



# Topics in Numerical Analysis II

## Computational Inverse Problems

Bangti Jin (b.jin@cuhk.edu.hk)

Chinese University of Hong Kong

November 16, 2022



# Outline

1  $\ell^1$  solvers

2 Greedy methods



# Review

finite-dimensional formulation

$$b = Ax^* + \eta,$$

- $x^* \in \mathbb{R}^p$ : the unknown signal
- $\eta \in \mathbb{R}^n$ : additive Gaussian noise;  $\epsilon = \|\eta\|$ : noise level
- $A \in \mathbb{R}^{n \times p}$ ,  $p \gg n$ : (normalized column), i.e.,  $\|A_i\| = 1$

The problem has infinitely many solutions (if it has one), which one shall we take ?

sparsity regularization

$$\frac{1}{2} \|Ax - b\|^2 + \alpha \|x\|_1$$



## restricted isometry property (RIP)

- RIP of order  $s$ , if  $\exists$  a  $\delta_s \in (0, 1)$  s.t.

$$(1 - \delta_s) \|c\|^2 \leq \|A_I c\|^2 \leq (1 + \delta_s) \|c\|^2 \quad \forall I \subset S, |I| \leq s.$$

with  $\delta_s$  being the smallest constant for which RIP holds

$$\delta_s := \inf \{ \delta : (1 - \delta) \|c\|^2 \leq \|A_I c\|^2 \leq (1 + \delta) \|c\|^2 \quad \forall |I| \leq s, \forall c \in \mathbb{R}^{|I|} \}$$

denoted by RIP  $(s, \delta_s)$



## Matrices satisfying RIP

- random matrices with i.i.d. Gaussian/Bernoulli with zero mean and variance  $1/p$   
RIP holds with **overwhelming probability** if

$$s \leq cn / \log(p/n)$$

- random matrices from Fourier ensemble: randomly select  $n$  rows from  $p \times p$  discrete Fourier transform matrix *uniformly at random*, and then re-normalized  
RIP holds with **overwhelming probability** if

$$s \leq cn / (\log(p))^6$$

DFT matrix is used in Magnetic Resonance Imaging



under certain conditions on the matrix  $\Psi$  and the true solution  $x^\dagger$ :

$$\|x^\dagger - x_\alpha^\delta\| \leq c\delta$$

conditions

- $\alpha$  by the discrepancy principle

$$\min \|x\|_{\ell^1} \quad \text{s.t.} \quad \|\Psi x - b^\delta\| \leq \delta$$

- with  $s = \|x^\dagger\|_0$ , the result holds on  $\delta_{3s} + 3\delta_{4s} < 2$
- $n$  is nearly of order  $s$ , i.e.,  $n \sim s$  up to log factors for random  $\Psi$
- the constant  $c$  depends on RIP constant
- the reconstruction error is of **the same order** as data error  $\delta$   
much better than the classical inverse problems  $\sim$  sublinear  
 $\Leftarrow$  **much stronger conditions**

there are other methods achieving similar errors

E.J Candes, J. Romberg, T. Tao. Commun. Pure Appl. Math. 59(8), 1207–1223, 2006



$\ell^1$  term is not differentiable, but a generalized derivative exists

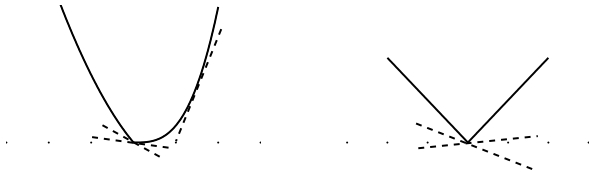
- a vector  $g \in \mathbb{R}^n$  is a **subgradient** of a convex function  $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$  at  $x^0$  if

$$f(x) - f(x^0) \geq \langle x - x^0, g \rangle \quad \forall x \in \text{dom}(f)$$

i.e.,

$$f(x) \geq f(x^0) + \langle x - x^0, g \rangle \quad \forall x \in \text{dom}(f)$$

- the set of subgradient at  $x^0$  is denoted by  $\partial f(x^0)$
- if  $f$  is differentiable at  $x^0$ , then it is identical with  $f'(x^0)$





the subdifferential of  $f(t) = |t|$

- at  $t \neq 0$ ,  $f$  is differentiable,  $\partial f(t) = \{f'(t)\}$ , i.e.,

$$\partial f(t) = \text{sign}(t), \quad t \neq 0$$

- at  $t = 0$ ,  $f(t)$  is not differentiable: any constant  $c$  s.t.

