Finding Low Rank Matrices by Convex Optimization

Maryam Fazel

Electrical Engineering
University of Washington, Seattle

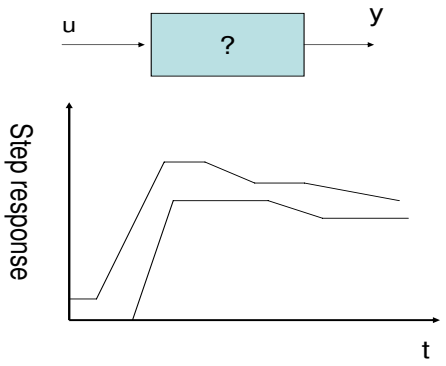
AIM workshop on convex algebraic geometry
9/24/09

Rank Minimization Problem

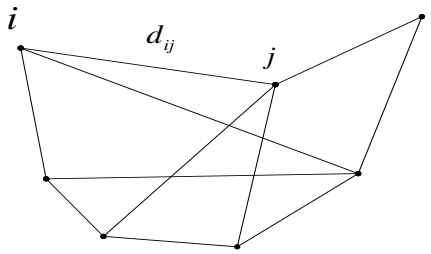
$$\begin{array}{ll} \text{minimize} & \mathbf{Rank}\, X \\ \text{subject to} & X \in \mathcal{C} \end{array}$$

where $X \in \mathbf{R}^{m \times n}$; $\mathcal{C}$ is convex set. Nonconvex problem, NP-hard in general.

system ID



distance geometry



min embedding dimension,
given pairwise distances

machine learning

users/movies database

$$\begin{bmatrix} 5 & ? & 8 & \dots & ? \\ ? & 10 & 3 & \dots & 5 \\ & & \ddots & & \\ 2 & 1 & ? & \dots & 6 \end{bmatrix}$$

movie recommendation based on few features

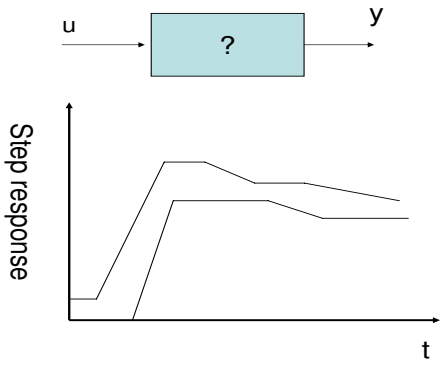
also: quadratic optimization (rank-1 solution);
low-rank matrix completion;
Sum-of-Squares decomposition (with a few squares)

Rank Minimization Problem

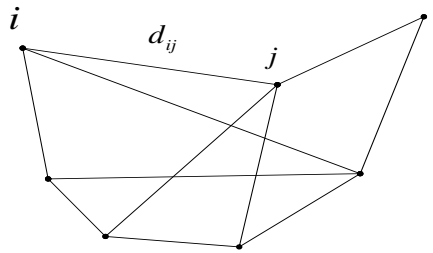
$$\begin{array}{ll} \text{minimize} & \mathbf{Rank}X \\ \text{subject to} & X \in \mathcal{C} \end{array}$$

where $X \in \mathbf{R}^{m\times n}$; $\mathcal{C}$ is convex set. Nonconvex problem, NP-hard in general.

system ID



distance geometry



min embedding dimension, given pairwise distances

machine learning

users/movies database

$$\begin{bmatrix} 5 & ? & 8 & \dots & ? \\ ? & 10 & 3 & \dots & 5 \\ & & \ddots & & \\ 2 & 1 & ? & \dots & 6 \end{bmatrix}$$

collaborative prediction with few features

special case: when $X = \begin{bmatrix} x_1 & & & \\ & x_2 & & \\ & & \ddots & \\ & & & x_n \end{bmatrix}$, $\mathbf{Rank}X = \#$ of nonzero x_i

find sparsest vector in $\mathcal{C}$

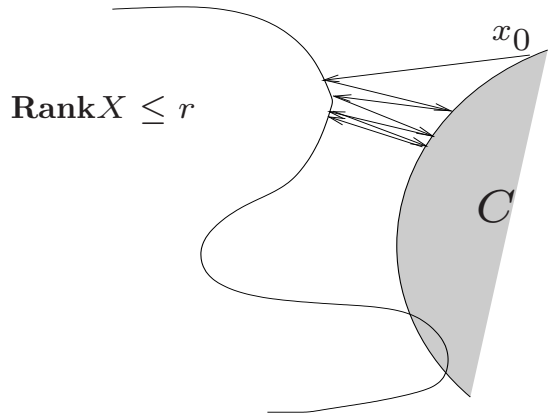
Overview: Solution Methods

Methods: either exact for very special cases, or heuristic:

- Special case: analytical solution via Singular Value Decomposition [Eckart,Young'36]:

$$\text{minimize } \mathbf{Rank}X, \quad \text{subject to } \|X - \hat{X}\| \leq \epsilon$$

