Performance of Sparse Decomposition Algorithms with Deterministic versus Random Dictionaries

Rémi Gribonval
EPI METISS
INRIA Rennes - Bretagne Atlantique

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

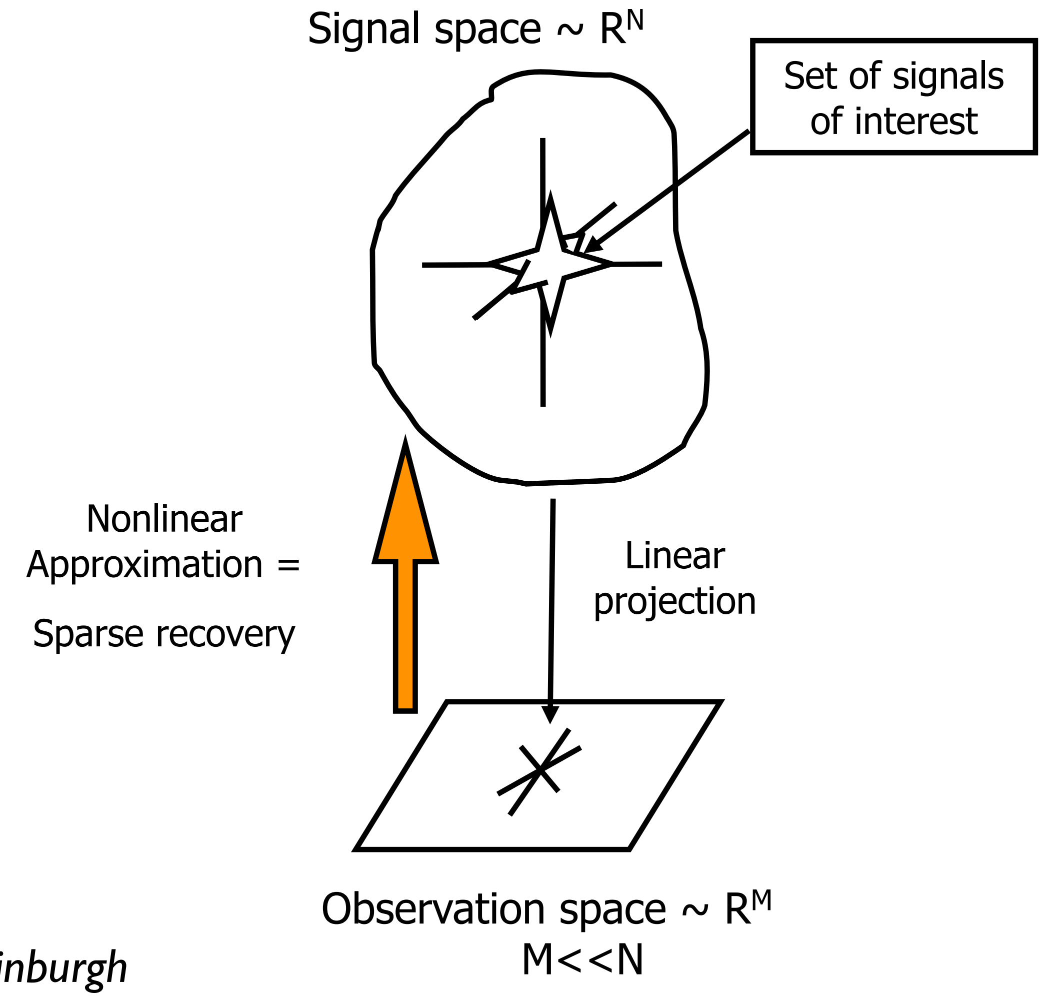
Slides

- http://www.irisa.fr/metiss/members/remi/talks/

Summary

- Session 1:
  - role of sparsity for compression and inverse problems
- Session 2:
  - Review of main algorithms & complexities
  - Success guarantees for L1 minimization to solve under-determined inverse linear problems
- Session 3:
  - Comparison of guarantees for different algorithms
  - Robust guarantees & Restricted Isometry Property
  - Explicit guarantees for various inverse problems

