The convex geometry of inverse problems

Benjamin Recht
Department of Computer Sciences
University of Wisconsin-Madison

Joint work with
Venkat Chandrasekaran
Pablo Parrilo
Alan Willsky

Linear Inverse Problems

- Find me a solution of

$$y = \Phi x$$

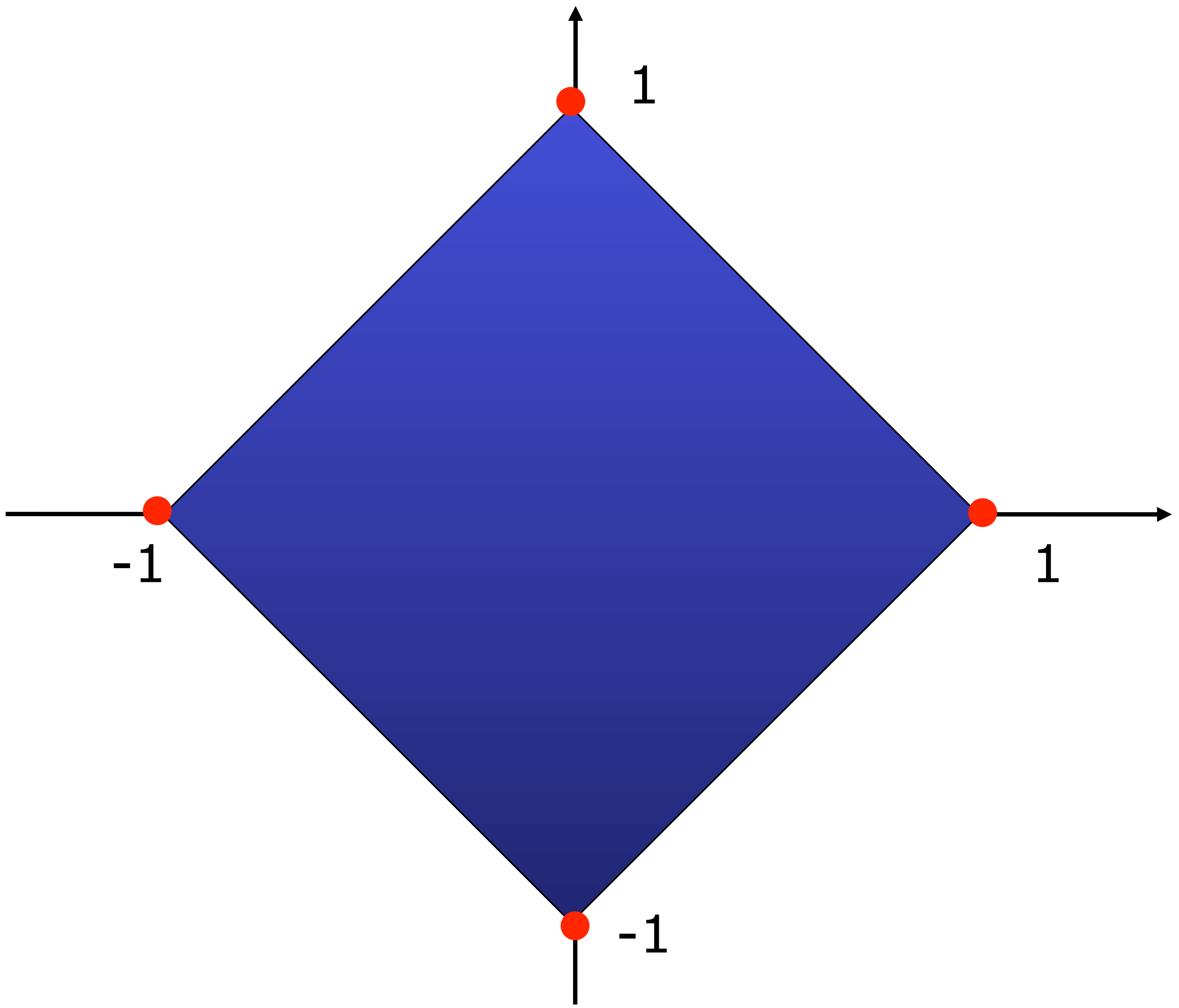
- Φ m x n, m<n
- Of the infinite collection of solutions, which one should we pick?
- Leverage structure:

Sparsity Rank Smoothness Symmetry

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

Sparsity

- 1-sparse vectors of Euclidean norm 1

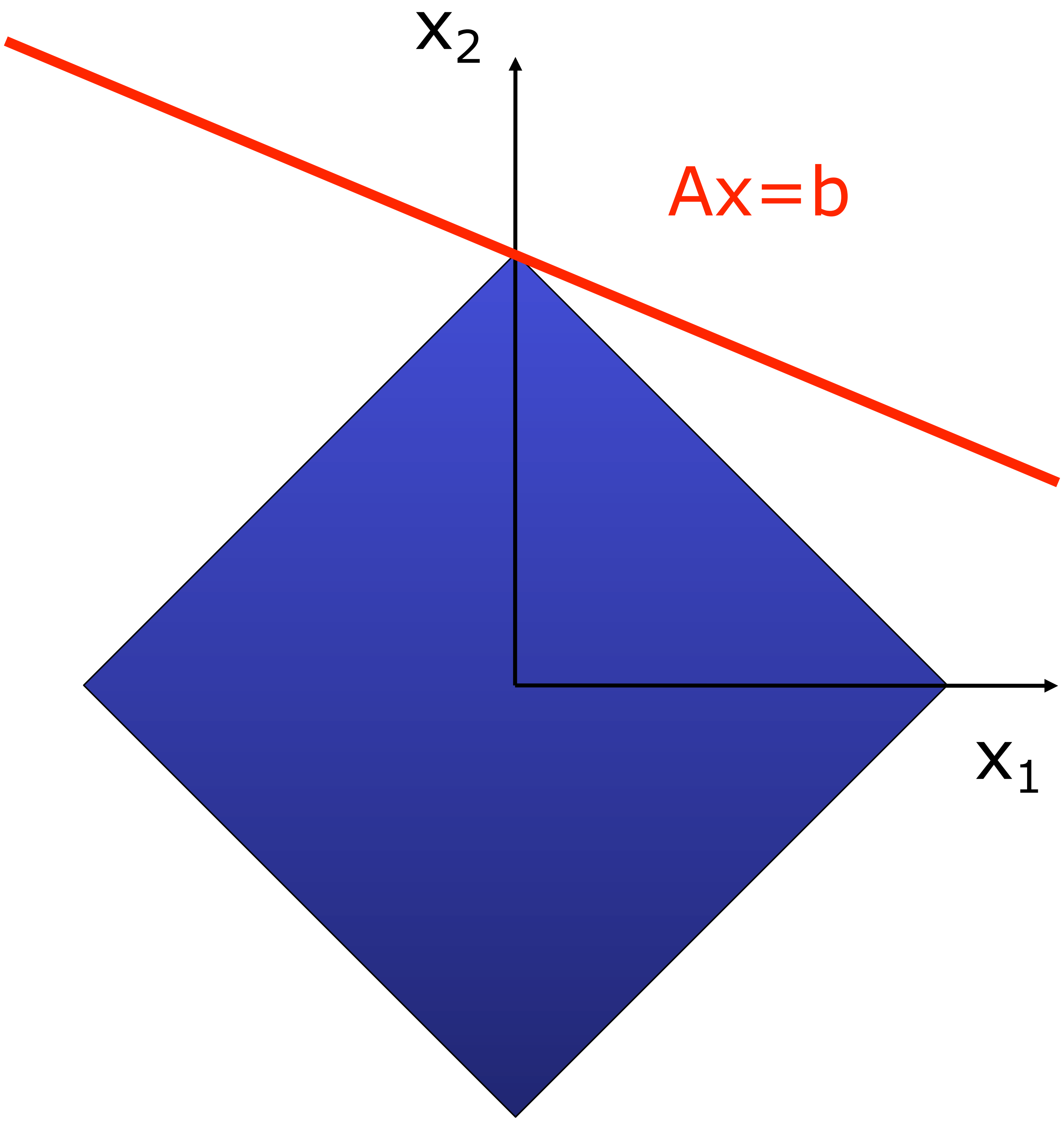


- Convex hull is the unit ball of the l_1 norm

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

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

$$\begin{array}{ll} \text{minimize} & \|x\|_1 = \sum_{i=1}^{n} |x_i| \\ \text{subject to} & Ax = b \end{array}$$



