Provable Non-convex Optimization for ML

Prateek Jain

Microsoft Research India

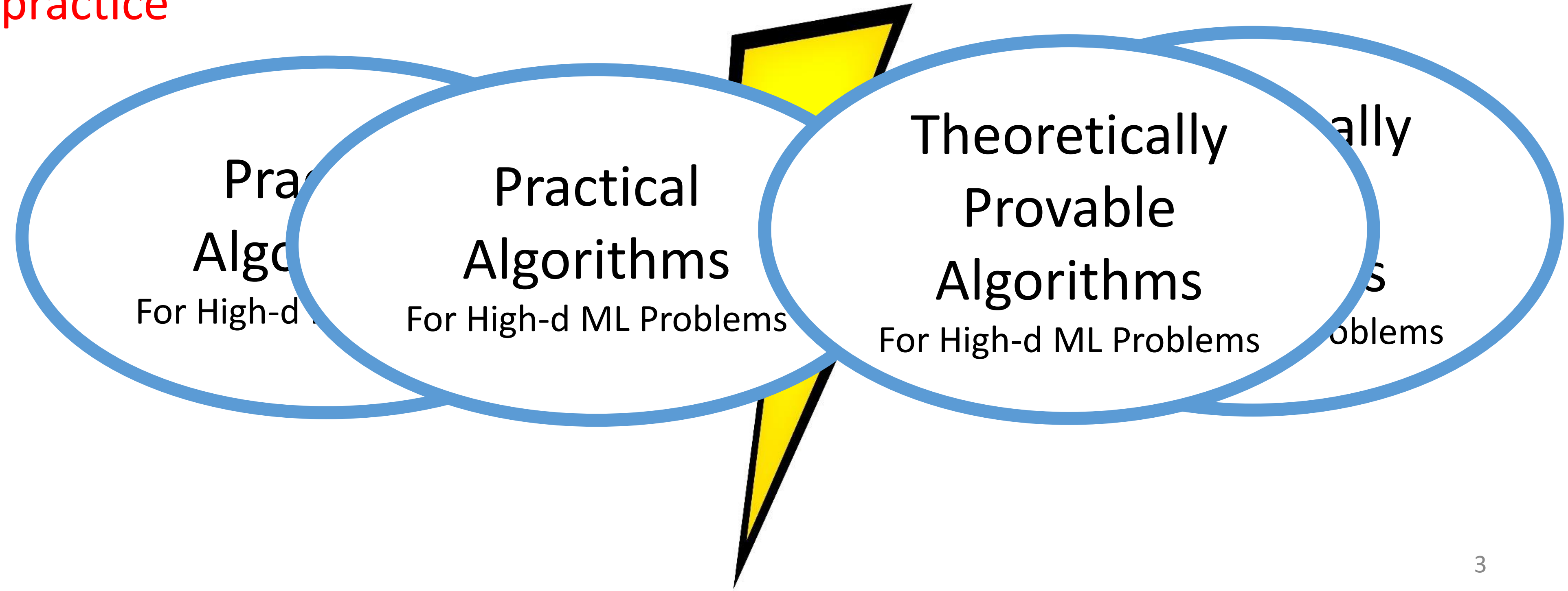
http://research.microsoft.com/en-us/people/prajain/

Overview

- High-dimensional Machine Learning
  - Many many parameters
  - Impose structural assumptions
- Requires solving non-convex optimization
  - In general NP-hard
  - No provable generic optimization tools

Overview

- Most popular approach: convex relaxation
  - Solvable in poly-time
  - Guarantees under certain assumptions
  - Slow in practice

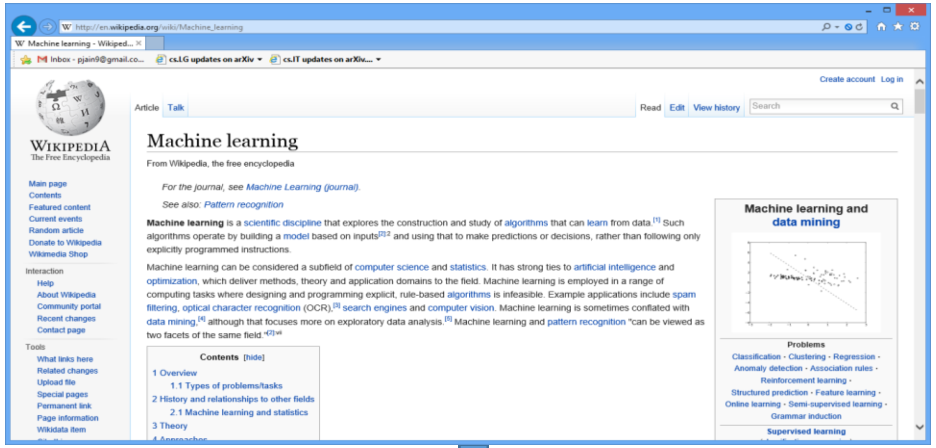


Practical Algorithms
For High-d ML Problems

Theoretically Provable Algorithms
For High-d ML Problems

Learning in Large No. of Dimensions

f(



Machine learning

From Wikipedia, the free encyclopedia

For the journal, see Machine Learning (journal).

See also: Pattern recognition

Machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data.[1] Such algorithms operate by building a model based on inputs[2]:2 and using that to make predictions or decisions, rather than following only explicitly programmed instructions.

Machine learning can be considered a subfield of computer science and statistics. It has strong ties to artificial intelligence and optimization, which deliver methods, theory and application domains to the field. Machine learning is employed in a range of computing tasks where designing and programming explicit, rule-based algorithms is infeasible. Example applications include spam filtering, optical character recognition (OCR),[3] search engines and computer vision. Machine learning is sometimes conflated with data mining,[4] although that focuses more on exploratory data analysis.[5] Machine learning and pattern recognition "can be viewed as two facets of the same field."[2]:vii

Contents [hide]

1 Overview
    1.1 Types of problems/tasks
2 History and relationships to other fields
    2.1 Machine learning and statistics
3 Theory

Machine learning and data mining

Problems

Classification · Clustering · Regression · Anomaly detection · Association rules · Reinforcement learning · Structured prediction · Feature learning · Online learning · Semi-supervised learning · Grammar induction

Supervised learning

) → {Learning, Optimization}

0	3	0	1	0	…	…	2	…	0	9

d

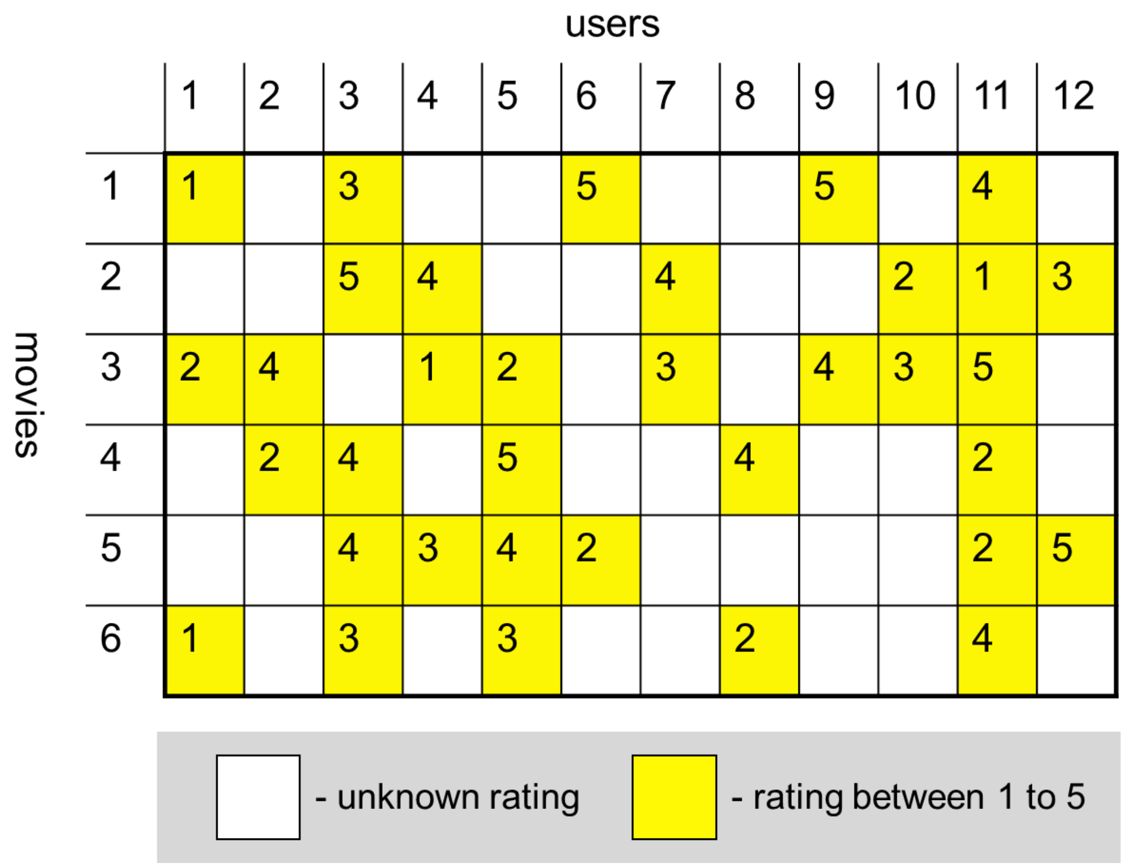
Linear Model

$$f(x) = \sum_{i} w_i x_i = \langle w, x \rangle$$

- w: d −dimensional vector
- No. of training samples: $n = O(d)$
  - For bi-grams: $n = 1000B$ documents!
- Prediction and storage: O(d)
  - Prediction time per query: 1000 secs
- Over-fitting

0	3	0	…	…	…	…	…	0	9	…	22	1

Another Example: 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
- No. of Parameters: $d_1 \times d_2$
  - $d_1 = 1M, d_2 = 10K$
  - $d_1 \times d_2 = 10B$

Key Issues

- Large no. of training samples required
- Large training time
- Large storage and prediction time

Learning with Structure

- Restrict the parameter space
- Linear classification/regression: $f(x) = \langle w, x \rangle$
    - Restrict no. of zeros in w to $s \ll d$