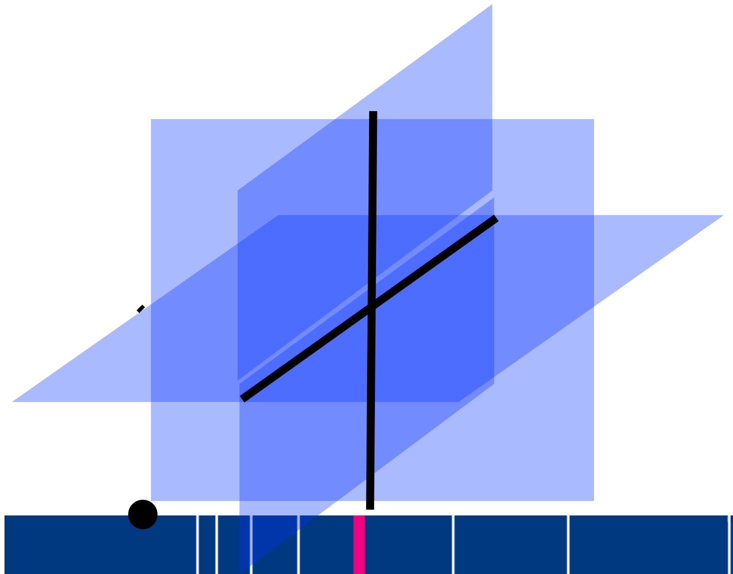
Inverse problems

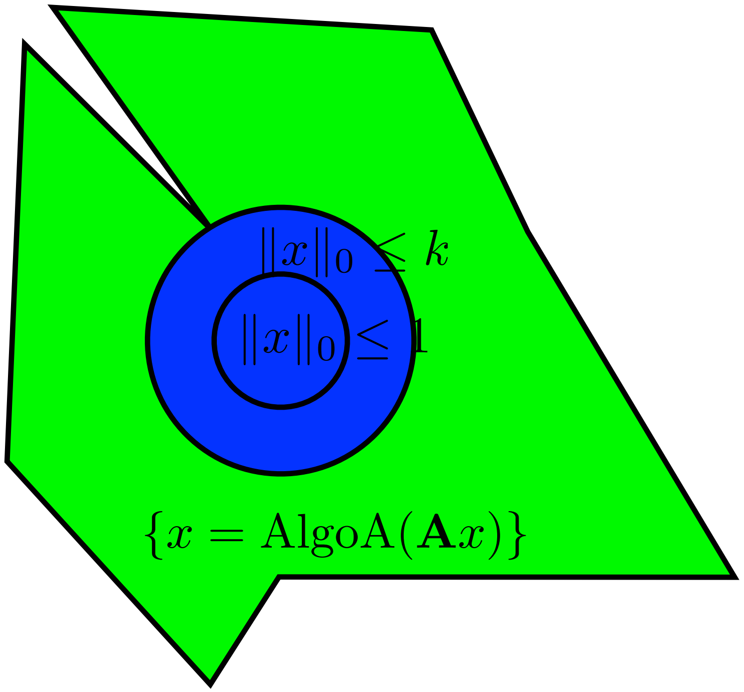


Courtesy: M. Davies, U. Edinburgh

Recovery analysis for inverse problem $\mathbf{b} = \mathbf{A}x$

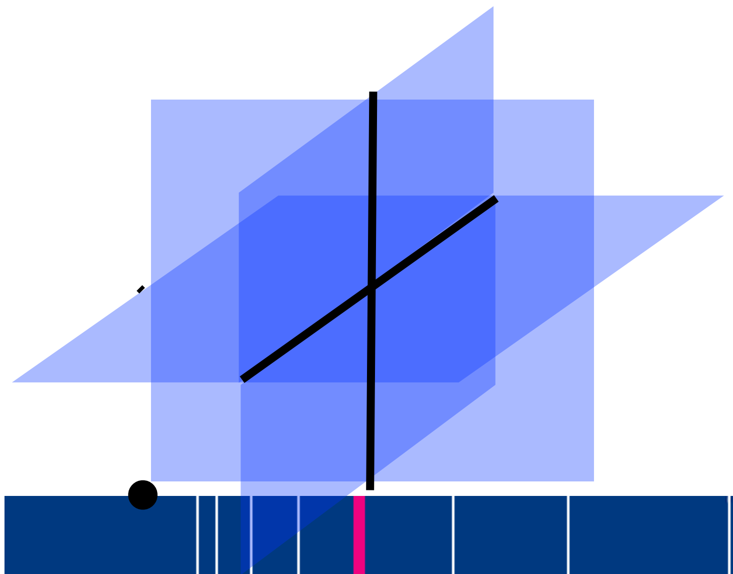
- Recoverable set for a given “inversion” algorithm
- Level sets of L0-norm
  - 1-sparse
  - 2-sparse
  - 3-sparse ...

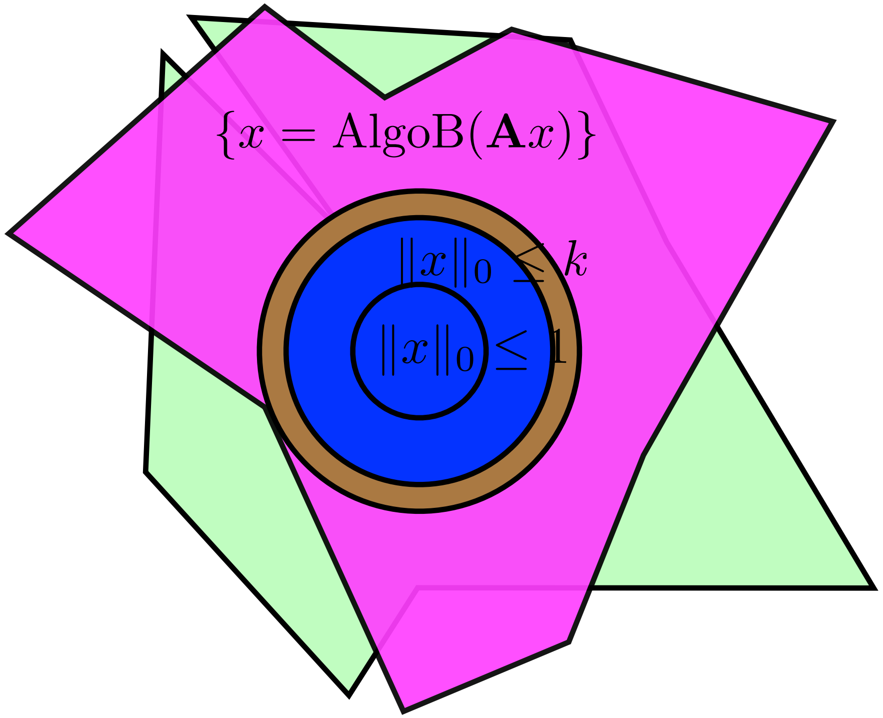




Recovery analysis for inverse problem $\mathbf{b} = \mathbf{A}x$

- Recoverable set for a given "inversion" algorithm
- Level sets of L0-norm
  - 1-sparse
  - 2-sparse
  - 3-sparse ...



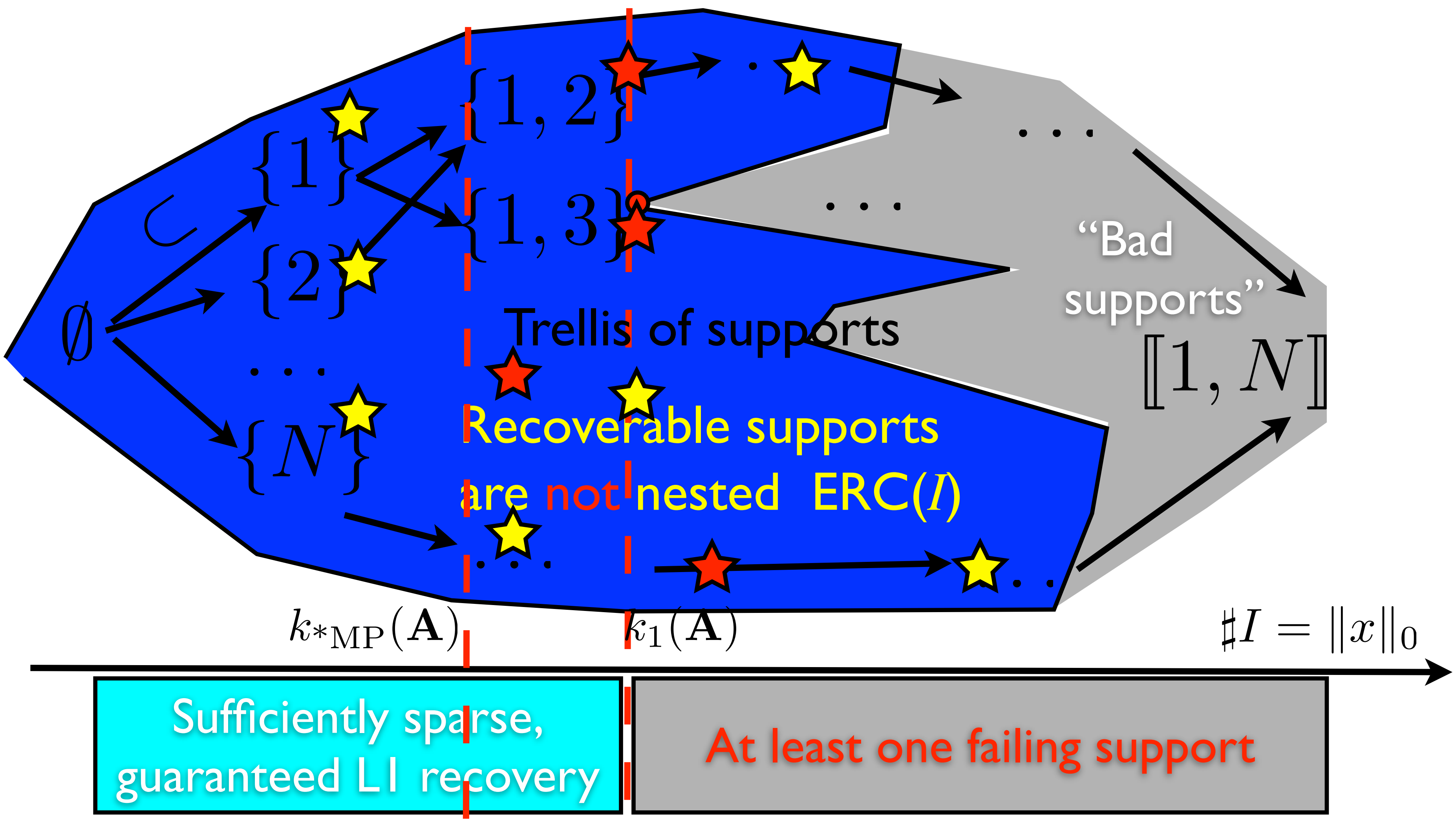


Some “simple” recovery conditions

Support	**Sparsity level**
“recoverable supports” = subsets $I \subset [\![1, N]\!]$ such that	“recoverable sparsity” = integers *k* such that
$\mathrm{supp}(x) := \{k, x_k \neq 0\} \subset I$	$\|x_0\|_0 \leq k$

