# Pursuit Algorithms for Sparse Representations

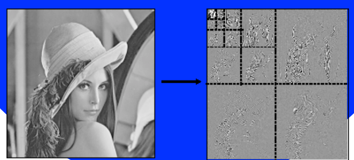
Rémi Gribonval
EPI METISS
INRIA Rennes - Bretagne Atlantique

remi.gribonval@inria.fr
http://www.irisa.fr/metiss/members/remi

# Summary

Notion of sparsity
(Fourier, wavelets, ...)

Compression
Representation
Description
Classification




Sparsity

Denoising
Blind source separation
Compressed sensing
...

Natural / traditional role

Novel indirect role

Sparsity = low cost (bits, computations, ...)
**Direct objective**

Sparisty = prior knowledge, regularization
**Tool for inverse problems**

# Overview

- Convex optimization algorithms
- Greedy algorithms
- Comparison of complexities
- Exact recovery conditions for Lp minimization

# Overall compromise

- Approximation quality

$$\|\mathbf{A}x - \mathbf{b}\|_2$$

- Ideal sparsity measure : $\ell^0$ "norm"

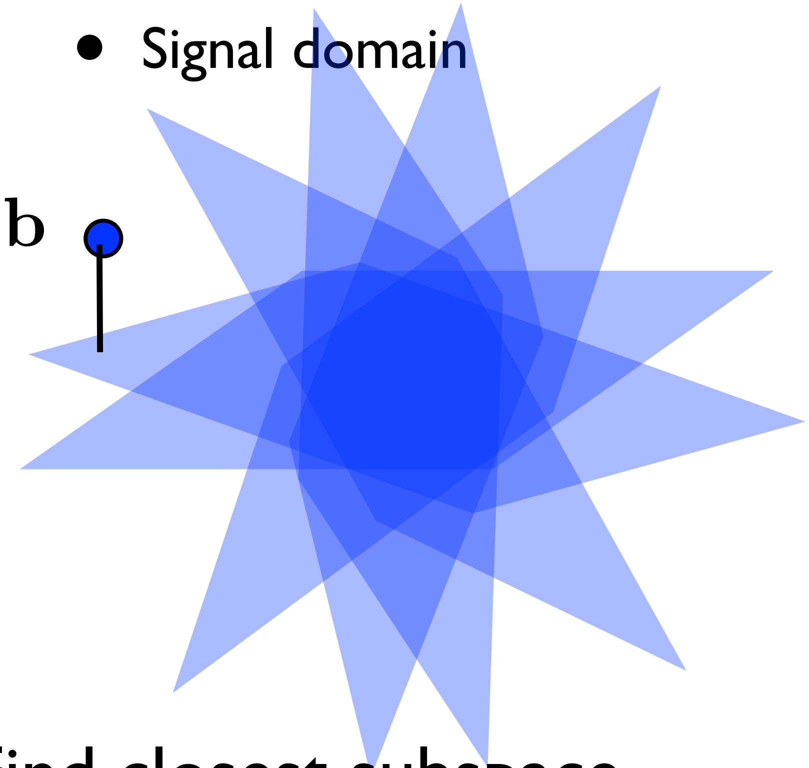
$$\|x\|_0 := \sharp\{n,\ x_n \neq 0\} = \sum_n |x_n|^0$$

- "Relaxed" sparsity measures

$$0 < p < \infty,\ \|x\|_p := \Big(\sum_n |x_n|^p\Big)^{1/p}$$

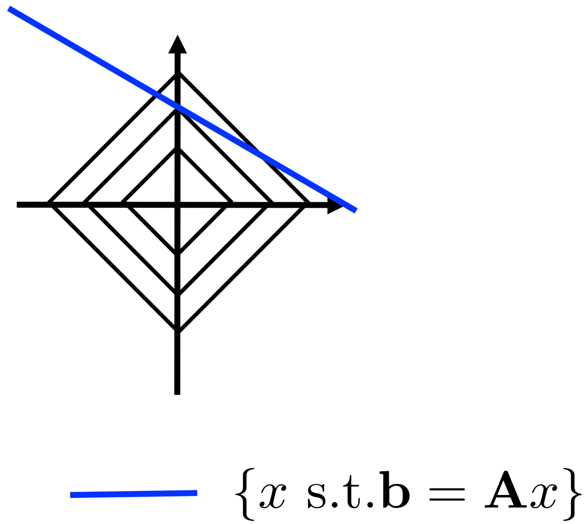
# Two geometric viewpoints

- Signal domain




Find closest subspace through correlations $\mathbf{A}^T\mathbf{b}$

- Coefficient domain




Find sparsest representation through convex optimization

# Algorithms for L1: Linear Programming

- L1 minimization problem of size *m* x *N*

Basis Pursuit (BP)
LASSO

$$\min_x \|x\|_1, \ \text{s.t.} \ \mathbf{A}x = \mathbf{b}$$

- Equivalent linear program of size *m* x 2*N*

$$\min_{z \geq 0} \mathbf{c}^T z, \ \text{s.t.} \ [\mathbf{A}, -\mathbf{A}]z = \mathbf{b}$$

$$\mathbf{c} = (c_i), \ c_i = 1, \forall i$$

# L1 regularization: Quadratic Programming

- L1 minimization problem of size *m* x *N*

Basis Pursuit Denoising (BPDN)

$$\min_{x} \frac{1}{2}\|\mathbf{b} - \mathbf{A}x\|_2^2 + \lambda\|x\|_1$$

- Equivalent quadratic program of size *m* x 2*N*

$$\min_{z \geq 0} \frac{1}{2}\|\mathbf{b} - [\mathbf{A}, -\mathbf{A}]z\|_2^2 + \mathbf{c}^T z$$

$$\mathbf{c} = (c_i),\ c_i = 1, \forall i$$

# Generic approaches vs specific algorithms

- Many algorithms for linear / quadratic programming
- Matlab Optimization Toolbox: `linprog` /qp
- But ...
  - The problem size is "doubled"
  - Specific structures of the matrix **A** can help solve BP and BPDN more efficiently
  - More efficient toolboxes have been developed
- CVX package (Michael Grant & Stephen Boyd):
  - http://www.stanford.edu/~boyd/cvx/

# Overview

- **Convex optimization algorithms**
- Greedy algorithms
- Comparison of complexities
- Exact recovery conditions for Lp minimization

# Example: orthonormal **A**

- Assumption : *m*=*N* and **A** is *orthonormal*

$$\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{Id}_N$$
$$\|\mathbf{b} - \mathbf{A}x\|_2^2 = \|\mathbf{A}^T\mathbf{b} - x\|_2^2$$