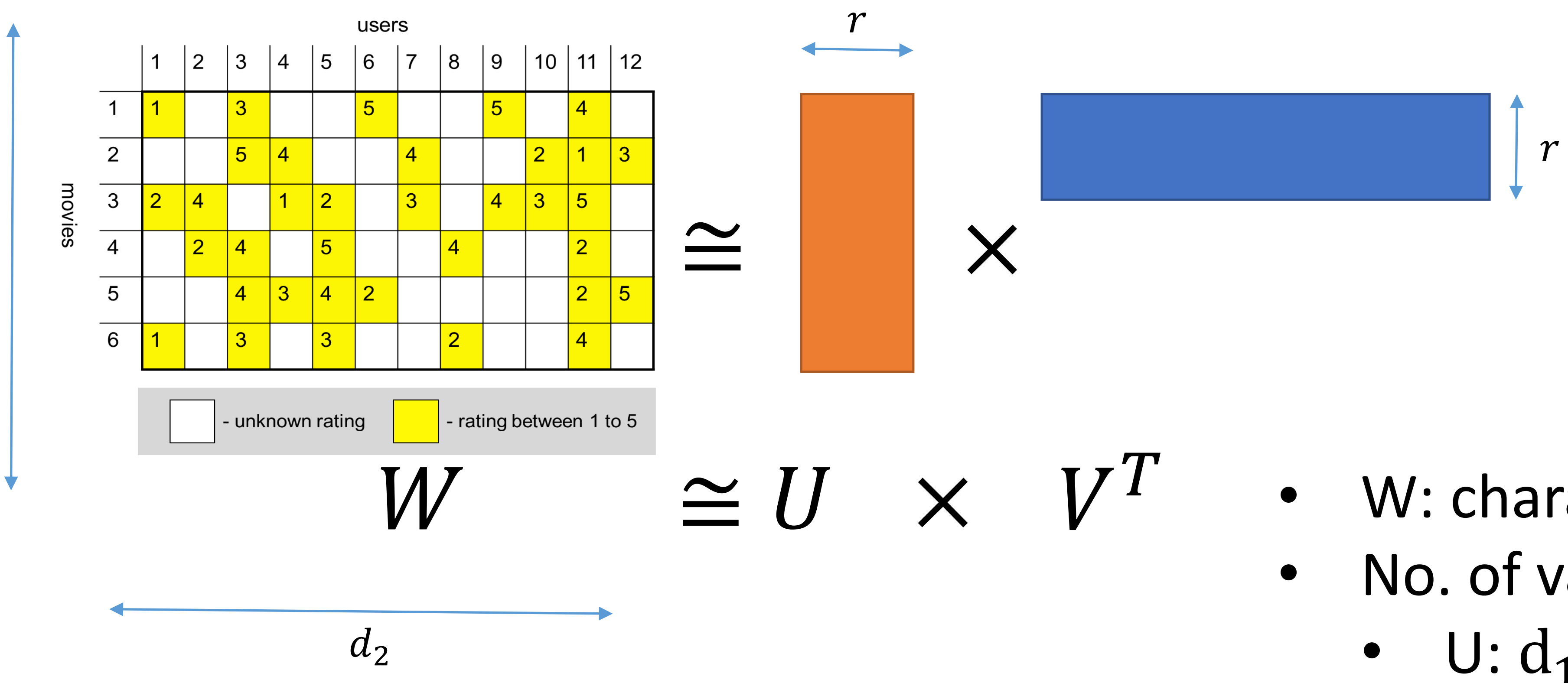
0	3	0	0	1	0	0	0	0	9

- Say $d = 1M, s = 100$

- Need to learn only $O(s \log d)$ parameters

Learning with Structure contd…

- Matrix completion:



	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

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

- W: characterized by U, V
- No. of variables:
  - U: $d_1 \times r = d_1 r$
  - V: $d_2 \times r = d_2 r$

Learning with structure

Data Fidelity Function

$$\min_{w} L(w)$$
$$s.t.\ \ w \in C$$

- Linear classification/regression
  - $C = \{w,\ ||w||_0 \leq s\}$
  - $s \log d \ll d$

- Matrix completion
  - $C = \{W, rank(W) \leq r\}$
  - $r(d_1 + d_2) \ll d_1 d_2$

- Comp. Complexity: NP-Hard
- $||w||_0$: Non-convex

- Comp. Complexity: NP-Hard
- $rank(W)$: Non-convex

Other Examples

- Low-rank Tensor completion
  - $C = \{W, \ \ tensor - rank(W) \leq r\}$
  - $r(d_1 + d_2 + d_3) \ll d_1 d_2 d_3$

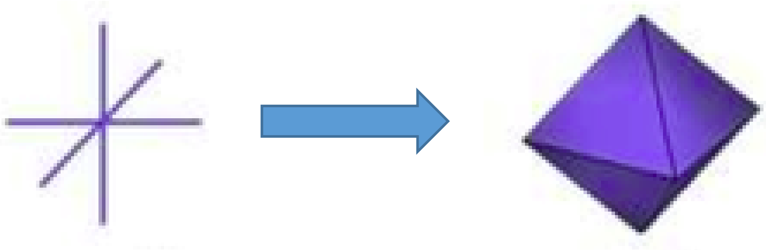
- Complexity: undecidable
- $tensor - rank(W)$: Non-convex

- Robust PCA
  - $C = \{W, W = L + S, rank(L) \leq r, ||S||_0 \leq s\}$
  - $r(d_1 + d_2) + s \log(d_1 + d_2) \ll d_1 d_2$

- Complexity: NP-Hard
- $rank(W), ||S||_0$: Non-convex

Convex Relaxations

- Linear classification/regression
  - $C = \{w, \; ||w||_0 \leq s\} \longrightarrow \tilde{C} = \{w, \quad ||w||_1 \leq \lambda(s)\}$
  - $||w||_1 \leq \sum_i w_i$



- Matrix completion
  - $C = \{W, rank(W) \leq r\} \longrightarrow \tilde{C} = \{W, \quad ||W||_* \leq \lambda(r)\}$
  - $||W||_* \leq \sum_i \sigma_i \, , W = U\Sigma V^T$

Convex Relaxations Contd…

- Low-rank Tensor completion
  - $C = \{W,\ tensor - rank(W) \leq r\} \longrightarrow \tilde{C} = \{W, ||W||_* \leq \lambda(r)\}$

- Robust PCA
  - $C = \{W, W = L + S, rank(L) \leq r, ||S||_0 \leq s\}$

$$\Big\downarrow$$

$$\tilde{C} = \{W, W = L + S, ||L||_* \leq \lambda(r), ||S||_1 \leq \lambda(s)\}$$

Convex Relaxation

- Advantage:
  - Convex optimization: Polynomial time
  - Generic tools available for optimization
  - Systematic analysis

- Disadvantage:
  - Optimizes over a much bigger set
  - Not scalable to large problems

This tutorial's focus

Don't Relax!

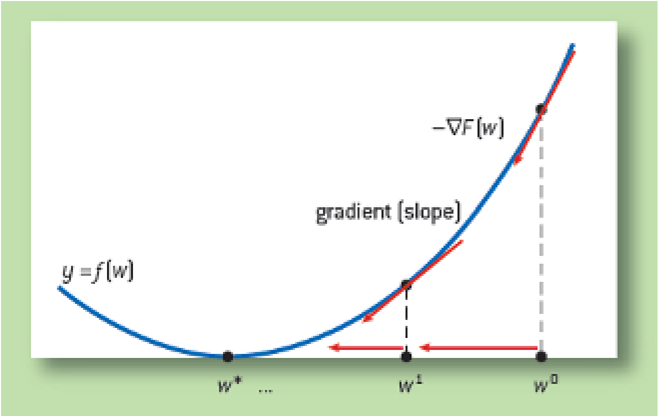
- Advantage: scalability
- Disadvantage: optimization and its analysis is much harder
  - Local minima problems

- Two approaches:
  - Projected gradient descent
  - Alternating minimization

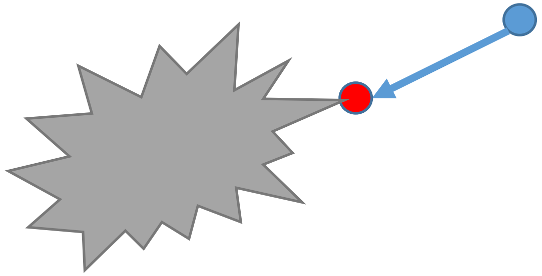
Approach 1: Projected Gradient Descent

$$\min_{w} L(w)$$
$$s.t.\ w \in C$$

- $w_{t+1} = w_t - \partial_{w_t} L(w_t)$



- $w_{t+1} = P_C(w_{t+1})$



Efficient Projection

- Sparse linear regression/classification
  - $C = \{w,\ ||w||_0 \leq s\}$
  - $supp\big(Proj_C(z)\big) = \{i_1, \dots, i_s\}$
  - $\left|z_{i_1}\right| \geq \left|z_{i_2}\right| \geq \cdots \geq |z_{i_d}|$
  - $O(d \log d)$

- Low-rank Matrix completion
  - $C = \{W, rank(W) \leq r\}$
  - SVD (top-r singular components)
  - $O(d_1 \cdot d_2 \cdot r)$

Approach 2: Alternating Minimization

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

- 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"
  - Generic technique, e.g., EM algorithms

