Pursuit Algorithms for Sparse Approximation

Rémi Gribonval
METISS project-team (audio signal processing, speech recognition, source separation)
INRIA, Rennes, France

IRISA
UNE UNITÉ DE RECHERCHE À LA POINTE DES SCIENCES
ET DES TECHNOLOGIES DE L'INFORMATION
ET DE LA COMMUNICATION

Ecole d'été en Traitement du Signal
Peyresq, Juillet 2009

UNIVERSITÉ DE RENNES 1 INSA CNRS
INRIA RENNES

Overview

- Complexity of ideal sparse approximation
- Convex optimization
- Greedy algorithms
- Nonconvex optimization ?

Ideal sparse approximation

- Input:
  m x N matrix $\mathbf{A}$, with $m < N$, m-dimensional vector $\mathbf{b}$

- Possible objectives:
  find the sparsest approximation within tolerance

$$\arg\min_x \|x\|_0, \text{ s.t.} \|\mathbf{b} - \mathbf{A}x\| \leq \epsilon$$

  find best approximation with given sparsity

$$\arg\min_x \|\mathbf{b} - \mathbf{A}x\|, \text{ s.t.} \|x\|_0 \leq k$$

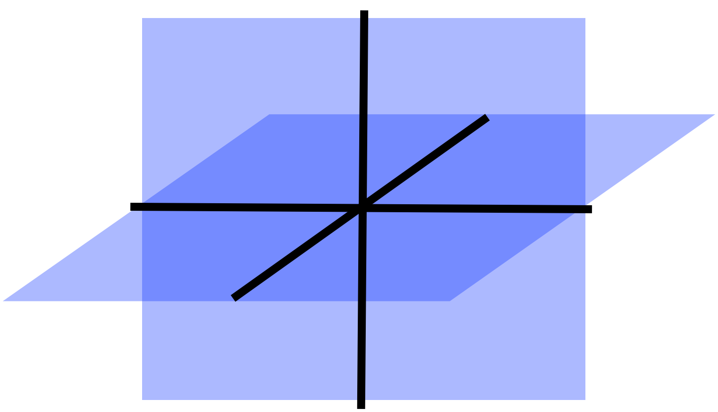
  find a solution x to

$$\|\mathbf{b} - \mathbf{A}x\| \leq \epsilon, \text{ and } \|x\|_0 \leq k$$

Geometric interpretation of sparse approximation

- Coefficient domain $\mathbb{R}^N$:
  - set Σ_k of sparse vectors

$\|x\|_0 \leq k$



$\binom{N}{k}$ subspaces

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

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

Combinatorial search!
Actual complexity ?

Complexity

Complexity

- **Polynomial algorithm**: given input of size *N*, compute output in cost *poly*(*N*)
- **Polynomial problem (is in P)**: there is a polynomial algorithm which can compute the solution to each instance of the problem
- **Example**:
  - problem: find the nearest neighbor to an *m*-dimensional vector from a collection of *N* such vectors
  - input size = $m \times (N+1)$
  - complexity = $O(Nm)$ [N distances in $\mathbb{R}^m$]

Complexity: NP

- **Decision problem:** of the type "does there exist *x* satisfying a given set of constraints"
- **Non-deterministic polynomial decision problems (in NP):** if there is a polynomial algorithm which can check for any instance of the problem if a candidate solution *x* satisfies the constraint.
    - **warning**: the algorithm is not required to *find* a solution. It merely has to *check* if a solution *x* (given by an "oracle") is acceptable.

IRISA

Complexity: NP-complete

- **Reduction**: every instance of Problem A can be transformed into an instance of Problem B in polynomial time *A "less complex" than B*
- **NP-hard problem**: Problem B such that *every* Problem A in NP can be reduced to B.
- **NP-complete problems:** NP-hard + in NP
- ***Fact***: there exists at least one NP-complete problem (satisfiability problem = SAT)

Complexity of sparse approximation

- **Step 1**: express it as a decision problem:
  - description of an instance

    m x *N* matrix **A,** *m*-dimensional vector **b**, parameters (ϵ, k)
  - size of an instance = approximately *mN*
  - decision problem: does there exists *x* such that

$$\|\mathbf{b} - \mathbf{A}x\| \leq \epsilon, \text{ and } \|x\|_0 \leq k$$

- **Step 2**: prove it is in NP. Indeed, one can check in polynomial time O(mN) whether a given *x* satisfies the constraints
- **Step 3**: reduce an existing problem to it to show it is NP-complete

IRISA

NP-completeness of sparse approximation

- Which known NP-complete problem?
  Exact-cover by 3-sets *[Davis & al 1997]*
  (other approach in [Natarajan 1995])
  - Description of an instance:
    - The integer interval $E = [\![1, 3k]\!]$
    - *A* collection of subsets of size 3

$$C = \{F_n, 1 \leq n \leq N\}, F_n \subset E, \sharp F_n = 3$$

  - Decision problem:
    - does there exist an exact cover (=disjoint partition) of *E* from elements of *C* ?

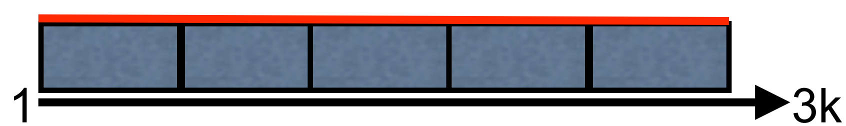
$$\exists ? \Lambda, \cup_{n \in \Lambda} F_n = E \qquad n \neq n' \in \Lambda \Rightarrow F_n \cap F_{n'} = \emptyset$$

IRISA

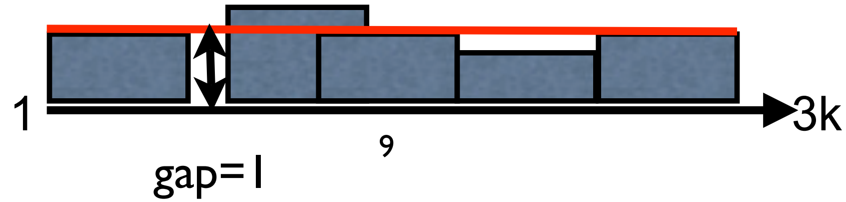
NP-completeness

- Reduction of 3-SETS to sparse approximation
  - $m=3k$
  - vector $\mathbf{b} = (b_i)_{i=1}^m \quad b_i = 1, \forall i$
  - matrix $\mathbf{A} = (a_{in})_{1\leq i\leq m, 1\leq n\leq N}$ $\qquad a_{in} = \begin{cases} 1, & i \in F_n \\ 0, & \text{otherwise} \end{cases}$
  - tolerance $\epsilon < 1$
- Exact cover implies existence of *x* such that

$$\|\mathbf{b} - \mathbf{A}x\| \leq \epsilon, \text{ and } \|x\|_0 \leq k$$



- Non-exact cover implies the opposite



Practical approaches: Optimization *principles*

Overall compromise

- Approximation quality

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

- Ideal sparsity measure : ℓ^0 "norm"