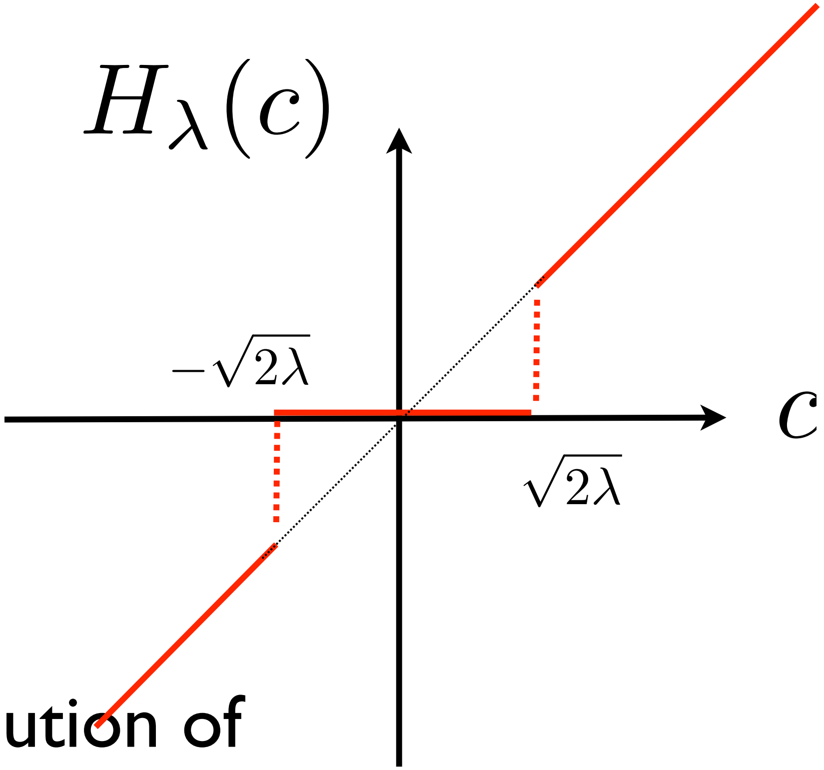
- Expression of BPDN criterion to be minimized

$$\sum_n \frac{1}{2}\big((\mathbf{A}^T\mathbf{b})_n - x_n\big)^2 + \lambda|x_n|^p$$

- Minimization can be done coordinate-wise

$$\min_{x_n} \frac{1}{2}\big(c_n - x_n\big)^2 + \lambda|x_n|^p$$

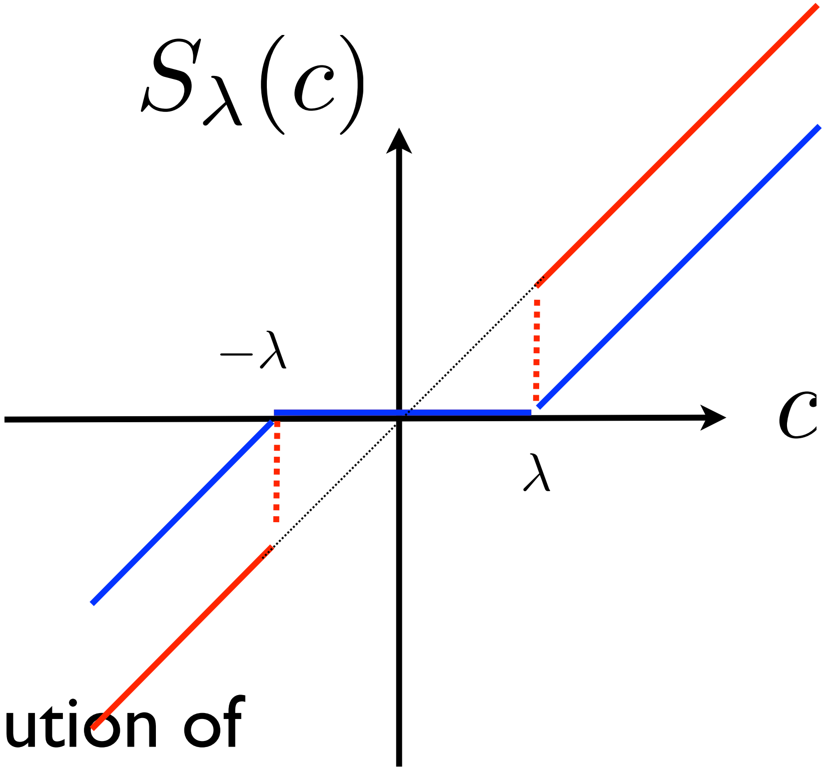
# Hard-thresholding (p=0)




- Solution of

$$\min_x \frac{1}{2}(c-x)^2 + \lambda \cdot |x|^0$$

# Soft-thresholding (p=1)




- Solution of

$$\min_x \frac{1}{2}(c-x)^2 + \lambda \cdot |x|$$

# Iterative thresholding

- Proximity operator

$$\Theta_\lambda^p(c) = \arg\min_x \frac{1}{2}(x-c)^2 + \lambda|x|^p$$

- Goal = compute

$$\arg\min_x \frac{1}{2}\|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda\|x\|_p^p$$

- Approach = iterative alternation between
  - ✦ gradient descent on fidelity term

$$x^{(i+1/2)} := x^{(i)} + \alpha^{(i)}\mathbf{A}^T(\mathbf{b} - \mathbf{A}x^{(i)})$$

  - ✦ thresholding

$$x^{(i+1)} := \Theta_{\lambda^{(i)}}^p(x^{(i+1/2)})$$

# Iterative Thresholding

- **Theorem** : *[Daubechies, de Mol, Defrise 2004, Combettes & Pesquet 2008]*
  - ✦ consider the iterates $x^{(i+1)} = f(x^{(i)})$ defined by the thresholding function, with $p \geq 1$

