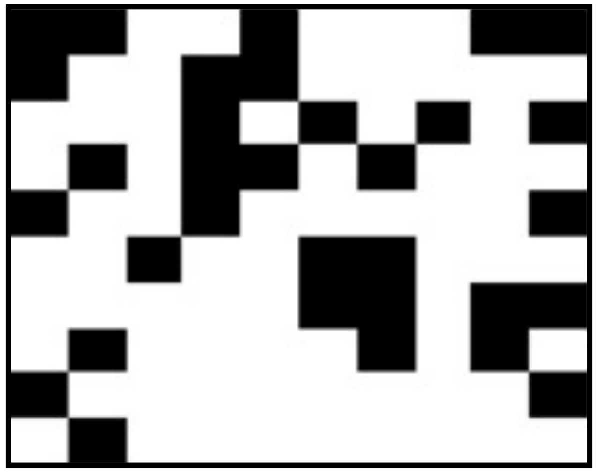
From Compressed Sensing to Matrix Completion and Beyond

Benjamin Recht
Department of Computer Sciences
University of Wisconsin-Madison

Abstract Setup: Matrix Completion

M =



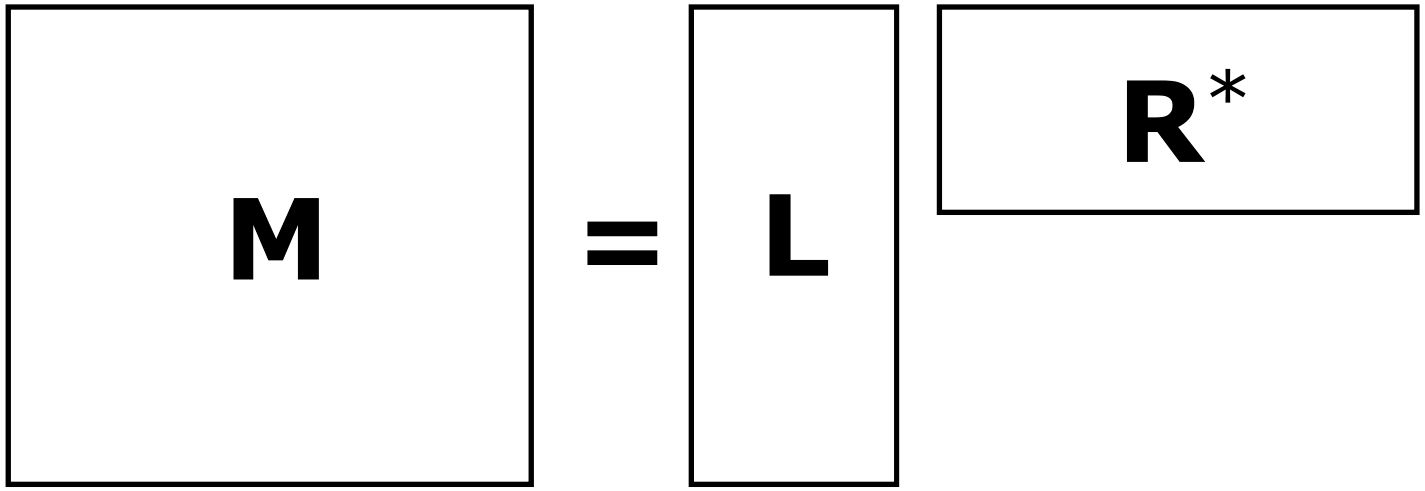
M_{ij} known for black cells

M_{ij} unknown for white cells

Rows index movies

Columns index users

- How do you fill in the missing data?



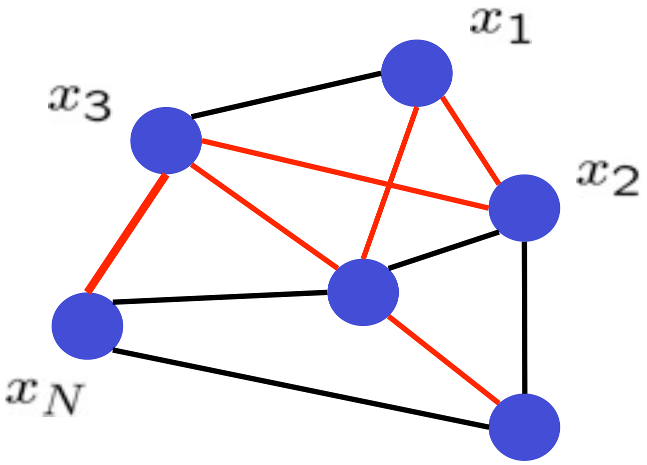
k x n   k x r   r x n

kn entries   r(k+n) entries

Recommender Systems

Euclidean Embedding

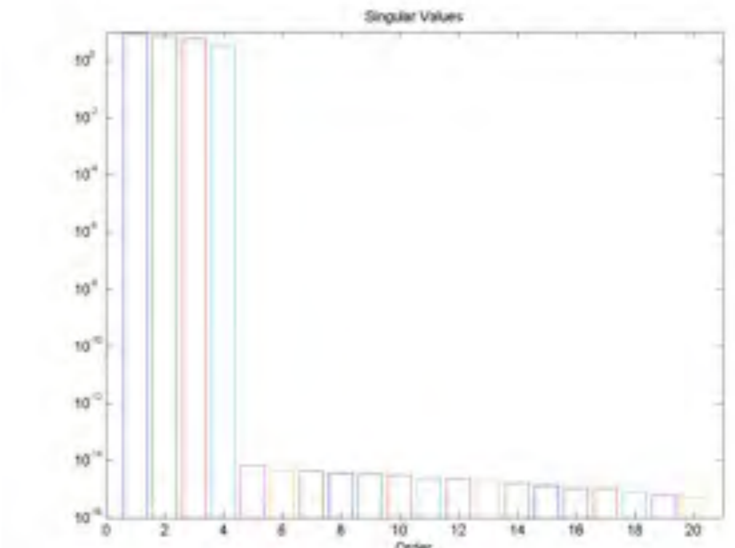
Multitask Learning

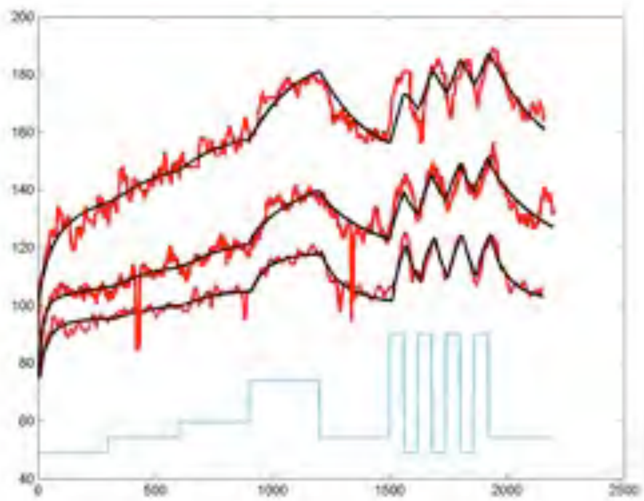


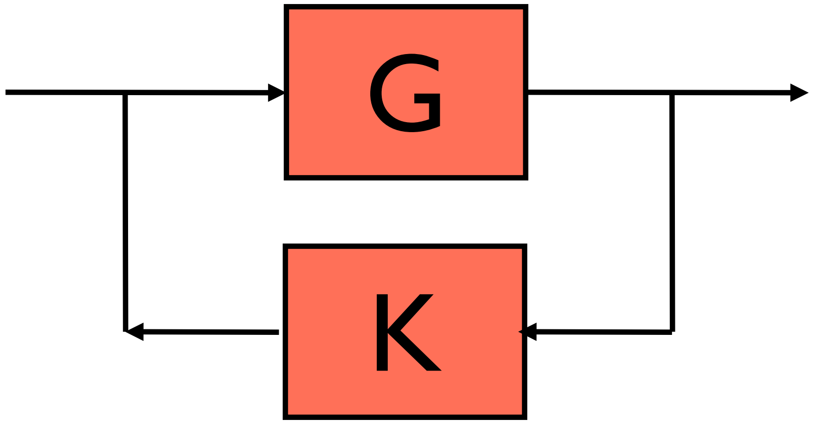
Rank of: Data Matrix

Gram Matrix

Matrix of Classifiers







Model Reduction

System Identification

Controller Design

Constraints involving the rank of the Hankel Operator, Matrix, or Singular Values

Affine Rank Minimization

- **PROBLEM:** Find the matrix of lowest rank that satisfies/approximates the underdetermined linear system

$$\Phi(X) = y \qquad \Phi : \mathbb{R}^{k \times n} \to \mathbb{R}^m$$

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & \Phi(X) = y \end{array}$$

- **NP-HARD:**
  - Reduce to MAXCUT
  - Hard to approximate
  - Exact algorithms are awful

