(Graphical) Derivation of the soft thresholding operator

Andersen Ang

Mathématique et recherche opérationnelle
UMONS, Belgium

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

First draft: August, 6, 2018
Last update : October 16, 2019

L_1 regularized least square

Given $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$, find $\mathbf{x} \in \mathbb{R}^n$ by solving

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

A regularized least square problem :

- $\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is the "data fitting" term
- $\frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is differentiable, so this part can be handled by simple method like gradient descent
- $\lambda\|\mathbf{x}\|_1$ is the regularizer, $\lambda \geq 0$ is regularization parameter
- The non-differentiable L_1 norm $\|\mathbf{x}\|_1$ is a sparsity inducing regularizer, it promotes sparsity of solution $\mathbf{x}$

Solving the L_1 regularized least square

Recall here : if no regularization, the gradient descent step can be views as the minimizer of a local quadratic model of f in the form as

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \left\{ \frac{1}{2t_k} \|\mathbf{x} - \big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\big)\|_2^2 \right\}$$

where

- $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$
- $t_k > 0$ is a suitable stepsize
- $\nabla f(\mathbf{x}_k)$ is gradient of f with respect to (w.r.t) $\mathbf{x}_k$
- k is iteration counter, $k = 1, 2, ...$

With the L_1 regularization term, the above problem is changed to

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}} \left\{ \underbrace{\frac{1}{2t_k} \|\mathbf{x} - \big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\big)\|_2^2 + \lambda \|\mathbf{x}\|_1}_{F} \right\}$$

Note that F is convex as it is sum of two convex functions, and F is "separable" : $F(\mathbf{x}) = \sum F_i(\mathbf{x}_i)$ and each F_i is a scalar problem

Separable problem

Let $\mathbf{y} = \mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)$, we have

$$\begin{aligned} \mathbf{x}_{k+1} &= \arg\min_{\mathbf{x}} \left\{ \frac{1}{2t_k} \|\mathbf{x} - \big(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)\big)\|_2^2 + \lambda \|\mathbf{x}\|_1 \right\} \\ &= \arg\min_{\mathbf{x}} \left\{ \frac{1}{2t_k} \|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{x}\|_1 \right\} \end{aligned}$$

Recall $\|\mathbf{x}\|_2^2 = \sum_i x_i^2, \ \|\mathbf{x}\|_1 = \sum_i |x_i|$ so

$$\frac{1}{2t_k} \|\mathbf{x} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{x}\|_1 = \sum_i \frac{1}{2t_k} (x_i - y_i)^2 + \lambda |x_i|$$

We have a coordinate wise scalar expression

$$x_i = \arg\min_{x} \left\{ \frac{1}{2t_k} (x - y_i)^2 + \lambda |x| \right\}$$

i.e. given y_i, find the minimizer x of the scalar function $\dfrac{1}{2t_k}(x - y_i)^2 + \lambda |x|$

The scalar problem

Consider the scalar function

$$f(x) = \frac{1}{2t}(x-y)^2 + \lambda|x|$$

It is convex :

- $|x|$ is not differentiable but convex
- $(x-y)^2$ is differentiable and convex

The minimizer of f, denoted as x^*, can be obtained by considering the optimality condition. As $|x|$ is not differentiable, we use the optimality condition of subgradient.

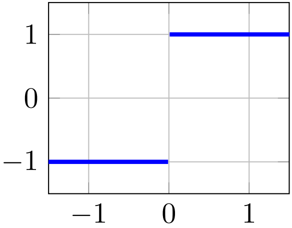
The minimum of a nondifferentiable function. A point x^* is a minimizer of a convex function f if and only if f is subdifferentiable at x^* and

$$0 \in \partial f(x^*).$$

Subgradients optimality condition

For the scalar function $f(x) = \dfrac{1}{2t_k}(x-y)^2 + \lambda|x|$

- The subdifferential of $|x|$ is $\text{sgn}(x)$



- The gradient of $(x-y)^2$ is $2(x-y)$

Hence $0 \in \partial f(x^*)$ becomes

$$0 \in \frac{1}{t}(x^* - y) + \lambda \text{sgn}(x^*)$$

For simplicity let $t = 1$, we have

$$0 \in x^* - y + \lambda \text{sgn}(x^*) \iff y = x^* + \lambda \text{sgn}(x^*)$$