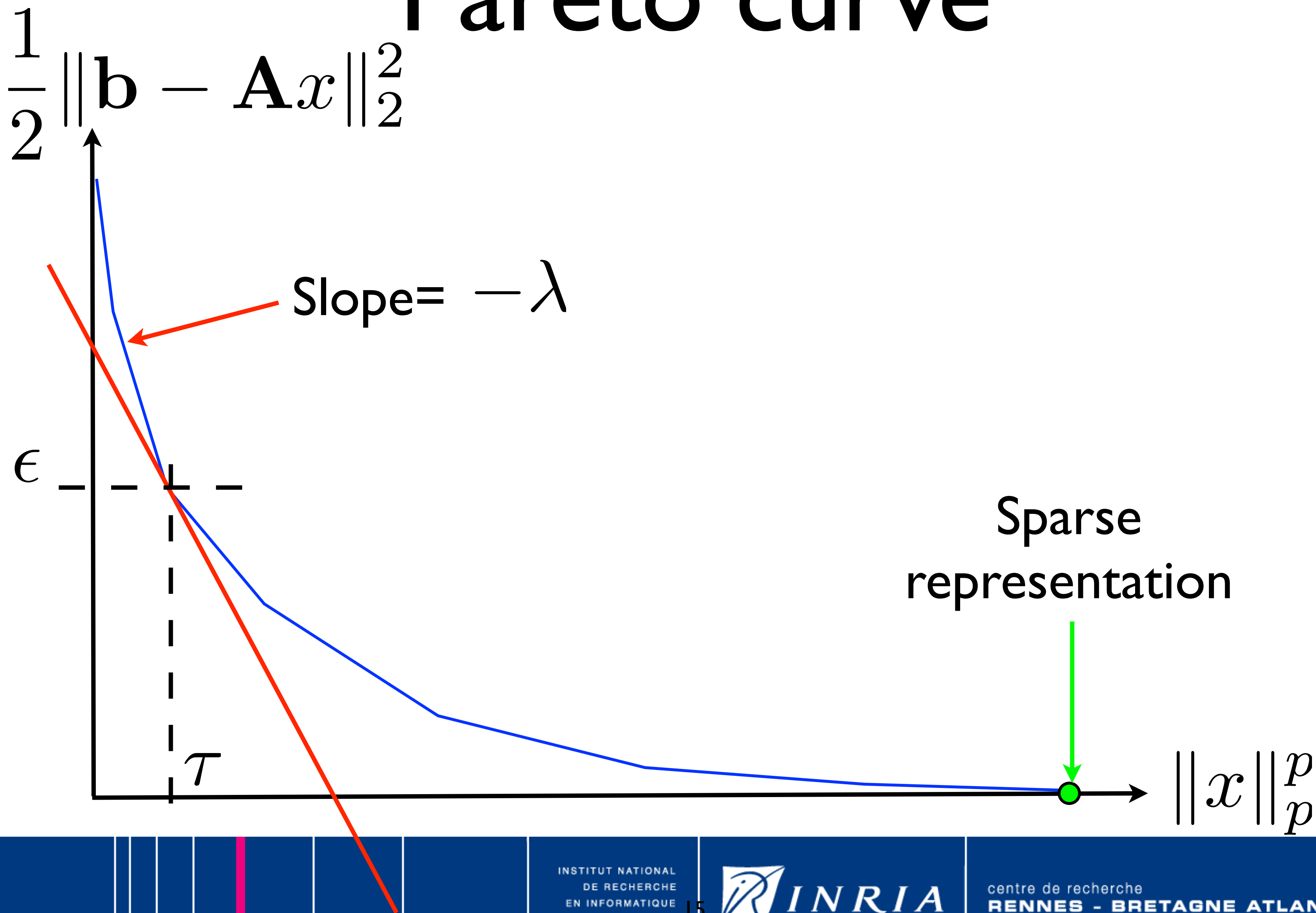
$$f(x) = \Theta^{p}_{\alpha\lambda}(x + \alpha \mathbf{A}^T(\mathbf{b} - \mathbf{A}x))$$

  - ✦ assume that $\forall x,\ \|\mathbf{A}x\|_2^2 \leq c\|x\|_2^2$ and $\alpha < 2/c$
  - ✦ then, the iterates converge strongly to a limit $x^\star$

$$\|x^{(i)} - x^\star\|_2 \rightarrow_{i \rightarrow \infty} 0$$

  - ✦ the limit $x^\star$is a global minimum of $\frac{1}{2}\|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda\|x\|_p^p$
  - ✦ if *p*>1, or if **A** is invertible, $x^\star$ is the *unique* minimum

# Pareto curve



# Path of the solution

- **Lemma**: let $x^\star$ be a local minimum of BPDN

$$\arg\min_x \frac{1}{2}\|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda\|x\|_1$$

- let *I* be its support
- Then $$\mathbf{A}_I^T(\mathbf{A}x^\star - \mathbf{b}) + \lambda \cdot \mathrm{sign}(x_I^\star) = 0$$

$$\|\mathbf{A}_{I^c}^T(\mathbf{A}x^\star - \mathbf{b})\|_\infty < \lambda$$

- In particular

$$x_I = (\mathbf{A}_I^T\mathbf{A}_I)^{-1}\left(\mathbf{A}_I^T\mathbf{b} - \lambda \cdot \mathrm{sign}(x_I)\right)$$

# Homotopy method

- Principle: track the solution $x^\star(\lambda)$ of BPDN along the Pareto curve
- Property:
  - ✦ solution is characterized by its sign pattern through

$$x_I = \left(\mathbf{A}_I^T \mathbf{A}_I\right)^{-1} \left(\mathbf{A}_I^T \mathbf{b} - \lambda \cdot \mathrm{sign}(x_I)\right)$$

  - ✦ for given sign pattern, dependence on $\lambda$ is affine
  - ✦ sign patterns are piecewise constant functions of $\lambda$
  - ✦ overall, the solution is piecewise affine
- Method = iteratively find breakpoints

# Overview

- Convex optimization algorithms
- **Greedy algorithms**
- Comparison of complexities
- Exact recovery conditions for Lp minimization

# Matching Pursuit (MP)

- Matching Pursuit (*aka* Projection Pursuit, CLEAN)
  - Initialization $\mathbf{r}_0 = \mathbf{b} \qquad i = 1$
  - Atom selection:

$$n_i = \arg\max_n \left|\mathbf{A}_n^T \mathbf{r}_{i-1}\right|$$

  - Residual update

$$\mathbf{r}_i = \mathbf{r}_{i-1} - (\mathbf{A}_{n_i}^T \mathbf{r}_{i-1})\mathbf{A}_{n_i}$$

- Energy preservation (Pythagoras theorem)

$$\|\mathbf{r}_{i-1}\|_2^2 = |\mathbf{A}_{n_i}^T \mathbf{r}_{i-1}|^2 + \|\mathbf{r}_i\|_2^2$$