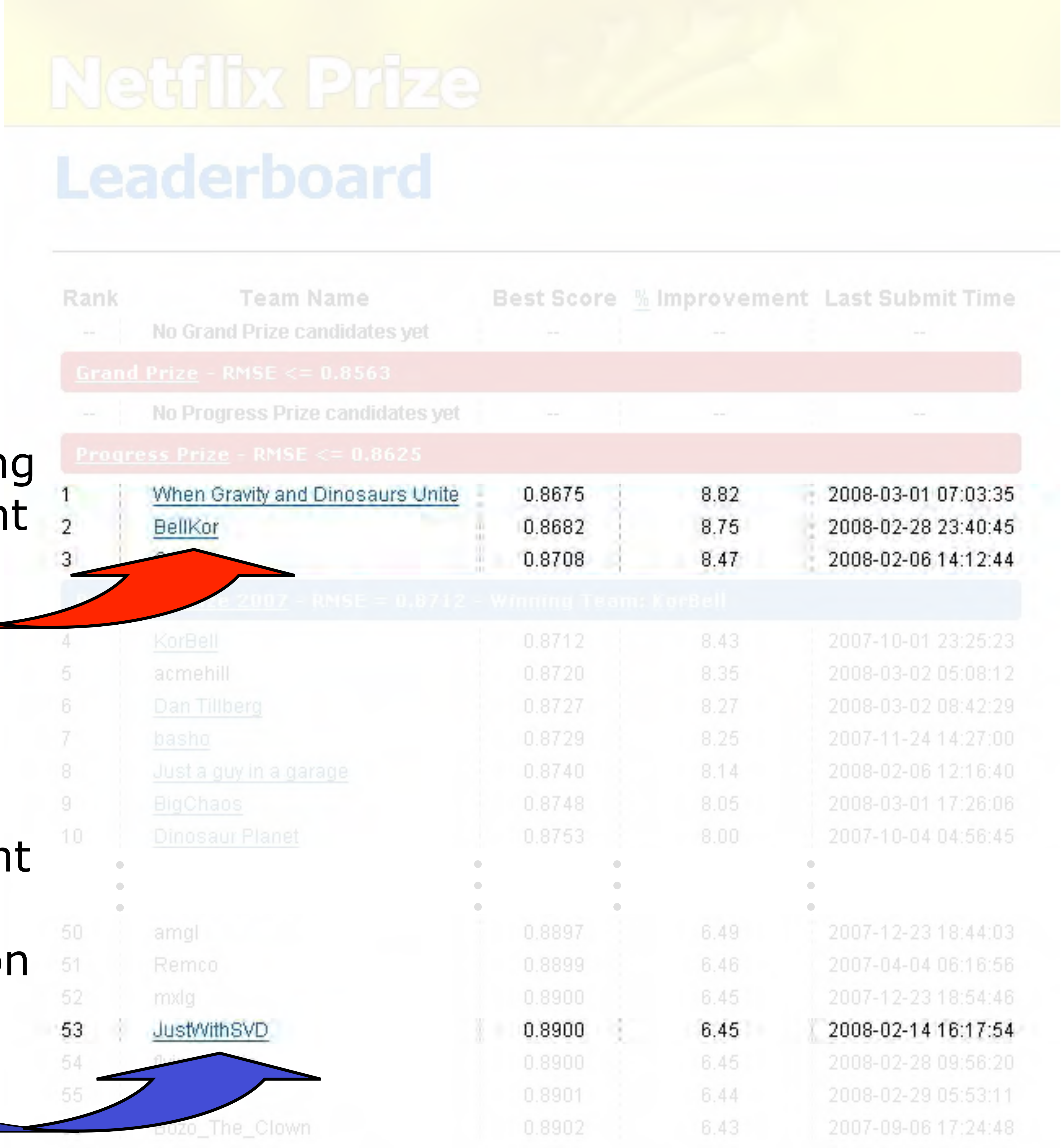
Heuristic: Gradient Descent

$$\text{minimize} \ \sum_{i=1}^{k}\sum_{a=1}^{r} L_{ia}^2 + \sum_{j=1}^{n}\sum_{a=1}^{r} R_{ja}^2 + \lambda \sum_{i,j}\left(\sum_{k} L_{ik}R_{jk} - M_{ij}\right)^2$$

- Just run gradient descent
- λ determines tradeoff between satisfying constraints and the size of the factors

Mixture of hundreds of models, including gradient descent

Gradient descent on low-rank parameterization



Netflix Prize

Leaderboard

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8675	8.82	2008-03-01 07:03:35
2	BellKor	0.8682	8.75	2008-02-28 23:40:45
3	[illegible]	0.8708	8.47	2008-02-06 14:12:44
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	acmehill	0.8720	8.35	2008-03-02 05:08:12
6	Dan Tillberg	0.8727	8.27	2008-03-02 08:42:29
7	basho	0.8729	8.25	2007-11-24 14:27:00
8	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
9	BigChaos	0.8748	8.05	2008-03-01 17:26:06
10	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
⋮		⋮	⋮	⋮
50	amgl	0.8897	6.49	2007-12-23 18:44:03
51	Remco	0.8899	6.46	2007-04-04 06:16:56
52	mxlg	0.8900	6.45	2007-12-23 18:54:46
53	JustWithSVD	0.8900	6.45	2008-02-14 16:17:54
54	[illegible]	0.8900	6.45	2008-02-28 09:56:20
55	[illegible]	0.8901	6.44	2008-02-29 05:53:11
[illegible]	Bozo_The_Clown	0.8902	6.43	2007-09-06 17:24:48

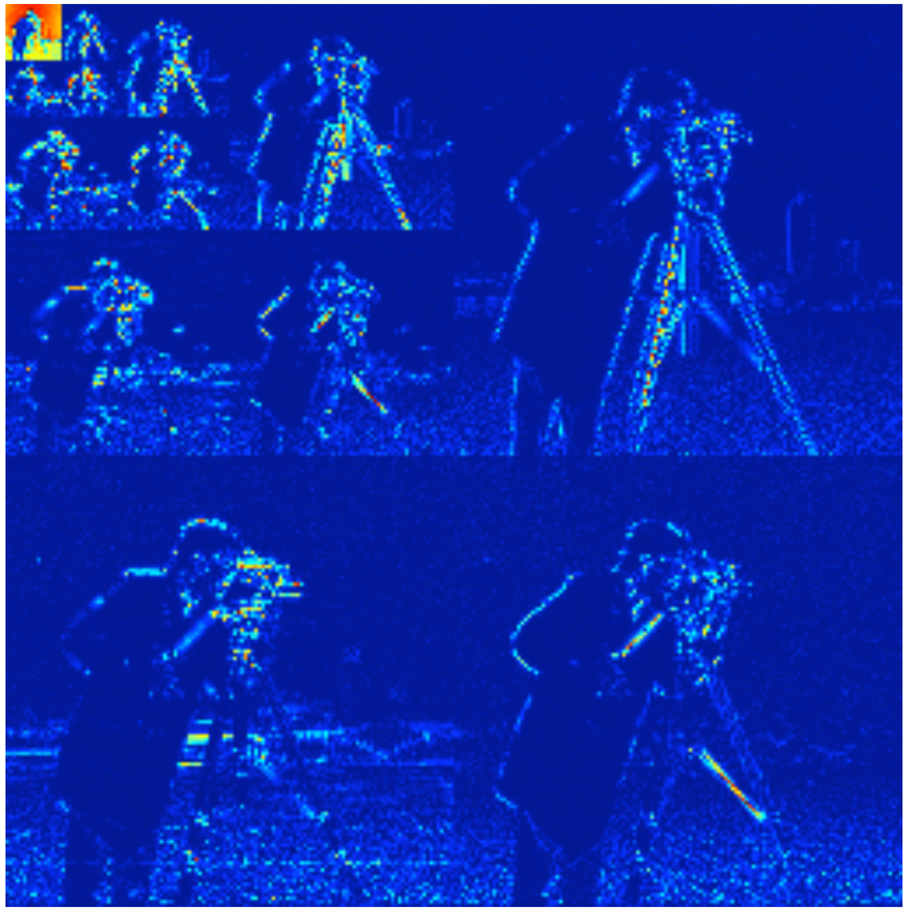
Low-rank Matrix Completion

- **PROBLEM:** Find the matrix of lowest rank has the specified entries

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\mathbf{X}) \\ \text{subject to} & X_{ij} = M_{ij} \quad \forall\ (i,j) \in \Omega \end{array}$$

- **When is this problem easy?**
  - Which algorithms?
  - Which sampling sets?
  - Which low-rank matrices?

Compressed Sensing



- Model: most of the energy is in few wavelet coefficients
- Use the fact that the image is sparse in wavelet basis to reduce number of measurements required for signal acquisition.
- ***decode using l_1 minimization***

Cardinality Minimization

- **PROBLEM:** Find the vector of lowest cardinality that satisfies/approximates the underdetermined linear system

$$\Phi x = y \qquad \Phi : \mathbb{R}^n \to \mathbb{R}^m$$

- **NP-HARD:**
  - Reduce to EXACT-COVER [Natarajan 1995]
  - Hard to approximate
  - Known exact algorithms require enumeration

