Iterative Hard Thresholding for Sparse/Low-rank Linear Regression

Prateek Jain
Microsoft Research, India



Ambuj Tewari
Univ of Michigan

Purushottam Kar
MSR, India

Praneeth Netrapalli
MSR, NE

Microsoft Research India

Our work

- Foundations
- Systems
- Applications
- Interplay of society and technology
- Academic and government outreach

Our vectors of impact

- Research impact
- Company impact
- Societal impact

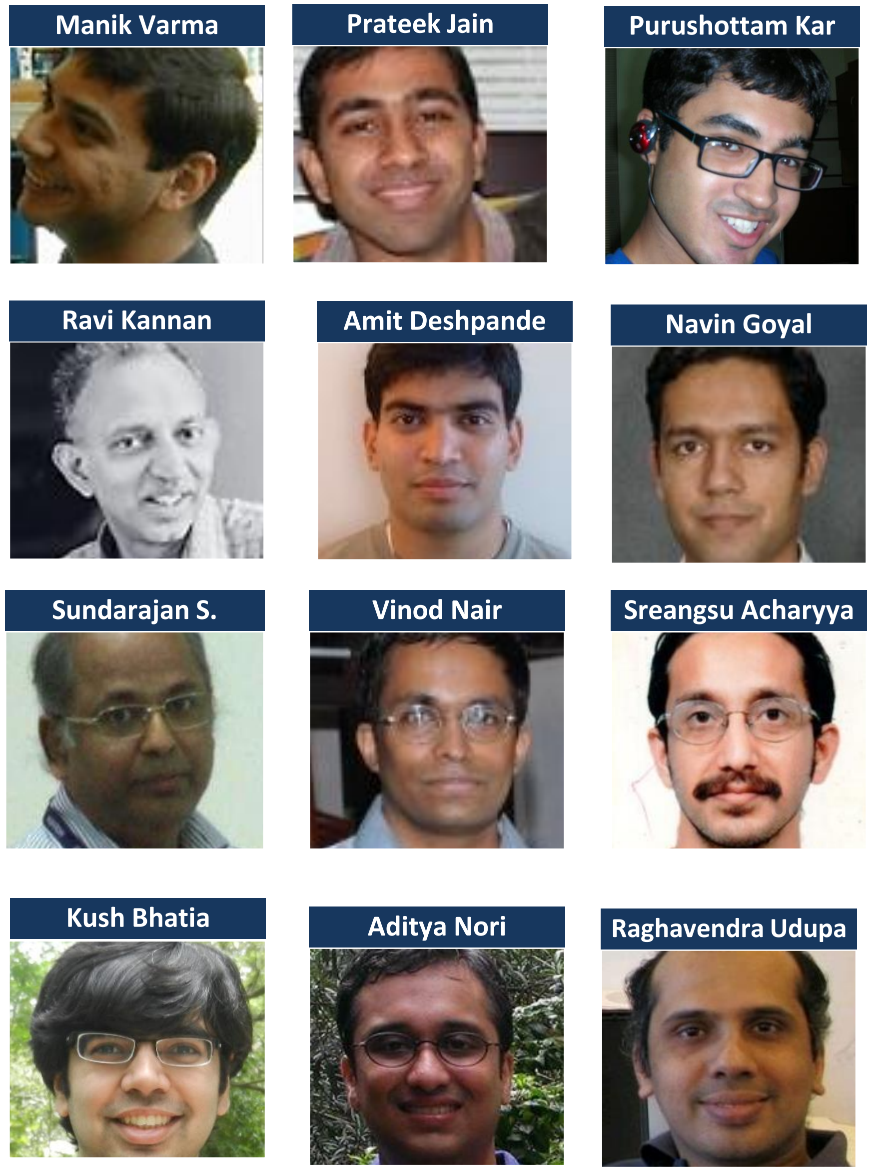
Machine Learning and Optimization @ MSRI

- *High-dimensional Learning & Optimization*
- *Extreme Classification*
- *Online Learning / Multi-armed Bandits*

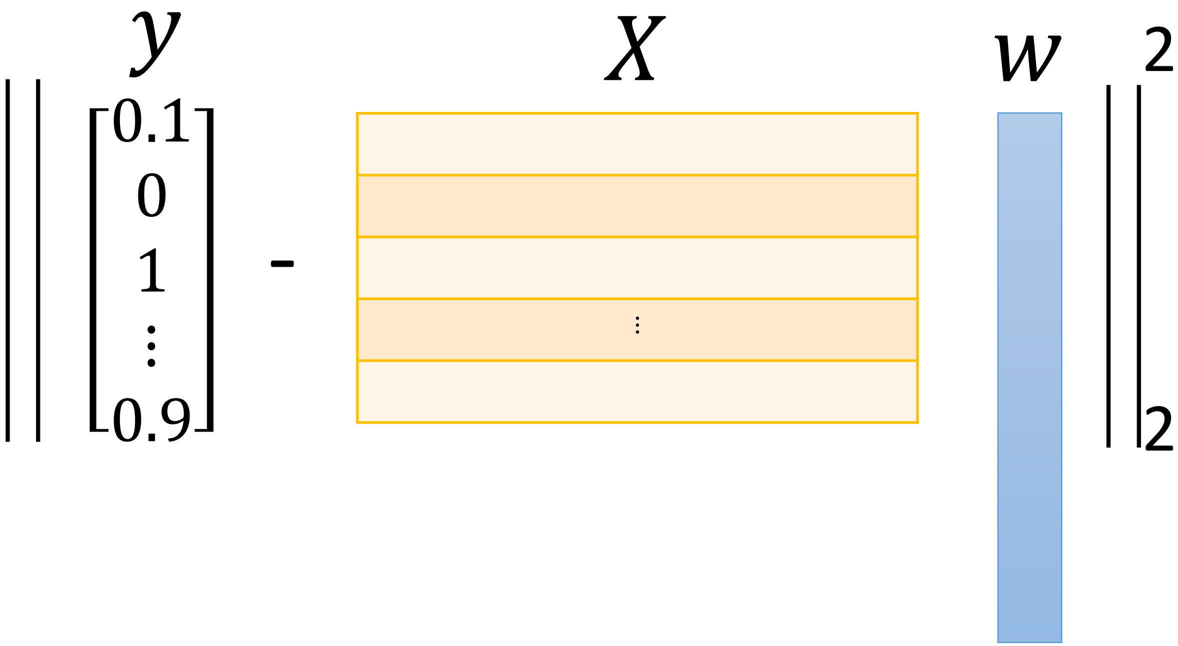
We are Hiring!