# Main properties

- Global energy preservation

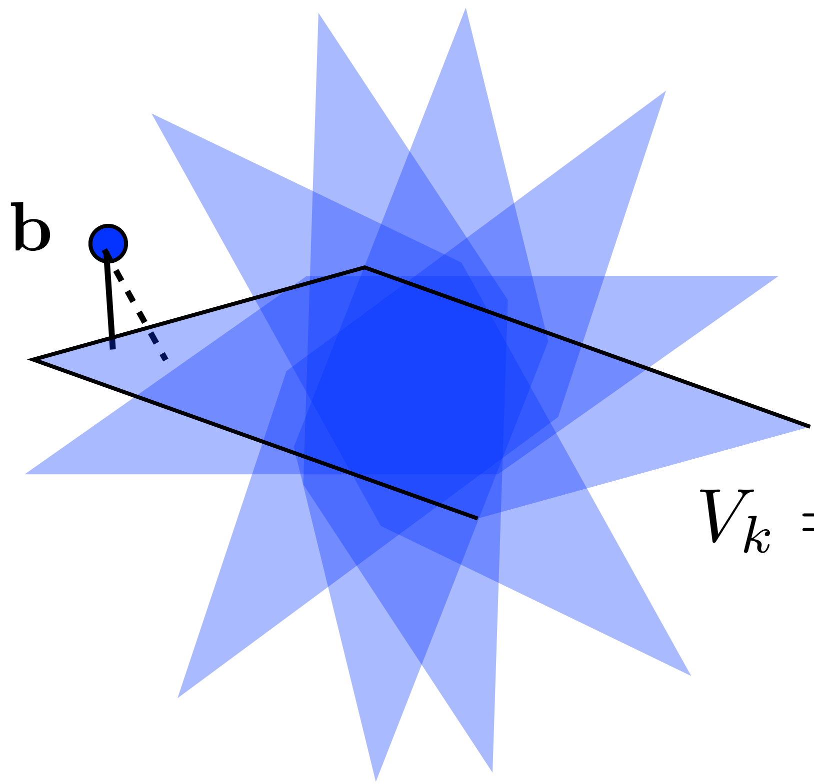
$$\|\mathbf{b}\|_2^2 = \|\mathbf{r}_0\|_2^2 = \sum_{i=1}^{k} |\mathbf{A}_{n_i}^T \mathbf{r}_{i-1}|^2 + \|\mathbf{r}_k\|_2^2$$

- Global reconstruction

$$\mathbf{b} = \mathbf{r}_0 = \sum_{i=1}^{k} (\mathbf{A}_{n_i}^T \mathbf{r}_{i-1}) \mathbf{A}_{n_i} + \mathbf{r}_k$$

- Strong convergence

$$\lim_{i \to \infty} \|\mathbf{r}_i\|_2 = 0$$




$$V_k = \mathrm{span}(\mathbf{A}_n, n \in \Lambda_k)$$

# Orthonormal MP (OMP)

- Observation: after k iterations $\mathbf{r}_k = \mathbf{b} - \sum_{i=1}^{k} \alpha_k \mathbf{A}_{n_i}$
- Approximant belongs to

$$V_k = \mathrm{span}(\mathbf{A}_n, n \in \Lambda_k)$$
$$\Lambda_k = \{n_i, 1 \leq i \leq k\}$$

- Best approximation from $V_k$ = orthoprojection

$$P_{V_k}\mathbf{b} = \mathbf{A}_{\Lambda_k}\mathbf{A}_{\Lambda_k}^{+}\mathbf{b}$$

- **OMP residual update rule** $\mathbf{r}_k = \mathbf{b} - P_{V_k}\mathbf{b}$

# OMP

- Same as MP, except residual update rule
  - Atom selection:

$$n_i = \arg\max_n |\mathbf{A}_n^T \mathbf{r}_{i-1}|$$

  - Index update $\Lambda_i = \Lambda_{i-1} \cup \{n_i\}$
  - *Residual update*

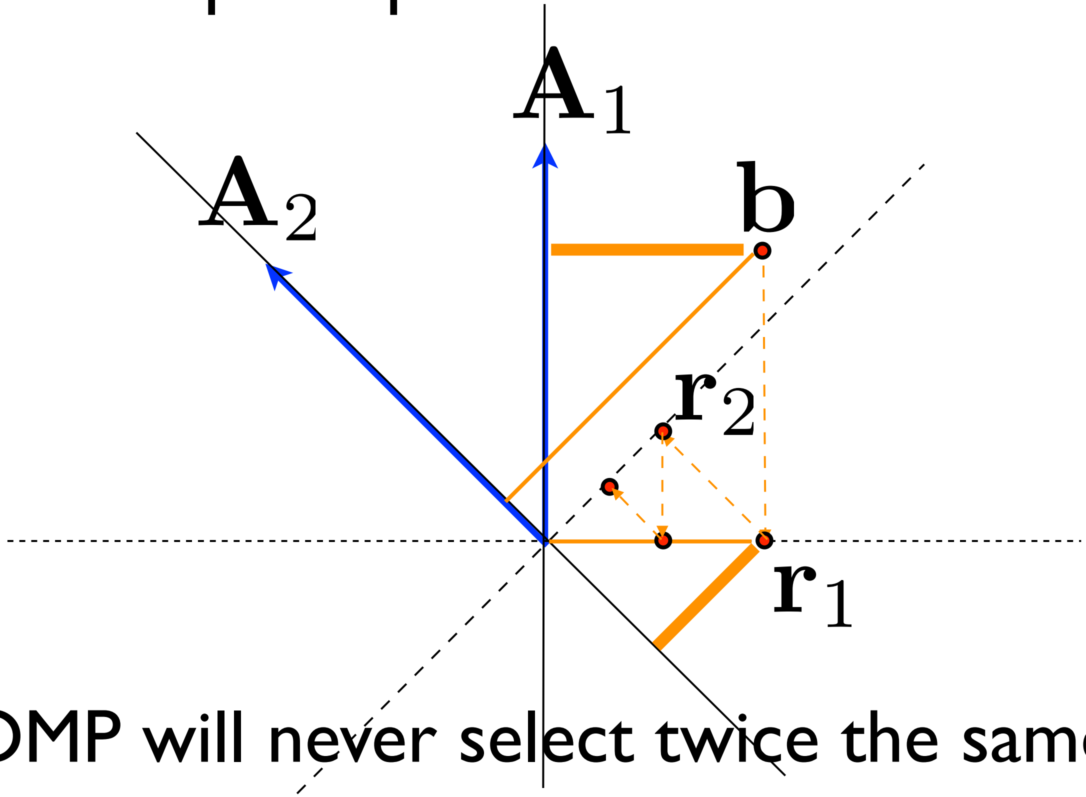
$$V_i = \mathrm{span}(\mathbf{A}_n, n \in \Lambda_i)$$

$$\mathbf{r}_i = \mathbf{b} - P_{V_i}\mathbf{b}$$

- Property : strong convergence $\lim_{i \to \infty} \|\mathbf{r}_i\|_2 = 0$

# Caveats (1)

- MP can pick up the same atom more than once




- OMP will never select twice the same atom

# Caveats (2)

- "Improved" atom selection does not necessarily improve convergence
- There exists two dictionaries **A** and **B**
  - Best atom from **B** at step i:

$$n_i = \arg\max_n |\mathbf{B}_n^T \mathbf{r}_{i-1}|$$

  - Better atom from **A**

$$|\mathbf{A}_{\ell_i}^T \mathbf{r}_{i-1}| \geq |\mathbf{B}_n^T \mathbf{r}_{i-1}|$$