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

- "Relaxed" sparsity measures

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

Lp norms / quasi-norms

- **Norms** when $1 \leq p < \infty$ = convex

$$\|x\|_p = 0 \Leftrightarrow x = 0$$

$$\|\lambda x\|_p = |\lambda| \|x\|_p, \forall \lambda, x$$

Triangle inequality $$\|x + y\|_p \leq \|x\|_p + \|y\|_p, \forall x, y$$

- **Quasi-norms** when $0 < p < 1$ = nonconvex

Quasi-triangle inequality

$$\|x + y\|_p \leq 2^{1/p} \big(\|x\|_p + \|y\|_p \big), \forall x, y$$

$$\|x + y\|_p^p \leq \|x\|_p^p + \|y\|_p^p, \forall x, y$$

- "*Pseudo*"-norm for p=0

$$\|x + y\|_0 \leq \|x\|_0 + \|y\|_0, \forall x, y$$

Optimization problems

- Approximation

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

- Sparsification

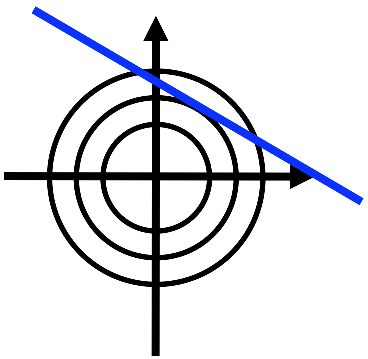
$$\min_x \|x\|_p \text{ s.t. } \|\mathbf{b} - \mathbf{A}x\|_2 \leq \epsilon$$

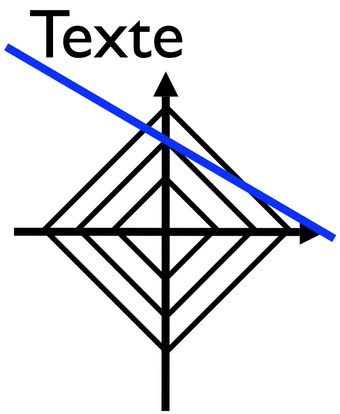
- Regularization

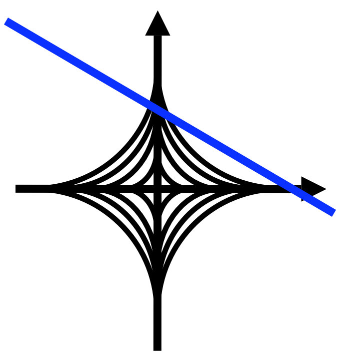
$$\min_x \frac{1}{2}\|\mathbf{b} - \mathbf{A}x\|_2 + \lambda\|x\|_p$$

L*p* “norms” level sets

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







Observation: *the minimizer is sparse*

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

Sparsity of L*1* minimizers

- Real-valued case
  - **A** = an *m* x *N* real-valued matrix
  - **b** = an *m*-dimensional real-valued vector
  - *X* = set of all minimum L*1* norm solutions to $\mathbf{A}x = \mathbf{b}$

$$\tilde{x} \in X \Leftrightarrow \|\tilde{x}\|_1 = \min \|x\|_1 \text{ s.t. } \mathbf{A}x = \mathbf{b}$$

- **Fact 1**: *X* is convex and contains a "sparse" solution

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

- Proof : exercice!

Sparsity of L*1* minimizers

- Real-valued case
  - **A** = an *m* x *N* real-valued matrix
  - **b** = an *m*-dimensional real-valued vector
  - *X* = set of al solutions to regularization problem

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

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

- **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!

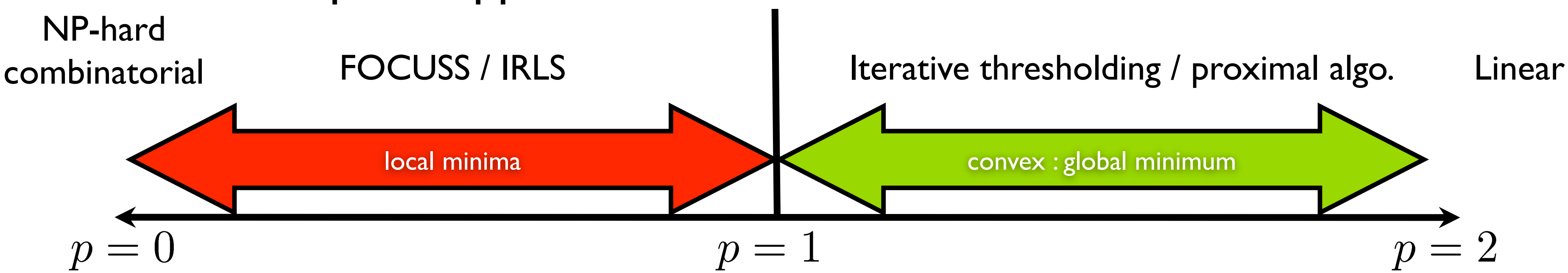
Sparsity of L*1* minimizers

- A word of caution: this **does not hold true in the complex-valued case**
- Counter example: there is a construction where
  - **A** = a 2 x *3* complex-valued matrix
  - **b** = a 2-dimensional complex-valued vector
  - the minimum L1 norm solution is unique and has 3 nonzero components

[E.Vincent, Complex Nonconvex Optimization l_p norm minimization for underdetermined source separation, Proc. ICA 2007.]

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



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, ...]

Algorithms for L1: Linear Programming

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

Basis Pursuit (BP)
LASSO

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

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

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

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

L1 regularization: Quadratic Programming

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

Basis Pursuit Denoising (BPDN)

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

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

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

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

Generic approaches vs specific algorithms

- There is a vast literature on linear / quadratic programming algorithms
- Can use linprog in Matlab
- 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

Optimization algorithms

Example: orthonormal **A**

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

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

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

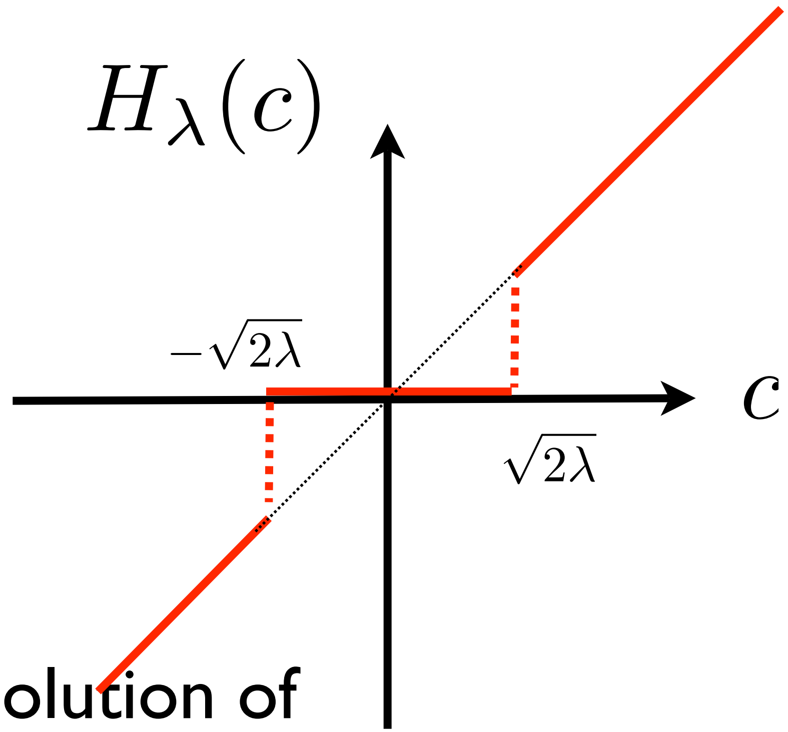
- Expression of BPDN criterion to be minimized