$x \in$ recoverable

Greedy vs L1: summary



Comparison between algorithms

- Recovery conditions based on number of nonzero components $\|x\|_0$ for $0 \leq p \leq q \leq 1$

$$k_{*\mathrm{MP}}(\mathbf{A}) \leq k_1(\mathbf{A}) \leq k_p(\mathbf{A}) \leq k_q(\mathbf{A}) \leq k_0(\mathbf{A}), \forall \mathbf{A}$$

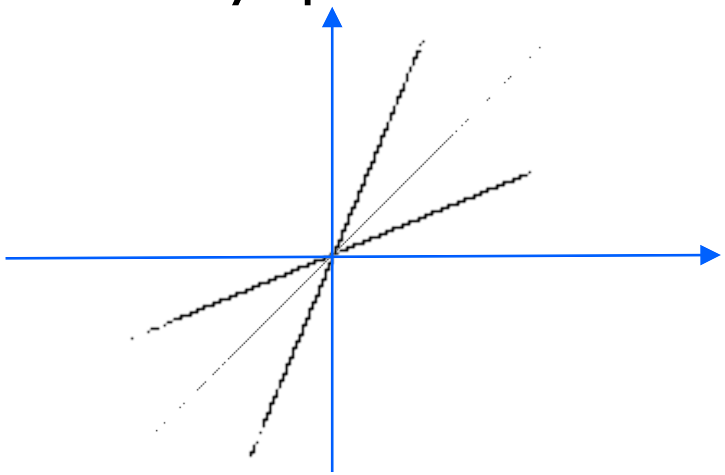
- **Warning** :
  - there often exists vectors beyond these critical sparsity levels, which are recovered
  - there often exists vectors beyond these critical sparsity levels, where the successful algorithm is not the one we would expect

[Gribonval & Nielsen, ACHA 2007]

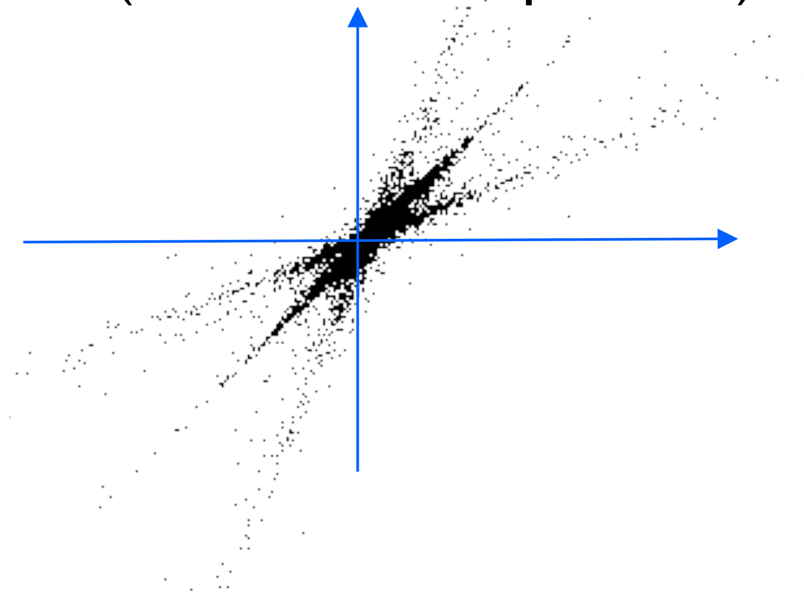
Stability and robustness

Need for stable recovery

Exactly sparse data



Real data (from source separation)



Formalization of stability

- Toy problem: exact recovery from $\mathbf{b} = \mathbf{A}x$
  - Assume sufficient sparsity $\|x\|_0 \leq k_p(\mathbf{A}) < m$
  - Wish to obtain $x_p^\star(\mathbf{b}) = x$

- Need to relax sparsity assumption
  - New benchmark = best k-term approximation

$$\sigma_k(x) = \inf_{\|y\|_0 \leq k} \|x - y\|$$

  - Goal = stable recovery = *instance optimality*

$$\|x_p^\star(\mathbf{b}) - x\| \leq C \cdot \sigma_k(x)$$

[Cohen, Dahmen & De Vore 2006]

Stability for Lp minimization

- Assumption: «stable Null Space Property»

NSP(k, ℓ^p,t)

$$\|z_{I_k}\|_p^p \leq t \cdot \|z_{I_k^c}\|_p^p \qquad \text{when} \quad z \in \mathcal{N}(\mathbf{A}), z \neq 0$$

- Conclusion: *instance optimality* for all *x*

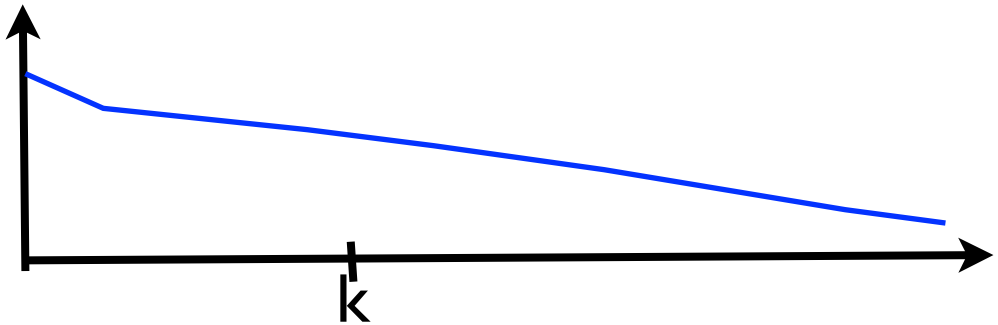
$$\|x_p^\star(\mathbf{b}) - x\|_p^p \leq C(t) \cdot \sigma_k(x)_p^p$$

$$C(t) := 2\frac{1+t}{1-t}$$

[Davies & Gribonval, SAMPTA 2009]

Reminder on 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 a fraction of that of the tail $\|z_{I_k}\|_p^p \leq t \cdot \|z_{I_k^c}\|_p^p$

All elements of the null space must be "flat"

Robustness

- Toy model = noiseless
- Need to account for noise
  - measurement noise
  - modeling error
  - numerical inaccuracies ...

