Topics in Numerical Analysis II
Computational Inverse Problems

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

Chinese University of Hong Kong

Oct 21, 2024

Outline

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 δ_s being the smallest constant for which RIP holds

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

denoted by RIP (s, δ_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 Ψ and the true solution $x^\dagger$:

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

conditions

- α by the discrepancy principle

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

- 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 A
- the constant c depends on RIP constant
- the reconstruction error is of the same order as data error δ
  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

ℓ^1 term is not differentiable, but a generalized derivative exists

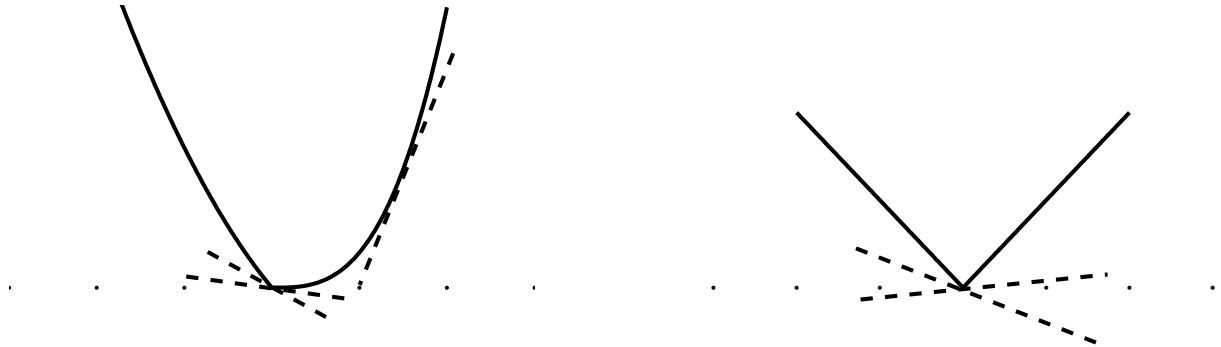
- a vector $g \in \mathbb{R}^n$ is a subgradient of a convex function $f(x) : \mathbb{R}^n \to \mathbb{R}$ at x^0 if

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

  i.e.,

$$f(x) \geq f(x_0) + \langle x - x^0, g \rangle \quad \forall x \in \mathrm{dom}(f)$$

- 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.,

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

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

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

$\Rightarrow -1 \leq c \leq 1$, 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

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

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

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

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

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

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

$\Rightarrow$

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

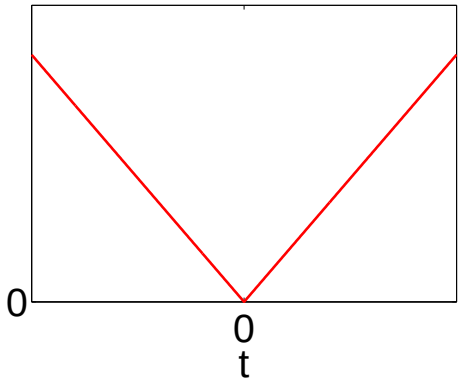
$$t - \alpha < 0 \Rightarrow s^* = 0$$

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

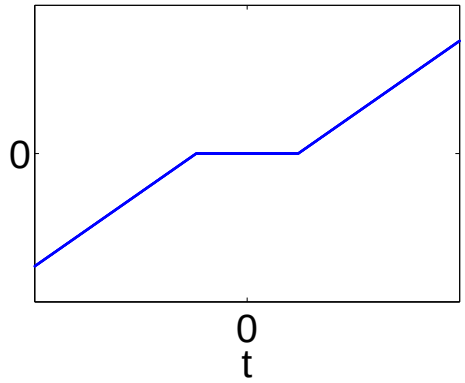
the optimality condition is

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

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



$|t|$



$S_\alpha(t)$

It shrinks the value, and zeros it if small

similar result holds for a general convex penalty (proximal operator) for any convex function ψ, we call

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

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

Reconstruction methods

How can we obtain such nice pictures numerically ?

- ℓ^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)

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

observations:

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

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

  the problem decouples into n one-dimensional problems

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

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

- 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

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

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

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

algorithm: simplified minimization problem:

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

approximate minimization problem

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

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

let

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

then

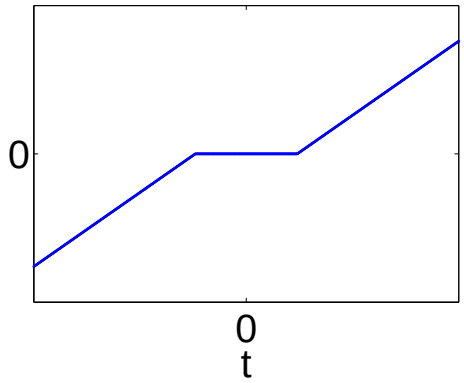
$$\begin{aligned} Q(x, x^k) + \alpha\|x\| &= \tfrac{\tau_k}{2}\|x - \bar{x}^{k+1}\|^2 + \alpha\|x\|_1 + \text{cnst} \\ &= \sum_i \left(\tfrac{\tau_k}{2}(x_i - \bar{x}_i^{k+1})^2 + \alpha|x_i|\right) + \text{cnst} \end{aligned}$$