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

- 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)

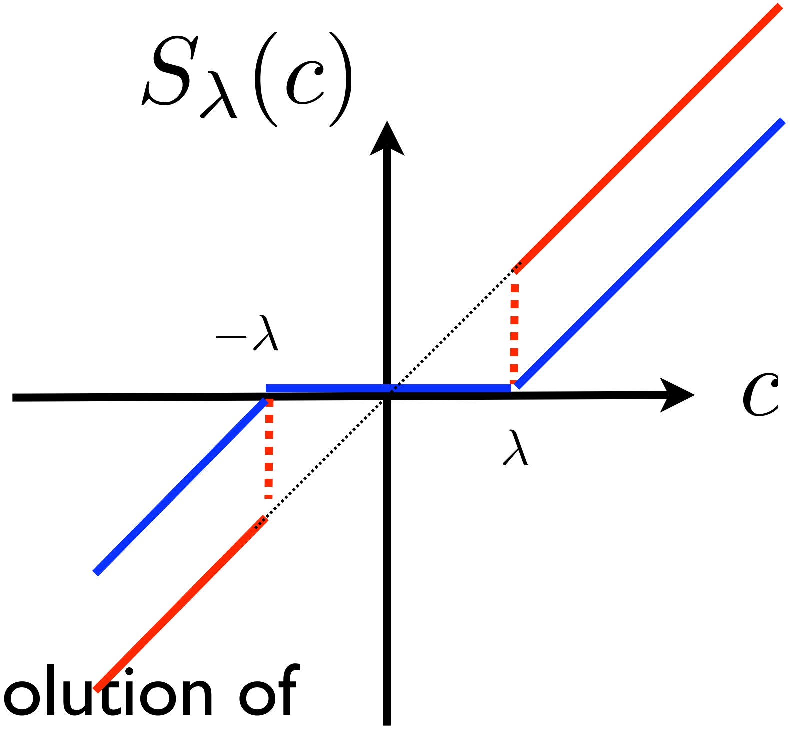


- Solution of

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

IRISA

Soft-thresholding (p=1)



- Solution of

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

Iterative thresholding

- Proximity operator

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

- Goal = compute

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

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

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

  - thresholding

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

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 $p \geq 1$

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

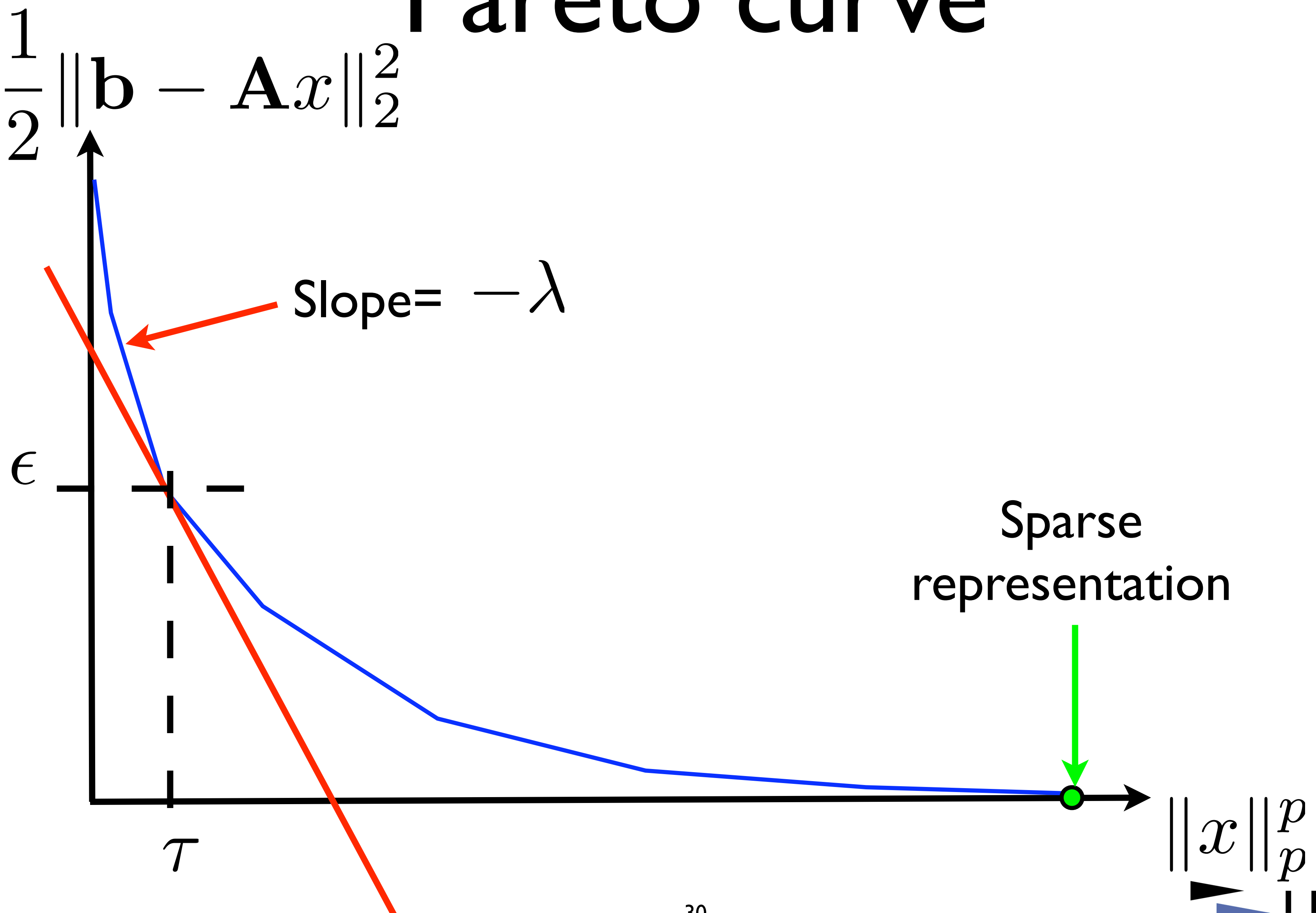
  - assume that $\forall x,\ \|\mathbf{A}x\|_2^2 \leq c\|x\|_2^2$ and $\alpha < 2/c$
  - then, the iterates converge strongly to a limit $x^\star$

$$\|x^{(i)} - x^\star\|_2 \to_{i\to\infty} 0$$

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

IRISA

Pareto curve



IRISA

Path of the solution

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

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

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

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

- In particular

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

Homotopy method

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

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

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

Greedy Algorithms

Greedy algorithms

- Observation: when **A** is orthormal,
  - the problem

$$\min_x \|\mathbf{b} - \mathbf{A}x\|_2^2 \text{ s.t. } \|x\|_0 \leq k$$

  - is equivalent to

$$\min_x \sum_n (\mathbf{A}_n^T \mathbf{b} - x_n)^2 \text{ s.t. } \|x\|_0 \leq k$$