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

- Goal: predict robust estimation

$$\|x_p^\star(\mathbf{b}) - x\| \leq C\|e\| + C'\sigma_k(x)$$

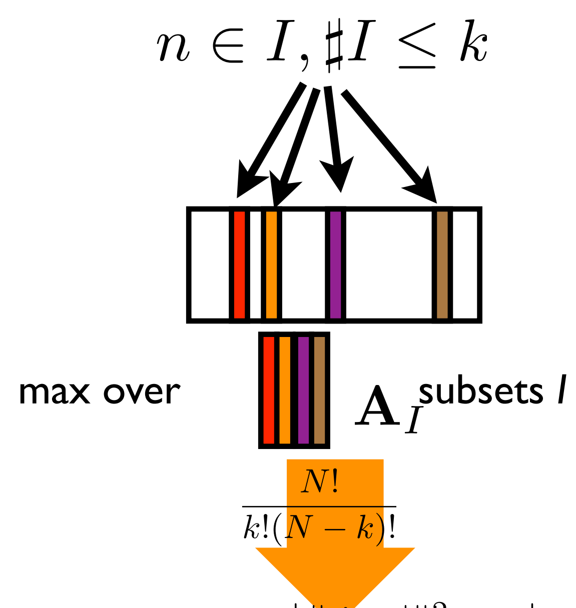
- Tool: restricted isometry property

Restricted Isometry Property

- Definition

$\mathbf{A}$

N columns



$$\delta_k := \sup_{\sharp I \leq k,\ c \in \mathbb{R}^k} \left| \frac{\|\mathbf{A}_I c\|_2^2}{\|c\|_2^2} - 1 \right|$$

- Computation ?
  - naively: combinatorial
  - **open question**: NP ? NP-complete ?

Stability & robustness from RIP

RIP(k, δ)

$$\delta_{2k}(\mathbf{A}) \leq \delta$$

[Candès 2008] $t := \sqrt{2}\delta/(1-\delta)$

NSP(k, ℓ^1,t)

$$\|z_{I_k}\|_1 \leq t \cdot \|z_{I_k^c}\|_1 \quad \text{when} \quad z \in \mathcal{N}(\mathbf{A}), z \neq 0$$

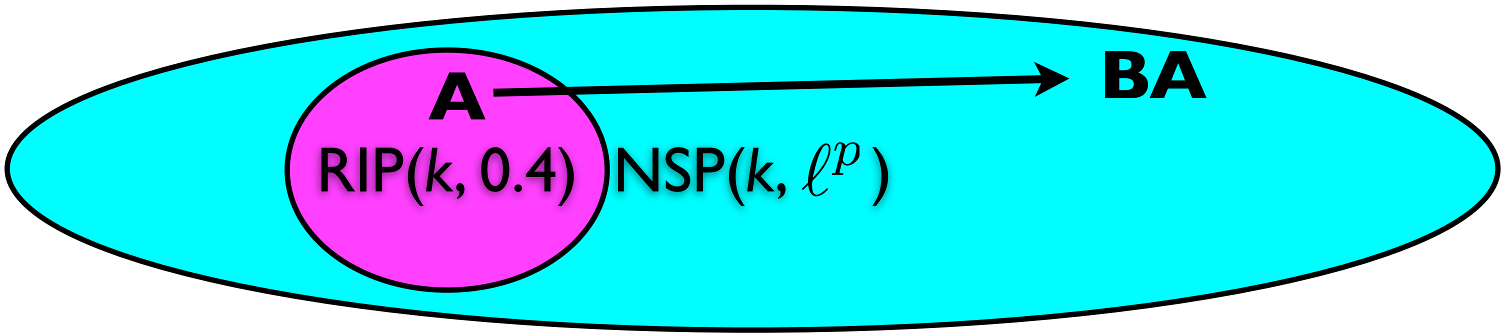
- Result: **stable + robust** L1-recovery under assumption that

$$\delta_{2k}(\mathbf{A}) < \sqrt{2} - 1 \approx 0.414$$

  - *Foucart-Lai 2008*: Lp with p<1, and $\delta_{2k}(\mathbf{A}) < 0.4531$
  - *Chartrand 2007, Saab & Yilmaz 2008*: other RIP condition for p<1
  - *G., Figueras & Vandergheynst 2006*: robustness with *f*-norms
  - *Needel & Tropp 2009, Blumensath & Davies 2009*: RIP for greedy algorithms

Is the RIP a sharp condition ?

- The Null Space Property
  - ✦ “algebraic” + sharp property for Lp, only depends on $\mathcal{N}(\mathbf{A})$
  - ✦ invariant by linear transforms $\mathbf{A} \rightarrow \mathbf{BA}$
- The RIP(k, δ) condition
  - ✦ “metric” ... and not invariant by linear transforms
  - ✦ predicts performance + robustness of several algorithms



[Davies & Gribonval, IEEE Inf.Th. 2009]

Remaining agenda

- Recovery conditions based on number of nonzero components $\|x\|_0$ $\quad 0 \leq p \leq q \leq 1$

$$k_{*\mathrm{MP}}(\mathbf{A}) \leq k_1(\mathbf{A}) \leq k_p(\mathbf{A}) \leq k_q(\mathbf{A}) \leq k_0(\mathbf{A}), \forall \mathbf{A}$$

- **Question**
  - ✦ what is the order of magnitude of these numbers ?
  - ✦ how do we estimate them in practice ?
- A first element:
  - ✦ if **A** is *m* x *N*, then $k_0(\mathbf{A}) \leq \lfloor m/2 \rfloor$
  - ✦ for almost all matrices (in the sense of Lebesgue measure in $\mathbb{R}^{mN}$) this is an equality

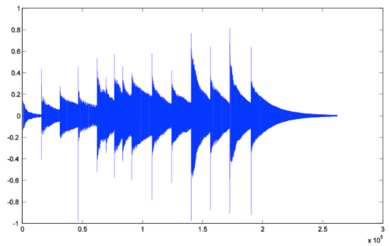
Explicit guarantees in various inverse problems

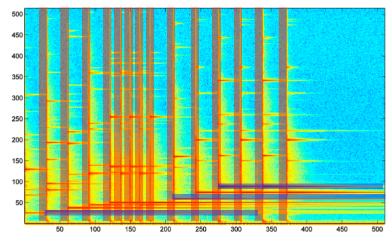
Scenarios

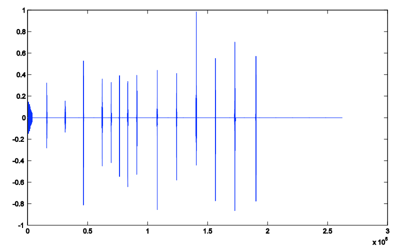
- Range of "choices" for the matrix **A**
  - ✦ Dictionary modeling structures of signals
    union of wavelets + curvelets + spikes
  - ✦ «Transfer function» from physics of inverse problem
    convolution operator / transmission channel
  - ✦ Designed Compressed Sensing matrix
    random Gaussian matrix
- Estimation of the recovery regimes
  - ✦ coherence for deterministic matrices
  - ✦ typical results for random matrices

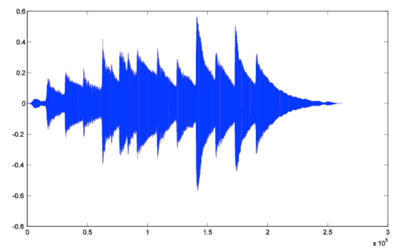
Multiscale Time-Frequency Structures

- Audio = superimposition of structures
- Example : glockenspiel









  - ✦ transients = short, small scale
  - ✦ harmonic part = long, large scale
- Gabor atoms $\left\{ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w \left(\frac{t-\tau}{s} \right) \, e^{2i\pi f t} \right\}_{s,\tau,f}$

