Provable Alternating Minimization Methods for Non-convex Optimization

Prateek Jain

Microsoft Research, India

Joint work with Praneeth Netrapalli, Sujay Sanghavi, Alekh Agarwal, Animashree Anandkumar, Rashish Tandon

Outline

- Alternating Minimization
  - Empirically successful
  - Very little theoretical understanding
- Three problems:
  - Low-rank Matrix Completion
  - Phase Retrieval
  - Dictionary Learning
- Open problems

Optimization over two variables

$$\min_{U,V} \quad f(U,V)$$

- Alternating Minimization:
  - Fix U, optimize for V
    $$V^t = arg \min_V f(U^t, V)$$
  - Fix V, optimize for U
    $$U^{t+1} = arg \min_U f(U, V^t)$$

- Generic technique
  - If each individual problem is "easy"
  - Forms basis for several generic algorithm techniques like EM algorithms

A few known ML-related applications

- EM algorithms
- Recommendation systems
- Dictionary Learning
- Low-rank matrix estimation
- Active Learning
- Phase Retrieval
- …..

Known Theoretical Results

- Known Results:
  - f: convex function jointly in U, V
  - f: smooth function in both U, V
  - Then, Alternating minimization converges to global optima
- Known counter-examples if either of the conditions not satisfied
  - Does not converge to correct solution even if f is not smooth
- In many practical problems: f is non-convex !!!!
  - But surprisingly method works very well in practice

Our Contribution

- Studied three important ML-related problems
  - Low-rank Matrix Completion (Recommendation systems)
  - Phase Retrieval (X-ray Crystallography)
  - Dictionary Learning (Image Processing)
- For all the problems
  - The underlying function f is non-convex
  - Alternating Minimization was known to be very successful
  - But there were some situations where the algorithm will not succeed
- We provide certain enhancements to the basic algorithm
- Provide first theoretical analysis under certain standard assumptions

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

$\cong$  $\times$

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

- M: characterized by U, V
- No. of variables:
  - U: $m \times k = mk$
  - V: $n \times k = nk$
- DoF: $mk + nk$

Low-rank Matrix Completion

$$\min_{X} \; Error_{\Omega}(X) = \sum_{(i,j)\in\Omega} \left(X_{ij} - M_{ij}\right)^2$$
$$s.t \qquad \mathbf{rank}(X) \le k$$

- Ω: set of known entries
- Problem is NP-hard in general
  - Two approaches:
    - Relax rank function to its convex surrogate (Trace-norm based method)
    - Use alternating minimization

Existing method: Trace-norm minimization

$$\min_{X} \sum_{(i,j)\in\Omega} \left(X_{ij} - M_{ij}\right)^2$$
$$s.t. \qquad ||X||_* \leq \lambda(k)$$

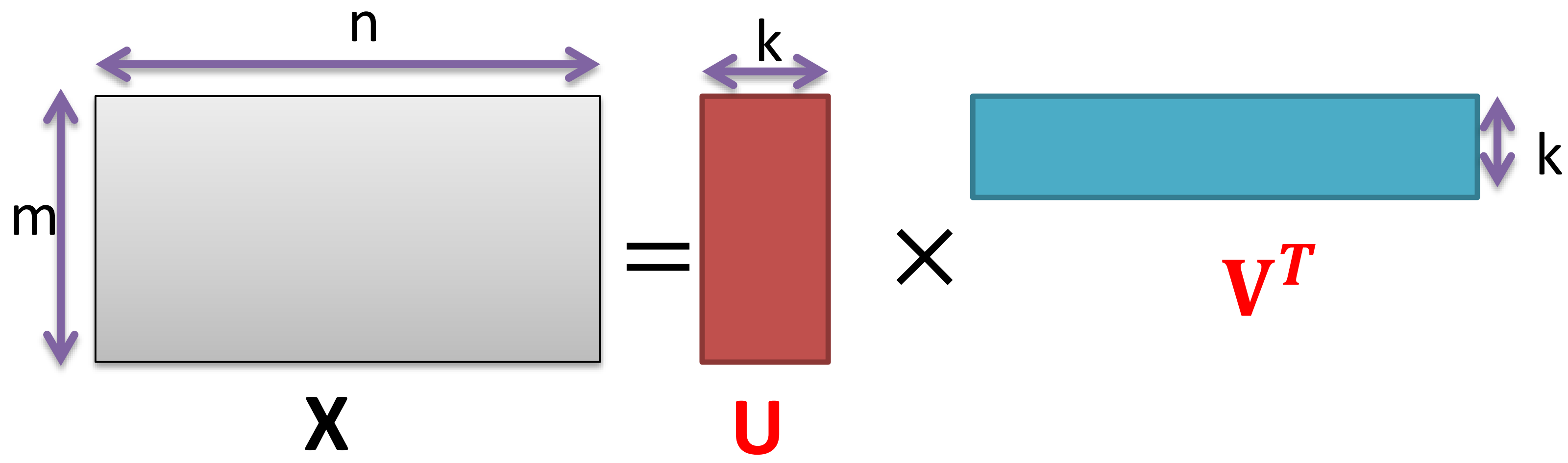
- $||X||_*$: sum of singular values
- Candes and Recht prove that above problem solves matrix completion (under assumptions on Ω and M)
- However, convex optimization methods for this problem don't scale well

