Inverse problems and sparse models (6/6)

Rémi Gribonval
INRIA Rennes - Bretagne Atlantique, France
remi.gribonval@inria.fr

Overview of the course

- **Introduction**
  - ✓ sparsity & data compression
  - ✓ inverse problems in signal and image processing
    - ✦ image deblurring, image inpainting,
    - ✦ channel equalization, signal separation,
    - ✦ tomography, MRI
  - ✓ sparsity & under-determined inverse problems
    - ✦ relation to subset selection problem
- **Pursuit Algorithms**
  - ✓ Greedy algorithms: Matching Pursuit & al
  - ✓ L1 minimization principles
  - ✓ L1 minimization algorithms
  - ✓ (Complexity of Pursuit Algorithms)

Overview of the course

- **Recovery guarantees for Pursuit Algorithms**
  - ✓ Well-posedness
  - ✓ Coherence *vs* Restricted Isometry Constant
  - ✓ Worked examples
  - ✓ Summary

Exercice at home

- Implement in Matlab / Scilab:
  - ✓ Matching Pursuit (MP), Orthonormal MP (OMP)
  - ✓ Basis Pursuit = L1 minimization [with CVX] (BP)
- Generate test problems
  - ✓ Create matrix A (random Gaussian, normalize columns)
  - ✓ Create k-sparse x and b=Ax
- Compute mp(b,A,k) / omp(b,A,k) / bp(b,A)
- Measure quality (SNR on x) & computation time
- Curves of success as function of sparsity k

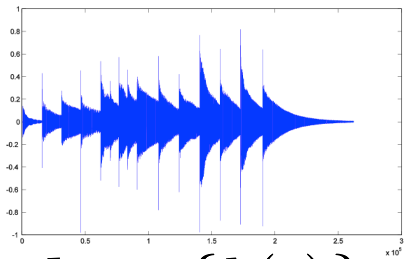
Recovery guarantees in various inverse problems

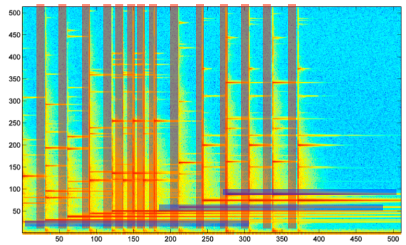
Scenarios

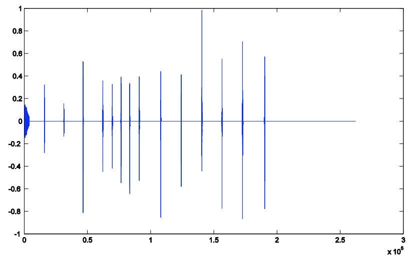
- Range of "choices" for the matrix **A**
  - Ex 1: Dictionary modeling structures of signals
    - **Constrained** choice = to fit the data.
    - *Ex: union of wavelets + curvelets + spikes*
  - Ex2: «Transfer function» from physics of inverse problem
    - **Constrained** choice = to fit the direct problem.
    - *Ex: convolution operator / transmission channel*
  - Ex3-4: Hybrid setting
  - *Ex5: Designed / chosen «Compressed Sensing» matrix*
    - «**Free**» design = to maximize recovery performance *vs* cost of measures
    - *Ex: random Gaussian matrix... or coded aperture, etc.*
- Estimation of the recovery regimes
  - ✓ coherence for deterministic matrices
  - ✓ typical results for random matrices

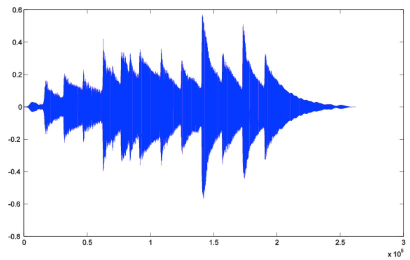
Example 1: Multiscale Time-Frequency Structures

- Audio = superimposition of structures









$$\mathbf{b} = \{b(t)\}_t \qquad x = \{x(s,\tau,f)\}_{s,\tau,f}$$

  - ✓ transients = short, small scale
  - ✓ harmonic part = long, large scale

- Gabor atoms $g_{s,\tau,f}(t) := \frac{1}{\sqrt{s}} w\left(\frac{t-\tau}{s}\right) e^{2i\pi f t}$
- Dictionary matrix: $\mathbf{A}_n = \{g_{s_n,\tau_n,f_n}(t)\}_t$

$$\boxed{\mathbf{A} = [\mathbf{A}_1 \ldots \mathbf{A}_N]}$$

Recovery conditions based on coherence

- **Convention**: normalized columns $\|\mathbf{a}_i\|_2 = 1$
- **Definition**: coherence of dictionary

$$\mu(\mathbf{A}) := \max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$$

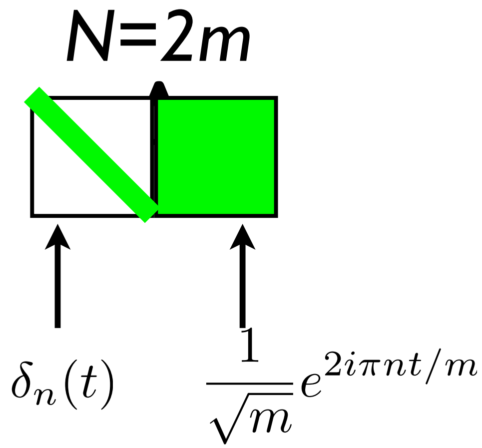
- **Theorem**:
  - ✓ Assume that

$$\boxed{k = \|x\|_0 < \frac{1}{2}(1 + 1/\mu)}$$

  - ✓ Then
    - ✦ x minimizes the L0 and **L1** norm among all solutions x' to the linear inverse problem $\mathbf{A}x' = \mathbf{A}x$
    - ✦ k steps of **OMP** performed on $\mathbf{b} = \mathbf{A}x$ recover x