- Let Λ_k index the *k* largest inner products

$$\min_{n \in \Lambda_k} |\mathbf{A}_n^T \mathbf{b}| \geq \max_{n \notin \Lambda_k} |\mathbf{A}_n^T \mathbf{b}|$$

  - an optimum solution is

$$x_n = \mathbf{A}_n^T \mathbf{b}, n \in \Lambda_k;\ \ x_n = 0, n \notin \Lambda_k$$

IRISA

Greedy algorithms

- Iterative algorithm (= *Matching Pursuit*)
  - Initialize a residual to $\mathbf{r}_0 = \mathbf{b}$ $\quad i = 1$
  - Compute all inner products

$$\mathbf{A}^T\mathbf{r}_{i-1} = (\mathbf{A}_n^T\mathbf{r}_{i-1})_{n=1}^N$$

  - Select the largest in magnitude

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

  - Compute an updated residual

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

  - If $i \geq k$ then stop, otherwise increment *i* and iterate

Dictionaries and atoms

- Convention on *m* x *N* matrix **A**
  - normalized columns: $\|\mathbf{A}_n\|_2 = 1, \forall n$
  - complete column span: $\mathrm{span}(\mathbf{A}_n, 1 \leq n \leq N) = \mathbb{R}^m$
  - in particular: $m \leq N$
- Vocabulary:
  - **A** is called a signal **dictionary**
  - columns are called **atoms**

Matching Pursuit (MP)

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

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

  - Residual update

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

- Energy preservation (Pythagoras theorem)

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

Main properties

- Global energy preservation

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

- Global reconstruction

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

- Strong convergence

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

Orthonormal MP (OMP)

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

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

- Best approximation from V_k = orthoprojection

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

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

OMP

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

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

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

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

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

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

IRISA

Weak Pursuits

- Sometimes the following optimization is too complex

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

- Weak selection : pick *any* atom such that

$$|\mathbf{A}_{n_i}^T \mathbf{r}_{i-1}| \geq t \sup_n |\mathbf{A}_n^T \mathbf{r}_{i-1}|$$

- Convergence is preserved [*Temlyakov*]

IRISA

Convergence rate

- Observation:
  - the quantity $\|\mathbf{r}\|_{\mathbf{A}} = \sup_n |\mathbf{A}_n^T \mathbf{r}|$ is a norm
  - by equivalence of all norms in finite dimension

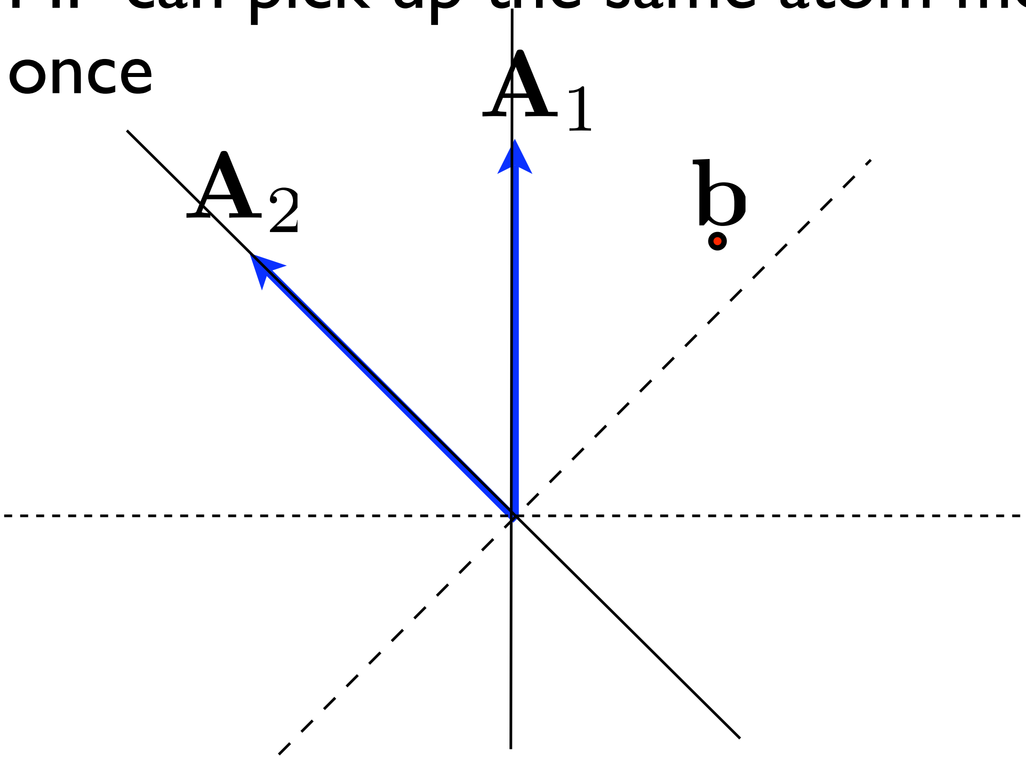
$$\exists c > 0, \forall \mathbf{r}, \|\mathbf{r}\|_{\mathbf{A}} \geq c\|\mathbf{r}\|_2$$

- At each iteration

$$\begin{aligned}\|\mathbf{r}_i\|_2^2 &\leq \|\mathbf{r}_{i-1}\|_2^2 - t^2\|\mathbf{r}_{i-1}\|_{\mathbf{A}}^2 \\ &\leq \|\mathbf{r}_{i-1}\|_2^2 - t^2c^2\|\mathbf{r}_{i-1}\|_2^2 \\ &\leq (1 - t^2c^2)^i\|\mathbf{r}_0\|_2^2\end{aligned}$$

Caveats (I)

