Inverse problems and sparse representations

Rémi Gribonval, DR INRIA
EPI METISS (Speech and Audio Processing)
INRIA Rennes - Bretagne Atlantique

remi.gribonval@inria.fr
http://www.irisa.fr/metiss/members/remi/talks
Further material on sparsity

• **Books with a Signal Processing perspective**

• **Review paper:**
  ✦ Bruckstein, Donoho, Elad, SIAM Reviews, 2009

• **Video lectures**
  ✦ E. Candès, MLSS’09
  ✦ F. Bach, NIPS 2009
  ✦ Sparsity in Machine Learning and Statistics SMLS’09
Structure of the course

• Session 1: Panorama
  ✓ sparsity: compression, inverse problems, learning
  ✓ introduction to compressed (random) sensing

• Session 2: Algorithms
  ✓ review of main algorithms & complexities

• Session 3: Guarantees for Deterministic vs Random dictionaries
  ✓ compared success guarantees for different algorithms
  ✓ robust guarantees & Restricted Isometry Property
  ✓ explicit guarantees for various inverse problems
Overview of Session 1

• Sparsity and compression of large-scale data
• Sparsity for source separation, inverse problems, and learning
• Sparse decomposition algorithms
  ✓ L1 minimisation
  ✓ Matching Pursuits
• Provably good algorithms
• Sparsity and compressed sensing
Sparsity & data compression
Large-scale data

- **Fact**: digital data = large volumes
  - 1 second stereo audio, CD quality = 1.4 Mbit
  - 1 uncompressed 10 M pixels picture = 240 Mbit

- **Need**: «concise» data representations
  - storage & transmission (volume / bandwidth) ...
  - manipulation & processing (algorithmic complexity)
Sparse representations

- Audio: time-frequency representations (MP3)
- Images: wavelet transform (JPEG2000)

Black = zero
Gray = zero
Mathematical expression

- Signal / image = high dimensional vector
  \[ y \in \mathbb{R}^N \]

- **Model** = linear combination of basis vectors (ex: *time-frequency atoms, wavelets*)
  \[ y \approx \sum_k x_k \varphi_k = \Phi x \]

- **Sparsity** = small L0 (quasi)-norm
  \[ \|x\|_0 = \sum_k |x_k|_0 = \text{card}\{k, x_k \neq 0\} \]
**Sparsity & compression**

- Full vector
  - $N$ entries
  - $= N$ floats
  - $y \approx \Phi \cdot x$

- Sparse vector
  - $k \ll N$ nonzero entries
  - $= k$ floats
  - + $k$ positions among $N$
  - $= \log_2 \left( \frac{N}{k} \right) \approx k \log_2 \frac{N}{k}$ bits

Key practical issues: choose dictionary
Sparsity & inverse problems
Example: image inpainting

Courtesy of: G. Peyré, Ceremade, Université Paris 9 Dauphine
Example: audio source separation

• « Softly as in a morning sunrise »
Inverse problems

- **Inverse problem**: exploit indirect or incomplete observation to reconstruct some data
  \[ z = My \]
  fewer equations than unknowns

- **Sparsity**: represent / approximate high-dimensional & complex data using few parameters
  \[ y \approx \Phi x \]
  few nonzero components
Blind Source Separation

• Mixing model: linear instantaneous mixture

\[
\begin{bmatrix}
y_{\text{right}}(t) \\
y_{\text{left}}(t)
\end{bmatrix} = A \begin{bmatrix} s_1(t) \\
s_2(t) \\
s_3(t) \end{bmatrix}
\]

• Source model: if disjoint time-supports ...

... then clustering to:
1- identify (columns of) the mixing matrix
2- recover sources
Blind Source Separation

- Mixing model: linear instantaneous mixture

\[
y_{\text{right}}(t) \begin{pmatrix} s_1(t) \\ s_2(t) \\ s_3(t) \end{pmatrix} = A \begin{pmatrix} y_{\text{left}}(t) \\ y_{\text{left}}(t) \end{pmatrix}
\]

- In practice ...

\[
y_{\text{left}}(t) \quad y_{\text{right}}(t)
\]
Time-Frequency Masking

- Mixing model in the time-frequency domain

\[
\begin{bmatrix}
Y_{\text{right}}(\tau, f) \\
Y_{\text{left}}(\tau, f)
\end{bmatrix} = \mathbf{A} \mathbf{S}(\tau, f)
\]

- And “miraculously” ...

... time-frequency representations of audio signals are (often) almost disjoint.
Inverse Problems & Sparsity: Mathematical foundations

• Bottleneck 1990-2000: fewer equations than unknowns
  \[ \mathbf{A} \mathbf{x}_0 = \mathbf{A} \mathbf{x}_1 \nRightarrow \mathbf{x}_0 = \mathbf{x}_1 \]

• Novelty 2001-2006:
  ✓ Uniqueness of sparse solution:
    ✦ if \( \mathbf{x}_0, \mathbf{x}_1 \) are “sufficiently sparse”,
    ✦ then \( \mathbf{A} \mathbf{x}_0 = \mathbf{A} \mathbf{x}_1 \Rightarrow \mathbf{x}_0 = \mathbf{x}_1 \)

✓ Recovery of \( \mathbf{x}_0 \) with practical algorithms
  ✦ Thresholding, Matching Pursuits, Minimisation of Lp norms \( p \leq 1, \ldots \)
Algorithmic principles for sparse approximation
Signal Processing Vocabulary

Unknown $x$
representation, sources, ...

Forward linear model $b \approx Ax$

Known linear system: $A$
dictionary, mixing matrix, sensing system...

Decomposition Reconstruction Separation

Observed data: $b$
signal, image, mixture of sources,...
Forward linear model

\[ b \approx A x \]

\[ y = X \beta \]

Known linear system:
dictionary, mixing matrix, sensing system...

\[ y = X \beta \]

Unknown representation, sources, ...

\[ \beta \]

Regression coeffs

Design matrix: \( A \)

Decomposition
Reconstruction
Separation

Observed data:
signal, image, mixture of sources, ...

Observation

lundi 12 septembre 2011
Ideal sparse approximation

• Input:
  \[ m \times N \text{ matrix } \mathbf{A}, \text{ with } m < N, \]  \[ m\text{-dimensional vector } \mathbf{b} \]

• Possible objectives:
  find the sparsest approximation within tolerance
  \[ \arg \min_{\mathbf{x}} \|\mathbf{x}\|_0, \ \text{s.t.} \|\mathbf{b} - \mathbf{A}\mathbf{x}\| \leq \epsilon \]
  find best approximation with given sparsity
  \[ \arg \min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|, \ \text{s.t.} \|\mathbf{x}\|_0 \leq k \]
  find a solution \( \mathbf{x} \) to
  \[ \|\mathbf{b} - \mathbf{A}\mathbf{x}\| \leq \epsilon, \text{ and } \|\mathbf{x}\|_0 \leq k \]
Geometric interpretation of sparse approximation

- Coefficient domain $\mathbb{R}^N$:
  - set $\sum_k$ of sparse vectors
  $$\|x\|_0 \leq k$$

- Set $A \sum_k = \binom{N}{k}$ subspaces in signal domain

- Ideal sparse approximation
  = find nearest subspace among $\binom{N}{k}$

Combinatorial search!
Actual complexity? NP-complete!
Practical approaches: Optimization *principles*
Overall compromise

• Approximation quality

\[ \| A x - b \|_2 \]

• Ideal sparsity measure: \( \ell^0 \) “norm”

\[ \| x \|_0 := \# \{ n, \ x_n \neq 0 \} = \sum_n |x_n|^0 \]

• “Relaxed” sparsity measures

\[ 0 < p < \infty, \ \| x \|_p := \left( \sum_n |x_n|^p \right)^{1/p} \]
Lp norms / quasi-norms

• **Norms** when $1 \leq p < \infty$ = convex
  \[ \|x\|_p = 0 \iff x = 0 \]
  \[ \|\lambda x\|_p = |\lambda|\|x\|_p, \forall \lambda, x \]
  
  Triangle inequality
  \[ \|x + y\|_p \leq \|x\|_p + \|y\|_p, \forall x, y \]