  - Residual update

$$\mathbf{r}_i = \mathbf{r}_{i-1} - (\mathbf{A}_{\ell_i}^T \mathbf{r}_{i-1}) \mathbf{A}_{\ell_i}$$

- Divergence! $\exists c > 0, \forall i, \|\mathbf{r}_i\|_2 \geq c$

# Stagewise greedy algorithms

- Principle = select *multiple* atoms at a time to accelerate the process
- Example of such algorithms
  - ✦ Morphological Component Analysis *[MCA, Bobin et al]*
  - ✦ Stagewise OMP *[Donoho & al]*
  - ✦ CoSAMP *[Needell & Tropp]*
  - ✦ ROMP *[Needell & Vershynin]*
  - ✦ Iterative Hard Thresholding *[Blumensath & Davies 2008]*

# Main greedy algorithms

$$\mathbf{b} = \mathbf{A}x_i + \mathbf{r}_i \qquad \mathbf{A} = [\mathbf{A}_1, \ldots \mathbf{A}_N]$$

| | Matching Pursuit | OMP | Stagewise |
|---|---|---|---|
| Selection | $\Gamma_i := \arg\max_n \lvert \mathbf{A}_n^T \mathbf{r}_{i-1} \rvert$ (Matching Pursuit and OMP) | | $\Gamma_i := \{n \mid \lvert \mathbf{A}_n^T \mathbf{r}_{i-1} \rvert > \theta_i\}$ |
| Update | $\Lambda_i = \Lambda_{i-1} \cup \Gamma_i$<br>$x_i = x_{i-1} + \mathbf{A}_{\Gamma_i}^+ \mathbf{r}_{i-1}$<br>$\mathbf{r}_i = \mathbf{r}_{i-1} - \mathbf{A}_{\Gamma_i} \mathbf{A}_{\Gamma_i}^+ \mathbf{r}_{i-1}$ | $\Lambda_i = \Lambda_{i-1} \cup \Gamma_i$<br>$x_i = \mathbf{A}_{\Lambda_i}^+ \mathbf{b}$<br>$\mathbf{r}_i = \mathbf{b} - \mathbf{A}_{\Lambda_i} x_i$ (OMP and Stagewise) | |

MP & OMP: *Mallat & Zhang 1993*
StOMP: *Donoho & al 2006* (similar to MCA, *Bobin & al 2006*)

# Summary

| | Global optimization | Iterative greedy algorithms |
|---|---|---|
| Principle | $\min_x \frac{1}{2}\|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda\|x\|_p^p$ | iterative decomposition $\mathbf{r}_i = \mathbf{b} - \mathbf{A}x_i$<br>• select new components<br>• update residual |
| Tuning quality/sparsity | regularization parameter $\lambda$ | stopping criterion<br>(nb of iterations, error level, ...)<br>$\|x_i\|_0 \geq k$ $\|\mathbf{r}_i\| \leq \epsilon$ |
| Variants | • choice of sparsity measure p<br>• optimization algorithm<br>• initialization | •selection criterion (weak, stagewise ...)<br>•update strategy (orthogonal ...) |

# Overview

- Convex optimization algorithms
- Greedy algorithms
- **Comparison of complexities**
- Exact recovery conditions for Lp minimization

# Complexity of IST

- Notation: $O(\mathbf{A})$ cost of applying $\mathbf{A}$ or $\mathbf{A}^T$
- Iterative Thresholding $f(x) = \Theta^p_{\alpha\lambda}(x + \alpha \mathbf{A}^T(\mathbf{b} - \mathbf{A}x))$
  - cost per iteration = $O(\mathbf{A})$
  - when **A** invertible, linear convergence at rate

$$\|x^{(i)} - x^\star\|_2 \lesssim C\beta^i \|x^\star\|_2 \qquad \beta \leq 1 - \frac{\sigma^2_{\min}}{\sigma^2_{\max}}$$

  - number of iterations guaranteed to approach limit within relative precision $\epsilon$

$$O(\log 1/\epsilon)$$

- Limit depends on choice of penalty factor $\lambda$, added complexity to adjust it

# Complexity of MP

- Number of iterations depends on stopping criterion

$$\|\mathbf{r}_i\|_2 \leq \epsilon, \|x_i\|_0 \geq k$$

- Cost of first iteration = atom selection (computation of all inner products) $O(\mathbf{A})$
- Naive cost of subsequent iterations = $O(\mathbf{A})$
- If "local" structure of dictionary *[Krstulovic & al, MPTK]*
  - ✦ subsequent iterations only cost $O(\log N)$

| | Generic $\mathbf{A}$ | Local $\mathbf{A}$ |
|---|---|---|
| k iterations | $O(k\mathbf{A}) \geq O(km)$ | $O(\mathbf{A} + k \log N)$ |
| $k \propto m$ | $O(m^2)$ | $O(m \log N)$ |

# Complexity of OMP

- Number of iterations depends on stopping criterion

$$\|\mathbf{r}_i\|_2 \leq \epsilon, \|x_i\|_0 \geq k$$

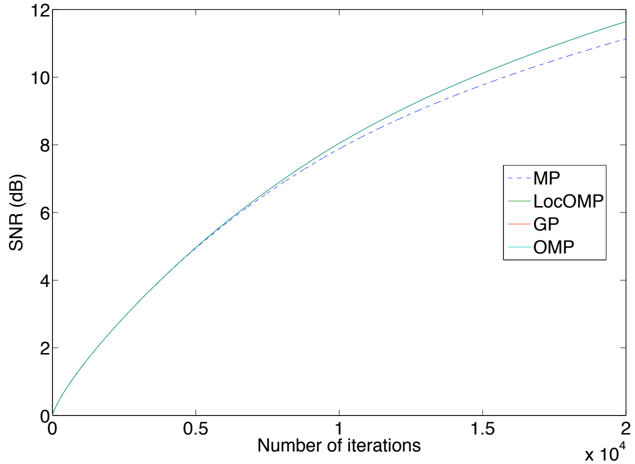
- Naive cost of iteration *i*
  - ✦ atom selection $O(\mathbf{A})$ + orthoprojection $O(i^3)$
- With iterative matrix inversion lemma
  - ✦ atom selection $O(\mathbf{A})$ + coefficient update $O(i)$
- If "local" structure of dictionary *[Mailhé & al, LocOMP]*
  - ✦ subsequent approximate iterations only cost $O(\log N)$

| | Generic **A** | Local **A** |
|---|---|---|
| k iterations | $O(k\mathbf{A} + k^2)$ | $O(\mathbf{A} + k \log N)$ |
| $k \propto m$ | $O(m^3)$ | $O(m \log N)$ |

# LoCOMP

- A variant of OMP for shift invariant dictionaries
  (Ph.D. thesis of Boris Mailhé, ICASSP09)