Caricature of two-scale Gabor dictionary

- **Dirac-Fourier dictionary** $\mathbf{A} = m$



- **Coherence** $\mu = 1/\sqrt{m}$

- **Sparsity thresholds** $k_{*\mathrm{MP}}(\mathbf{A}) \geq 0.5\sqrt{m}$

Example 2: convolution operator

- Deconvolution problem with spikes

$$b = h \star x + e$$

- ✓ Matrix-vector form $\mathbf{b} = \mathbf{A}x + \mathbf{e}$ with **A** = Toeplitz or circulant matrix $[\mathbf{A}_1, \ldots, \mathbf{A}_N]$

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

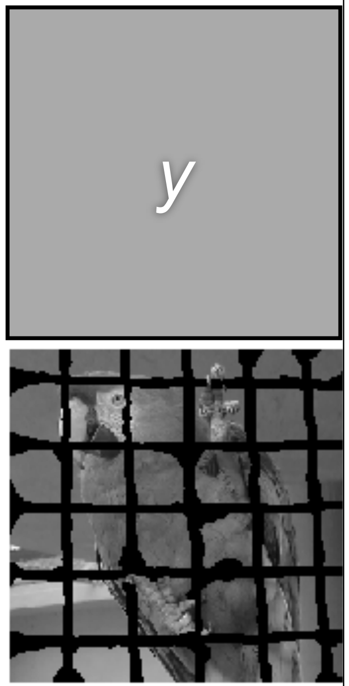
by convention $\|\mathbf{A}_n\|_2^2 = \sum_i h(i)^2 = 1$

- ✓ 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 guarantees assuming distance between spikes *[Dossal*

Example 3: Inpainting Problem



- Unknown image with N pixels

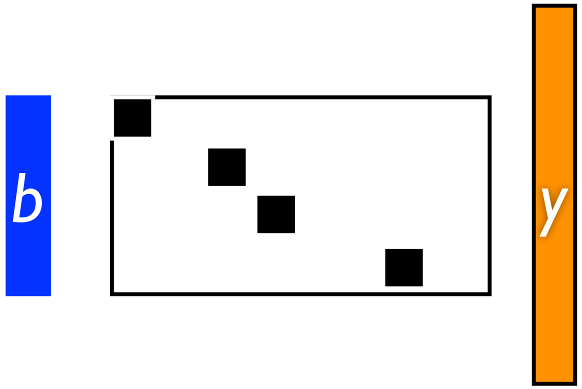
$$\mathbf{y} \in \mathbb{R}^N$$

- Partially observed image:
  - ✓ *m* < N observed pixels

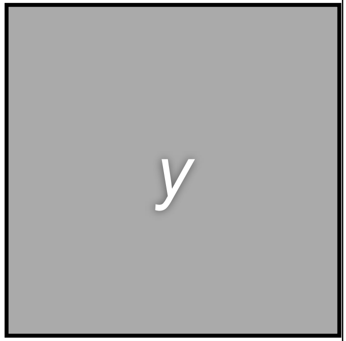
$$b[\vec{p}] = y[\vec{p}],\ \vec{p} \in \mathrm{Observed}$$

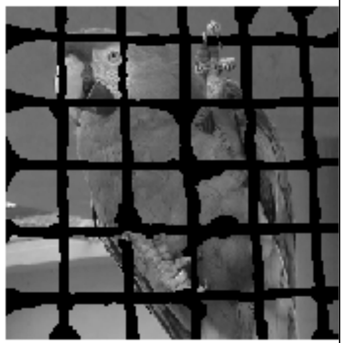
- Measurement matrix

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



Example 3: Inpainting Problem





- Unknown image with N pixels

$$\mathbf{y} \in \mathbb{R}^N$$

- Sparse Model in wavelet domain
  - wavelets coefficients are sparse $\quad x \approx \mathbf{\Phi}^T \mathbf{y}$
  - sparse representation of unknown image $\quad \mathbf{y} \approx \mathbf{\Phi} x$
- Measurement matrix