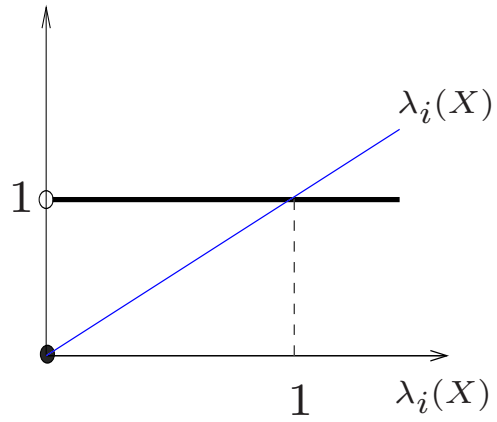
- Heuristic local methods
  - alternating projections [e.g.,Grigoriadis&Beran'00]
  - factorization, alternating LMIs [e.g.,Iwasaki'99]



- A simple convex heuristic: **Trace minimization** (when $X \succeq 0$)

Trace Minimization, $X \succeq 0$

$$\begin{array}{ll} \text{minimize} & \mathbf{Tr}\, X \\ \text{subject to} & X \in \mathcal{C}, \\ & X = X^T \succeq 0 \end{array}$$



($\mathbf{Rank} X$: # of non-zero λ_i's, $\mathbf{Tr}\, X = \sum_i \lambda_i$)

- convex problem, used often in practice [e.g.,Pare'00,Beck'96,'99]
- extension to general matrices: sum of singular values (nuclear norm)
- variation: iterative weighted trace minimization

Nuclear Norm Minimization

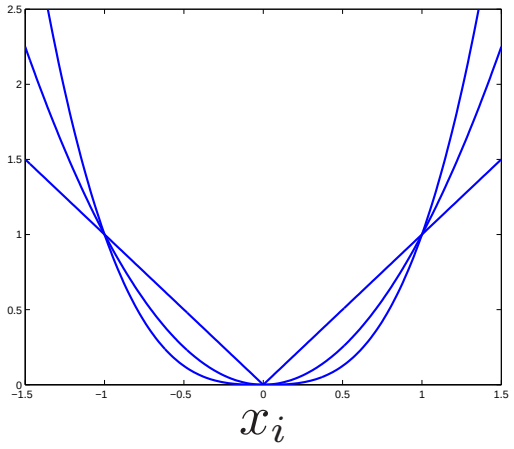
Extension: minimizing sum of singular values:

$$\begin{array}{ll} \text{minimize} & \|X\|_* \\ \text{subject to} & X \in \mathcal{C} \end{array}$$

$\|X\|_* = \sum_{i=1}^{n} \sigma_i(X)$ is called nuclear norm of X, where $\sigma_i(X) = \sqrt{\lambda_i(X^T X)}$;

dual norm of $\|X\| = \sigma_{\max}(X)$

- for $X = \mathbf{diag}(x)$, reduces to minimizing $\|x\|_1 = \sum |x_i|$; well-known ℓ_1 heuristic for finding sparse vectors
- useful in generalizing sparse recovery and ℓ_1 results to matrix rank

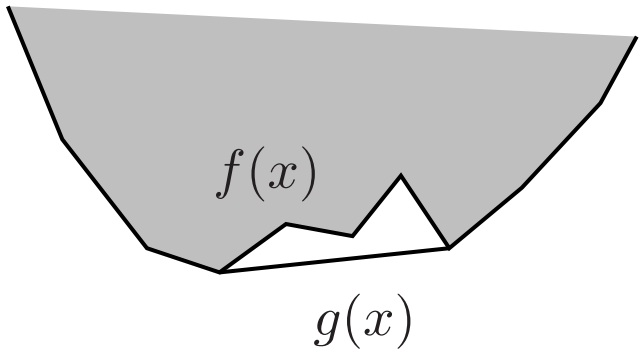


Convex Envelope of Rank

Theorem. $\|X\|_*$ is the convex envelope of $\mathbf{Rank} X$ on $\{X \in \mathbf{R}^{m\times n} \mid \|X\| \leq 1\}$.
[Fazel,Hindi,Boyd'01]

Def. *convex envelope* of $f : C \to \mathbf{R}$ is largest convex function g s.t. $g(x) \leq f(x)$ for all $x \in C$.

'best' convex lower approximation



From Compressed Sensing to Matrix Rank Minimization

Compressed sensing: a framework for measurement/recovery of **sparse** signals
[Candés,Tao'04; Donoho'04; Wakin,et al'06; Baraniuk,et al'07; many others. . .]

- signal $x \in \mathbf{R}^n$ has k nonzeros (k-sparse)
- measurements: $b = Ax, \quad A \in \mathbf{R}^{p \times n}, \quad p \ll n$
- underdetermined. . . but for some classes of random A, with very high probability k-sparse solution is unique and coincides with min ℓ_1 norm solution
- recovery: ℓ_1 minimization

From Compressed Sensing to Matrix Rank Minimization

Compressed sensing: a framework for measurement/recovery of **sparse** signals
[Candés,Tao'04; Donoho'04; Wakin,et al'06; Baraniuk,et al'07; many others. . .]