- Interns
- PostDocs
- Applied Researchers
- Full-time Researchers

- *Ranking & Recommendation*



Learning in High-dimensions

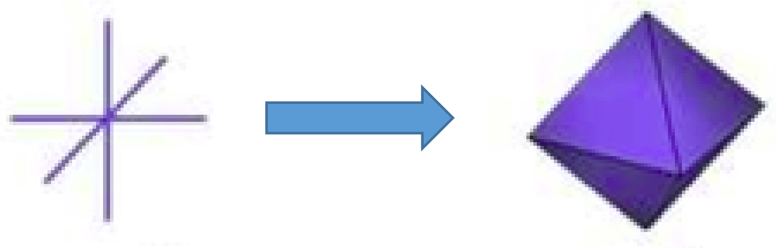


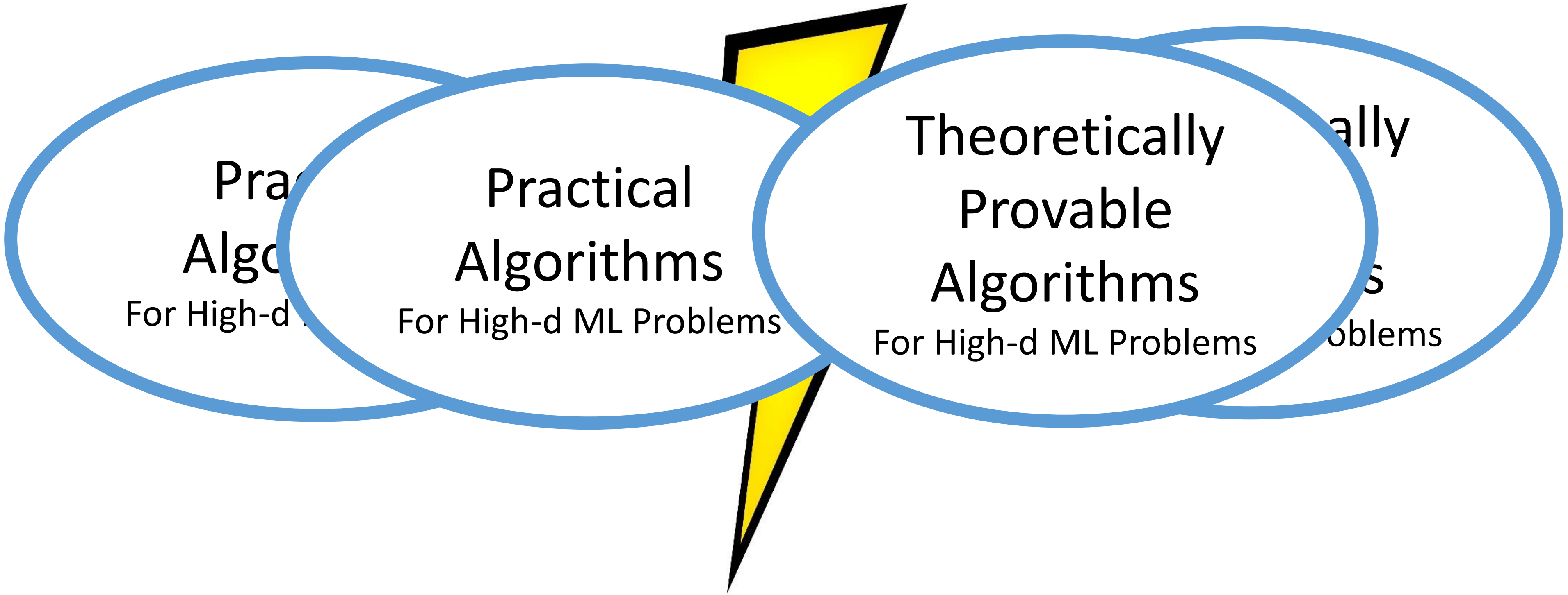
- Need to solve: $\min_{w} ||y - Xw||_2^2 \; s.t. \, w \in C$
- C can be:
  - Set of sparse vectors
  - Set of group-sparse vectors
  - Set of low-rank matrices