$$|t| = f(t) \geq f(0) + c(t - 0) = ct \quad \forall t \in \mathbb{R}$$

$$\Rightarrow -1 \leq c \leq 1, \text{ i.e. } (\partial|t|)(0) = [-1, 1]$$

Hence,  $\partial|t|$

$$\partial(|t|) = \begin{cases} 1, & t > 0, \\ -1, & t < 0, \\ [-1, 1], & t = 0. \end{cases}$$

property

- $x^*$  is a minimizer to  $f$  if and only if  $0 \in \partial f(x^*)$
- sum rules (under certain mild conditions)





one-dimensional example: fixed  $t$

$$f(s) = \frac{1}{2}(t-s)^2 + \alpha|s|$$

the function is strictly convex  $\exists!$  a unique minimizer

$$f(s) = \frac{1}{2}(t-s)^2 + \alpha|s| = \begin{cases} \frac{1}{2}(t-s)^2 + \alpha s, & s > 0 \\ \frac{1}{2}(t-s)^2 - \alpha s, & s \leq 0 \end{cases}$$

suppose  $t > 0$  and the minimum is achieved at  $s^* > 0$ , then

$$s^* = t - \alpha > 0, \quad f(s^*) = \frac{1}{2}\alpha^2 + \alpha(t - \alpha)$$

$$s^* = 0, \quad f(s^*) = \frac{1}{2}t^2$$

$\Rightarrow$

$$t - \alpha \geq 0 \Rightarrow s^* = t - \alpha$$

$$t - \alpha < 0 \Rightarrow s^* = 0$$

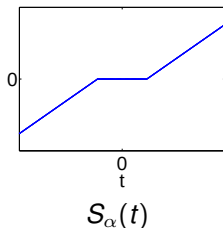
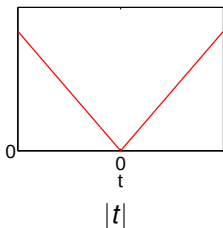


$$s = S_{\alpha}(t) = \begin{cases} t - \alpha, & t > \alpha \\ 0, & |t| \leq \alpha \\ t + \alpha, & t < -\alpha \end{cases}$$

the optimality condition is

$$0 \in (s - t) + \partial\alpha|s|, \quad \text{i.e.} \quad t \in s + \alpha\partial|s|,$$

$\Rightarrow$  soft thresholding operator  $S_{\alpha}(t) = (\partial\alpha|\cdot| + I)^{-1}(t)$



It shrinks the value, and zeros it if small

similar result holds for a general convex penalty ([proximal operator](#))  
for any convex function  $\psi$ , we call

$$\arg_{s \in \mathbb{R}} \min \frac{1}{2} \|s - t\|^2 + \psi(s)$$

the proximal operator  $\text{prox}_\psi(t)$



# Reconstruction methods

How can we obtain such nice pictures numerically ?

- $\ell^1$  regularization & nonconvex counterparts
- greedy algorithms



# Convex approach – l1 penalty

popular approach: basis pursuit or lasso Chen et al 1998, Tibshirani 1996

$$\min_{x \in \mathbb{R}^p} J_\alpha(x) = \frac{1}{2} \|Ax - b\|^2 + \alpha \|x\|_1,$$

- nonsmooth but convex optimization problem
- efficient solver, solid theory
- many existing solvers: gradient projection, coordinate descent algorithm etc.
- theory: good prediction, support recovery (under RIP etc.)



# iterative soft thresholding

iterative soft thresholding Daubechies-defrise-de Mol 2005

an iterative algorithm for computing the solution by surrogate function approach (majorization-minimization, optimization transfer)

$$J_{\alpha}(x) = \frac{1}{2} \|Ax - b\|^2 + \alpha \|x\|_1,$$

observations:

- if  $K = I$ , the problem is easy

$$J_{\alpha}(x) = \sum_i \left( \frac{1}{2} (x_i - b_i)^2 + \alpha |x_i| \right)$$

the problem decouples into  $n$  one-dimensional problems

- if  $K$  is orthonormal, i.e.,  $A^*A = I$ ,

$$J_{\alpha}(x) = \frac{1}{2} \|x - A^*b\|^2 + \alpha \|x\|_1$$



- the presence of an operator  $A \Rightarrow$  surrogate function
- coupling  $f(x) = \frac{1}{2}\|Ax - b\|^2 \approx$  1st-order Taylor expansion ...

