

Provable Alternating Minimization Methods for Non-convex Optimization

Prateek Jain
Microsoft Research

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
- Theoretical Results
- Empirical Results

Optimization over two variables

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

- Alternating Minimization:

- Fix U, optimize for V

$$V^t = \mathit{arg} \min_V f(U^t, V)$$

- Fix V, optimize for U

$$U^{t+1} = \mathit{arg} \min_U f(U, V^t)$$

- 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

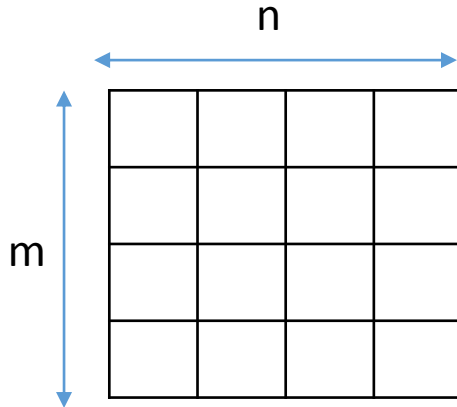
Low-rank Matrix Completion

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

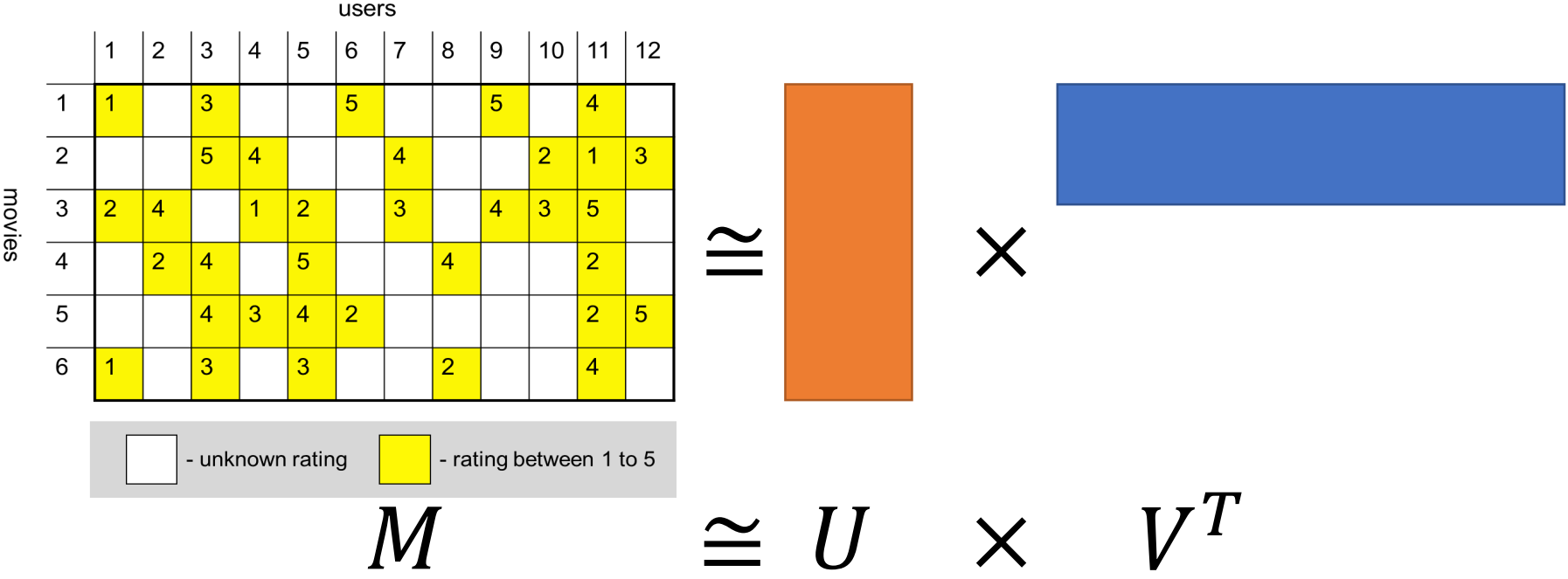
- **Task:** Complete ratings matrix
- Applications: recommendation systems, PCA with missing entries

Low-rank Assumption



- Rank: no. of independent columns or rows
- Rank- k , $k \ll (m, n)$
- Rank: no. of latent factors in the matrix

Low-rank



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

Low-rank Matrix Completion

$$\begin{aligned} \min_X \quad & Error_{\Omega}(X) = \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2 \\ \text{s.t.} \quad & \mathbf{rank}(X) \leq k \end{aligned}$$

