Sparsity Regularization

Bangti Jin

Course “Inverse Problems & Imaging”

Outline

problem setup

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., $\|A_i\| = 1$

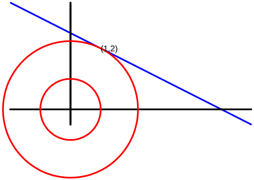
The problem has infinitely many solutions (if it has one), which one shall we take ?

insights from "exact data"

toy example: find a "reasonable" solution to the problem

$$x_1 + 2x_2 = 5$$

There are infinitely many solutions.

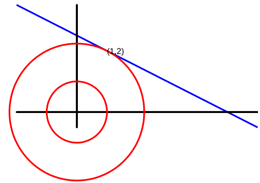


Which one shall we take ?
convention: least-squares Gauss 1809

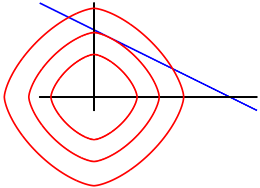
$$\begin{aligned} &\min |x_1|^2 + |x_2|^2 \\ &\text{s.t.} \quad x_1 + 2x_2 = 5 \end{aligned}$$

generalized "minimum-energy" solution

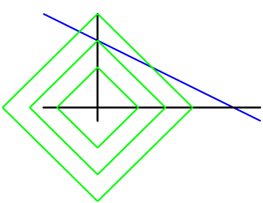
$$\begin{aligned} &\min |x_1|^p + |x_2|^p, 0 \leq p \leq 2 \\ &\text{s.t.} \quad x_1 + 2x_2 = 5 \end{aligned}$$



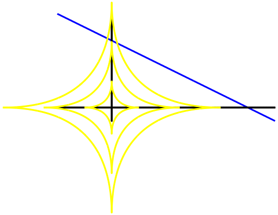
$p = 2$



$p = 3/2$



$p = 1$



$p = 1/2$

in case of noisy data: Tikhonov regularization

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

two possible choices of $\psi(x)$ (convexity ...)

- classical Tikhonov regularization

$$\psi(x) = \frac{1}{2}\|x\|_2^2 =: \frac{1}{2}\sum_i |x_i|^2$$

- sparsity regularization

$$\psi(x) = \|x\|_p^p =: \frac{1}{p}\sum_i |x_i|^p, \ p \in [0, 1]$$

- general analogues ...

in case of noisy data: Tikhonov regularization

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

assumption: i.i.d. additive Gaussian noise on the data

$$b_i=b_i^\dagger+\xi_i, \quad \xi_i\sim N(0,\sigma^2)$$

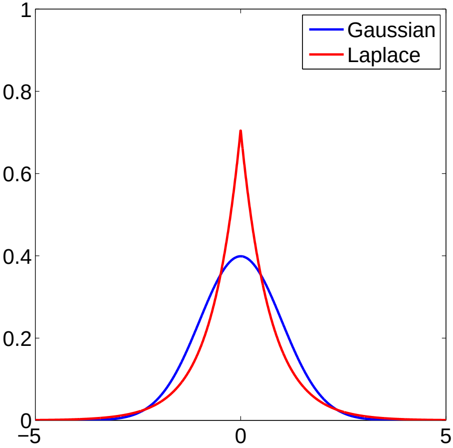
$\Rightarrow$likelihood

$$p(b|x)\propto e^{-\frac{1}{2\sigma^2}(Ax-b)^2}$$

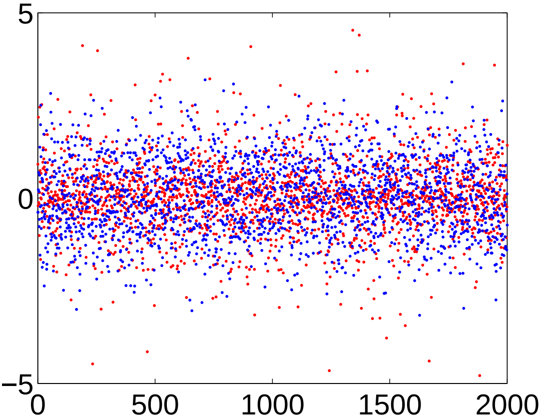
assumption: prior knowledge

$$p(x)\propto e^{-\lambda\psi(x)}$$

- classical Tikhonov regularization $\Leftrightarrow$ Gaussian prior distribution
- sparsity regularization $\Leftrightarrow$ Laplace distribution



Gaussian v.s. Laplace distribution



Gaussian v.s. Laplace

The energy can be more general:

$$\psi(x) = \tilde{\psi}(Wx),$$

under certain transformation, e.g., wavelet, framelet, curvelet, shearlet ...

The discussions below extend to these more complex cases natural idea for sparse solution is to penalize the number of unknowns

$$\frac{1}{2}\|Ax - b\|^2 + \alpha\|x\|_0$$

where

$$\|x\|_0 = \#(\text{nonzeros in } x)$$

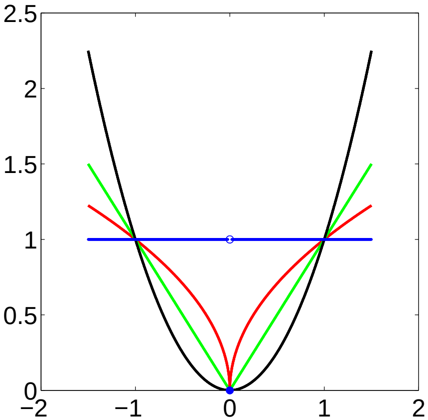
conceptually intuitive, but computationally very challenging: approximations:

- bridge penalty

$$\|x\|_q^q = \sum |x_i|^q, \quad q \in (0, 1)$$

- l1 penalty

$$\|x\|_1 = \sum |x_i|$$


2.5
2
1.5
1
0.5
0
−2
−1
0
1
2

- The l0 penalty is the genuine choice, but VERY challenging there are different ways to approximate it ...
- lq is an approximation, and there are many others
- especially l1 is very popular, since l1 is convex
- further, there is a solid theory

notation:

- $S = \{1, \ldots, p\}$
- $I \subset S$, x_I: subvector consisting of entries of x indexed by $i \in I$
- $I \subset S$, A_I: submatrix consisting of columns of A indexed by $i \in I$

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)

- RIP $(s, \delta_s) \Rightarrow$

$$1-\delta_s \leq \lambda_{min}(A_I^* A_I) \leq \lambda_{max}(A_I^* A_I) \leq 1+\delta_s$$

  the submatrix A_I is fairly well-conditioned

- RIP is difficult to compute

under certain conditions on the matrix A and the true solution x^*:

$$\|x^* - x_\alpha\| \leq C\epsilon$$

conditions

- the result holds on $\delta_{3s} + 3\delta_{4s} < 2$
- n is nearly of order s, i.e., $n \geq s$ up to log factors
- the reconstruction error is of the same order as data error ϵ much better than the classical inverse problems $\sim$ sublinear $\Leftarrow$ much stronger conditions

there are some other methods that also achieves the similar errors

convex function

convex functions: $f(x)$ is convex over its domain $\mathrm{dom}(f)$ if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad \forall \lambda \in [0,1], x_1, x_2 \in \mathrm{dom}(f)$$

- f is concave if $-f$ is convex
- f is strictly convex if

  $$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda) f(x_2) \quad \forall \lambda \in (0,1), x_1 \neq x_2 \in \mathrm{dom}(f),$$

- if f differentiable

  $$f(x_2) \geq f(x_1) + (\nabla f(x_1), x_2 - x_1)$$

  first-order Taylor exp. is a global under-estimator

how to verify:

- by definition
- if f is twice differential: convex $\equiv f'' \geq 0$

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

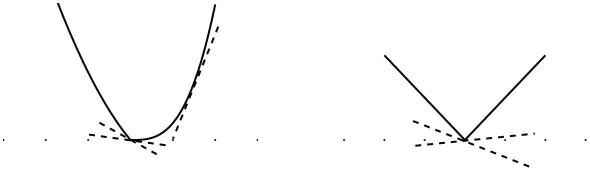
- a vector $g \in \mathbb{R}^n$ is a subgradient of a convex function $f(x) : \mathbb{R}^n \to \mathbb{R}$ at x^0 if

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

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.,

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

- 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

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

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

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

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

$\Rightarrow$

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

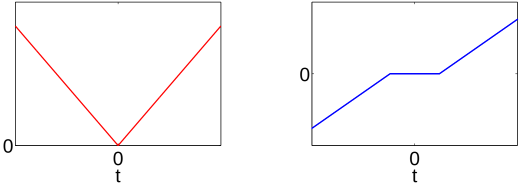
$$t - \alpha < 0 \Rightarrow s^* = 0$$

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

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



$|t|$ $S_\alpha(t)$

It shrinks the value, and zeros it if small

Convex approach – l1 penalty

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

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

How can we obtain such nice pictures numerically ?

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)

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

observations:

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

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

  the problem decouples into n one-dimensional problems
- if K is orthonormal, i.e., $A^*A = I$,

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

- 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

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

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

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

algorithm: simplified minimization problem:

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

UCL CENTRE FOR INVERSE PROBLEMS

approximate minimization problem

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

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

let

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

then

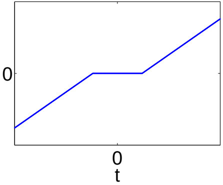
$$\begin{aligned} Q(x, x^k) + \alpha\|x\| &= \tfrac{\tau_k}{2}\|x - \bar{x}^{k+1}\|^2 + \alpha\|x\|_1 + \text{cnst} \\ &= \sum_i \left(\tfrac{\tau_k}{2}(x_i - \bar{x}_i^{k+1})^2 + \alpha|x_i|\right) + \text{cnst} \end{aligned}$$

The one-dimensional optimization problem

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

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

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



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

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

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

iterative thresholding iteration 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

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:

$$x^{k+1} = x^k - (B^k)^{-1} g^k, \quad B^k(x^k - x^{k-1}) = g^k - g^{k-1} \text{(quasi-Newton relation)}$$

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 $k \geq 1$, $t_1 = 1$

$$\begin{aligned} x^k &= S_\alpha(z^k - A^*(Az^k - b)) \\ t_{k+1} &= \frac{1 + \sqrt{1 + 4t_k^2}}{2} \\ z^{k+1} &= x^k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1}) \end{aligned}$$

"extrapolated" point z^k

UCL CENTRE FOR INVERSE PROBLEMS

observation:

$$\boxed{\text{One-dimensional problem can be solved easily !}}$$

The presence of the operator K 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

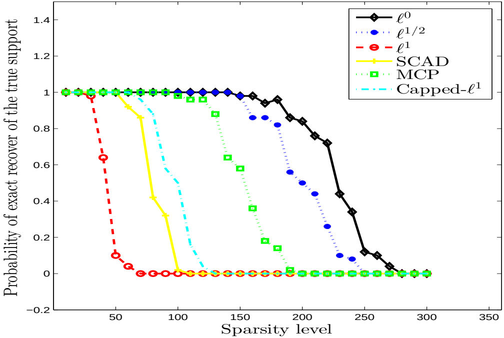
$$W_k = \mathrm{diag}(|x_i^k|^{-1})$$

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

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

what is beyond

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



setting: Gaussian Ψ 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

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