- Non-convex
- Comp. Complexity: NP-Hard

Overview

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





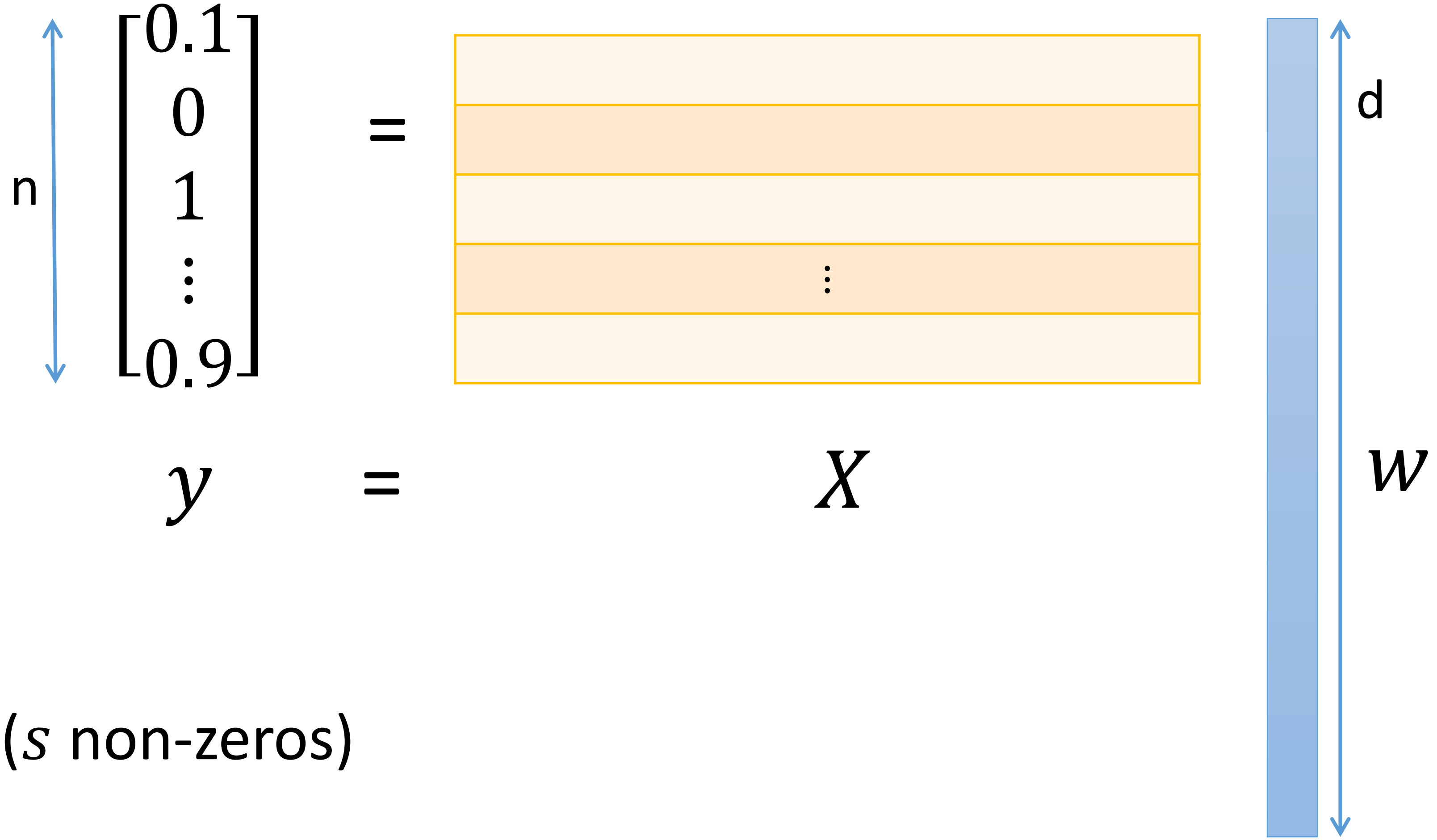
Results

- Sparse Regression [Garg & Khandekar. ICML 2008; J., Kar, Tewari. NIPS14]
  - $L_0 -$ constraint
- Matrix Completion/Regression [J., Netrapalli, Sanghavi. STOC 2013; Hardt & Wooters. COLT 2014]
  - Low-rank constraint
- Robust Regression [Loh & Wainwright. NIPS 2013 ; Bhatia, J., Kar. Submitted, 2015]
- Tensor Factorization and Completion [Anandkumar et al. Arxiv 2012; J., Oh. NIPS14]
  - Low-tensor rank constraint
- Dictionary Learning [Agarwal et al. COLT 2014; Arora et al. COLT 2014]
  - Non-convex bilinear form + Sparsity constraint
- Phase Sensing [Netrapalli, J., Sanghavi. NIPS13 ; Candes et al. Arxiv'2014]
  - System of quadratic equations
- Low-rank matrix approximation [Bhojanapalli, J., Sanghavi. SODA15]

Outline

- Sparse Linear Regression
  - Lasso
- Iterative Hard Thresholding
  - Our Results
- Low-rank Matrix Regression
- Low-rank Matrix Completion
- Conclusions

Sparse Linear Regression



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

Sparse Linear Regression

$$\min_w ||y - Xw||^2$$
$$s.t.\ \ ||w||_0 \leq s$$

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

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

Convex Relaxation

$$\min_w ||y - Xw||^2$$
$$s.t. \;\; ||w||_0 \leq s$$

- Relaxed Problem:

$$\min_w ||y - Xw||^2 \quad s.t. \;\; ||w||_1 \leq \mu(s) \quad \Longleftrightarrow \quad \min_w ||y - Xw||^2 + \lambda ||w||_1$$

Non-differentiable

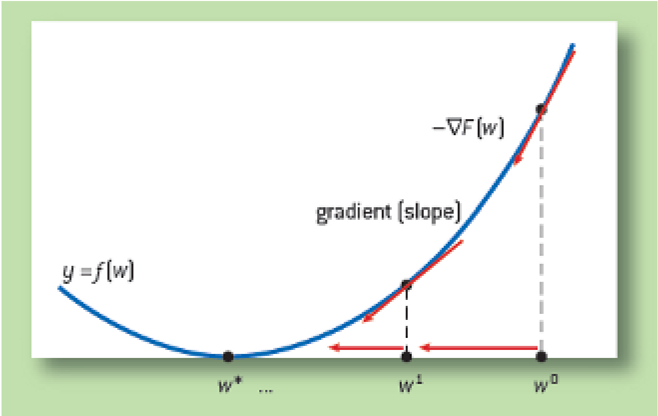
Lasso Problem

- $||w||_1 = \sum_i |w_i|$
  - Known to promote sparsity
- Pros: a) Principled approach, b) Solid theoretical guarantees
- Cons: Slow to optimize

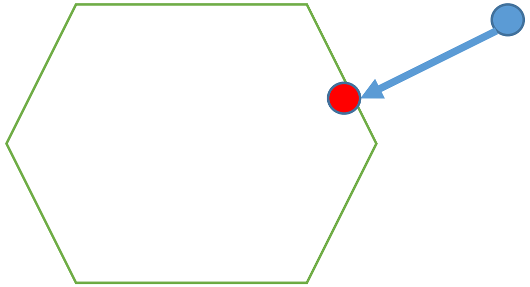
Our Approach : Projected Gradient Descent

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

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



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

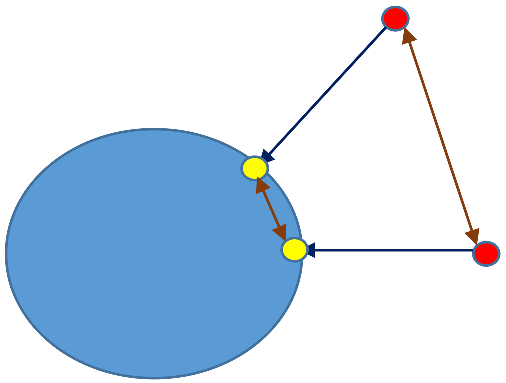


[Jain, Tewari, Kar’2014]

Projection onto L_0 ball?

$$\underset{z}{\begin{bmatrix} 0.1 \\ 0.01 \\ 1 \\ -1 \\ 0.9 \\ -0.1 \end{bmatrix}} \xrightarrow{\text{sort}} \begin{bmatrix} 1 \\ -1 \\ 0.9 \\ 0.1 \\ -0.1 \\ 0.01 \end{bmatrix} \xrightarrow{\text{Hard Thresholding}} \underset{P_3(z)}{\begin{bmatrix} 1 \\ -1 \\ 0.9 \\ 0 \\ 0 \\ 0 \end{bmatrix}}$$