Deterministic matrices and coherence

- **Lemma**
  - Assume normalized columns $\|\mathbf{A}_i\|_2 = 1$
  - Define **coherence** $\mu = \max_{i \neq j} |\mathbf{A}_i^T \mathbf{A}_j|$
  - Consider index set I of size $\sharp I \leq k$
  - Then for any coefficient vector $c \in \mathbb{R}^I$

$$1 - (k-1)\mu \leq \frac{\|\mathbf{A}_I c\|_2^2}{\|c\|_2^2} \leq 1 + (k-1)\mu$$

  - In other words

$$\delta_{2k} \leq (2k-1)\mu$$

Consequence

- Since $\delta_{2k} \leq \mu \cdot (2k-1)$ we obtain $\delta_{2k} \leq \delta$ as soon as

$$k < (1 + \delta/\mu)/2$$

- Combining with best known RIP condition for stable L1 recovery $\delta \approx 0.4531$

$$k_1(\mathbf{A}) \geq \lfloor (1 + 0.4531/\mu)/2 \rfloor$$

- In fact, can prove with other techniques that

$$k_0(\mathbf{A}) \geq k_1(\mathbf{A}) \geq \lfloor (1 + 1/\mu)/2 \rfloor$$

[G. Nielsen 2003]

Example: convolution operator

- Deconvolution problem $y = h \star s + e$
  - ✦ Matrix-vector form $\mathbf{b} = \mathbf{A}x + \mathbf{e}$ with **A** = Toeplitz or circulant matrix $[\mathbf{A}_1, \ldots, \mathbf{A}_N]$

$$\mathbf{A}_n(i) = h(i-n)$$

by convention $\|\mathbf{A}_n\|_2^2 = \sum_i h(i)^2 = 1$

  - ✦ Coherence = autocorrelation, can be large

$$\mu = \max_{n \neq n'} \mathbf{A}_n^T \mathbf{A}_{n'} = \max_{\ell \neq 0} h \star \tilde{h}(\ell)$$

  - ✦ Recovery guarantees
    - ✤ Worst case = close spikes, usually difficult and not robust
    - ✤ Stronger results assuming distance between spikes *[Dossal 2005]*

Example: image inpainting

Courtesy of: G. Peyré, Ceremade, Université Paris 9 Dauphine

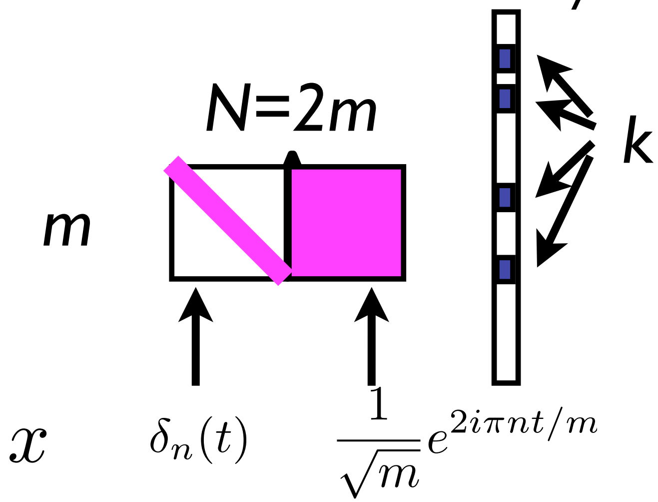
$$\mathbf{b} = \mathbf{M}y = \mathbf{M}\mathbf{\Phi}x$$

Inpainting

$$y = \mathbf{\Phi}x$$

Coherence vs RIP

- **Deterministic** matrix, such as Dirac-Fourier dictionary



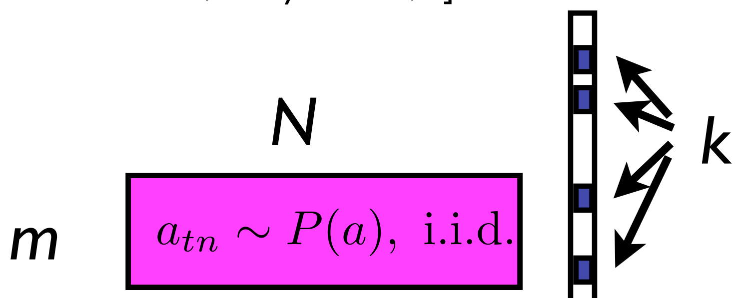
- **Coherence**

$$\mu = 1/\sqrt{m}$$

$\mathbf{A}$

- "**Generic**" (random) dictionary

[Candès & al 2004, Vershynin 2006, ...]



- **Isometry constants**

if $\quad m \geq Ck \log N/k$

then $\quad P(\delta_{2k} < \sqrt{2}-1) \approx 1$

Recovery regimes

$$k_1(\mathbf{A}) \approx 0.914\sqrt{m}$$
$$k_{*\mathrm{MP}}(\mathbf{A}) \geq 0.5\sqrt{m}$$

$$k_1(\mathbf{A}) \approx \frac{m}{2e \log N/m}$$ with high probability

[Elad & Bruckstein 2002] *[Donoho & Tanner 2009]*

Compressed sensing

- Approach = acquire some data *y* with a limited number *m* of (linear) measures, modeled by a measurement matrix $\mathbf{b} \approx \mathbf{K}y$
- Key hypotheses
  - Sparse model: the data can be sparsely represented in some dictionary $y \approx \mathbf{\Phi}x \qquad \sigma_k(x) \ll \|x\|$
  - The overall matrix $\mathbf{A} = \mathbf{K}\mathbf{\Phi}$ leads to robust + stable sparse recovery, e.g. $\delta_{2k}(\mathbf{A}) \ll 1$
- Reconstruction = sparse recovery algorithm

Key constraints to use Compressed Sensing

- Availability of sparse model= dictionary $\mathbf{\Phi}$
  - should fit well the **data**, not always granted. E.g.: cannot aquire white Gaussian noise!
  - require appropriate *choice* of dictionary, or **dictionary learning from training data**
- Measurement matrix $\mathbf{K}$
  - must be associated with **physical sampling process** (hardware implementation)
  - should guarantee **recovery** from $\mathbf{K}\mathbf{\Phi}$ hence incoherence
  - should ideally enable fast algorithms through **fast computation** of $\mathbf{K}y, \mathbf{K}^T\mathbf{b}$

Remarks

- Worthless if high-res. sensing+storage = cheap

i.e., not for your personal digital camera!