- signal $x \in \mathbf{R}^n$ has k nonzeros (k-sparse)
- measurements: $b = Ax, \quad A \in \mathbf{R}^{p \times n}, \quad p \ll n$
- underdetermined. . . but for some classes of random A, with very high probability k-sparse solution is unique and coincides with min ℓ_1 norm solution
- recovery: ℓ_1 minimization

What if object of interest is a low-rank matrix?

Examples: matrix completion from partial data (e.g., *Netflix problem*)

Hankel system identification

When does nuclear norm heuristic work?

Consider affine rank minimization problem:

$$\begin{array}{ll} \text{minimize} & \mathbf{Rank} X \\ \text{subject to} & \mathcal{A}(X) = b \end{array}$$

$\mathcal{A} : \mathbf{R}^{m \times n} \to \mathbf{R}^p$ is a linear map, $b \in \mathbf{R}^p$.

Equivalently, $\mathbf{Avec}(X) = b$, $\mathbf{A} \in \mathbf{R}^{p \times mn}$

Several key concepts from compressed sensing can be generalized:

- **restricted isometry** of $\mathcal{A}$ (this talk)
- **spherical section** property for nullspace of $\mathcal{A}$ (this talk)
- incoherence properties [Candés,Recht'08]

Nuclear norm minimization "works" under these conditions...

A dictionary:

parsimony concept	cardinality	rank
Hilbert space norm	Euclidean	Frobenius
sparsity inducing norm	ℓ_1	nuclear
dual norm	ℓ_∞	operator
convex optimization	linear programming	semidefinite programming

(Frobenius norm: $\|X\|_F = \sum_{i,j} X_{ij}^2 = \sum_i \sigma_i^2(X)$)

general "recipe":

- give a deterministic condition on $\mathcal{A}$ for heuristic to give exact solution
- sometimes condition may be hard to check. . .
- invoke randomness of problem data: random $\mathcal{A}$ satisfy the condition *with high probability*

Restricted Isometry Property (RIP)

Quantify behavior of $\mathcal{A}$ when restricted to set $\{X \text{ s.t. } \mathbf{Rank} X \leq r\}$

Definition. For linear map $\mathcal{A} : \mathbf{R}^{m \times n} \to \mathbf{R}^p$, the restricted isometry constant δ_r is the smallest number such that

$$1 - \delta_r \leq \frac{\|\mathcal{A}(X)\|}{\|X\|_F} \leq 1 + \delta_r$$

holds for all matrices X of rank up to r.

- vector case: bounds on $\|Ax\|/\|x\|$ for all k-sparse x
- means: all submatrices of A with k columns are "well-conditioned"
- matrix case: RIP is more complicated to interpret. . .

Two Recovery Results (using RIP)

Let $\mathbf{Rank} X_0 = r$ and $b = \mathcal{A}(X_0)$, and

$$\hat{X} := \arg\min_X \|X\|_* \qquad \text{subject to} \quad \mathcal{A}(X) = b$$

Theorem 1. (uniqueness) *If $\delta_{2r} < 1$, X_0 is the only matrix of rank at most r satisfying $\mathcal{A}(X) = b$.*

Two Recovery Results (using RIP)

Let $\mathbf{Rank}X_0 = r$ and $b = \mathcal{A}(X_0)$, and

$$\hat{X} := \arg\min_X \|X\|_* \qquad \text{subject to} \quad \mathcal{A}(X) = b$$

Theorem 1. (uniqueness) *If $\delta_{2r} < 1$, X_0 is the only matrix of rank at most r satisfying $\mathcal{A}(X) = b$.*

Proof: if there is an $X \neq X_0$ with $\mathcal{A}(X) = b$ and $\mathbf{Rank}X = r$, then $\mathcal{A}(X - X_0) = 0$. but $X_0 - X$ has rank at most $2r$, contradicting $\delta_{2r} < 1$.

Two Recovery Results (using RIP)

Let $\mathbf{Rank} X_0 = r$ and $b = \mathcal{A}(X_0)$, and

$$\hat{X} := \arg\min_X \|X\|_* \qquad \text{subject to} \quad \mathcal{A}(X) = b$$

Theorem 1. (uniqueness) *If $\delta_{2r} < 1$, X_0 is the only matrix of rank at most r satisfying $\mathcal{A}(X) = b$.*

Theorem 2. (exact recovery) *If $\delta_{5r} < 1/10$, then $\hat{X} = X_0$.*

What linear maps satisfy the RIP? iid Gaussian, iid Bernoulli,. . .