The one-dimensional optimization problem

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

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

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



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

$$\bar{x}^{k+1} = x^k - \tau_k^{-1} A^*(Ax^k - b) \qquad \text{(gradient descent)}$$
$$x^{k+1} = S_{\tau_k^{-1}\alpha}(\bar{x}^{k+1}) \qquad \text{(thresholding)}$$

iterative soft thresholding is a (nonlinear) gradient descent method

$\Rightarrow$ 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 $D^k = (B_k)^{-1} = \tau_k I$ and

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

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

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

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

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 A messes things up, so we update the solution componentwise

$$
\begin{aligned}
x_1^k &\in \arg\min J_\alpha(x_1, x_2^{k-1}, \cdots, x_p^{k-1}) \\
x_2^k &\in \arg\min J_\alpha(x_1^k, x_2, x_3^{k-1}, \cdots, x_p^{k-1}) \\
&\vdots \\
x_p^k &\in \arg\min J_\alpha(x_1^k, x_2^k, \cdots, x_{p-1}^k, x_k)
\end{aligned}
$$

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

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

i.e.

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

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

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

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 reweighed 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_α 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

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

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

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

primal-dual active set algorithm
key ingredients

- active set $\mathcal{A}_* = \{i : x_i^* \neq 0\}$
- primal variable x^*
- dual variable d^*, "derivative" of the penalty term

key steps:

- to determine active set $\mathcal{A}_*$ from (x^*, d^*)
- then update x^* on $\mathcal{A}_*$ only, and update d^*

both steps can be deduced from the optimality condition

PDAS: the ℓ^1-penalty

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

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

$$A^tAx^* + d^* = A^ty$$
$$d^* \in \partial\alpha\|x^*\|_1 \quad \text{(subdifferential)}$$

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_α and $x^* = S_\alpha(x^* + d^*)$

1st observation: active set $\mathcal{A}_*$ from (x^*, d^*)

let $\mathcal{A}_* = \{i : x_i^* \neq 0\}$. Then

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

$$\mathcal{A}_*^+ =: \{i : x_i^* + d_i^* > \alpha\}$$
$$\mathcal{A}_*^- =: \{i : x_i^* + d_i^* < -\alpha\}$$
$$\mathcal{A}_* =: \mathcal{A}_*^+ \cup \mathcal{A}_*^-, \mathcal{I}_* = \mathcal{A}_*^c$$

$\mathcal{A}_*$ can be determined from both primal and dual variables

from the first optimality equation $A^t A x^* + d^* = A^t y$

2nd observation: determine (x^*, d^*) from active set $\mathcal{A}_*$

suppose we have found active set $\mathcal{A}_*$. Let $\tilde{y} = A^t y$, then

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

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:

$$\mathcal{A}^+_{k+1} = \{i \in S : x^k_i + d^k_i > \alpha\}$$
$$\mathcal{A}^-_{k+1} = \{i \in S : x^k_i + d^k_i < -\alpha\}$$
$$\mathcal{A}_{k+1} = \mathcal{A}^+_{k+1} \cup \mathcal{A}^-_{k+1}, \quad \mathcal{I}_{k+1} = \mathcal{A}^c_{k+1}$$

Then update the iteration

$$x^{k+1}_{\mathcal{I}_{k+1}} = 0,$$
$$d^{k+1}_{\mathcal{A}_{k+1}} = \alpha[1^t_{\mathcal{A}^+_{k+1}}, -1^t_{\mathcal{A}^-_{k+1}}]^t,$$
$$A^t_{\mathcal{A}_{k+1}} A_{\mathcal{A}_{k+1}} x^{k+1}_{\mathcal{A}_{k+1}} = \tilde{y}_{\mathcal{A}_{k+1}} - d^{k+1}_{\mathcal{A}_{k+1}}$$
$$d^{k+1}_{\mathcal{I}_{k+1}} = \tilde{y}_{\mathcal{I}_{k+1}} - A^t_{\mathcal{I}_{k+1}} A_{\mathcal{A}_{k+1}} x^{k+1}_{\mathcal{A}_{k+1}}.$$

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

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

$\Rightarrow$ local superlinear convergence

Theorem (one-step convergence)

Let (x^*, d^*) be a solution to the KKT system. If $A_{\mathcal{A}_*}$ 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 $\mathcal{A}_k = \mathcal{A}_{k+1}$ or $k \geq J$
- linear system solver
  To avoid the ill-posedness of

$$A^t_{\mathcal{A}_{k+1}} A_{\mathcal{A}_{k+1}} x^{k+1}_{\mathcal{A}_{k+1}} = \tilde{y}_{\mathcal{A}_{k+1}} - d^{k+1}_{\mathcal{A}_{k+1}}$$

  it can be replaced by a regularized system

$$(A^t_{\mathcal{A}_{k+1}} A_{\mathcal{A}_{k+1}} + \beta I_{\mathcal{A}_{k+1}}) x^{k+1}_{\mathcal{A}_{k+1}} = \tilde{y}_{\mathcal{A}_{k+1}} - d^{k+1}_{\mathcal{A}_{k+1}}$$

  linear system solver: direct solver or conjugate gradient method
- initial guess: continuation/path following strategy

continuation strategy on α:

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

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" α:

- Bayesian information criterion

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

- discrepancy principle

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

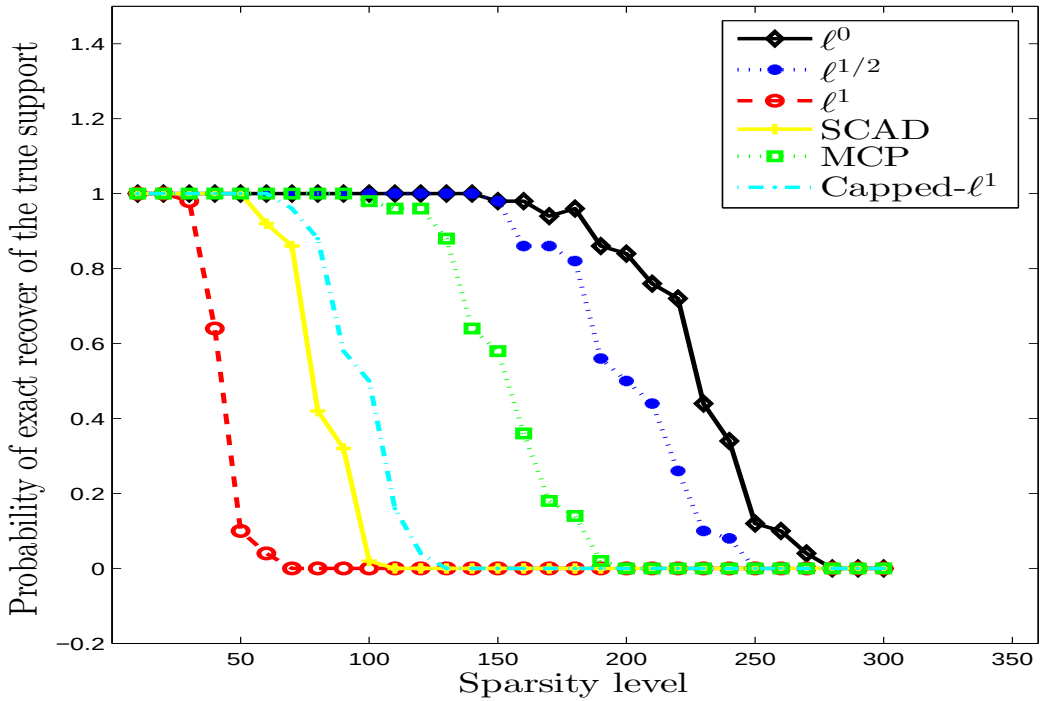
- L-curve criterion

The sampling matrix A is random Gaussian of size 500×1000 and x^* has 50 nonzero entries with Gaussian noise i.d.d.

method	CPU time	l2 error	$\|x_\alpha\|_0$
PDAS	**4.38e-1**	**6.9e-6**	**74**
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 A and noise, 500×1000, $DR = 10^3$, $\sigma = 0.01$

- with 500 data points: the ℓ^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 $k = 1, 2, \ldots, s$
  - $i_k = \arg\max_i |(A^t b_r)_i|$
  - $T^k = T^{k-1} \cup \{i_k\}$
  - $x_{T^k} = A^\dagger_{T^k} b$
  - $b_r = b - Ax$

characterization

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

orthogonality $\Rightarrow$ orthogonal matching pursuit

$$A^t_{T^k} b_r = 0$$

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 $i_k \in T^\dagger := \mathrm{supp}(x^\dagger)$

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 $i \in T^\dagger$:

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

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

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

Hence

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

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

- $T^{k-1} = \{i_1, \ldots, i_{k-1}\}$
- $b_r = b - A_{T^{k-1}} A^{\dagger}_{T^{k-1}} b$
- to show $i_k = \arg\max_i |(A_i, b_r)| \in T^{\dagger}$
- $b_r = A_{T^{\dagger}} x_{T^{\dagger}} - A_{T^{k-1}} (A^{\dagger}_{T^{k-1}} b) \in \mathrm{span}(A_{T^{\dagger}})$
- $b_r = A_{T^{\dagger}} x'_{T^{\dagger}}$
- 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 $T^0 = \mathrm{supp}(H_s(A^t b))$ (the largest s entries), $b_r = b - Ax_{T^0}$
- iterate $k = 1, 2$, until stop
  - $\tilde{T}^k = T^{k-1} \cup \mathrm{supp}(H_s(A^t b_r))$
  - $\tilde{x}_{\tilde{T}^k} = A^{\dagger}_{\tilde{T}^k} b$, $\tilde{x}_{(\tilde{T}^k)^c}$=0
  - $T^k = \mathrm{supp}(H_s(\tilde{x}))$
  - $x_{T^k} = A^{\dagger}_{T^k} b$
  - $b_r = b - Ax^k$

compressive sampling matching pursuit (CoSaMP) Needell-Tropp 2009

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

iterative hard thresholding Blumensath-Davies 2009

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

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

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