- Worth it whenever
  - High-res. = impossible (no miniature sensor, e.g, certain wavelength)
  - Cost of each measure is high
    - Time constraints [fMRI]
    - Economic constraints [well drilling]
    - Intelligence constraints [furtive measures]?
  - Transmission is lossy

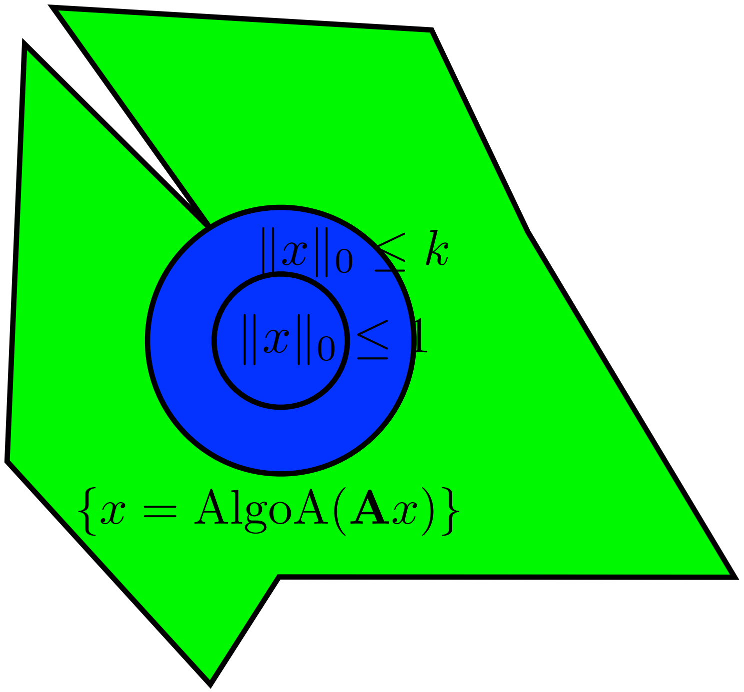
  (robust to loss of a few measures)

Excessive pessimism ?

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

- Recoverable set for a given "inversion" algorithm
- Level sets of L0-norm
- Worst case

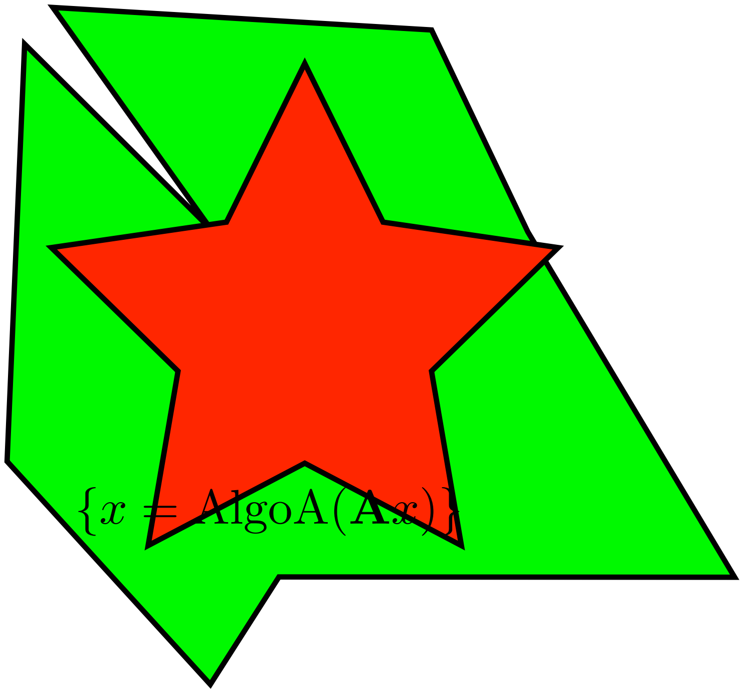
= too pessimistic!



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

- Recoverable set for a given "inversion" algorithm
- Level sets of L0-norm
- Worst case
  = too pessimistic!
- Finer "structures" of x
  $\mathrm{support}(x), \mathrm{sign}(x)$

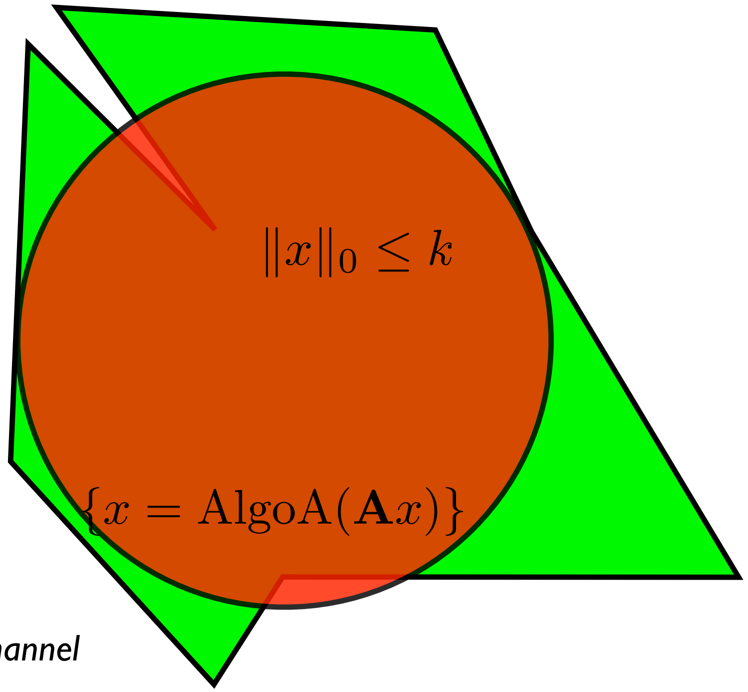
*Borup, G. & Nielsen ACHA 2008, **A** = Wavelets U Gabor, recovery of infinite supports for analog signals*



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

- Recoverable set for a given "inversion" algorithm
- Level sets of L0-norm
- Worst case
  = too pessimistic!
- Finer "structures" of x
  $$\mathrm{support}(x), \mathrm{sign}(x)$$

Borup, G. & Nielsen ACHA 2008, ***A*** *= Wavelets U Gabor, recovery of infinite supports for analog signals*



- Average/typical case

G., Rauhut,, Schnass & Vandergheynst, JFAA 2008, "Atoms of all channels, unite! Average case analysis of multichannel sparse recovery using greedy algorithms".

Average case analysis ?

$P(x_0) \xrightarrow{\text{draw ground truth}} x_0 \xrightarrow{\text{direct model}} \mathbf{b} := \mathbf{A}x_0$

