

Lecture 12: CS395T Numerical Optimization for Graphics and AI — Proximal Gradient Descents

Qixing Huang
The University of Texas at Austin
huangqx@cs.utexas.edu

1 Introduction

1.1 Proximal Mapping

The proximal mapping of a (convex) function $h(\mathbf{x})$ is given by

$$\text{prox}_h(\mathbf{x}) = \underset{\mathbf{u}}{\operatorname{argmin}} \left(h(\mathbf{u}) + \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|^2 \right).$$

The following are some examples:

- When $h(\mathbf{x}) = 0$, then $\text{prox}_h(\mathbf{x}) = \mathbf{x}$.
- When $h(\mathbf{x}) = Id_C$, where Id_C is the indicator function on C . Then

$$\text{prox}_h(\mathbf{x}) = \underset{\mathbf{u} \in C}{\operatorname{argmin}} \|\mathbf{u} - \mathbf{x}\|^2.$$

- When $h(\mathbf{x}) = t\|\mathbf{x}\|_1$ for some positive $t > 0$, then $\text{prox}_h(\mathbf{x})$ is a shrinkage operator defined as

$$\text{prox}_h(\mathbf{x})_i = \begin{cases} x_i - t & x_i > t \\ 0 & |x_i| \leq t \\ x_i + t & x_i < -t \end{cases}$$

1.2 Proximal Gradient Method

We are interested in minimizing an objective function of the following form:

$$f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x}).$$

Here $g(\mathbf{x})$ is a nice convex objective function, e.g., smooth and easy to optimize. $h(\mathbf{x})$ is also convex, but it maybe not that nice, e.g., non-differentiable and non-smooth. However, we assume optimizing $\text{prox}_h(\mathbf{x})$ is inexpensive. One such example is

$$\underset{\mathbf{x}}{\operatorname{argmin}} \|A\mathbf{x} - \mathbf{b}\|^2 + \lambda\|\mathbf{x}\|_1$$

Proximal gradient methods admit the following form:

$$\mathbf{x}^{(k)} = \text{prop}_{t_k h} \left(\mathbf{x}^{(k-1)} - t_k \nabla g(\mathbf{x}^{(k-1)}) \right),$$

where t_k is called the step-size, which is either a constant or determined by line-search.

We can understand the proximal update as follows:

$$\begin{aligned}
\mathbf{x}^{(k)} &= \text{prop}_{t_k h} \left(\mathbf{x}^{(k-1)} - t_k \nabla g(\mathbf{x}^{(k-1)}) \right) \\
&= \underset{\mathbf{u}}{\text{argmin}} \ h(\mathbf{u}) + \frac{1}{2t} \|\mathbf{u} - \mathbf{x}^{(k-1)} + t \nabla g(\mathbf{x}^{(k-1)})\|^2 \\
&= \underset{\mathbf{u}}{\text{argmin}} \ (h(\mathbf{u}) + g(\mathbf{x}^{(k-1)}) + \nabla g(\mathbf{x}^{(k-1)})^T (\mathbf{u} - \mathbf{x}^{(k-1)}) + \frac{1}{2t} \|\mathbf{u} - \mathbf{x}^{(k-1)}\|^2)
\end{aligned}$$

In other words, $\mathbf{x}^{(k)}$ minimizes $h(\mathbf{u})$ and a local quadratic approximation of $g(\mathbf{u})$ in the neighborhood of $\mathbf{x}^{(k-1)}$.

Example I. When $h(\mathbf{x}) = 0$, Then proximal gradient method becomes gradient method:

$$\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - t_k \nabla g(\mathbf{x}^{(k-1)}).$$

Example II. When $h(\mathbf{x}) = Id_C$, then

$$\mathbf{x}^{(k)} = \text{proj}_C(\mathbf{x}^{(k-1)} - t_k \nabla g(\mathbf{x}^{(k-1)})).$$

Example III. Interactive soft-thresholding where $h(\mathbf{x}) = \|\mathbf{x}\|_1$, i.e., minimize $g(\mathbf{x}) + \|\mathbf{x}\|_1$:

$$\mathbf{x}^{(k)} = \text{prox}_{t_k h}(\mathbf{x}^{(k-1)} - t_k \nabla g(\mathbf{x}^{(k-1)})),$$

where

$$\text{prox}_{th}(u)_i = \begin{cases} x_i - t & x_i \geq t \\ 0 & |x_i| \leq t \\ x_i + t & x_i \leq -t \end{cases}$$

2 Proximal Gradient Algorithm

The proximal gradient iteration can be written as $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} - t_k G_{t_k}(\mathbf{x}^{(k-1)})$ where

$$G_t(\mathbf{x}) = \frac{1}{t} \left(\mathbf{x} - \text{prox}_{th}(\mathbf{x} - t \nabla g(\mathbf{x})) \right).$$

from sub-gradient definition of prox

$$G_t(\mathbf{x}) \in \nabla g(\mathbf{x}) + \partial h(\mathbf{x} - t G_t(\mathbf{x})).$$

In other words, $G_t(\mathbf{x}) = 0$ if and only of \mathbf{x} minimizes $f(\mathbf{x}) = g(\mathbf{x}) + h(\mathbf{x})$.