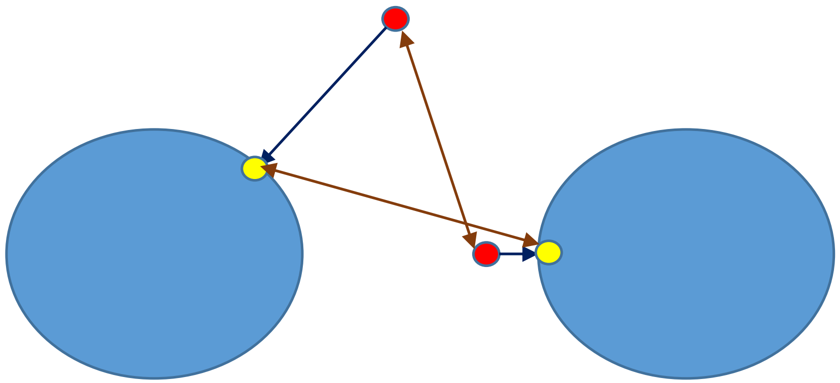
Convex-projections vs Non-convex Projections



$$||P_C(a) - P_C(b)|| \leq ||a - b||$$

C: convex set

1st order Optimality condition



$$||P_C(a) - a|| \leq ||u - a||, \qquad \forall\, u \in C$$

C: non-convex set

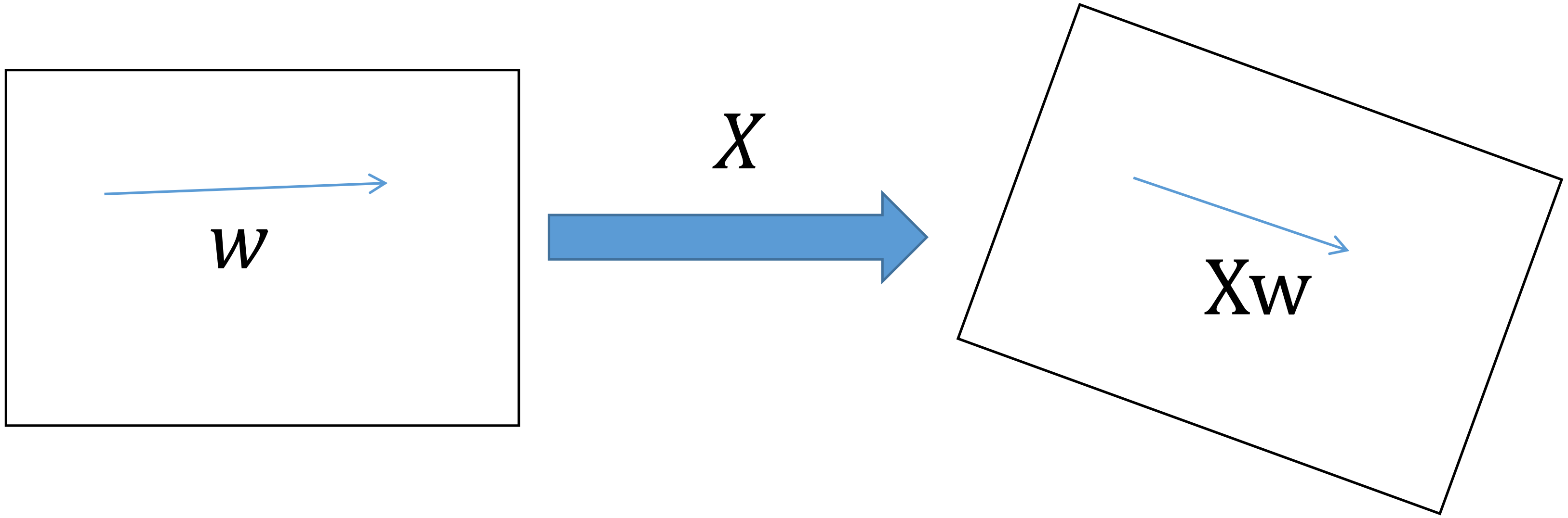
0-th order Optimality condition

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

Restricted Isometry Property (RIP)

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

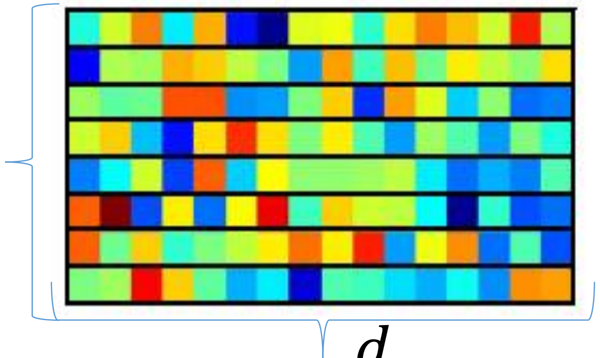
$$(1-\delta_s)||\mathbf{w}||^2 \leq ||\mathbf{Xw}||^2 \leq (1+\delta_s)||\mathbf{w}||^2$$



Popular RIP Ensembles

X

$n = O(\frac{s}{\delta_s^2} \log \frac{d}{s})$



d

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

Proof under RIP

Assume: $y = Xw^*, \quad \min_{w, ||w||_0 \leq s} f(w) = ||y - Xw||^2 = ||X(w - w^*)||^2$

Recall: $w_{t+1} = P_s(w_t - X^T X(w_t - w^*))$

Hard Thresholding

$$||w_{t+1} - \left(w_t - X^T X(w_t - w^*)\right)|| \leq ||w^* - \left(w_t - X^T X(w_t - w^*)\right)||$$

$I: supp(w_t) \cup supp(w_{t+1}) \cup supp(w^*), \quad |I| \leq 3s$

$$||w_{t+1} - \left(w_t - X_I^T X_I(w_t - w^*)\right)||^2 \leq ||w^* - \left(w_t - X_I^T X_I(w_t - w^*)\right)||^2$$

Triangle inequality

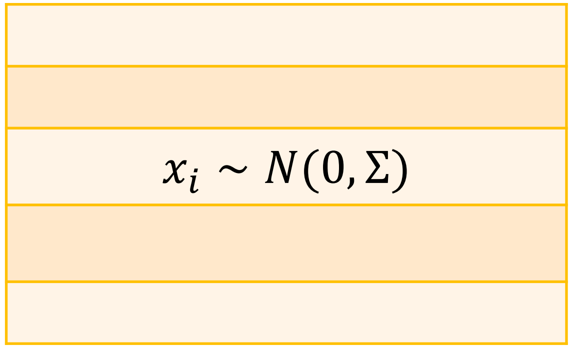
$$||w_{t+1} - w^*|| \leq 2||(I - X_I^T X_I)(w_t - w^*)||$$
$$\leq 2\delta_{3s}||w_t - w^*||$$

RIP

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

What if RIP is not possible?