Guaranteed minimum rank solution via nuclear norm

Theorem. *Fix $0 < \delta < 1$; pick $\mathbf{A} \in \mathbf{R}^{p \times mn}$ "randomly" (e.g., iid Gaussian). For any rank r there exist constants $c_0, c_1 > 0$ such that $\delta_r \leq \delta$ with probability at least $1 - e^{-c_1 p}$, if*

$$p \geq c_0 \underbrace{r(m+n)}_{\text{low-rank dim}} \log \underbrace{(mn)}_{\text{ambient dim}}$$

meaning: given p random constraints on low-rank matrix X_0,

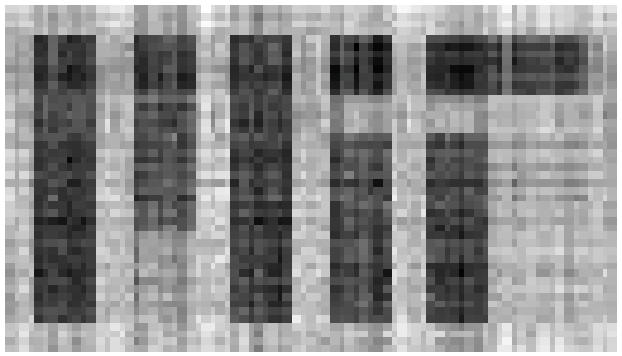
$$\mathbf{A}_{p \times mn} \mathbf{vec}(X_0) = b$$

nuclear norm heuristic finds the correct solution with very high probability! (c_0, c_1 depend only on δ)

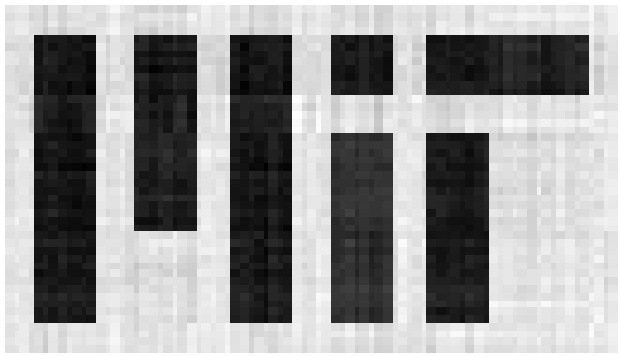
[Recht,Fazel,Parrilo'07]

