EE 588: Optimization for the information and data sciences

University of Southern California

Lecturer: Mahdi Soltanolkotabi November 2018

Lecture on Proximal Methods

1 Introduction

From last lectures, we know that

- For non-smooth, but Lipshitz functions $\implies$ Convergence result: $\frac{1}{\sqrt{t}}$
- For smooth functions $\implies$ Convergence result: $\frac{1}{t^2}$

Question: How do we optimize an objective of the following form?

$$f(\boldsymbol{x}) + g(\boldsymbol{x}) \tag{1}$$

where $f(\boldsymbol{x})$ is a smooth and $g(\boldsymbol{x})$ is a non-smooth function.
One answer: Proximal mapping algorithms.

2 Proximal Mapping

2.1 Extended Real-Valued Functions

Proximal mapping methods work with extended real-valued functions. For a given function $f(\boldsymbol{x})$, its extended real-valued function is

$$\tilde{f}(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & \boldsymbol{x} \in \text{dom } f \\ \infty & \boldsymbol{x} \notin \text{dom } f \end{cases} \tag{2}$$

So, $\tilde{f}(\boldsymbol{x})$ has an extended domain and range that is $\tilde{f} : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$.

For example, the extended indicator function is

$$\mathbb{I}_\Omega(\boldsymbol{x}) = \begin{cases} 1 & \boldsymbol{x} \in \Omega \\ \infty & \boldsymbol{x} \notin \Omega \end{cases} . \tag{3}$$

2.2 Proximal mappings associated with convex function

Let g be an extended real-valued convex function on $\mathbb{R}^n$, the proximal mapping is defined as

$$\text{prox}_g(\boldsymbol{x}) = \underset{\boldsymbol{y}}{\text{argmin}} \ \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}^2 + g(\boldsymbol{y}) \tag{4}$$

where $\text{prox}_g(\boldsymbol{x})$ is also known as proximal operator. One interpretation of the proximal operator is that it smooths the function $g(\boldsymbol{x})$.

Some properties of proximal mapping:

- It is a function.
- It is strongly convex $\implies$ a unique optimal solution.
- Subgradient characterization: $\text{prox}_g(\boldsymbol{x}) \iff \boldsymbol{x} - \partial g(\boldsymbol{y})$

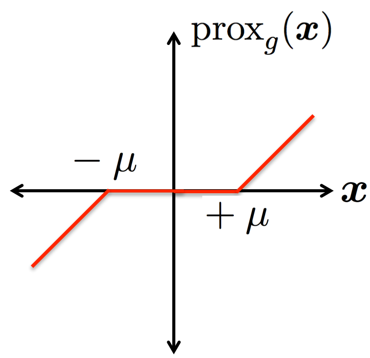
Examples:

- $g(\boldsymbol{x}) = 0$: $\text{prox}_g(\boldsymbol{x}) = x$.
- $\mathbb{I}_{\mathcal{C}}(\boldsymbol{x})$ where $\mathcal{C}$ is a convex set: $\text{prox}_{\mathbb{I}_{\mathcal{C}}}(\boldsymbol{x}) = \underset{\boldsymbol{y}}{\text{argmin}}\ \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell_2}^2 + \mathbb{I}_{\mathcal{C}}(\boldsymbol{y}) = \underset{\boldsymbol{y}\in\mathcal{C}}{\text{argmin}}\ \|\boldsymbol{x}-\boldsymbol{y}\|_{\ell_2}^2$

  $= \mathcal{P}_{\mathcal{C}}(\boldsymbol{x})$ (projection on set $\mathcal{C}$)
- $\mathbb{I}_{\mathcal{R}+}(\boldsymbol{x})$: $(\text{prox}_{\mathbb{I}_{\mathcal{R}+}}(\boldsymbol{x}))_i = \begin{cases} (\boldsymbol{x})_i + \mu & (\boldsymbol{x})_i \leq -\mu \\ 0 & (\boldsymbol{x})_i < 0 \end{cases} = \max((\boldsymbol{x})_i, 0)$

  where $(\boldsymbol{x})_i$ denotes i-th element of $\boldsymbol{x}$.

  i.e., proximal operator sets all negative entries to zero and keeps other entries the same.
- $g(\boldsymbol{x}) = \frac{\mu}{2}||\boldsymbol{x}||_2^2$: $\text{prox}_g(\boldsymbol{x}) = \frac{x}{1+\mu}$.
- $g(\boldsymbol{x}) = \mu||\boldsymbol{x}||_1$: $\text{prox}_g(\boldsymbol{x}) = \begin{cases} (\boldsymbol{x})_i + \mu & (\boldsymbol{x})_i \leq -\mu \\ 0 & |(\boldsymbol{x})_i| < \mu \\ (\boldsymbol{x})_i - \mu & (\boldsymbol{x})_i \geq +\mu \end{cases}$ (see figure below)



Lemma 1 *If* $\boldsymbol{u} = \text{prox}_g(\boldsymbol{x})$ *and* $\boldsymbol{v} = \text{prox}_g(\boldsymbol{y})$ *then* $(\boldsymbol{u}-\boldsymbol{v})^T(\boldsymbol{x}-\boldsymbol{y}) \geq \|\boldsymbol{u}-\boldsymbol{v}\|_{\ell_2}^2$.

