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

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:

$\boxed{\text{Sparsity}}$ $\boxed{\text{Rank}}$ $\boxed{\text{Smoothness}}$ $\boxed{\text{Symmetry}}$

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

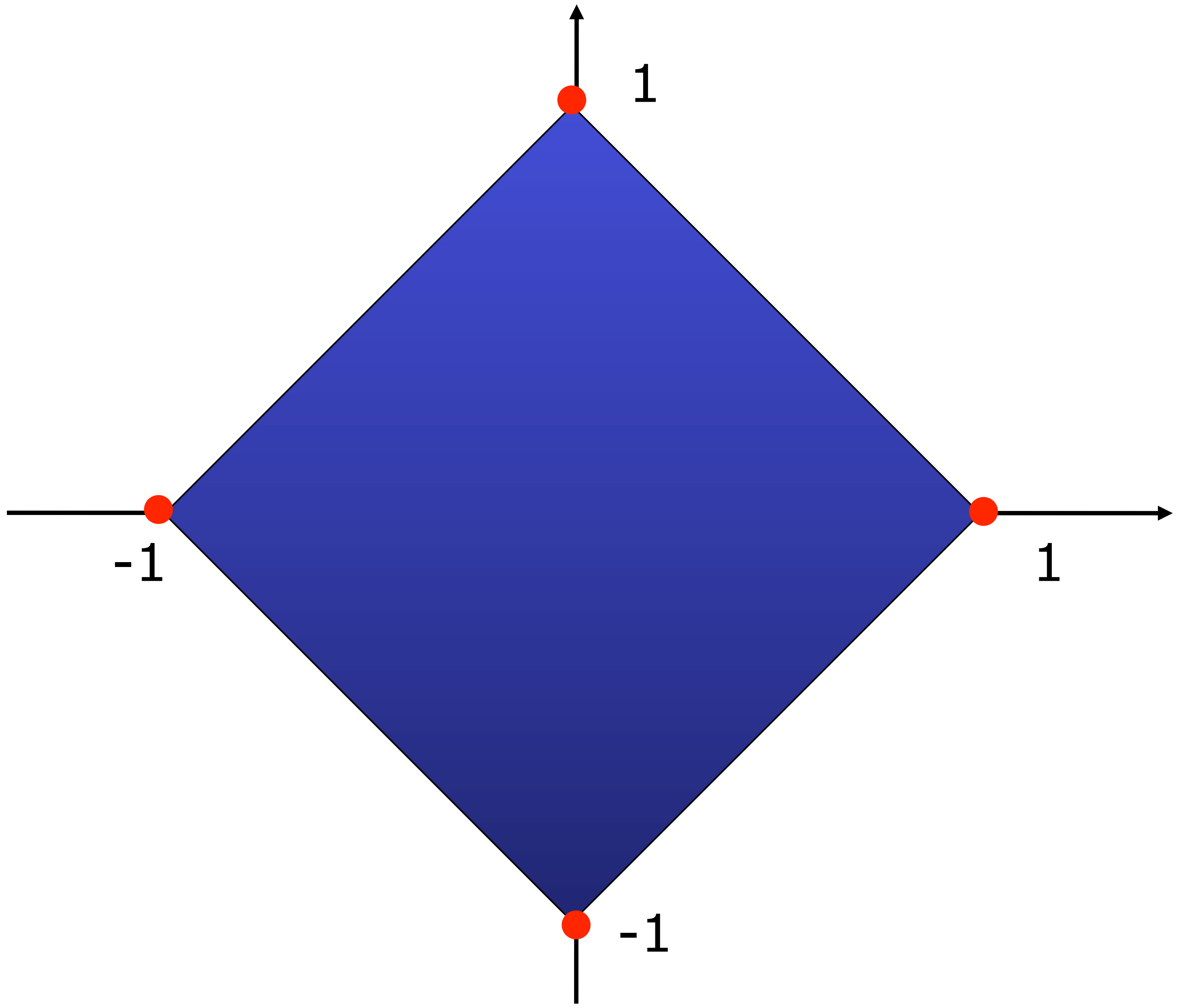
Sparsity

- 1-sparse vectors of Euclidean norm 1

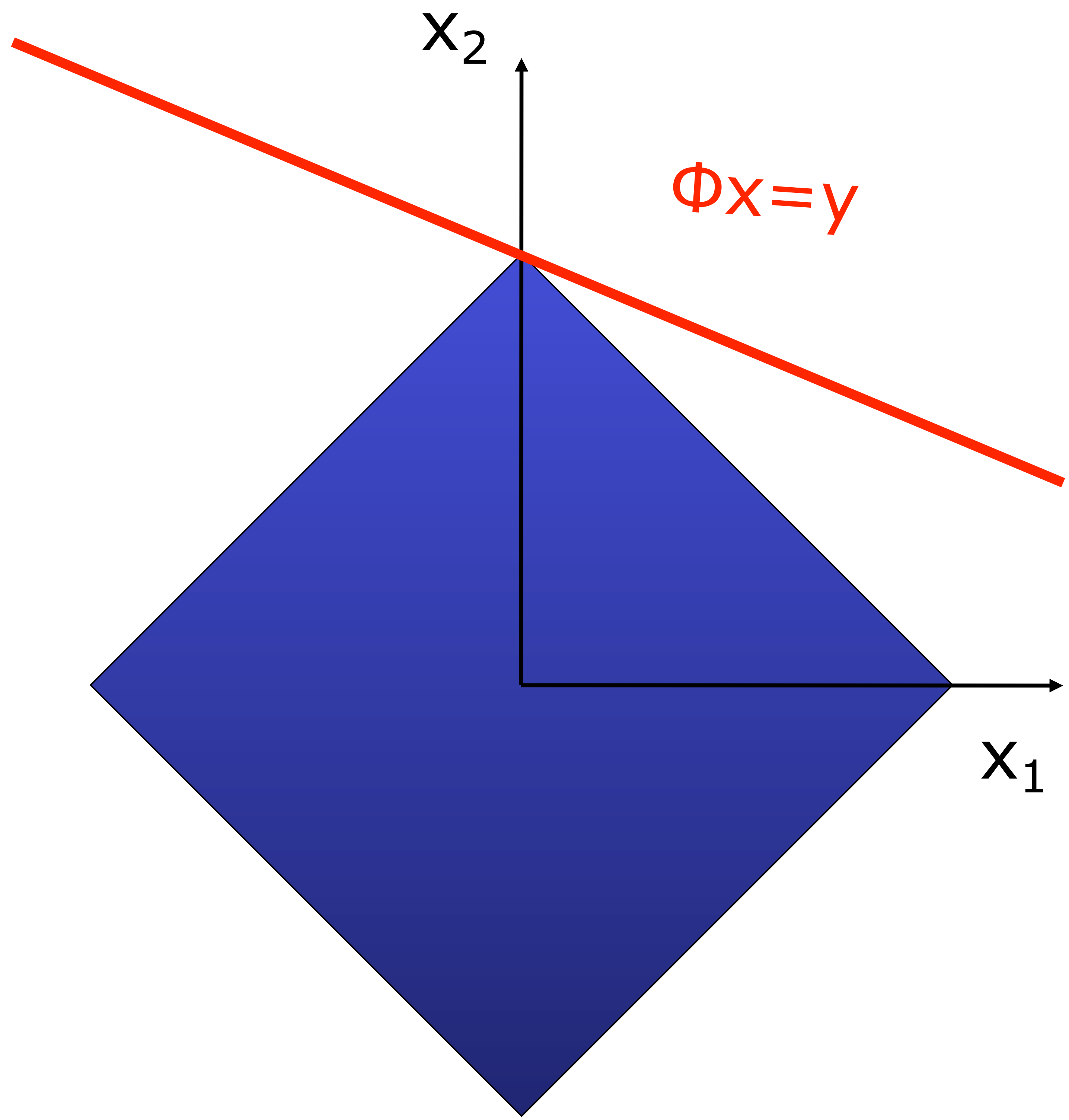
- Convex hull is the unit ball of the l_1 norm