**Fig. 1**. SNR depending on the number of iterations




$N$ = 5.10$^5$ samples, $k$= 20 000 iterations

**Table 3**. CPU time per iteration (s)

| Iteration | MP | LocOMP | GP | OMP |
|---|---|---|---|---|
| First ($i = 0$) | 3.4 | 3.4 | 3.4 | 3.5 |
| Begin ($i \approx 1$) | 0.028 | 0.033 | 3.4 | 3.4 |
| End ($i \approx I$) | 0.028 | 0.050 | 40.5 | 41 |
| Total time | 571 | 854 | $4.50 \cdot 10^5$ | $4.52 \cdot 10^5$ |

- Implementation in MPTK in progress for larger scale experiments

# Software ?

- Matlab (simple to adapt, medium scale problems):
  - **Thousands** of unknowns, few seconds of computations
  - L1 minimization with an available toolbox
    ➡ http://www.l1-magic.org/ (Candès et al.), CVX, ...
  - Iterative thresholding
    ➡ http://www.morphologicaldiversity.org/ (Starck et al.), FISTA, NESTA, ...
  - Matching Pursuits
    ➡ sparsify (Blumensath), GPSR, ...
- SMALLbox (to be released soon): unified API for several Matlab toolboxes

Sparse Models, Algorithms and Learning for Large-scale data SMALL

- MPTK : C++, large scale problems
  - **Millions** of unknowns, few minutes of computation
  - specialized for local + shift-invariant dictionaries
  - built-in multichannel
    ➡ http://mptk.irisa.fr

# Overview

- Convex optimization algorithms
- Greedy algorithms
- Comparison of complexities
- **Exact recovery conditions for Lp minimization**

# Usual sparsity measures

- L0-norm $\|x\|_0 := \sum_k |x_k|^0 = \sharp\{k, x_k \neq 0\}$

$$\sharp\{k, x_k \neq 0\} = \text{support}(x)$$

- L*p*-norms $\|x\|_p^p := \sum_k |x_k|^p, 0 \leq p \leq 1$

- Constrained minimization

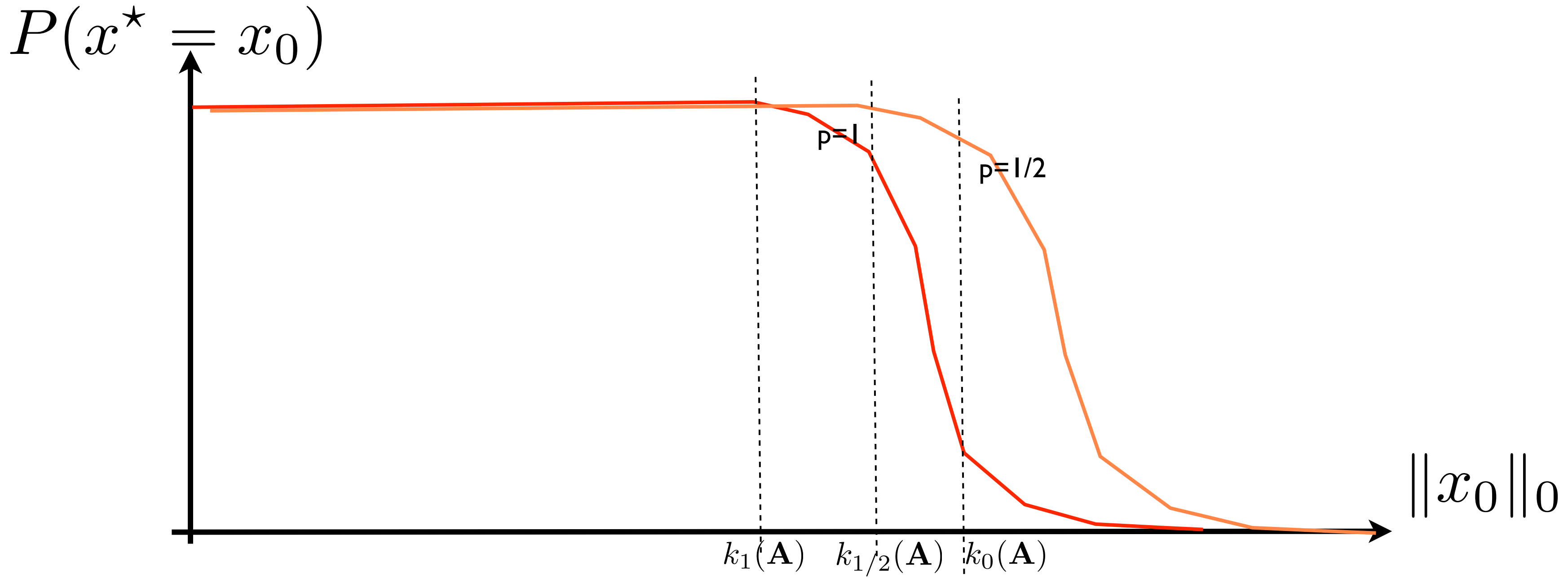
$x_p^\star \in \arg\min_x \|x\|_p$ subject to $\mathbf{b} = \mathbf{A}x$

# Empirical observation : L$p$ versus L1

$$x_0 \xrightarrow{\hspace{2cm}} \mathbf{b} := \mathbf{A}x_0 \xrightarrow{\hspace{2cm}} x_p^\star = \arg \min_{\mathbf{A}x = \mathbf{A}x_0} \|x\|_p$$

reference direct model inverse problem

Typical observation *(e.g. Chartrand 2007)* +extrapolation



# Proved Equivalence between L0 and L1

- "Empty" theorem : assume that $\mathbf{b} = \mathbf{A}x_0$
  - ✦ if $\|x_0\|_0 \leq k_0(\mathbf{A})$ then $x_0 = x_0^\star$
  - ✦ if $\|x_0\|_0 \leq k_1(\mathbf{A})$ $x_0 = x_1^\star$
- Content = estimation of $k_0(\mathbf{A})$ and $k_1(\mathbf{A})$
  - ✦ *Donoho & Huo 2001 : pair of bases, coherence*
  - ✦ *Donoho & Elad 2003, Gribonval & Nielsen 2003 : dictionary, coherence*
  - ✦ *Candes, Romberg, Tao 2004 : random dictionaries, restricted isometry constants*
  - ✦ *Tropp 2004 : idem for Orthonormal Matching Pursuit, cumulative coherence*
- What about $x_p^\star, 0 \leq p \leq 1$ ?