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

Existing method: Trace-norm minimization

$$\begin{aligned} \min_X \quad & \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2 \\ \text{s.t.} \quad & \|X\|_* \leq \lambda(k) \end{aligned}$$

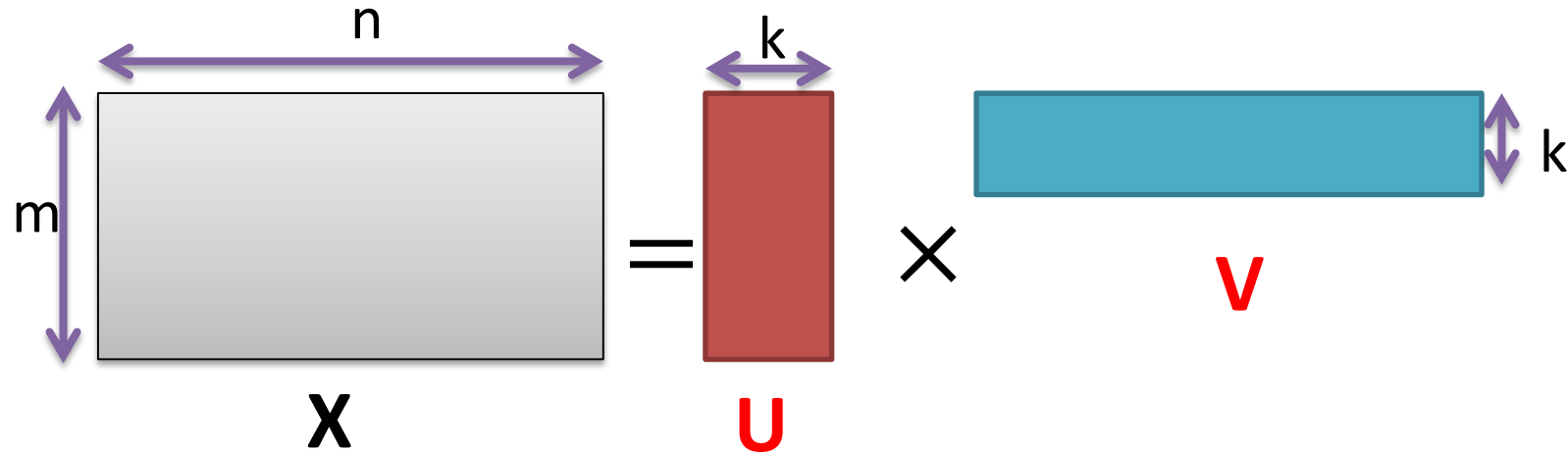
- $\|X\|_*$: sum of singular values
- Candes and Recht prove that above problem solves matrix completion (under assumptions on Ω and M)
- However, convex optimization methods for this problem don't scale well

Alternating Minimization

$$\min_X \text{Error}_\Omega(X) = \sum_{(i,j) \in \Omega} (X_{ij} - M_{ij})^2$$

s.t. $\text{rank}(X) \leq k$

- If X has rank- k :



$$V^{t+1} = \min_V \text{Error}_\Omega(U^t, V)$$

$$U^{t+1} = \min_U \text{Error}_\Omega(U, V^{t+1})$$

Our Results [JNS'13]

- Initialization:
 - $\text{SVD}(P_{\Omega}(M), k)$
- Alternating minimization is similar to power method with an “error term”
 - Power-method: a popular method to compute singular value decomposition
- 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 $O(\log(\frac{1}{\epsilon}))$ steps

Proof Sketch

Assume Rank-1 case, i.e., $M = u^* v^{*T}$

Fixing u , update for v is given by:

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

Method Term

- Observation: if $u = u^*$ then $Error = 0$

Problems:

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

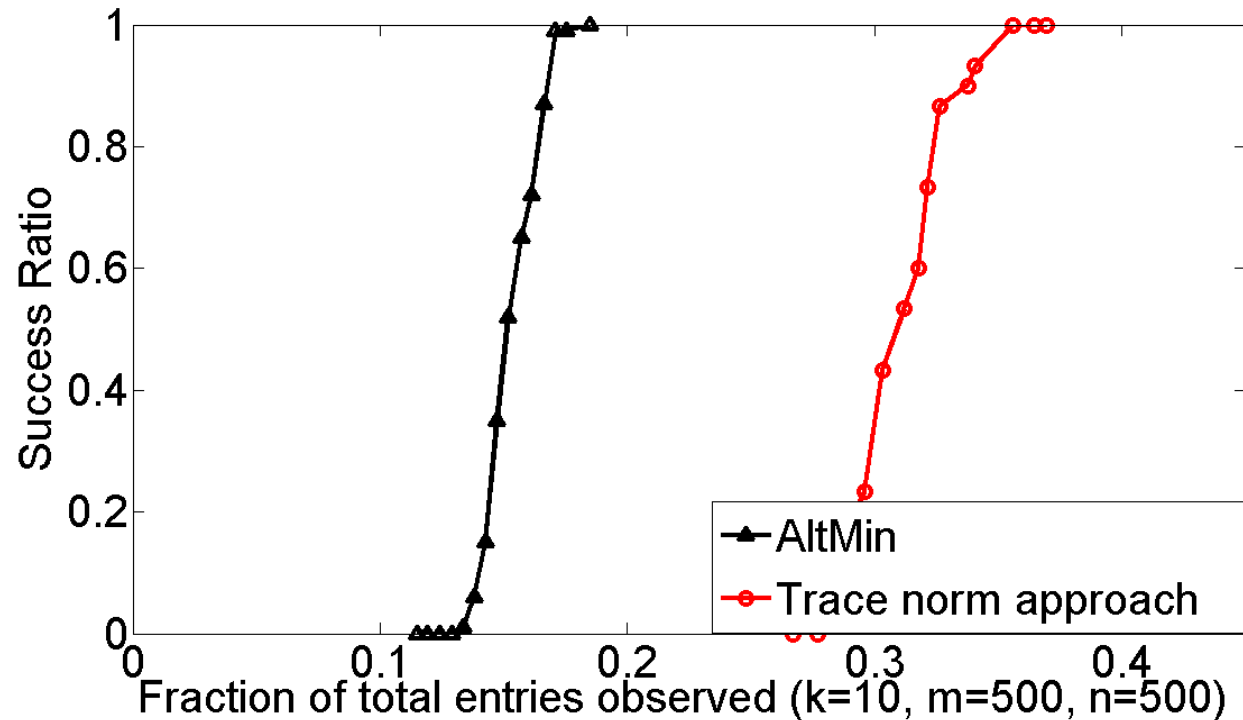
Tools:

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 $O(\log(\frac{1}{\epsilon}))$ 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 larger than k
Assumptions: random sampling and incoherence	Similar assumption
$ \Omega = O(k^7 \beta^6 d \log(d))$ $d = m + n$	$ \Omega = O(k d \log(d))$ $d = m + n$