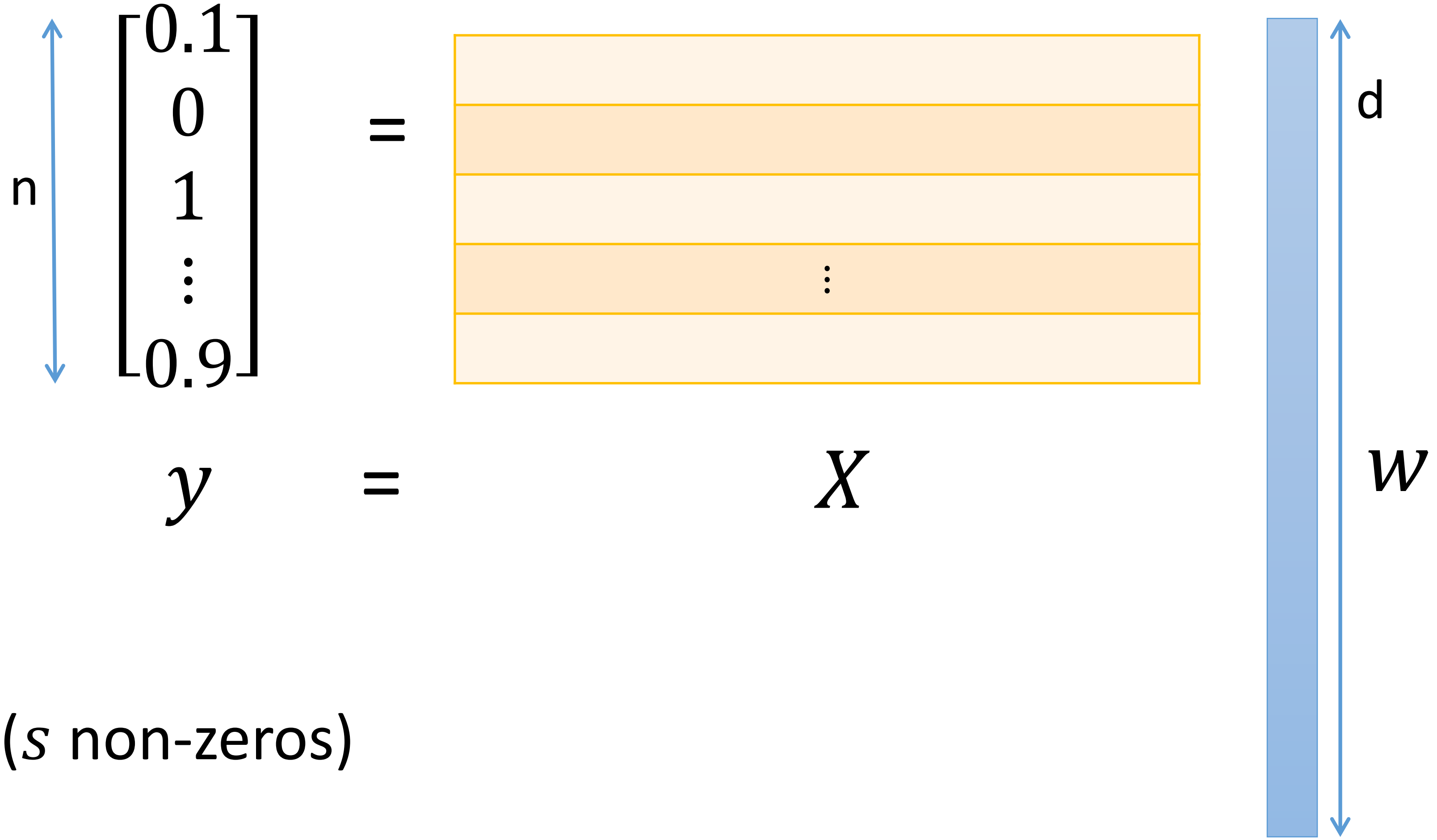
Results for Several Problems

- Sparse regression [Jain et al.'14, Garg and Khandekar'09]
  - Sparsity
- Robust Regression [Bhatia et al.'15]
  - Sparsity+output sparsity
- Dictionary Learning [Agarwal et al.'14]
  - Matrix Factorization + Sparsity
- Phase Sensing [Netrapalli et al.'13]
  - System of Quadratic Equations
- Vector-value Regression [Jain & Tewari'15]
  - Sparsity+positive definite matrix

Results Contd…

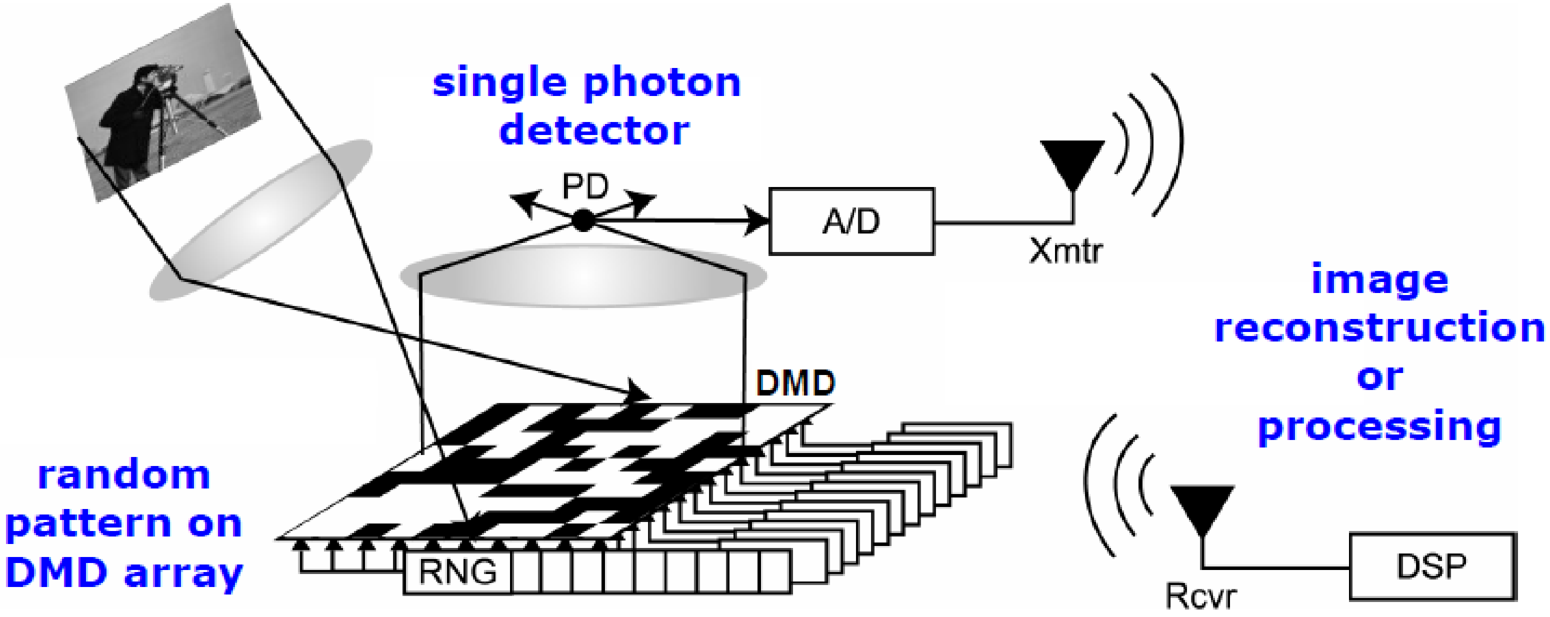
- Low-rank Matrix Regression [Jain et al.'10, Jain et al.'13]
    - Low-rank structure
- Low-rank Matrix Completion [Jain & Netrapalli'15, Jain et al.'13]
    - Low-rank structure
- Robust PCA [Netrapalli et al.'14]
    - Low-rank ∩ Sparse Matrices
- Tensor Completion [Jain and Oh'14]
    - Low-tensor rank
- Low-rank matrix approximation [Bhojanapalli et al.'15]
    - Low-rank structure

Sparse Linear Regression



- But: $n \ll d$
- w: s −sparse (s non-zeros)

Motivation: Single Pixel Camera

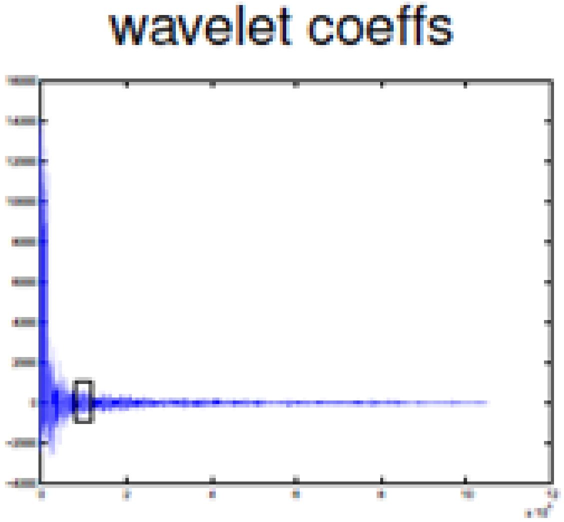


- For 1Megapixel image, 1Million measurements would be required

Picture taken from Baranuik et al.

Sparsity

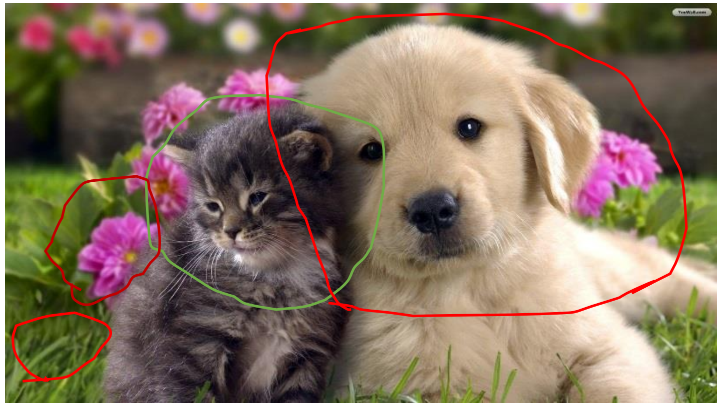
1 megapixel image



- Most images are ***sparse*** in wavelet transform space
  - Typically around 2.5% coefficients are significant

Picture taken from Baranuik et al.

Motivation: Multi-label Classification

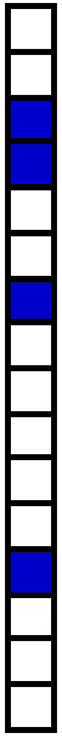


- Formulate as C 1-vs-all binary problems
  - Learn $\mathbf{w_i}, 1 \leq i \leq C$ s.t. prediction is $sign(\mathbf{w_i} \cdot \mathbf{z}\ )$
- Imagenet has 20,000 categories
- Problem: Train 20,000 SVM’s
  - Prediction time: O(20,000 · d)

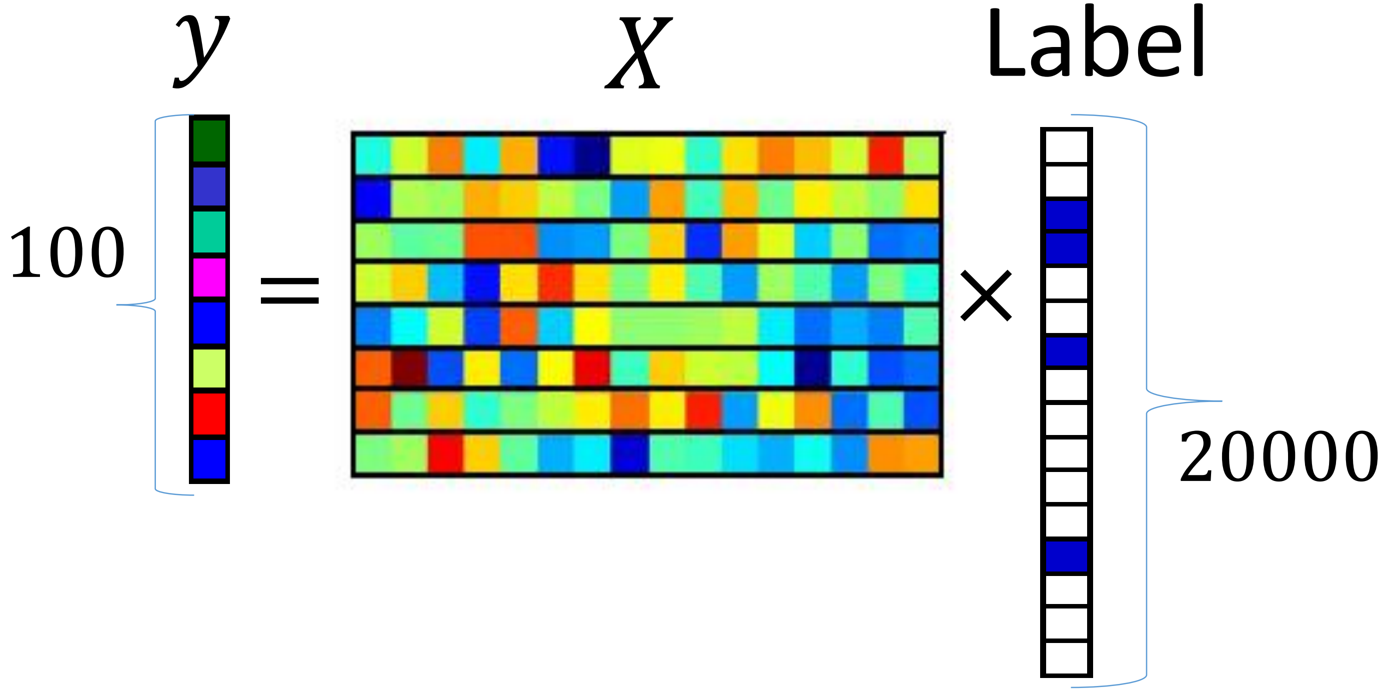
Sparsity

- Typically an image has only $5 - 10$ objects

Label



Compressive Sensing of Labels



- Learn 100 classifiers/regression functions
- Use Recovery algorithms to map back to label space
- Proposed by Hsu et al and then later pursued by several works