• **Quasi-norms** when $0 < p < 1$ = nonconvex
  \[ \|x + y\|_p \leq 2^{1/p} (\|x\|_p + \|y\|_p), \forall x, y \]
  
  Quasi-triangle inequality
  \[ \|x + y\|^p_p \leq \|x\|^p_p + \|y\|^p_p, \forall x, y \]

• **"Pseudo"-norm for p=0**
  \[ \|x + y\|_0 \leq \|x\|_0 + \|y\|_0, \forall x, y \]
Optimization problems

• Approximation

$$\min_x \|b - Ax\|_2 \quad \text{s.t.} \quad \|x\|_p \leq \tau$$

• Sparsification

$$\min_x \|x\|_p \quad \text{s.t.} \quad \|b - Ax\|_2 \leq \varepsilon$$

• Regularization

$$\min_x \frac{1}{2}\|b - Ax\|_2^2 + \lambda \|x\|_p$$
Lp “norms” level sets

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

Observation: the minimizer is sparse

$\{x \text{ s.t. } b = Ax\}$
Global Optimization: from Principles to Algorithms

- **Optimization principle**

\[
\min_x \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_p^p
\]

- **Sparse representation**
- **Sparse approximation**

- NP-hard combinatorial
  - FOCUSS / IRLS
  - Iterative thresholding / proximal algo.
  - Linear

- **Local minima**
- **Convex: global minimum**

- **Lasso** [Tibshirani 1996], **Basis Pursuit (Denoising)** [Chen, Donoho & Saunders, 1999]
- Linear/Quadratic programming (interior point, etc.)
- Iterative / proximal algorithms [Daubechies, de Frise, de Mol 2004, Combettes & Pesquet 2008, ...]

\[\lambda \to 0 \quad Ax = b\]
\[\lambda > 0 \quad Ax \approx b\]

\(\lambda \to 0\) if \(p = 0\)
\(\lambda > 0\) if \(p \geq 1\)
Greedy Algorithms
Greedy algorithms

- Observation: when $A$ is orthonormal,
  - the problem
    \[
    \min_x \| b - A x \|_2^2 \quad \text{s.t.} \quad \| x \|_0 \leq k
    \]
  - is equivalent to
    \[
    \min_x \sum_n (a_n^T b - x_n)^2 \quad \text{s.t.} \quad \| x \|_0 \leq k
    \]

- Let $\Lambda_k$ index the $k$ largest inner products
  \[
  \min_n |a_n^T b| \geq \max_{n \notin \Lambda_k} |a_n^T b|
  \]
  - an optimum solution is
    \[
    x_n = a_n^T b, \quad n \in \Lambda_k; \quad x_n = 0, \quad n \notin \Lambda_k
    \]
Greedy algorithms

- Iterative algorithm (= Matching Pursuit)
  ✓ Initialize a residual to $r_0 = b$  \( i = 1 \)
  ✓ Compute all inner products
    \[
    A^T r_{i-1} = (a_n^T r_{i-1})_{n=1}^N
    \]
  ✓ Select the largest in magnitude
    \[
    n_i = \text{arg max}_n |a_n^T r_{i-1}|
    \]
  ✓ Compute an updated residual
    \[
    r_i = r_{i-1} - (a_{n_i}^T r_{i-1}) a_{n_i}
    \]
  ✓ If $i \geq k$ then stop, otherwise increment $i$ and iterate
Matching Pursuit (MP)

- Matching Pursuit (aka Projection Pursuit, CLEAN)
  - Initialization \( r_0 = b \quad i = 1 \)
  - Atom selection:
    \[
    n_i = \arg \max_n |a_n^T r_{i-1}|
    \]
  - Residual update
    \[
    r_i = r_{i-1} - (a_{n_i}^T r_{i-1}) a_{n_i}
    \]

- Energy preservation (Pythagoras theorem)
  \[
  \|r_{i-1}\|_2^2 = |a_{n_i}^T r_{i-1}|^2 + \|r_i\|_2^2
  \]
# Summary

## Global optimization

<table>
<thead>
<tr>
<th>Principle</th>
<th>Iterative greedy algorithms</th>
</tr>
</thead>
</table>
| <br>\[
\min_x \frac{1}{2} \|Ax - b\|^2 + \lambda \|x\|^p_p
\] | iterative decomposition \[\mathbf{r}_i = b - Ax_i\] |
| Tuning quality/sparsity | stopping criterion <br>\[\|x_i\|_0 \geq k \quad \|\mathbf{r}_i\| \leq \epsilon\] |
| Variants | • choice of sparsity measure \(p\) <br>• optimization algorithm <br>• initialization | • selection criterion (weak, stagewise ...) <br>• update strategy (orthogonal ...) |
Provably good algorithms
Inverse problems

Signal space $\sim \mathbb{R}^N$

Set of signals of interest

Nonlinear Approximation = Sparse recovery

Linear projection

Observation space $\sim \mathbb{R}^M$

$M \ll N$

Courtesy: M. Davies, U. Edinburgh
Recovery analysis for inverse problem $b = Ax$

- Recoverable set for a given “inversion” algorithm

$$\{ x = \text{Algo1}(Ax) \}$$
Recovery analysis for inverse problem $b = Ax$

- Recoverable set for a given “inversion” algorithm
- Level sets of L0-norm

\[
\{ x = \text{Algo1}(Ax) \}
\]
Recovery analysis for inverse problem $b = Ax$

- Recoverable set for a given "inversion" algorithm
- Level sets of $L_0$-norm
  \[ \{ x = \text{Algo1}(Ax) \} \]
  \[ \| x \|_0 \leq 1 \]
Recovery analysis for inverse problem $b = Ax$

- Recoverable set for a given “inversion” algorithm
- Level sets of L0-norm
  - 1-sparse
  - 2-sparse
Recovery analysis for inverse problem $b = Ax$

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

\[ \|x\|_0 \leq k \]
\[ \{ x = \text{Algol}(A \cdot x) \} \]
Recovery analysis for inverse problem $b = Ax$

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

\[
\{ x = \text{Algo2}(Ax) \}
\]

\[
\|x\|_0 \leq k
\]

\[
\|x\|_0 \leq 1
\]
Recovery analysis for inverse problem $b = Ax$

- Recoverable set for a given “inversion” algorithm
- Level sets of L0-norm
  - 1-sparse
  - 2-sparse
  - 3-sparse ...
Recovery analysis for inverse problem $b = Ax$

- Recoverable set for a given “inversion” algorithm
- Level sets of L0-norm
  - 1-sparse
  - 2-sparse
  - 3-sparse ...
Equivalence between L0, L1, OMP

• Theorem: assume that \( \mathbf{b} = \mathbf{A} \mathbf{x}_0 \)

\[
\begin{align*}
\text{✓ if} & \quad \| \mathbf{x}_0 \|_0 \leq k_0(\mathbf{A}) \quad \text{then} \quad \mathbf{x}_0 = \mathbf{x}_0^* \\
\text{✓ if} & \quad \| \mathbf{x}_0 \|_0 \leq k_1(\mathbf{A}) \quad \text{then} \quad \mathbf{x}_0 = \mathbf{x}_1^*
\end{align*}
\]

where \( x_p^* = \arg\min_{\mathbf{A}x = \mathbf{A}x_0} \| x \|_p \)

• Donoho & Huo 01: pair of bases, coherence
• Donoho & Elad, Gribonval & Nielsen 2003: dictionary, coherence
• Tropp 2004: Orthonormal Matching Pursuit, cumulative coherence
• Candes, Romberg, Tao 2004: random dictionaries, restricted isometry constants
Compressed sensing
Example: tomography

- MRI from incomplete data

[Candès, Romberg & Tao]

Model / knowledge
The (unknown) wavelet transform is sparse

Data to be captured
Tomography = incomplete projection
Measured observations (incomplete FFT)

Analog domain
Digital domain

\[ y = \Phi x \]

\[ z = \text{My} \]

\[ \min \| x \|_1 \text{ s.t. } z = M\Phi x \]