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

$$\mathbf{A} := \mathbf{M}\mathbf{\Phi}$$

Example 3: image inpainting

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

Wavelets

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

Inpainting

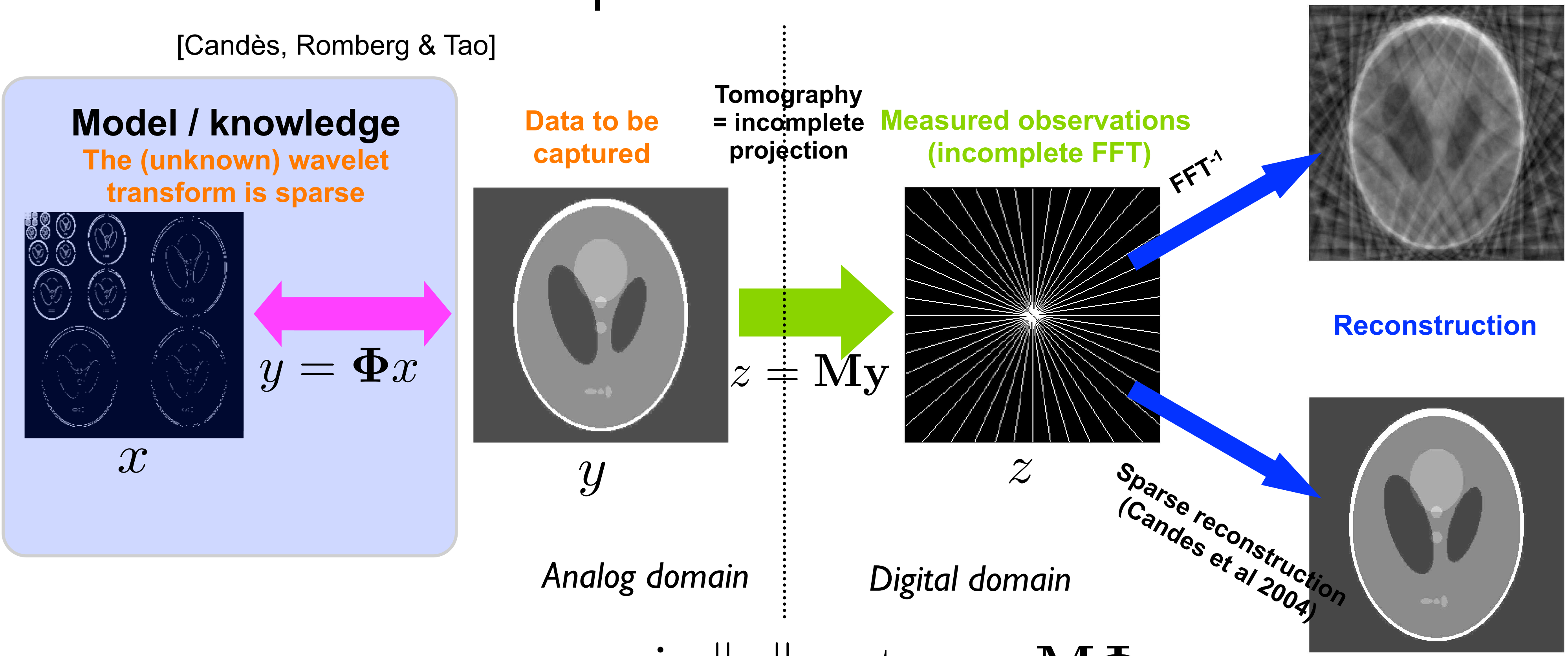
$$\mathbf{b} = \mathbf{M} y = \mathbf{M}\mathbf{\Phi} x$$

$$\boxed{\mathbf{A} = \mathbf{M}\mathbf{\Phi}}$$

Example 4: tomography

- MRI from incomplete data

[Candès, Romberg & Tao]



$$\min \|x\|_1 \text{ s.t. } z = \mathbf{M}\mathbf{\Phi} x$$

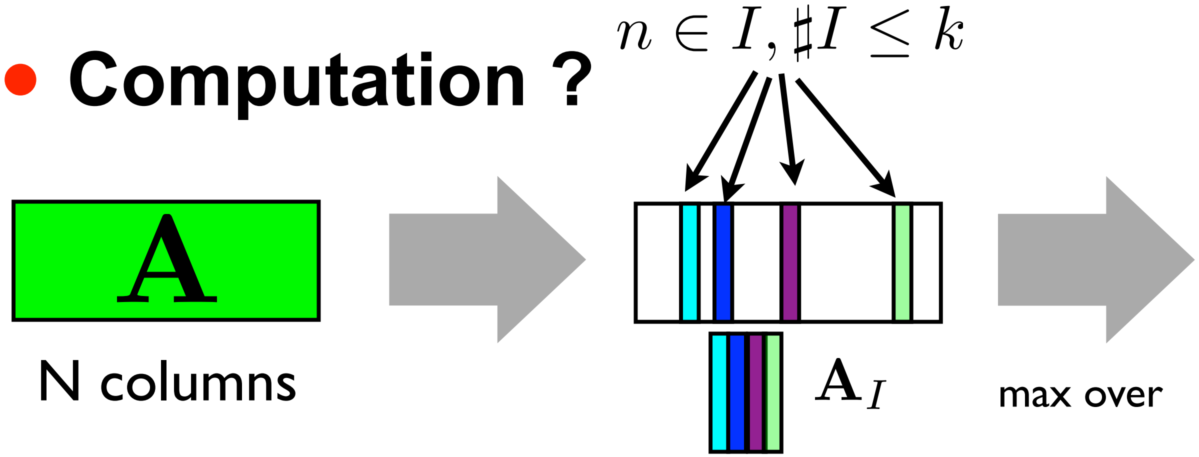
$$\boxed{\mathbf{A} = \mathbf{M}\mathbf{\Phi}}$$

Restricted Isometry Constants (RIC)

- **Definition**: smallest δ_k such that for any k-sparse x

$$1-\delta_k \leq \frac{\|\mathbf{A}x\|_2^2}{\|x\|_2^2} \leq 1+\delta_k$$

- **Computation ?**



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

max over $\frac{N!}{k!(N-k)!}$ subsets *I*

- **NP-complete** [Kloiran & Zouzias 2011, Tillmann & Pfetsch 2012, Bandeira & al 2012]

Recovery conditions based on RIC

- **Definition**: RIC of dictionary of order 2k
  - ✓ for any 2k-sparse vector z

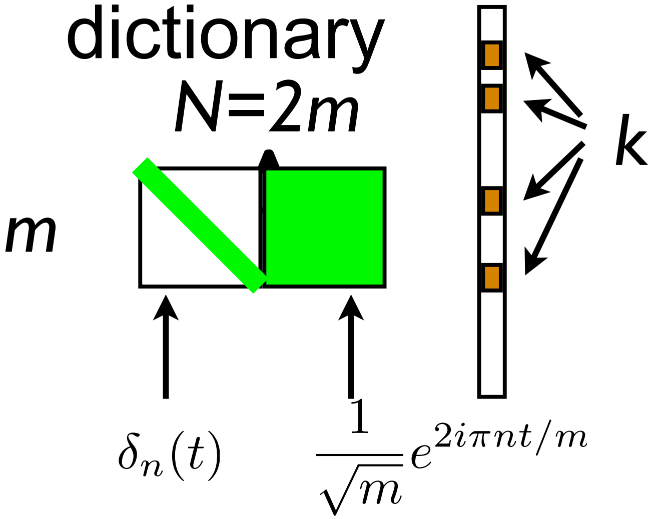
$$(1-\delta_{2k})\|z\|_2^2 \leq \|\mathbf{A}z\|_2^2 \leq (1+\delta_{2k})\|z\|_2^2$$