- **HEURISTIC:** Replace cardinality with l_1 norm
- *Compressed Sensing*

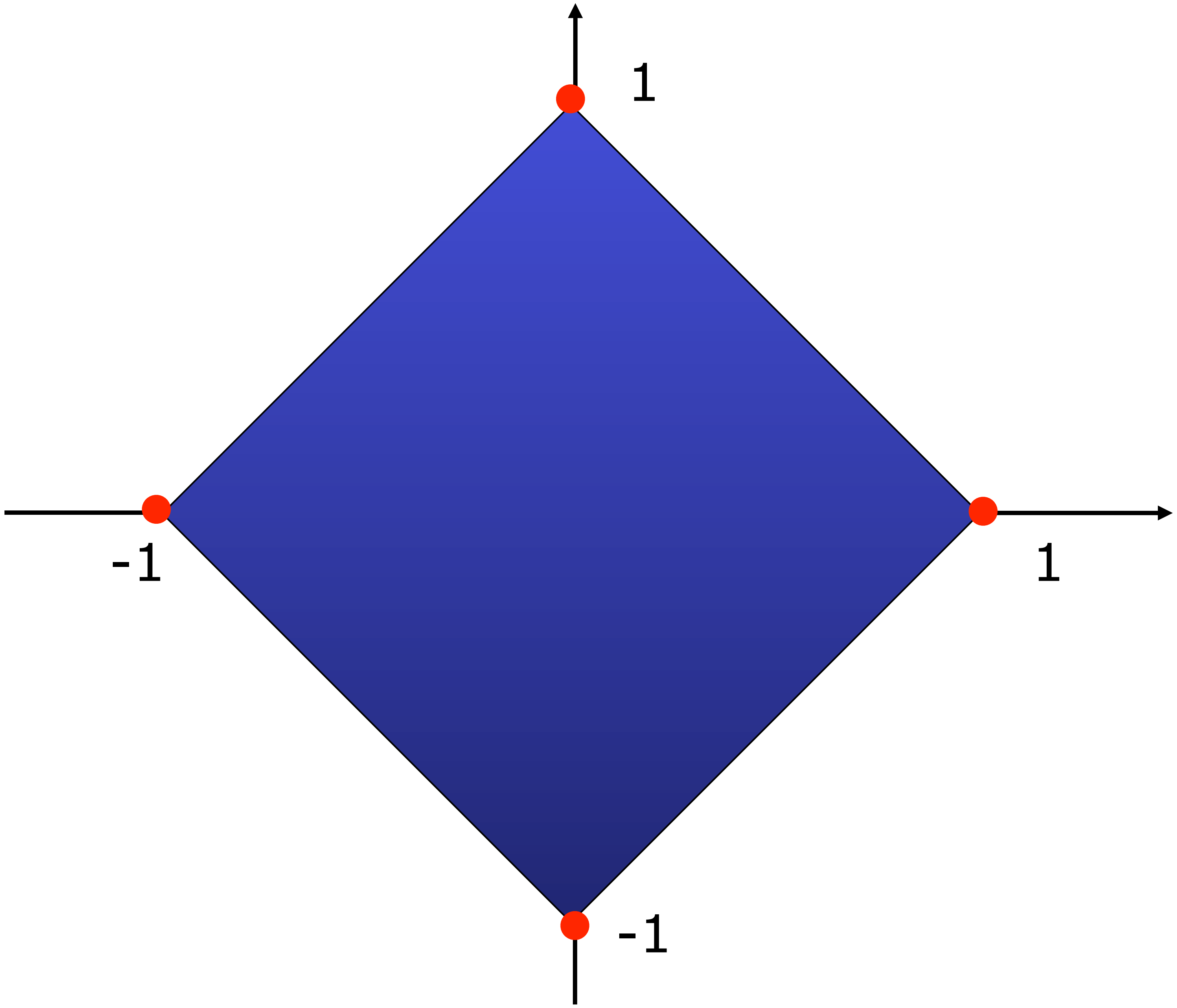
Sparsity

- 1-sparse vectors of Euclidean norm 1

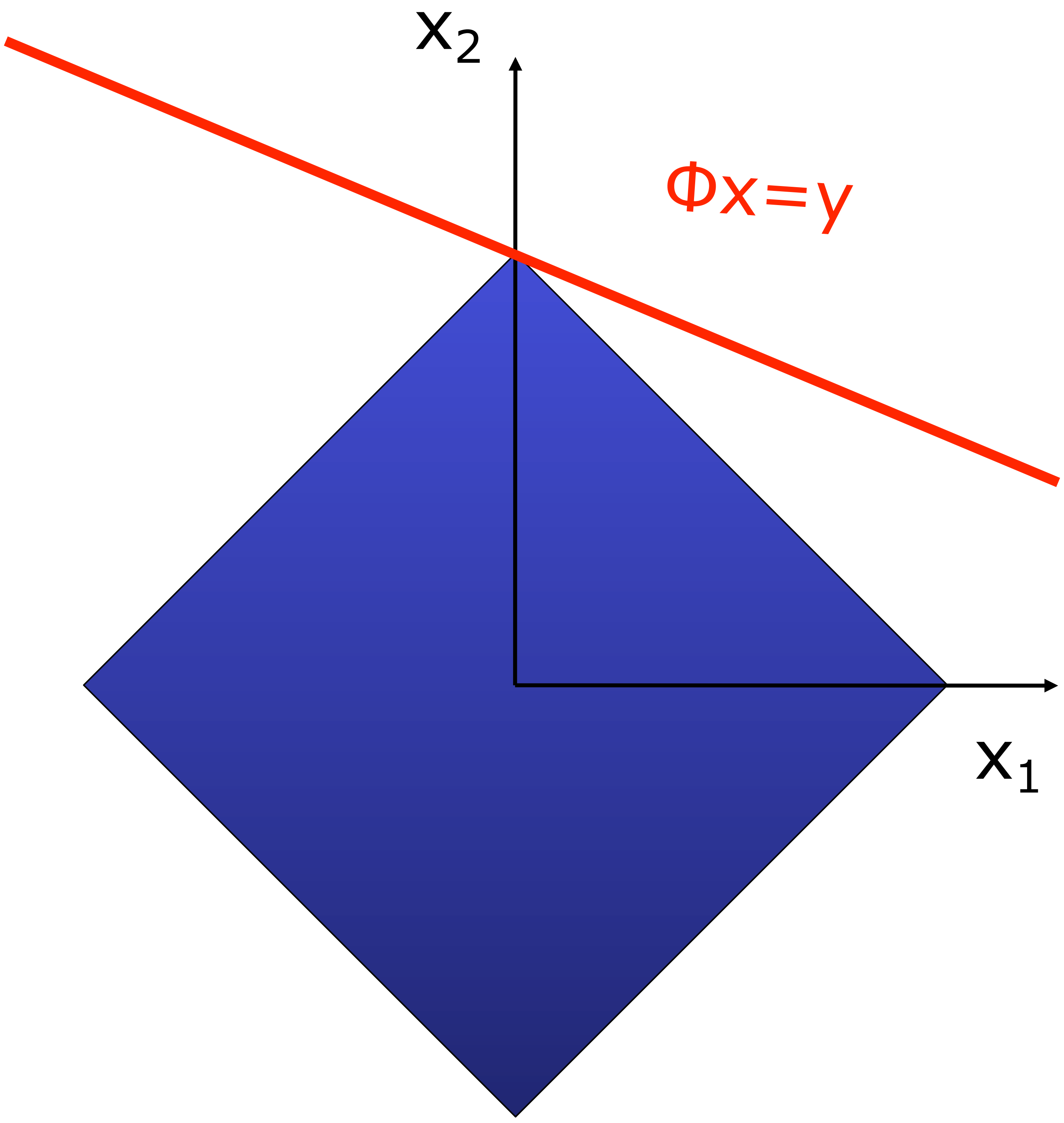
- Convex hull is the unit ball of the l_1 norm

$$\{x \ : \ \|x\|_1 \leq 1\}$$

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$



$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & \Phi x = y \end{array}$$



Compressed Sensing: Candes, Romberg, Tao, Donoho, Tanner, Etc...

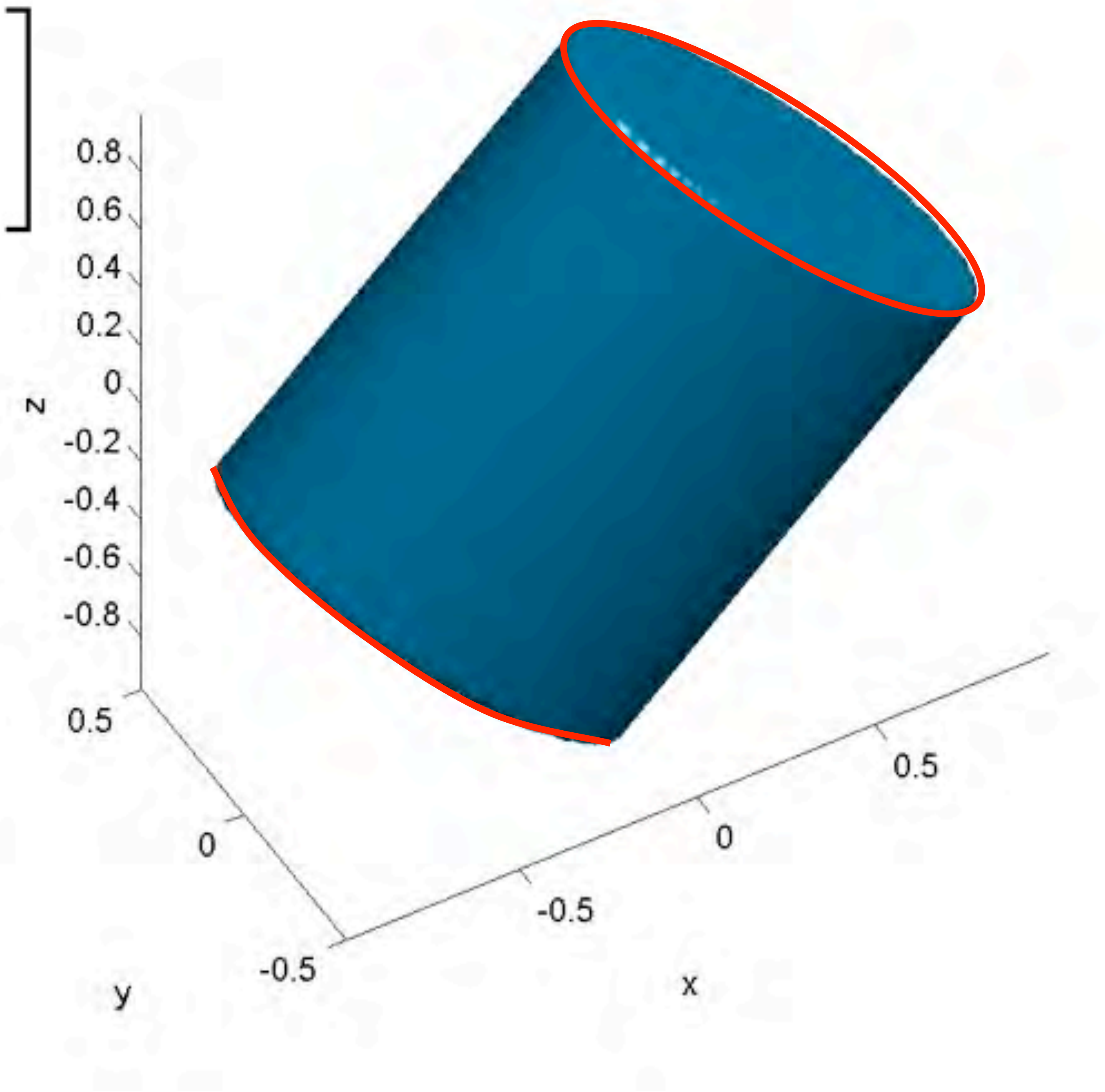
Rank

- 2x2 matrices
- plotted in 3d

$$\begin{bmatrix} x & y \\ y & z \end{bmatrix}$$

— rank 1

$x^2 + z^2 + 2y^2 = 1$



Convex hull:

$$\{X \ : \ \|X\|_* \leq 1\}$$

$$\|X\|_* = \sum_i \sigma_i(X)$$

Which Algorithm?