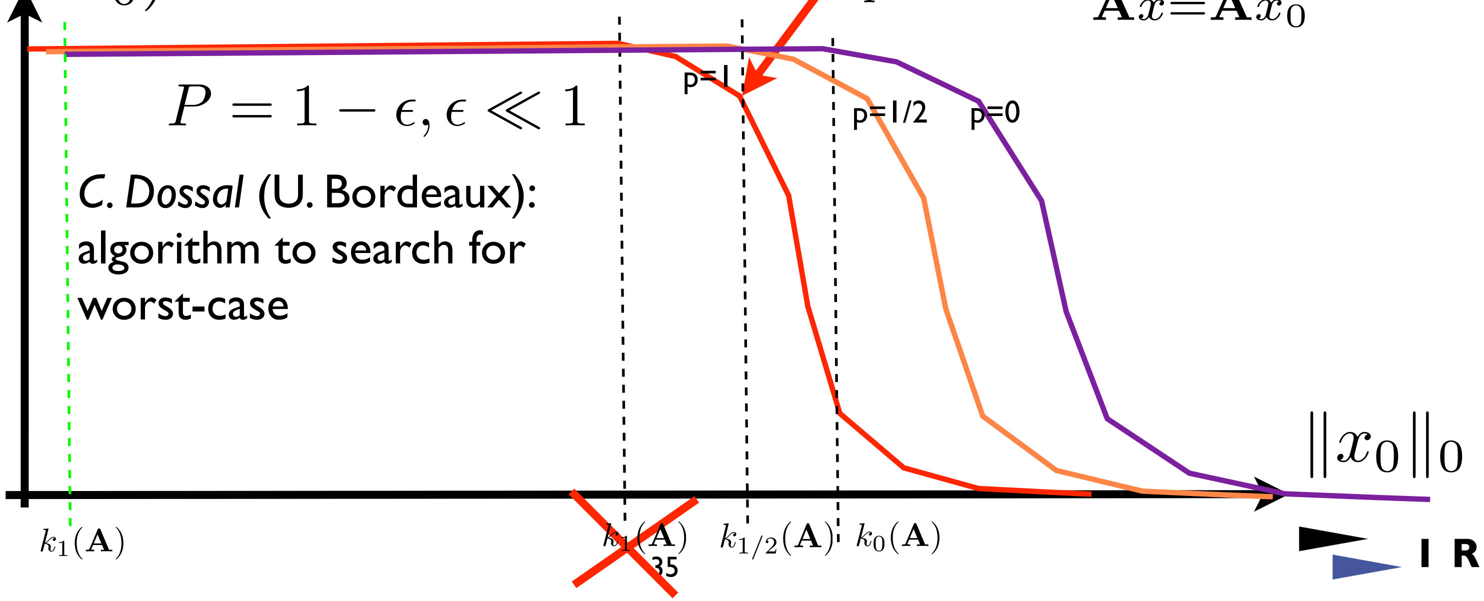
inverse problem

$x_p^\star = \arg \min_{\mathbf{A}x=\mathbf{A}x_0} \|x\|_p$

Bayesian! Favorable priors?

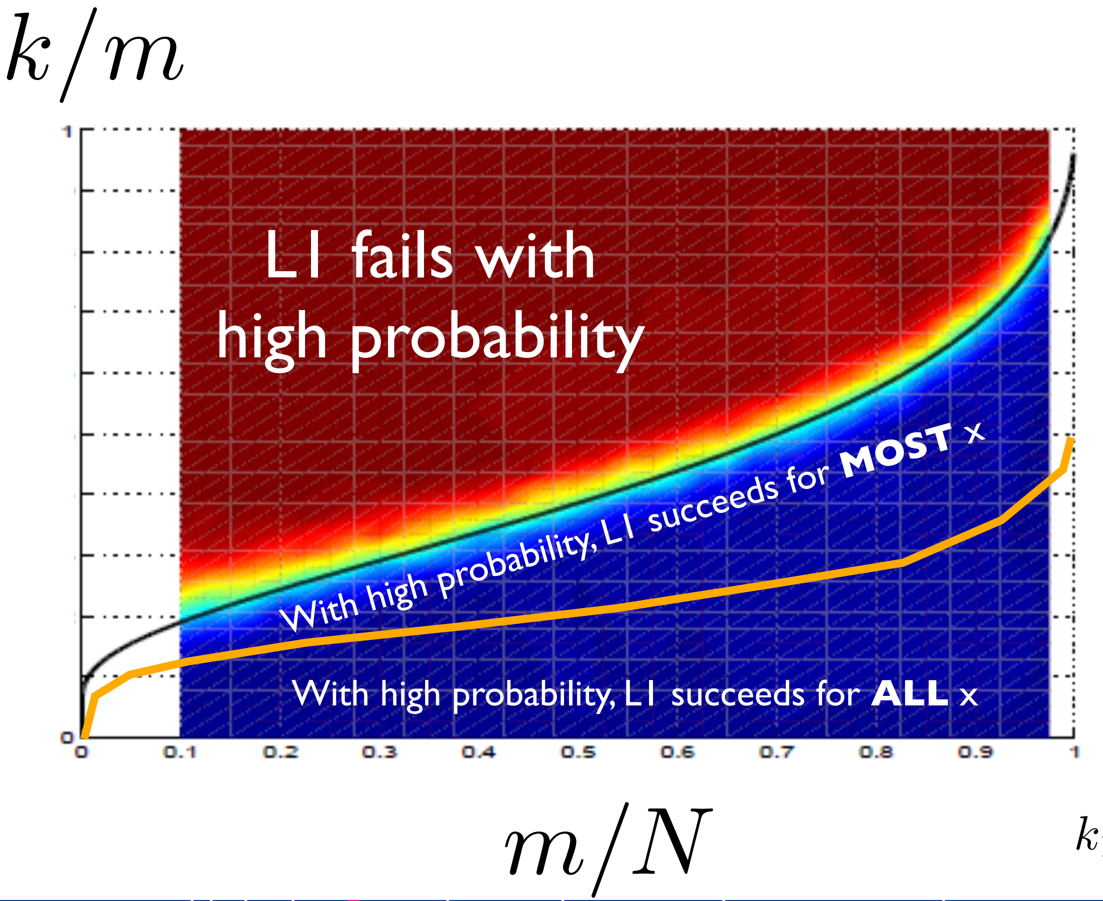
Typical observation

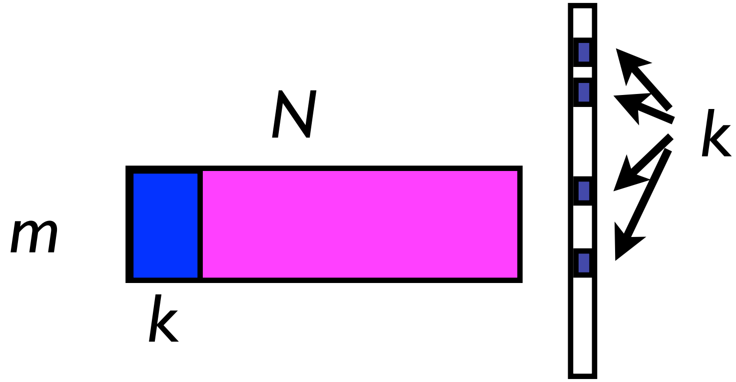
$P(x^\star = x_0)$



IRISA

Phase transitions





$$m \geq Ck \log N/k$$

$$P(\delta_{2k} < \sqrt{2} - 1) \approx 1$$

$$k_1(\mathbf{A}) \approx \frac{m}{2e \log N/m}$$

Conclusions

- Sparsity helps solve ill-posed inverse problems (more unknowns than equations).
- If the solution is sufficiently sparse, any reasonable algorithm will find it (even simple thresholding!).
- Computational efficiency is still a challenge, but problem sizes up to 1000 x 10000 already tractable efficiently.
- Theoretical guarantees are mostly worst-case, empirical recovery goes far beyond but is not fully understood!
- Challenging practical issues include:
  - choosing / learning / designing dictionaries;
  - designing feasible compressed sensing hardware.

Thanks to

- F. Bimbot, G.Gonon, S.Krstulovic, B. Roy
- A. Ozerov, S. Lesage, B. Mailhé
- M. Nielsen, L. Borup (Aalborg Univ.)
- P.Vandergheynst, R. Figueras, P. Jost, K. Schnass (EPFL)
- H. Rauhut (U.Vienna)
- M. Davies (U. Edinburgh)
- and several other collaborators ...