given the current guess  $x^k$

$$\begin{aligned} f(x) &= \frac{1}{2}\|A(x - x^k) + Ax^k - b\|^2 \\ &= \frac{1}{2}\|A(x - x^k)\|^2 + \langle A(x - x^k), Ax^k - b \rangle + \frac{1}{2}\|Ax^k - b\|^2 \\ &\leq \frac{\tau_k}{2}\|x - x^k\|^2 + \langle x - x^k, A^*(Ax^k - b) \rangle + \frac{1}{2}\|Ax^k - b\|^2 \\ &:= Q(x, x^k) \end{aligned}$$

it is easy to verify that

$$Q(x^k, x^k) = f(x^k), \quad Q'(x^k, x^k) = f'(x^k)$$

and further

$$Q(x, x^k) \geq f(x) \quad \text{if } \tau_k \geq \|A\|^2$$

algorithm: simplified minimization problem:

$$x^{k+1} \in \arg \min Q(x, x^k) + \alpha \|x\|_1$$



## approximate minimization problem

$$x^{k+1} = \arg \min Q(x, x^k) + \alpha \|x\|_1$$

$$\begin{aligned} Q(x, x^k) + \alpha \|x\|_1 &= \frac{\tau_k}{2} \|x - x^k\|^2 - \langle A^*(Ax^k - b), x - x^k \rangle + \alpha \|x\|_1 \\ &= \frac{\tau_k}{2} \|x - (x^k - \tau^{-1} A^*(Ax^k - b))\|^2 + \alpha \|x\|_1 \\ &\quad - \frac{1}{2\tau_k} \|A^*(Ax^k - b)\|^2 \end{aligned}$$

let

$$\bar{x}^{k+1} = x^k - \tau_k^{-1} A^*(Ax^k - b)$$

then

$$\begin{aligned} Q(x, x^k) + \alpha \|x\|_1 &= \frac{\tau_k}{2} \|x - \bar{x}^{k+1}\|^2 + \alpha \|x\|_1 + \text{cnst} \\ &= \sum_i \left( \frac{\tau_k}{2} (x_i - \bar{x}_i^{k+1})^2 + \alpha |x_i| \right) + \text{cnst} \end{aligned}$$



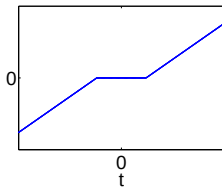


## The one-dimensional optimization problem

$$\frac{1}{2}(s - t)^2 + \alpha|s|$$

the solution  $S_\alpha(t)$  is given by

$$S_\alpha(t) = \begin{cases} t - \alpha, & t > \alpha \\ 0, & |t| \leq \alpha \\ t + \alpha, & t < -\alpha \end{cases}$$



It shrinks the value, and zeros it if small



iterative soft thresholding Daubechies-Defrise-De Mol, 2005

given initial guess  $x^0$ , update the solution iteratively by

$$\bar{x}^{k+1} = x^k - \tau^{-1} A^* (Ax^k - b) \quad (\text{gradient descent})$$

$$x^{k+1} = S_{\tau^{-1}\alpha}(\bar{x}^{k+1}) \quad (\text{thresholding})$$

iterative thresholding iteration is a (nonlinear) gradient descent method

⇒ the convergence is slow ...

- adaptive choice of step size can improve convergence ...
- primal dual active set (PDAS) algorithm  
PDAS = Newton method, for a class of convex optimization

Hintermuller-Ito-Kunisch 2002



choice I: Cauchy step size Cauchy 1847

$$\tau_k = \arg \min_{\tau > 0} \|A(x^k - \tau A^*(Ax^k - b)) - b\|$$

i.e.,

$$\tau_k = \frac{\|A^*(Ax^k - b)\|^2}{\|AA^*(Ax^k - b)\|^2} = \frac{\|d^k\|^2}{\|Ad^k\|^2}$$