- **Theorem**:
  - ✓ Assume that $\|x\|_0 \leq k$ and $\delta_{2k} < \sqrt{2}-1 \approx 0.414$

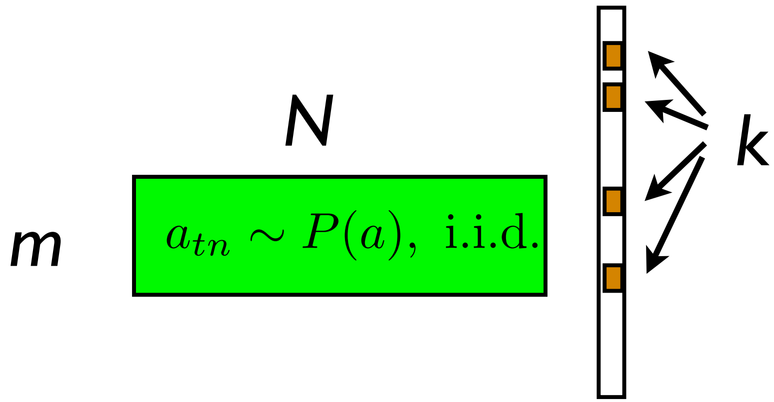
    Restricted Isometry Property (RIP)
  - ✓ Then
    - ✦ x minimizes the L0 and **L1** norm among all solutions x' to the linear inverse problem $\mathbf{A}x' = \mathbf{A}x$

Coherence vs RIP

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



$\mathbf{A}$



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

$m \geq Ck \log N/k$

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

Recovery regimes

$k_1(\mathbf{A}) \approx 0.914\sqrt{m}$

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

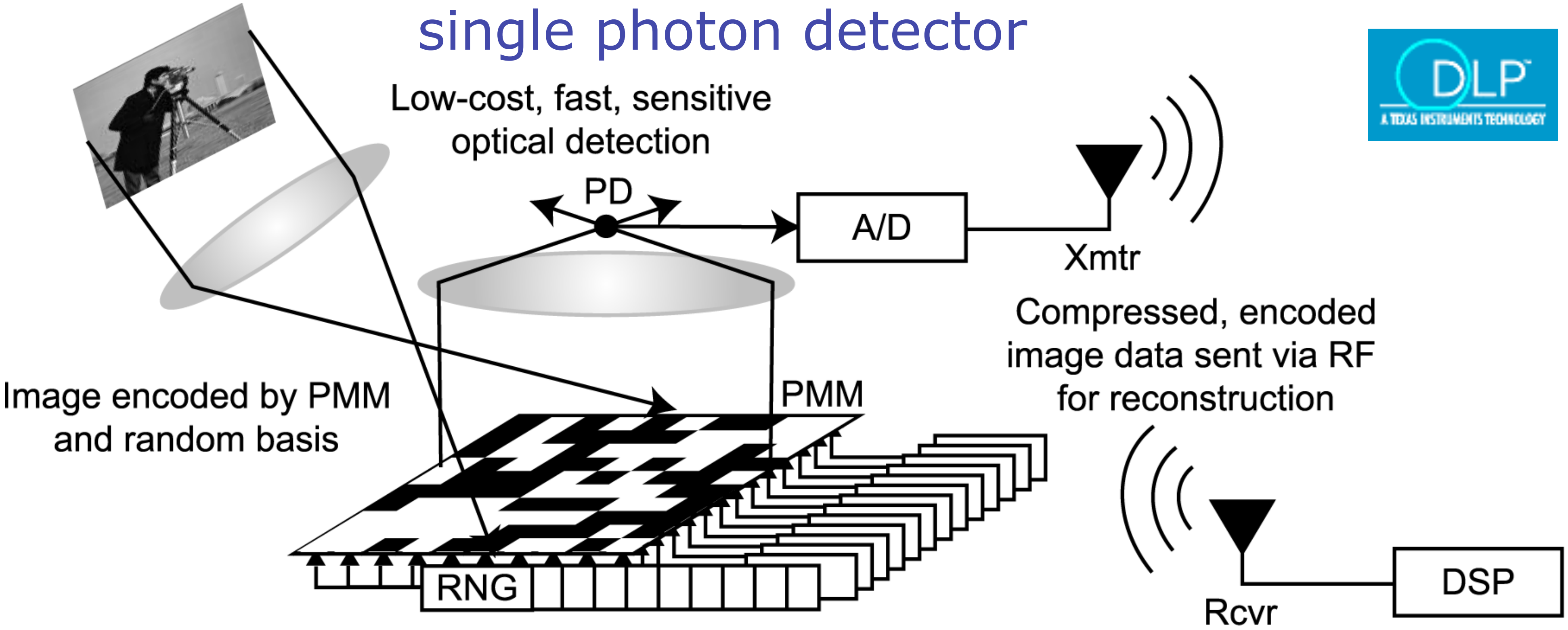
$k_1(\mathbf{A}) \approx \frac{m}{2e \log N/m}$

with high probability for *Gaussian* **A**

[Elad & Bruckstein 2002]

[Donoho & Tanner 2009]