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

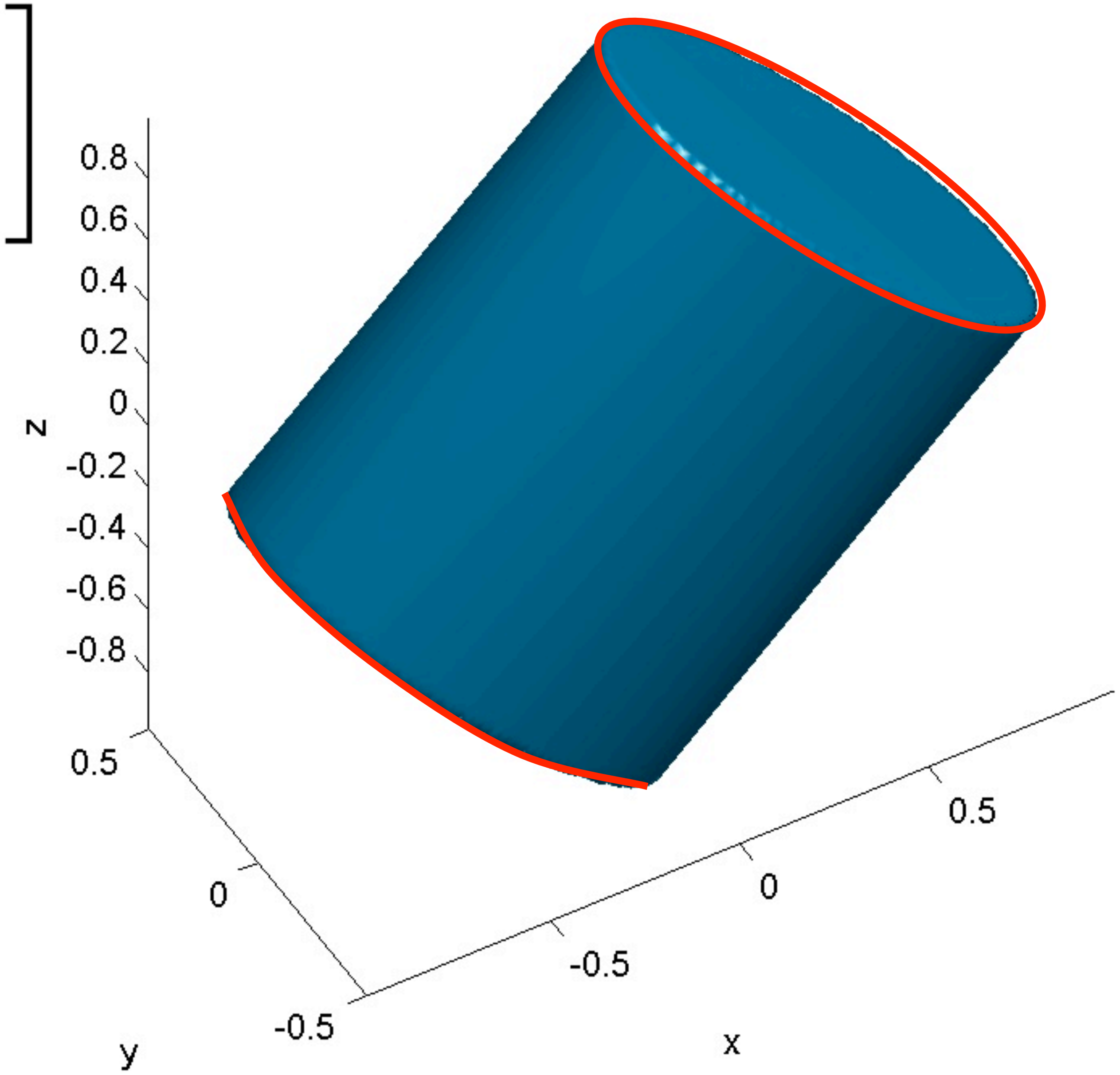
Rank

- 2x2 matrices
- plotted in 3d

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

— rank 1

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



Convex hull:

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

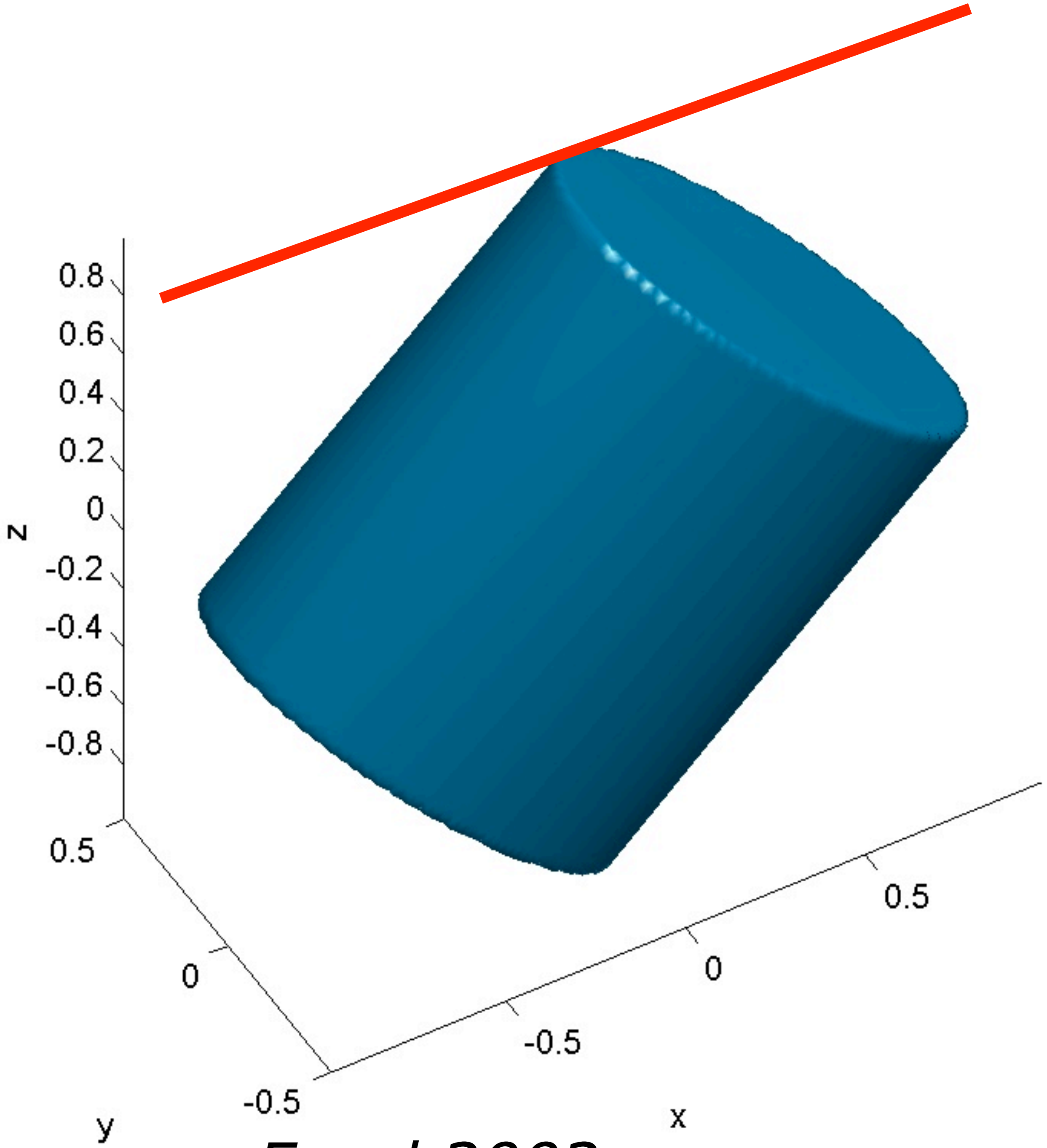
$$\|X\|_* = \sum_i \sigma_i(X)$$

- 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



Fazel 2002.

R, Fazel, and Parrilo 2007

Rank Minimization/Matrix Completion

Integer Programming

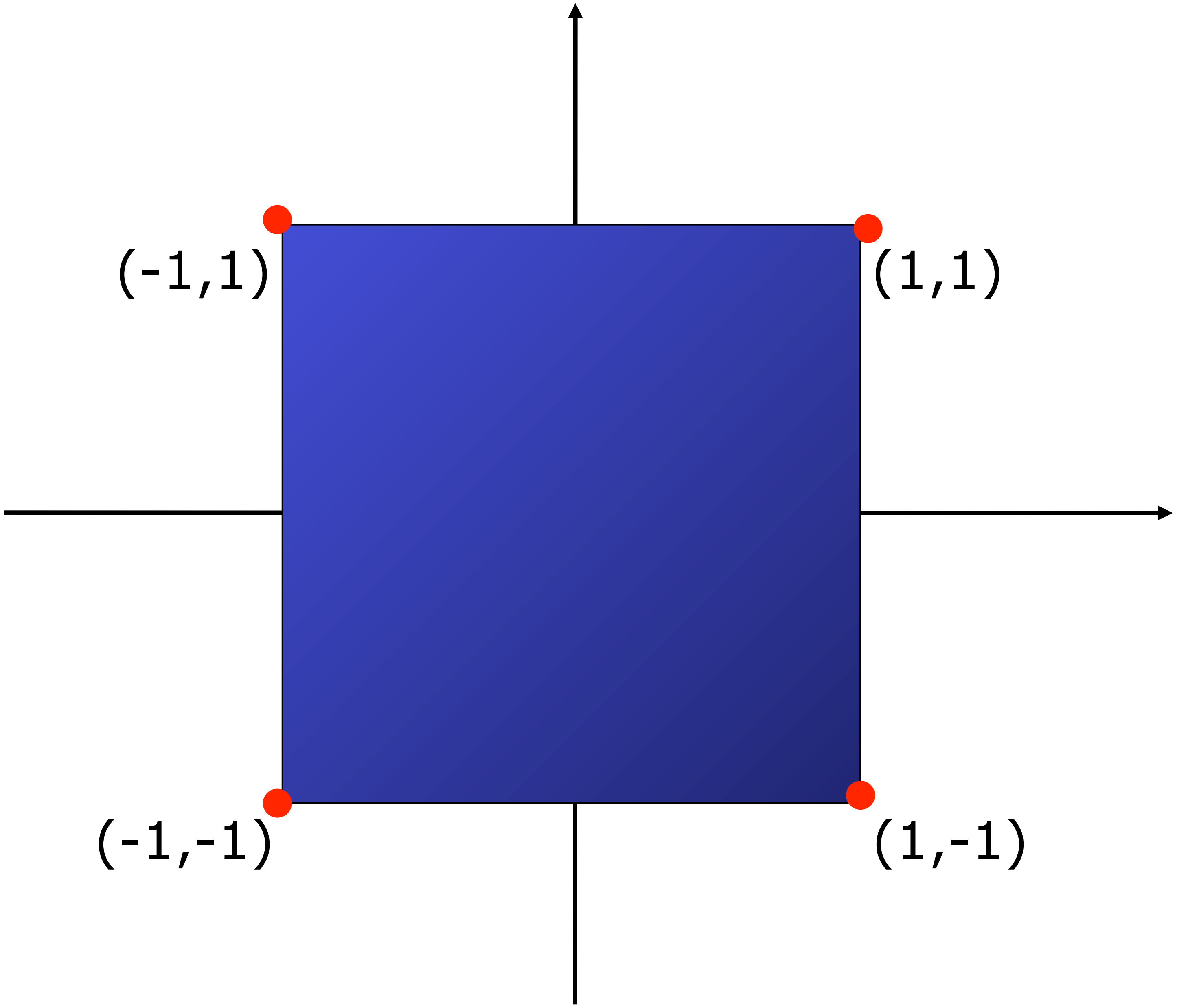
- Integer solutions:

  all components of x are ±1

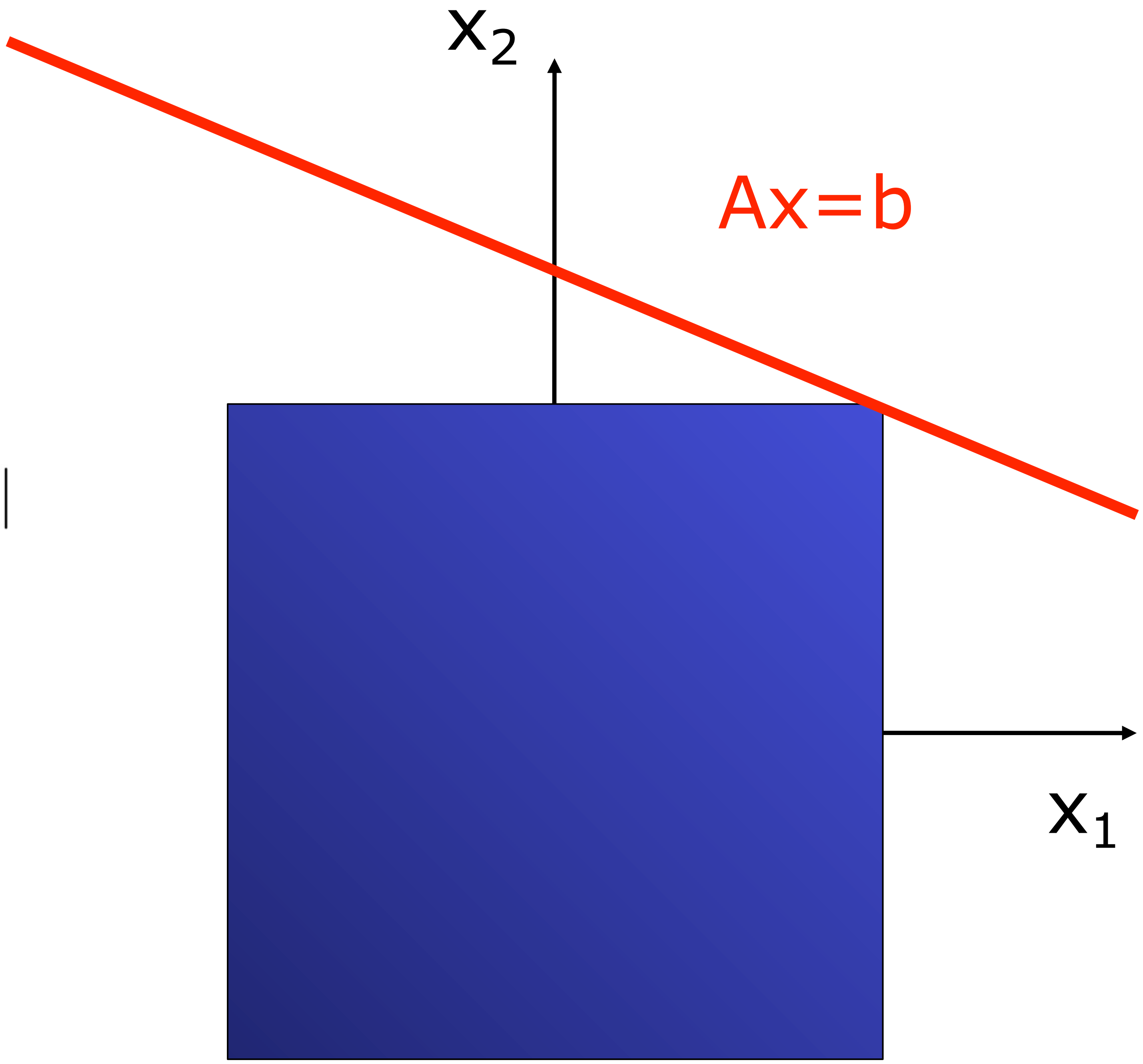
- Convex hull is the unit ball of the l_1 norm