Sparse Linear Regression

$$\min_{w} ||y - Xw||^2$$
$$\boxed{s.t. \ \ ||w||_0 \leq s}$$

- $||y - Xw||^2 = \sum_i (y_i - \langle x_i, w \rangle)^2$
- $||w||_0$: number of non-zeros

- NP-hard problem in general ☹
    - L_0: non-convex function

Non-convexity of Low-rank manifold

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

Convex Relaxation

$$\min_{w} ||y - Xw||^2$$
$$s.t.\ \ ||w||_0 \leq s$$

- Relaxed Problem:

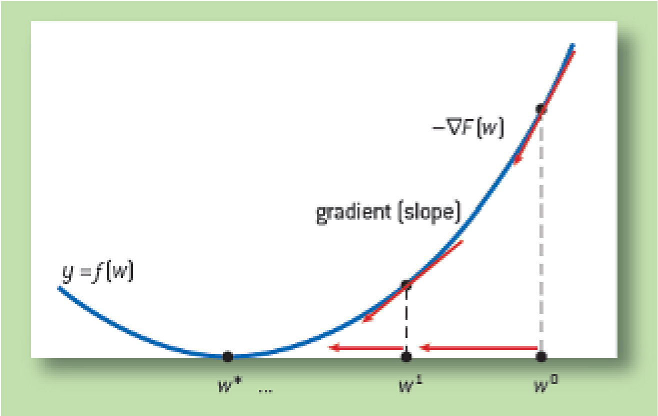
$$\min_{w} ||y - Xw||^2$$
$$s.t.\ \ ||w||_1 \ \leq s$$

- $||w||_1 = \sum_i |w_i|$
  - Known to promote sparsity
- Pros: a) Principled approach, b) Captures correlations between features
- Cons: Slow to optimize

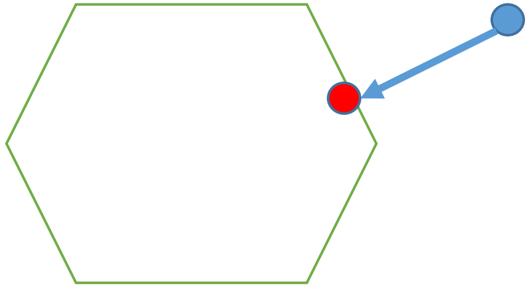
Our Approach : Projected Gradient Descent

$$\min_{w} f(w) = ||y - Xw||^2$$
$$s.t. \;\; ||w||_0 \leq s$$

- $w_{t+1} = w_t - \partial_{w_t} f(w_t)$



- $w_{t+1} = P_s(w_{t+1})$



[Jain, Tewari, Kar’2014]

Projection onto L_0 ball?

$$\min_x ||x - z||_2^2$$
$$s.t.\ ||x||_0 \leq s$$

Important Properties

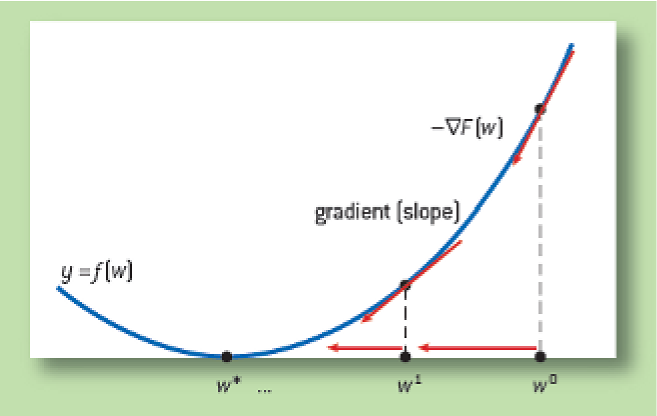
A Stronger Result?

$$||P_s(z) - z||_2^2 \leq \frac{d - s}{d - s^*} ||P_{s^*}(z) - z||_2^2$$

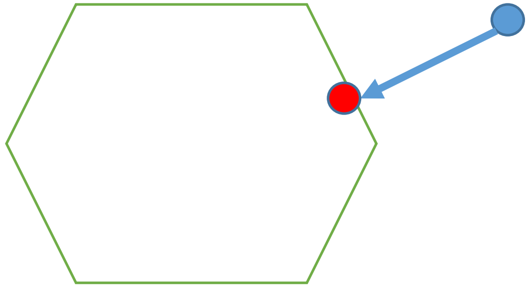
Our Approach : Projected Gradient Descent

$$\min_{w} f(w) = ||y - Xw||^2$$
$$s.t.\ \ ||w||_0 \leq s$$

- $w_{t+1} = w_t - \partial_{w_t} f(w_t)$



- $w_{t+1} = P_s(w_{t+1})$



[Jain, Tewari, Kar’2014]

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!!!

Convex-Projected Gradient Descent Proof?

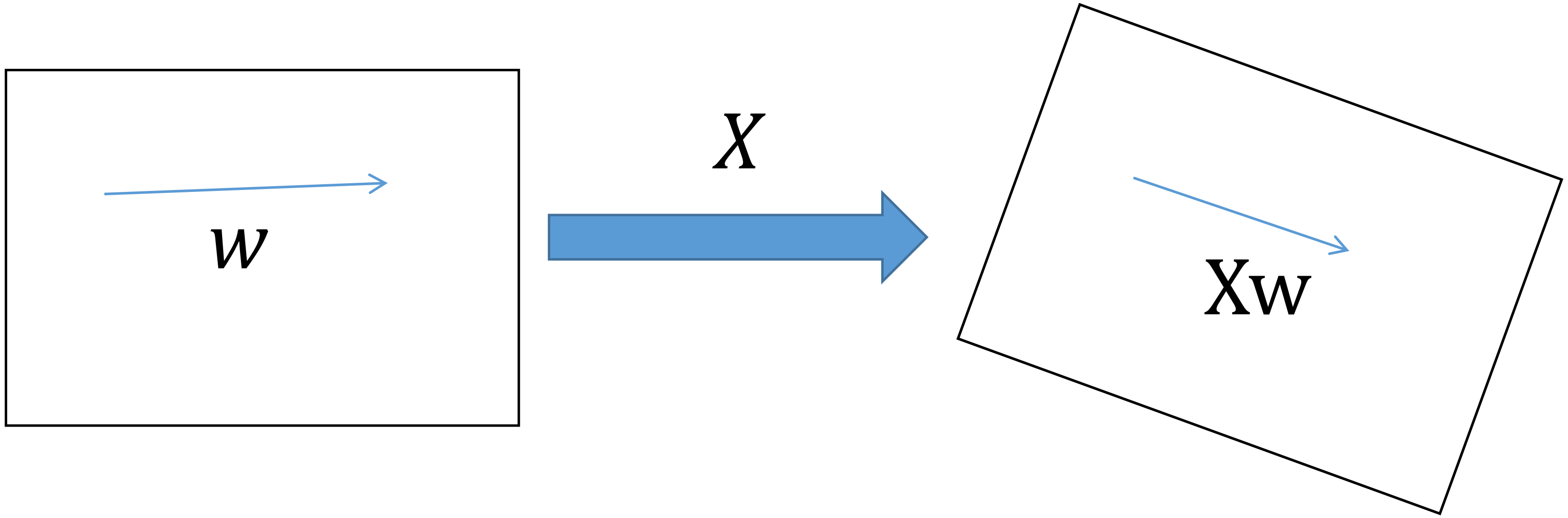
- Let $f(w) = ||X(w - w^*)||_2^2$
- Let $\alpha \cdot I_{d\times d} \preccurlyeq X^T X \preccurlyeq L \cdot I_{d\times d}$
- Let $w_{t+1} = P_C(w_t - \eta\, g_t), \quad g_t = X^T X(w_t - w^*), \;\; \eta = \frac{1}{L}$
- C: convex set and $w^* \in C$

$$||w_{t+1} - w^*|| \leq \left(1 - \frac{\alpha}{L}\right) ||w_t - w^*||$$

Restricted Isometry Property (RIP)

- X satisfies RIP if, for all **sparse** vectors Φ acts as an Isometry
- Formally: For all s-sparse $\boldsymbol{w}$

$$(1-\delta_s)||\mathbf{w}||^2 \leq ||\mathbf{Xw}||^2 \leq (1+\delta_s)||\mathbf{w}||^2$$



Proof under RIP

- Let $f(w) = ||X(w - w^*)||_2^2$
- Let $\delta_{3s} \leq \frac{1}{2}$
- Let $w_{t+1} = P_C(w_t - \eta\, g_t), \quad g_t = X^T X(w_t - w^*), \;\; \eta = 1$
- C: L_0 ball with s non-zeros and $w^* \in C$

$$||w_{t+1} - w^*|| \leq \frac{3}{4} ||w_t - w^*||$$

[Blumensath & Davies’09, Garg & Khandekar’09]

Variations

- Fully corrective version:

$$u_{t+1} = P_C(w_t - \eta\, g_t)$$
$$w_{t+1} = \arg\min_{w} f(w)\,, \qquad s.t.\quad \mathrm{supp(w)} = \mathrm{supp(u)}$$

- Two stage algorithms:

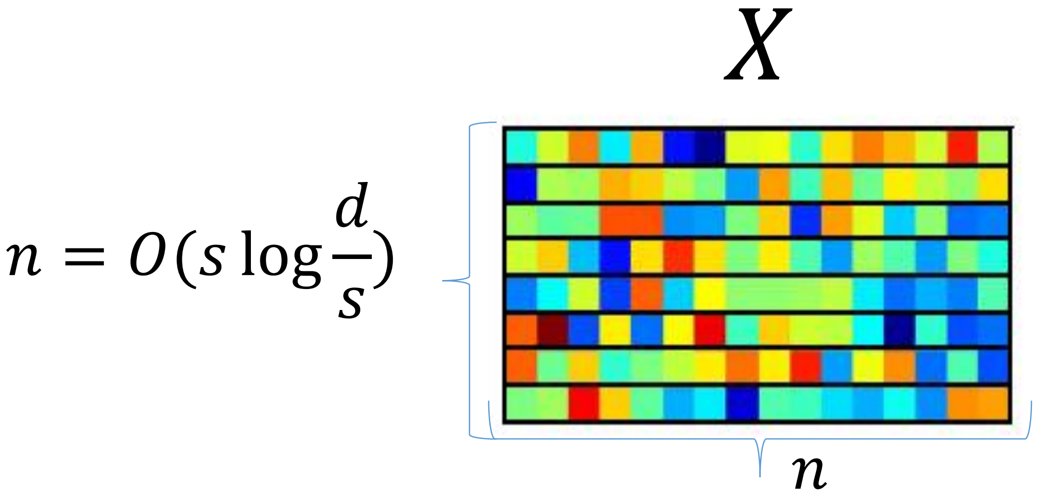
Summary so far…

- High-dimensional problems
  - $n \ll d$
- Need to impose structure on w
- Sparsity
  - Projection easy!
  - Projected Gradient works (if RIP is satisfied)
  - Several variants exist