Affine Rank Minimization:

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(X) \\ \text{subject to} & \Phi(X) = y \end{array}$$

Convex Relaxation:

$$\begin{array}{ll} \text{minimize} & \|X\|_* = \sum_{i=1}^{k} \sigma_i(X) \\ \text{subject to} & \Phi(X) = y \end{array}$$

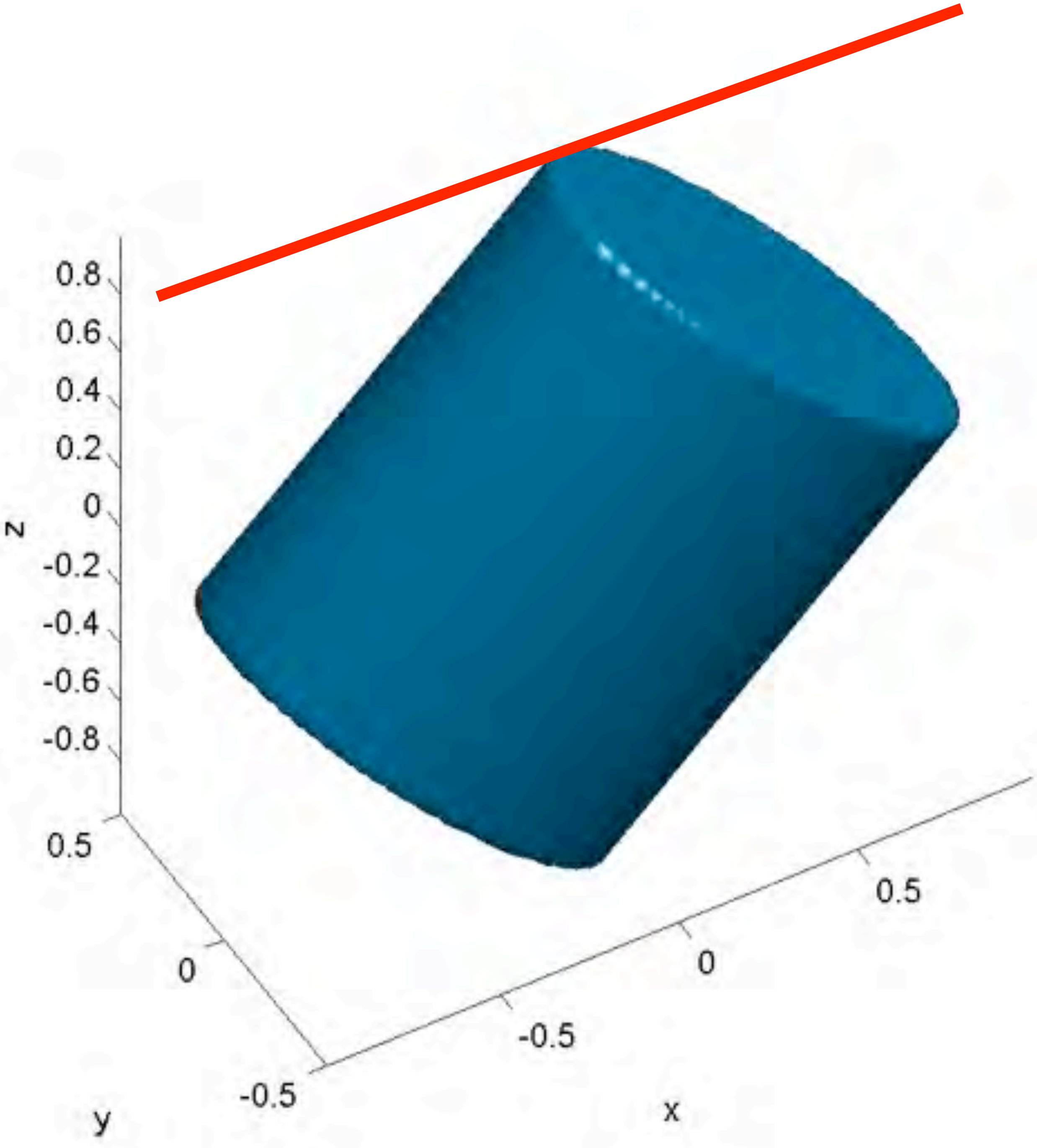
- *Nuclear Norm Heursistic.* Proposed by Fazel (2002).
- Nuclear norm is the "numerical rank" in numerical analysis
- The "trace heuristic" from controls if **X** is p.s.d.

- 2x2 matrices
- plotted in 3d

$$\left\| \begin{bmatrix} x & y \\ y & z \end{bmatrix} \right\|_* \leq 1$$

$$\|X\|_* = \sum_i \sigma_i(X)$$

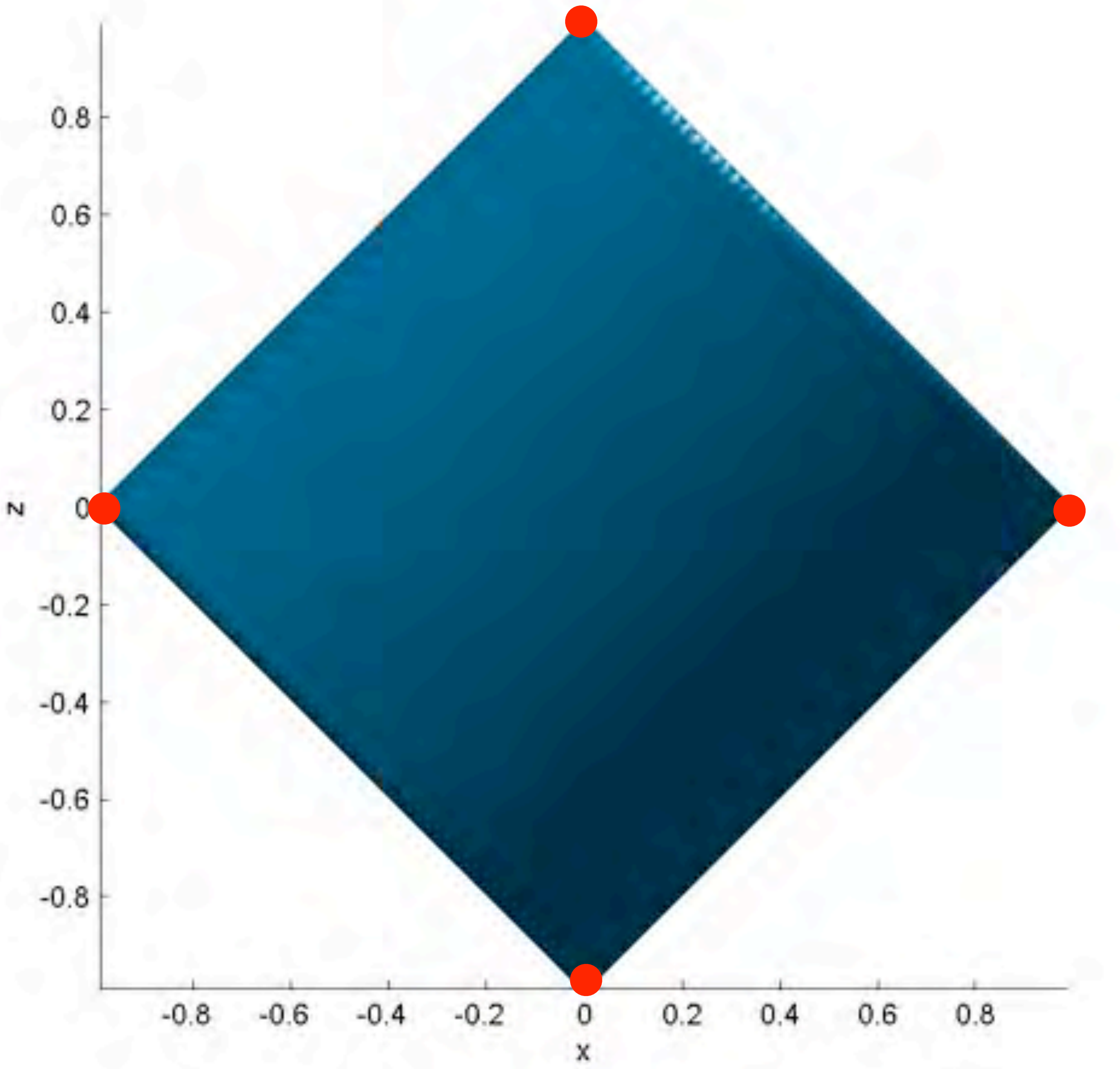
Nuclear Norm Heuristic



- 2x2 matrices
- plotted in 3d

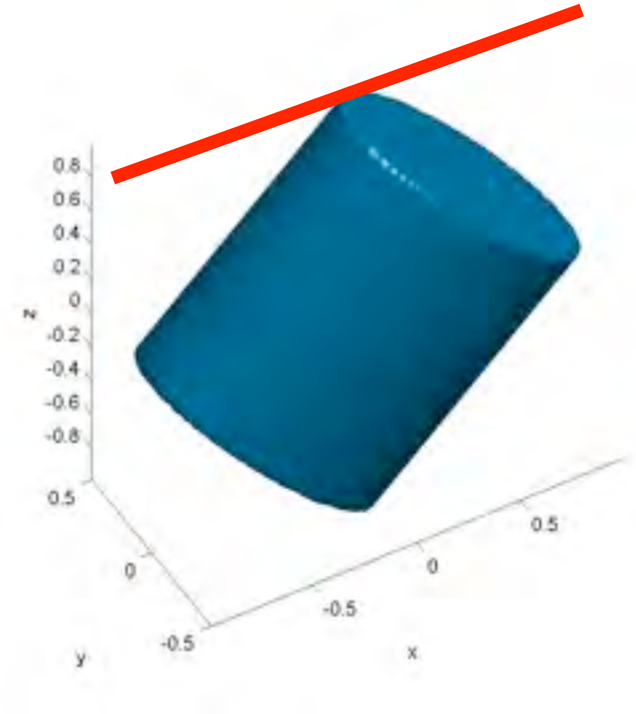
$$\left\| \begin{bmatrix} x & 0 \\ 0 & z \end{bmatrix} \right\|_* \leq 1$$

- Projection onto x-z plane is l_1 ball



Nuclear Norm minimization

$$\begin{array}{ll} \text{minimize} & \|X\|_* = \sum_{i=1}^k \sigma_i(X) \\ \text{subject to} & \Phi(X) = y \end{array}$$



Low-rank parameterization

$$X = U\Sigma V^*$$

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}(\|L\|_F^2 + \|R\|_F^2) \\ \text{subject to} & \Phi(LR^*) = y \end{array}$$

$$L = U\Sigma^{1/2}$$

$$R = V\Sigma^{1/2}$$

Method of Multipliers

"The Blog Heuristic"

$$\text{minimize} \sum_{i=1}^{k}\sum_{a=1}^{r} L_{ia}^2 + \sum_{j=1}^{n}\sum_{a=1}^{r} R_{ja}^2 + \lambda\|\Phi(LR^*) - y\|_2^2$$

First theory result

$$\Phi(X) = y \qquad \Phi : \mathbb{R}^{k\times n} \to \mathbb{R}^m$$

- If m > $c_0 r(k+n-r)\log(kn)$, the heuristic succeeds for most maps Φ.

Recht, Fazel, and Parrilo. 2007.