Alternating Minimization

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

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

- 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($P_\Omega(M),\ k$)

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. $|\Omega| = O(k^7 n \log n \ \beta^6)$
    - $\beta = \sigma_1/\sigma_k$
  - 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

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

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

Problems:

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

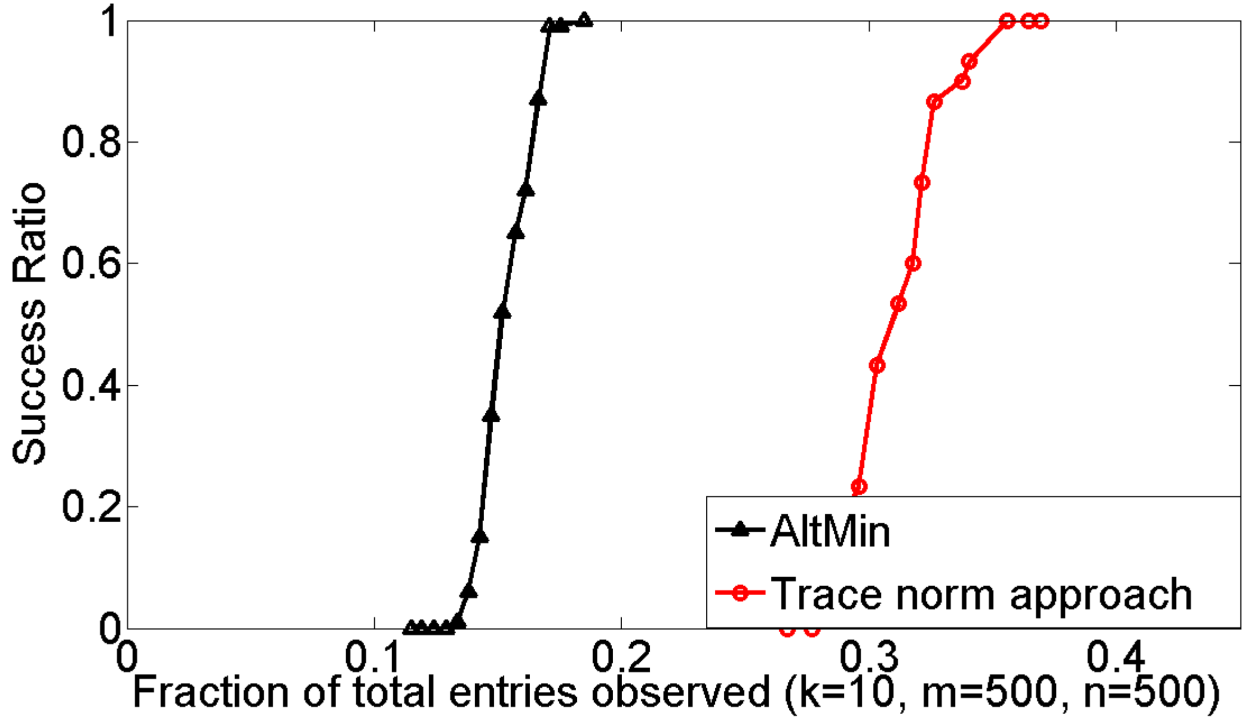
Tools:

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

Alternating Minimization	Trace-Norm Minimization
$\|\|M - UV^T\|\|_F \leq \epsilon \, \|\|M\|\|_F$ after $O(\log\left(\frac{1}{\epsilon}\right))$ steps	Requires $\mathrm{O}(\log\left(\frac{1}{\epsilon}\right))$ 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
$\|\boldsymbol{\Omega}\| = O\left(k^7 \beta^6 d \log^2(d)\right)$ $d = m + n$	$\|\boldsymbol{\Omega}\| = O\left(k \, d \log^2(d)\right)$ $d = m + n$

Empirical Performance

- Generated 100 low-rank matrix completion problems:
  - Vary fraction of total entries observed
  - Success: $||M - X|| \leq .1\ ||M||$



- Variants of alternating minimization form important component of the winning entry for Netflix Challenge

Comparison to Keshavan'12

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

$$|\Omega| = O(k\beta^8(m+n)\log(m+n))$$

  - Ours:

$$|\Omega| = O(k^7\beta^6(m+n)\log(m+n))$$

- Recent work of Hardt & Wooters improve bounds to:

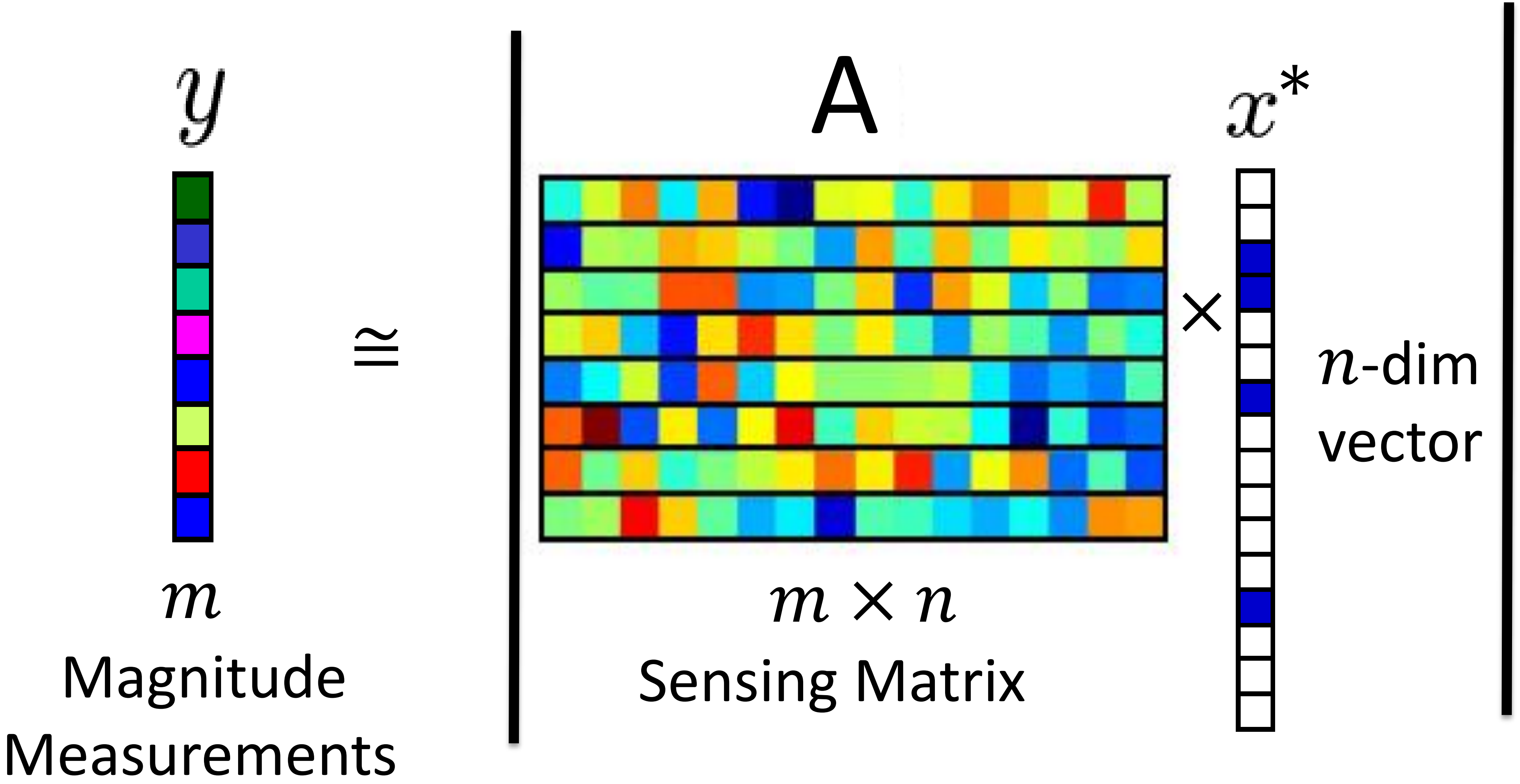
$$|\Omega| = O(poly(k)\log\beta\ (m+n)\log(m+n))$$