Reconstruction

Sparse reconstruction (Candès et al 2004)
Classical Shannon Sampling

- « Sample first, think and compress afterwards »

\[ z = M y \]

A/D conversion
High resolution
Analog domain Digital domain

\( M = \) sampling operator
Classical Shannon Sampling

- « Sample first, think and compress afterwards »

\[ y = \Phi x \]

Sparse model

A/D conversion

High resolution

Analog domain

Digital domain
Classical Shannon Sampling

- « Sample first, think and compress afterwards »

- \[ y = \Phi x \]

- Analog domain
  - Sparse model
  - A/D conversion
  - High resolution

- Digital domain
  - Compression (~costly)
  - Decoding (~cheap)

JPEG / MP3
Compressed Sensing

- First model the data, then sample & compress

\[ y = \Phi x \]
Compressed Sensing

- First model the data, then sample & compress

\[ y = \Phi x \]

A/D conversion and compression (~cheap)

\[ z = M y \]

Sparse model

Analog domain

Digital domain
Compressed Sensing

- First model the data, then sample & compress

Analog domain

Sparse model

\[ y = \Phi x \]

A/D conversion and compression

\( y \rightarrow z = My \)

(≈ cheap)

Sparse recovery (≈ costly)

Digital domain

\[ \min \|x\|_1, \text{ subject to } z = K\Phi x \]
Conditions of success of Compressed Sensing

- Knowledge: transform domain where data is sparse

- "Incoherence" between measurement domain and sparse domain (uncertainty principle à la Heisenberg)
  - time domain / frequency domain
  - spatial domain / frequency domain
  - random measurements!

- Sufficiently many measures \( m \geq Ck \log_2 \frac{N}{k} \)
  - necessary
  - sufficient (with random Gaussian measures)
Why the log factor?

- Full vector

\[ N \text{ entries} = N \text{ floats} \]

\[ y \approx \Phi \cdot x \]

- Sparse vector

\[ k \ll N \text{ nonzero entries} = k \text{ floats} \]

+ k positions among N

\[ = \log_2 \left( \binom{N}{k} \right) \approx k \log_2 \frac{N}{k} \text{ bits} \]

Key practical issues: choose dictionary
Summary

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

Compression
Representation
Description
Classification

Natural / traditional role

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

Denoising
Blind source
separation
Compressed
sensing
...

Novel indirect role

Sparisty = prior knowledge, regularization
Tool for inverse problems
Example: single-pixel camera, Rice University

- Single photon detector
- Low-cost, fast, sensitive optical detection
- Image encoded by PMM and random basis
- Random pattern on DMD array
- Compressed, encoded image data sent via RF for reconstruction
- Image reconstruction
Pursuit Algorithms for Sparse Representations

Rémi Gribonval, DR INRIA
EPI METISS (Speech and Audio Processing)
INRIA Rennes - Bretagne Atlantique

remi.gribonval@inria.fr
http://www.irisa.fr/metiss/members/remi/talks
Structure of the course

• Session 1: Panorama
  ✓ sparsity: compression, inverse problems, learning
  ✓ introduction to compressed (random) sensing

• Session 2: Algorithms
  ✓ review of main algorithms & complexities

• Session 3: Guarantees for Deterministic vs Random dictionaries
  ✓ compared success guarantees for different algorithms
  ✓ robust guarantees & Restricted Isometry Property
  ✓ explicit guarantees for various inverse problems
Summary

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

Natural / traditional role
Sparsity = low cost (bits, computations, ...)
Direct objective

Novel indirect role
Sparsity = prior knowledge, regularization
Tool for inverse problems

Denoising
Blind source separation
Compressed sensing
...
Overview of Session 2

- Convex & nonconvex optimization principles
- Convex & nonconvex optimization algorithms
- Greedy algorithms
- Comparison of complexities
Overall compromise

- Approximation quality
  \[ \| A x - b \|_2 \]

- Ideal sparsity measure: \( \ell^0 \) “norm”
  \[ \| x \|_0 := \# \{ n, \ x_n \neq 0 \} = \sum_n |x_n|^0 \]

- “Relaxed” sparsity measures
  \[ 0 < p < \infty, \ \| x \|_p := \left( \sum_n |x_n|^p \right)^{1/p} \]
Two geometric viewpoints

- Low-dim signal domain
- High-dim coeff domain

Find closest subspace through correlations $A^T b$

Find sparsest representation through (convex) optimization

\{x \text{ s.t.} b = Ax\}
Algorithms for L1:
Linear Programming

• L1 minimization problem of size $m \times N$

$$\min_{x} \|x\|_1, \text{ s.t. } Ax = b$$

• Equivalent linear program of size $m \times 2N$

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

$$c = (c_i), \quad c_i = 1, \forall i$$
L1 regularization: Quadratic Programming

• L1 minimization problem of size $m \times N$

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

Basis Pursuit Denoising (BPDN)

• Equivalent quadratic program of size $m \times 2N$

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

\[
\mathbf{c} = (c_i), \quad c_i = 1, \forall i
\]
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):
Overview

Convex & nonconvex optimization principles
Convex & nonconvex optimization algorithms
Greedy algorithms
Comparison of complexities
What if $A$ is orthonormal?

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

  $$A^T A = A A^T = \text{Id}_N$$

  $$\| b - A x \|^2_2 = \| A^T b - x \|^2_2$$

- **Expression of BPDN criterion to be minimized**

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

- **Minimization can be done coordinate-wise**

  $$\min_{x_n} \frac{1}{2} \left( c_n - x_n \right)^2 + \lambda |x_n|^p$$
Hard-thresholding (p=0)

\[ H_\lambda(c) = \begin{cases} c & \text{if } |c| \geq \sqrt{2\lambda} \\ 0 & \text{if } |c| < \sqrt{2\lambda} \end{cases} \]

- Solution of

\[
\min_x \frac{1}{2} (c - x)^2 + \lambda \cdot |x|^0
\]
Soft-thresholding (p=1)

\[ S_\lambda(c) \]

\[ \min_x \frac{1}{2} (c - x)^2 + \lambda \cdot |x| \]
Iterative thresholding

- **Proximity operator**
  
  $$\Theta^p_\lambda(c) = \arg \min_x \frac{1}{2} (x - c)^2 + \lambda |x|^p$$

- **Goal = compute**
  
  $$\arg \min_x \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_p^p$$

- **Approach = iterative alternation between**
  
  - gradient descent on fidelity term
    
    $$x^{(i+1/2)} := x^{(i)} + \alpha^{(i)} A^T (b - Ax^{(i)})$$
  
  - thresholding
    
    $$x^{(i+1)} := \Theta^p_{\lambda(i)}(x^{(i+1/2)})$$
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 \( p \geq 1 \)

\[
f(x) = \Theta^p_\alpha \lambda (x + \alpha A^T (b - Ax))
\]

✓ assume that \( \forall x, \|Ax\|_2^2 \leq c\|x\|_2^2 \) and \( \alpha < 2/c \)

✓ then, the iterates converge strongly to a limit \( x^* \)

\[
\|x^{(i)} - x^*\|_2 \to_{i \to \infty} 0
\]

✓ the limit \( x^* \) is a global minimum of \( \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_p^p \)

✓ if \( p > 1 \), or if \( A \) is invertible, \( x^* \) is the *unique* minimum
Iterative Thresholding: convex penalties

• Strong convergence to global minimum

• Accelerated convergence: Nesterov schemes
  ✓ see e.g. Beck & Teboulle 2009;

• Many variants of iterative thresholding
  ✓ depends on properties of penalty terms
    ♦ smoothness
    ♦ strong convexity
    ♦ etc.
  ✓ see course by L. Vandenberghe
Iterative Thresholding: nonconvex penalties

- **Example**: Iterative Hard Thresholding for L0
  - keep components above threshold
  - *or rather* keep $k$ largest components
    - [IHT: Blumensath & Davies 2009]

- **More generally**, with *nonconvex* cost functions
  - Possible ‘spurious’ local minima
  - Convergence: fixed point, under certain assumptions
  - Limit = global min: under certain assumptions (RIP)