- Number of measurements c_0 $r(k+n-r)$ $\log(kn)$

constant

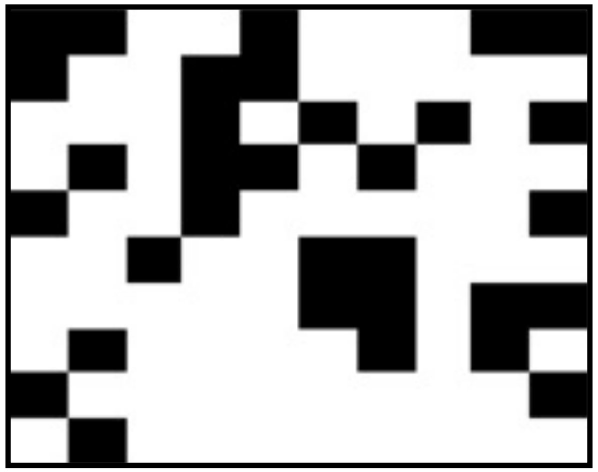
intrinsic dimension

ambient dimension

- **Approach:** Show that a random Φ is nearly an isometry on the manifold of low-rank matrices.
- Stable to noise in measurement vector *y* and returns as good an answer as a truncated SVD of the true *X*.

Low-rank Matrix Completion

$\mathbf{M}$ =



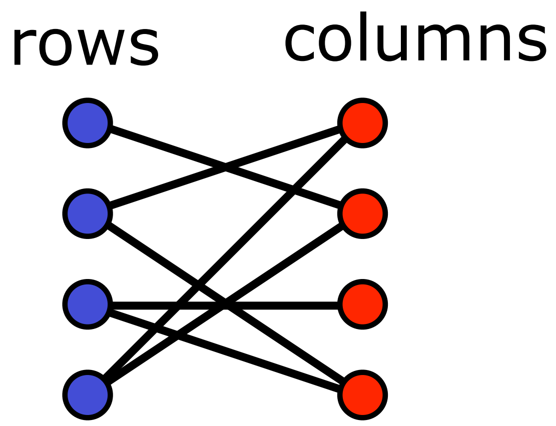
M_{ij} known for black cells
M_{ij} unknown for white cells

- How do you fill in the missing data?

$$\begin{array}{ll} \text{minimize} & \operatorname{rank}(\mathbf{X}) \\ \text{subject to} & X_{ij} = M_{ij} \quad \forall\ (i,j) \in \Omega \end{array}$$

Which Sampling Sets?

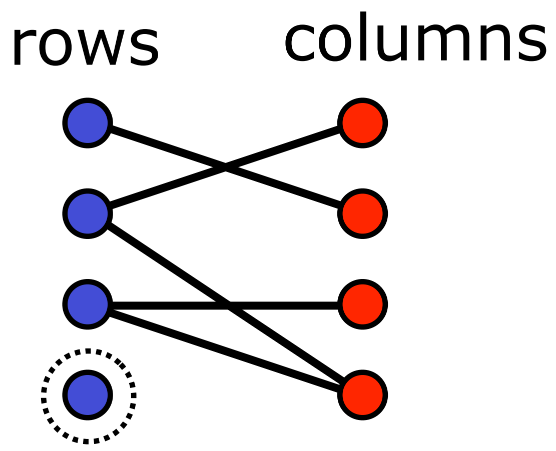
$$\Omega = \begin{bmatrix} 0 & * & 0 & 0 \\ * & 0 & 0 & * \\ 0 & 0 & * & * \\ * & * & 0 & 0 \end{bmatrix}$$



- **Row-column graph**
  - Vertices: indexed by rows and columns
  - Edge if that entry is in Ω

Which Sampling Sets?

$$\Omega = \begin{bmatrix} 0 & * & 0 & 0 \\ * & 0 & 0 & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

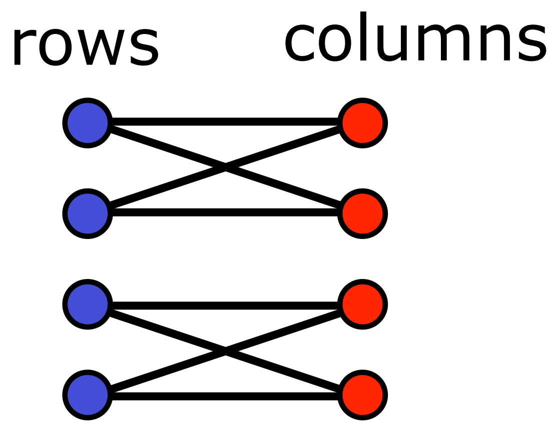


- **Row-column graph: all vertices must be observed**

- M = xy*. If you miss row 4, cannot determine x$_4$.

Which Sampling Sets?

$$\Omega = \begin{bmatrix} * & * & 0 & 0 \\ * & * & 0 & 0 \\ 0 & 0 & * & * \\ 0 & 0 & * & * \end{bmatrix}$$

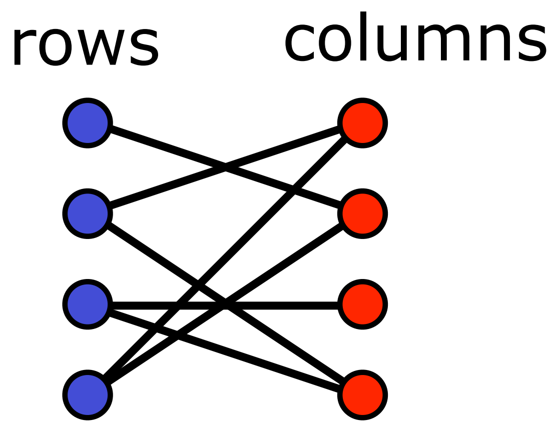


- **Row-column graph: must be connected**
- If M = xy*, cannot distinguish between

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ -x_3 \\ -x_4 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ -y_3 \\ -y_4 \end{bmatrix}$$

Which Sampling Sets?

$$\mathbf{X} = \begin{bmatrix} 0 & * & 0 & 0 \\ * & 0 & 0 & * \\ 0 & 0 & * & * \\ * & * & 0 & 0 \end{bmatrix}$$



- **Row-column graph: must have at least** r(n+k-r) **edges**

- The dimension of the manifold of rank r, k x n matrices is r(n+k-r)

If we can choose the samples...

- Generically, first r rows and r columns are sufficient:

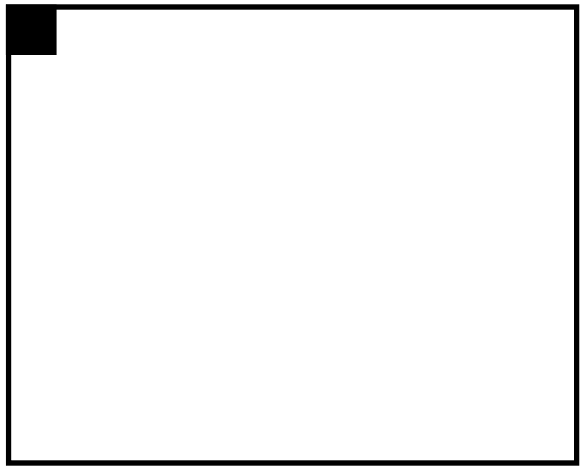
$$M = \left[\begin{array}{cc} A & B \\ C & CA^{-1}B \end{array} \right]$$

- [Frieze, Kannan, Vempala 1998, Drineas, Kannan, Mahoney 2003, etc.]: sample proportional to norms of columns. *Low-rank matrix approximations.*

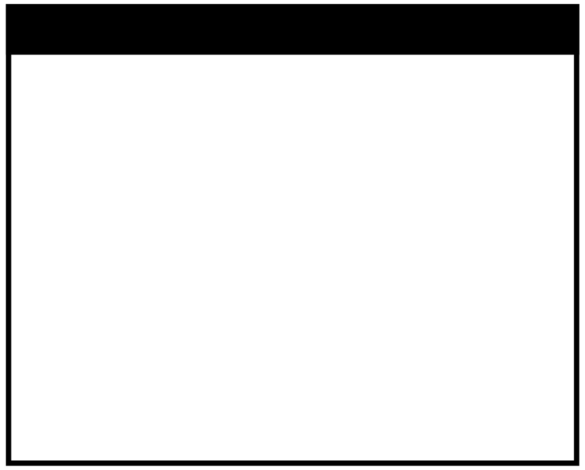
If we can't choose the samples...

- Most sets with more than 2*rn*β *log(n)* entries have at least one entry for every row and column, the row-column graph is connected.
- [Achloptas, McSherry 2004]: random sampling sufficient to obtain an additive error approximation to

Which matrices?

X =  $(= e_1 e_1^*)$

- Any subset of entries that misses the (1,1) component tells you nothing!

X =  $(= e_1 v^*)$

- Still need to see the entire first row

- Want each entry to provide nearly the same amount of information

Incoherence

- Let *U* be a subspace of $\mathbb{R}^n$ of dimension r and $\mathbf{P}_U$ be the orthogonal projection onto *U*. Then the *coherence* of *U* (with respect to the standard basis $\mathbf{e}_i$) is defined to be

$$\mu(U) \equiv \frac{n}{r} \max_{1 \le i \le n} \|\mathbf{P}_U \mathbf{e}_i\|^2.$$

- $\mu(U) \geq 1$
  - e.g., span of r columns of the Fourier transform
- $\mu(U) \leq n/r$
  - e.g., any subspace that contains a standard basis element
- $\mu(U) = O(1)$
  - sampled from the uniform distribution with r > log n

Incoherence

- Let *U* be a subspace of $\mathbb{R}^n$ of dimension r and $\mathbf{P}_U$ be the orthogonal projection onto *U*. Then the *coherence* of *U* (with respect to the standard basis $\mathbf{e}_i$) is defined to be

$$\mu(U) \equiv \frac{n}{r} \max_{1 \le i \le n} \|\mathbf{P}_U \mathbf{e}_i\|^2.$$

- $\mu(U) \geq 1$
  - [illegible]
- $\mu(U) \leq n/r$
  - e.g., any subspace that contains a standard basis element
- [illegible]
  - sampled from the uniform distribution with r > log n

$\mu(U)$ small means *leverage scores* are uniform.

$$p_i = \|\mathbf{P}_U \mathbf{e}_i\|^2$$

[Drineas, Mahoney, Muthukrishnan 2006]: uniform row/column sampling gives exact reconstruction.