  - But use a modified and more complicated version of AltMin

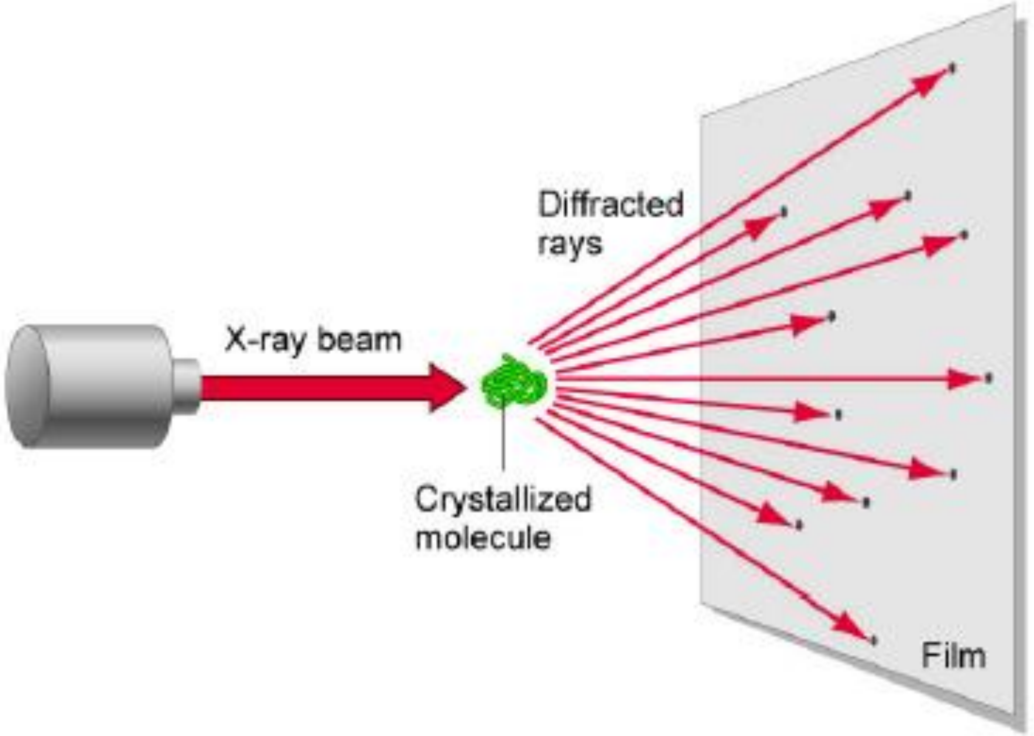
Recap

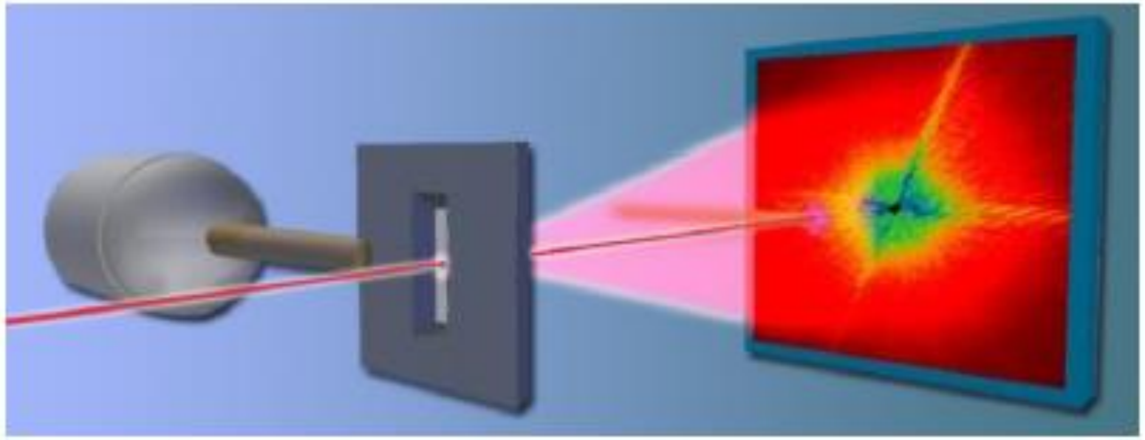
- Study Alternating Minimization method for:
  - Low-rank Matrix Completion
  - Low-rank Matrix Sensing
- The objective function in these problems is non-convex
- Provide convergence to the global optima guarantees
  - Use similar assumptions as existing methods
  - But slightly worse no. of measurements (or entries)

Phase Retrieval-Problem Definition



Motivation (X-ray Crystallography)