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

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



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

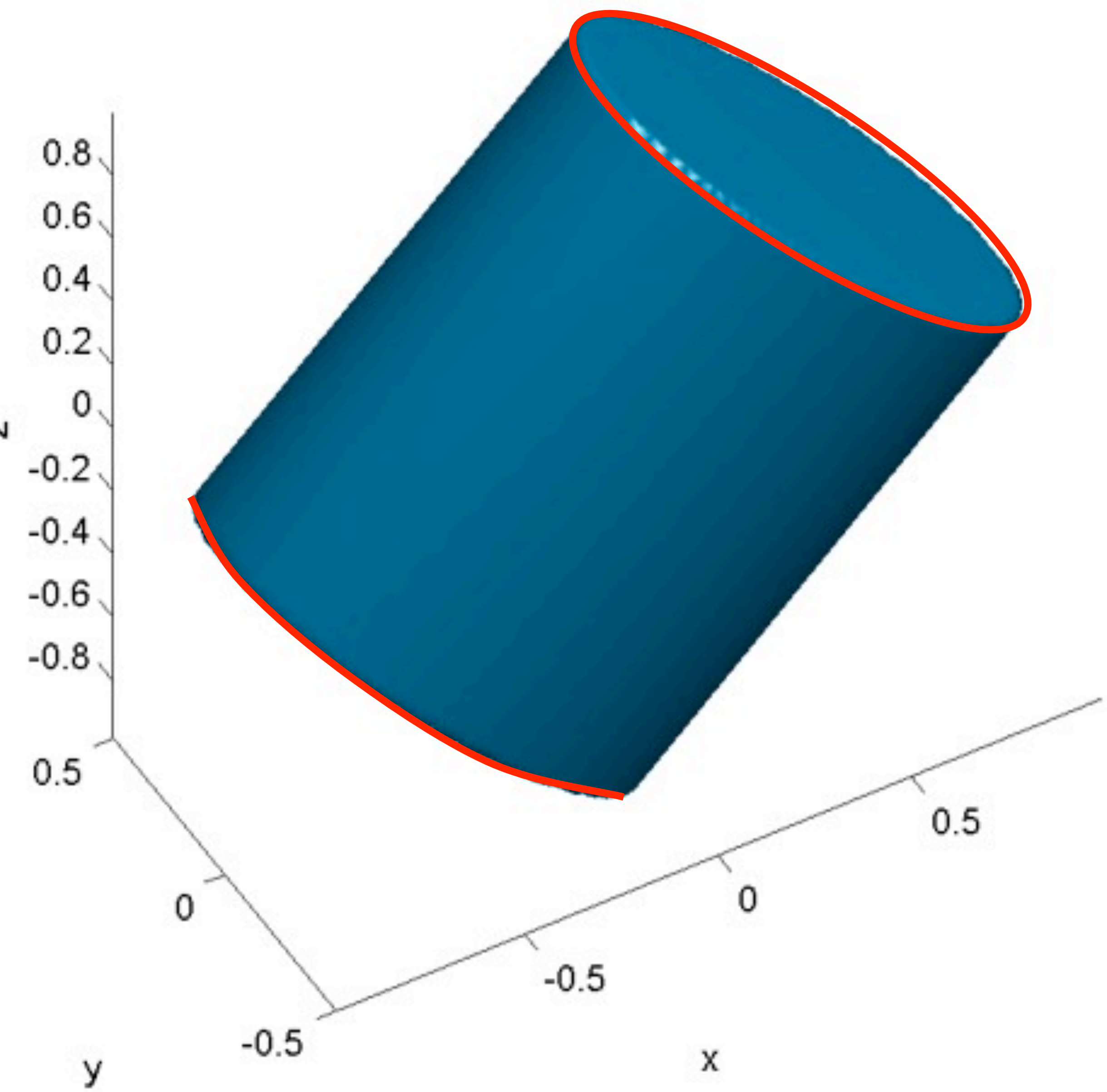


Compressed Sensing: Candes, Romberg, Tao, Donoho, Tanner, Etc...

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:

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

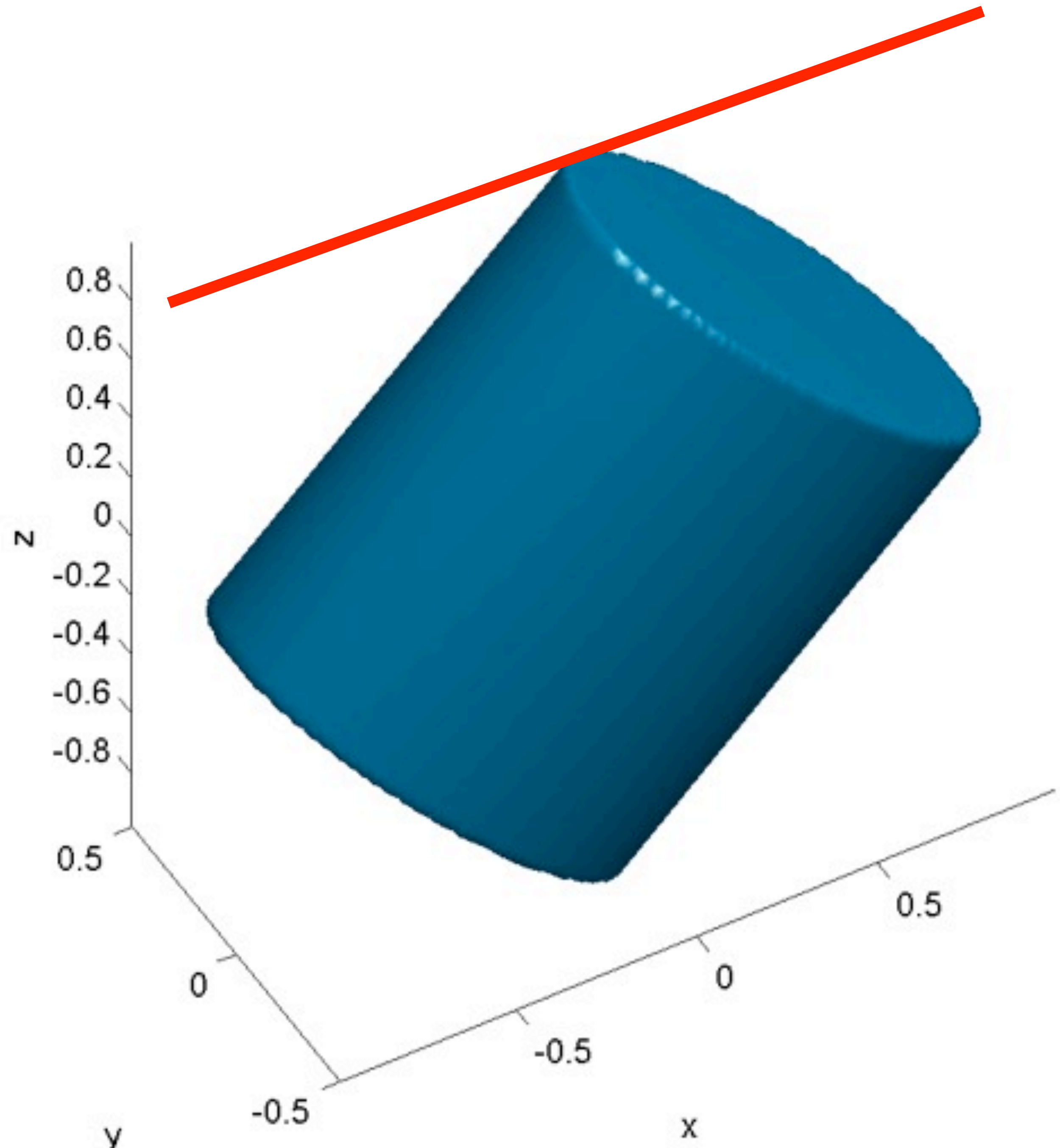
$$\|X\|_* = \sum_i \sigma_i(X)$$

- 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

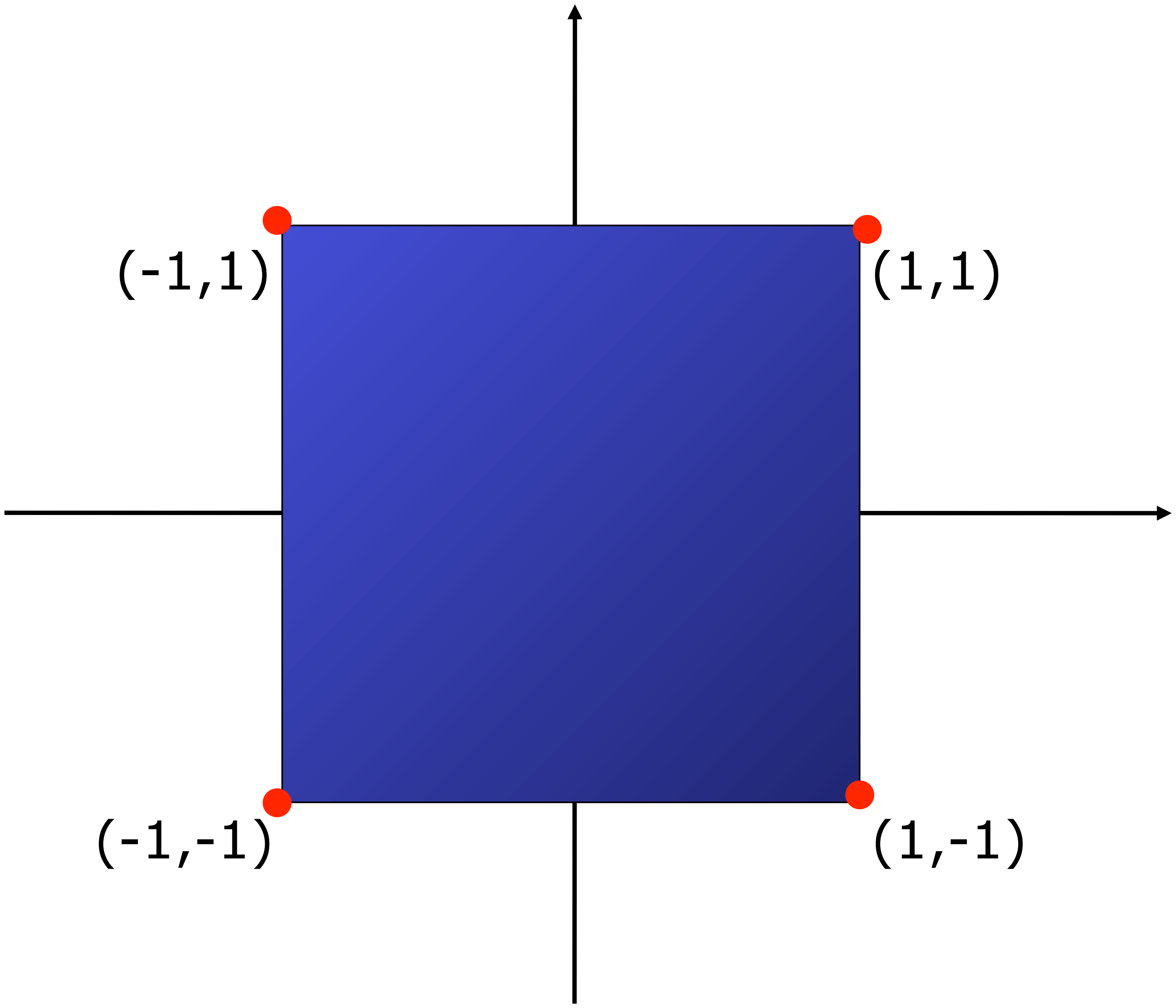
Integer Programming

- Integer solutions:
  all components of x are ±1

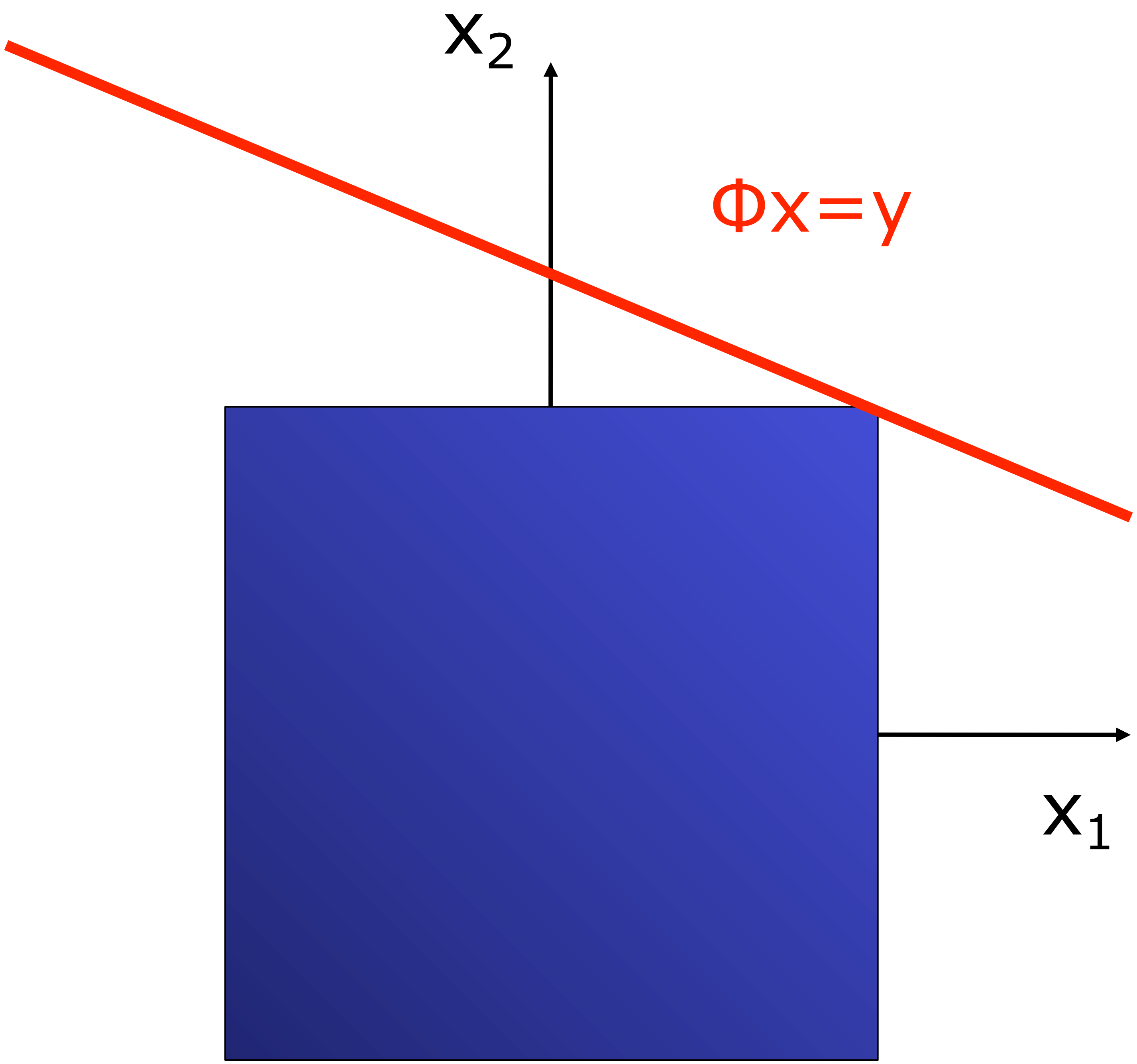
- Convex hull is the
  unit ball of the l_1 norm

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

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



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



Donoho and Tanner 2008
Mangasarian and Recht. 2009.