The end

remi.gribonval@inria.fr
www.irisa.fr/metiss/gribonval

The Bayesian bit: L1 minimization and the Laplacian distribution

Bayesian modeling

- Observation : $\mathbf{b} = \mathbf{A}x$
- "True" Bayesian model $P(x_k) \propto \exp(-f(|x_k|))$
- Maximum likelihood estimation

$$\max_x \prod_k P(x_k) \Leftrightarrow \min_x \sum_k f(|x_k|)$$

- L1 minimization equivalent to MAP with Laplacian model

$$\hat{P}(x_k) \propto \exp(-|x_k|)$$

- Does L1 minimization fit Laplacian data ?

L1 minimization for Laplacian data ...

- Gaussian matrix

$$\mathbf{A} \in \mathbb{R}^{m \times N} \quad N = 128 \quad 1 \leq m \leq 100$$

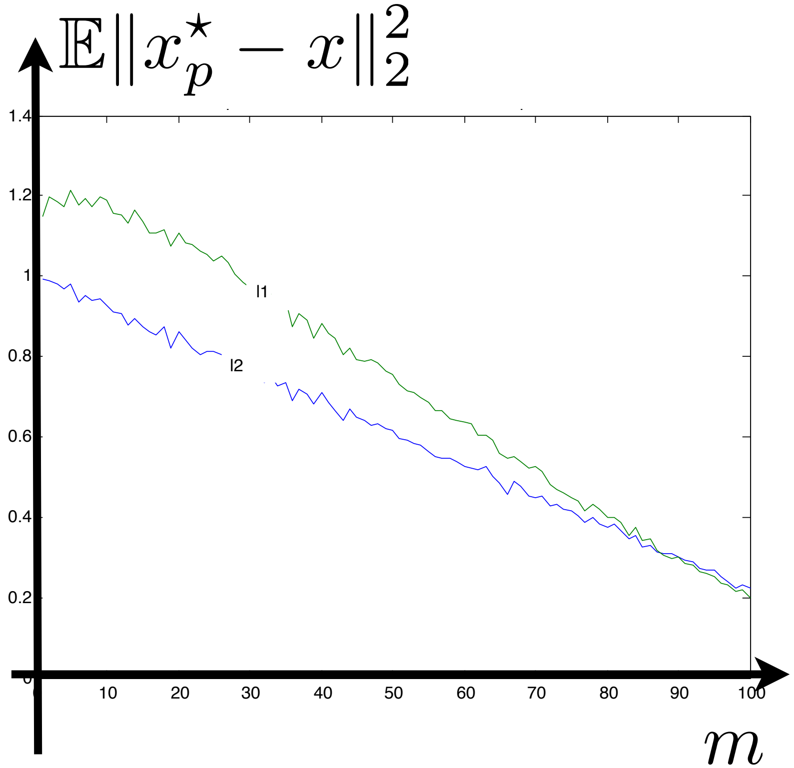
- Laplacian data, 500 draws

$$x \in \mathbb{R}^N \longrightarrow \mathbf{b} = \mathbf{A}x$$

- Reconstruction L1 or L2

$$x_p^\star := \arg\min \|x\|_p, \; p = 1, 2$$

= ML with Laplacian / Gaussian prior

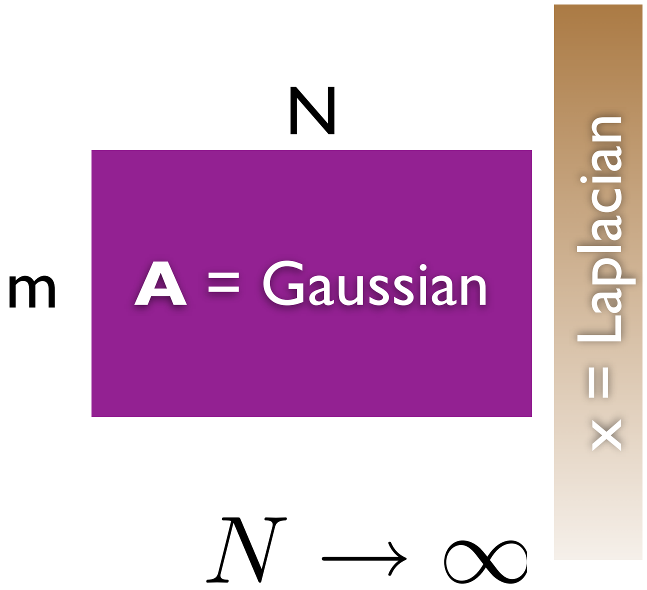


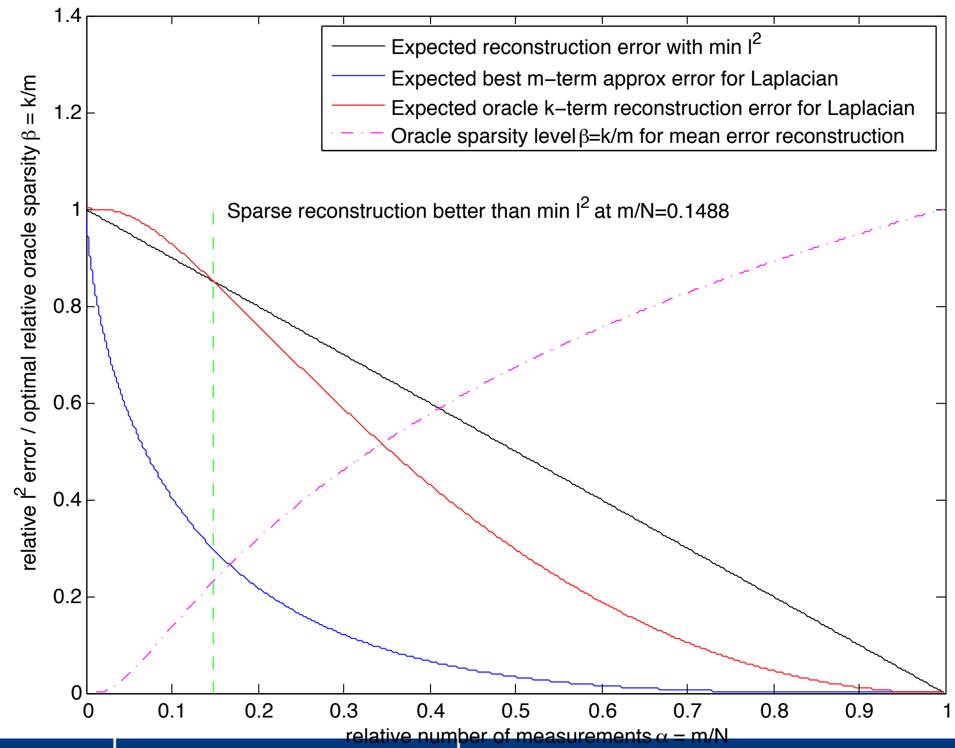
cf also Seeger and Nickish, ICML 2008

MAP is bad when the model fits the data!
Mikolova 2007, Inverse Problems and Imaging

Sparse recovery for Laplacian data ?

- Asymptotic analysis with "oracle" sparse estimation





work in progress, G. & Davies