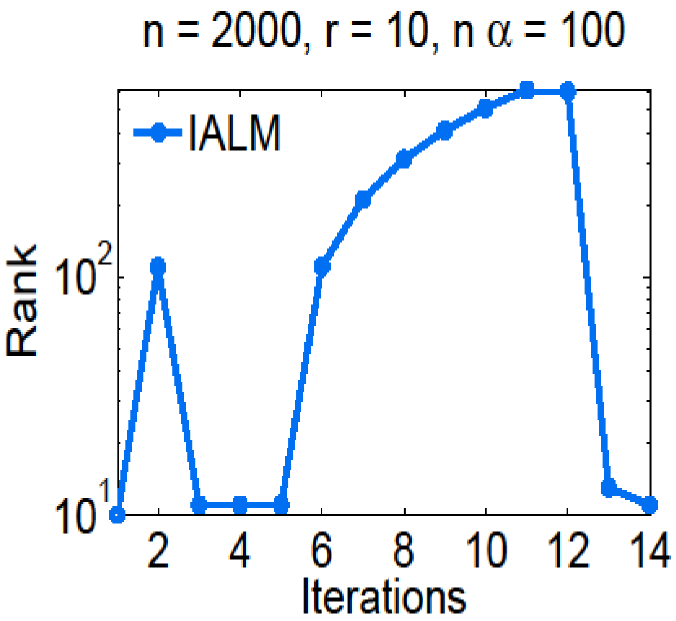
$$||L - L_t||_\infty \leq \frac{2\mu^2 r}{n}\left(\sigma_{k+1} + \left(\frac{1}{2}\right)^t \sigma_k\right)$$

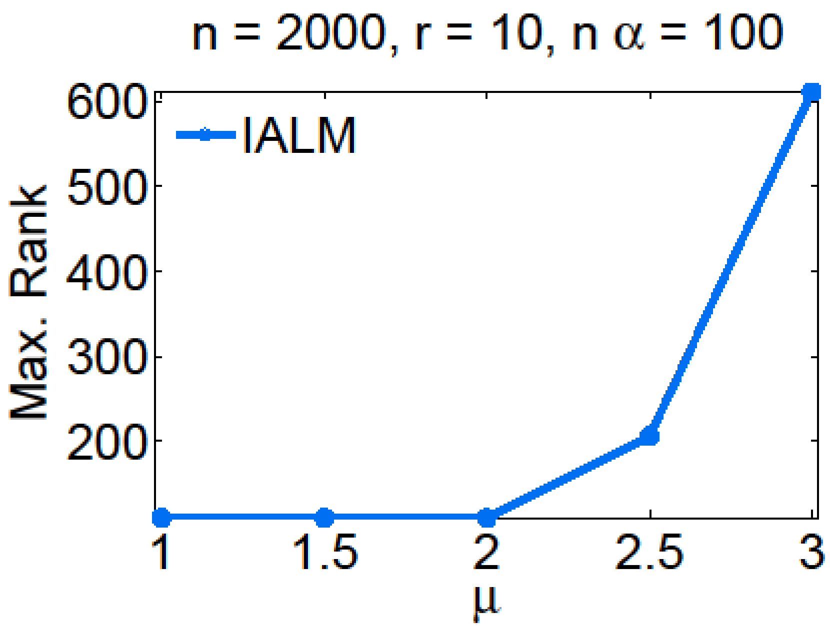
- L: rank-r i.e. $\sigma_{r+1} = 0$
- After $\mathrm{T} = \log(\frac{1}{\epsilon})$ steps of r-th stage: $||L - L_T||_\infty \leq \epsilon$
- Computation complexity: $O(n^2 r^2 \log \frac{1}{\epsilon})$
  - $O(r \log \frac{1}{\epsilon})$ more expensive than PCA
- Require conditions similar to Chandrasekharan et al'2009

[NUSAJ'14]

Empirical Results







Empirical Results

Original Image

PCA

Convex RPCA

Non-Convex RPCA

Runtime:

- Convex RPCA: 3500s
- NcRPCA: 118s

Summary

- Main message: non-convex projected gradient descent converges
  - If underlying functions has special structure
- Problems considered:
  - RIP/RSC based function optimization
  - Matrix completion
  - Robust PCA
- Provable guarantees
  - Significantly faster than the convex-surrogate based methods
  - Empirical results match the theoretical observation

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!