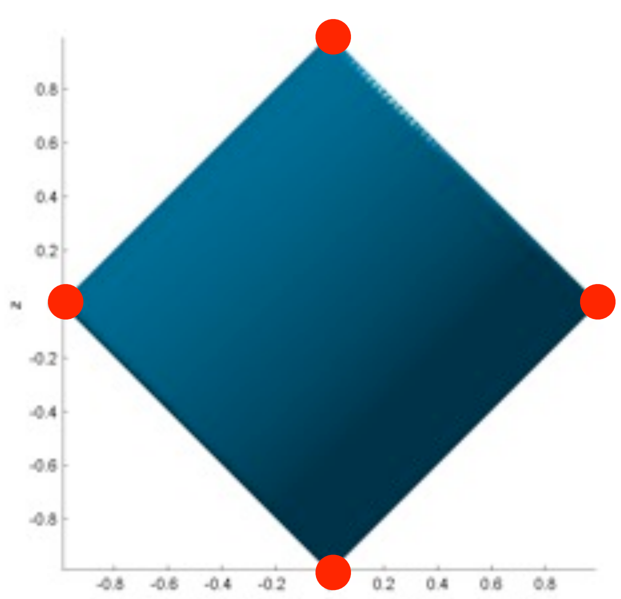
Parsimonious Models

rank

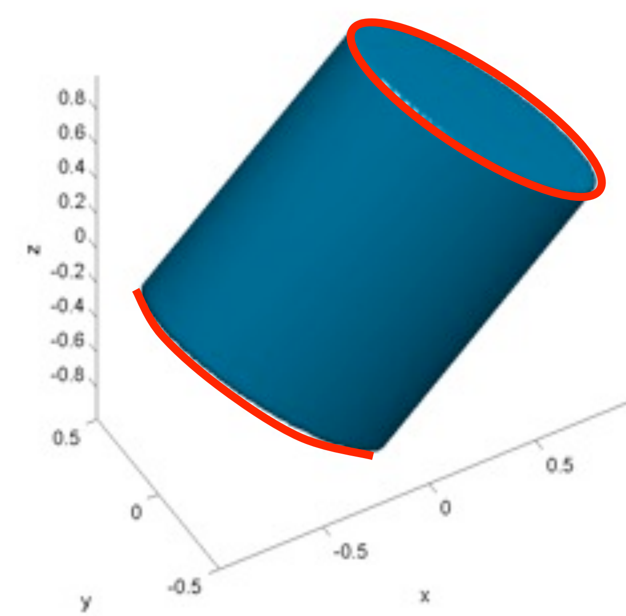
$$x = \sum_{k=1}^{r} w_k \alpha_k$$

model weights atoms

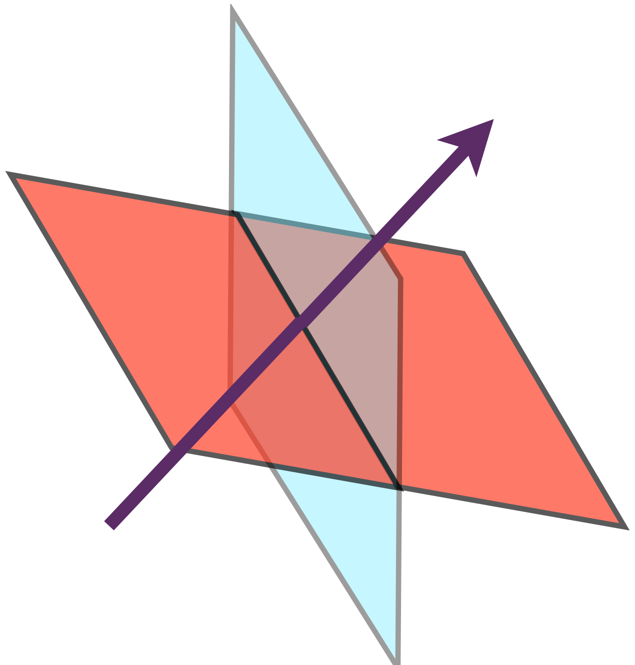
- Search for best linear combination of fewest atoms
- “rank” = fewest atoms needed to describe the model

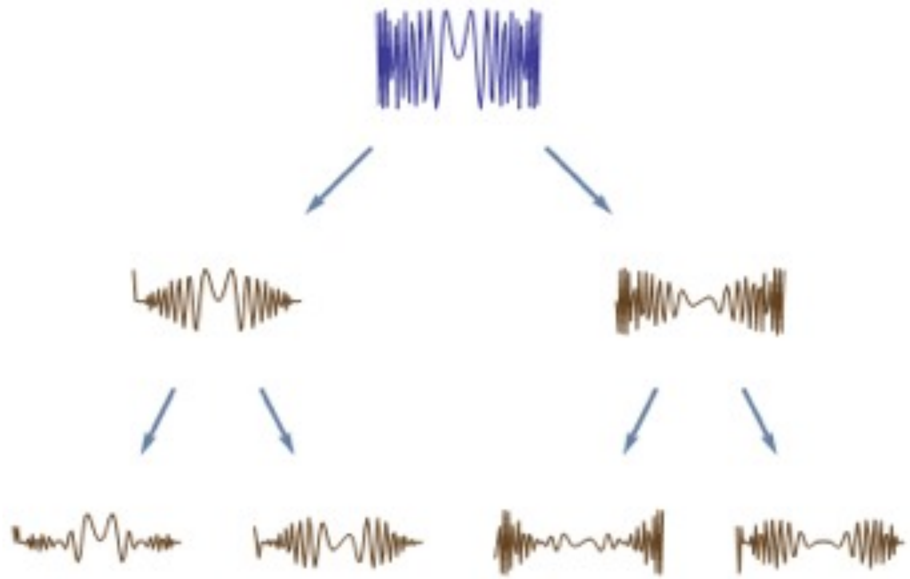


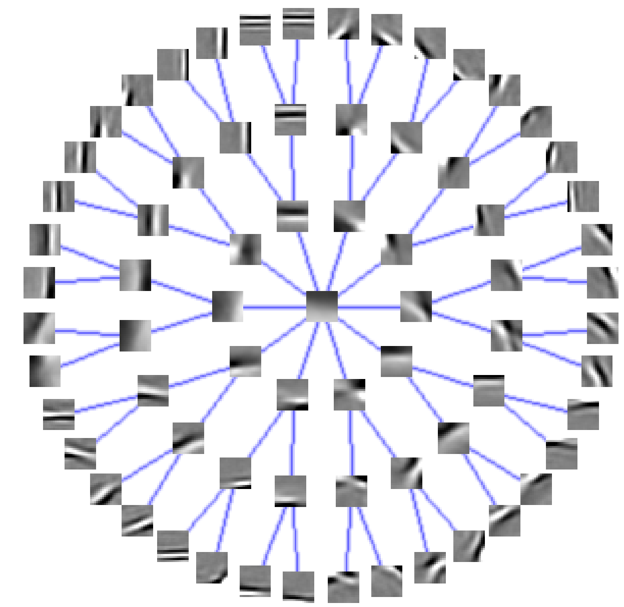
$$\|x\|_{\mathcal{A}} \equiv \inf_{(w,\alpha)} \sum_{k=1}^{r} |w_k|$$



Union of Subspaces







- 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

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

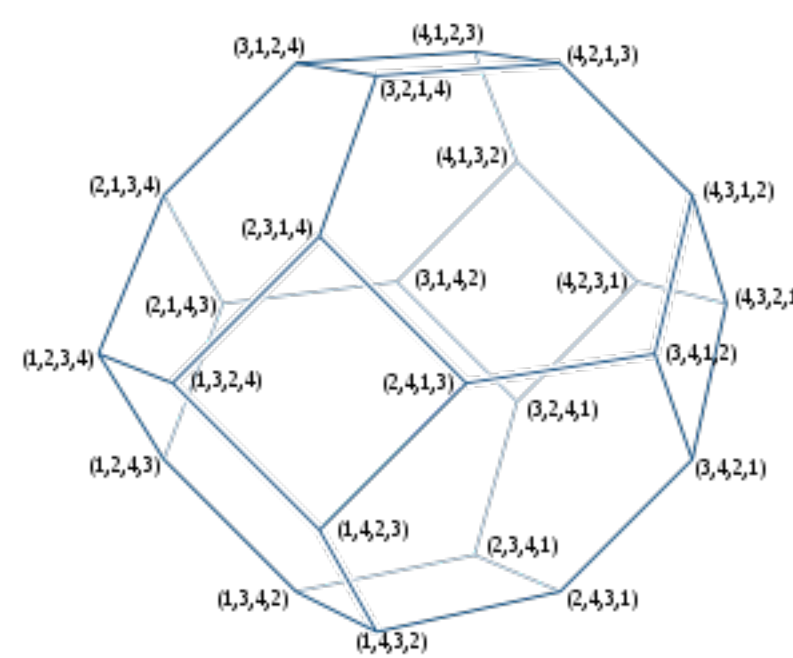
- Proposed by Jacob, Obozinski and Vert (2009).

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] $\longrightarrow$

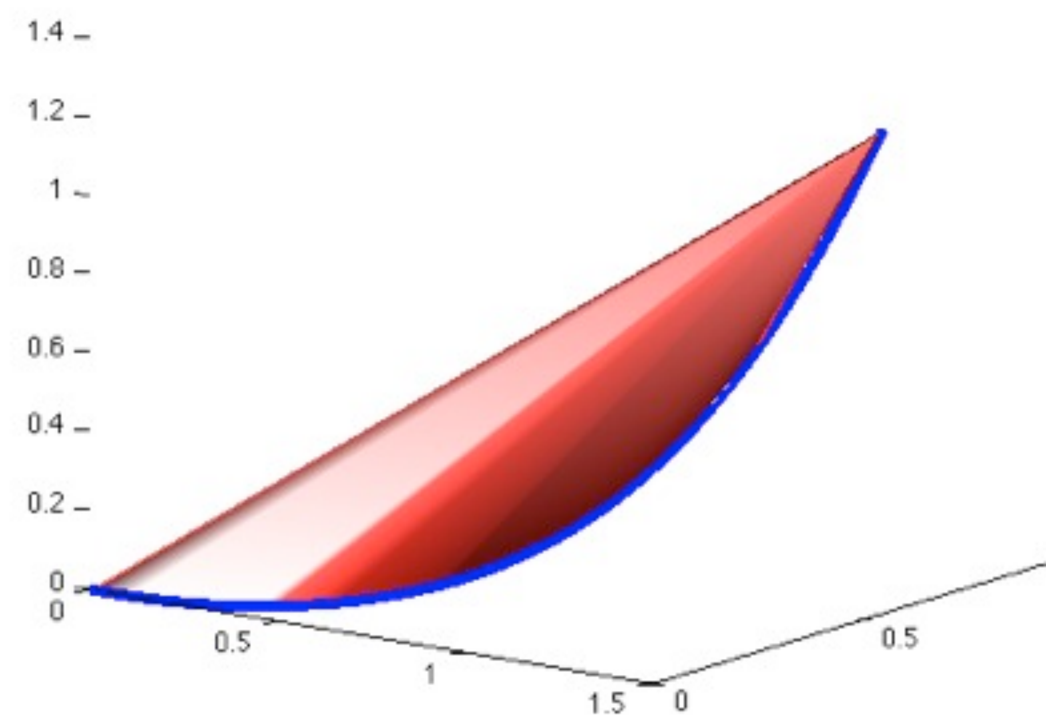


Moment Curve

- Curve of $[1,t,t^2,t^3,t^4,\ldots]$, $t \in T$, some basic set.
- System Identification, Image Processing, Numerical Integration, Statistical Inference...