- **Pruning strategies**:
  - ex: keep 2$k$ components, project, keep $k$ components
    - ex: CoSAMP [Needell & Tropp 2008], ALPS [Cevher 2011], ...
Pareto curve

\[ \frac{1}{2} \left\| \mathbf{b} - \mathbf{A} \mathbf{x} \right\|_2^2 \]

Slope = $-\lambda$

Sparse representation

\[ \left\| \mathbf{x} \right\|_p^p \]
Path of the solution

- **Lemma**: let $x^*$ be a local minimum of BPDN
  
  $\arg \min_{x} \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1$

- let $I$ be its support
- Then
  
  $A^T_I (Ax^* - b) + \lambda \cdot \text{sign}(x^*_I) = 0$
  
  $\|A^T_{Ic} (Ax^* - b)\|_\infty < \lambda$

- In particular
  
  $x_I = (A^T_I A_I)^{-1} (A^T_I b - \lambda \cdot \text{sign}(x_I))$
Homotopy method

- **Principle**: track the solution \( x^*(\lambda) \) of BPDN along the Pareto curve

- **Property**: [Fuchs 97, 05; Osborne 2000]
  - solution is characterized by its sign pattern through
    \[
    x_I = (A_I^T A_I)^{-1} (A_I^T b - \lambda \cdot \text{sign}(x_I))
    \]
  - 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*
  - [Osborne 2000; Efron & al 2004]
Piecewise Linear Path

![Graph showing piecewise linear paths with weights and regularization parameter.](image)

Courtesy: F. Bach
Overview

Convex & nonconvex optimization principles
Convex & nonconvex optimization algorithms
Greedy algorithms
Comparison of complexities
Matching Pursuit with Time-Frequency Atoms

- Audio = superimposition of structures
- Example: glockenspiel

- Transients = short, small scale
- Harmonic part = long, large scale

Gabor atoms

\[
\{ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w \left( \frac{t - \tau}{s} \right) e^{2\pi i ft} \}_{s,\tau,f}
\]
Matching Pursuit with Time-Frequency Atoms

- Audio = superimposition of structures
- Example: glockenspiel

- Transients = short, small scale
- Harmonic part = long, large scale

- Gabor atoms

\[ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w \left( \frac{t - \tau}{s} \right) e^{2i\pi ft} \]
Matching Pursuit with Time-Frequency Atoms

• Audio = superimposition of structures

• Example: glockenspiel

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

• Gabor atoms
\[
\left\{ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w \left( \frac{t - \tau}{s} \right) e^{2i\pi ft} \right\}_{s,\tau,f}
\]
Matching Pursuit with Time-Frequency Atoms

- Audio = superimposition of structures
- Example: glockenspiel

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

- Gabor atoms
  \[ g_{s,\tau,f}(t) = \frac{1}{\sqrt{s}} w\left(\frac{t - \tau}{s}\right) e^{2\pi i ft} \]
Matching Pursuit (MP)

[Friedman & Stuetzle 81; Mallat & Zhang 93]

- Matching Pursuit (aka Projection Pursuit, CLEAN)
  - Initialization \( r_0 = b \)  \( i = 1 \)
  - Atom selection: (assuming normed atoms: \( \| A_n \|_2 = 1 \) )
    \[ n_i = \arg \max_n | A_n^T r_{i-1} | \]
  - Residual update
    \[ r_i = r_{i-1} - (A_{n_i}^T r_{i-1}) A_{n_i} \]

- Energy preservation (Pythagoras theorem)
  \[ \| r_{i-1} \|_2^2 = | A_{n_i}^T r_{i-1} |^2 + \| r_i \|_2^2 \]
Main properties

• Global energy preservation

\[ \|b\|_2^2 = \|r_0\|_2^2 = \sum_{i=1}^{k} |A_{n_i}^T r_{i-1}|^2 + \|r_k\|_2^2 \]

• Global reconstruction

\[ b = r_0 = \sum_{i=1}^{k} (A_{n_i}^T r_{i-1}) A_{n_i} + r_k \]

• Strong convergence (assuming full-rank dictionary)

\[ \lim_{i \to \infty} \|r_i\|_2 = 0 \]
\[ V_k = \text{span}(A_n, n \in \Lambda_k) \]
Orthonormal MP (OMP)

[Mallat & Zhang 93, Pati & al 94]

• Observation: after k iterations
  \[ r_k = b - \sum_{i=1}^{k} \alpha_{k} A_{n_i} \]

• Approximant belongs to
  \[ V_k = \text{span}(A_n, n \in \Lambda_k) \]
  \[ \Lambda_k = \{n_i, 1 \leq i \leq k\} \]

• Best approximation from \( V_k \) = orthoprojection
  \[ P_{V_k} b = A_{\Lambda_k} A_{\Lambda_k}^+ b \]

• OMP residual update rule
  \[ r_k = b - P_{V_k} b \]
OMP

- Same as MP, except residual update rule
  - Atom selection:
    \[ n_i = \arg \max_n |A_n^T r_{i-1}| \]
  - Index update: \[ \Lambda_i = \Lambda_{i-1} \cup \{n_i\} \]
  - Residual update
    \[ V_i = \text{span}(A_n, n \in \Lambda_i) \]
    \[ r_i = b - P_{V_i} b \]

- Property: strong convergence \[ \lim_{i \to \infty} \|r_i\|_2 = 0 \]
Caveats (1)

- MP can pick up the same atom more than once
- OMP will never select twice the same atom
Caveats (1)

• MP can pick up the same atom more than once

• OMP will never select twice the same atom
Caveats (1)

• MP can pick up the same atom more than once

• OMP will never select twice the same atom
Caveats (1)

- MP can pick up the same atom more than once
- OMP will never select twice the same atom
Caveats (1)

- MP can pick up the same atom more than once
- OMP will never select twice the same atom
Caveats (1)

- MP can pick up the same atom more than once
- OMP will never select twice the same atom
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 $i$:
    $$n_i = \arg \max_n |B_n^T r_{i-1}|$$
  ✓ Better atom from $A$
    $$|A_{\ell_i}^T r_{i-1}| \geq |B_n^T r_{i-1}|$$
  ✓ Residual update
    $$r_i = r_{i-1} - (A_{\ell_i}^T r_{i-1}) A_{\ell_i}$$

• Divergence!
  $$\exists c > 0, \forall i, \|r_i\|_2 \geq c$$
Stagewise greedy algorithms

- **Principle**
  - select *multiple* atoms at a time to accelerate the process
  - possibly *prune out* some atoms at each stage

- **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]
Overview of greedy algorithms

\[ b = Ax_i + r_i \quad \quad \quad A = [A_1, \ldots, A_N] \]

<table>
<thead>
<tr>
<th></th>
<th>Matching Pursuit</th>
<th>OMP</th>
<th>Stagewise</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selection</strong></td>
<td>( \Gamma_i := \arg \max_n</td>
<td>A_n^T r_{i-1}</td>
<td>)</td>
</tr>
<tr>
<td><strong>Update</strong></td>
<td>( \Lambda_i = \Lambda_{i-1} \cup \Gamma_i )</td>
<td>( \Lambda_i = \Lambda_{i-1} \cup \Gamma_i )</td>
<td>( \Lambda_i = \Lambda_{i-1} \cup \Gamma_i )</td>
</tr>
<tr>
<td></td>
<td>( x_i = x_{i-1} + A_{\Gamma_i}^+ r_{i-1} )</td>
<td>( x_i = A_{\Lambda_i}^+ b )</td>
<td>( x_i = A_{\Lambda_i}^+ b )</td>
</tr>
<tr>
<td></td>
<td>( r_i = r_{i-1} - A_{\Gamma_i} A_{\Gamma_i}^+ r_{i-1} )</td>
<td>( r_i = b - A_{\Lambda_i} x_i )</td>
<td>( r_i = b - A_{\Lambda_i} x_i )</td>
</tr>
</tbody>
</table>