Empirical Performance

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

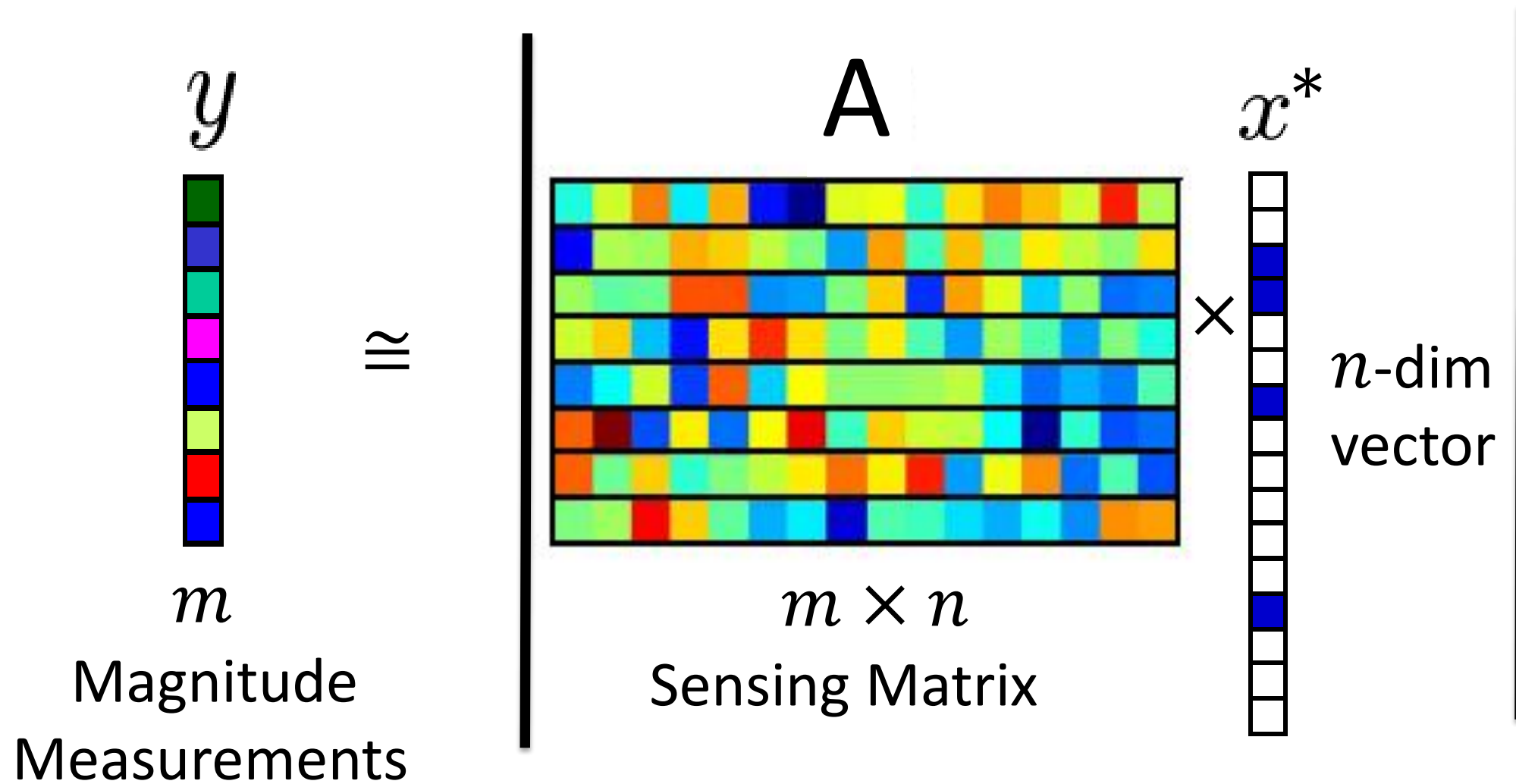


- Variants of alternating minimization form important component of the winning entry for Netflix Challenge

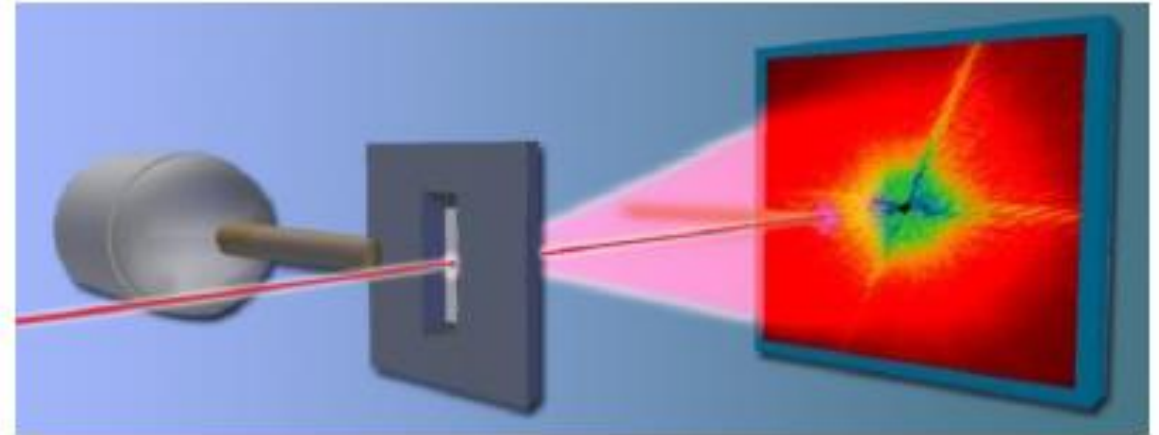
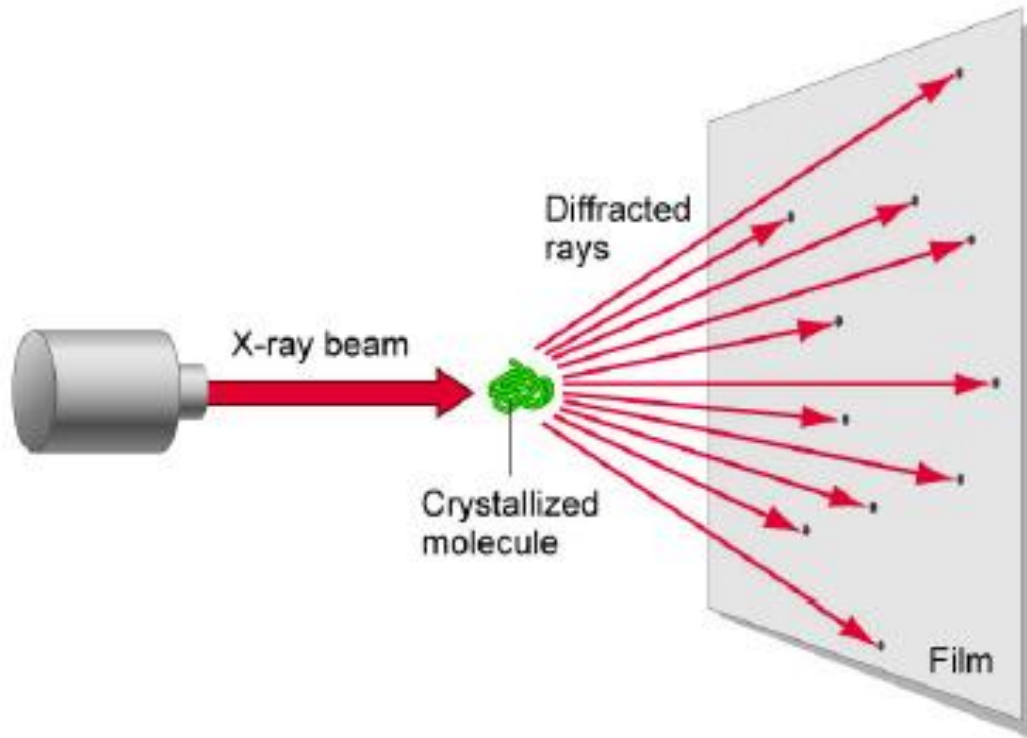
Recap