## Choice II: Barzilai-Borwein rule Barzilai-Borwein 1988

- to use preceding two iterates to decide the step size

general quasi-Newton method:

$$x^{k+1} = x^k - \underbrace{(B^k)^{-1}}_{\text{approx. inv. Hessian}} g^k$$

with quasi-Newton relation (or secant equation)

$$B^k(x^k - x^{k-1}) = g^k - g^{k-1}$$

select  $D^k = \tau_k I$  and

$$x^{k+1} = x^k - D^k g^k$$

to mimic the quasi-Newton method (in least-squares sense)

$$\min \|(x^k - x^{k-1}) - \tau(g^k - g^{k-1})\|$$

$\Rightarrow$

$$\tau_k = \frac{\langle x^k - x^{k-1}, g^k - g^{k-1} \rangle}{\|g^k - g^{k-1}\|^2}$$



fast iterative shrinkage-thresholding algorithm Nesterov 1980s, Beck-Teboulle 2008

$x^{-1} = x^0$ ,  $z^1 = x^0$ , and for  $k \geq 1$ ,  $t_1 = 1$

$$x^k = S_\alpha(z^k - A^*(Az^k - b))$$

$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$

$$z^{k+1} = x^k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$$

“extrapolated” point  $z^k$



observation:

One-dimensional problem can be solved easily !

The presence of the operator  $K$  messes things up, so we update the solution componentwise

$$x_1^k \in \arg \min J_\alpha(x_1, x_2^{k-1}, \dots, x_p^{k-1})$$

$$x_2^k \in \arg \min J_\alpha(x_1^k, x_2, x_3^{k-1}, \dots, x_p^{k-1})$$

$$\vdots$$

$$x_p^k \in \arg \min J_\alpha(x_1^k, x_2^k, \dots, x_{p-1}^k, x_k)$$

theoretically P. Tseng, 2001

- The sequence have a subsequence converging to the minimizer.
- The sequence of function value to the minimum.

revived interest in statistics Friedman et al 2007



simple case  $\alpha = 0$ ,  $f(x) = \frac{1}{2} \|Ax - b\|^2$   
minimizing over  $x_i$ , with all  $x_j$ ,  $j \neq i$  fixed

$$0 = \nabla_i f(x) = A_i^*(Ax - b) = A_i^*(A_{-i}x_{-i} + A_ix_i - b)$$

i.e.

$$x_i = \frac{A_i^*(b - A_{-i}x_{-i})}{A_i^*A_i}$$

coordinate descent repeats this for  $i = 1, 2, \dots, \dots$

$$x_i = \frac{A_i^*r}{\|A_i\|^2} + x_i^{old}$$

with  $r = y - Ax \Rightarrow O(n)$  operation per cycle



## l1 problem

minimization over  $x_i$ , with  $x_j, j \neq i$  fixed

$$0 = A_i^* A_i x_i + A_i^* (A_{-i} x_{-i} - b) + \alpha s_i$$

$$s_i \in \partial |x_i|$$

$$x_i = S_{\alpha/\|A_i\|^2} \left( \frac{A_i^* (b - A_{-i} x_{-i})}{\|A_i\|^2} \right)$$





# iteratively reweighted least-squares (IRLS)

Another viewpoint: recall for the quadratic penalty

$$\frac{1}{2} \|Ax - b\|^2 + \frac{\alpha}{2} \|x\|^2$$

the optimal solution  $x_\alpha$  satisfies the following optimality system

$$A^*(Ax_\alpha - b) + \alpha x_\alpha = 0$$

i.e.,

$$(A^*A + \alpha I)x_\alpha = A^*b$$



To take advantage of the quadratic problem, we rewrite the l1 problem as (given current estimate  $x^k$ )

$$\frac{1}{2} \|Ax - b\|^2 + \alpha x^t W_k x,$$

with

$$W_k = \text{diag}(|x_i^k|^{-1})$$

Then this gives the iterative scheme (+ regularization with small  $\epsilon > 0$ ):

$$W_k = 2\text{diag}((|x_i^k| + \epsilon)^{-1}),$$
$$x^k = (A^*A + W_k)^{-1} A^*b.$$



## primal-dual active set algorithm key ingredients

- active set  $A_* = \{i : x_i^* \neq 0\}$
- primal variable  $x^*$
- dual variable  $d^*$ , “derivative” of the penalty term