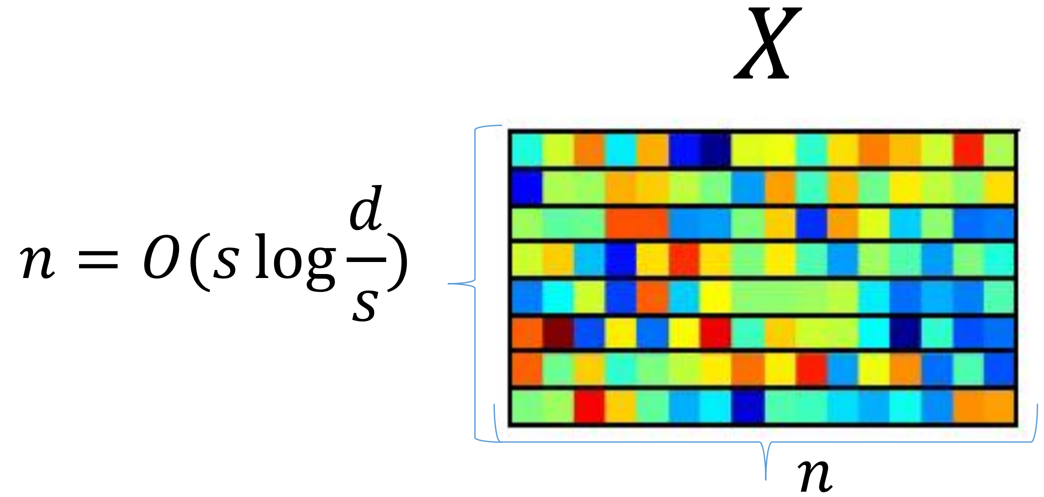
Which Matrices Satisfy RIP?

$$(1-\delta_s)||\mathbf{w}||^2 \leq ||\mathbf{Xw}||^2 \leq (1+\delta_s)||\mathbf{w}||^2, \qquad ||w||_0 \leq s$$

- Several ensembles of random matrices
  - Large enough m
- For example: $\mathrm{n} = O(s\log\frac{d}{s})$
  - $\mathrm{X}_{ij} \sim D$
  - D: 0-mean distribution
    - Bounded fourth moment



Popular RIP Ensembles



- Most popular examples:
  - $X_{ij} \sim N(0,1/\sqrt{m})$
  - $X_{ij} = +\frac{1}{\sqrt{m}}\left(w.p.\frac{1}{2}\right) and\ -\frac{1}{\sqrt{m}}\ (w.p.\frac{1}{2})$

Proof of RIP for Gaussian Ensemble

- $X \in R^{n \times d}$
- $X_{ij} \sim \frac{1}{\sqrt{n}} N(0,1)$
- $n \geq \left(\frac{1}{\delta_s^2}\right) s \log d$
- Then, X satisfies RIP at s-sparsity with constant δ_s

Other structures?

- Group sparsity
- Tree sparsity
- Union of subspaces (polynomially many subspaces)

- Projection easy for each one of these problems
- Gaussian matrices satisfy RIP (because union of small no. of subspaces)

General Result

- Let $f(w) = ||X(w - w^*)||_2^2$
- Let $w_{t+1} = P_C(w_t - \eta\, g_t), \quad g_t = X^T X(w_t - w^*),\ \ \eta = \frac{1}{(1+\delta_{3s})}$
- C: Any non-convex set and $w^* \in C$

$$(1 - \delta_s)||\mathbf{w}||^2 \leq ||\mathbf{Xw}||^2 \leq (1 + \delta_s)||\mathbf{w}||^2, \qquad w \in C$$

$$||w_{t+1} - w^*|| \leq \frac{3}{4}||w_t - w^*||$$

Proof?

But what if RIP is not possible?

Statistical Guarantees

$$y_i = \langle x_i, w^* \rangle + \eta_i$$

- $x_i \sim N(0, \Sigma)$
- $\eta_i \sim N(0, \sigma^2)$
- $w^*: s$ −sparse

$$|| \hat{w} - w^* || \leq \frac{\sigma \cdot \kappa \cdot \sqrt{s \log d}}{\sqrt{n}}$$

- $\kappa = \lambda_1(\Sigma)/\lambda_d(\Sigma)$

[Jain, Tewari, Kar'2014]

Proof?

- $f(w) = \frac{1}{2}||X(w - w^*)||^2$
- $X = [x_1; x_2; \ldots; x_n]$
- $x_i \sim N(0, \Sigma), \alpha \cdot I_{d\times d} \preccurlyeq \Sigma \preccurlyeq L \cdot I_{d\times d}$
- $w_{t+1} = P_s(w_t - \eta\, g_t), L = \frac{2}{3L}$
- $s = \left(\frac{L}{\alpha}\right)^2 s^*$

$$||w_{t+1} - w^*||_2^2 \leq \left(1 - \frac{\alpha}{10 \cdot L}\right) ||w_t - w^*||_2^2$$

Proof?

General Result for Any Function

- $f: R^d \to R$
- f: satisfies RSC/RSS, i.e.,

$$\alpha_s \cdot I_{d\times d} \preccurlyeq H(w) \preccurlyeq L_s \cdot I_{d\times d}, \qquad if, ||w||_0 \leq s$$

- IHT and several similar algorithm guarantee:

$$f(w_T) \leq f(w^*) + \epsilon$$

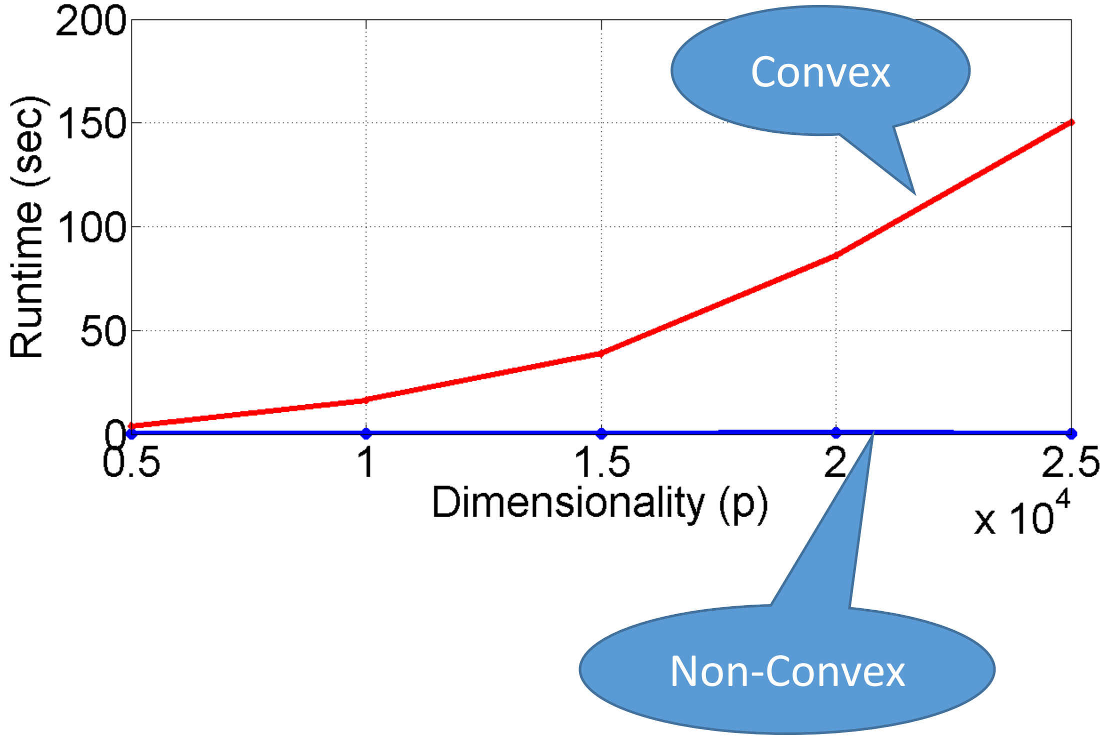
After $T = O(\frac{\log\left(\frac{f(w^0)}{\epsilon}\right)}{\log(1-\frac{L_s}{\alpha_s})})$ steps

- If $||w^*|| \leq s^*$ and $s \geq 10\frac{L_s^2}{\alpha_s^2}s^*$

[Jain, Tewari, Kar’2014]

Theory and Practice

$$y_i = \langle x_i, w^* \rangle + \eta_i$$



- $x_i \sim N(0, \Sigma), \eta_i \sim N(0, \sigma^2)$
- $w^*: s$ −sparse
- Number of iterations: $\log(\frac{1}{\epsilon})$

$$|| \widehat{w} - w^* || \leq \epsilon + \frac{\sigma\kappa\sqrt{s \log d}}{\sqrt{n}}$$

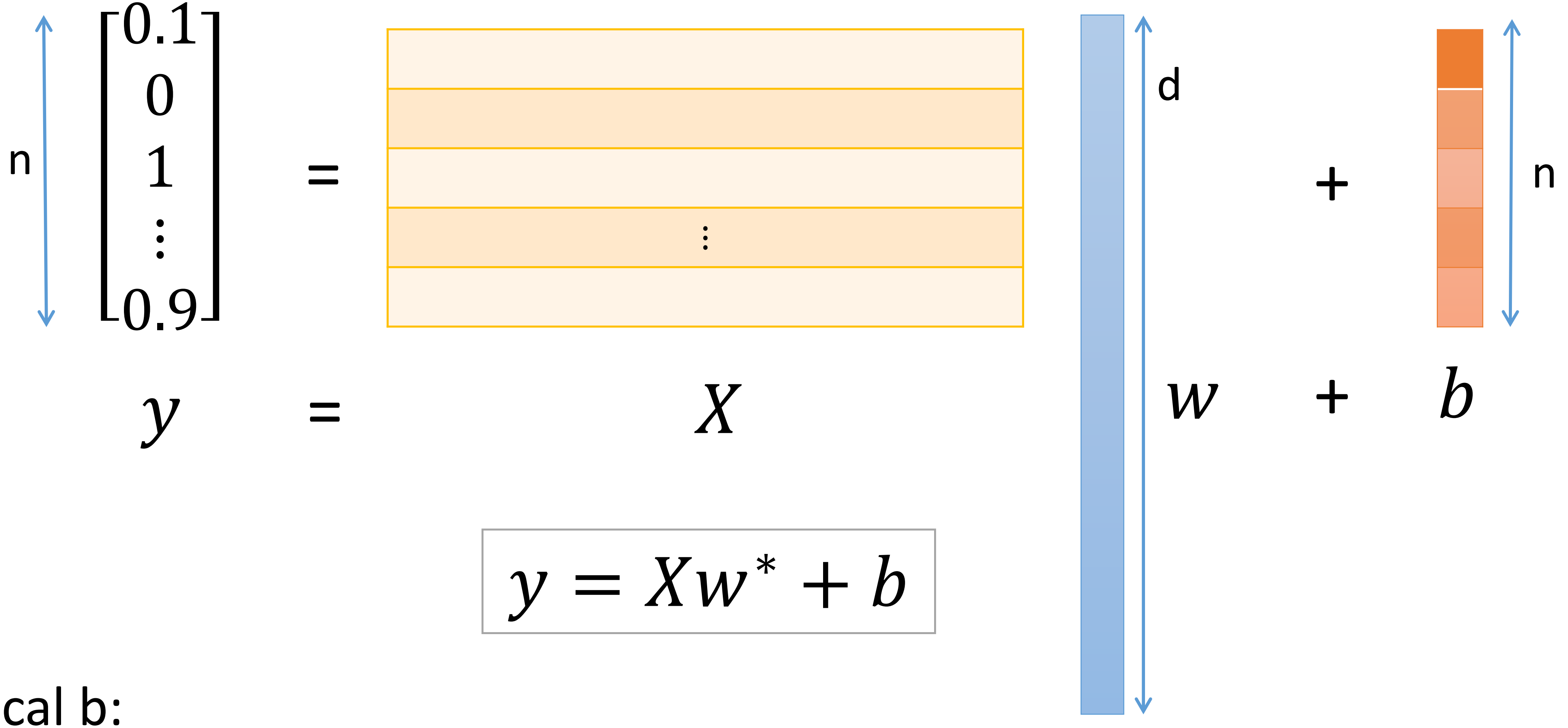
- $\kappa = \lambda_1(\Sigma)/\lambda_d(\Sigma)$