- $y_i = \langle x_i, w^* \rangle$
- $x_i \sim N(0, \Sigma)$



$$\Sigma = \begin{bmatrix} 1 & 1-\epsilon & 0 \\ 1-\epsilon & 1 & 0 \\ 0 & 0 & I_{d-2 \times d-2} \end{bmatrix}$$

- Eigenvalues of $\Sigma = 2-\epsilon,\ \epsilon$
- $\delta_s \geq \delta_2 = 1-\epsilon$
- So, $\delta_s < \frac{1}{2}$ doesn't hold even for infinite samples
  - Problem is solvable for $O(d)$ samples using standard regression

Iterative Hard Thresholding: Larger Sparsity

- $w_1 = 0$
- For t=1, 2, …
  - $w_{t+1} = P_{s'}(w_t - \eta \nabla_w f(w_t))$

- $s' \geq s$

Stronger Projection Guarantee

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

- d: dim of z
- $s \leq s'$

Statistical Guarantees

$$\min_{w} f(w) = ||y - Xw||^2$$

Statistically: $n \geq \frac{\sigma^2 \cdot s \log d}{\epsilon^2}$

Known Computation Lower-bound: $\frac{\kappa \cdot \sigma^2 \cdot s \log d}{\epsilon^2}$

- $w^*: s -$sparse

$$|| \hat{w} - w^* || \leq \epsilon, \qquad n \geq \frac{\kappa^2 \cdot \sigma^2 \cdot s \log d}{\epsilon^2}$$

Same as Lasso

- $\kappa = \lambda_1(\Sigma)/\lambda_d(\Sigma)$
- Recall, $w_{t+1} = P_{s'}(w_t - \eta \nabla_w f(w))$
  - $s' = \kappa^2 s$

[J., Tewari, Kar'2014]

General Result for Any Function

- $f: R^d \rightarrow R$
- 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 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

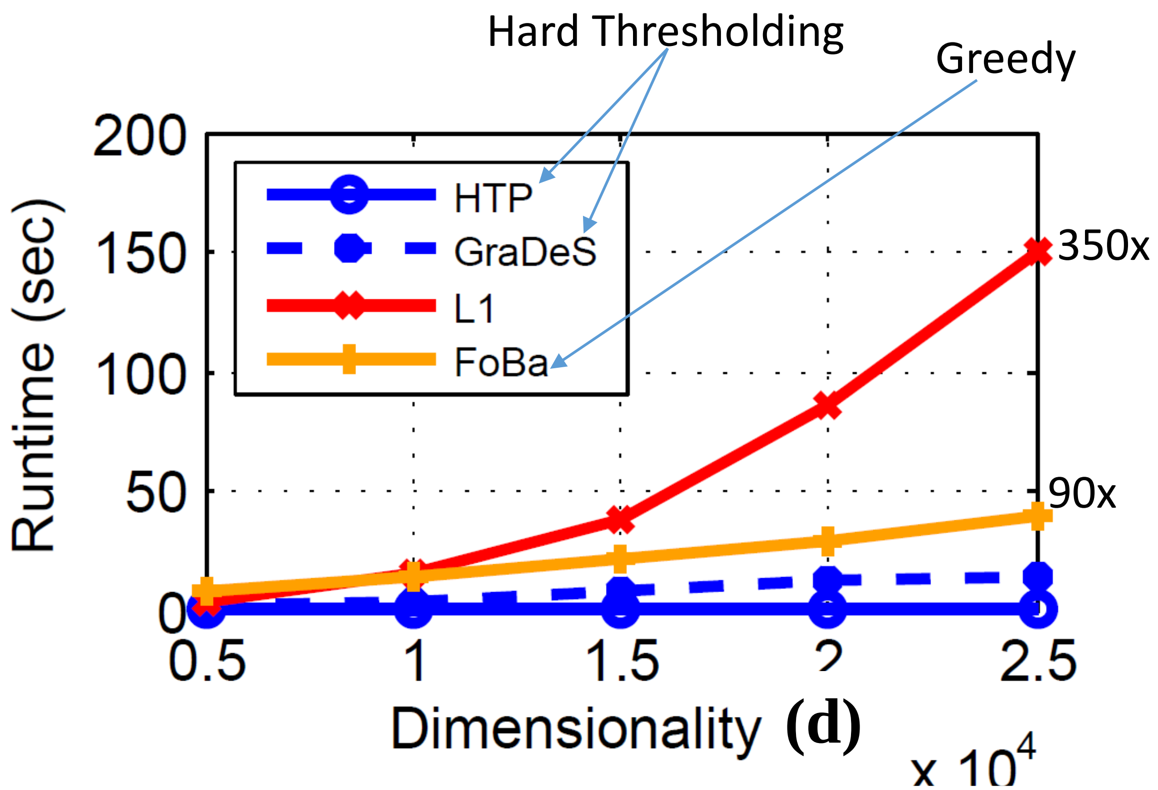
- If $||w^*||_0 \leq s$ and $s' \geq 10\,\frac{L_{s'}^2}{\alpha_{s'}^2}\,s$

[J., Tewari, Kar'2014]

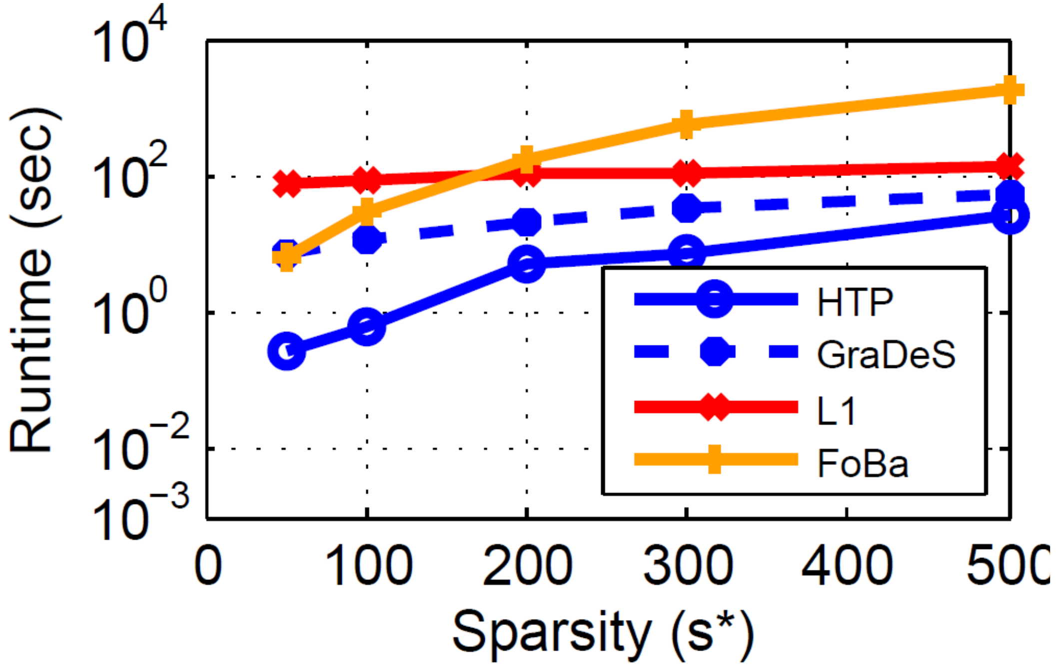
Extension to other Non-convex Procedures