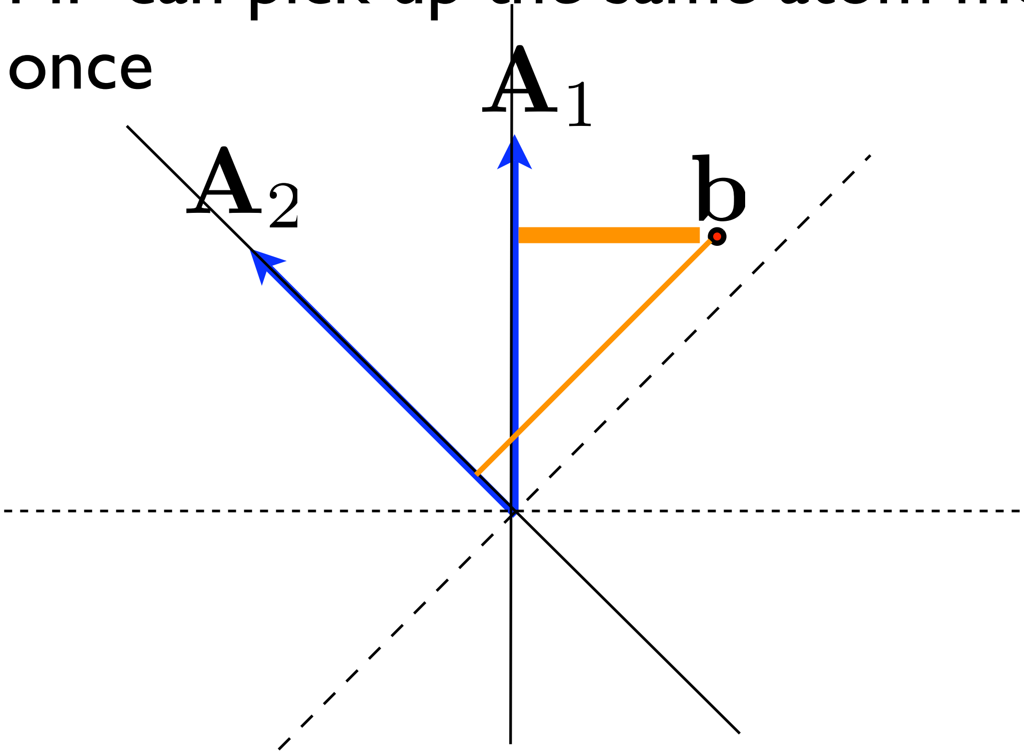
- MP can pick up the same atom more than once



- OMP will never select twice the same atom

Caveats (1)

- MP can pick up the same atom more than once

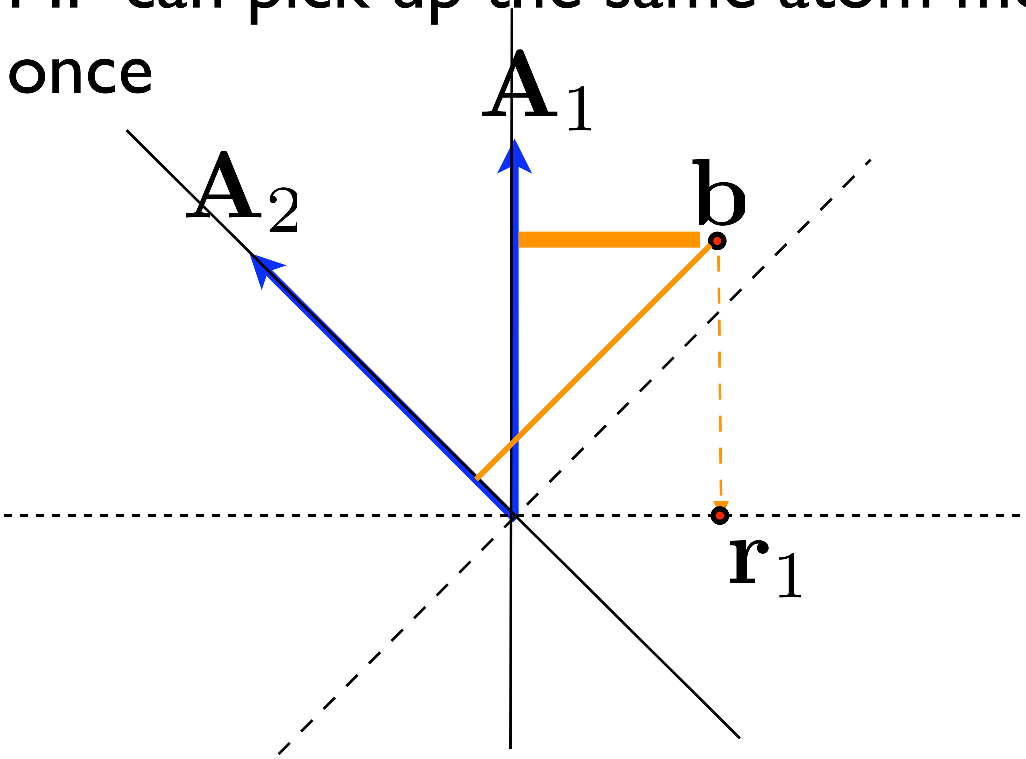


- OMP will never select twice the same atom

IRISA

Caveats (I)

- MP can pick up the same atom more than once

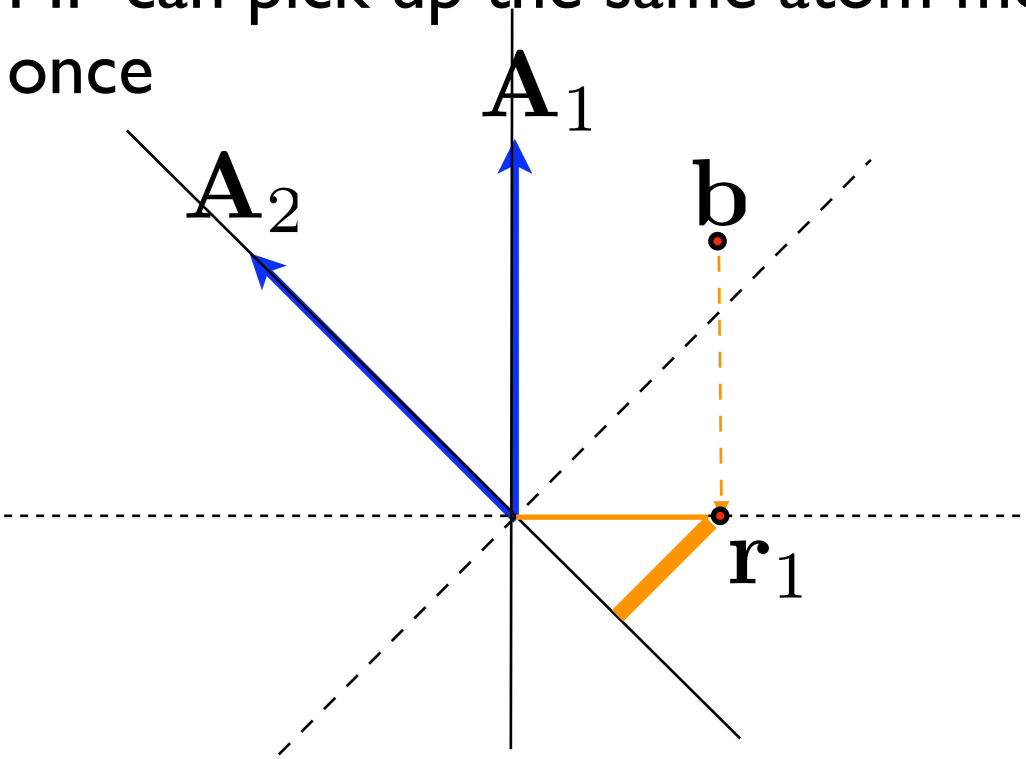


- OMP will never select twice the same atom

IRISA

Caveats (I)