Bounds for Matrix Completion

- Suppose **X** is k x n (k≤n) has rank *r* and has row and column spaces with incoherence bounded above by μ. Then the nuclear norm heuristic recovers **X** from most subsets of entries Ω with cardinality at least

$$|\Omega| \geq C\mu n^{6/5}\ r\ \log(n)$$

Candès and Recht. 2008

special case extensions:

$$|\Omega| \geq C\mu^2 n\ r\ \log^6(n)$$

[Candès and Tao 2009]
stronger assumptions

$$|\Omega| > C' n log(n)$$

[Keshavan et al, 2009]
rank = o(1), σ_1/σ_r bounded

$$|\Omega| \geq 32\mu\, r(n+k)\ \log^2(2n)$$

[Gross et al 2009,
Recht 2009,
Gross2009]

Recent Extensions

- Noise robustness
  - Candes-Plan, Keshavan *et al* 2009, Lounici *et al*, Neghaban and Wainwright 2010

- Deconvolving Sparse and Low-rank matrices
  - Chandrasekaran *et al* 2009, Wright *et al* 2009

- Fast algorithms
  - First order methods - Cai et al, Ma et al, Toh et al, Ji et al, etc...
  - "Generalized Blog Heuristic" - Lee et al, Recht and Re

Linear Inverse Problems

- Find me a solution of

$$y = \Phi x$$

- Φ m x n, m<n

- Of the infinite collection of solutions, which one should we pick?

- Leverage structure:

Sparsity

Rank

Smoothness

Symmetry

- How do we design algorithms to solve underdetermined systems problems with priors?

Parsimonious Models

rank

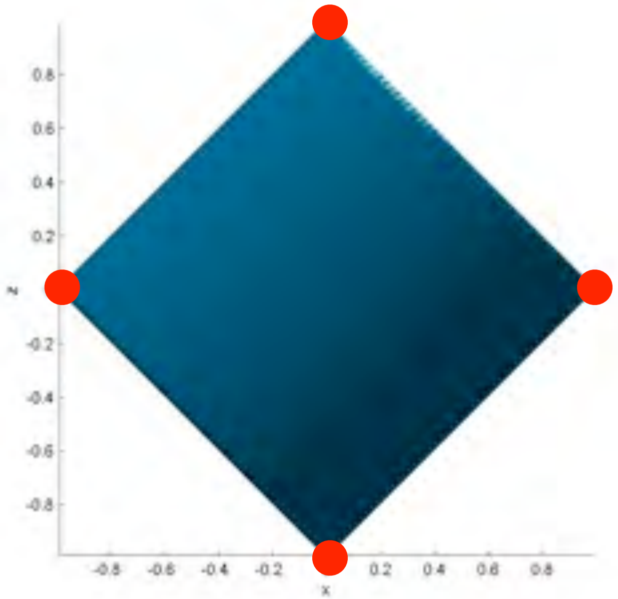
$$x = \sum_{k=1}^{r} w_k \alpha_k$$

model

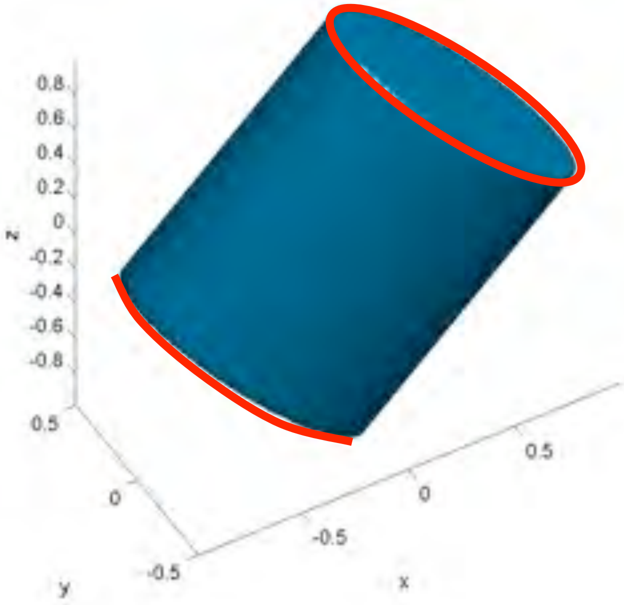
weights

atoms

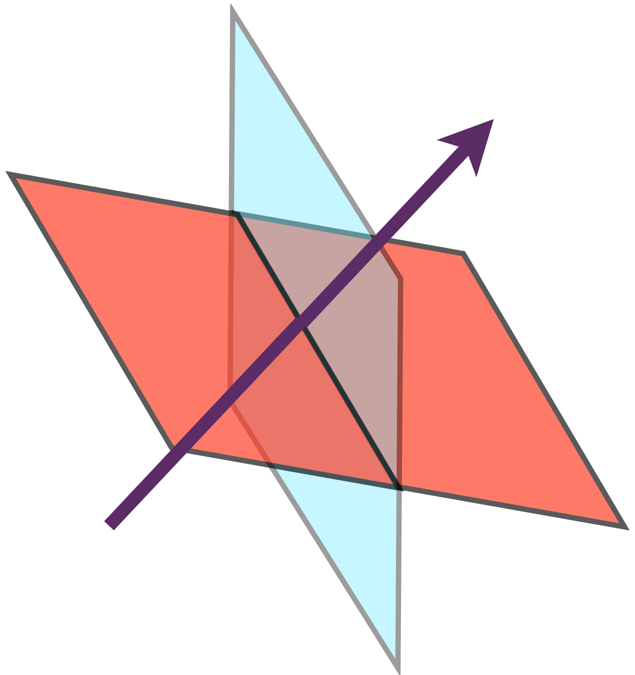
- Search for best linear combination of fewest atoms
- “rank” = fewest atoms needed to describe the model

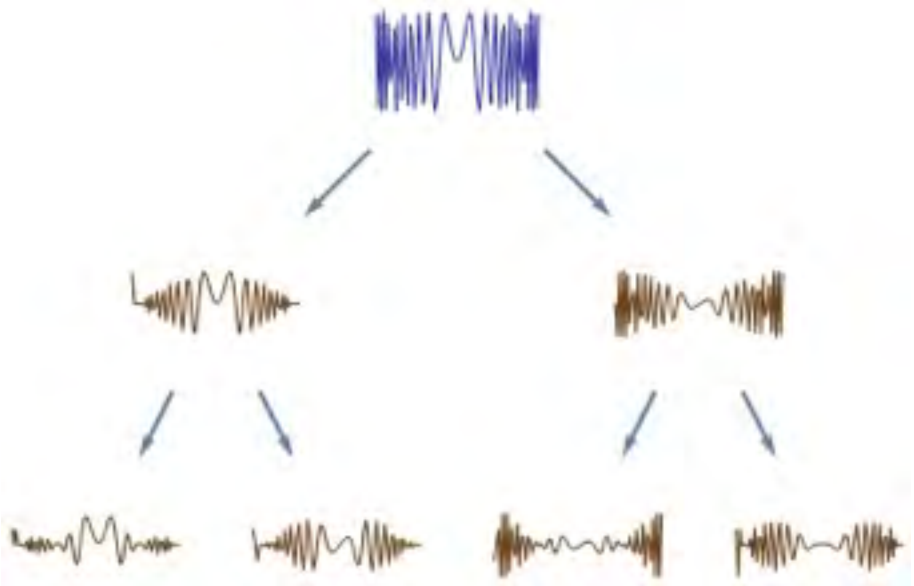


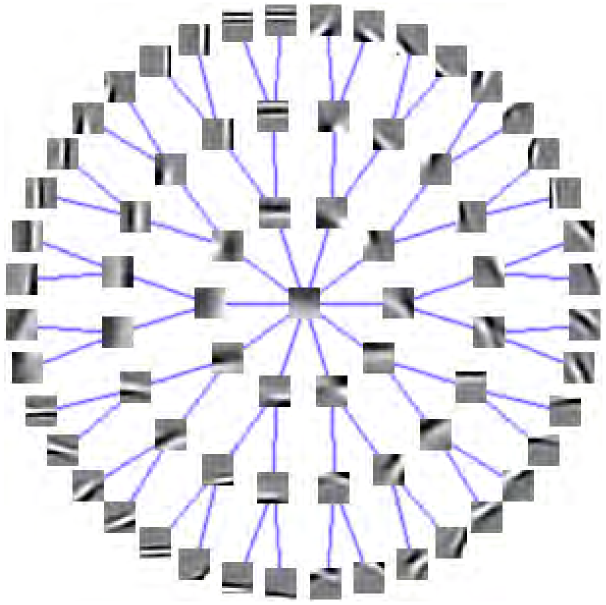
$$\|x\|_{\mathcal{A}} \equiv \inf_{(w,\alpha)} \sum_{k=1}^{r} |w_k|$$



Model Based Compressive Sensing







- X has structured sparsity: linear combination of elements from a set of subspaces $\{U_g\}$.
- Atomic set: unit norm vectors living in one of the U_g

$$\|x\|_{\mathcal{G}} = \inf \left\{ \sum_{g \in G} \|w_g\| \ : \ x = \sum_{g \in G} w_g, \ \ w_g \in U_g \right\}$$

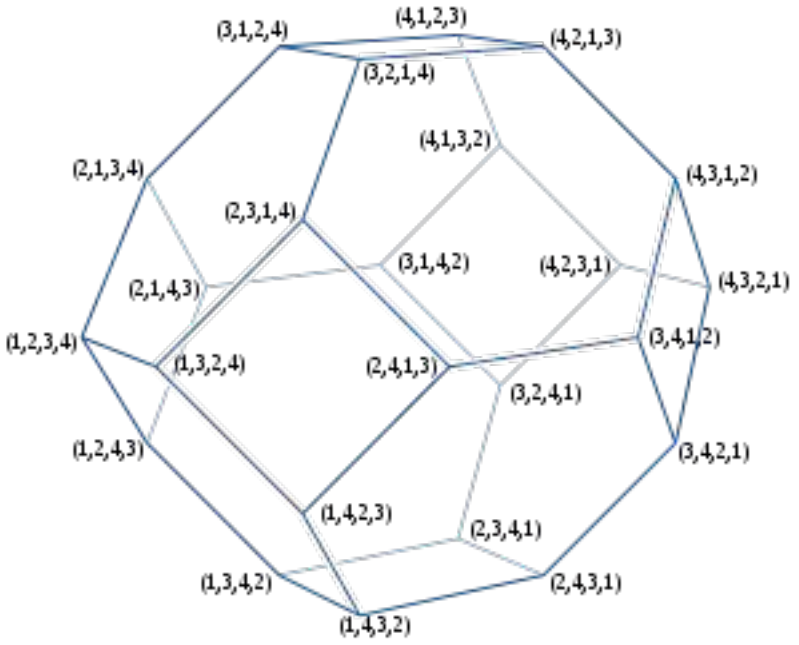
- Proposed by Jacob, Obozinski and Vert (2009).