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

Global optimization

Iterative greedy algorithms

<table>
<thead>
<tr>
<th>Principle</th>
<th>$\min_x \frac{1}{2} |Ax - b|_2^2 + \lambda |x|^p_p$</th>
<th>Iterative decomposition $r_i = b - Ax_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tuning quality/sparsity</td>
<td>regularization parameter $\lambda$</td>
<td>stopping criterion (nb of iterations, error level, ...)$|x_i|_0 \geq k \quad |r_i| \leq \varepsilon$</td>
</tr>
</tbody>
</table>
| Variants | • choice of sparsity measure $p$
• optimization algorithm
• initialization | • selection criterion (weak, stagewise ...)
• update strategy (orthogonal ...)|
Overview

Convex & nonconvex optimization principles
Convex & nonconvex optimization algorithms
Greedy algorithms
Comparison of complexities
Complexity of IST

• Notation: $O(A)$ cost of applying $A$ or $A^T$

• Iterative Thresholding
  
  $f(x) = \Theta_{\alpha \lambda}^p (x + \alpha A^T (b - A x))$

  ✓ cost per iteration $= O(A)$

  ✓ when $A$ invertible, linear convergence at rate

  $\|x^{(i)} - x^*\|_2 \lesssim C \beta^i \|x^*\|_2$

  $\beta \leq 1 - \frac{\sigma^2_{\min}}{\sigma^2_{\max}}$

  ✓ 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
  \[ \|r_i\|_2 \leq \epsilon, \|x_i\|_0 \geq k \]

- Cost of first iteration = atom selection
  \[ O(A) \]
  (computation of all inner products)

- Naive cost of subsequent iterations =
  \[ O(A) \]

- If “local” structure of dictionary \([Krstulovic \text{ & al, MPTK}]\)
  ✓ subsequent iterations only cost
  \[ O(\log N) \]

<table>
<thead>
<tr>
<th></th>
<th>Generic ( A )</th>
<th>Local ( A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( k ) iterations</td>
<td>( O(kA) \geq O(km) )</td>
<td>( O(A + k \log N) )</td>
</tr>
<tr>
<td>( k \propto m )</td>
<td>( O(m^2) )</td>
<td>( O(m \log N) )</td>
</tr>
</tbody>
</table>
Complexity of OMP

- Number of iterations depends on stopping criterion
  \[ \|r_i\|_2 \leq \epsilon, \|x_i\|_0 \geq k \]

- Naive cost of iteration \(i\)
  - atom selection \(O(A)\) + orthoprojection \(O(i^3)\)

- With iterative matrix inversion lemma
  - atom selection \(O(A)\) + coefficient update \(O(i^2)\)

- If “local” structure of dictionary [Mailhé & al, LocOMP]
  ✓ subsequent approximate iterations only cost \(O(\log N)\)

<table>
<thead>
<tr>
<th></th>
<th>Generic (A)</th>
<th>Local (A)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(k) iterations</td>
<td>(O(kA + k^3))</td>
<td>(O(A + k \log N))</td>
</tr>
<tr>
<td>(k \propto m)</td>
<td>(O(m^3))</td>
<td>(O(m \log N))</td>
</tr>
</tbody>
</table>
LoCOMP

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

Fig. 1. SNR depending on the number of iterations

\[ N = 5 \cdot 10^5 \text{ samples, } k = 20,000 \text{ iterations} \]

<table>
<thead>
<tr>
<th>Iteration</th>
<th>MP</th>
<th>LocOMP</th>
<th>GP</th>
<th>OMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>First ((i = 0))</td>
<td>3.4</td>
<td>3.4</td>
<td>3.4</td>
<td>3.5</td>
</tr>
<tr>
<td>Begin ((i \approx 1))</td>
<td>0.028</td>
<td>0.033</td>
<td>3.4</td>
<td>3.4</td>
</tr>
<tr>
<td>End ((i \approx I))</td>
<td>0.028</td>
<td>0.050</td>
<td>40.5</td>
<td>41.0</td>
</tr>
<tr>
<td>Total time</td>
<td>571</td>
<td>854</td>
<td>4.50 \cdot 10^5</td>
<td>4.52 \cdot 10^5</td>
</tr>
</tbody>
</table>

• 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/](http://www.l1-magic.org/) (Candès, Romberg et al.), cvx, ...
  - Iterative thresholding
    ➡ [http://www.morphologicaldiversity.org/](http://www.morphologicaldiversity.org/) (Starck et al.), FISTA, NESTA, ...
  - Matching Pursuits
    ➡ sparsify (Blumensath), GPSR, ...

- **SMALLbox ()**: unified API for Matlab toolboxes

- **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](http://mptk.irisa.fr)
Performance of Sparse Decomposition Algorithms with Deterministic versus Random Dictionaries

Rémi Gribonval, DR INRIA
EPI METISS (Speech and Audio Processing)
INRIA Rennes - Bretagne Atlantique

remi.gribonval@inria.fr
http://www.irisa.fr/metiss/members/remi/talks
Structure of the course

- **Session 1: Panorama**
  - role of sparsity for compression, inverse problems, and learning
  - introduction to compressed (random) sensing

- **Session 2: Algorithms**
  - Review of main algorithms & complexities

- **Session 3: Guarantees for Deterministic & Random dictionaries**
  - compared success guarantees for different algorithms
  - robust guarantees & Restricted Isometry Property
  - explicit guarantees for various inverse problems
Inverse problems

Signal space $\sim \mathbb{R}^N$

Set of signals of interest

Nonlinear Approximation = Sparse recovery

Linear projection

Observation space $\sim \mathbb{R}^M$

$M \ll N$

Courtesy: M. Davies, U. Edinburgh
Exact recovery conditions for $L^p$
Proved Equivalence between L0 and L1

• “Empty” theorem: assume that $b = Ax_0$

  - if $\|x_0\|_0 \leq k_0(A)$ then $x_0 = x_0^*$
  - if $\|x_0\|_0 \leq k_1(A)$ then $x_0 = x_1^*$

• Content = estimation of $k_0(A)$ and $k_1(A)$
  - 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 $x_p^*, 0 \leq p \leq 1$?
General sparsity measures

- **Lp-norms**
  \[ \|x\|_p^p := \sum_k |x_k|^p, \quad 0 \leq p \leq 1 \]

- **f-norms!**
  \[ \|x\|_f := \sum_k f(|x_k|) \]

- **Constrained minimization**
  \[ x_f^* = x_f^*(b, A) \in \arg \min_x \|x\|_f \quad \text{subject to} \quad b = Ax \]

When do we have \( x_f^*(Ax_0, A) = x_0 \)?
Null space

- Null space = kernel

\[ z \in \mathcal{N}(A) \iff A z = 0 \]

- Particular solution vs general solution
  - Particular solution
    \[ A x = b \]
  - General solution
    \[ A x' = b \iff x' - x \in \mathcal{N}(A) \]
Recoverable supports: the “Null Space Property” (1)

- **Theorem 1** \([\text{Donoho & Huo 2001 for } L_1, \text{ G. & Nielsen 2003 for } L_p \& \text{ more}]\)
  - Assumption 1: sub-additivity (for quasi-triangle inequality)
    \[
f(a + b) \leq f(a) + f(b), \forall a, b
    \]
  - Assumption 2: «Null Space Property»

**NSP**

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

- Conclusion: \(x^*_f\) 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 \(I\)
NSP is necessary

\begin{itemize}
  \item Notations
    \begin{itemize}
    \item index set \( I \)
    \item vector \( z \)
    \item restriction \( z_I = (z_i)_{i \in I} \)
    \end{itemize}
  \item Assume there exists \( z \in \mathcal{N}(A) \) with
    \[ \|z_I\|_f > \|z_{I^c}\|_f \]
  \item Define \( b := Az_I = A(-z_{I^c}) \)
  \item The vector \( z_I \) is supported in \( I \) but is not the minimum norm representation of \( b \)
\end{itemize}
NSP is sufficient

- Assume quasi-triangle inequality
  \[ \forall x, y \| x + y \|_f \leq \| x \|_f + \| y \|_f \]

- Consider \( x \) with support set \( I \) and \( x' \) with \( Ax' = Ax \)

- Denote \( z := x' - x \in \mathcal{N}(A) \) and observe
  \[
  \| x' \|_f = \| x + z \|_f = \| (x + z)_I \|_f + \| (x + z)_{I^c} \|_f \\
  = \| x + z_I \|_f + \| z_{I^c} \|_f \\
  \geq \| x \|_f - \| z_I \|_f + \| z_{I^c} \|_f
  \]