- IHT-Fully Corrective
  - HTP [Foucart'12]
- CoSAMP [Tropp & Neadell'2008]
- Subspace Pursuit [Dai & Milenkovic'2008]
- OMPR [J., Tewari, Dhillon'2010]
- Partial hard thresholding and two-stage family [J., Tewari, Dhillon'2010]

Empirical Results

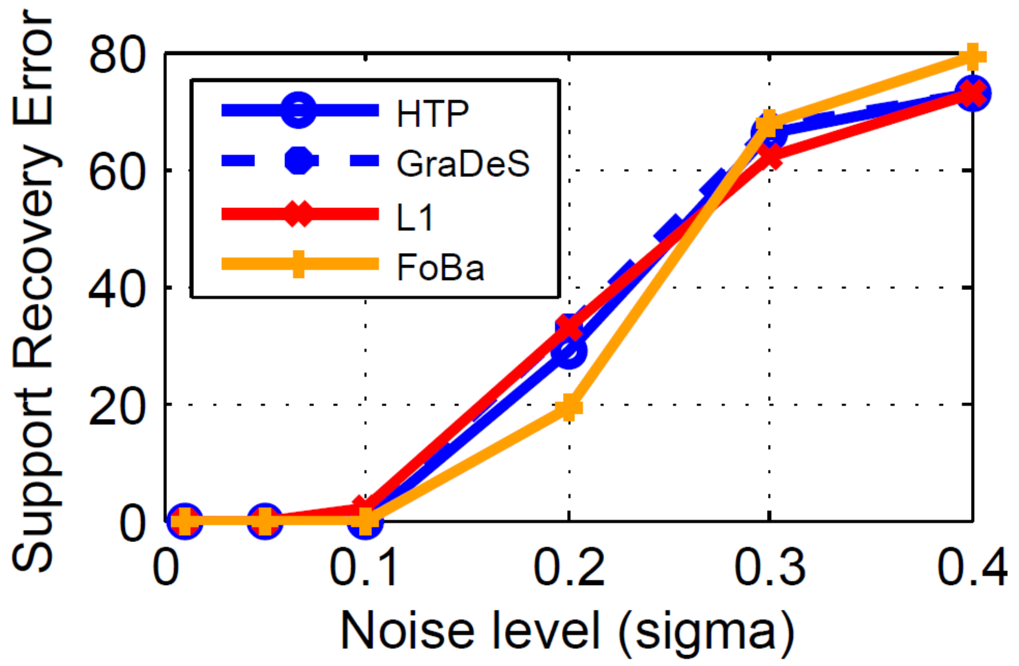


$$n = 2 \cdot s \cdot \log(d)\,, s = 300, \kappa = 1$$

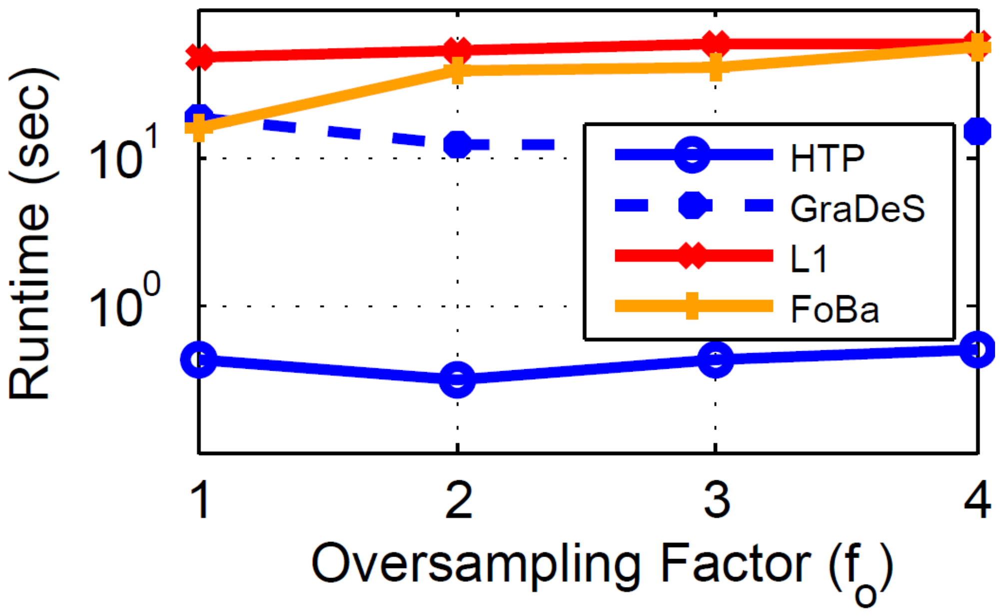


$$n = 2 \cdot s \cdot \log(d)\,, d = 20{,}000, \kappa = 1$$

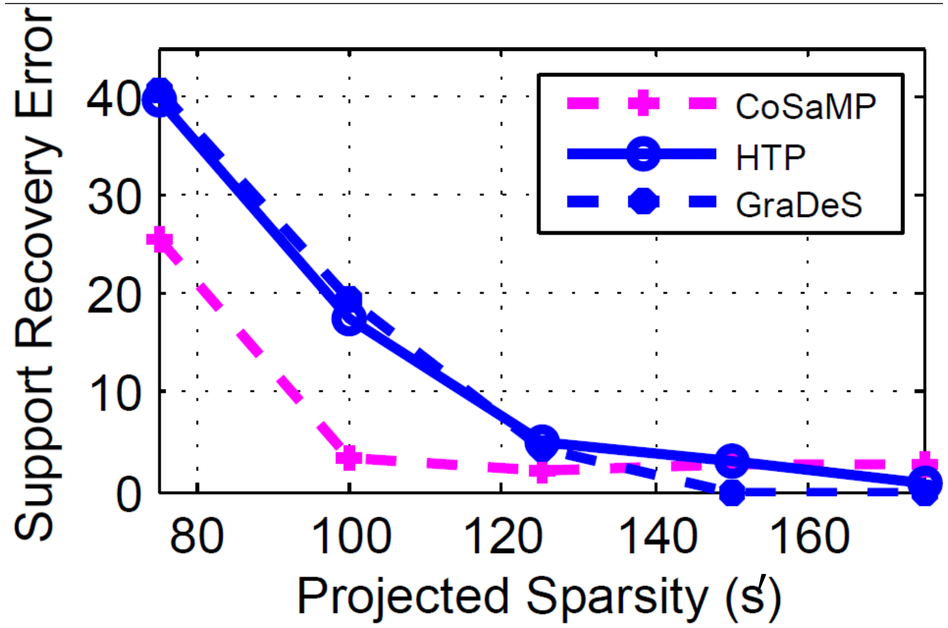
More Empirical Results



$$n = f_0 \cdot s \cdot \log d$$



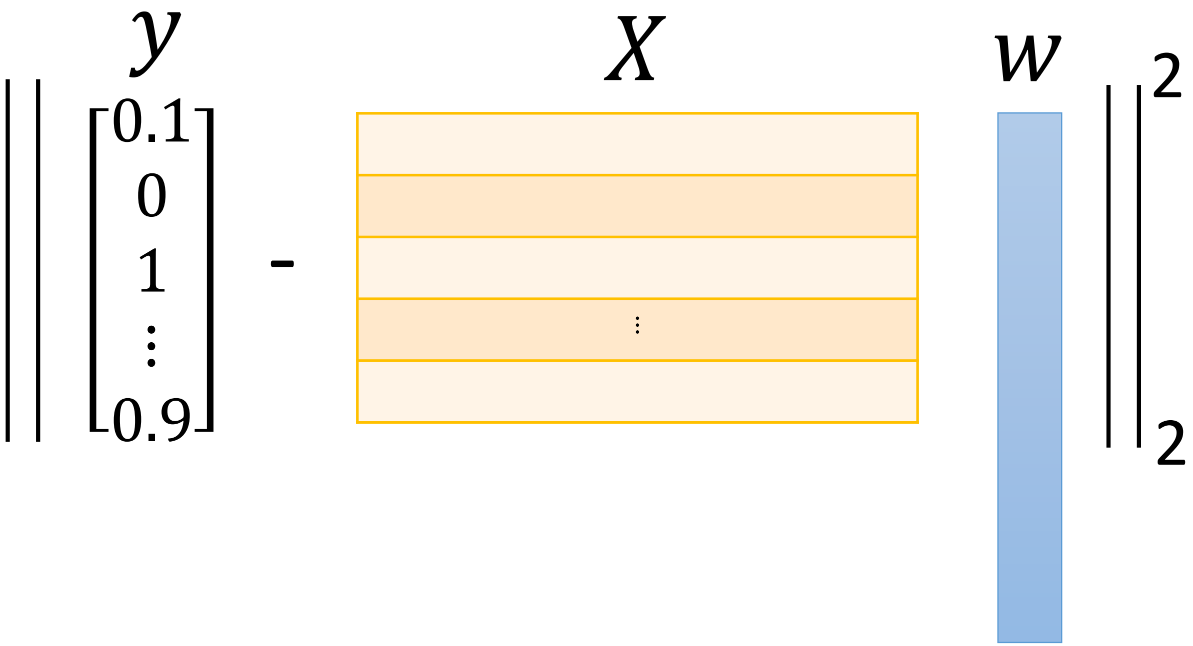
Empirical Results: poor condition number



$$n = 2 \cdot s \cdot \log(d)\,, s = 50, d = 20{,}000$$

$$\kappa = 50$$

Low-rank Matrix Regression



- W: $d_1 \times d_2$ matrix
- $rank(W) = r \ll \min(d_1, d_2)$

Low-rank Matrix Regression

$$\min_{W} f(W) = ||y - X \cdot W||^2$$
$$s.t.\ \ rank(W) \leq r$$

- Convex relaxation: $rank(W) \Rightarrow\ ||W||_*$
  - $||W||_* =$ sum of singular values of W
  - Several interesting results: [Recht et al.'2007, Negahban et al'2009]…

- Projected Gradient Descent:
  - $W_{t+1} = P_k\big(W_t - \eta \nabla_W f(W_t)\big), \forall t$
  - $k \geq r$