$$\sum_{k=1}^{r} \alpha_k \begin{bmatrix} 1 \\ e^{\phi_k i} \\ e^{2\phi_k i} \\ e^{3\phi_k i} \end{bmatrix} \begin{bmatrix} 1 \\ e^{\phi_k i} \\ e^{2\phi_k i} \\ e^{3\phi_k i} \end{bmatrix}^*$$

$$\boxed{\mu_t \sim \mathbb{E}[e^{t\phi i}]}$$



- Convex hull is characterized by linear matrix inequalities (Toeplitz psd, Hankel psd, etc)

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.

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

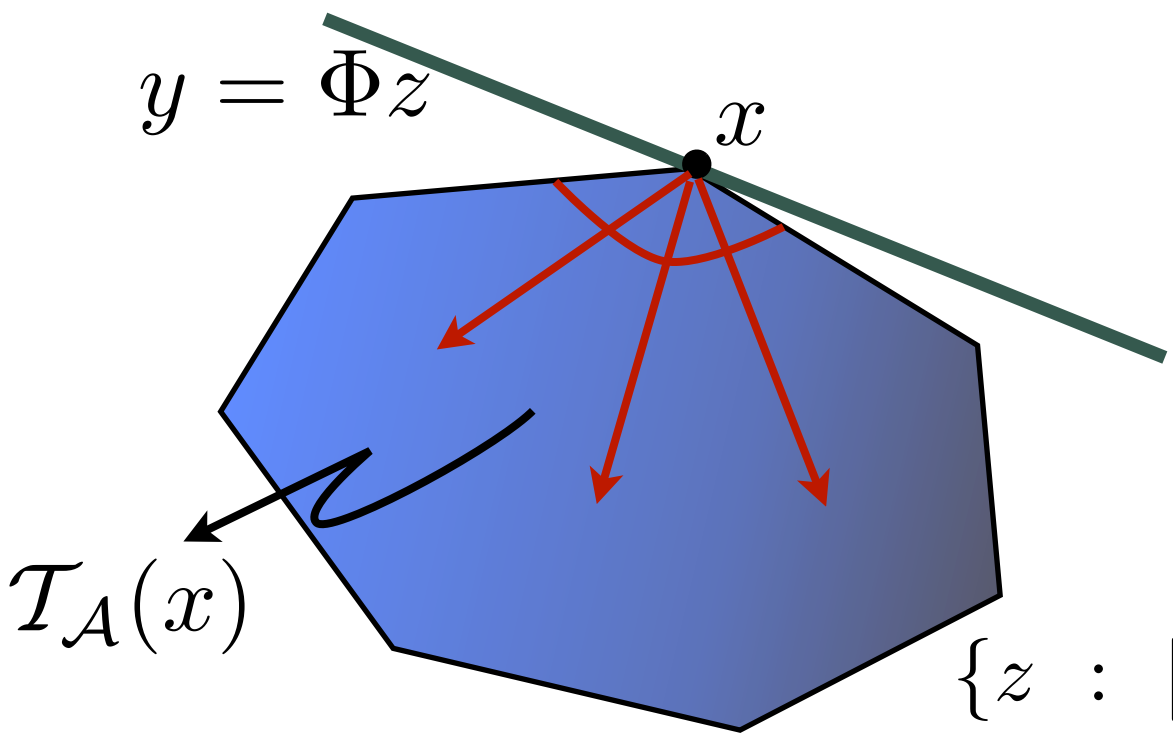
$$\|f - f_n\|_{\mathcal{L}_2} \leq \frac{c_0 \|f\|_{\mathcal{A}}}{\sqrt{n}}$$

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

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

Tangent Cones

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



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

- 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.*

$$\boxed{g \sim \mathcal{N}(0, I_n)}$$

- **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*.

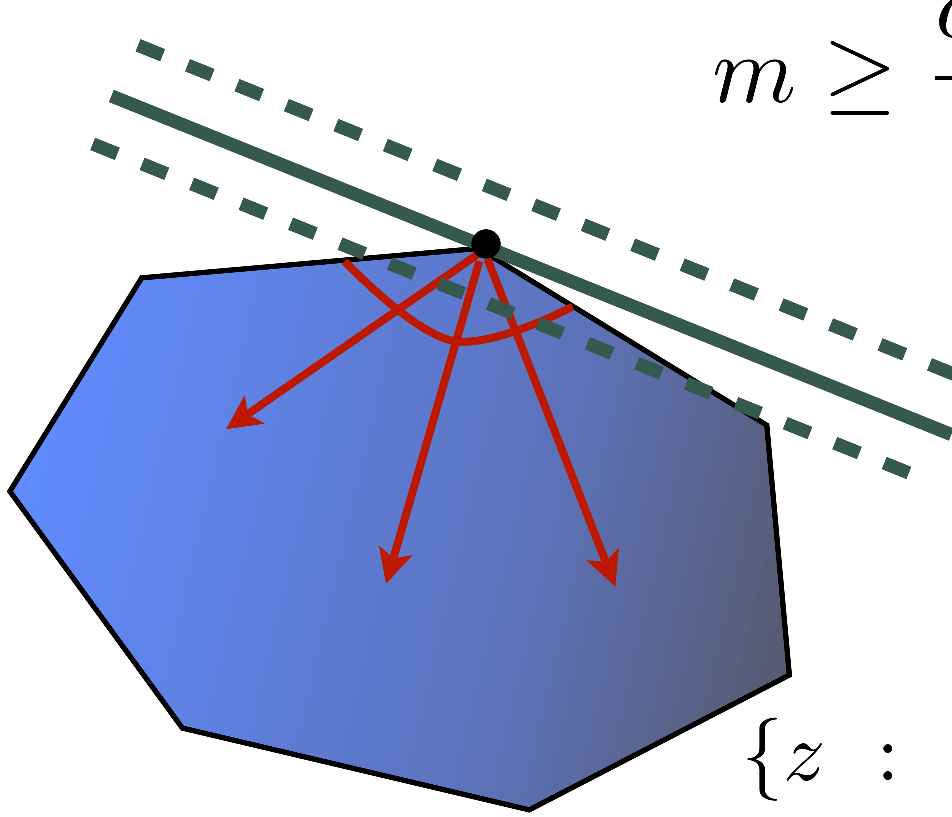
Robust Recovery

- 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

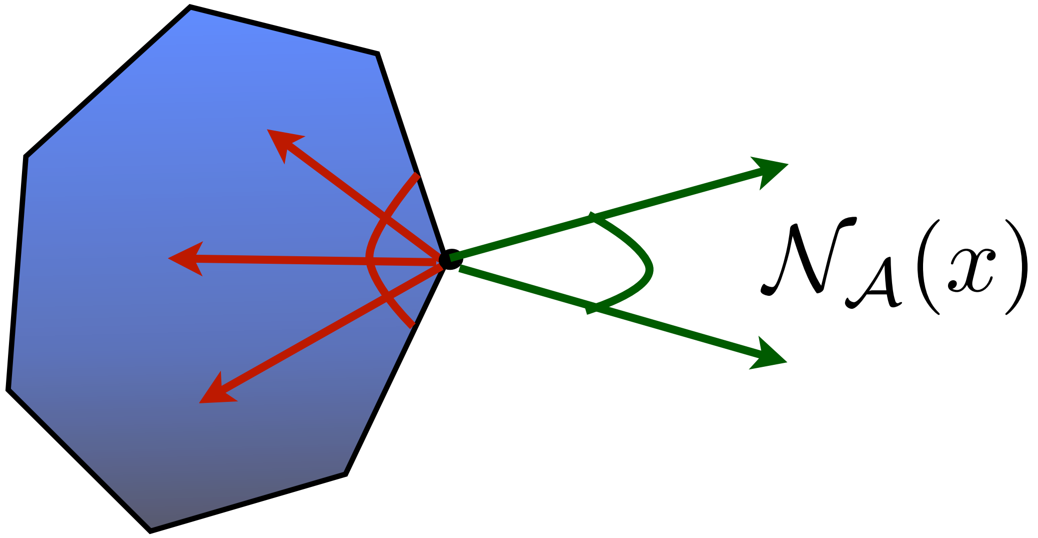
$$\begin{aligned} w(C) &= \mathbb{E}\left[\max_{\substack{v\in C\\ \|v\|=1}} \langle v, g\rangle\right] \\ &\leq \mathbb{E}\left[\max_{\substack{v\in C\\ \|v\|\leq 1}} \langle v, g\rangle\right] \\ &= \mathbb{E}\left[\min_{u\in C^*} \|g-u\|\right] \end{aligned}$$

- C^* is the polar cone.