## key steps:

- to determine active set  $A_*$  from  $(x^*, d^*)$
- then update  $x^*$  on  $A_*$  only, and update  $d^*$

both steps can be deduced from the optimality condition



## PDAS: the $\ell^1$ -penalty

$$\frac{1}{2} \|Kx - y\|^2 + \alpha \|x\|_1$$

nondifferentiable  $\Rightarrow$  the Newton method is not directly applicable.  
there are different ways to derive the method  
starting point: optimality system

$$\begin{aligned} K^t Kx^* + d^* &= K^t y \\ d^* &\in \partial \alpha \|x^*\|_1 \quad (\text{subdifferential}) \end{aligned}$$

the first equation is good, but the second relation is inconvenient for numerical treatment. We rewrite it as a nonlinear equation

$$d^* \in \partial \alpha \|x^*\|_1 \Leftrightarrow x^* = S_\alpha(x^* + d^*)$$



By the definition of  $S_\alpha$  and  $x^* = S_\alpha(x^* + d^*)$

1st observation: active set  $A_*$  from  $(x^*, d^*)$

let  $A_* = \{i : x_i^* \neq 0\}$ . Then

$$x_i^* > 0 \Leftrightarrow x_i^* + d_i^* > \alpha, \quad x_i^* < 0 \Leftrightarrow x_i^* + d_i^* < -\alpha$$

$$A_*^+ =: \{i : x_i^* + d_i^* > \alpha\}$$

$$A_*^- =: \{i : x_i^* + d_i^* < -\alpha\}$$

$$A_* =: A_*^+ \cup A_*^-, I_* = A_*^c$$

$A_*$  can be determined from both primal and dual variables



from the first optimality equation  $K^t Kx^* + d^* = K^t y$

2nd observation: determine  $(x^*, d^*)$  from active set  $A_*$

suppose we have found active set  $A_*$ . Let  $y = K^t y$ , then

$$\begin{aligned}x_{I_*}^* &= 0, \\d_{A_*} &= \alpha[1_{A_*^+}^t, -1_{A_*^-}^t]^t (= \partial \|x_{A_*}\|_1) \\K_{A_*}^t K_{A_*} x_{A_*}^* &= \tilde{y}_{A_*} - d_{A_*}^* \\d_{I_*} &= \tilde{y}_{I_*} - K_{I_*}^t K_{A_*} x_{A_*}^*.\end{aligned}$$

The active set uniquely determines the primal and dual variables.



PDAS: tries to approximate active set by primal and dual variables.

## primal dual active set algorithm

define active/inactive set:

$$A_{k+1}^+ = \{i \in S : x_i^k + d_i^k > \alpha\}$$

$$A_{k+1}^- = \{i \in S : x_i^k + d_i^k < -\alpha\}$$

$$A_{k+1} = A_{k+1}^+ \cup A_{k+1}^-, \quad l_{k+1} = A_{k+1}^c$$

Then update the iteration

$$x_{l_{k+1}}^{k+1} = 0,$$

$$d_{A_{k+1}}^{k+1} = \alpha [1_{A_{k+1}^+}^t, -1_{A_{k+1}^-}^t]^t,$$

$$K_{A_{k+1}}^t K_{A_{k+1}} x_{A_{k+1}}^{k+1} = \tilde{y}_{A_{k+1}} - d_{A_{k+1}}^{k+1}$$

$$d_{l_{k+1}}^{k+1} = \tilde{y}_{l_{k+1}} - K_{l_{k+1}}^t K_{A_{k+1}} x_{A_{k+1}}^{k+1}.$$



## overall algorithm

- given initial guess
- define active and inactive set using primal-dual variables
- update solution by solving a regularized system on active set
- check stopping criterion

What can we expect of the algorithm ?





it is equivalent to semismooth Newton method to solve

$$\frac{1}{2} \|Kx - y\|^2 + \alpha \|x\|_1$$

⇒ **local** **superlinear** convergence

### Theorem (one-step convergence)

*Let  $(x^*, d^*)$  be a solution to the KKT system. If  $K_{A_*}$  is of full column rank, and  $(x^0, d^0)$  is close to  $(x^*, d^*)$ . Then  $(x^1, d^1) = (x^*, d^*)$ .*



## implementation

- stopping condition  $A_k = A_{k+1}$  or  $k \geq J$
- linear system solver  
To avoid the ill-posedness of

$$K_{A_{k+1}}^t K_{A_{k+1}} x_{A_{k+1}}^{k+1} = \tilde{y}_{A_{k+1}} - d_{A_{k+1}}^{k+1}$$

it can be replaced by a regularized system

$$(K_{A_{k+1}}^t K_{A_{k+1}} + \beta I_{A_{k+1}}) x_{A_{k+1}}^{k+1} = \tilde{y}_{A_{k+1}} - d_{A_{k+1}}^{k+1}$$

linear system solver: direct solver or conjugate gradient method

- initial guess: continuation/path following strategy



continuation strategy on  $\alpha$ :

$$\Lambda = \{\alpha_0, \alpha_1, \dots, \alpha_L\}, \quad \alpha_i = \alpha_0 \rho^i, \quad \rho \in (0, 1)$$

compute  $(x_{\alpha_0}, d_{\alpha_0})$ , and use it for the initial guess for  $(x_{\alpha_1}, d_{\alpha_1})$  etc ...