# Exact recovery: Lp minimization

# Null space

- Null space = kernel

$$z \in \mathcal{N}(\mathbf{A}) \Leftrightarrow \mathbf{A}z = 0$$

- Particular solution *vs* general solution
  - ✦ particular solution

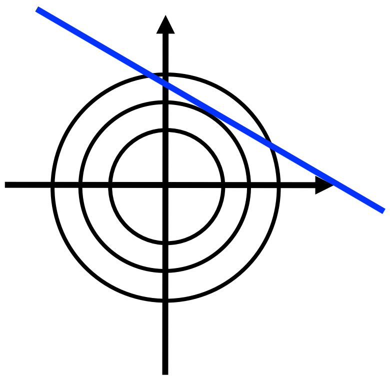
$$\mathbf{A}x = \mathbf{b}$$

  - ✦ general solution

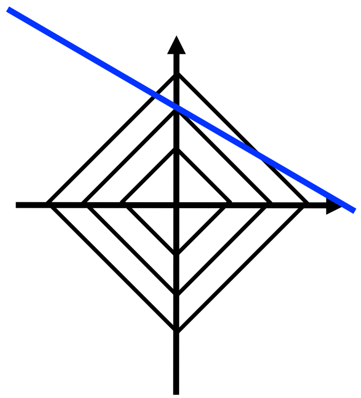
$$\mathbf{A}x' = \mathbf{b} \Leftrightarrow x' - x \in \mathcal{N}(\mathbf{A})$$

# L$p$ “norms” level sets

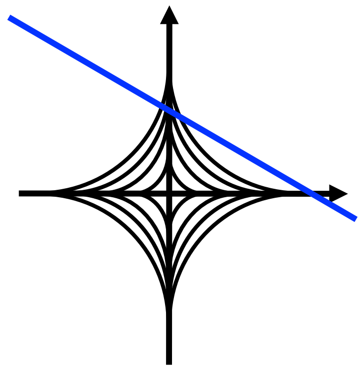
- Strictly convex when $p$>1



- Convex $p$=1



- Nonconvex $p$<1



**Observation**: *the minimizer is sparse*

$\{x \text{ s.t.} \mathbf{b} = \mathbf{A}x\}$

# Exact recovery: necessary condition

- Notations
  - index set *I*
  - vector *z*
  - restriction $$z_I = (z_i)_{i \in I}$$
- Assume there exists $z \in \mathcal{N}(\mathbf{A})$ with
$$\|z_I\|_f > \|z_{I^c}\|_f$$
- Define $$\mathbf{b} := A z_I = A(-z_{I^c})$$
- The vector $z_I$ is supported in *I* but is *not* the minimum norm representation of $\mathbf{b}$

# Exact recovery: sufficient condition

- Assume quasi-triangle inequality

$$\forall x, y \|x+y\|_f \leq \|x\|_f + \|y\|_f$$

- Consider *x* with support set *I* and *x'* with $\mathbf{A}x' = \mathbf{A}x$
- Denote $z := x' - x \in \mathcal{N}(\mathbf{A})$ and observe

$$\begin{aligned} \|x'\|_f = \|x+z\|_f &= \|(x+z)_I\|_f + \|(x+z)_{I^c}\|_f \\ &= \|x+z_I\|_f + \|z_{I^c}\|_f \\ &\geq \|x\|_f - \|z_I\|_f + \|z_{I^c}\|_f \end{aligned}$$

- Conclude:

If $\|z_{I^c}\|_f > \|z_I\|_f$ when $z \in \mathcal{N}(\mathbf{A})$ then $I$ is recoverable

# Recoverable supports : the "Null Space Property" (1)

- **Theorem 1** [*Donoho & Huo 2001 for L1, G. & Nielsen 2003 for Lp*]
  - ✦ Assumption 1: sub-additivity (for quasi-triangle inequality)

$$f(a+b) \leq f(a) + f(b), \forall a, b$$

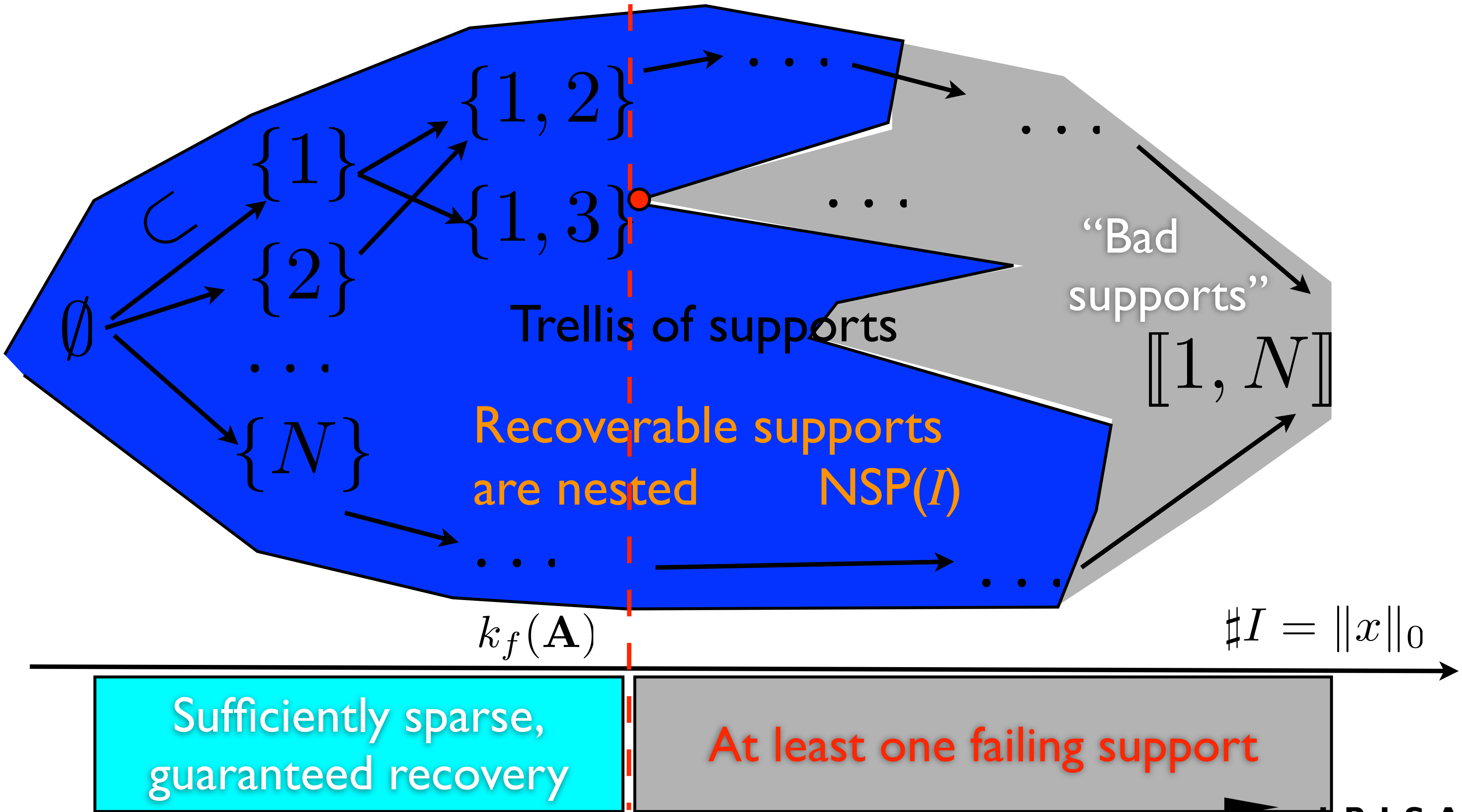
  - ✦ Assumption 2:

NSP

$$\|z_I\|_f < \|z_{I^c}\|_f \text{ when } \quad z \in \mathcal{N}(\mathbf{A}), z \neq 0$$

  - ✦ Conclusion: $x_f^\star$ recovers every $x$ supported in $I$
  - ✦ The result is sharp: if NSP fails on support *I* there is **at least one failing vector** x supported in *I*

# From “recoverable” supports to “sparse” vectors



# Recoverable sparsity levels: the "Null Space Property" (2)

- Corollary 1 [*Donoho & Huo 2001 for L1, G. Nielsen 2003 for Lp*]
  - ✦ Definition :

    $I_k =$ index of $k$ largest components of z
  - ✦ Assumption :

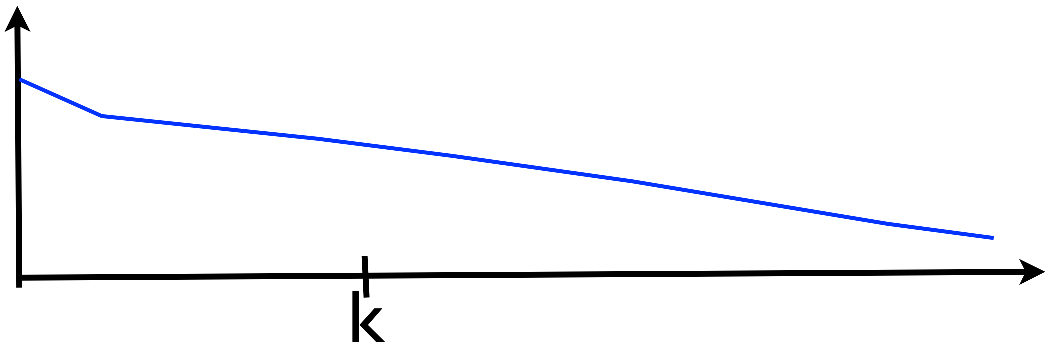
NSP

$$\|z_{I_k}\|_f < \|z_{I_k^c}\|_f \quad \text{when } z \in \mathcal{N}(\mathbf{A}), z \neq 0$$

  - ✦ Conclusion: $x_f^\star$ recovers every $x$ with $\|x\|_0 \leq k$
  - ✦ The result is sharp: if NSP fails there is **at least one failing vector** x with $\|x\|_0 = k$

# Interpretation of NSP

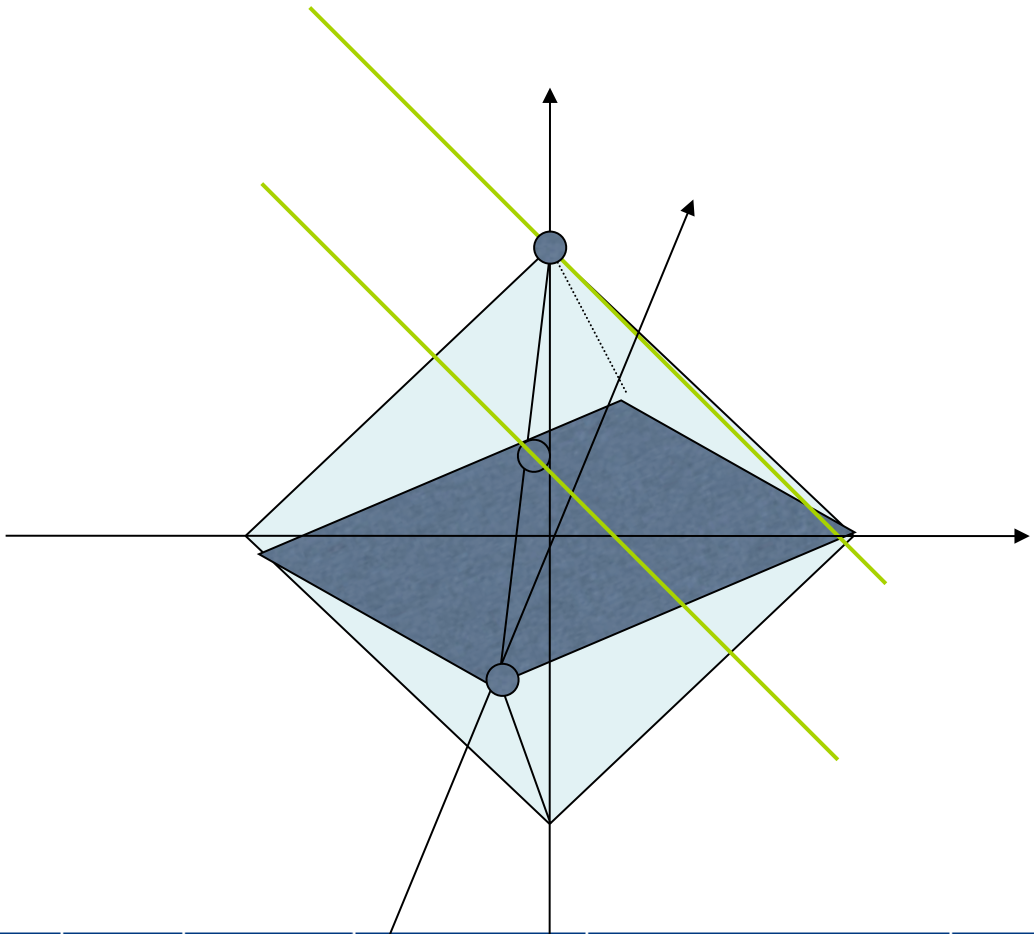
- Geometry in coefficient space:
  - ✦ consider an element z of the Null Space of A
  - ✦ order its entries in decreasing order




  - ✦ the mass of the largest k-terms should not exceed that of the tail $\|z_{I_k}\|_f < \|z_{I_k^c}\|_f$

All elements of the null space must be rather "flat"

# Geometric picture



# Summary

- Review of main algorithms & complexities
- Success guarantees for L1 minimization to solve under-determined inverse linear problems
- Next time:
  - success guarantees for greedy algorithms
  - robust guarantees
  - practical conditions to check guarantees