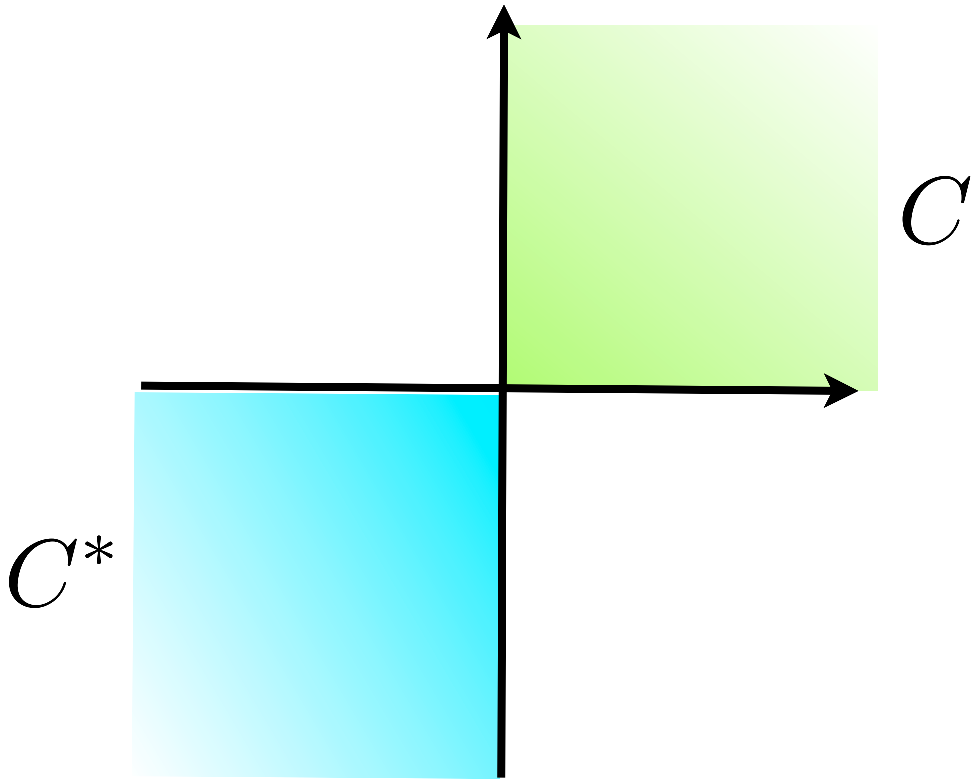
$$C^* = \{w \ : \ \langle w, z\rangle \leq 0 \ \forall\, z\in C\}$$

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

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



Dual Widths



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

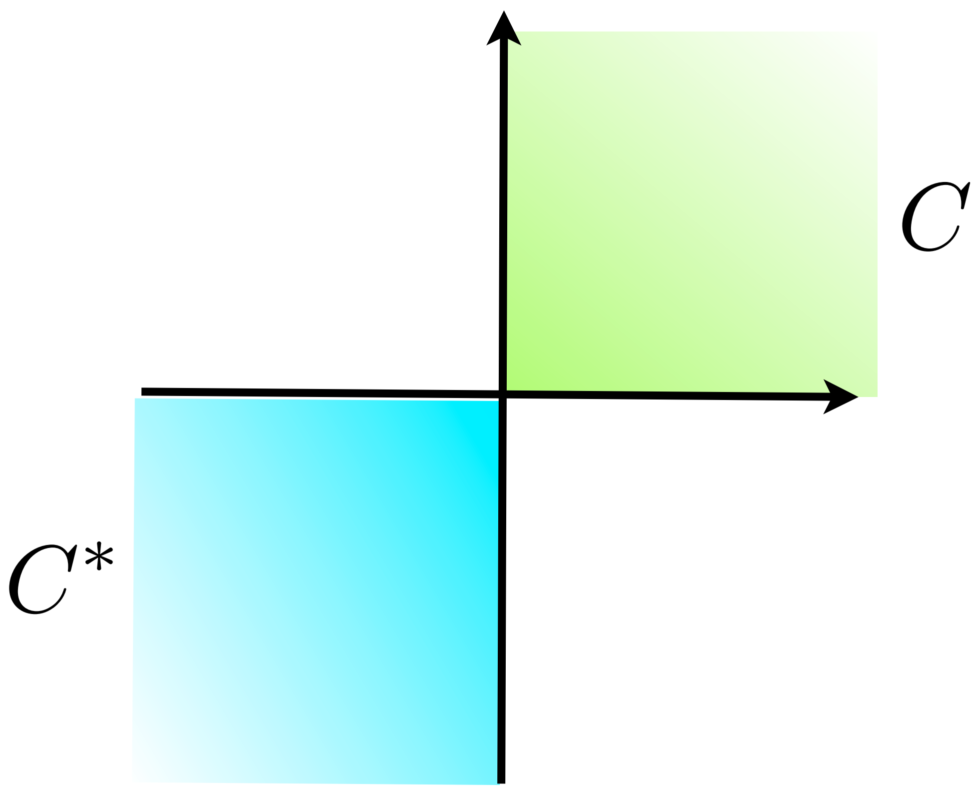
FACT:
$$x = \Pi_C(x) + \Pi_{C^*}(x)$$
$$\langle \Pi_C(x), \Pi_{C^*}(x)\rangle = 0$$

Proposition: $w(C)^2 + w(C^*)^2 \leq n$

$$\begin{aligned} w(C)^2 \leq \mathbb{E}_g\left[\mathrm{dist}(g, C^*)^2\right] &= \mathbb{E}_g\left[\|\Pi_C(g)\|^2\right] \\ &= \mathbb{E}_g\left[\|g\|^2 - \|\Pi_{C^*}(g)\|^2\right] \\ &= n - \mathbb{E}_g\left[\|\Pi_{C^*}(g)\|^2\right] \\ &= n - \mathbb{E}_g\left[\mathrm{dist}(g, C)^2\right] \leq n - w(C^*)^2 \end{aligned}$$

Symmetry I - self duality

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



$$w(C) = w(C^*)$$

$$+$$

$$w(C)^2 + w(C^*)^2 \leq n$$

$$\Downarrow$$

$$w(C)^2 \leq n/2$$

Spectral Norm Ball

- How many measurements to recover a unitary matrix?