- $P_k(Z) = U_k \Sigma_k V_k^T$ where $Z = U \Sigma V^T$

Statistical Guarantees

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

- $x_i \sim N(0, \Sigma) \in R^d$
- $\eta_i \sim N(0, \sigma^2)$
- $W^* \in R^{d_1 \times d_2}, \quad rank(W^*) = r$

$$|| \widehat{W} - W^* ||_2 \leq \frac{\sigma \cdot \kappa \cdot \sqrt{r(d_1 + d_2) \log(d_1 + d_2)}}{\sqrt{n}}$$

- $\kappa = \frac{\lambda_1(\Sigma)}{\lambda_d(\Sigma)}, \qquad k = \kappa^2 r$

[J., Tewari, Kar'2014]

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

$$\min_{W} \sum_{(i,j)\in\Omega} \left(W_{ij} - M_{ij}\right)^2$$

$$s.t \quad \mathbf{rank}(W) \leq r$$

Ω: set of known entries

- Special case of low-rank matrix regression
- However, assumptions required by the regression analysis not satisfied

Guarantees

- Projected Gradient Descent:
  - $W_{t+1} = P_r\big(W_t - \eta \nabla_W f(W_t)\big), \qquad \forall t$
- Show ϵ-approximate recovery in $\log \frac{1}{\epsilon}$ iterations
- Assuming:
  - M: incoherent
  - Ω: uniformly sampled
  - $|\Omega| \geq n \cdot r^5 \cdot \log^3 n$
- First near linear time algorithm for **exact** Matrix Completion with finite samples

[J., Netrapalli'2015]