[Jain, Tewari, Kar’2014]

Summary so far…

- High-dimensional problems
  - $n \ll d$
- Need to impose structure on w
- Sparsity
  - Projection easy!
  - Projected Gradient works (if RIP is satisfied)
  - Several variants exist
- RIP/RSC style proof works for subgaussian data
- Other structures also allowed

Robust Regression



$$\boxed{y = Xw^* + b}$$

Typical b:

a) Deterministic error : $||w - w^*|| \leq ||b||$

b) Gaussian error : $||w - w^*|| \leq \frac{||b||}{\sqrt{n}}$

Robust Regression

- $||b||_0 \leq \beta \cdot n$
  - We want β to be a constant
- Entries of b can be unbounded!
  - $||b||_2$ can be arbitrarily large

- Still we want: $||w - w^*|| = 0$

RR Algorithm

- $S_0 = \{1, 2, \ldots, n\}$
- For t=0, 1, ….
  - $w_{t+1} = \arg\min ||X_{S_t} w - y_{S_t}||_2^2$
  - $r_{t+1} = y - X w_{t+1}$
  - $S_{t+1} = Top(|r_{t+1}|, \beta \cdot n)$

- Algorithm: was vaguely proposed by Legendre-1805

Result

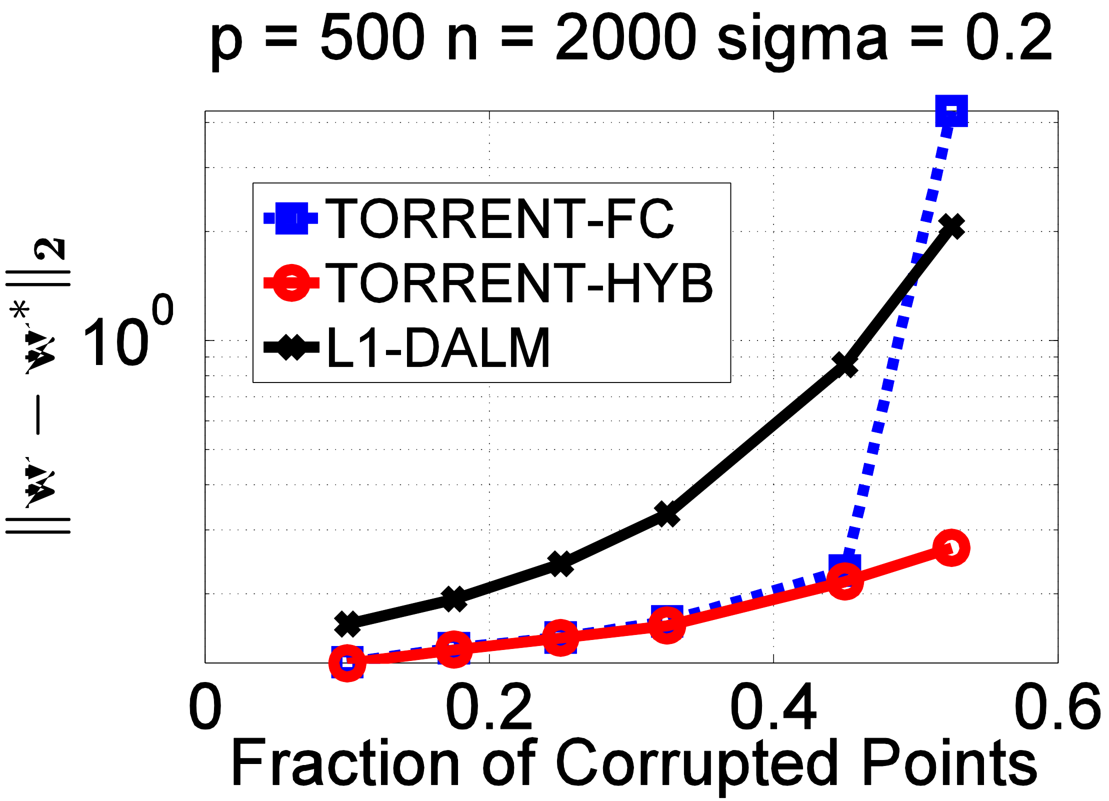
- $y = Xw^* + b$
- $||b||_0 \leq \beta \cdot n$
- $\beta \leq \frac{1}{100}$
- $n \geq d \log d$
- $X_{ij} \sim N(0,1)$

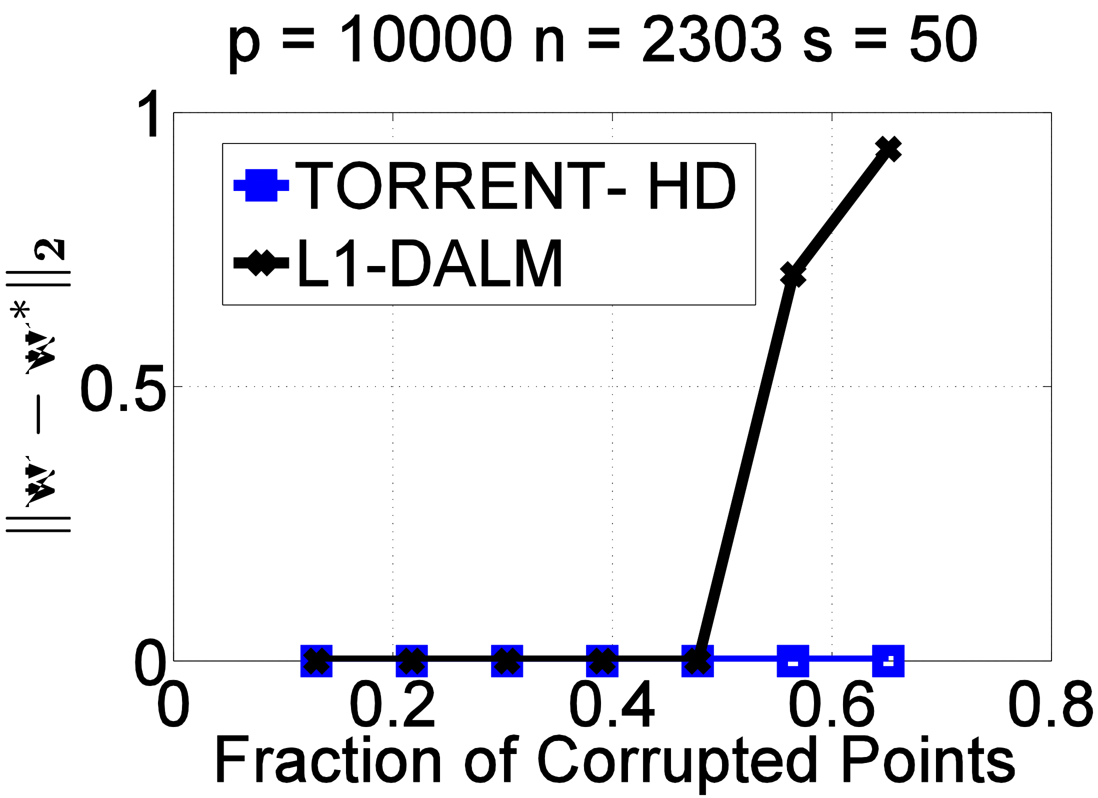
$$||b_{S_{t+1}}||_2 \leq \frac{9}{10} \, ||b_{S_t}||_2$$

Proof?

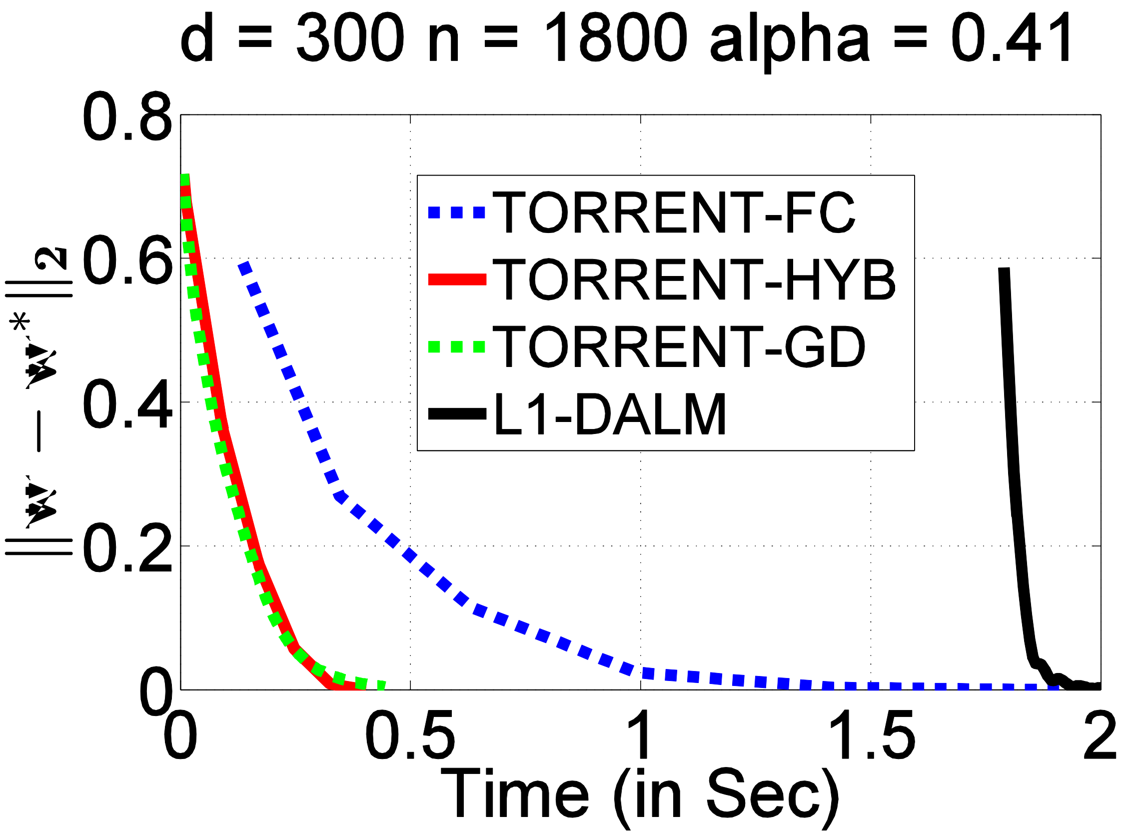
Proof?