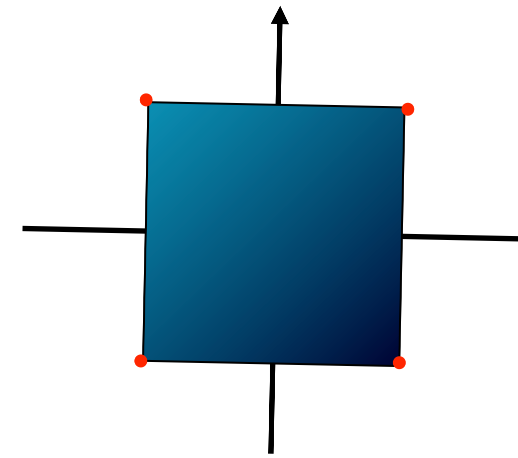
$$\mathcal{T}_{\mathcal{A}}(U) = S - P$$

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

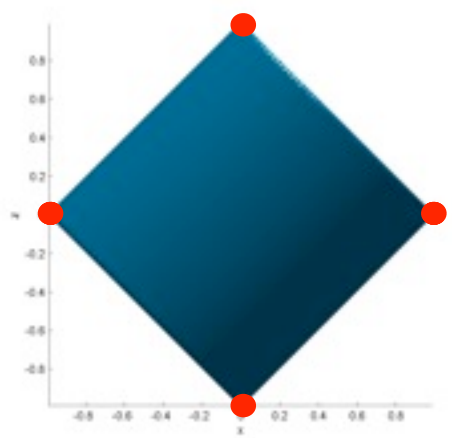
- These two sets are orthogonal, thus

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

Re-derivations



- Hypercube: $m \geq n/2$

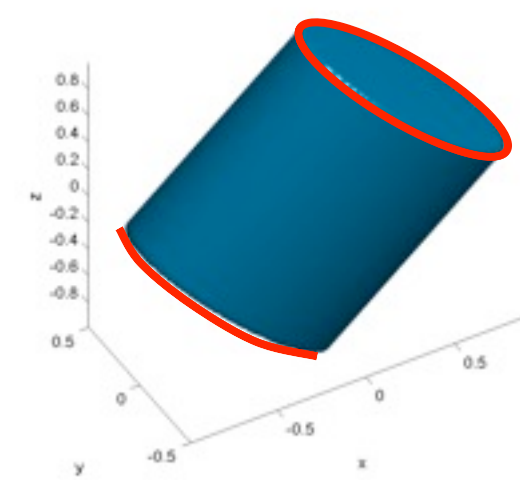


- Sparse Vectors, n vector, sparsity s<0.25n

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

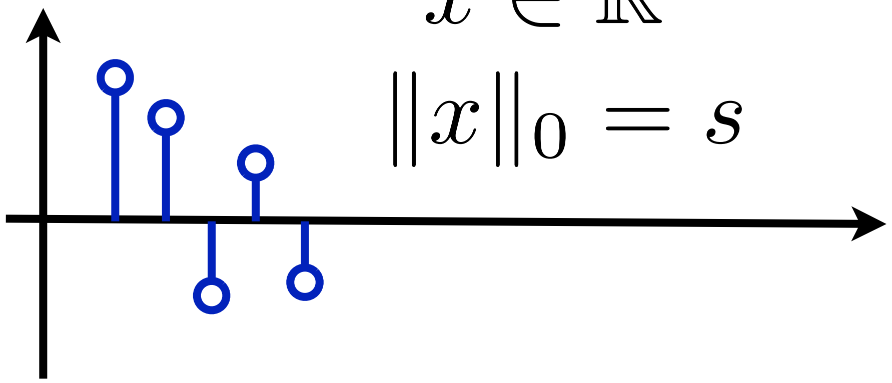
- Block sparse, M groups (possibly overlapping), maximum group size B, k active groups

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



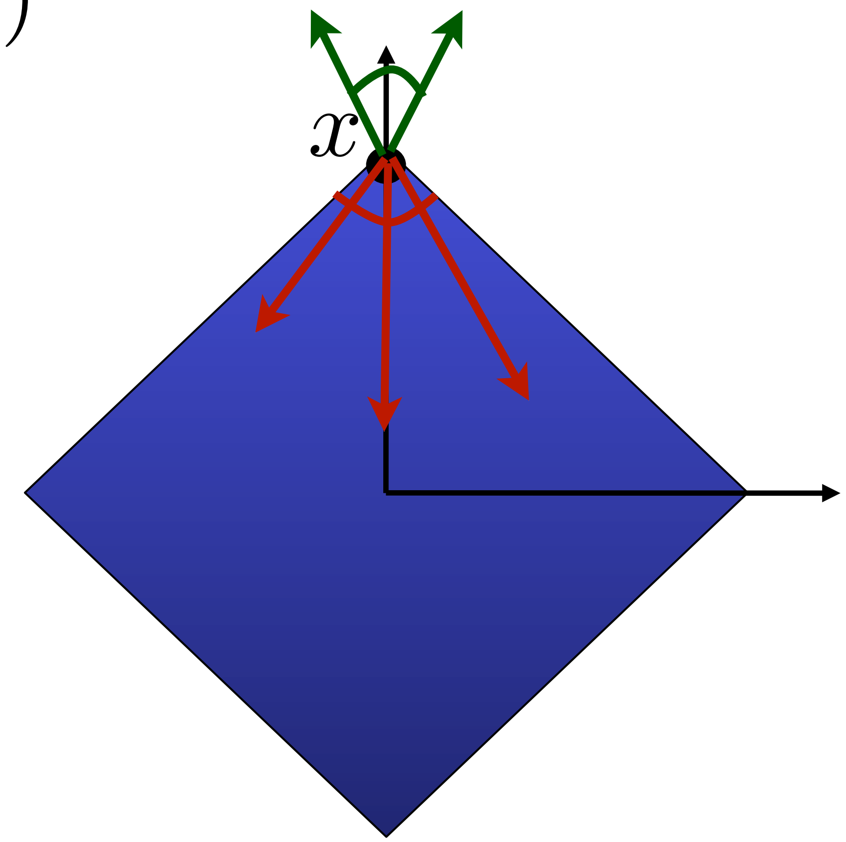
- Low-rank matrices: n_1 x n_2, (n_1<n_2), rank r

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

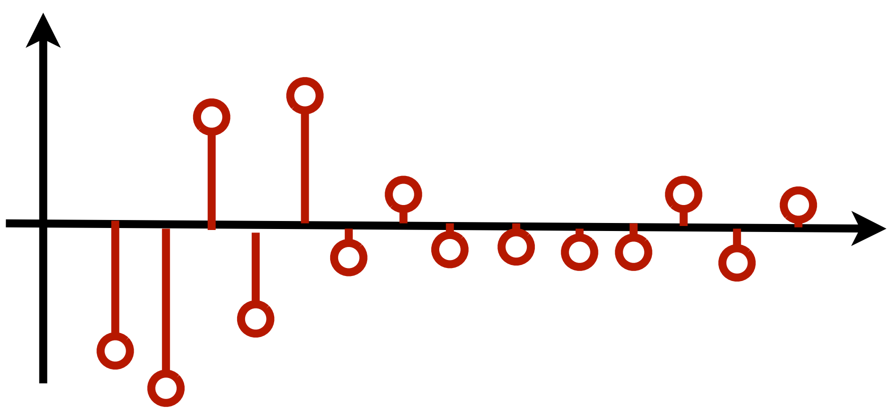


$$x \in \mathbb{R}^n$$
$$\|x\|_0 = s$$

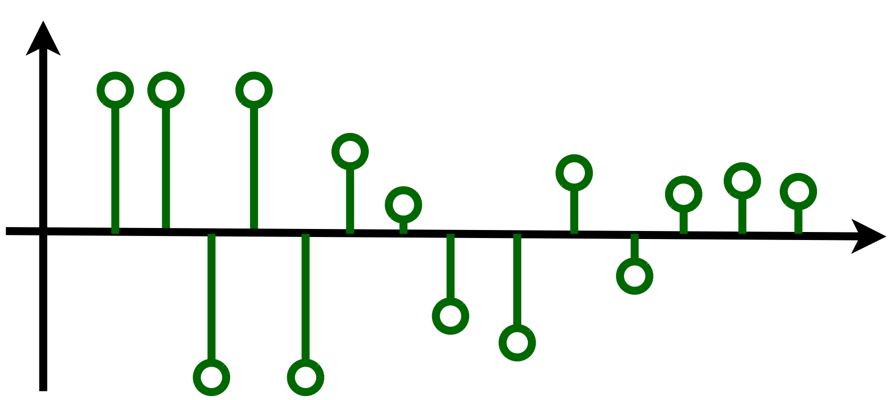
$$T = \operatorname{supp}(x) \subseteq \{1, \ldots, n\}$$
$$T^c = \{1, \ldots, n\} \setminus \operatorname{supp}(x)$$
$$\sigma_T = \operatorname{sign}(x_T)$$