Numerical Example

total pixels=46x81=3726, rank=5



700 constraints

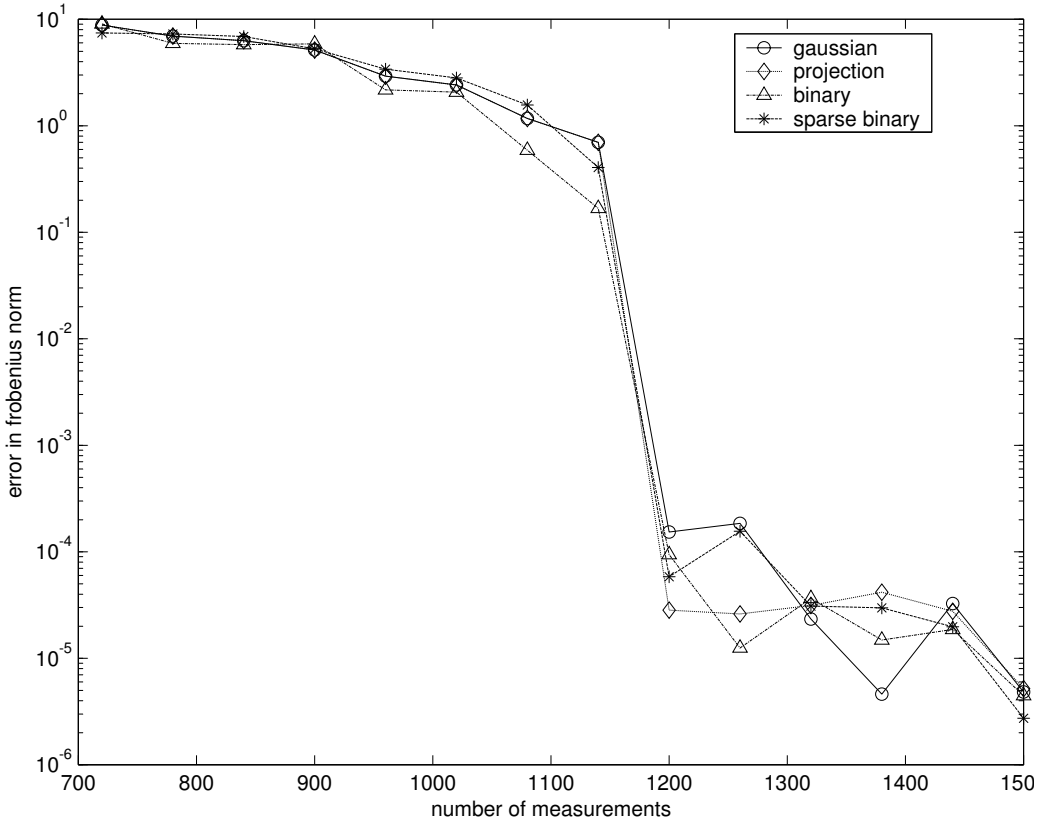


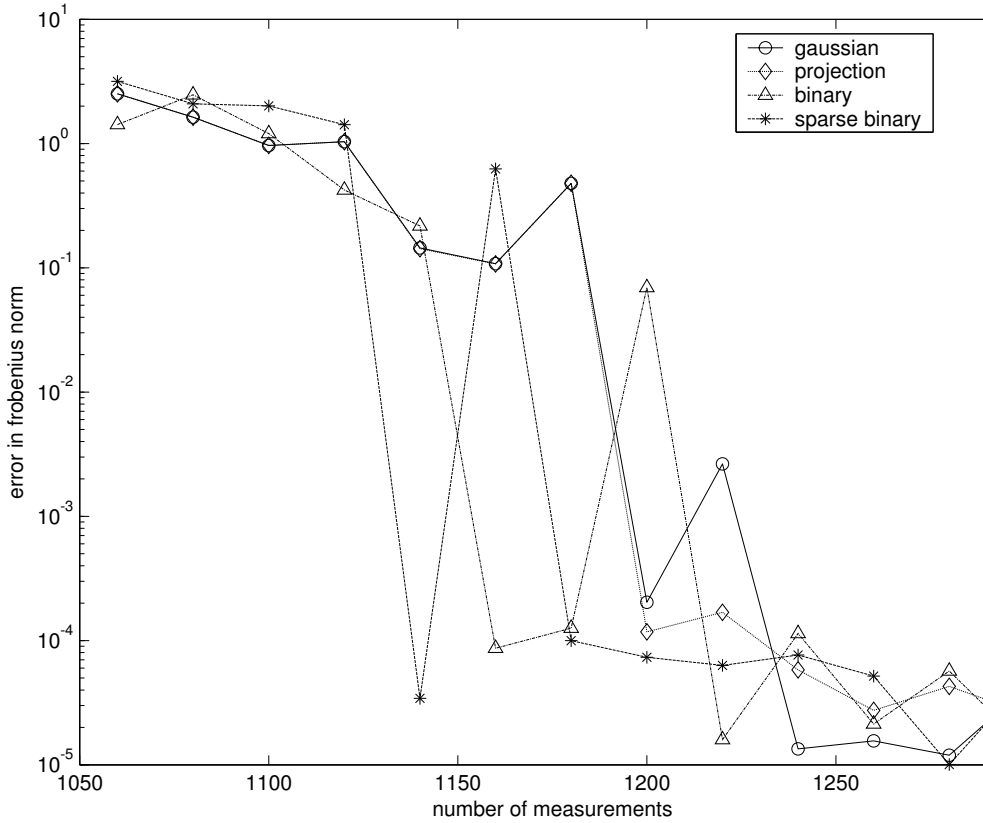
1100 constraints

1250 constraints

- random Gaussian measurements
- solve SDP (just using SeDuMi here)

Recovery errors:



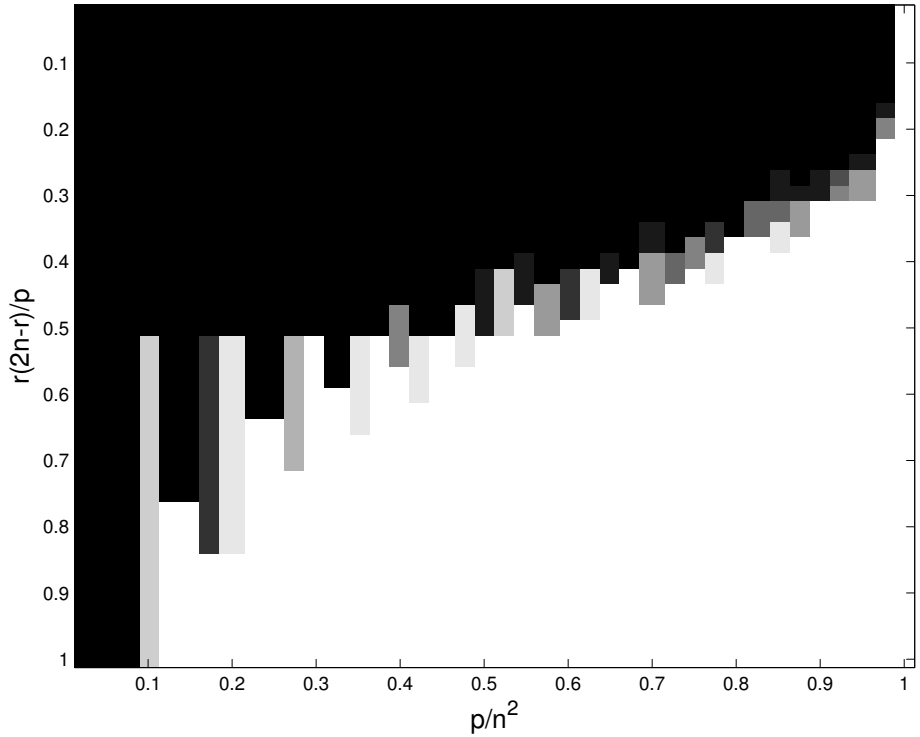


- sharp transition near 1200 measurements ($\approx 2r(m+n-r)$)
- theoretical bound is conservative