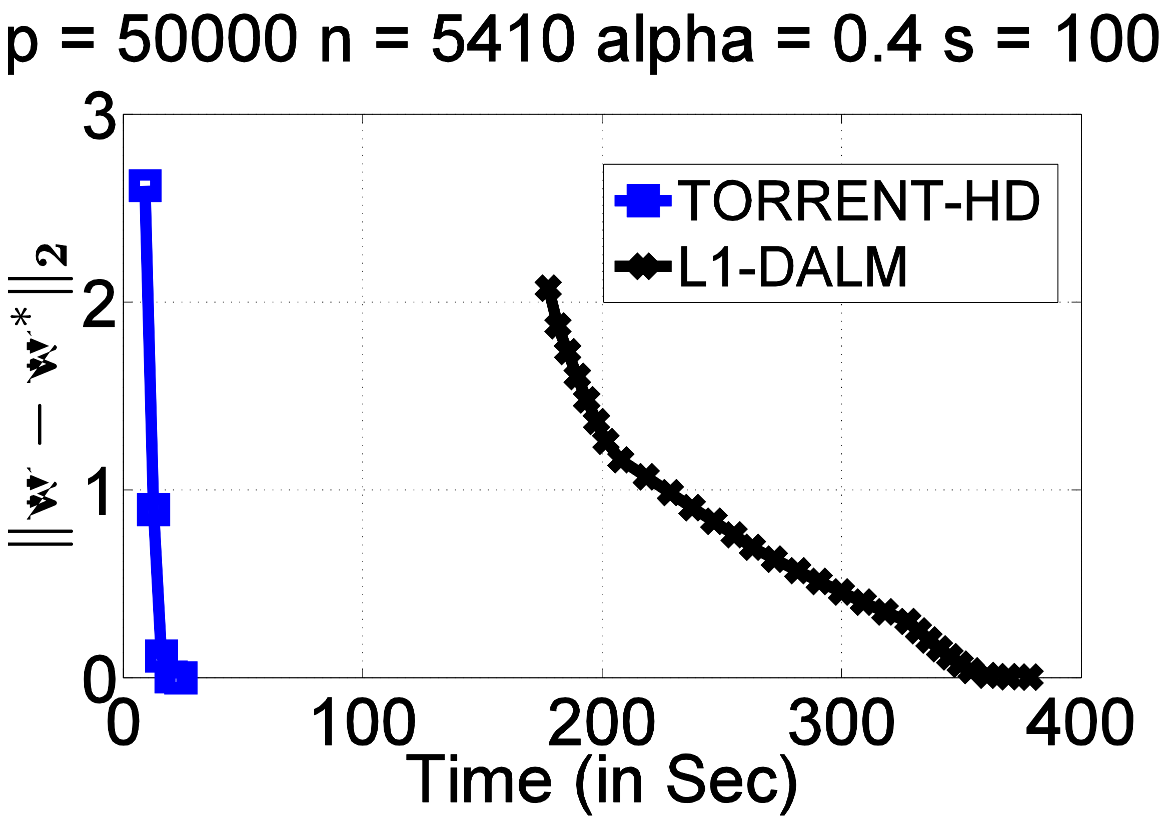
Empirical Results



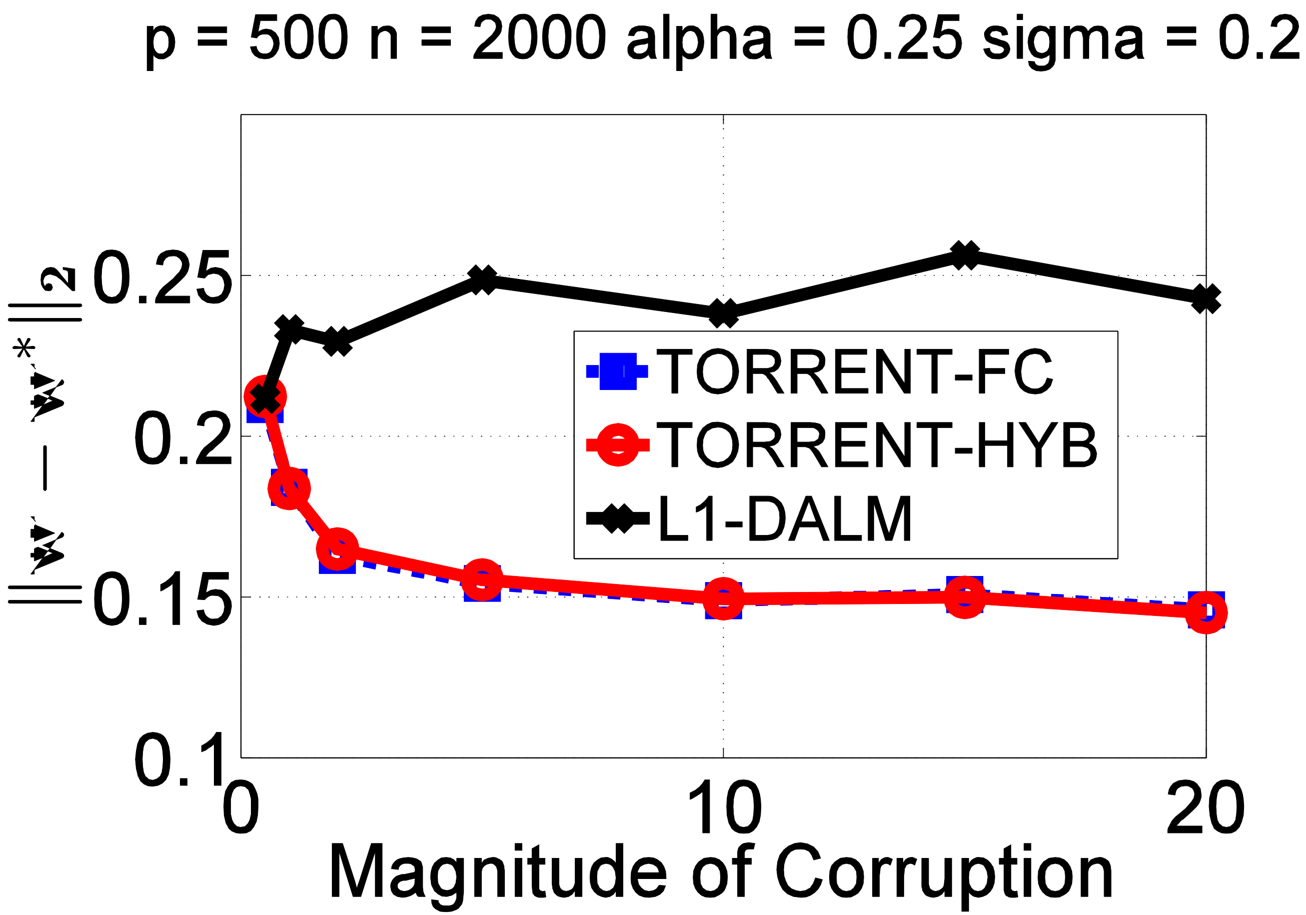


Empirical Results


d = 300 n = 1800 alpha = 0.41


p = 50000 n = 5410 alpha = 0.4 s = 100

Empirical Results



One-bit Compressive Sensing

- Compressed Sensing:

$$\mathbf{y} = \mathrm{X}\mathbf{w}^*$$

  - Require to know $\mathbf{y}$ exactly
  - In practice, finite bit representation, some quantization required
- One-bit CS: extreme quantization

$$\mathbf{y} = sign(\mathrm{X}\mathbf{w}^*)$$

  - Easily implementable through comparators
- Results in two categories:
  - Support Recovery [HB12, GNJN13]
  - Approximate Recovery [PV12, GNJN13]

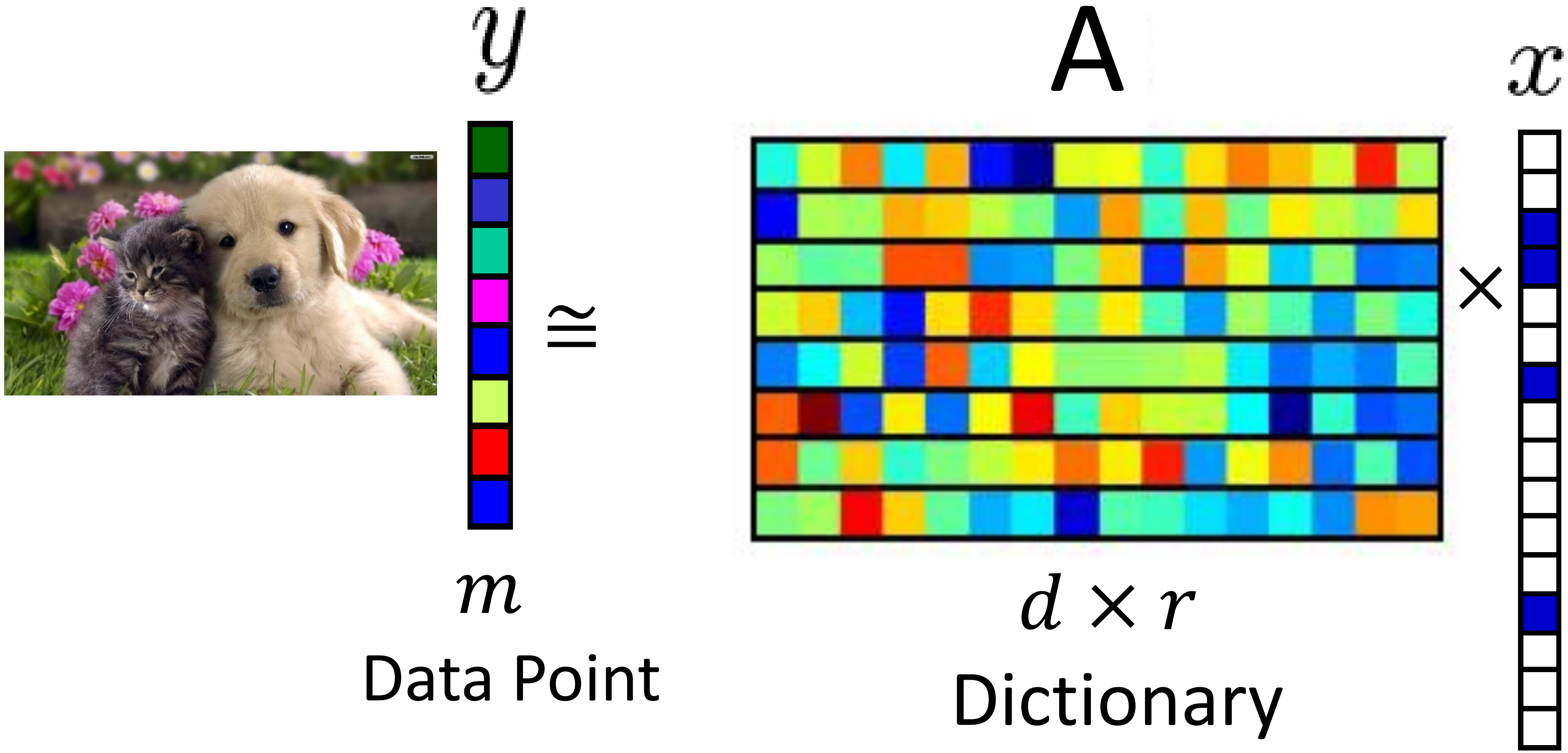
Phase-Retrieval

- Another extreme:

$$\mathbf{y} = |X\mathbf{w}^*|$$

  - Useful in several imaging applications
  - A field in itself
- Ideas from sparse-vector and low-rank matrix estimation [C12, NJS13]