Permutation Matrices

- X a sum of a few permutation matrices
- Examples: Multiobject Tracking (Huang et al), Ranked elections (Jagabathula, Shah)

- Convex hull of the permutation matrices: Birkhoff Polytope of doubly stochastic matrices
- *Permutahedra*: convex hull of permutations of a fixed vector.

[1,2,3,4] →



Atomic Norms

- Given a basic set of *atoms,* $\mathcal{A}$, define the function

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 \ : \ x \in t\mathrm{conv}(\mathcal{A})\}$$

- When $\mathcal{A}$ is centrosymmetric, we get a norm

$$\|x\|_{\mathcal{A}} = \inf\{\sum_{a \in \mathcal{A}} |c_a| \ : \ x = \sum_{a \in \mathcal{A}} c_a a\}$$

IDEA:
$$\begin{array}{ll} \text{minimize} & \|z\|_{\mathcal{A}} \\ \text{subject to} & \Phi z = y \end{array}$$

- When does this work?
- How do we solve the optimization problem?
- **A:** *Chandrasekaran, Recht, Willsky, and Parrilo 2010*

Atomic Norm Decompositions

- Propose a natural convex heuristic for enforcing prior information in inverse problems

- Bounds for the linear case: heuristic succeeds for most sufficiently large sets of measurements

- Stability without restricted isometries

- Standard program for computing these bounds: distance to normal cones

- Approximation schemes for computationally difficult priors

Extensions…

- Width Calculations for more general structures
- Recovery bounds for structured measurement matrices (application specific)
- Incorporating stochastic noise models
- Understanding of the loss due to convex relaxation and norm approximation
- Scaling generalized shrinkage algorithms to massive data sets

Acknowledgements

- See:

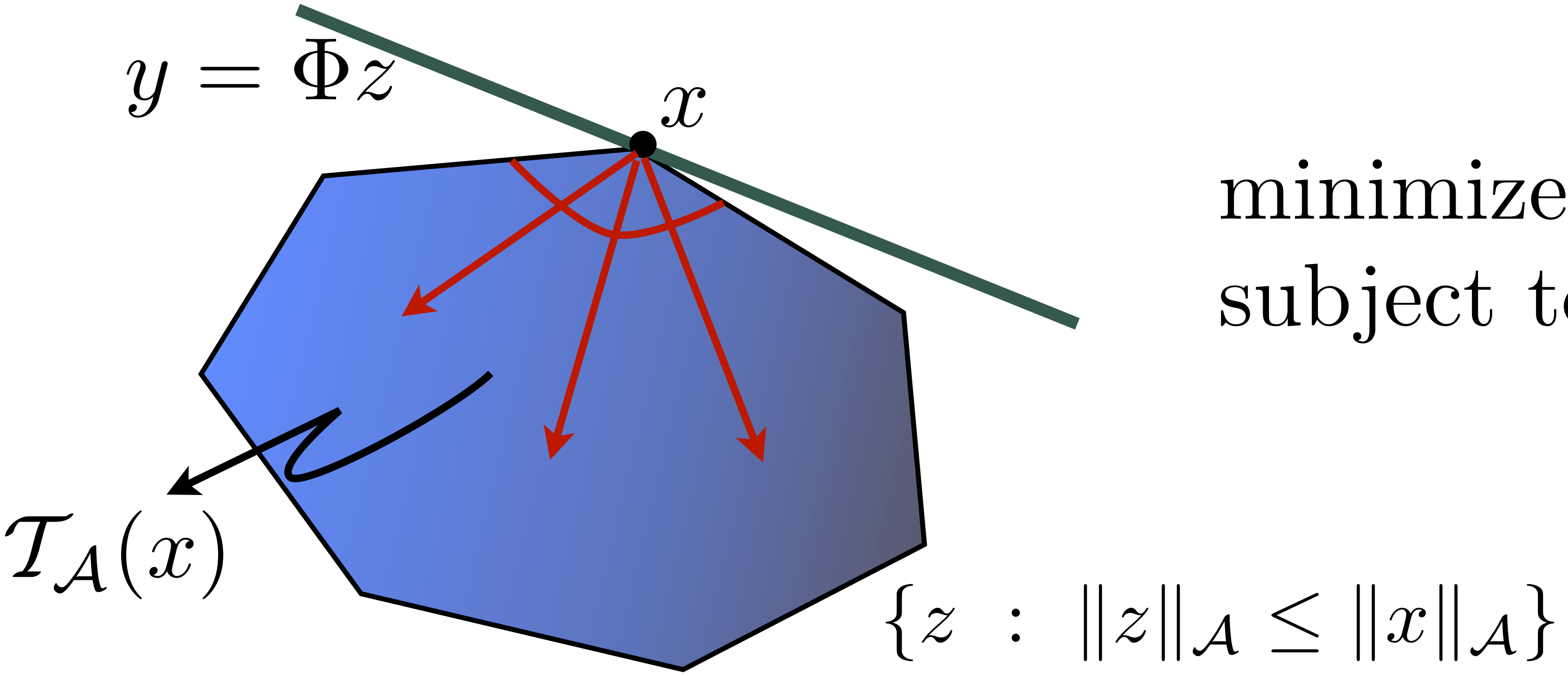
  http://pages.cs.wisc.edu/~brecht/publications.html

  for all references

- Results developed in collaboration with Venkat Chandrasekaran, Pablo Parrilo, and Alan Willsky (MIT), Jason Lee (Stanford), Ruslan Salakhutdinov (MIT), Nati Srebro (TTI), Joel A. Tropp (Caltech), and Christopher Re (UW-Madison).

Tangent Cones

- Set of directions that decrease the norm from x form a cone:
  $\mathcal{T}_{\mathcal{A}}(x) = \{d \ : \ \|x + \alpha d\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}} \text{ for some } \alpha > 0\}$


$y = \Phi z$
x
$\mathcal{T}_{\mathcal{A}}(x)$
$\{z \ : \ \|z\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}}\}$

$$\begin{array}{ll} \text{minimize} & \|z\|_{\mathcal{A}} \\ \text{subject to} & \Phi z = y \end{array}$$

- x is the unique minimizer if the intersection of this cone with the null space of Φ equals {0}

Gaussian Widths

- When does a random subspace, *U*, intersect a convex cone *C* at the origin?

- **Gordon 88:** with high probability if

$$\mathrm{codim}(U) \geq w(C)^2$$

- Where $w(C) = \mathbb{E}\left[\max_{x \in C \cap \mathbb{S}^{n-1}} \langle x, g \rangle\right]$ is the *Gaussian width*
(*g* is a normal Gaussian random vector.)

- **Corollary:** For inverse problems: if Φ is a random Gaussian matrix with m rows, need $m \geq w(\mathcal{T}_{\mathcal{A}}(x))^2$ for recovery of *x*.

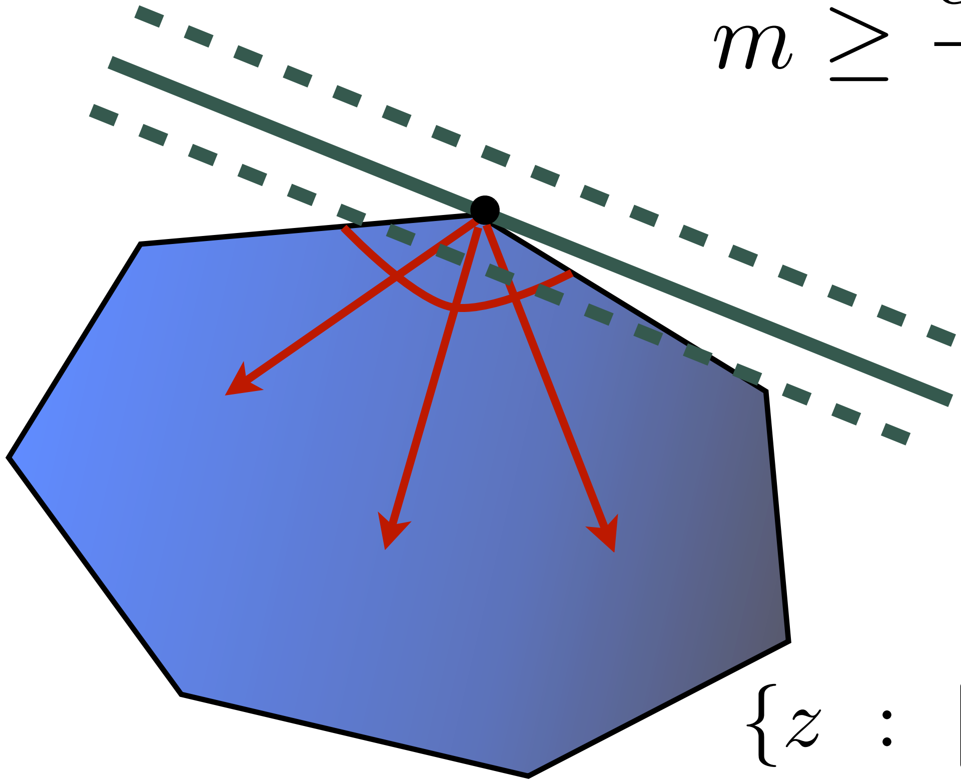
Robust Recovery

- Suppose we observe $y = \Phi x + w$ $\qquad \|w\|_2 \leq \delta$

$$\begin{array}{ll} \text{minimize} & \|z\|_{\mathcal{A}} \\ \text{subject to} & \|\Phi z - y\| \leq \delta \end{array}$$