Example : single-pixel camera, Rice University



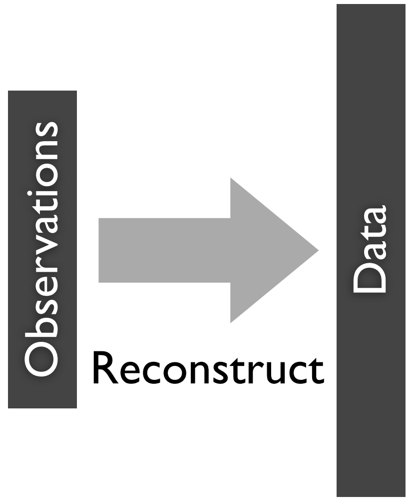
Random pattern on DMD array

image reconstruction

Summary

Inverse problems

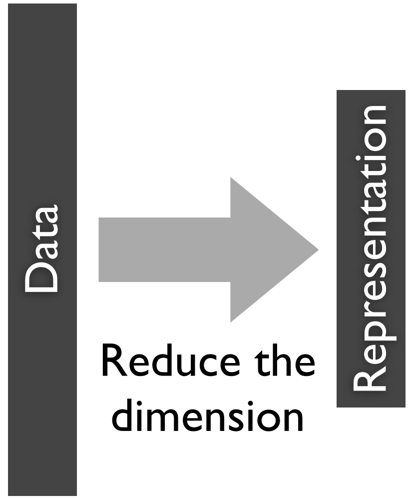
- **Inverse problem** : exploit indirect or incomplete obervation to recontruct some data



$$z = \mathbf{M}\mathbf{y}$$

fewer equations than unknowns

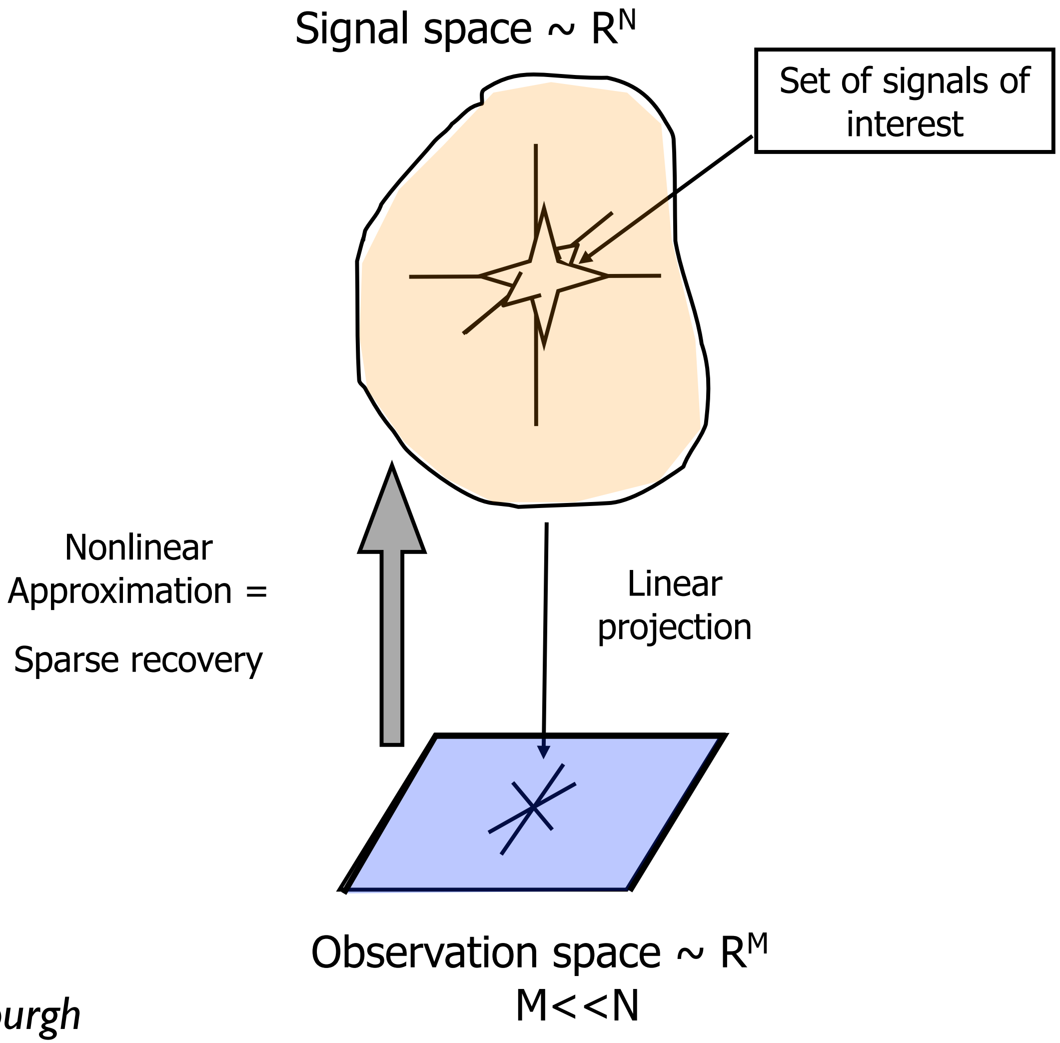
- **Sparsity** : represent / approximate high-dimensional & complex data using few parameters



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

few nonzero components

Inverse problems



Courtesy: M. Davies, U. Edinburgh

Linear inverse problems

- **Definition**: a problem where a high-dimensional vector must be estimated from its low dimensional projection
- **Generic form:**

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

observation/measure   unknown   noise

projection matrix

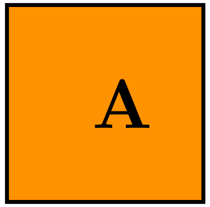
- ✓ m observations / measures $\mathbf{b} \in \mathbb{R}^m$
- ✓ N unknowns $\mathbf{y} \in \mathbb{R}^N$

$$\mathbf{A} \in \mathbb{R}^{m \times N}$$

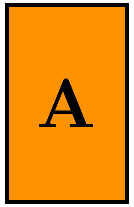
Classes of linear inverse problems

- **Determined:** the matrix **A** is square and invertible
  - ✓ Unique solution to $\mathbf{b} = \mathbf{A}\mathbf{y}$
  - ✓ Linear function of observations