- Conclude:
  If \( \| z_{I^c} \|_f > \| z_I \|_f \) when \( z \in \mathcal{N}(A) \) then \( I \) is recoverable
From “recoverable” supports to “sparse” vectors

\[ \emptyset \subseteq \{1\} \subseteq \{2\} \subseteq \{1, 2\} \subseteq \ldots \subseteq \{1, N\} \subseteq \ldots \]
From “recoverable” supports to “sparse” vectors

Trellis of supports

Recoverable supports are nested

NSP(I)

“Bad supports”

[1, N]
From “recoverable” supports to “sparse” vectors

Trellis of supports

Recoverable supports are nested

NSP(I)

#I = ||x||_0
From “recoverable” supports to “sparse” vectors

Recoverable supports are nested

NSP(I)

Sufficiently sparse, guaranteed recovery

At least one failing support

Trellis of supports

Recoverable supports

“Bad supports”
Recoverable sparsity levels: the “Null Space Property” (2)

- **Corollary 1** [Donoho & Huo 2001 for L1, G. Nielsen 2003 for Lp]
  - Definition: $I_k = \text{index of } k \text{ largest components of } z$
  - Assumption:

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

- Conclusion: $x_f^* \text{ recovers every } x \text{ with } \| x \|_0 \leq k$
- 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

$$\| z I_k \|_f < \| z I^c_k \|_f$$

Null space vectors must be “flat”, not sparse
NSP for L0: Identifiability of sparse representations

- Case of L0: **identifiability**
  - L0 min = **guaranteed unique sparsest solution** if
    - elements in null space have at least 2k+1 nonzeros
    - equivalently: every 2k columns are linearly independent
  - the mass of the largest k-terms should not exceed that of the tail
    \[
    k^* = \| \tilde{z}_I \|_0 < \| \tilde{z}_{I^C} \|_0
    \]
Stability and robustness
Need for stable recovery

Exactly sparse data  Real data (from source separation)
Formalization of stability

• Toy problem: exact recovery from \( \mathbf{b} = \mathbf{A} \mathbf{x} \)
  ✓ Assume sufficient sparsity \( \| \mathbf{x} \|_0 \leq k_p(\mathbf{A}) < m \)
  ✓ Wish to obtain \( x^*_p(\mathbf{b}) = \mathbf{x} \)

• Need to relax sparsity assumption
  ✓ New benchmark = best k-term approximation
  \[
  \sigma_k(\mathbf{x}) = \inf_{\| \mathbf{y} \|_0 \leq k} \| \mathbf{x} - \mathbf{y} \| \\
  \]
  ✓ Goal = stable recovery = instance optimality

\[
\| x^*_p(\mathbf{b}) - \mathbf{x} \| \leq C \cdot \sigma_k(\mathbf{x})
\]

[Cohen, Dahmen & DeVore 2006]
Stability for Lp minimization

• Assumption: «stable Null Space Property»

$\text{NSP}(k, \ell_p^p, t)$

\[ \| z_{I_k^C} \|_p^p \leq t \cdot \| z_{I_k^C} \|_p^p \] when $z \in \mathcal{N}(A)$, $z \neq 0$

• Conclusion: instance optimality for all $x$

\[ \| x^*_p(b) - x \|_p^p \leq C(t) \cdot \sigma_k(x)_p^p \]

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

[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

$$\|z_{I_k}\|_p^p \leq t \cdot \|z_{I_k^c}\|_p^p$$

Null space vectors must be “flat”, not sparse
Robustness

• Toy model = noiseless
  \[ \mathbf{b} = \mathbf{A}\mathbf{x} \]

• Need to account for noise
  ✓ measurement noise
  ✓ modeling error
  ✓ numerical inaccuracies ...

• Goal: predict robust estimation
  \[ \| x_p^* (\mathbf{b}) - \mathbf{x} \| \leq C \| e \| + C' \sigma_k (\mathbf{x}) \]

• Tool: restricted isometry property
Restricted Isometry Property

- **Definition**
  - $A$ with $N$ columns

- **Computation**
  - Naively: combinatorial
  - Open question: NP? NP-complete?
Stability & robustness from RIP

RIP\((k, \delta)\)

\[ \delta_{2k}(A) \leq \delta \]

[Candès 2008]

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

NSP\((k, \ell^1,t)\)

\[ \|z_{I_k}\|_{1} \leq t \cdot \|z_{I^c_k}\|_{1} \quad \text{when} \quad z \in \mathcal{N}(A), \ z \neq 0 \]

- Result: **stable + robust** $L^1$-recovery under assumption that
  \[ \delta_{2k}(A) < \sqrt{2} - 1 \approx 0.414 \]

  ✓ Foucart-Lai 2008: $L^p$ with $p<1$, and $\delta_{2k}(A) < 0.4531$
  ✓ Chartrand 2007, Saab & Yilmaz 2008: other RIP condition for $p<1$
  ✓ G., Figueras & Vandergheynst 2006: robustness with $f$-norms
  ✓ Needell & Tropp 2009, Blumensath & Davies 2009: RIP for greedy algorithms
Is the RIP a sharp condition?

- The Null Space Property
  ✓ “algebraic” + sharp property for $L^p$, only depends on $\mathcal{N}(A)$

[Davies & Gribonval, IEEE Inf. Th. 2009]
Is the RIP a sharp condition?

- The Null Space Property
  - “algebraic” + sharp property for $L^p$, only depends on $\mathcal{N}(A)$
  - invariant by linear transforms $A \rightarrow BA$

[Davies & Gribonval, IEEE Inf. Th. 2009]
Is the RIP a sharp condition?

- The Null Space Property
  ✓ “algebraic” + sharp property for $L^p$, only depends on $\mathcal{N}(A)$
  ✓ invariant by linear transforms $A \rightarrow BA$

- The RIP($k$, $\delta$) condition

[Davies & Gribonval, IEEE Inf. Th. 2009]
Is the RIP a sharp condition?

- The Null Space Property
  ✓ “algebraic” + sharp property for Lp, only depends on $\mathcal{N}(A)$
  ✓ invariant by linear transforms $A \rightarrow BA$

- The RIP($k$, $\delta$) condition
  ✓ “metric” ... and not invariant by linear transforms

[Davies & Gribonval, IEEE Inf. Th. 2009]
Is the RIP a sharp condition?

- The Null Space Property
  ✓ “algebraic” + sharp property for $L_p$, only depends on $\mathcal{N}(A)$
  ✓ invariant by linear transforms $A \rightarrow BA$

- The RIP($k, \delta$) condition
  ✓ “metric” ... and not invariant by linear transforms
  ✓ predicts performance + robustness of several algorithms

[Davies & Gribonval, IEEE Inf. Th. 2009]
Comparison between algorithms

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

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

• 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

Proof

[Gribonval & Nielsen, ACHA 2007]
Remaining agenda

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

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

• Question
  ✓ what is the order of magnitude of these numbers ?
  ✓ how do we estimate them in practice ?