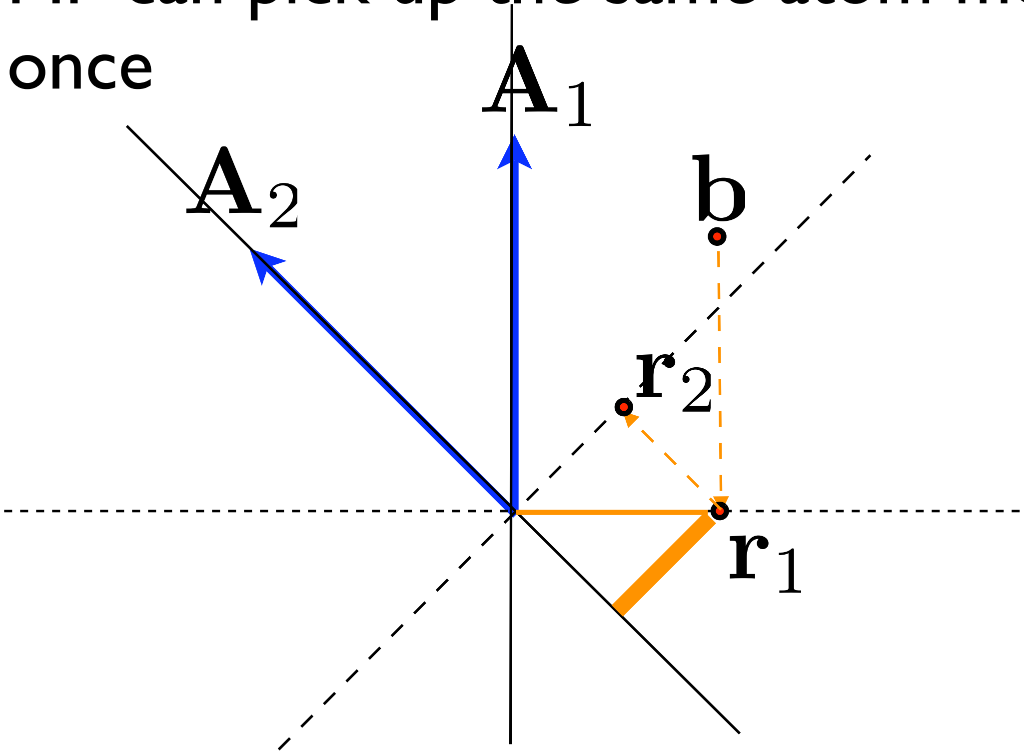
- MP can pick up the same atom more than once



- OMP will never select twice the same atom

Caveats (I)

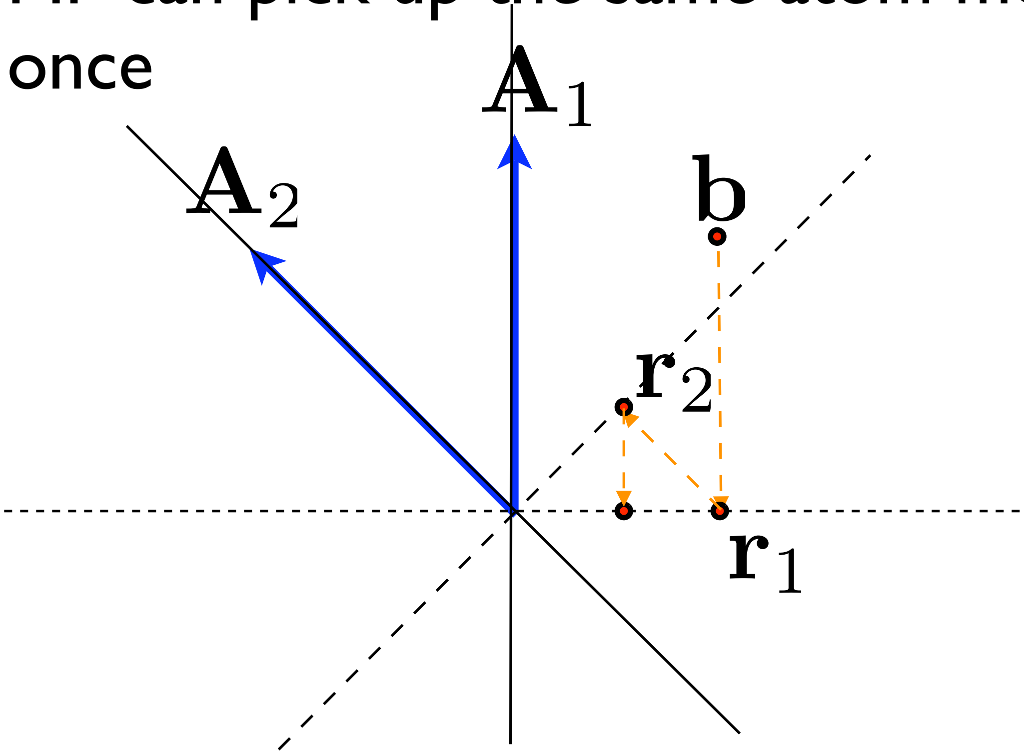
- MP can pick up the same atom more than once



- OMP will never select twice the same atom

Caveats (1)