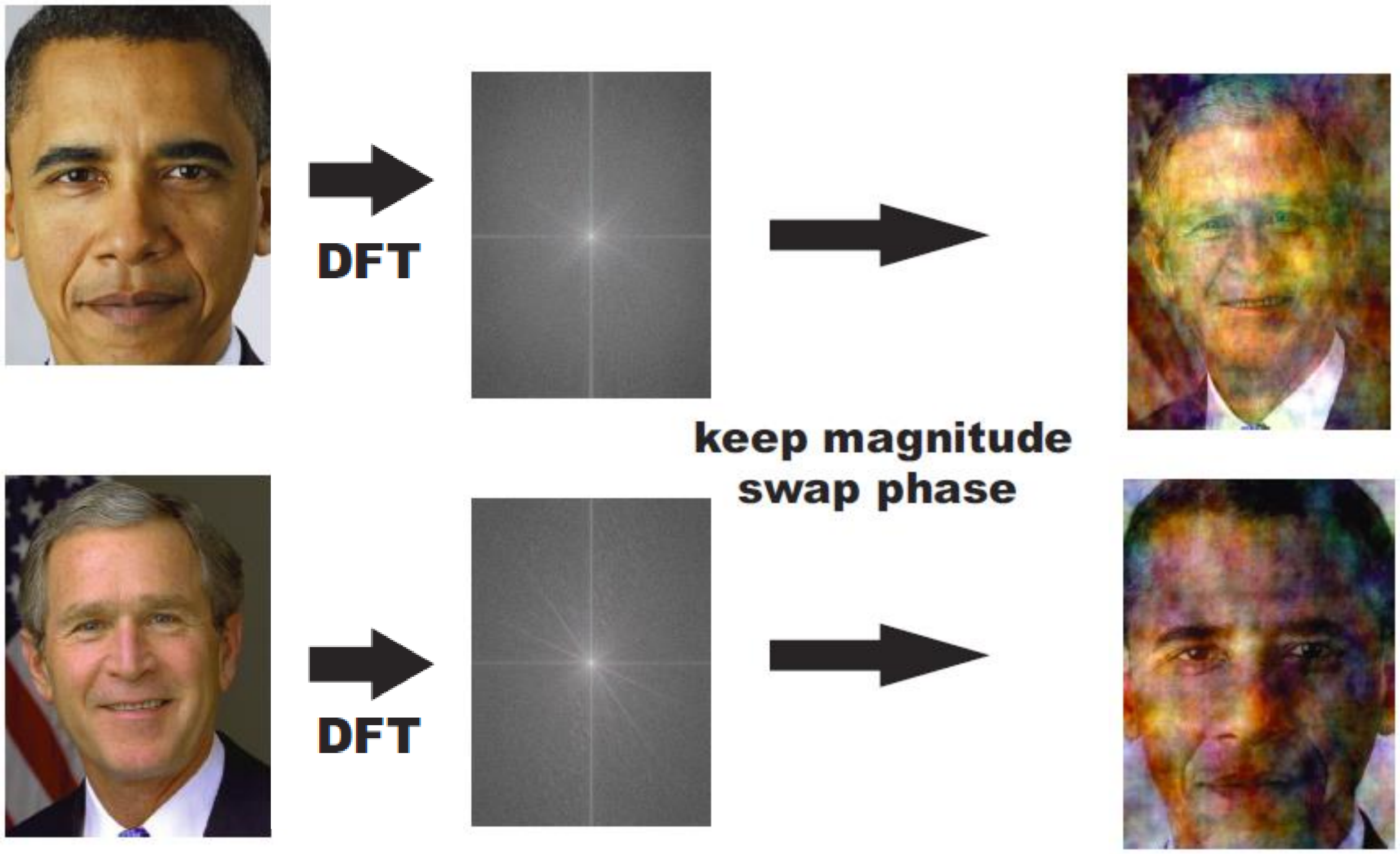




Problem: Detectors record intensities only

- Magnitude measurements only; Phase Missing

Importance of Phase



Slide from Candes'ISIT2013

Phase Retrieval

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

- Only magnitudes of measurements available
- Goal: Recovery x^* i.e., given A, y

$$Find \quad x \quad s.t. \qquad y_i = |\langle a_i, x \rangle| \; \forall i$$

$$Find \quad x \quad s.t. \qquad y_i^2 = \langle a_i a_i^T, x x^T \rangle \; \forall i$$

PhaseLift

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

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

Alternating Minimization

$$Find \quad x \quad s.t. \qquad y_i = |\langle a_i, x\rangle| \;\; \forall 1 \le i \le m$$

- Let say phase of measurements is known
  - $P_{ii}^* = Phase(\langle a_i, x^* \rangle)$
- Then the problem is: $Find \quad x \;\; s.t. \quad P^* y = Ax$
  - Simple system of linear equation
- Make P also as a variable

$$Find \quad x\,, P \quad s.t. \qquad P_{ii} \cdot y_i = \langle a_i, x\rangle \;\; \forall 1 \le i \le m$$

Alternating Minimization

- A variant was proposed by Gerchberg and Saxton in 1972
  - Random initialization
- Heavily used in practice
- However, no analysis for last 41 years
- Our contributions:
  - Better initialization
  - Provide first theoretical analysis for the algorithm
    - Results hold in “certain settings”

Our Modification

- Input: A, y
- Initialize: $x_0 = Largest\ Eigenvector\ (\sum_i y_i^2 a_i a_i^T)$
- For t=1 to T
  - $P_t = Phase(A\,x_t)$
  - $x_{t+1} = arg\min_x ||P_t y - Ax||^2$
- EndFor
- Output: x_T