• A first element:
  ✓ if $A$ is $m \times N$, then $k_0(A) \leq \lfloor m/2 \rfloor$
  ✓ 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 $\mathbf{A}$
  ✓ Dictionary modeling structures of signals
    ✦ Constrained choice = to fit the data.
    ✦ Ex: union of wavelets + curvelets + spikes
  ✓ «Transfer function» from physics of inverse problem
    ✦ Constrained choice = to fit the direct problem.
    ✦ Ex: convolution operator / transmission channel
  ✓ 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
Multiscale Time-Frequency Structures

- Audio = superimposition of structures

\[ b = \{ b(t) \}_t \quad \text{and} \quad x = \{ x(s, \tau, f) \}_{s, \tau, f} \]

- Transients = short, small scale
- Harmonic part = long, large scale

- Gabor atoms

\[ g_{s, \tau, f}(t) := \frac{1}{\sqrt{s}} w \left( \frac{t - \tau}{s} \right) e^{2i\pi ft} \]

- Dictionary matrix:

\[ A_n = \{ g_{s_n, \tau_n, f_n}(t) \}_t \]

\[ A = [A_1 \ldots A_N] \]
Multiscale Time-Frequency Structures

• Audio = superimposition of structures

\[ b = \{ b(t) \}_t \quad \text{and} \quad x = \{ x(s, \tau, f) \}_{s,\tau,f} \]

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

• Gabor atoms

\[ g_{s,\tau,f}(t) := \frac{1}{\sqrt{s}} w \left( \frac{t - \tau}{s} \right) e^{2i\pi ft} \]

• Dictionary matrix:

\[ A_n = \{ g_{s_n,\tau_n,f_n}(t) \}_t \]

\[ A = [A_1 \ldots A_N] \]
Deterministic matrices and coherence

- **Lemma**
  - Assume normalized columns $\|A_i\|_2 = 1$
  - Define coherence $\mu = \max_{i \neq j} |A_i^T A_j|$
  - Consider index set $I$ of size $\|I\| \leq k$
  - Then for any coefficient vector $c \in \mathbb{R}^I$
    - $1 - (k - 1)\mu \leq \frac{\|A_I c\|_2^2}{\|c\|_2^2} \leq 1 + (k - 1)\mu$
  - In other words
    - $\delta_{2k} \leq (2k - 1)\mu$

$mardi 13 septembre 2011$
Coherence vs RIP

- **Deterministic** matrix, such as Dirac-Fourier dictionary
- **Coherence**
  \[ \delta_n(t) = \frac{1}{\sqrt{m}} e^{2i\pi nt/m} \]
  \[ \mu = 1/\sqrt{m} \]

- **“Generic”** (random) dictionary
  [Candès & al 2004, Vershynin 2006, ...]

- **Isometry constants**
  if \( m \geq Ck \log N/k \)
  then \( P(\delta_{2k} < \sqrt{2} - 1) \approx 1 \)

**Recovery regimes**

- \( k_1(A) \approx 0.914 \sqrt{m} \)
- \( k_{*\text{MP}}(A) \geq 0.5 \sqrt{m} \)
- \( k_1(A) \approx \frac{m}{2e \log N/m} \) with high probability for Gaussian \( A \)

[Elad & Bruckstein 2002] [Donoho & Tanner 2009]
Example: convolution operator

- Deconvolution problem with spikes
  \[ b = h \star x + e \]
  ✓ Matrix-vector form \( b = A x + e \) with \( A = \text{Toeplitz or circulant matrix} \ [A_1, \ldots, A_N] \)
  \[ A_n(i) = h(i - n) \] by convention \( \|A_n\|_2^2 = \sum_i h(i)^2 = 1 \)
  ✓ Coherence = autocorrelation, can be large
    \[ \mu = \max_{n \neq n'} A_n^T A_{n'} = \max_{\ell \neq 0} h \star \tilde{h}(\ell) \]
  ✓ Recovery guarantees
    ✦ Worst case = close spikes, usually difficult and not robust
    ✦ Stronger guarantees assuming distance between spikes [Dossal 2005]
  ✓ Algorithms: exploit fast convolution to apply \( A \) and adjoint.
Example: image inpainting

Wavelets

\[ y = \Phi x \]

Image

Inpainting

Result

\[ b = My = M\Phi x \]

\[ A = M\Phi \]
Compressed sensing

• Approach = acquire some data $y$ with a limited number $m$ of (linear) measures, modeled by a measurement matrix $M : b \approx My$

• Key hypotheses
  ✓ Sparse model: the data can be sparsely represented in a known dictionary $\Phi : y \approx \Phi x$, with $\sigma_k(x) \ll \|x\|$
  ✓ Overall matrix $A = M\Phi$ is «incoherent»: $\delta_{2k}(A) \ll 1$

leading to robust + stable sparse recovery

• Reconstruction = sparse recovery algorithm
Compressed Sensing: key requirements

- **Sparse model** = dictionary $\Phi$
  - ✓ need to «fit» the data
    - ✦ does not always exist: e.g. white Gaussian noise cannot be sparsified!
  - ✓ dictionary design:
    - ✦ expert knowledge
    - ✦ dictionary selection from a library (wavelets, curvelets, Gabor, ...)
    - ✦ dictionary learning

- **Measurement matrix** $\mathcal{M}$
  - ✓ physically feasible: hardware implementation!
  - ✓ recovery guarantees: incoherence of $\mathcal{M}\Phi$

- **Efficiency**:
  - ✓ fast computation of $\mathcal{M}\Phi y, (\mathcal{M}\Phi)^T b$
Compressed Sensing: when is it worth it?

- **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]?
    - ✦ Constraints on data flow
  - ✓ Transmission is lossy
    (CS=robust to loss of a few measures)
Excessive pessimism ?
Recovery analysis

- Recoverable set for a given “inversion” algorithm
- Level sets of L0-norm
- Worst case = too pessimistic!

\[ b = A x \]
Recovery analysis

\[ b = Ax \]

- Recoverable set for a given “inversion” algorithm
- Level sets of L0-norm
- Worst case = too pessimistic!
- Finer “structures” of x

Borup, G. & Nielsen ACHA 2008, \( A = \text{Wavelets U Gabor,} \)
recovery of infinite supports for analog signals
\[ \text{support}(x), \text{sign}(x) \]
Fuchs 2005; Zhao & Yu 2006; Zou 2006; Yuan & Lin 2007;
Wainwright 2009;
L1 recovery beyond $k_1(A)$
Recovery analysis

- Recoverable set for a given “inversion” algorithm
- Level sets of L0-norm
- Worst case = too pessimistic!
- Finer “structures” of x
  - Borup, G. & Nielsen ACHA 2008, $A =$ Wavelets U Gabor, recovery of infinite supports for analog signals
  - support$(x)$, sign$(x)$
  - Fuchs 2005; Zhao & Yu 2006; Zou 2006; Yuan & Lin 2007; Wainwright 2009;
- Average/typical case
  - G., Rauhut, Schnass & Vandergheynst, JFAA 2008, “Atoms of all channels, unite! Average case analysis of multichannel sparse recovery using greedy algorithms”.

\[ b = Ax \]

\[ \|x\|_0 \leq k \]

\[ \{ x = \text{AlgoA}(Ax) \} \]
Average case analysis?

Typical observation

\[ P(x^* = x_0) \]

\[ x_0 \xrightarrow{\text{direct model}} b := Ax_0 \]

Inverse problem

\[ x_p^* = \arg \min_{Ax = Ax_0} \| x \|_p \]

\[ \| x_0 \|_0 \]
Average case analysis?

\[ x_0 \rightarrow \text{direct model} \rightarrow b := A x_0 \]

Typical observation

\[ P(x^* = x_0) \]

\[ x^*_p = \arg \min_{A x = Ax_0} \| x \|_p \]

Dossal, Peyré, Fadili 2009: heuristic algorithm to find worst-case
Average case analysis?

Typical observation

\[ P(x^* = x_0) \]

\[ P = 1 - \epsilon, \epsilon \ll 1 \]

Dossal, Peyré, Fadili 2009: heuristic algorithm to find worst-case

\[ x_p^* = \arg \min_{Ax = Ax_0} \|x\|_p \]

\[ b := Ax_0 \]

direct model

inverse problem

\[ \|x_0\|_0 \]

\[ k_0(A), k_{1/2}(A), k_1(A) \]
Average case analysis?