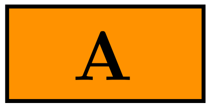
$$\mathbf{y} = \mathbf{A}^{-1}\mathbf{b}$$



- **Over-determined:** more equations than unknowns
  - ✓ Unique solution to $\mathbf{b} = \mathbf{A}\mathbf{y}$:
  - ✓ Linear function of observations
  - ✓ with pseudo-inverse $\mathbf{y} = \mathbf{A}^{\dagger}\mathbf{b}$



- **Under-determined:** fewer equations than unknowns
  - ✓ **Infinitely many solutions to $\mathbf{b} = \mathbf{A}\mathbf{y}$**
  - ✓ **Need to choose one?**



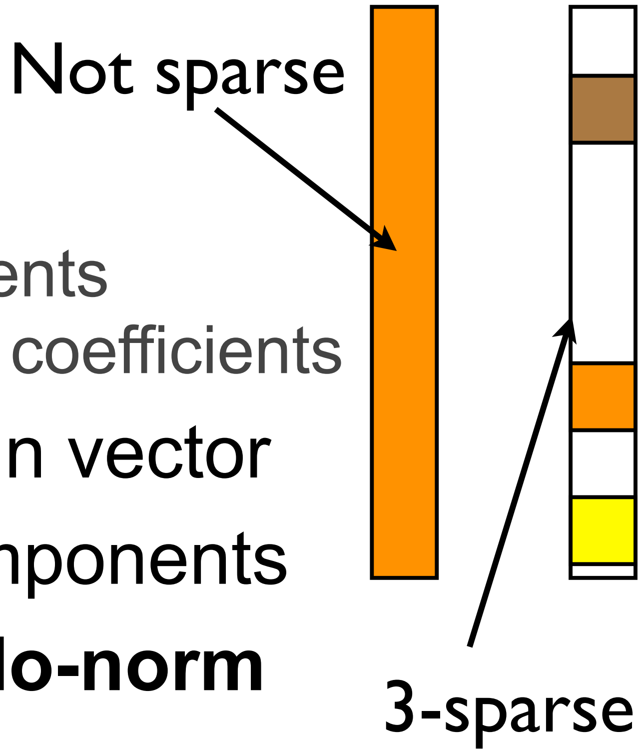
Inverse Problems & Sparsity: Mathematical foundations

- Bottleneck 1990-2000 : under-determined = fewer equations than unknowns = ill-posed

$$\mathbf{A}x_0 = \mathbf{A}x_1 \not\Rightarrow x_0 = x_1$$

- Novelty 2001-2006 :
  - ✓ **Uniqueness** of sparse solution:
    - ✦ if x_0, x_1 are "sufficiently sparse" (in an appropriate «domain»),
    - ✦ then $\mathbf{A}x_0 = \mathbf{A}x_1 \Rightarrow x_0 = x_1$
  - ✓ Recovery of x_0 with **efficient** algorithms
    - ✦ Thresholding, Matching Pursuits, Minimisation of Lp norms p<=1,...

Sparsity: definition



- A vector is
  - ✓ **sparse** if it has (many) zero coefficients
  - ✓ ***k*-sparse** if it has *at most k* nonzero coefficients
- Symbolic representation as column vector
- **Support** = indices of nonzero components
- Sparsity measured with **L0 pseudo-norm**

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

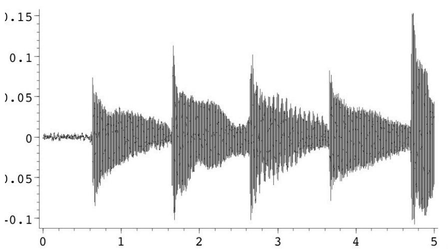
Convention here

$$a^0 = 1(a > 0); 0^0 = 0$$

- *In french:*
  - ✦ sparse -> *«creux», «parcimonieux»*
  - ✦ sparsity, sparseness -> *«parcimonie»,* ~~*«sparsité»*~~

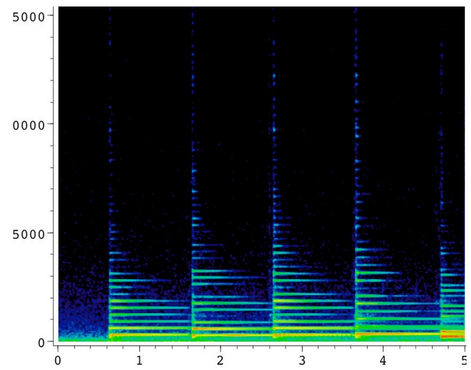
Notion of sparse representation

- Audio : time-frequency representations (MP3)



ANALYSIS

SYNTHESIS



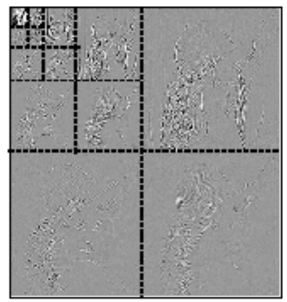
Black
= zero

- Images : wavelet transform (JPEG2000)