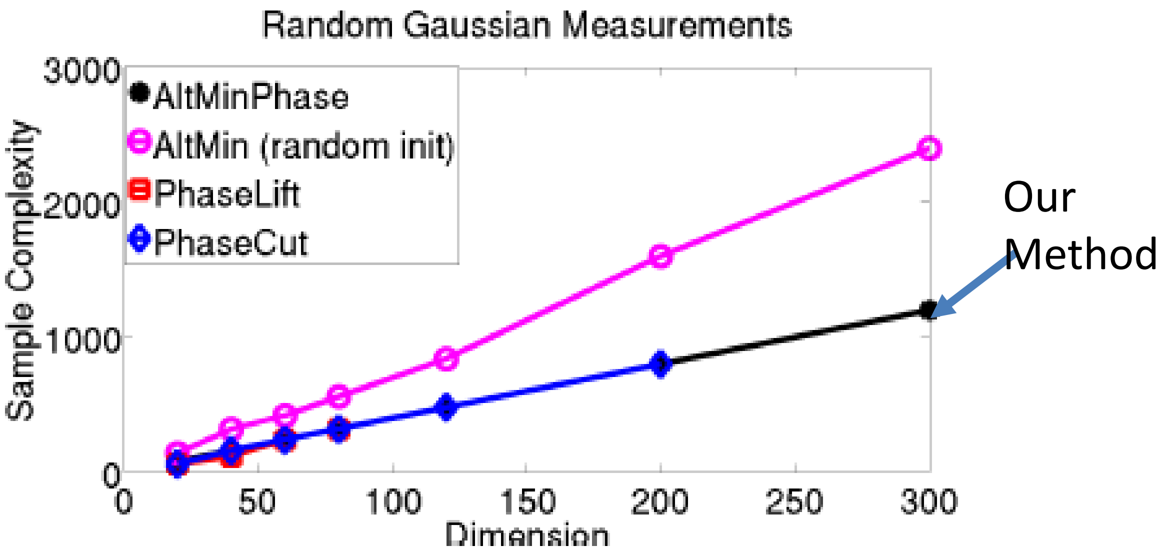
Measurement vector

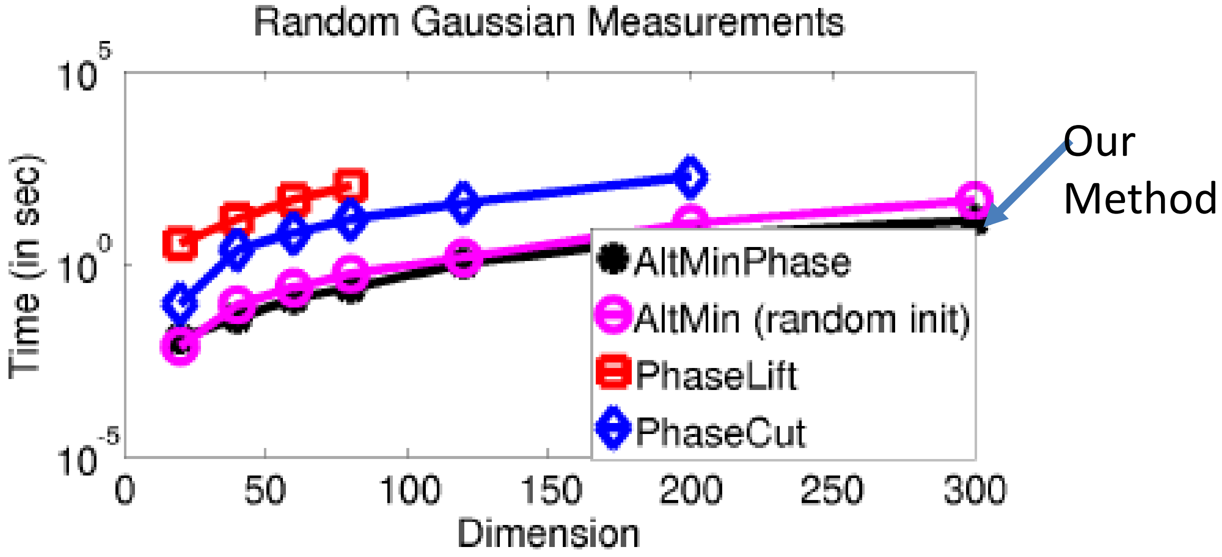
Measurements
$y_i = |\langle a_i, x\rangle|$

Our Results [JNS'13]

- Assumptions:
  - a_i: Gaussian distributed
  - $m = O(n \log^3 n/\epsilon)$
    - m: number of measurements, n: dimensionality
- Alternating minimization recovers $\hat{x}$
  - $|| \hat{x} - x^*||_2 \leq \epsilon||x^*||$
  - Number of iteration: $\log(\frac{1}{\epsilon})$
  - First analysis for alternating minimization for Phase Retrieval
- Assumptions similar to existing methods (convex relaxation based)
  - $m = O(n)$ suffices
  - Typically no. of iterations: $1/\sqrt{\epsilon}$