- 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

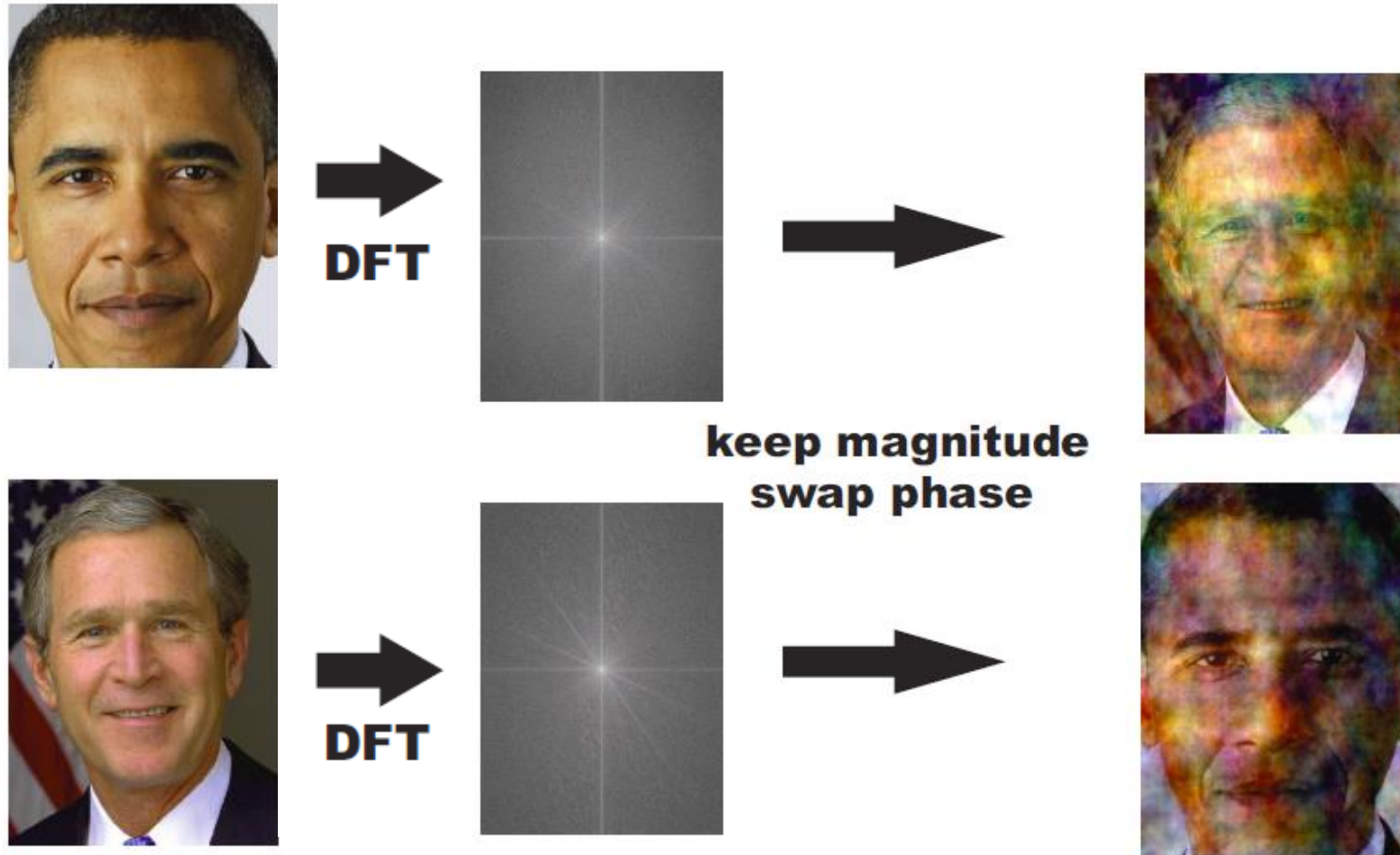


Motivation (X-ray Crystallography)