- MP can pick up the same atom more than once

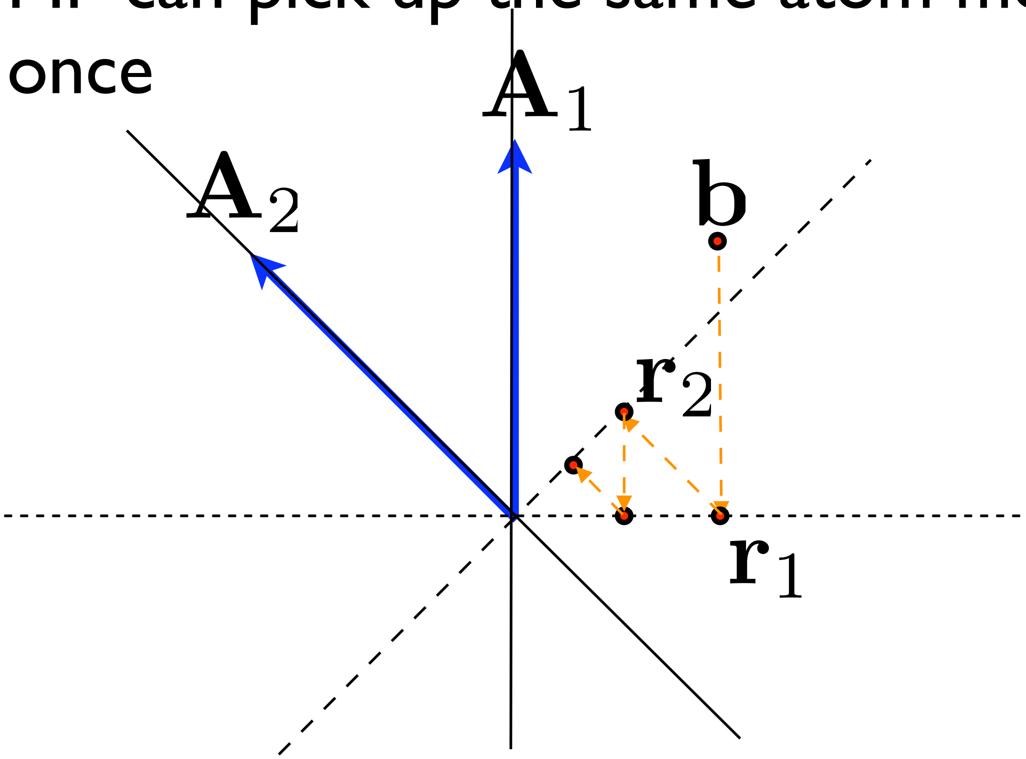


- OMP will never select twice the same atom

IRISA

Caveats (1)

- MP can pick up the same atom more than once



- OMP will never select twice the same atom

IRISA

Caveats (2)

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

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

  - Better atom from **A**

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

  - Residual update

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

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

IRISA

Stagewise greedy algorithms

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

IRISA

Main greedy algorithms

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

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

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

Summary

	Global optimization	Iterative greedy algorithms
Principle	$\min_x \frac{1}{2}\|\mathbf{A}x - \mathbf{b}\|_2^2 + \lambda\|x\|_p^p$	iterative decomposition $\mathbf{r}_i = \mathbf{b} - \mathbf{A}x_i$ • 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 ...)

Complexity of IST

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

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

  - number of iterations guaranteed to approach limit within relative precision ϵ

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

- Limit depends on choice of penalty factor λ, added complexity to adjust it

Complexity of MP

- Number of iterations depends on stopping criterion $\|\mathbf{r}_i\|_2 \leq \epsilon, \|x_i\|_0 \geq k$
- Cost of first iteration = atom selection $O(\mathbf{A})$ (computation of all inner products)
- Naive cost of subsequent iterations = $O(\mathbf{A})$
- If "local" structure of dictionary *[Krstulovic & al, MPTK]*
  - ✦ subsequent iterations only cost $O(\log N)$

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

IRISA

Complexity of OMP

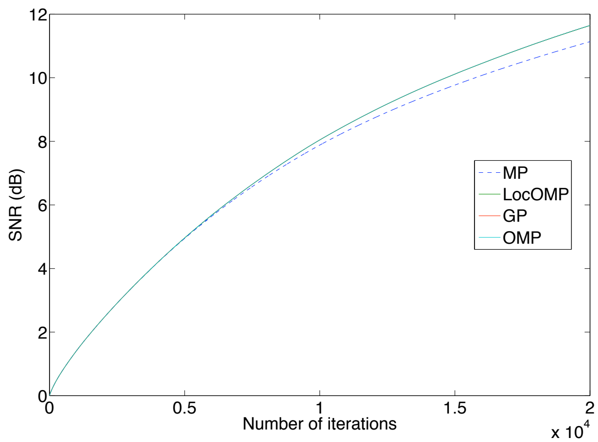
- Number of iterations depends on stopping criterion $\|\mathbf{r}_i\|_2 \leq \epsilon, \|x_i\|_0 \geq k$
- Naive cost of iteration *i*
  - atom selection $O(\mathbf{A})$ + orthoprojection $O(i^3)$
- With iterative matrix inversion lemma
  - atom selection $O(\mathbf{A})$ + coefficient update $O(i)$
- If "local" structure of dictionary *[Mailhé & al, LocOMP]*
  - subsequent approximate iterations only cost $O(\log N)$

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

LoCOMP

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

Fig. 1. SNR depending on the number of iterations



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

Table 3. CPU time per iteration (s)

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

- Implementation in MPTK in progress for larger scale experiments, collaboration with T. Blumensath

Some algorithms / software on the market

- Matlab (simple to adapt, medium scale problems):
  - L1 minimization with an available toolbox
    - http://www.l1-magic.org/ (Candès et al.), ...
  - iterative thresholding
    - http://www.morphologicaldiversity.org/ (Starck et al.)
- MPTK : C++, large scale problems
  - optimized Matching Pursuit
  - millions of unknowns, a few minutes of computation
  - several time-frequency dictionaries
  - builtin multichannel
    - http://mptk.irisa.fr
- More on http://www.dsp.rice.edu/cs

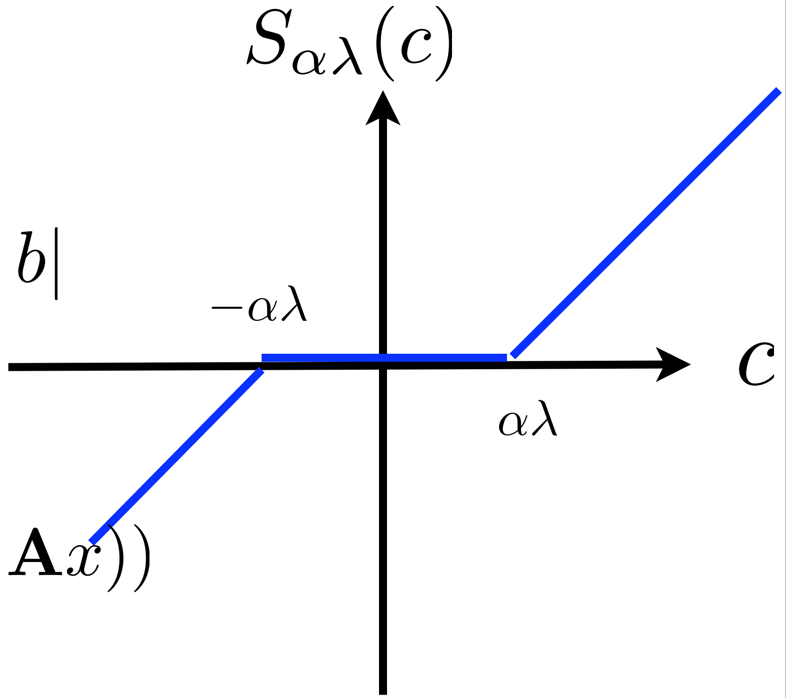
Appendix

Iterative Soft Thresholding (IST)