Empirical Results



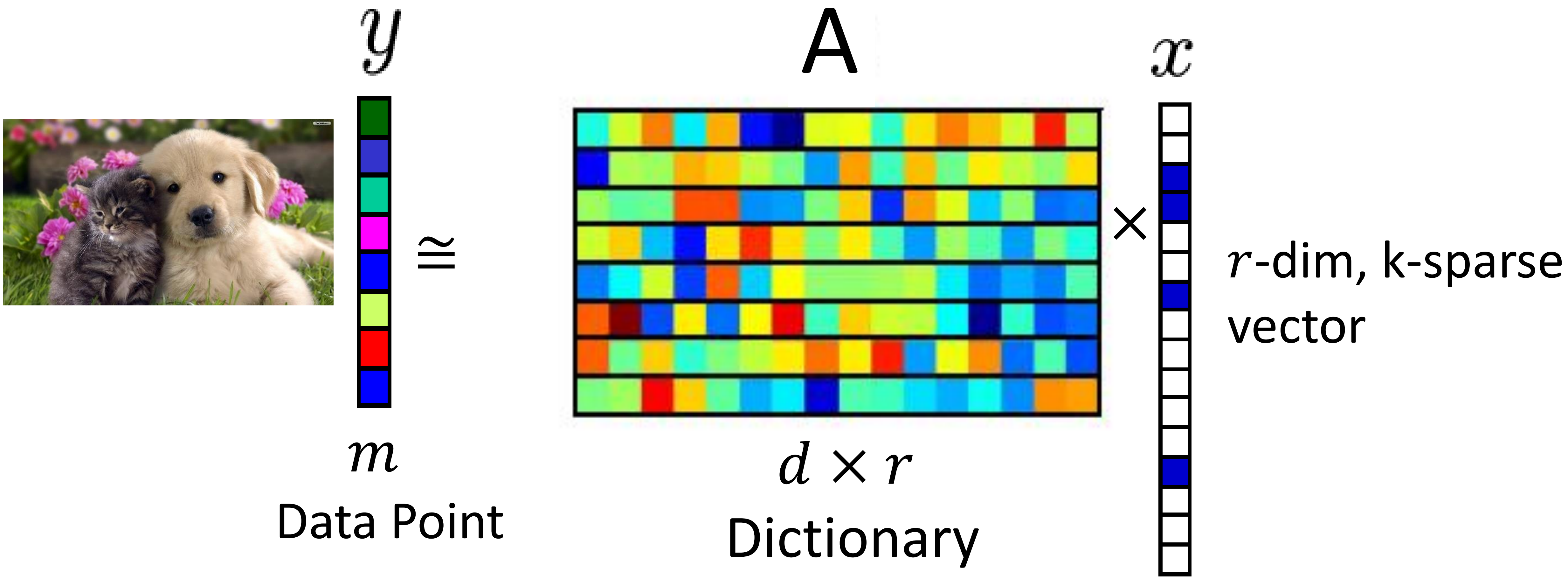


- Smaller is better

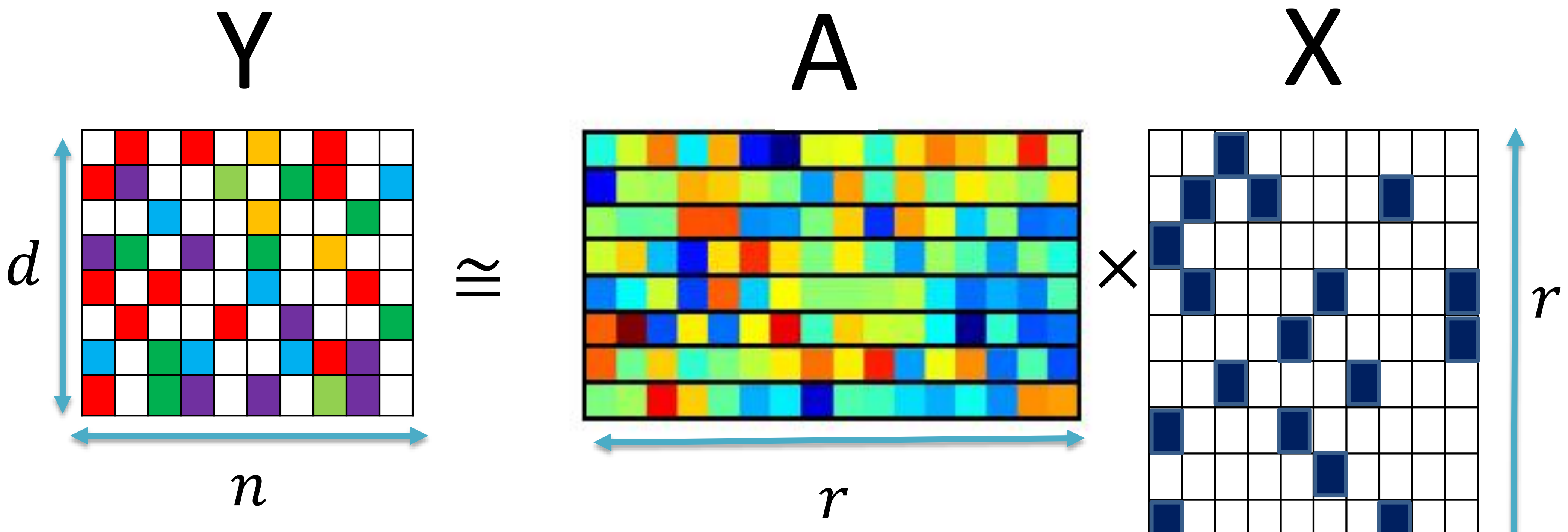
Summary

- Given:
  - Measurements:
$$y_i = |\langle a_i, x_* \rangle|, 1 \le i \le m,\ x_* \in C^n$$
  - Measurement matrix:
$$A = [a_1 a_2 \dots a_m]$$
$$a_i \sim N(0, I)$$
- Recover x_*
- Alternating minimization with proper initialization require:
$$m = O(n \log^3 \frac{n}{\epsilon})$$
- Open problem: use more realistic Fourier measurements

Dictionary Learning



Dictionary Learning



- Overcomplete dictionaries: $r \gg d$
- Goal: Given Y, compute A, X
  - Using small number of samples n

Existing Results

- Generalization error bounds [VMB'11, MPR'12, MG'13, TRS'13]
  - But assumes that the optimal solution is reached
  - Do not cover exact recovery with finite many samples
- Identifiability of A, X [HS'11]
  - Require exponentially many samples
- Exact recovery [SWW'12]
  - Restricted to square dictionary ($d = r$)
  - In practice, overcomplete dictionary ($d \ll r$) is more useful

Generating Model

- Generate dictionary A
  - Assume A to be incoherent, i.e., $\langle A_i, A_j \rangle \leq \mu/\sqrt{d}$
  - $r \gg d$
- Generate random samples $X = [x_1, x_2, \ldots, x_n] \in R^{d \times n}$
  - Each x_i is k-sparse
- Generate observations: $Y = AX$

Algorithm

- Typically practical algorithm: alternating minimization
  - $X_{t+1} = argmin_X ||Y - A_t X||_F^2$
  - $A_{t+1} = argmin_A ||Y - A X_{t+1}||_F^2$
- Initialize A_0
  - Using clustering+SVD method of [AAN'13] or [AGM'13]

Results [AAJNT'13]

