Optimization Theory and Algorithm	**Lecture 12 - 06/04/2021**

Lecture 12

Lecturer:Xiangyu Chang *Scribe: Xiangyu Chang*

Edited by: Xiangyu Chang

Example 1 *Let us consider the LASSO problem again:*

$$\min_{\mathbf{x}} \left\{ \frac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_1 \right\}. \tag{1}$$

- $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|_2^2$ *is convex and* β*-smooth.*
- $b(\mathbf{x}) = \lambda\|\mathbf{x}\|_1$ *is convex and non-smooth.*
- *Proximal Operator:*

$$\begin{aligned} prox_{\gamma\|\mathbf{x}\|_1}(\mathbf{z}) &= \arg\min_{\mathbf{x}} \left\{ \frac{1}{2\gamma}\|\mathbf{x}-\mathbf{z}\|_2^2 + \lambda\|\mathbf{x}\|_1 \right\} & (2)\\ &= \arg\min_{\mathbf{x}} \sum_{i=1}^{n} \left\{ \frac{1}{2\gamma}(x_i - z_i)^2 + \lambda|x_i| \right\}. & (3) \end{aligned}$$

- *For each* $i = 1, \ldots, n$*, we need to solve*

$$\arg\min_{x_i} \left\{ \frac{1}{2\gamma}(x_i - z_i)^2 + \lambda|x_i| \right\} := \phi(x_i). \tag{4}$$

If $x_i > 0$*, then set* $\phi'(x_i) = 0$ *it has* $x_i^* = \mathbf{z}_i - \gamma\lambda$*. If* $x_i < 0$*, then* $x_j^* = \mathbf{z}_i + \gamma\lambda$*. If* $x_j = 0$*, then we need* $0 \in \partial\phi(0)$*. So,* $0 \in \frac{1}{\gamma}(0 - z_i) + \lambda\partial|0|$*. So,* $\frac{z_i}{\gamma\lambda} \in \partial|0| = [-1, 1]$*. Thus,*

$$x_i = \begin{cases} \mathbf{z}_i - \gamma\lambda & z_i > \gamma\lambda, \\ 0, & |z_i| \le \gamma\lambda, \\ \mathbf{z}_i + \gamma\lambda & z_i < -\gamma\lambda. \end{cases} \tag{5}$$

This is called the soft thresholding function.

- $prox_{\gamma\|\mathbf{x}\|_1}(\mathbf{z}) = sign(\mathbf{z})(|\mathbf{z}| - \gamma\lambda)_+$.
- *Go back to LASSO.* $\beta = \lambda_{\max}(A^\top A)$, $\nabla f(\mathbf{x}) = A^\top(A\mathbf{x} - \mathbf{b})$.
- *Algorithm:*

$$\begin{aligned} \mathbf{z}^t &= \mathbf{x}^t - \frac{1}{\lambda_{\max}(A^\top A)} A^\top(A\mathbf{x} - \mathbf{b}) = \left(I - \frac{A^\top A}{\lambda_{\max}(A^\top A)}\right)\mathbf{x}^t + \frac{A^\top \mathbf{b}}{\lambda_{\max}(A^\top A)}, \\ \mathbf{x}^{t+1} &= prox_{\frac{\lambda}{\lambda_{\max}(A^\top A)}\|\mathbf{x}\|_1}(\mathbf{z}^t) = sign(\mathbf{z}^t)\left(|\mathbf{z}^t| - \frac{\lambda}{\lambda_{\max}(A^\top A)}\right)_+. \end{aligned}$$

0.1 Convergence Theory

Let us define $\mathbf{x}^+ = prox_{\gamma g}(\mathbf{x} - \gamma\nabla f(\mathbf{x}))$ and $G_\gamma(\mathbf{x}) = \frac{1}{\gamma}(\mathbf{x} - \mathbf{x}^+)$.

Insights: if $\mathbf{x}^+ = \mathbf{x} - \gamma\nabla f(\mathbf{x})$, then $\frac{1}{\gamma}(\mathbf{x} - \mathbf{x}^+) = \nabla f(\mathbf{x})$. We hope G_γ has similar behaviours with $\nabla f(\mathbf{x})$.

Lemma 1 *let* $h(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$*,* $g(\mathbf{x})$ *is convex and non-smooth,* $f(\mathbf{x})$ *is smooth.* $\mathbf{x}^*$ *is a local minimal point of* h*, then*

$$-\nabla f(\mathbf{x}^*) \in \partial g(\mathbf{x}^*). \tag{6}$$

Lemma 2 $G_\gamma(\mathbf{x}) = 0$ *if and only if* $0 \in \partial h(\mathbf{x})$*. This implies* $\mathbf{x}$ *is a global minimum.*

Proof 1 *We know that* $\mathbf{x}^+$ *minimizes*

$$\frac{1}{2\gamma}\|\mathbf{z} - (x - \gamma f(\mathbf{x}))\|^2 + g(\mathbf{z})$$

by definition of proximal operator. In terms of optimality conditions for this problem, this means

$$0 \in \frac{1}{\gamma}(\mathbf{x}^+ - (x - \gamma f(\mathbf{x})) + \partial g(\mathbf{x}^+) = -G_\gamma(\mathbf{x}) + f(\mathbf{x}) + \partial g(\mathbf{x}^+)$$

or equivalently $G_\gamma(\mathbf{x}) \in f(\mathbf{x}) + \partial g(\mathbf{x}^+)$*. If* $\mathbf{x} = \mathbf{x}^+$*, and hence* $G_\gamma(\mathbf{x}) = G_\gamma(\mathbf{x}^+) = 0$*, this implies*

$$0 \in f(\mathbf{x}^+) + \partial g(\mathbf{x}^+) = \partial f(\mathbf{x}^+) + \partial g(\mathbf{x}^+) = \partial h(\mathbf{x}^+)$$

so $\mathbf{x}^+$ *is a minimizer of* h*.*

References