- Problem:** Detectors record intensities only
- Magnitude measurements only; Phase Missing

Importance of Phase



Phase Retrieval

$$y_i = |\langle a_i, x_* \rangle|, \quad 1 \leq i \leq m,$$
$$x_* \in \mathbb{C}^n$$

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

$$\textit{Find } x \textit{ s.t. } y_i = |\langle a_i, x \rangle| \forall i$$

$$\textit{Find } x \textit{ s.t. } y_i^2 = \langle a_i a_i^T, x x^T \rangle \forall i$$

PhaseLift

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

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

Alternating Minimization

Find x *s.t.* $y_i = |\langle a_i, x \rangle| \quad \forall 1 \leq i \leq m$

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

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

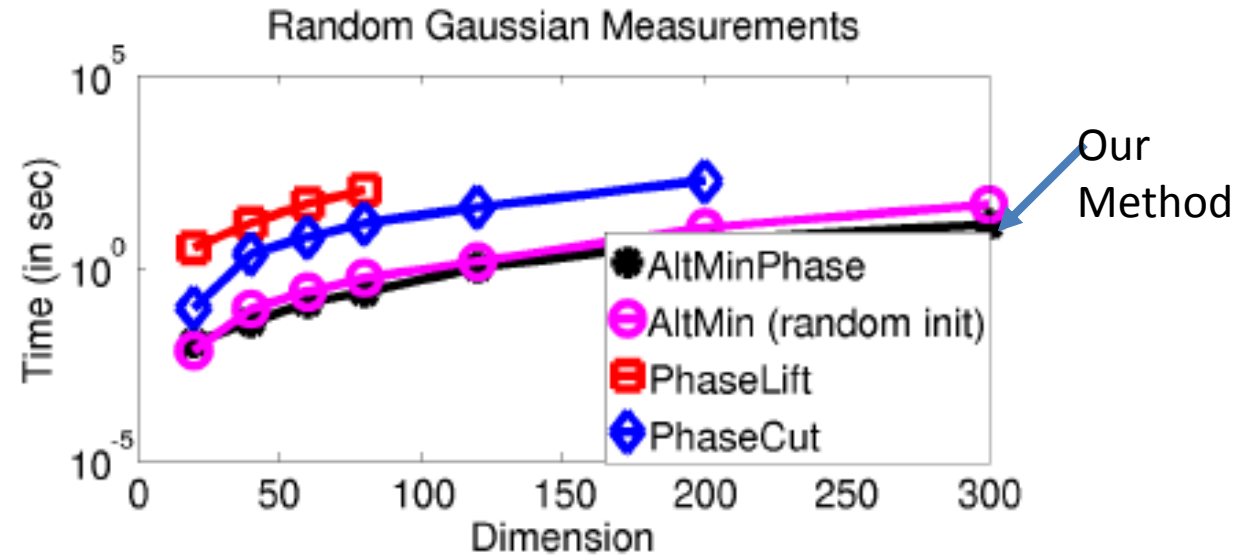
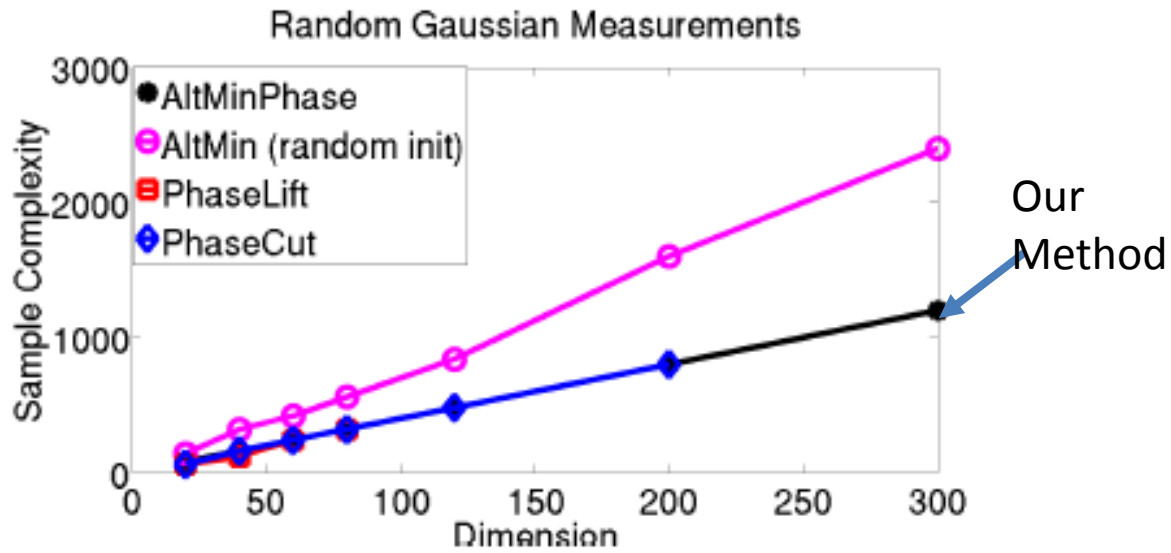
Alternating Minimization