ORIGINAL
128, 129, 125, 64, 65,

ANALYSIS

SYNTHESIS



TRANSFORM COEFFICIENTS
4123, -12.4, -96.7, 4.5,

Gray
= zero

Mathematical expression of the sparsity assumption

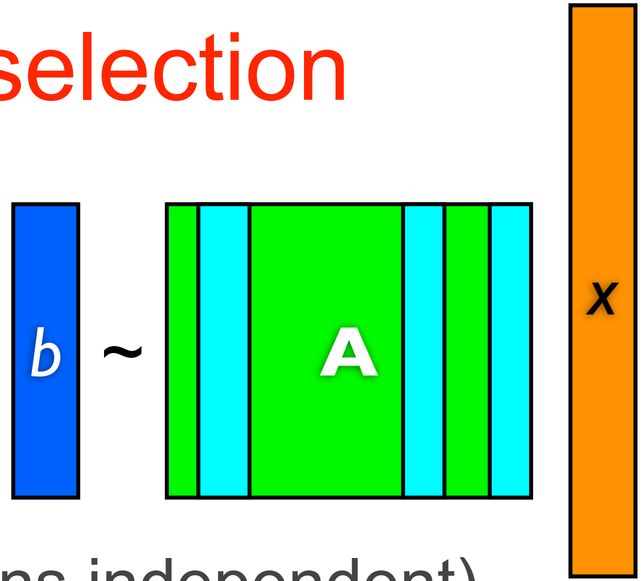
- Signal / image = high dimensional vector

$$y \in \mathbb{R}^N$$

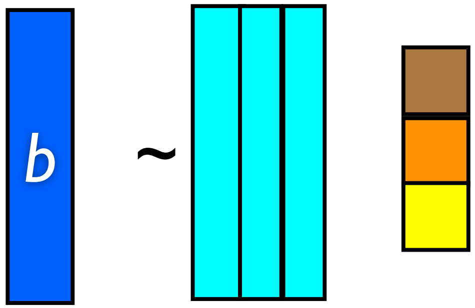
- Definition:
  - ✓ **Atoms**: basis vectors $\varphi_k \in \mathbb{R}^N$
    - ✦ ex: time-frequency atoms, wavelets
  - ✓ **Dictionary:**
    - ✦ collection of atoms $\{\varphi_k\}_{1 \leq k \leq K}$
    - ✦ matrix $\mathbf{\Phi} = [\varphi_k]_{1 \leq k \leq K}$ which columns are the atoms
- Sparse **signal model** = combination of few atoms

$$y \approx \sum_k x_k \varphi_k = \mathbf{\Phi} x$$

Sparsity and subset selection



- Under-determined system
  - ✓ Infinitely many solutions
- If vector is sparse:
  - ✓ If support is known (and columns independent)
    - ✦ nonzero values characterized by (over)determined linear problem
  - ✓ **If support is unknown**
    - ✦ Main issue = finding the support!
    - ✦ This is the **subset selection problem**

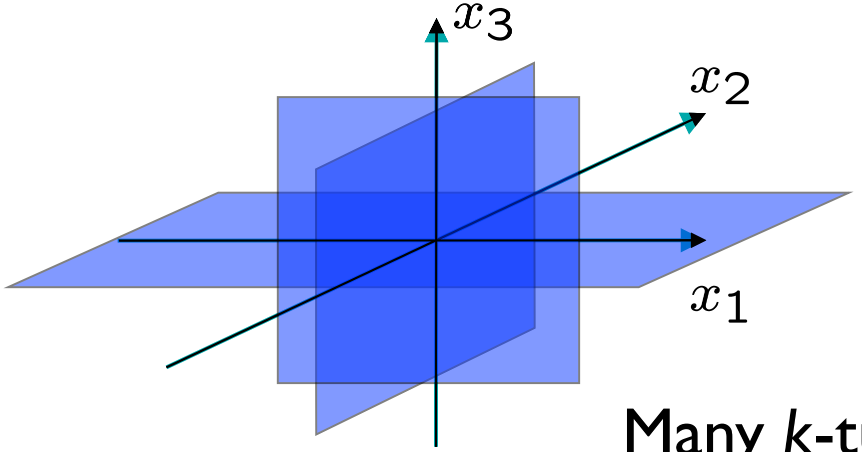


- Objectives of the course
  - ✦ **Well-posedness** of subset selection
  - ✦ Efficient subset selection algorithms = **pursuit algorithms**
  - ✦ **Stability guarantees** of pursuits

Complexity of Ideal Sparse Approximation

- Naive: Brute force search

$$\min_x \|\mathbf{b} - \mathbf{A}x\|_2 \text{ s.t. support}(x) = I$$



$$\begin{bmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{bmatrix} \cdot \begin{pmatrix} x_1 \\ x_3 \end{pmatrix} = \mathbf{b}$$

$$\begin{pmatrix} x_1 \\ x_3 \end{pmatrix} = \begin{bmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{bmatrix}^{-1} \cdot \mathbf{b}$$

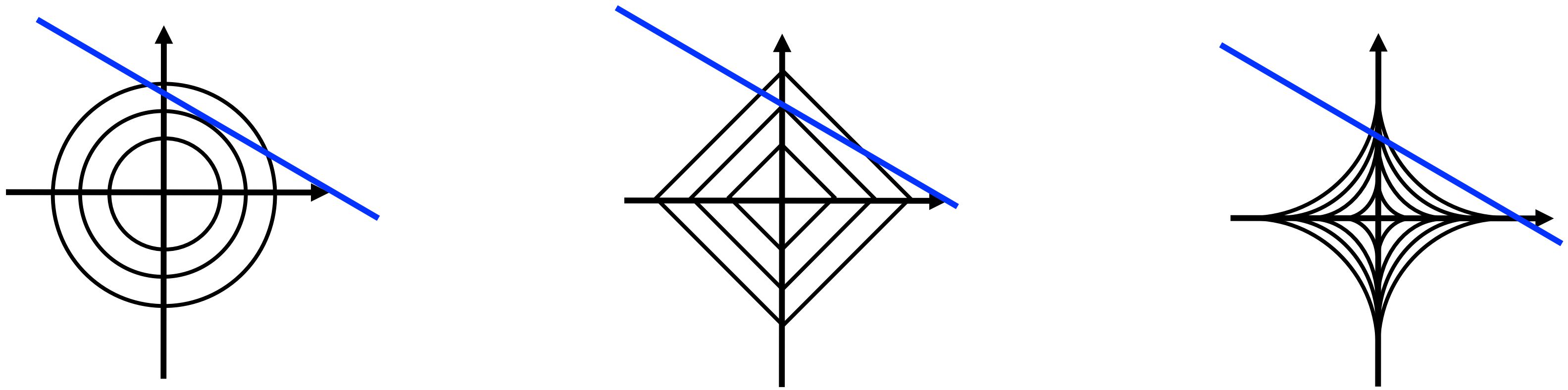
Many *k*-tuples to try!