Phase transition: series of experiments for various n, r, p

$n = 40$, p runs from 0 to n^2. for fixed (n, p), r covers all values satisfying $r(2n - r) \leq p$. for each (n, p, r), generate 10 random cases for Y_0, solve

$$\begin{array}{ll} \text{min} & \|X\|_* \\ \text{subject to} & \mathbf{Avec}(X) = \mathbf{Avec}(Y_0) \end{array}$$



(analogous to [Donoho,Tanner'05])

Noise and approximation error bounds

Consider noisy constraints and approximately low-rank X

$$y = \mathcal{A}(X) + z, \qquad \|z\|_2 \leq \beta$$

Theorem. *If $\delta_{5r} < 1/10$, then*

$$\|\hat{X} - X\|_F \leq \frac{c_0}{\sqrt{r}}\|X - X_r\|_* + c_1\beta$$

and if $\beta = 0$, we also have

$$\|\hat{X} - X\|_* \leq c_2\|X - X_r\|_*$$

where X_r is the sum of the first r SVD terms of X, and $c_0,\ c_1,\ c_2$ are constants (that depend only on the isometry constants).

[Fazel,Candès,Recht,Parrilo'08]

Comments about RIP

- RIP not invariant w.r.t multiplying with an invertible map:

  $\mathcal{G}\mathcal{A}(X) = \mathcal{G}b$, but RIP constants of $\mathcal{A}$ and $\mathcal{G}\mathcal{A}$ are different

- intuitively: ability to recover X from $\mathcal{A}(X)$ depends only on nullspace

- bound on measurements can be very loose... e.g., $m = n$, $r = \alpha n$, $\Rightarrow n^2 \log n$

Comments about RIP

- RIP not invariant w.r.t multiplying with an invertible map:

  $\mathcal{G}\mathcal{A}(X) = \mathcal{G}b$, but RIP constants of $\mathcal{A}$ and $\mathcal{G}\mathcal{A}$ are different

- intuitively: ability to recover X from $\mathcal{A}(X)$ depends only on nullspace

- bound on measurements can be very loose... e.g., $m = n$, $r = \alpha n$, $\Rightarrow n^2 \log n$

direct **conditions on the nullspace:**

- a (nec. and suff.) property [Recht,Xu,Hassibi'08]: for all elements with 'rank r plus higher rank' decomposition, nuclear norm of low-rank part is smaller

- we consider simpler sufficient condition: *spherical section property*

Matrix spherical section property

Definition. a subspace $V \subset \mathbf{R}^{m\times n}$ satisfies **Δ-spherical section property** if

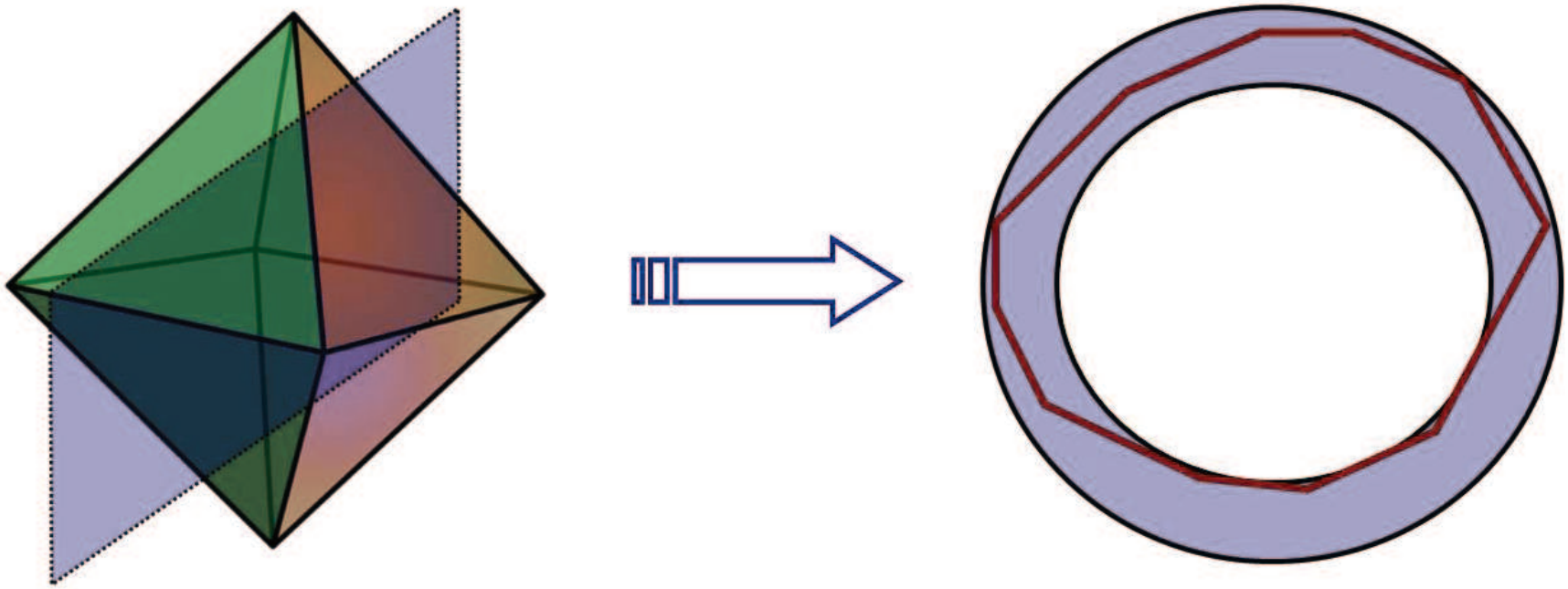
$$\frac{\|Z\|_*}{\|Z\|_F} \geq \sqrt{\Delta}, \quad \text{for all } Z \in V,\ Z \neq 0$$