Dictionary Learning



$y \cong A \times x$

m

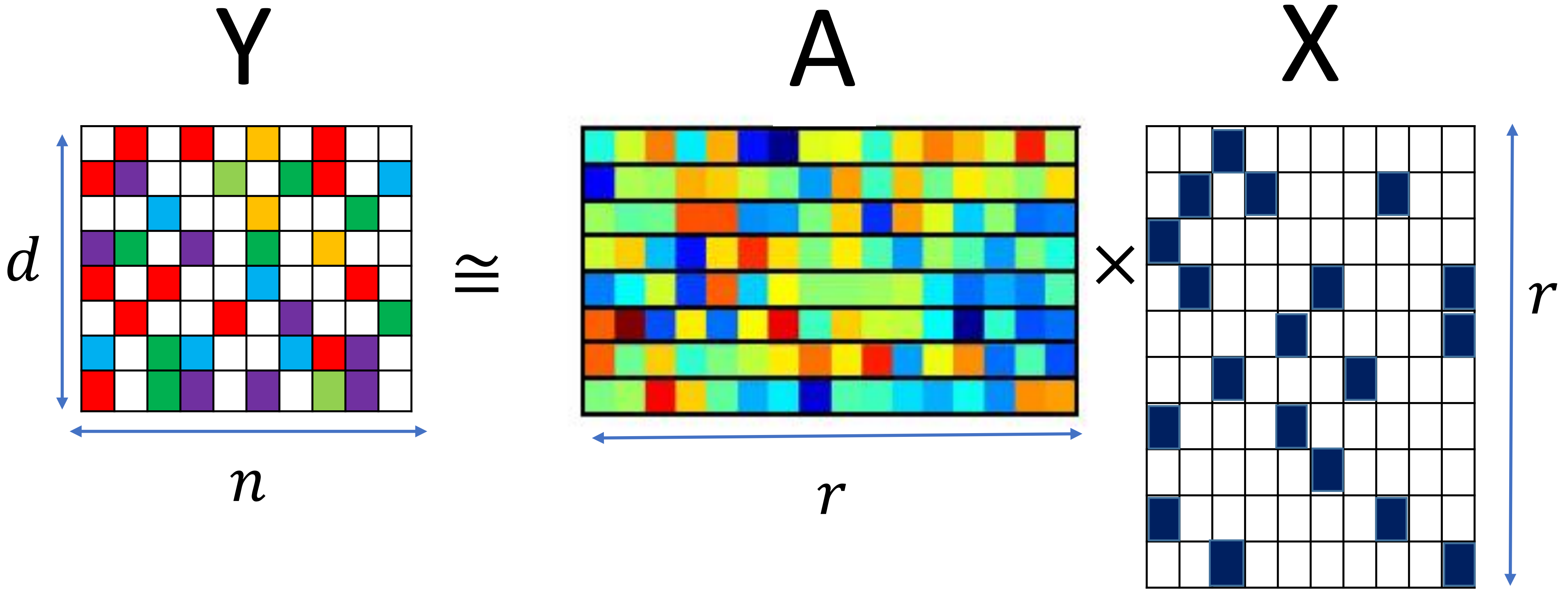
Data Point

$d \times r$

Dictionary

r-dim, k-sparse vector

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., $\langle A_i, A_j \rangle \le \mu/\sqrt{d}$
  - $r \gg d$
- 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

- Typically practical algorithm: alternating minimization
  - $X_{t+1} = argmin_X ||Y - A_t X||_F^2$
  - $A_{t+1} = argmin_A ||Y - AX_{t+1}||_F^2$
- 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:

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

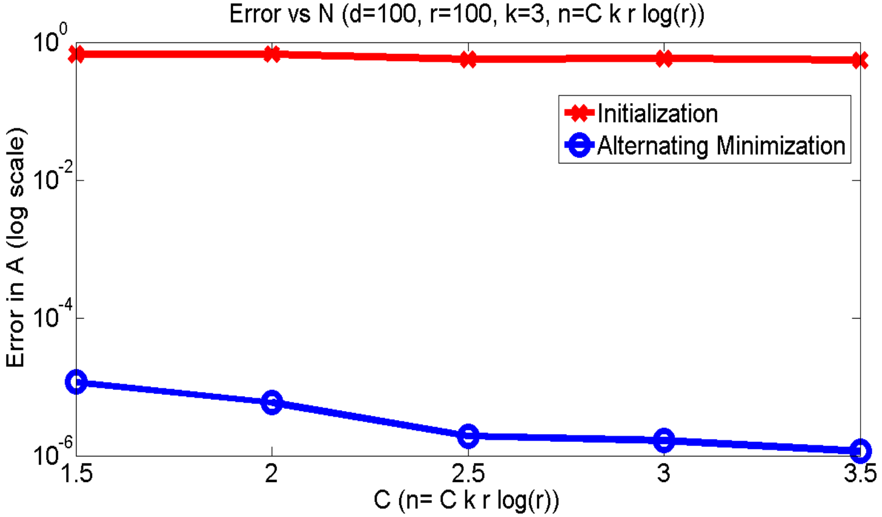
Proof Sketch

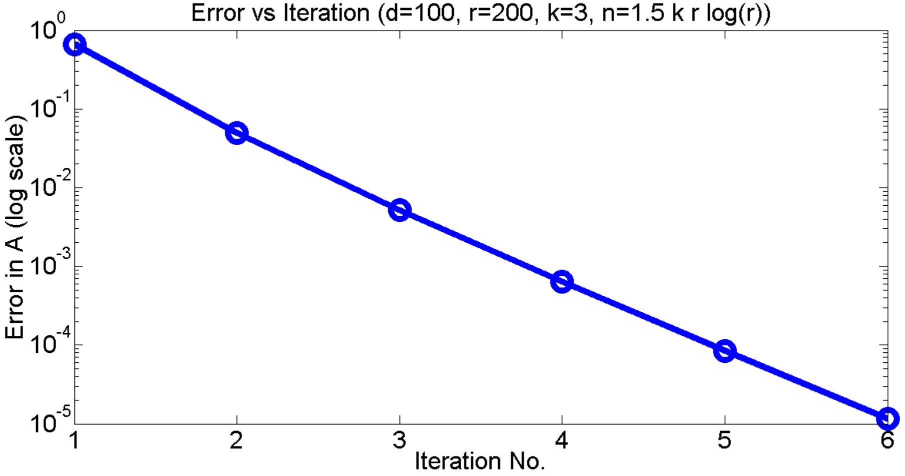
- Initialization step ensures that:

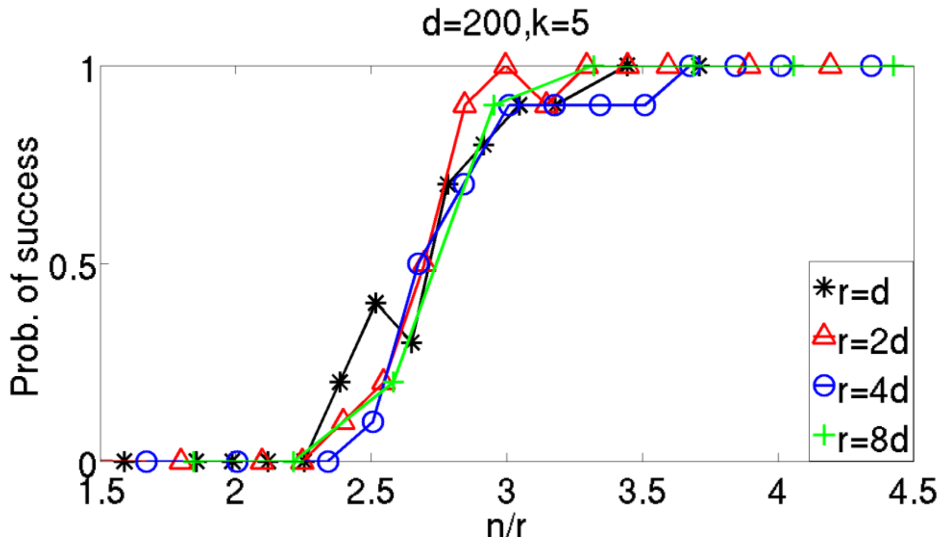
$$||A^i - A_0^i|| \leq \frac{1}{k^2}$$

- 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

- Consider high-dimensional structured problems
  - Sparsity
  - Block sparsity
  - Tree-based sparsity
  - Error sparsity
- Iterative hard thresholding style method
  - Practical/easy to implement
  - Fast convergence
- RIP/RSC/subGaussian data: Provable guarantees

http://research.microsoft.com/en-us/people/prajain/

Purushottam Kar

PostDoc
MSR, India

Kush Bhatia

Research Fellow
MSR, India

Ambuj Tewari

Asst. Prof.
Univ of Michigan

Next Lecture

- Low-rank Structure
    - Matrix Regression
    - Matrix Completion
    - Robust PCA
- Low-rank Tensor Structure
    - Tensor completion

Block-sparse Signals

$$\mathbf{y}_1 = \Phi_1\mathbf{x}_1, \mathbf{y}_2 = \Phi_2\boldsymbol{x}_2, \dots, \mathbf{y}_r = \Phi_r\mathbf{x}_r$$

- Total no. of measurements: $O(r \cdot k \cdot \log n)$
- Correlated signals: $\mathrm{J} = |x_1 \cup x_2 \dots x_r| \leq k \cdot r$
- Method--- Group norms: $L_{2,1}$ or $L_{2,\infty}$
- Improvement in sample complexity if

$$J \ll k \cdot r$$