To determine stepsize t in

$$\mathbf{x}^+ = \mathbf{x} - t G_t(\mathbf{x}).$$

Start with some $t = \hat{t}$; repeat $t := \beta t$ (with $0 < \beta < 1$) until

$$g(\mathbf{x} - t G_t(\mathbf{x})) \leq g(\mathbf{x}) - t \nabla g(\mathbf{x})^T G_t(\mathbf{x}) + \frac{t}{2} \|G_t(\mathbf{x})\|^2.$$

The inequality is motivated from the convergence analysis, which will be described next.

3 Convergence of Proximal Gradient Method

Assumptions.

- $\nabla g(\mathbf{x})$ is Lipschitz continuous with constant $L > 0$

$$\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y}$$

- optimal value f^* is finite and attained at \mathbf{x}^* (not necessarily unique).

Claim. We show that $f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*)$ decreases at least as fast as $1/k$ if

- if step size $t_k = 1/L$ is used
- if backtrack line search is used

Proof. To prove this we will start with some properties regarding $g(\mathbf{x})$:

- affine lower bound from convexity:

$$g(\mathbf{y}) \geq g(\mathbf{x}) + \nabla g(\mathbf{x})^T(\mathbf{y} - \mathbf{x}).$$

- quadratic upper bound from Lipschitz property

$$g(\mathbf{y}) \leq g(\mathbf{x}) + \nabla g(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2.$$

Let $\mathbf{v} = \mathbf{y} - \mathbf{x}$. The proof of this straight-forward using

$$g(\mathbf{y}) - g(\mathbf{x}) = \nabla g(\mathbf{x})^T \mathbf{v} + \int_0^1 (\nabla g(\mathbf{x} + t\mathbf{v}) - \nabla g(\mathbf{x}))^T \mathbf{v} dt.$$

A consequence of this is that the line search inequality

$$g(\mathbf{x} - tG_t(\mathbf{x})) \leq g(\mathbf{x}) - t\nabla g(\mathbf{x})G_t(\mathbf{x}) + \frac{t}{2}\|G_t(\mathbf{x})\|^2 \quad (1)$$

is satisfied for $0 \leq t \leq \frac{1}{L}$. This means back-tracking at \hat{t} terminates at $t \geq \min(\hat{t}, \beta/L)$.

If the line search inequality (1) holds, then for all \mathbf{z} ,

$$f(\mathbf{x} - tG_t(\mathbf{x})) \leq f(\mathbf{z}) + G_t(\mathbf{x})^T(\mathbf{x} - \mathbf{z}) - \frac{t}{2}\|G_t(\mathbf{x})\|^2. \quad (2)$$

Using (2), we can obtain the progress in one iteration:

$$f(\mathbf{x}^+) - f(\mathbf{x}^*) \leq \frac{1}{2t}(\|\mathbf{x} - \mathbf{x}^*\|^2 - \|\mathbf{x}^+ - \mathbf{x}^*\|^2).$$

Analysis with fixed stepsize. We will show that $f(\mathbf{x}^{(k)}) - f(\mathbf{x}^*) = O(\frac{1}{k})$. In fact,

$$\begin{aligned} \sum_{i=1}^k (f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) &\leq \sum_{i=1}^k \frac{1}{2t} (\|\mathbf{x}^{(i-1)} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{(i)} - \mathbf{x}^*\|^2) \\ &\leq \frac{1}{2t} \sum_{i=1}^k (\|\mathbf{x}^{(i-1)} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{(i)} - \mathbf{x}^*\|^2) \\ &= \frac{1}{2t} (\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2). \end{aligned}$$

Since $f(\mathbf{x}^{(i)})$ is non-increasing,

$$f(\mathbf{x}^{(k)}) - f^* \leq \frac{1}{2kt} \|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2.$$

This means when $t = \frac{1}{L}$, $f(\mathbf{x}^{(k)}) - f^* = O(\frac{L}{2k})$.

Analysis with line search. The derivation is quite similar:

$$\begin{aligned}
\sum_{i=1}^k (f(\mathbf{x}^{(i)}) - f(\mathbf{x}^*)) &\leq \sum_{i=1}^k \frac{1}{2t_i} (\|\mathbf{x}^{(i-1)} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{(i)} - \mathbf{x}^*\|^2) \\
&\leq \frac{1}{2t_{\min}} \sum_{i=1}^k (\|\mathbf{x}^{(i-1)} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{(i)} - \mathbf{x}^*\|^2) \\
&= \frac{1}{2t_{\min}} (\|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2 - \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2).
\end{aligned}$$

Since $f(\mathbf{x}^{(i)})$ is non-increasing,

$$f(\mathbf{x}^{(k)}) - f^* \leq \frac{1}{2kt_{\min}} \|\mathbf{x}^{(0)} - \mathbf{x}^*\|^2.$$

□