- If $\hat{x}$ is an optimal solution, then $\|x - \hat{x}\| \leq \dfrac{2\delta}{\epsilon}$ provided that

$$m \geq \frac{c_0 w(\mathcal{T}_{\mathcal{A}}(x))^2}{(1-\epsilon)^2}$$



$$\|\Phi z - y\| \leq \delta$$

$$\{z \ : \ \|z\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}}\}$$

Duality

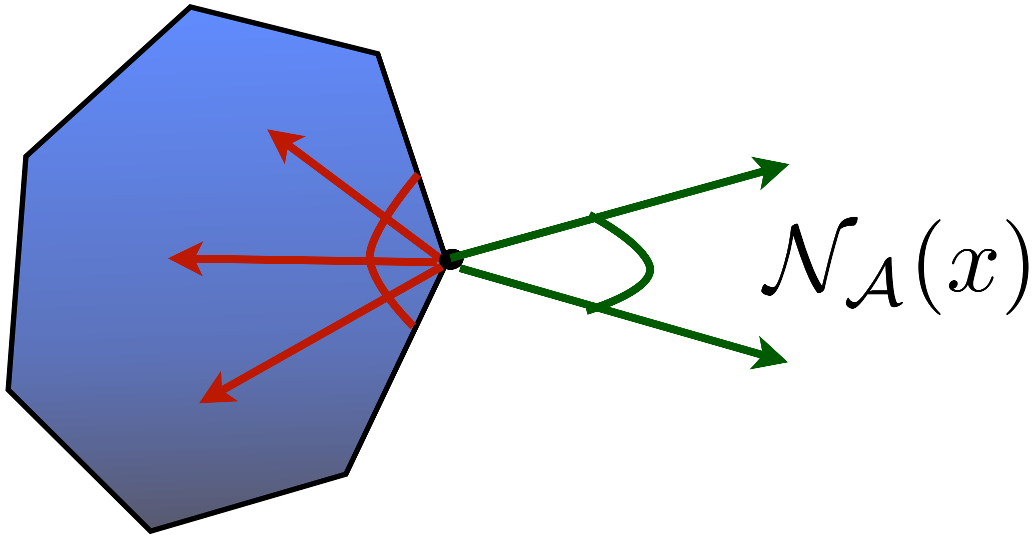
$$\begin{aligned} w(C) &= \mathbb{E}\left[\max_{\substack{v\in C\\ \|v\|=1}} \langle v, g\rangle\right] \\ &\leq \mathbb{E}\left[\max_{\substack{v\in C\\ \|v\|\leq 1}} \langle v, g\rangle\right] \\ &= \mathbb{E}\left[\min_{u\in C^*} \|g-u\|\right] \end{aligned}$$

- C^* is the polar cone.

$$C^* = \{w \;:\; \langle w, z\rangle \leq 0 \;\forall\, z \in C\}$$

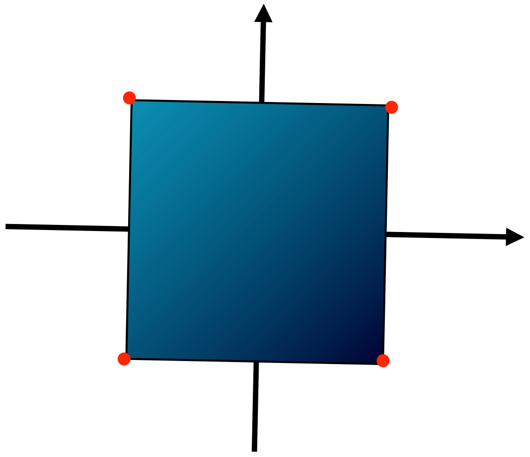
$$\mathcal{T}_{\mathcal{A}}(x)^* = \mathcal{N}_{\mathcal{A}}(x)$$

- $\mathcal{N}_{\mathcal{A}}(x)$ is the *normal cone*. Equal to the cone induced by the subdifferential of the atomic norm at *x*.



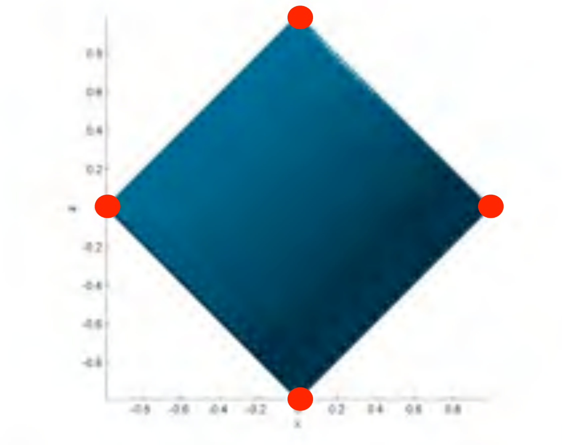
Re-derivations

- Hypercube: $m \geq n/2$



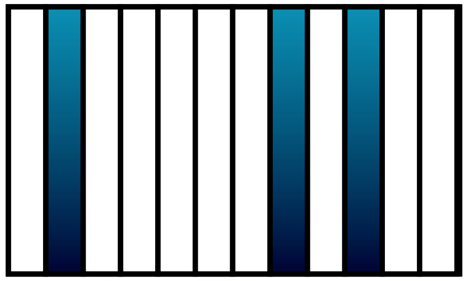
- Sparse Vectors, n vector, sparsity s<0.25n

$$m \geq 2s\left(\log\left(\frac{n-s}{s}\right)+1\right)$$



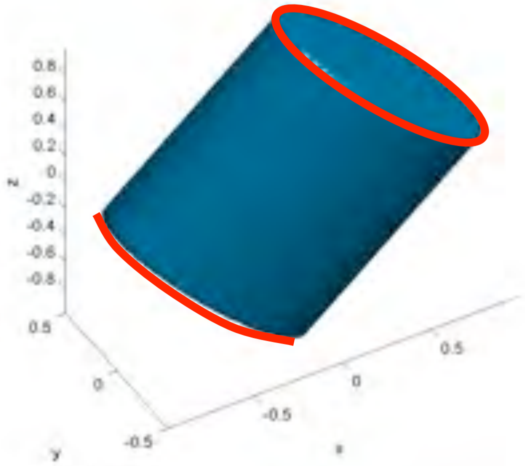
- Block sparse, M groups (possibly overlapping), maximum group size B, k active groups

$$m \geq 2k\left(\log\left(M-k\right)+B\right)+k$$



- Low-rank matrices: n_1 x n_2, (n_1<n_2), rank r

$$m \geq 3r(n_1+n_2-r)$$



General Cones

- **Theorem:** Let *C* be a nonempty cone with polar cone *C**. Suppose C* subtends normalized solid angle μ. Then

$$w(C) \leq 3\sqrt{\log\left(\frac{4}{\mu}\right)}$$

- **Proof Idea:** The expected distance to C* can be bounded by the expected distance to a spherical cap
- *Isoperimetry*: Out of all subsets of the sphere with the same measure, the one with the smallest neighborhood is the spherical cap
- The rest is just integrals...

Symmetric Polytopes

- **Corollary:** For a vertex-transitive (i.e., “symmetric”) polytope with p vertices, O(log p) Gaussian measurements are sufficient to recover a vertex via convex optimization.

- For n x n permutation matrix: m = O(n log n)
- For n x n cut matrix: m = O(n)
  - (Semidefinite relaxation also gives m = O(n))

Algorithms

$$\text{minimize}_z \quad \|\Phi z - y\|_2^2 + \mu \|z\|_{\mathcal{A}}$$

- Naturally amenable to projected gradient algorithm:

$$\boxed{z_{k+1} = \Pi_{\eta\mu}(z_k - \eta \Phi^* r_k)}$$

residual $\quad r_k = \Phi z_k - y$

"shrinkage" $\quad \Pi_\tau(z) = \arg\min_u \tfrac{1}{2}\|z-u\|^2 + \tau \|u\|_{\mathcal{A}}$

- Similar algorithm for atomic norm constraint

- Same basic ingredients for ALM, ADM, Bregman, Mirror Prox, etc... how to compute the shrinkage?

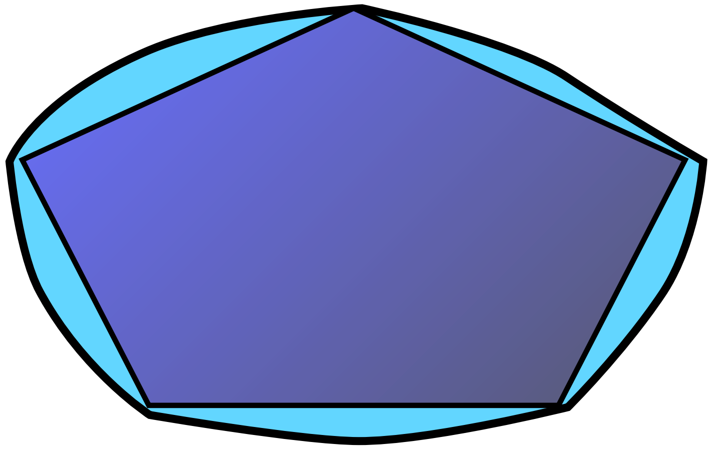
Relaxations

$$\|v\|_{\mathcal{A}}^{*} = \max_{a \in \mathcal{A}} \langle v, a \rangle$$

- Dual norm is efficiently computable if the set of atoms is polyhedral or semidefinite representable

$$\mathcal{A}_1 \subset \mathcal{A}_2 \implies \|x\|_{\mathcal{A}_1}^{*} \leq \|x\|_{\mathcal{A}_2}^{*} \quad \text{and} \quad \|x\|_{\mathcal{A}_2} \leq \|x\|_{\mathcal{A}_1}$$

- Convex relaxations of atoms yield approximations to the norm



NB! tangent cone gets wider

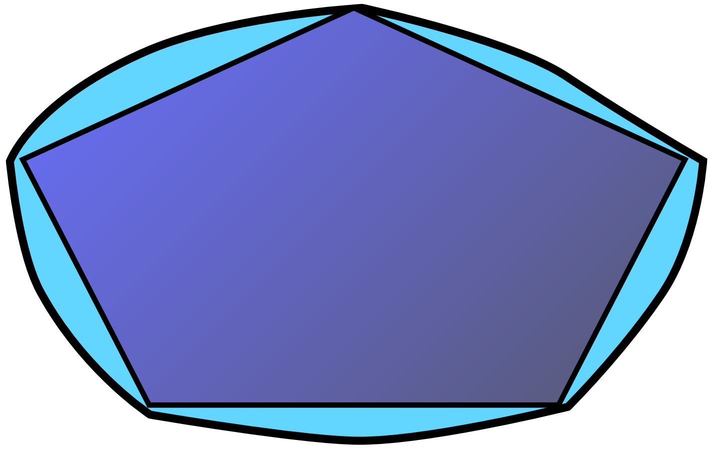
- Hierarchy of relaxations based on *θ-Bodies* yield progressively tighter bounds on the atomic norm

Theta Bodies

- Suppose $\mathcal{A}$ is an *algebraic variety*

$$\mathcal{A} = \{x \ : \ f(x) = 0 \ \forall f \in I\}$$

$$\|v\|_{\mathcal{A}}^{*} = \max_{a \in \mathcal{A}} \langle v, a \rangle \le \tau$$



$$q = h + g$$

$$\underbrace{h(x) \ge 0 \ \forall x}_{\text{positive everywhere}} \qquad \underbrace{g \in I}_{\text{vanishes on atoms}}$$

- *Relaxation:* restrict h to be sum of squares.
- Gives a lower bound on atomic norm
- Solvable by semidefinite programming (*Gouveia, Parrilo, and Thomas, 2010*)