- Assumptions:
  - A is $\mu -$ incoherent ($\langle A_i, A_j \rangle \leq \mu / \sqrt{d}, ||A_i|| = 1$)
  - $1 \leq \left|X_{ij}\right| \leq 100$
  - Sparsity: $k \leq \frac{d^{\frac{1}{6}}}{\mu^{\frac{1}{3}}}$ (better result by AGM'13)
  - $n \geq O(r^2 \log r)$
- After $\log(\frac{1}{\epsilon})$-steps of AltMin:

$$||A_T^i - A^i||_2 \leq \epsilon$$

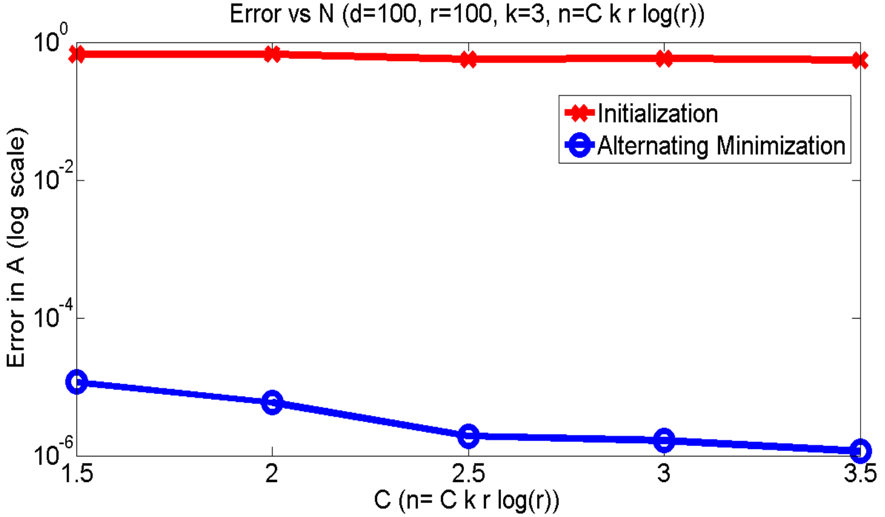
Proof Sketch

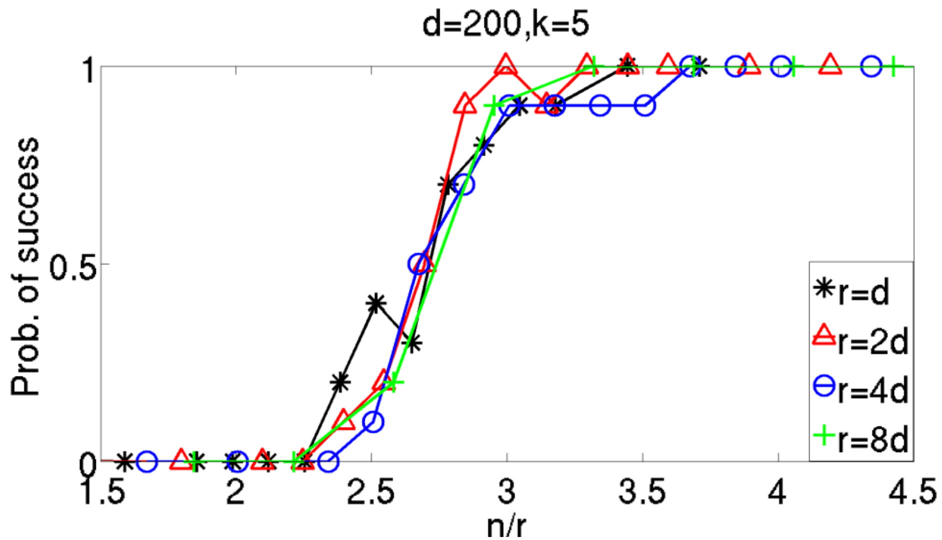
- Initialization step ensures that:

$$||A^i - A_0^i|| \leq \frac{1}{k^2}$$

- Lower bound on each element of X_{ij} + above bound:
  - $supp(x_i)$ is recovered exactly
  - Robustness of compressive sensing!
- A_{t+1} can be expressed exactly as:
  - $A_{t+1} = A + Error_{(A_t, X_t)}$
  - Use randomness in $supp(X_t)$

Simulations





Emirically: $n = O(r)$

Known result: $n = O(r^2 \log r)$

Summary

- Studied three problems
  - Low-rank matrix estimation
    - Recommendation systems, matrix sensing
  - Phase Retrieval
    - Important problem in x-ray crystallography; several other applications
  - Dictionary Learning
- Alternating Minimization
  - Empirically successful
  - Rigorous analysis was unknown
- Our contribution
  - Good initialization
  - Rigorous theoretical guarantees
    - Setting similar to that of existing theoretical results

Future Work

- Low-rank MC:
  - Remove dependence on condition number
- Phase Sensing:
  - $m = O(n \log^3 n) \Rightarrow m = O(n)$?
  - Better measurement scheme?

Future Work Contd…

- Dictionary Learning:
  - Efficient solution for $k = O(d)$ (best known solution for $k = O(\sqrt{d})$)
  - Sample complexity: $n = O(r^2 \log r) \Rightarrow n = O(r \log r)$?
- Explore Alt-Min as a generalized approach for a whole class of problems
  - Tensor completion (Ongoing Project)
  - Generalized analysis of AltMin (Ongoing Project)
  - General EM-method

Thank You!!!