$$\{x \;:\; \|x\|_\infty \leq 1\}$$

$$\|x\|_\infty = \max_i |x_i|$$

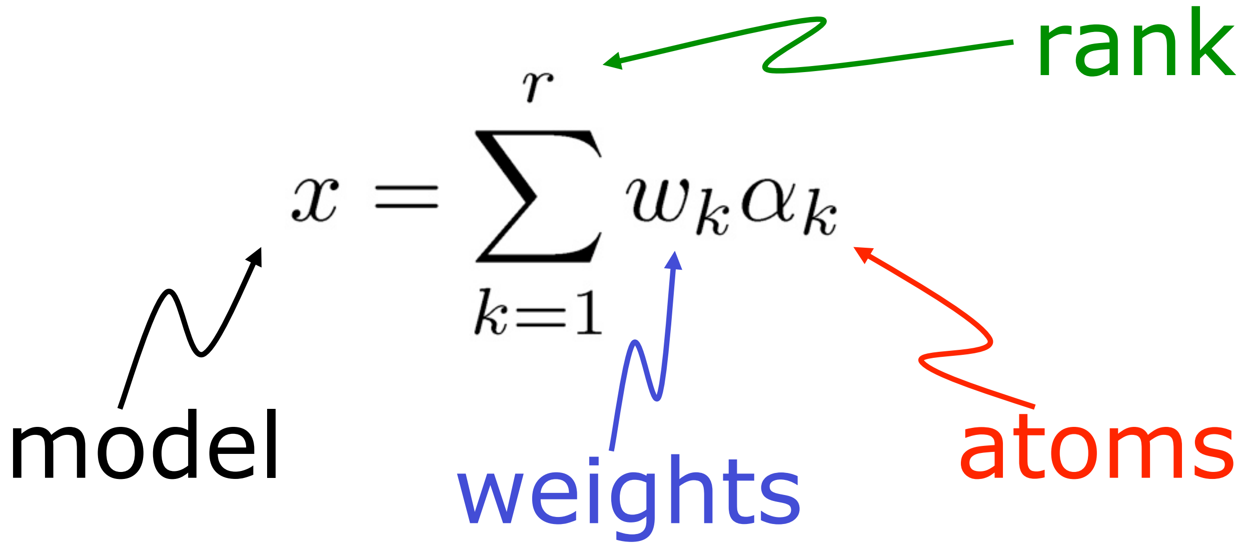


$$
\begin{array}{ll}
\text{minimize} & \|x\|_\infty = \max_i |x_i| \\
\text{subject to} & Ax = b
\end{array}
$$



Donoho and Tanner 2008
Mangasarian and Recht. 2009.

