Solving L_1 regularized Least Squares by Reweighted Least Squares

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

manshun.ang@umons.ac.be Homepage: angms.science

First draft : August 5, 2020
Last update : August 6, 2020

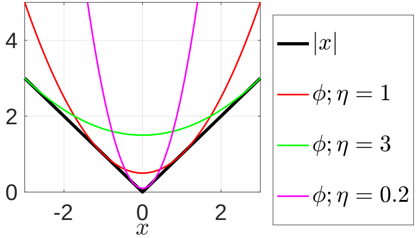
Reweighted Least squares on absolute value

- The absolute value $|x|$ is a nonsmooth function.
- The function can be approximate by the parametric quadratic function

$$\phi(x;\eta) = \frac{1}{2\eta}x^2 + \frac{1}{2}\eta,$$

for some $\eta \geq 0$.

- The idea is to use a quadratic function to approximate the nonsmooth absolute value.



The advantage of replacing $|x|$ by ϕ is that ϕ is smooth and it is a quadratic function, easy to deal with.

Small details about ϕ

- The optimal η for $\phi(x,\eta)$ at a specific point $x = x_0$ is $\eta = |x_0|$: this can be shown by completing the square. The key is to treat $x^2 = |x|^2$.

$$\begin{aligned}
\phi(x;\eta) &= \frac{1}{2\eta}x^2 + \frac{1}{2}\eta \\
&= \frac{1}{2\eta}|x|^2 + \frac{1}{2}\eta \\
&= \frac{1}{2\eta}|x|^2 + \frac{1}{2\eta}\eta^2 \\
&= \frac{1}{2\eta}(|x|^2 + \eta^2) \\
&= \frac{1}{2\eta}(|x|^2 - 2|x|\eta + \eta^2 + 2|x|\eta) \\
&= \frac{1}{2\eta}(|x| - \eta)^2 + |x|.
\end{aligned}$$

  Now ϕ is minimized on η at $\eta = |x|$, which gives $\phi(x;\eta) = |x|$.

- Furthermore, why we have to restrict $\eta \geq 0$ in ϕ: if $\eta < 0$, it flips the quadratic function along the x-axis.

Application to L_1-regularized Least squares

- L_1-regularized least squares

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

  Here the optimization variable is $\mathbf{x} \in \mathbb{R}^n$.

- As $\|\mathbf{x}\|_1 = \sum |x_i|$, using the ϕ function on each component in $\mathbf{x}$, we arrive at an equivalent problem

$$\min_{\eta}\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 + \lambda\Big(\sum_{i=1}^{n} \frac{x_i^2}{2\eta_i} + \frac{\eta_i}{2}\Big).$$

  Here the optimization variables are $\mathbf{x} \in \mathbb{R}^n$ and $\eta = [\eta_1, \dots, \eta_n] \in \mathbb{R}^n$.

- The new problem has two variables, it can solved using coordinate descent.

Solving the new problem

$$\min_{\eta} \min_{\mathbf{x}} \ f(\mathbf{x}, \eta) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\Big(\sum_{i=1}^{n} \frac{x_i^2}{2\eta_i} + \frac{\eta_i}{2}\Big).$$

- The optimum of f with respect to η_i is $|x_i|$.
- f is convex in η_i so the update $\eta_i = |x_i|$ gives the global minimizer.
- To derive the optimal $\mathbf{x}$ of f, we can use the 1st-order optimality condition: $\nabla_{\mathbf{x}} f = 0$. First, the gradient is

$$\begin{aligned} \nabla_{\mathbf{x}} f &= \mathbf{A}^\top \mathbf{A}\mathbf{x} - \mathbf{A}^\top \mathbf{b} + \lambda \mathsf{Diag}(\eta)^{-1}\mathbf{x} \\ &= \Big(\mathbf{A}^\top \mathbf{A} + \lambda \mathsf{Diag}(\eta)^{-1}\Big)\mathbf{x} - \mathbf{A}^\top \mathbf{b}. \end{aligned}$$

  Then $\nabla_{\mathbf{x}} f = 0$ gives

$$\mathbf{x} = \Big(\mathbf{A}^\top \mathbf{A} + \lambda \mathsf{Diag}(\eta)^{-1}\Big)^{-1} \mathbf{A}^\top \mathbf{b}.$$

  As f is convex in $\mathbf{x}$, this gives the global minimizer.

The two-line algorithm

Algorithm 1: RLS: Reweighted Least Squares

Result: Solution to L_1-regularized least sqaures

Initialize $\eta_i = |x_i|$;

for $k = 1, 2, \ldots$ **do**

$$\mathbf{x} = \left(\mathbf{A}^\top \mathbf{A} + \lambda \text{Diag}(\eta)^{-1}\right)^{-1} \mathbf{A}^\top \mathbf{b};$$

$\quad \eta_i = |x_i|$;

end

Note : the terms $\mathbf{A}^\top \mathbf{A}$, $\mathbf{A}^\top b$ should be precomputed before the main loop.

RLS compared with ISTA and FISTA

RLS may perform worse than proximal gradient methods as RLS does not have a "hard projection step".

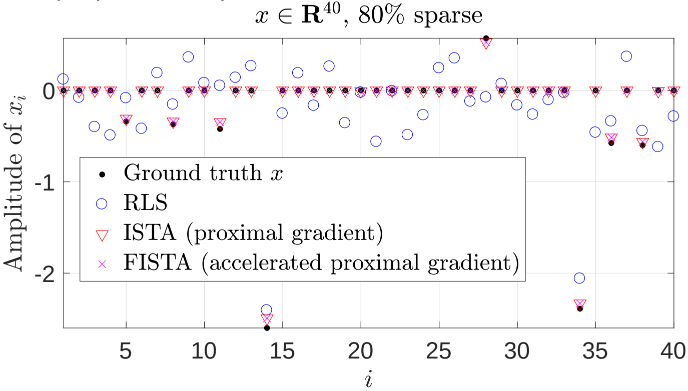


Figure: x produced by the algorithms with same initialization and number of iterations

Links to the details of proximal gradient algorithm, convergence of ISTA and FISTA.

RLS compared with ISTA and FISTA

RLS may perform worse than proximal gradient methods as RLS does not have a "hard projection step".

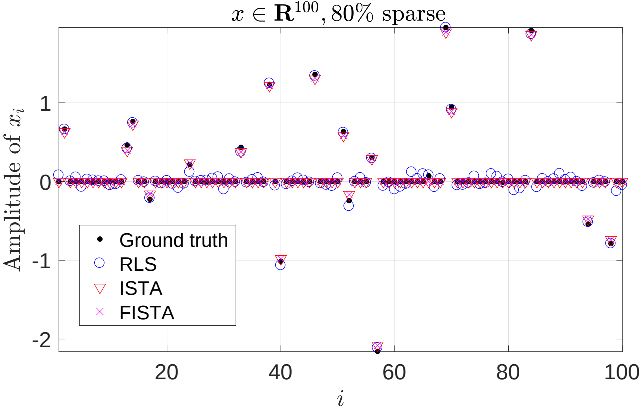


Figure: x produced by the algorithms with same initialization and number of iterations

Links to the details of proximal gradient algorithm, convergence of ISTA and FISTA.

Last page - summary

- Approximating absolute value by a parametric quadratic function

$$\phi(x;\eta) = \frac{1}{2\eta}x^2 + \frac{1}{2}\eta, \quad \eta \geq 0.$$

- Application to L_1-regularized least squares
- RLS algorithm

Reference : Francis Bach's blog

End of document