$\Rightarrow$  the sequence of minimizers can be used for choosing "optimal"  $\alpha$ :

- Bayesian information criterion

$$\min_{\alpha \in \Lambda} \left\{ BIC(\alpha) := \ln \frac{\|Kx_\alpha - y\|^2}{n} + \frac{\ln n}{n} \|x_\alpha\|_0 \right\}$$

- discrepancy principle

$$\|Kx_{\alpha_k} - y\| \leq \epsilon$$

- L-curve criterion



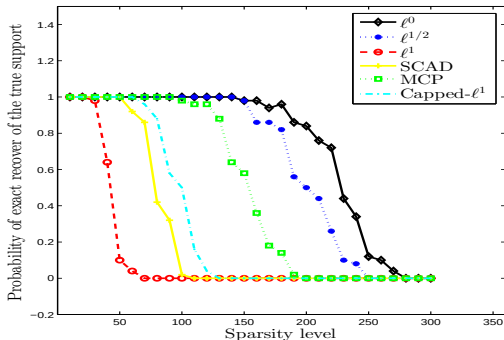
The sampling matrix  $\Psi$  is random Gaussian of size  $500 \times 1000$  and  $x^*$  has 50 nonzero entries with Gaussian noise i.d.d.

| method      | CPU time       | l2 error      | $\ x_\alpha\ _0$ |
|-------------|----------------|---------------|------------------|
| <b>PDAS</b> | <b>4.38e-1</b> | <b>6.9e-6</b> | <b>74</b>        |
| LARS        | 7.33e-1        | 6.8e-6        | 74               |
| FISTA       | 7.70e-1        | 8.3e-6        | 96               |



what is beyond

- nonlinear forward operators (many medical imaging problems)
- structured sparsity patterns
- total variation regularization
- nonconvex penalties (can be efficiently solved)



setting: Gaussian  $\Psi$  and noise,  $500 \times 1000$ ,  $DR = 10^3$ ,  $\sigma = 0.01$

- with 500 data points: the  $\ell^1$  allows exact support recovery only if solution is **very sparse**
- nonconvex models allow recovering far more nonzeros



Like the classical case, apart from penalty, one can also have iterative methods – greedy methods

- orthogonal matching pursuit (OMP)
- compressive sampling matching pursuit (CoSaMP)
- subspace pursuit (SP)
- iterative hard thresholding (IHT)
- ...

$$\min_x \|Ax - b\|_2^2 \quad \text{s.t. } \|x\|_0 \leq s$$

with an upper bound on the sparsity level  $s$

main idea: estimate the support of the signal sequentially / recursively



## orthogonal matching pursuit Tropp 2004

- input  $s, A, b$
- initialize  $x = 0, T = \emptyset, b_r = b$
- iterate  $k = 1, 2, \dots, s$ 
  - $i_k = \arg \max_i |(A^t b_r)_i|$
  - $T^k = T^{k-1} \cup \{i_k\}$
  - $x_{T^k} = A_{T^k}^\dagger b$
  - $b_r = b - Ax$

## characterization

$$A_{T^k}^\dagger b = \arg \min \|A_{T^k} x_{T^k} - b\|^2$$

orthogonality  $\Rightarrow$  **orthogonal** matching pursuit

$$A_{T^k}^t b_r = 0$$





If the matrix satisfies

$$\mu < 1/2s$$

Then the OMP is guaranteed to exactly recover all  $s$ -sparse  $x$  from the measurement vector  $b$ .