Parsimonious Models



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

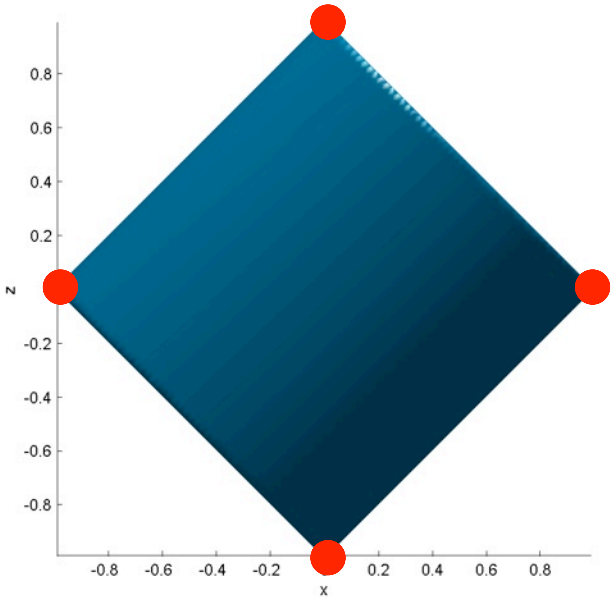
rank

model

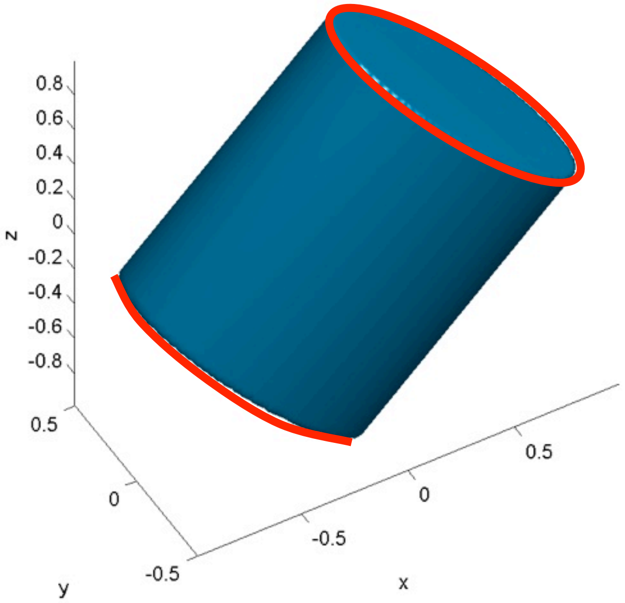
weights

atoms

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



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

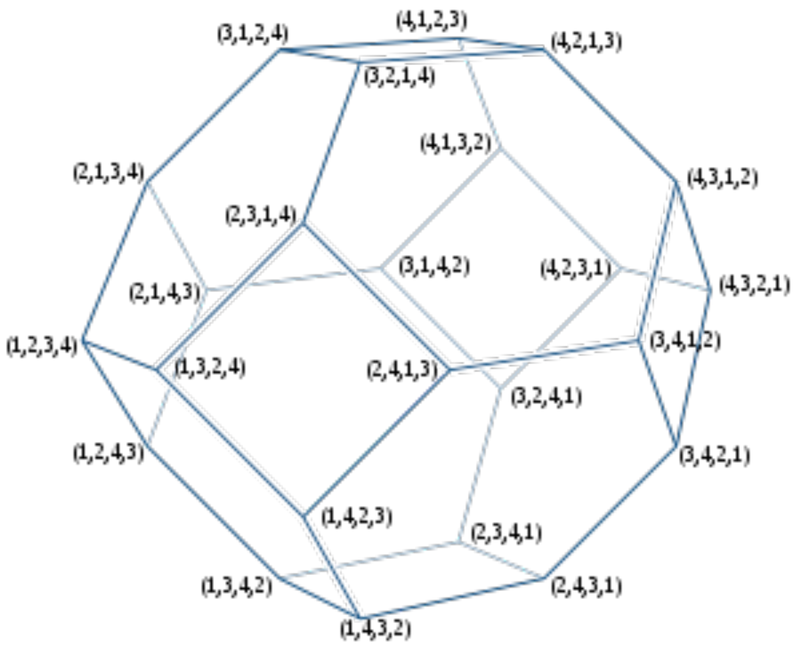


Permutation Matrices

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

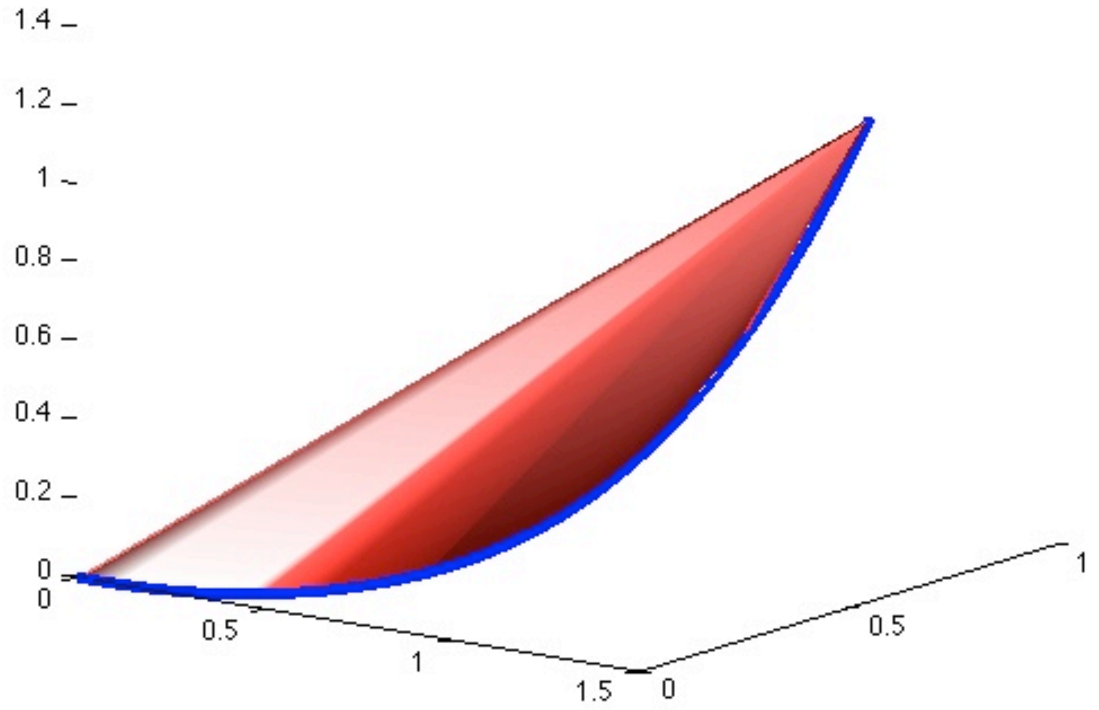
- Convex hull of the permutation matrices: Birkhoff Polytope of doubly stochastic matrices
- *Permutahedra*: convex hull of permutations of a fixed vector.

[1,2,3,4] →



Moment Curve

- Curve of $[1,t,t^2,t^3,t^4,\ldots]$
- System Identification, Image Processing, Numerical Integration, Statistical Inference...
- Convex hull is characterized by linear matrix inequalities



Cut Matrices

- Sums of rank-one sign matrices:

$$X = \sum_i p_i X_i \qquad X_i = x_i x_i^* \qquad X_{ij} = \pm 1$$

- Collaborative Filtering (Srebro et al), Clustering in Genetic Networks (Tanay et al), Combinatorial Approximation Algorithms (Frieze and Kannan)

- Convex hull is the *cut polytope*. Membership is NP-hard to test

- Semidefinite approximations of this hull to within constant factors.

Tensors

- X a low-rank tensor (multi-index array)
- Examples: Polynomial equations, computer vision, differential equations, statistics, chemometrics,...
- Convex hull of rank-1 tensors leads to a “tensor nuclear norm ball”
- Everything involving tensors is intractable to compute (in theory...)
- But heuristics work unreasonably well: why?

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?

Atomic norms in sparse approximation

- Greedy approximations

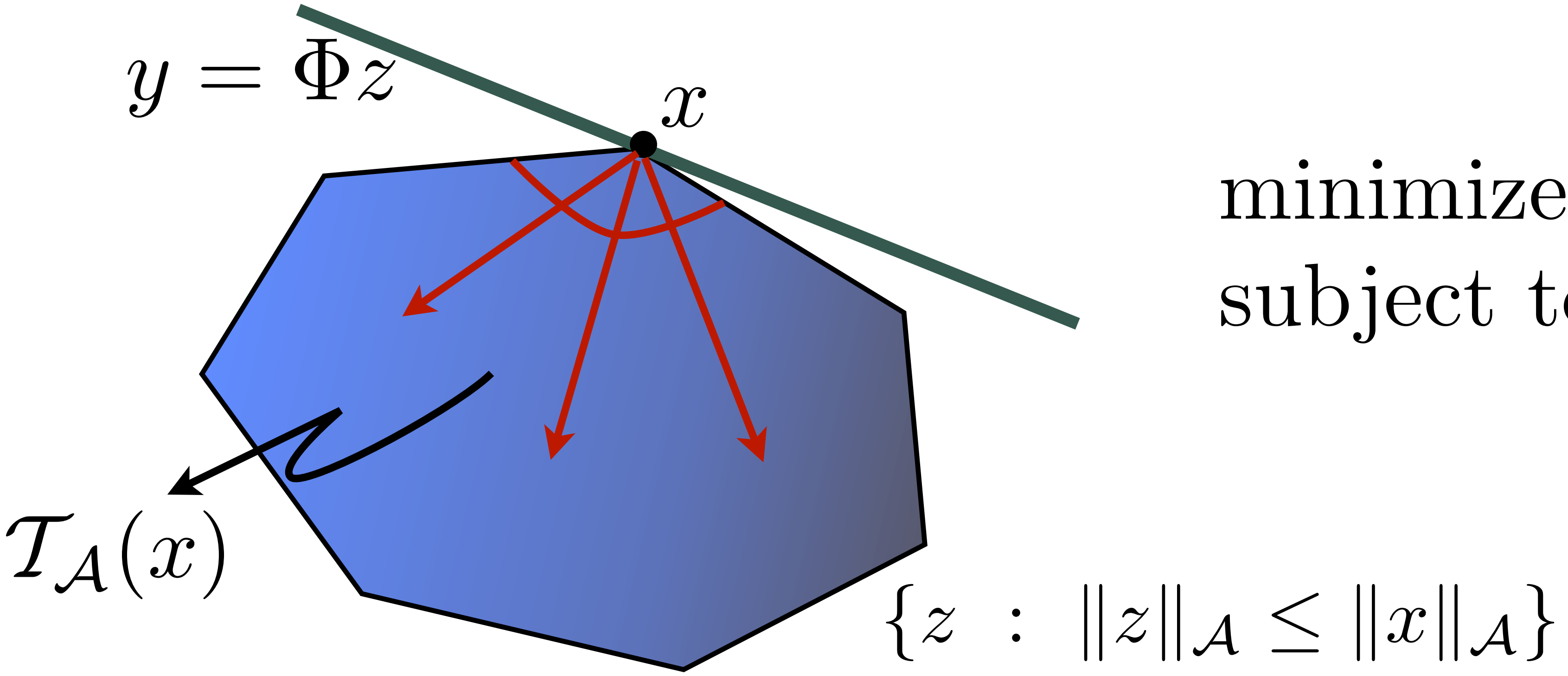
$$\|f - f_n\|_{\mathcal{L}_2} \leq \frac{c_0 \|f\|_{\mathcal{A}}}{\sqrt{n}}$$

- Best *n* term approximation to a function *f* in the convex hull of $\mathscr{A}$ in Banach space.