- 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 ($x_0 = eig(\sum_i y_i a_i a_i^T, 1)$)
 - Provide first theoretical analysis for the algorithm
 - Results hold in “certain settings”

Our Results [JNS'13]

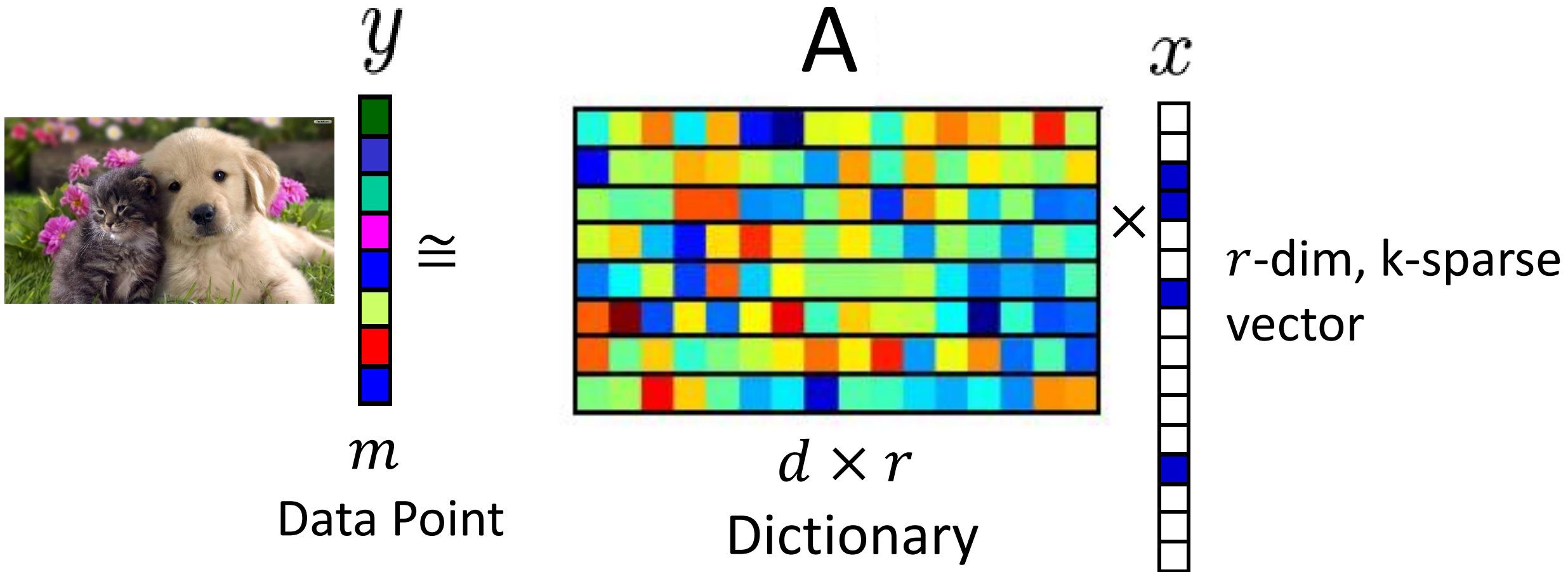
- Assumptions:
 - a_i : Gaussian distributed
 - $m = O(n \log^3 n)$
 - m : number of measurements, n : dimensionality
- Alternating minimization recovers \hat{x}
 - $\|\hat{x} - x^*\|_2 \leq \epsilon \|x^*\|$
 - Number of iteration: $\log(\frac{1}{\epsilon})$
 - First analysis for alternating minimization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