- Δ lower bounds $\mathbf{Rank}Z$: Δ large $\Rightarrow$ V doesn't include low rank matrices

- vector case: for subspaces of $\mathbf{R}^n$ [Kashin'77],[Gluskin,Garnaev'84],. . .
- arises in different areas: low-dim embedding, n-widths in approximation theory, etc.
- used in compressed sensing e.g., [Kashin,Temlyakov'07],[Zhang'08]

vector case: consider $\frac{\|x\|_1}{\|x\|_2} \geq \sqrt{\Delta}$

Δ large means if unit ℓ_1 norm ball is cut by subspace V, the intersection looks **"spherical"** (ℓ_2 norm is not too big)



intuitively: random subspaces should have large Δ. . .

Exact recovery results

Let $r = \mathbf{Rank} X_0$, $b = \mathcal{A}(X_0)$, $m = \min\{m, n\}$. Consider relaxation

$$\hat{X} := \arg\min_X \|X\|_* \qquad \text{subject to} \quad \mathcal{A}(X) = b$$

Let nullspace have the Δ-spherical section property.

Exact recovery results

Let $r = \mathbf{Rank} X_0$, $b = \mathcal{A}(X_0)$, $m = \min\{m, n\}$. Consider convex problem

$$\hat{X} := \arg\min_X \|X\|_* \qquad \text{subject to} \quad \mathcal{A}(X) = b$$

Let nullspace have the Δ-spherical section property.

Theorem. (uniqueness) *If $r < \frac{\Delta}{2}$, X_0 is the only matrix of rank at most r satisfying the constraints.*

Exact recovery results

Let $r = \mathbf{Rank} X_0$, $b = \mathcal{A}(X_0)$, $m = \min\{m, n\}$. Consider relaxation

$$\hat{X} := \arg\min_X \|X\|_* \qquad \text{subject to} \quad \mathcal{A}(X) = b$$

Let nullspace have the Δ-spherical section property.

Theorem. (uniqueness) *If $r < \frac{\Delta}{2}$, X_0 is the only matrix of rank at most r satisfying the constraints.*

Theorem. (exact recovery) *If $r < \frac{\Delta}{6}$ then $\hat{X} = X_0$.*

more generally: *If* $r < \min\{\frac{3m}{4} - \sqrt{\frac{9m^2}{16} - \frac{m}{4\Delta}}, \frac{m}{2}\}$, *then* $\hat{X} = X_0$.

[Dvijotham,Fazel'09]

proof outline: (generalize idea from [Vavasis'09])

$$\|X_0\|_* \geq \|X_0 + Z\|_* = \| \begin{bmatrix} \Sigma + \bar{Z}_{11} & \bar{Z}_{12} \\ \bar{Z}_{21} & \bar{Z}_{22} \end{bmatrix} \|_* \quad \geq \quad \|\Sigma + \bar{Z}_{11}\|_* + \|\bar{Z}_{22}\|_*$$

$$\geq \quad \|\Sigma\|_* - \|\bar{Z}_{11}\|_* + \|\bar{Z}_{22}\|_*$$

so "bad" Z's in nullspace satisfy $\|\bar{Z}_{22}\|_* \leq \|\bar{Z}_{11}\|_*$

now if Δ-spherical section rules out these Z's for certain rank r, any matrix up to that rank can be recovered. . .

consider

$$\begin{array}{ll} \text{maximize} & \frac{\|Z\|_*}{\|Z\|_F} \\ \text{subject to} & \|Z_{22}\|_* \leq \|Z_{11}\|_* \end{array}$$

where $Z_{11} \in \mathbf{R}^{r \times r}$, solve analytically (by symmetry and duality).