- **Theorem** *(Davies et al, Natarajan)*

  Solving the L0 optimization problem is **NP-complete**

Lp "norms" level sets

- Strictly convex when $p>1$
- Convex $p=1$
- Nonconvex $p<1$



Observation: *the minimizer is sparse*

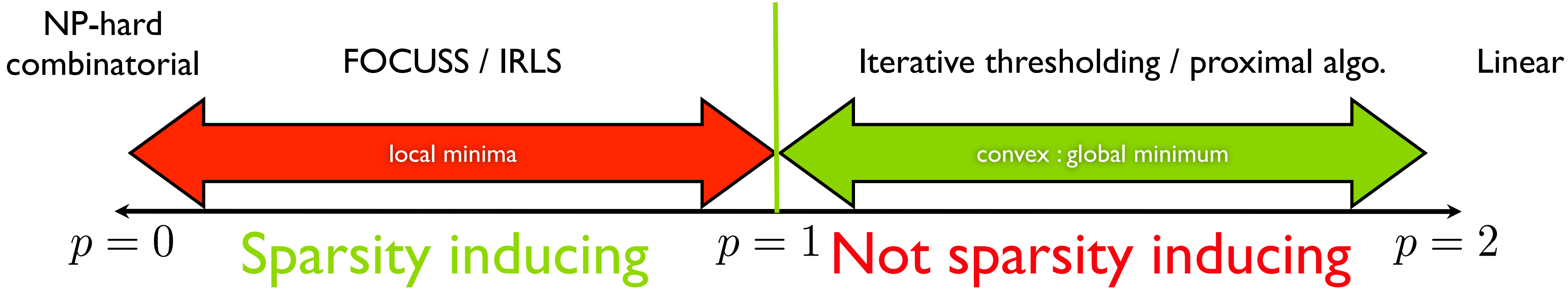
 $\{x \text{ s.t.} \mathbf{b} = \mathbf{A}x\}$

Global Optimization : from Principles to Algorithms

- Optimization principle

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

  - ✓ Sparse representation $\quad \lambda \to 0 \quad \mathbf{A}x = \mathbf{b}$
  - ✓ Sparse approximation $\quad \lambda > 0 \quad \mathbf{A}x \approx \mathbf{b}$



Lasso [Tibshirani 1996], *Basis Pursuit (Denoising)* [Chen, Donoho & Saunders, 1999]
Linear/Quadratic programming (interior point, etc.)
Homotopy method [Osborne 2000] / Least Angle Regression [Efron &al 2002]
Iterative / proximal algorithms [Daubechies, de Frise, de Mol 2004, Combettes & Pesquet 2008, Beck & Teboulle 2009 ...]

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 $\mathbf{r}_i = \mathbf{b} - \mathbf{A}x_i$ • select new components • update residual
Tuning quality/sparsity	regularization parameter λ	stopping criterion (nb of iterations, error level, ...) $\|x_i\|_0 \geq k$ $\|\mathbf{r}_i\| \leq \epsilon$
Variants	• choice of sparsity measure p • **optimization algorithm** • initialization	•selection criterion (weak, stagewise ...) •update strategy (orthogonal ...)

Notions of Kruskal rank / spark
Well-posedness of L0 problem

- **Definition:** Kruskal rank K-rank(**A**):
  - ✓ maximal L such that ***every L columns linearly indep.***
- **Definition**: spark(**A**)
  - ✓ size of minimal set of linearly dependent columns
- **Property:** $\text{K-rank}(\mathbf{A}) = \text{spark}(\mathbf{A}) - 1 \leq \text{rank}(\mathbf{A})$
- **Theorem**: let $\mathbf{b} := \mathbf{A}x$
  - ✓ if x is *k*-sparse with $2k \leq \text{K-rank}(\mathbf{A})$
  - ✓ then x is the **unique** k-sparse vector satisfying $\mathbf{b} = \mathbf{A}x$
  - ✓ hence $x = \arg \min_{x' | \mathbf{b} = \mathbf{A}x'} \|x\|_0$

Recovery conditions based on coherence

- **Convention**: normalized columns $\|\mathbf{a}_i\|_2 = 1$
- **Definition**: coherence of dictionary

$$\mu(\mathbf{A}) := \max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$$

- **Theorem**:
  - ✓ Assume that

$$\boxed{k = \|x\|_0 < \frac{1}{2}(1 + 1/\mu)}$$

  - ✓ Then
    - ✦ x minimizes the L0 and **L1** norm among all solutions x' to the linear inverse problem $\mathbf{A}x' = \mathbf{A}x$
    - ✦ k steps of **OMP** performed on $\mathbf{b} = \mathbf{A}x$ recover x

Conclusions

- Sparsity: prior to solve **ill-posed inverse problems**
- If solution sufficiently sparse, **reasonable algorithms are guaranteed to find it**
- **Computational efficiency still a challenge**
  - problem sizes up to 1000 x 10000 efficiently tractable.
- Theoretical guarantees are **mostly worst-case**
  - Empirical recovery goes far beyond, but is not fully understood.
- Challenging practical issues include:
  - ✓ choosing / learning / designing dictionaries;
  - ✓ exploiting structures beyond sparsity;
  - ✓ designing feasible compressed sensing hardware.

Hot Topics, not covered in this course

- Structured sparsity: group LASSO, etc.
- Analysis vs synthesis sparsity
- Combinatorial algorithms: submodular functions, etc.
- Approximate Message Passing algorithms