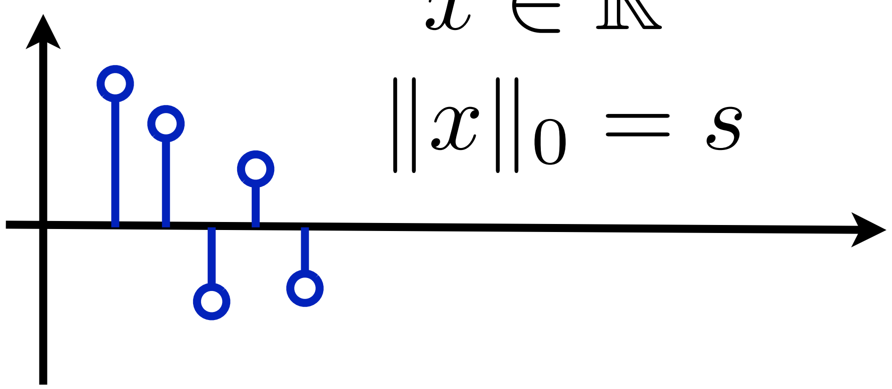


$$\mathcal{T}_{\mathcal{A}}(x) = \{z \in \mathbb{R}^n \ : \ \langle \sigma_T, z_T \rangle \leq \|z_{T^c}\|_1\}$$



$$\mathcal{N}_{\mathcal{A}}(x) = \{u \in \mathbb{R}^n \ : \ u_T = t\sigma_T, \ \ \|u_{T^c}\|_\infty \leq t \ \text{ for some } \ t \geq 0\}$$





$$x \in \mathbb{R}^n$$
$$\|x\|_0 = s$$

$$T = \operatorname{supp}(x) \subseteq \{1, \dots, n\}$$
$$T^c = \{1, \dots, n\} \setminus \operatorname{supp}(x)$$
$$\sigma_T = \operatorname{sign}(x_T)$$

$$\mathcal{N}_{\mathcal{A}}(x) = \{u \in \mathbb{R}^n \ : \ u_T = t\sigma_T, \ \|u_{T^c}\|_\infty \le t \ \text{ for some } \ t \ge 0\}$$

Given: $g \sim \mathcal{N}(0, I_n)$ Find a nearby: $u(g) \in \mathcal{N}_{\mathcal{A}}(x)$

$$\boxed{u_i(g) = \begin{cases} g_i & i \in T^c \\ \sigma_i \|g_{T_c}\|_\infty & i \in T \end{cases}}$$

$$\begin{aligned} w(\mathcal{T}_{\mathcal{A}}(x))^2 \le \mathbb{E}[\|u(g) - g\|^2] &= \mathbb{E}[\|u_T(g) - g_T\|^2] \\ &= \mathbb{E}[\|u_T(g)\|^2] + \mathbb{E}[\|g_T\|^2] \\ &= s\mathbb{E}[\|g_{T^c}\|_\infty^2] + s \\ &\le 2s\log(n-s) + 2s \end{aligned}$$

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 $\qquad r_k = \Phi z_k - y$

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

- Similar algorithm for atomic norm constraint

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

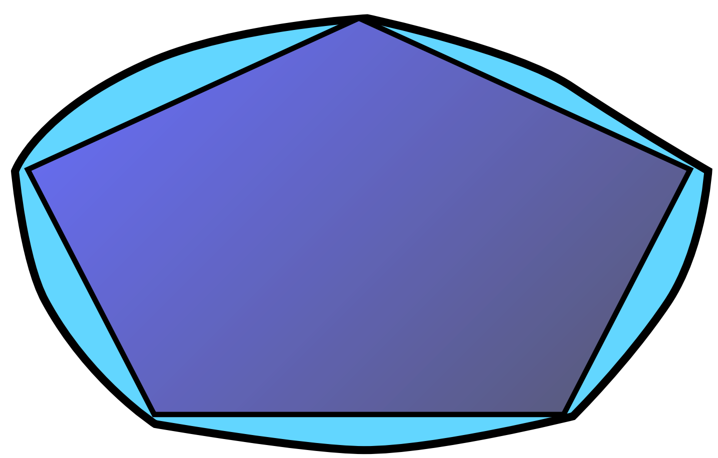
Relaxations

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

- 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

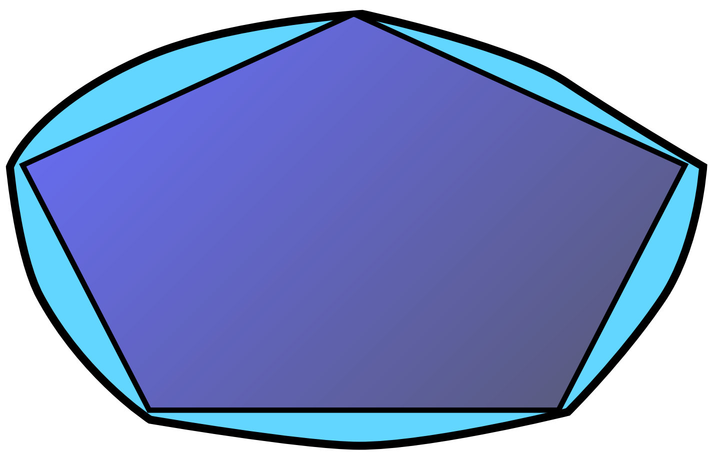
- Hierarchy of relaxations based on *θ-Bodies* yield progressively tighter bounds on the atomic norm

Theta Bodies

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

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

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



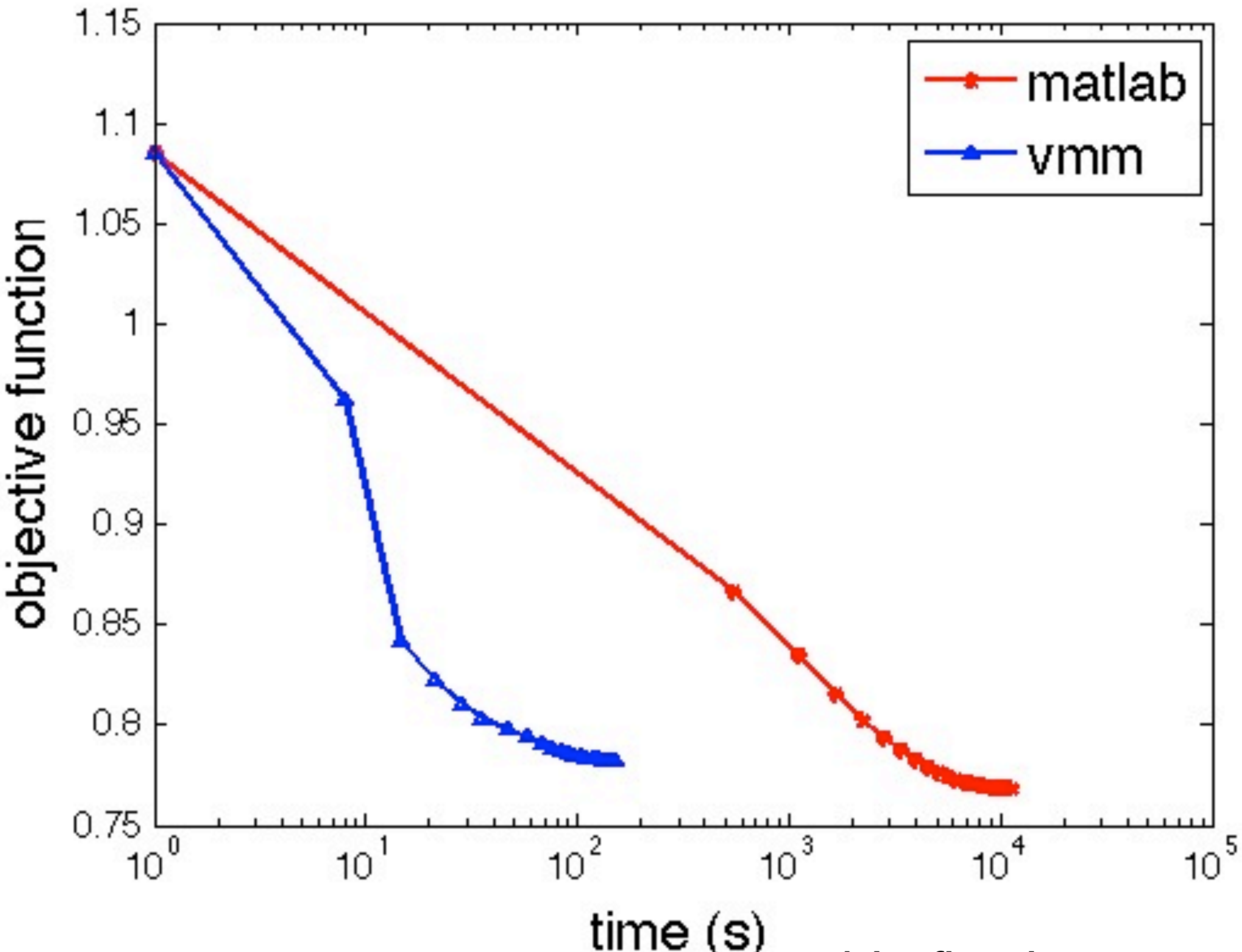
$$q = h + g$$

$h(x) \geq 0 \ \forall x$ — positive everywhere

$g \in I$ — 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*)

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.
- Submitted to VLDB11 with Christopher Ré

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

- Algorithms and 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