- Maurey, Jones, and Barron (1980s-90s)
- Devore and Temlyakov (1996)

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

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

- **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*.

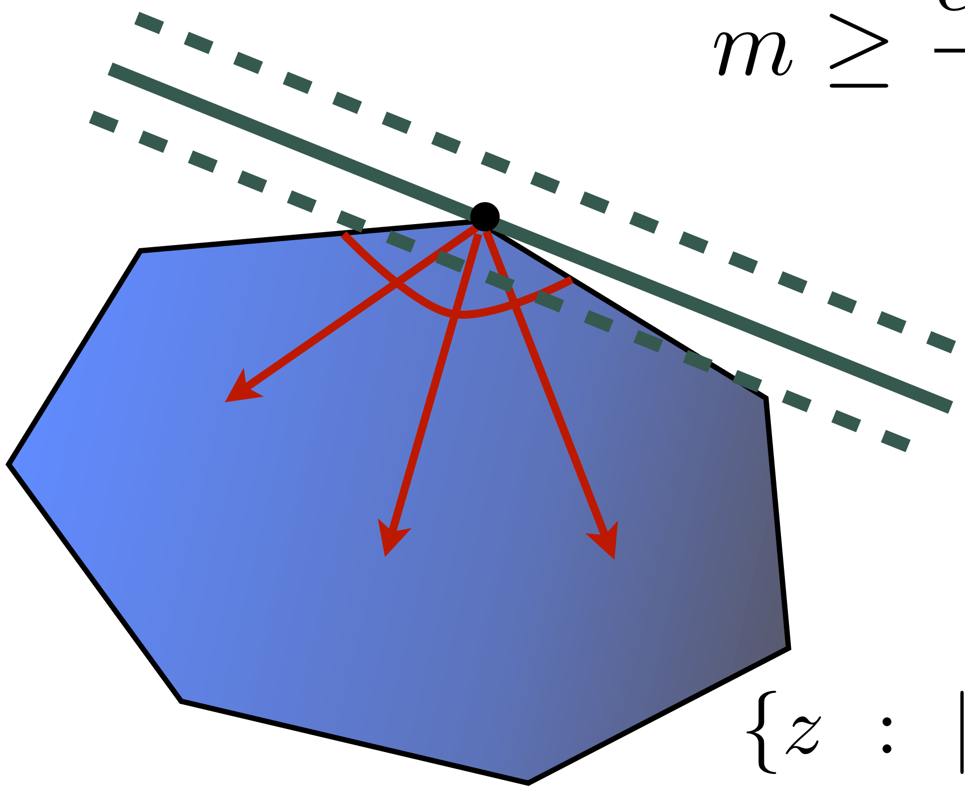
Robust Recovery

- 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

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



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

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

What can we do with Gaussian widths?

- Used by Rudelson & Vershynin for analyzing sharp bounds on the RIP for special case of sparse vector recovery using l_1.

- For a k-dim subspace S, $w(S)^2 = k$.

- Computing width of a cone C not easy in general

- Main property we exploit: symmetry and duality (inspired by Stojnic 09)

Duality

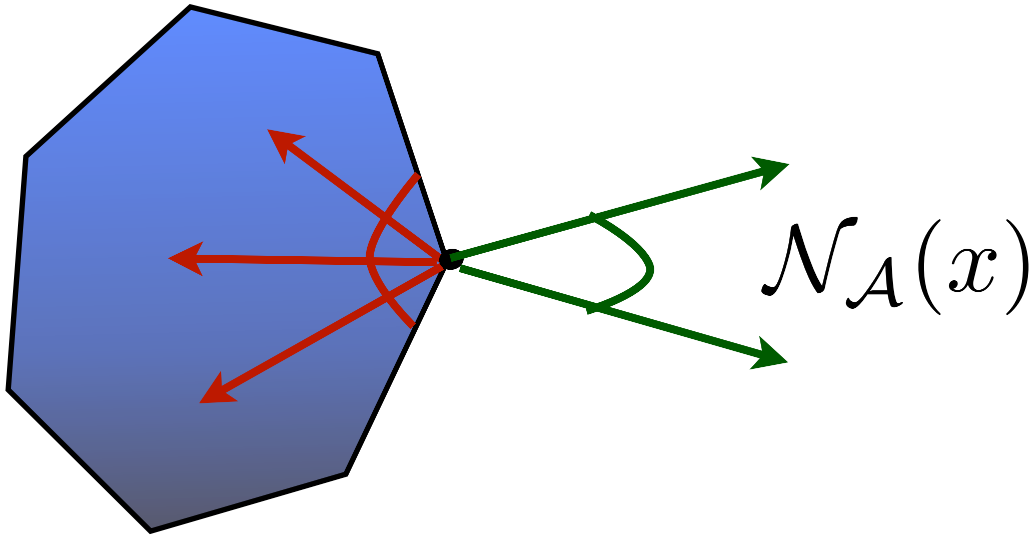
$$\begin{aligned} w(C) &= \mathbb{E}\left[\max_{\substack{v\in C\\ \|v\|=1}} \langle v, g\rangle\right] \\ &\leq \mathbb{E}\left[\max_{\substack{v\in C\\ \|v\|\leq 1}} \langle v, g\rangle\right] \\ &= \mathbb{E}\left[\min_{u\in C^*} \|g-u\|\right] \end{aligned}$$

- C^* is the polar cone.

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

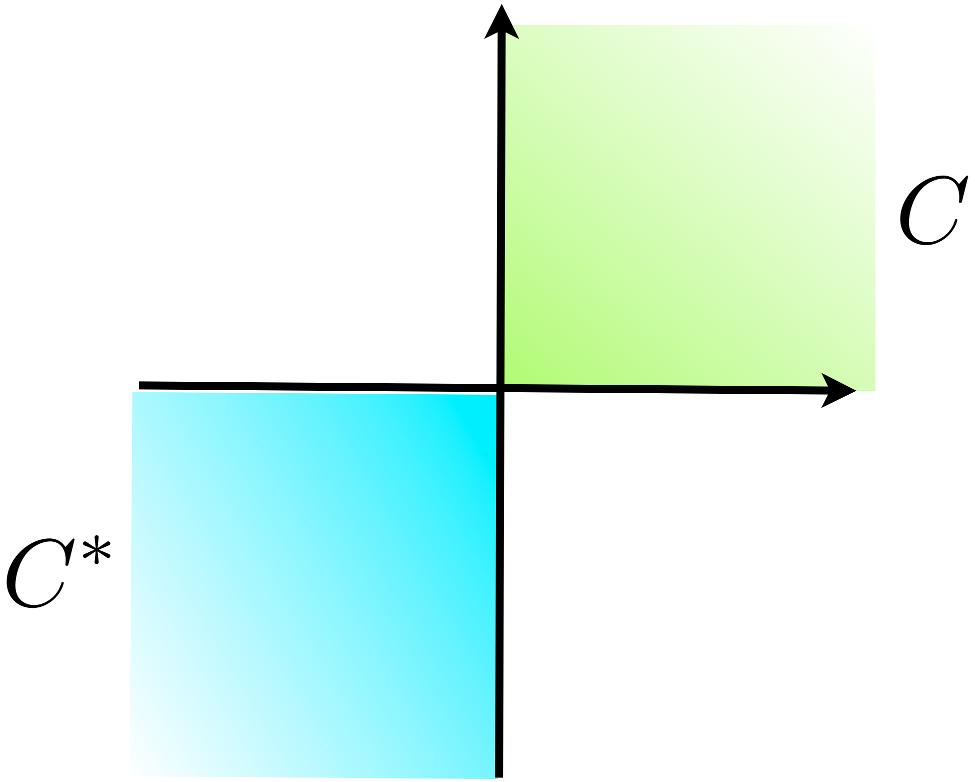
$$\mathcal{T}_{\mathcal{A}}(x)^* = \mathcal{N}_{\mathcal{A}}(x)$$