The key idea: to show  $i_k \in T^\dagger := \text{supp}(x^\dagger)$



step 1: to show  $i_1 \in T^\dagger$ ,  $i_1 = \arg \max_i |(A_i, y)|$

■  $b = Ax = \sum_{j \in T^\dagger} A_j x_j$

$$\forall i : |(A_i, b)| = |(A_i, \sum_{j \in T^\dagger} A_j x_j)| = |\sum_{j \in T^\dagger} x_j (A_i, A_j)|$$

■ if  $i \notin T^\dagger$

$$|(A_i, b)| \leq \sum_{j \in T^\dagger} |x_j| |(A_i, A_j)| \leq \mu \sum_{j \in T^\dagger} |x_j| = \mu \sqrt{s} \|x\|_2$$

i.e.

$$\max_{i \notin T^\dagger} |(A_i, b)| \leq \mu \sqrt{s} \|x\|_2$$



■ if  $i \in T^\dagger$ :

$$|(A_i, b)| \geq |x_i(A_i, A_i)| - \left| \sum_{j \neq i} x_j(A_j, A_i) \right| \geq |x_i| - \mu \sum_{j \neq i} |x_j|$$

■  $\max_{i \in T^\dagger} \geq \|x\|_2 / \sqrt{s}$

■ hence

$$\max_{i \in T^\dagger} |(A_i, b)| \geq \frac{\|x\|_2}{\sqrt{s}} - \mu \sqrt{s} \|x\|_2$$

Now suppose  $\mu < 1/2s$ ,  $\frac{1}{\sqrt{s}} \|x\|_2 \geq 2\mu \sqrt{s} \|x\|_2$ , i.e.

$$\max_{i \in T^\dagger} |(A_i, b)| \geq \frac{1}{\sqrt{s}} \|x\|_2 - \mu \sqrt{s} \|x\|_2 > \mu \sqrt{s} \|x\|_2 \geq \max_{i \notin T^\dagger} |(A_i, b)|$$

Hence

$$i_1 \in T^\dagger.$$



the  $k$ th iteration: Let  $i_1, \dots, i_{k-1}$  be the indices in the first  $k - 1$  iterations

- $T^{k-1} = \{i_1, \dots, i_{k-1}\}$
- $b_r = b - A_{T^{k-1}} A_{T^{k-1}}^\dagger b$
- to show  $i_k = \arg \max_i |(A_i, b_r)| \in T^\dagger$
- $b_r = A_{T^\dagger} x_{T^\dagger} - A_{T^{k-1}} (A_{T^{k-1}}^\dagger b) \in \text{span}(A_{T^\dagger})$
- $b_r = A_{T^\dagger} x'_{T^\dagger}$
- use the same argument  $i_k : i_k \in T^\dagger$  and  $i_k \notin T^{k-1}$



## subspace pursuit Dai-Milenkovic 2009

- input  $s, A, b$
- initialize  $T^0 = \text{supp}(H_s(A^t b))$  (the largest  $s$  entries),  
 $b_r = b - Ax_{T^0}$
- iterate  $k = 1, 2$ , until stop
  - $\tilde{T}^k = T^{k-1} \cup \text{supp}(H_s(A^t b_r))$
  - $\tilde{x}_{\tilde{T}^k} = A_{\tilde{T}^k}^\dagger b, \tilde{x}_{(\tilde{T}^k)^c} = 0$
  - $T^k = \text{supp}(H_s(\tilde{x}))$
  - $x_{T^k} = A_{T^k}^\dagger b$
  - $b_r = b - Ax^k$



## compressive sampling matching pursuit (CoSaMP) Needell-Tropp 2009

- input  $s, A, b$
- initialize  $x^0 = 0, b_r = b$
- iterate  $k = 1, 2, \dots$ 
  - $\tilde{T}^k = T^{k-1} \cup \text{supp}(H_{2s}(A^t b_r))$
  - $\tilde{x} = A_{\tilde{T}^k}^\dagger b$
  - $x^k = \text{supp}(H_s(\tilde{x}))$  ( $T^k = \text{supp}(H_s(\tilde{x}))$ )
  - $b_r = b - Ax^k$



## iterative hard thresholding Blumensath-Davies 2009

- input  $s, A, b$
- initialize  $x^0 = 0$
- iterate  $k = 1, 2, \dots$

$$x^k = H_s(x^{k-1} + A^t(b - Ax^{k-1}))$$

gradient descent + hard thresholding (pick the  $s$  largest component)

more generally

$$x^k = H_s(x^{k-1} + \mu A^t(b - Ax^{k-1}))$$



## references

- E Candes, J. Romberg, T Tao, CPAM 2006
- I. Daubechies et al, CPAM 2005
- S Wright, R Nowak, M Figueiredo, ITSP 2009
- Q. Fan, Y Jiao, X. Lu, ITSP 2014
- B Jin, P Maass, IP, 2012