Empirical Results

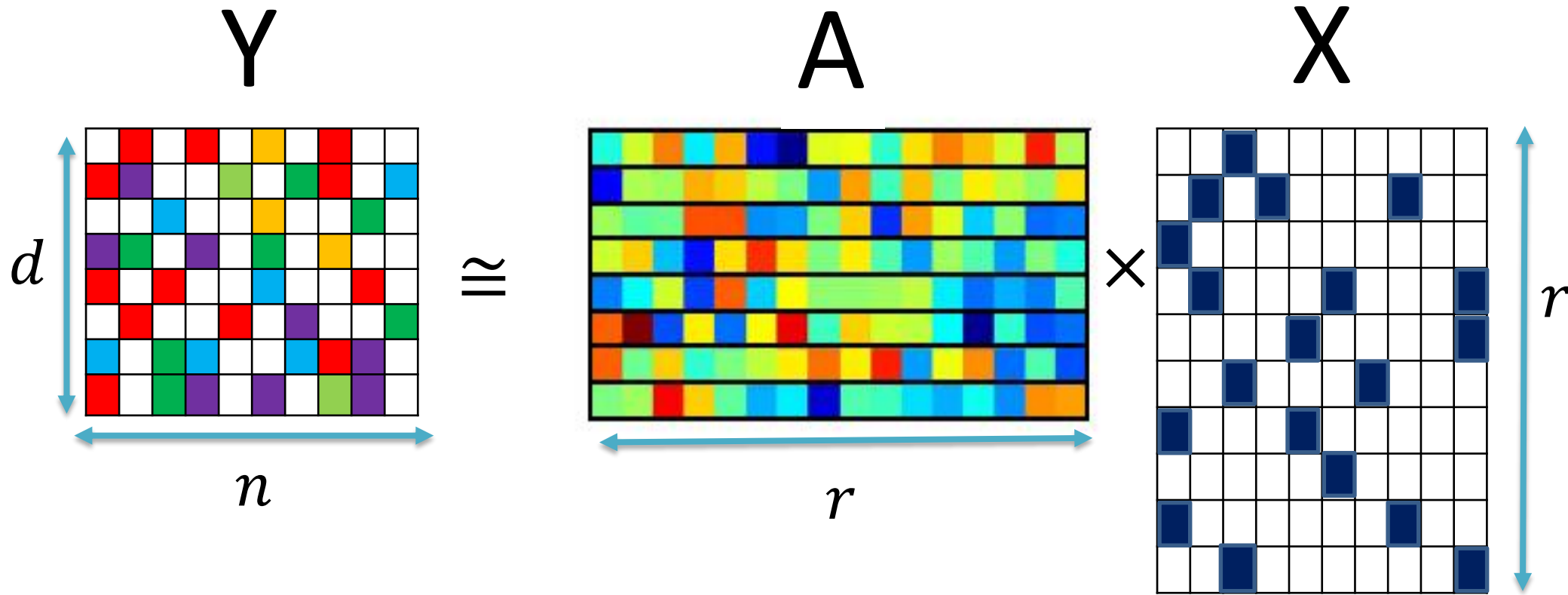


- Smaller is better

Dictionary Learning



Dictionary Learning



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

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 $X = [x_1, x_2, \dots, x_n] \in R^{d \times n}$
 - Each x_i is k -sparse
- Generate observations: $Y = AX$

Algorithm

- Existing provable methods for square dictionaries only
 - $r = d$
- Typically practical algorithm: alternating minimization
 - $X_{t+1} = \operatorname{argmin}_X \|Y - A_t X\|_F^2$
 - $A_{t+1} = \operatorname{argmin}_A \|Y - A_{t+1} X\|_F^2$
- Initialize A_0
 - Using clustering+SVD method of [AAN'13] or [AGM'13]

Results [AAJNT'13]

- Assumptions:
 - A is μ – incoherent ($\langle A_i, A_j \rangle \leq \mu/\sqrt{d}$, $\|A_i\| = 1$)
 - $1 \leq |X_{ij}| \leq 100$
 - Sparsity: $k \leq \frac{d^{\frac{1}{6}}}{\mu^{\frac{1}{3}}}$ (better result by AGM'13)
 - $n \geq O\left(\frac{r^2}{s} \log r\right)$
- After $\log\left(\frac{1}{\epsilon}\right)$ -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:
 - $\text{supp}(x_i)$ is recovered **exactly**
 - Robustness of compressive sensing!
- A_{t+1} can be expressed exactly as:
 - $A_{t+1} = A + \text{Error}_{(A_t, X_t)}$
 - Use randomness in $\text{supp}(X_t)$

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

- Improve sample complexity bound for low-rank MC
 - Remove dependence on condition number
- Dictionary Learning
 - Efficient solution for $k = O(r)$ (best known solution for $k = O(\sqrt{r})$)
- 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!!!