Tale of two Lemmas

- Lemma 1: Perturbation bound with L_∞ bounds

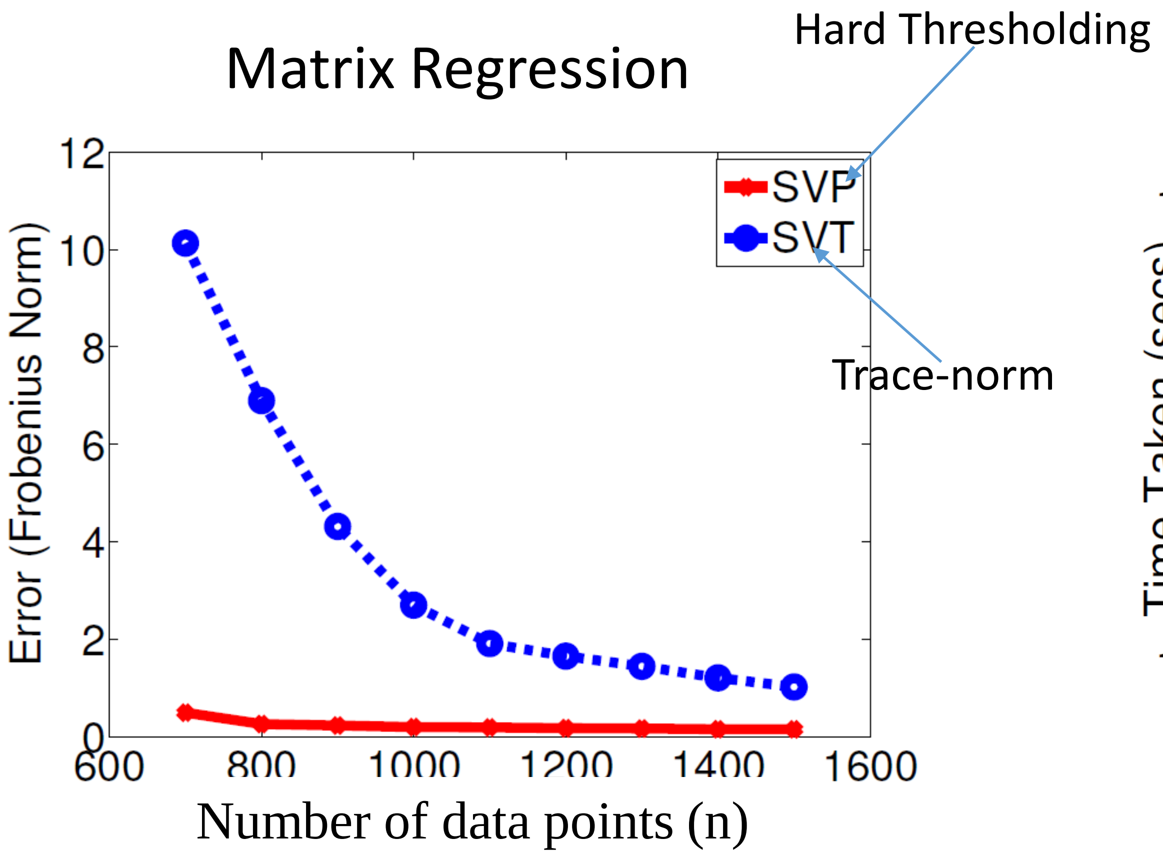
$$||P_r(M+E_t) - M||_\infty \leq .5\ ||E_t||_\infty$$

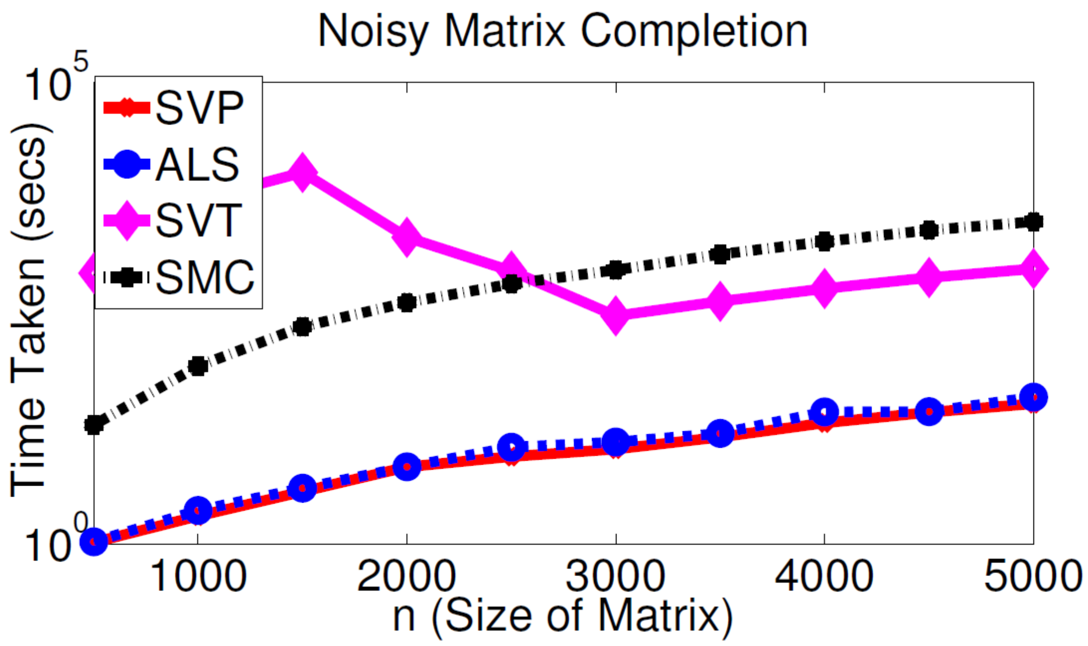
  - Standard bounds only give: $||P_r(M+E_t) - M||_2 \leq 2||E_t||_2$
  - M: incoherent
  - E_t: zero-mean with small variance

- Lemma 2: Davis-Kahn style result for matrix perturbation
  - If $\sigma_{k+1}(M) < .25\ \sigma_k(M)$ and $||E_t||_F \leq .25\ \sigma_k(M)$

$$||P_k(M+E_t) - P_k(M)||_F \ \leq c(\sqrt{k}\ ||E_t||_2 + ||E||_F)$$

Empirical Results





$$r = 100, |\Omega| = 5\, r\, n \log n$$

Summary

- High-dimensional problems
  - $n \ll d$
- Need to impose structure on w
- Typical structures
  - Sparsity
  - Low-rank
  - Low-rank+Sparsity
- Non-convex sets but easy projection
- Proof of convergence (linear rate)
  - Under suitable generative model assumptions

Future Work

- Generalized theory for such provable non-convex optimization
- Performance analysis on different models
- Empirical comparisons on “real-world” datasets

Questions?