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



Symmetry I - self duality

- Self dual cones - orthant, positive semidefinite cone, second order cone
- Gaussian width = half the dimension of the cone



$$x = \Pi_C(x) + \Pi_{C^*}(x)$$

$$\langle \Pi_C(x), \Pi_{C^*}(x) \rangle = 0$$

$$\mathbb{E}[\inf_{u \in C^*} \|g - u\|_2^2] = \mathbb{E}[\|\Pi_C(g)\|_2^2] = \mathbb{E}[\|\Pi_{C^*}(g)\|_2^2]$$

Spectral Norm Ball

- How many measurements to recover a unitary matrix?

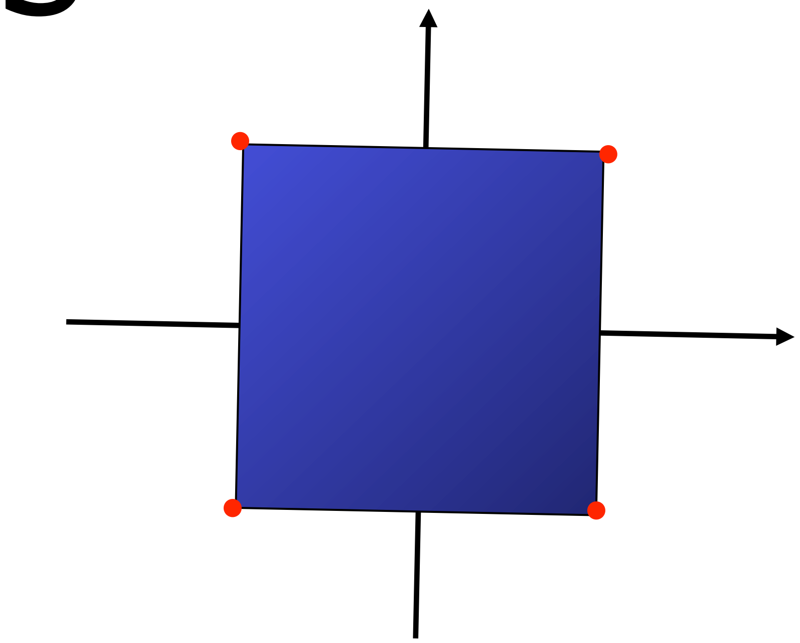
$$\mathcal{T}_{\mathcal{A}}(U) = S - P$$

- Tangent cone is skew-symmetric matrices minus the positive semidefinite cone.

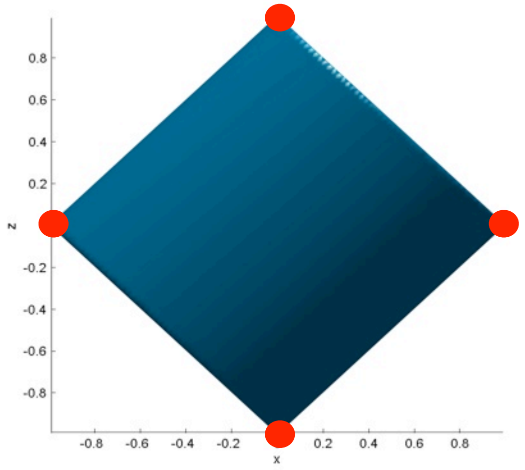
- These two sets are orthogonal, thus

$$w(\mathcal{T}_{\mathcal{A}}(U))^2 \leq \binom{n-1}{2} + \frac{1}{2}\binom{n}{2} = \frac{3n^2 - n}{4}$$

Re-derivations



- Hypercube: $m \geq n/2$
  - (orthant is self dual, or direct integration)

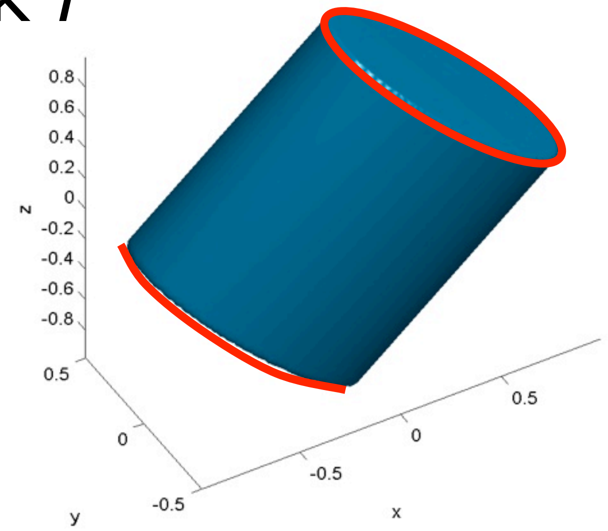


- Sparse Vectors, n vector, sparsity s

$$m \geq (2s+1)\log(n-s)$$

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

$$m \geq 3r(n_1+n_2-r)+2n_1$$



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...

Symmetry II - 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 $$r_k = \Phi z_k - y$$

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

- Similar algorithm for atomic norm constraint

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

Shrinkage

$$\Pi_\tau(z) = \arg\min_u \tfrac{1}{2}\|z-u\|^2 + \tau\|u\|_{\mathcal{A}}$$
$$\Lambda_\tau(z) = \arg\min_{\|v\|_{\mathcal{A}}^* \leq \tau} \tfrac{1}{2}\|z-v\|^2$$

$$\boxed{z = \Pi_\tau(z) + \Lambda_\tau(z)}$$

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

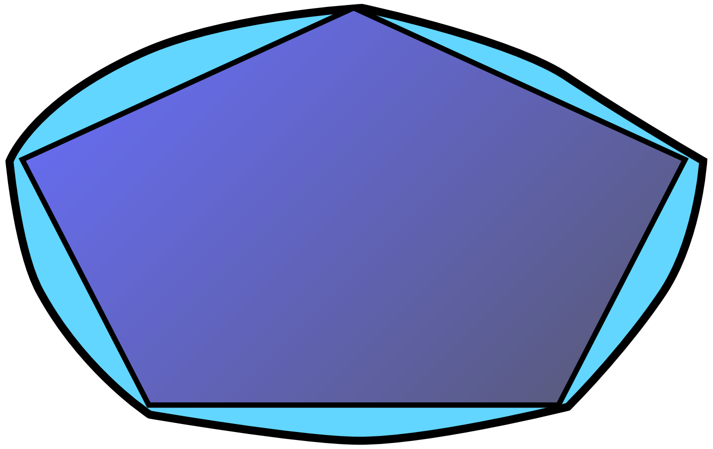
Relaxations

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

- 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 polyhedral (Sherali-Adams) or semi-definite (Positivstellensatz) approximations to atomic sets yield progressively tighter bounds on the atomic norm