- **Theorem** : assume
  - consider the iterates $x^{(i+1)} = f(x^{(i)})$ defined by the soft thresholding function
$$f(x) = S_{\alpha\lambda}(x + \alpha \mathbf{A}^T(\mathbf{b} - \mathbf{A}x))$$
  - assume that $a\|x\|_2^2 \leq \|\mathbf{A}x\|_2^2 \leq b\|x\|_2^2, \forall x$ $\quad 0 < a \leq b < \infty$
  - whenever $\alpha = 2/(b+a)$ the iterates converge geometrically in L2 norm to the unique local minimum $x^\star$ of the BPDN optimization problem
  - for $\alpha = 2/(b+a)$ the rate is
$$\|x^{(i)} - x^\star\|_2 \leq \left(\frac{b-a}{b+a}\right)^i \|x^{(0)} - x^\star\|_2$$

IRISA

Convergence of IST (I)



- Soft thresholding satisfies

$$|S_{\alpha\lambda}(a) - S_{\alpha\lambda}(b)| \leq |a-b|$$

- Recall that

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

- Therefore for any *x,y*

$$\begin{aligned} \|f(x) - f(y)\|_q &\leq \|x - y - \alpha\mathbf{A}^T\mathbf{A}(x-y)\|_q \\ &= \|(\mathbf{Id} - \alpha\mathbf{A}^T\mathbf{A})(x-y)\|_q \\ &\leq \|\mathbf{Id} - \alpha\mathbf{A}^T\mathbf{A}\|_{q\to q} \cdot \|x-y\|_q \end{aligned}$$

Convergence of IST (2)

- Assume that for some $1 \leq q \leq \infty$

$$\beta := \|\mathbf{Id} - \alpha \mathbf{A}^T \mathbf{A}\|_{q \to q} < 1$$

- Fixed point theorem (contracting iterations):
  - the sequence $x^{(i)}$ converges in the *p*-norm to the *unique* solution of the fixed point equation

$$x^\star = f(x^\star) = S_\mu(x^\star + \alpha \mathbf{A}^T(\mathbf{b} - \mathbf{A}x^\star))$$

- The convergence is geometric with rate β

$$\|x^{(i)} - x^\star\|_q \leq \beta^i \|x^{(0)} - x^\star\|_q$$

Convergence of IST (3)

- Set q=2. By assumption, in the sense of symmetric matrices

$$a\mathbf{Id} \leq \mathbf{A}^T\mathbf{A} \leq b\mathbf{Id}$$

$$(1-\alpha b)\mathbf{Id} \leq \mathbf{Id} - \alpha\mathbf{A}^T\mathbf{A} \leq (1-\alpha a)\mathbf{Id}$$

- The condition $\beta = \|\mathbf{Id} - \alpha\mathbf{A}^T\mathbf{A}\|_{2\to 2} < 1$ is equivalent to $\max(|1-\alpha b|, |1-\alpha a|) < 1$

$$0 < \alpha < 2/b$$

- The optimum is reached for $\alpha = \dfrac{2}{b+a}$

$$\beta = \frac{b-a}{b+a}$$

IRISA

Proof of the Lemma

- $\mathbf{A}_I$ =matrix with columns of $\mathbf{A}$ indexed by *I*
- The restricted vector $x_I^\star$ is a local minimum of

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

- Since $x_I^\star$ has no zero entry, the objective function is smooth at $x_I^\star$ and its gradient must be zero

$$\mathbf{A}_I^T(\mathbf{A}_I x_I^\star - \mathbf{b}) + \lambda \cdot \mathrm{sign}(x_I^\star) = 0$$

- A similar analysis yields the second condition

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

Limit of IST (2)

- $x^\star$ = any local minimum of BPDN
- *I* = support of $x^\star$
- For indices in I we have

$$\alpha \mathbf{A}_I^T(\mathbf{b} - \mathbf{A}x^\star) = \alpha\lambda \mathrm{sign}(x_I^\star)$$
$$x_I^\star + \alpha \mathbf{A}_I^T(\mathbf{b} - \mathbf{A}x^\star) = (|x_I^\star| + \alpha\lambda)\mathrm{sign}(x_I^\star)$$
$$S_{\alpha\lambda}(x_I^\star + \alpha \mathbf{A}_I^T(\mathbf{b} - \mathbf{A}x^\star)) = |x_I^\star|\mathrm{sign}(x_I^\star) = x_I^\star$$

- For indices not in I we have

$$\begin{aligned} S_{\alpha\lambda}(x_{I^c}^\star + \alpha \mathbf{A}_{I^c}^T(\mathbf{b} - \mathbf{A}x^\star)) &= S_{\alpha\lambda}(\alpha \mathbf{A}_{I^c}^T(\mathbf{b} - \mathbf{A}x^\star)) \\ &= 0 = x_{I^c}^\star \end{aligned}$$

- Therefore $x^\star$ is *the unique* fixed point

IRISA

Limit of IST (3)

- We conclude that

$$x^\star = f(x^\star) = S_{\alpha\lambda}(x^\star + \alpha\mathbf{A}^T(\mathbf{b} - \mathbf{A}x^\star))$$

  - $x^\star$ was any local minimum of BPDN
  - it must be the *unique* fixed point
  - therefore, there is a unique local minimum of BPDN, which is the limit of IST.

IRISA

Homotopy method

$$x_I = (\mathbf{A}_I^T \mathbf{A}_I)^{-1} \left(\mathbf{A}_I^T \mathbf{b} - \lambda \cdot \mathrm{sign}(x_I)\right)$$
$$x_{I^c} = 0$$

- For any sign pattern *s*, define $x^\star(\lambda, s)$ as above, which varies affinely with λ
- If $\|\mathbf{A}_{I(s)^c}^T(\mathbf{A}x^\star(\lambda, s) - \mathbf{b})\|_\infty < \lambda$ then
  - the strict inequality remains true for λ' close to λ, meaning that in a neighborhood of λ the solution to BPDN is indeed $x^\star(\lambda, s)$
  - the sign pattern is therefore piecewise constant
  - breakpoint occur where $\|\mathbf{A}_{I(s)^c}^T(\mathbf{A}x^\star(\lambda, s) - \mathbf{b})\|_\infty = \lambda$

Homotopy algorithm

- For $\lambda > \|\mathbf{A}^T \mathbf{b}\|_\infty$ the solution is $x^\star = 0$ with sign pattern $s_0 = 0$; set $\lambda_0 = \infty$ and *k*=0
- Determine the next breakpoint: λ_{k+1} is the largest value of $\lambda < \lambda_k$ such that either
  - a component of $x^\star_{I_k}(\lambda, s_k)$ vanishes
  - a component violates the inequality

$$\|\mathbf{A}^T_{I_k^c}(\mathbf{A}x^\star(\lambda, s_k) - \mathbf{b})\|_\infty < \lambda$$

- Determine the sign pattern s_{k+1} for $\lambda \lesssim \lambda_k$
  - some components may go to zero
  - some new components may enter