Approximately low rank matrices

Theorem. *Let $X \in \mathbf{R}^{m \times n}$ be a general matrix (not necessarily low rank), and $1 \leq r \leq m$. If*

$$r < min\left\{\frac{m}{6}, \frac{\Delta}{24}\right\},$$

then

$$\|\hat{X} - X\|_* \leq 4\|X - X_r\|_*.$$

where X_r is the sum of the first r SVD terms of X.

Approximately low rank matrices

Theorem. *Let $X \in \mathbf{R}^{m \times n}$ be a general matrix (not necessarily low rank), and $1 \le r \le m$. If*

$$r < min\left\{\frac{m}{6}, \frac{\Delta}{24}\right\},$$

then

$$\|\hat{X} - X\|_* \le 4\|X - X_r\|_*.$$

where X_r is the sum of the first r SVD terms of X.

more generally: if $r < \min\left\{\frac{c-2}{c+2}\frac{m}{2}, \frac{\Delta(c-2)^2}{6c^2}, \right\}$

then $\|\hat{X} - X\|_* \le c\|X - X_r\|_*.$

When does spherical section property hold? iid Gaussian. Others?

[Dvijotham,Fazel'09]

Probabilistic result for square matrices

if $\mathcal{A}$ from random Gaussian ensemble, $p = \mu n^2$, all X of rank up to αn are recovered with probability at least

$$1 - \exp\left(-\frac{\kappa n^2}{2(1+\sqrt{\alpha})^2}\right),$$

if $c - \sqrt{\alpha} - \kappa \geq 0$, and

$$\mu > 1 - \left(\frac{c - \sqrt{\alpha} - \kappa}{1+\sqrt{\alpha}}\right)^2.$$

for **large** n, $c \approx 8/3\pi \approx 0.85$.

(can also get the general case)

for general (non-square) matrices:

$m = \gamma n$ (γ=*aspect ratio*), all X of rank up to $\frac{\alpha}{6n}$ are recovered with same probability if $c\phi(\gamma) - \sqrt{\alpha\gamma} - \kappa \geq 0$, and

$$\mu > 1 - \left(\frac{c\phi(\gamma) - \sqrt{\alpha\gamma} - \kappa}{1 + \sqrt{\alpha}} \right)^2,$$

(where $\phi(\gamma)$ comes from integrating the distribution of singular values of Gaussian matrices, explicitly known [Bai'99]).

Algorithms for minimizing $\|X\|_*$

Semidefinite program and its dual:

$$\begin{array}{ll} \min\limits_{X,Y,Z} & \mathbf{Tr}\, Y + \mathbf{Tr}\, Z \\ \text{s.t.} & \begin{bmatrix} Y & X \\ X^T & Z \end{bmatrix} \succeq 0 \\ & \mathcal{A}(X) = b \end{array} \qquad \begin{array}{ll} \max\limits_{z} & b^T z \\ \text{s.t.} & \begin{bmatrix} I_m & \mathcal{A}^*(z) \\ \mathcal{A}^*(z)^T & I_n \end{bmatrix} \succeq 0. \end{array}$$

- custom interior point methods (e.g., [Liu,Vandenberghe'08])
- subgradient methods, proximal gradients (e.g., [Ma,Goldfarb,Chen'09])
- singular value thresholding (SVT) [Cai,Candés,Shen'08]
- low-rank factorization (e.g., SDPLR [Burer,Monteiro'05])

- other (non-SDP) algorithms: greedy algorithms [Lee,Bresler'09], special case algorithms for matrix completion [Keshavan,Oh,Montanari'09],. . .

Conclusions

- Rank minimization problem is NP-hard in general; many applications
- Convex heuristic: Nuclear norm minimization, variations
- For affine rank minimization with certain (random) constraints: theoretical guarantees for exact solution
- A rich generalization of vector sparsity and compressed sensing theory; has opened the door to new set of applications and new links between areas

Conclusions

A rich generalization of vector sparsity and compressed sensing theory; has opened the door to new set of applications and new links between areas

- low-rank matrix completion: e.g., recommendation systems
  [Candès,Recht'08;Candès,Tao'09;Keshavan,et al'09]
- low-rank+sparse decompositions: e.g., graphical models; matrix rigidity theory
  [Chandrasekaran,et al'09]
- graph problems: e.g., some max-clique problems [Ames,Vavasis'09]
- Hankel rank minimization and system identification
- ...and more to come!

Future directions

extend nullspace analysis:

- analyze "structured" versions of $\mathcal{A}$ arising in practice
- non-Gaussian ensembles; tighter bounds
- robustness to noise
- probabilistic analysis for the PSD case

- Algorithms for nuclear norm minimization, large scale
- Other notions of parsimony
- New applications

Acknowledgements

Student: Dvijotham Krishnamurthy (U Washington)

Collaborators: Pablo Parrilo (MIT), Ben Recht (Caltech/U Wisconsin), Emmanuel Candes (Caltech), Stephen Boyd (Stanford), Haitham Hindi (PARC)

Funding: NSF