Maxnorm Algorithms

$$\|X\|_{\mathcal{A}} = \inf\left\{\|\sigma\|_1 \;:\; X = \sum\nolimits_j \sigma_j u_j v_j' \text{ where } \|u_j\|_\infty = 1 \text{ and } \|v_j\|_\infty = 1\right\}$$

- Atomic set of rank one sign matrices

$$\|X\|_{\max} := \inf\left\{\|U\|_{2,\infty}\|V\|_{2,\infty} \;:\; X = UV^*\right\}$$

- Semidefinite relaxation

$$\|X\|_{\max} \le \|X\|_{\mathcal{A}} \le 1.8\|X\|_{\max}$$

- Grothendieck's inequality

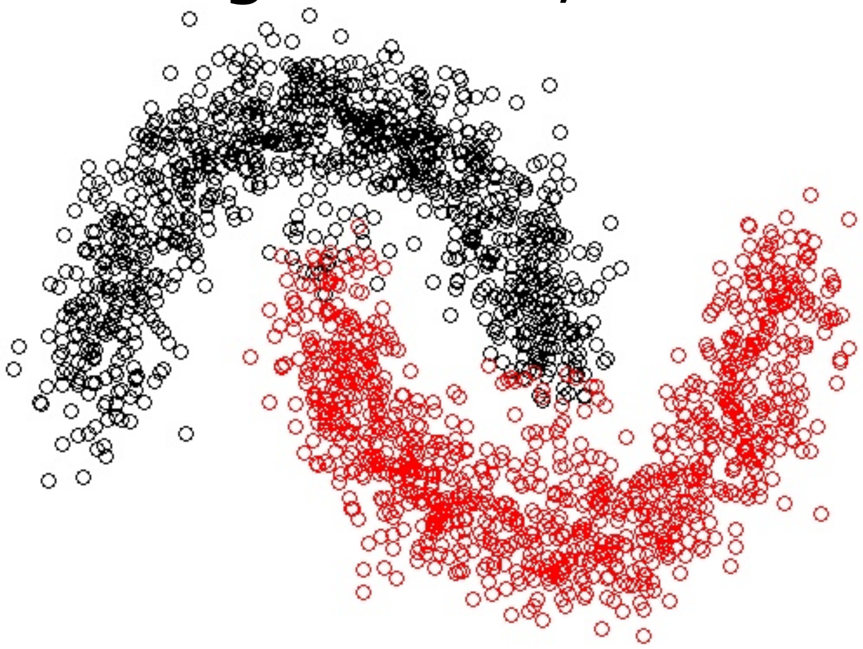
- Fast algorithms based on projection/shrinkage: Lee, R., Salakhutdinov, Srebro, Tropp (NIPS2010)
  - Key ingredients: semidefinite programming, low-rank embedding, projected gradient, stochastic approximation

Maxnorm vs Tracenorm

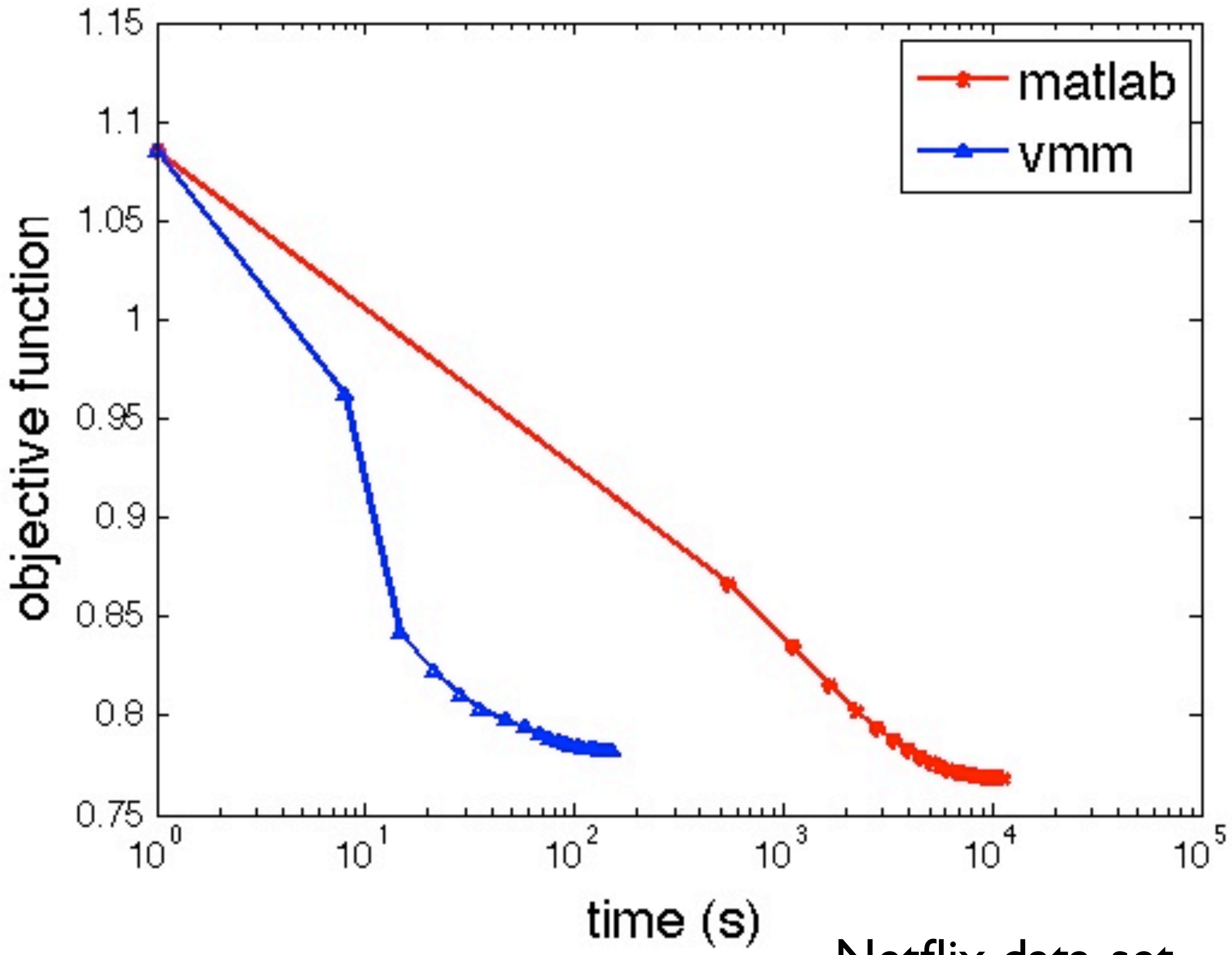
- Better Generalization in theory (Srebro 05, and width arguments)

- More stable and better prediction in practice (for example, significantly better performance on collaborative filtering data sets)

- Extensions to spectral clustering, graph approximation algorithms, etc.



Scaling up



Netflix data-set
100M examples
17770 rows
480189 columns

- Exploiting geometric structure in multicore data analysis
- Clever parallelization of incremental gradient algorithms, cache alignment, etc.
- In preparation for SIGMOD11 with Christopher Re

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)
- Understanding of the loss due to convex relaxation and norm approximation
- Scaling generalized shrinkage algorithms to massive data sets