Expressing x^* as a function of y

- Equation $y = x^* + \lambda \text{sgn}(x^*)$ expresses y as a function of x^*
- As y is given and we want to find x^*. The goal is to express x^* as a function of y.
- This can be done by swapping the xy axes of the plot of $y = x + \lambda \text{sgn}(x)$. For simplicity let $\lambda = 1$, then

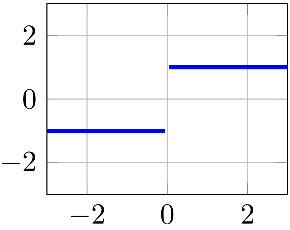


Figure: $\text{sgn}(x)$

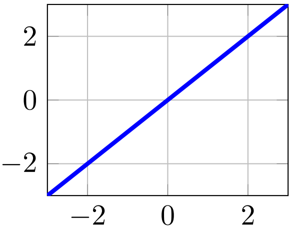


Figure: x

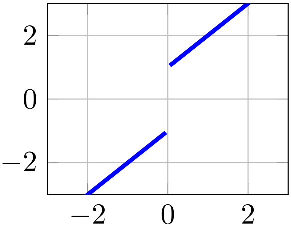
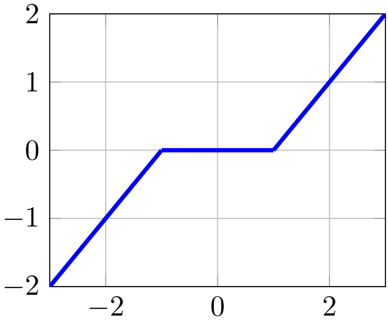


Figure: $x + \lambda \text{sgn}(x), \lambda = 1$

The soft thresholding operator

Swaps the axes, we get the soft thresholding operator $\mathcal{T}$



$$\mathcal{T}(x) = \mathsf{sgn}(x)(|x| - 1)_+ = \mathsf{sgn}(x)\max(|x| - 1, 0)$$

In general case with threshold λ we have

$$\mathcal{T}_\lambda(x) = \mathsf{sgn}(x)(|x| - \lambda)_+$$

In MATLAB : `sign(x).*(max(abs(x)-lambda,0));`

Iterative Shrinkage Thresholding Algorithm (ISTA)

The L_1-regularized least square

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

Can be solved by the following update rule

$$\mathbf{x}_{k+1} = \mathcal{T}(\mathbf{x}_k)$$

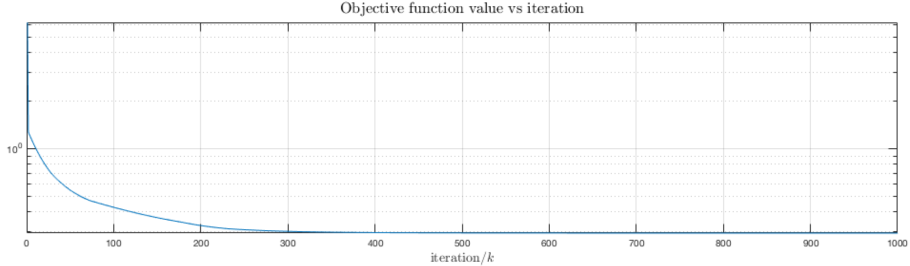
where $\mathcal{T}$ is the soft thresholding operator applied on $\mathbf{x}$ componentwise

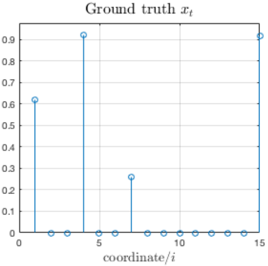
$$[\mathcal{T}(\mathbf{x}_k)]_i = \mathsf{sgn}\Big([\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)]_i\Big)\Big(|[\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)]_i| - \lambda t_k\Big)_+$$

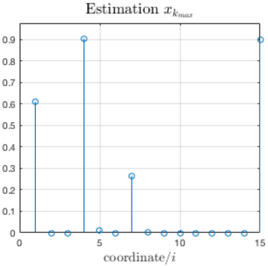
The algorithm using such update rule is called *Iterative Shrinkage Thresholding Algorithm* (ISTA).

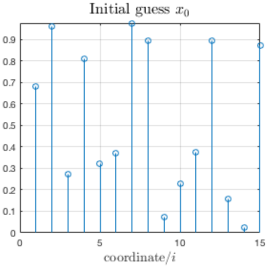
- Note. In the view point of proximal operator, ISTA is an example of proximal gradient update. ISTA is nothing no more than just the proximal gradient update applied on the L_1-regularized least square problem.

Illustration (MATLAB code)









End of document