Lemma 2 $\|\text{prox}_g(\boldsymbol{x}) - \text{prox}_g(\boldsymbol{x})\|_{\ell_2} \leq \|\boldsymbol{x}-\boldsymbol{y}\|_{\ell_2}$.

2.3 Two Proximal Mapping Algorithms: ISTA and FISTA

Iterative Shrinkage Thresholding Algorithm (ISTA)
Goal: Minimizing $f(\boldsymbol{x}) + g(\boldsymbol{x})$ where $f(\boldsymbol{x})$: smooth and $g(\boldsymbol{x})$: non-smooth

1. Pick step size μ and initial guess $\boldsymbol{x}_0$
2. Repeat for $t \geq 0$

$$\boldsymbol{x}_{t+1} = \operatorname{prox}_{\mu g}\left(\boldsymbol{x}_t - \mu \nabla f(\boldsymbol{x}_t)\right). \tag{5}$$

How did we get iteration stated in (5)?
Gradient descend iteration for minimizing function $f(\boldsymbol{x})$:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \mu \nabla f(\boldsymbol{x}_t) \tag{6}$$

Equivalently written in proximal mapping form:

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \underset{\boldsymbol{y}}{\operatorname{argmin}}\ \mu \nabla f(\boldsymbol{x}_t)^T(\boldsymbol{y} - \boldsymbol{x}_t) + \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{x}_t\|_{\ell_2}^2 \\ &= \underset{\boldsymbol{y}}{\operatorname{argmin}}\ \nabla f(\boldsymbol{x}_t)^T(\boldsymbol{y} - \boldsymbol{x}_t) + \frac{1}{2\mu}\|\boldsymbol{y} - \boldsymbol{x}_t\|_{\ell_2}^2 \end{aligned} \tag{7}$$

For minimizing $f(\boldsymbol{x}) + g(\boldsymbol{x})$ we can write:

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \underset{\boldsymbol{y}}{\operatorname{argmin}}\ g(\boldsymbol{y}) + \nabla f(\boldsymbol{x}_t)^T(\boldsymbol{y} - \boldsymbol{x}_t) + \frac{1}{2\mu}\|\boldsymbol{y} - \boldsymbol{x}_t\|_{\ell_2}^2 \\ &= \underset{\boldsymbol{y}}{\operatorname{argmin}}\ g(\boldsymbol{y}) + \frac{1}{2\mu}\|\boldsymbol{y} - (\boldsymbol{x}_t - \mu \nabla f(\boldsymbol{x}_t))\|_{\ell_2}^2 \\ &= \operatorname{prox}_{\mu g}\left(\boldsymbol{x}_t - \mu \nabla f(\boldsymbol{x}_t)\right). \end{aligned} \tag{8}$$

Theorem 3 *If $f(\boldsymbol{x})$ is L-smooth and $g(\boldsymbol{x})$ is closed and convex, for a fixed step size $\mu = \frac{1}{L}$, we have the following convergence result for ISTA algorithm,*

$$(f(\boldsymbol{x}_t) + g(\boldsymbol{x}_t)) - (f(\boldsymbol{x}^*) + g(\boldsymbol{x}^*)) \leq \frac{L\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_{\ell_2}^2}{2t} \tag{9}$$

Proof: See Theorem 3.1 in [**?**].

Theorem 4 *If $f(\boldsymbol{x})$ is m-convex and L-smooth, and $g(\boldsymbol{x})$ is closed and convex, for a fixed step size $\mu < \frac{2}{L+m}$, we have the following convergence result for ISTA algorithm,*

$$\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|_{\ell_2} \leq \left(\frac{L-m}{L+m}\right)\|\boldsymbol{x}_t - \boldsymbol{x}^*\|_{\ell_2}. \tag{10}$$

Fast Iterative Shrinkage Thresholding Algorithm (FISTA)
Goal: Minimizing $f(\boldsymbol{x}) + g(\boldsymbol{x})$ where $f(\boldsymbol{x})$: smooth and $g(\boldsymbol{x})$: non-smooth

1. Pick step size μ and initial guess $\boldsymbol{x}_0 = \boldsymbol{x}_{-1}$
2. Repeat for $t \geq 0$

$$\begin{aligned} \beta &= \frac{t-1}{t+2} \\ \boldsymbol{z}_t &= \boldsymbol{x}_t - \beta(\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) \\ \boldsymbol{x}_{t+1} &= \operatorname{prox}_{\mu g}\left(\boldsymbol{z}_t - \mu \nabla f(\boldsymbol{z}_t)\right). \end{aligned} \tag{11}$$

Convergence result for FISTA: $(f(\boldsymbol{x}_t) + g(\boldsymbol{x}_t)) - (f(\boldsymbol{x}^*) + g(\boldsymbol{x}^*))$ decreases as fast as $\frac{1}{t^2}$.