Typical observation

\[ P(x^* = x_0) \]

\[ P = 1 - \epsilon, \epsilon \ll 1 \]

Dossal, Peyré, Fadili 2009: heuristic algorithm to find worst-case
Phase transitions for Gaussian $\mathbf{A}$

With high probability, $L_1$ succeeds for

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

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

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

[Donoho & Tanner 2009]
Phase transitions for Gaussian $\mathbf{A}$

$$k/m$$

With high probability, $L_1$ fails with high probability

With high probability, $L_1$ succeeds for most $x$

With high probability, $L_1$ succeeds for all $x$

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

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

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

[Donoho & Tanner 2009]
Conclusions

• Sparsity: prior to solve **ill-posed inverse problems**
• If solution sufficiently sparse, **reasonable algorithms are guaranteed to find it** (even one step thresholding!).
• **Computational efficiency still a challenge**
  ✓ problem sizes up to 1000 x 10000 already 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 tutorial

- Structured sparsity: group LASSO, etc.
- Combinatorial algorithms: submodular functions, etc.
- Approximate Message Passing algorithms
- Analysis vs synthesis sparsity
- Dictionary learning
- Low-rank matrices & sparsity
Thanks to

- F. Bimbot, S. Krstulovic, A. Ozerov, S. Lesage, B. Mailhé, S. Arberet, P. Sudhakar
- M. Nielsen, L. Borup (Aalborg Univ.)
- P. Vandergheynst, R. Figueras, P. Jost, K. Schnass (EPFL)
- M. Davies (U. Edinburgh), M. Elad (Technion), M. Plumbley (QMUL)
- H. Rauhut (U. Vienna)
- and several other collaborators...
The end

remi.gribonval@inria.fr
www.irisa.fr/metiss/gribonval
Lp “norms” level sets

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

Observation: the minimizer is sparse

\[ \{ x \text{ s.t. } b = Ax \} \]
Sparsity of L1 minimizers

• Real-valued case
  ✓ \( A = \) an \( m \times N \) real-valued matrix
  ✓ \( b = \) an \( m \)-dimensional real-valued vector
  ✓ \( X = \) set of all minimum L1 norm solutions to \( A x = b \)

\[ \tilde{x} \in X \iff \|\tilde{x}\|_1 = \min \|x\|_1 \text{ s.t. } Ax = b \]

• Fact 1: \( X \) is convex and contains a “sparse” solution

\[ \exists x_0 \in X, \|x_0\|_0 \leq m \]

• Proof: exercice!
Sparsity of L1 minimizers

• Real-valued case
  ✓ $A$ = an $m \times N$ real-valued matrix
  ✓ $b$ = an $m$-dimensional real-valued vector
  ✓ $X$ = set of all solutions to regularization problem

$$\mathcal{L}(x) := \frac{1}{2} \|Ax - b\|_2^2 + \lambda \|x\|_1$$

$$\tilde{x} \in X \iff \mathcal{L}(\tilde{x}) = \min_x \mathcal{L}(x)$$

• Fact 2: $X$ is a convex set and contains a “sparse” solution

$$\exists x_0 \in X, \|x_0\|_0 \leq m$$

• Proof : exercice, using Fact 1!
Sparsity of L1 minimizers

• A word of caution: this **does not hold true in the complex-valued case**

• Counter example: there is a construction where
  ✓ $A = a \times 3$ complex-valued matrix
  ✓ $b = a$ 2-dimensional complex-valued vector
  ✓ the minimum L1 norm solution is unique and has 3 nonzero components

L1 vs Lp
Lp better than L1 (1)

• **Theorem 2** [G. Nielsen 2003]
  ✓ Assumption 1: **sub-additivity** of sparsity measures $f, g$
  $$f(a + b) \leq f(a) + f(b), \forall a, b$$
  ✓ Assumption 2: the function $t \mapsto \frac{f(t)}{g(t)}$ is non-increasing
  ✓ Conclusion: $k_g(A) \leq k_f(A), \forall A$

  *Minimizing $\|x\|_f$ can recover vectors which are less sparse than required for guaranteed success when minimizing $\|x\|_g$*
Lp better than L1 (2)

• **Example**
  - sparsity measures
    \[ f(t) = t^p, \quad g(t) = t^q, \quad 0 \leq p \leq q \leq 1 \]
  - sub-additivity
    \[ |a + b|^p \leq |a|^p + |b|^p, \quad \forall a, b, \quad 0 \leq p \leq 1 \]
  - function
    \[ \frac{f(t)}{g(t)} = t^{p-q} \text{ is non-increasing} \]
  - therefore
    \[ k_1(A) \leq k_q(A) \leq k_p(A) \leq k_0(A), \quad \forall A \]
Lp better than L1: proof

1) Since $f/g$ non-decreasing:
$$z_1 \geq z_2 \geq 0$$

2) Similarly
$$z_1 \geq \ldots \geq z_N \geq 0$$

3) Conclusion: if NSP$(g,t,k)$ then NSP$(f,t,k)$
Lp better than L1 (3)

- At sparsity levels where L1 is guaranteed to “succeeds”, all Lp p\leq 1 is also guaranteed to succeed

\[ P(x^* = x_0) \]

\[ x_p^* = \arg \min_{Ax = Ax_0} \|x\|_p \]

Highly sparse representations are independent of the (admissible) sparsity measure
Lp better than L1 \((4)\)

- Lp \(p<1\) can succeed where L1 fails
  - How much improvement? Quantify \(k_p(\mathbf{A})\)?
- Lp \(p<1\): nonconvex, has many local minima
  - Better recovery with Lp *principle*, what about *algorithms*?

\[
P(x^* \equiv x_0)
\]

\[
x_p^* = \text{arg} \min_{\mathbf{A}x = \mathbf{A}x_0} \|x\|_p
\]

Lp better than L1: compressed sensing with fewer measurements?
When does $\delta_{2k}(A) < \delta$ imply $k \leq k_p(A)$?
When does $\delta_{2k}(A) < \delta$ imply $k \leq \kappa_p(A)$?
When does $\delta_{2k}(A) < \delta$ imply $k \leq k_p(A)?$
When does $\delta_{2k}(A) < \delta$ imply $k \leq k_p(A)$?
When does $\delta_{2k}(A) < \delta$ imply $k \leq k_p(A)$?
When does $\delta_{2k}(A) < \delta$ imply $k \leq k_p(A)$?
When does $\delta_{2k}(A) < \delta$ imply $k \leq k_p(A)$?
When does $\delta_{2k}(A) < \delta$ imply $k \leq k_p(A)$?
L1 minimization for Laplacian data?
Bayesian modeling

- Observation: \( b = Ax + n \)
- Prior model: \( \hat{P}(x_k) \propto \exp(-f(|x_k|)) \)
- Probability vs «energy»
  \[
  \max_x \prod_k P(x_k) \iff \min_x \sum_k f(|x_k|)
  \]
- L1 minimization «equivalent to MAP with Laplacian model»
  \( \hat{P}(x_k) \propto \exp(-|x_k|) \)
Experiment: L1 minimization for Laplacian data ...

- Gaussian matrix

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

\[ N = 128 \quad 1 \leq m \leq 100 \]

- Laplacian data, 500 draws

\[ x \in \mathbb{R}^{N} \quad \rightarrow \quad \mathbf{b} = \mathbf{A}x \]

- Reconstruction L1 or L2

\[ x_p^* := \arg \min \|x\|_p, \quad p = 1, 2 \]

= ML with Laplacian / Gaussian prior

\[ \mathbb{E}\|x_p^* - x\|_2^2 \]

MAP is bad when the model fits the data!

Mikolova 2007, Inverse Problems and Imaging

cf also Seeger and Nickish, ICML 2008

\[ m \]

\[ m \leq 100 \]
When to rely on sparse methods?

- **Laplacian distribution:**
  \[ P(x_i) \propto \exp(-\lambda |x_i|)^{-1} \]
  - ✓ peak at the origin
  - ✓ not heavy tailed
  - ✓ \( x \) not well approximated by sparse vector

- **Cauchy distribution:**
  \[ P(x_i) \propto (1 + \lambda x_i^2)^{-1} \]
  - ✓ no peak at the origin
  - ✓ heavy tailed (large values)
  - ✓ \( x \) very well approximated by sparse vector

Sparse methods always provide sparse estimates
Lp “norms” level sets

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

Property: the minimizer is sparse

\[ \{ x \ \text{s.t.} \ b = Ax \} \]
Empirical observation: $L^p$ versus $L^1$

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

$$ x_0 \rightarrow b := Ax_0 \rightarrow x^*_p = \arg \min_{Ax = Ax_0} \| x \|_p $$

$$ P(x^* = x_0) $$

Reference: direct model
Inverse problem

Typical observation: $p=1/2$ vs $p=1$