Lecture 17: 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 $h(\boldsymbol{x})$ is given by

$$\mathrm{prox}_h(\boldsymbol{x}) = \underset{\boldsymbol{u}}{\mathrm{argmin}}\ \big(h(\boldsymbol{u}) + \frac{1}{2}\|\boldsymbol{u}-\boldsymbol{x}\|^2\big).$$

The following are some examples:

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

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

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

$$\mathrm{prox}_h(\boldsymbol{x})_i = \begin{cases} x_i - t & x_i > t \\ 0 & |x_i| \le t \\ x_i + t & x_i < -t \end{cases}$$

1.2 Proximal Gradient Method

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

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

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

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

Proximal gradient methods admit the following form:

$$\boldsymbol{x}^{(k)} = \mathrm{prop}_{t_k h}\Big(\boldsymbol{x}^{(k-1)} - t_k\nabla g(\boldsymbol{x}^{(k-1)})\Big),$$

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:

$$\begin{aligned}
\boldsymbol{x}^{(k)} &= \mathrm{prop}_{t_k h}\Big(\boldsymbol{x}^{(k-1)} - t_k \nabla g(\boldsymbol{x}^{(k-1)})\Big) \\
&= \operatorname*{argmin}_{\boldsymbol{u}} \; h(\boldsymbol{u}) + \frac{1}{2t}\|\boldsymbol{u} - \boldsymbol{x}^{(k-1)} + t\nabla g(\boldsymbol{x}^{(k-1)})\|^2 \\
&= \operatorname*{argmin}_{\boldsymbol{u}} \; \big(h(\boldsymbol{u}) + g(\boldsymbol{x}^{(k-1)}) + \nabla g(\boldsymbol{x}^{(k-1)})^T(\boldsymbol{u} - \boldsymbol{x}^{(k-1)}) + \frac{1}{2t}\|\boldsymbol{u} - \boldsymbol{x}^{(k-1)}\|^2\big)
\end{aligned}$$

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

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

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

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

$$\boldsymbol{x}^{(k)} = \mathrm{proj}_C\big(\boldsymbol{x}^{(k-1)} - t_k \nabla g(\boldsymbol{x}^{(k-1)})\big).$$

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

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

where

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

2 Proximal Gradient Algorithm

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

$$G_t(\boldsymbol{x}) = \frac{1}{t}\Big(\boldsymbol{x} - \mathrm{prox}_{th}\big(\boldsymbol{x} - t\nabla g(\boldsymbol{x})\big)\Big).$$

from sub-gradient definition of prox

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

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

To determine stepsize t in

$$\boldsymbol{x}^+ = \boldsymbol{x} - tG_t(\boldsymbol{x}).$$

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

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

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

3 Convergence of Proximal Gradient Method

Assumptions.

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

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

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

Claim. We show that $f(\boldsymbol{x}^{(k)}) - f(\boldsymbol{x}^\star)$ 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 $g(\boldsymbol{x})$:

- affine lower bound from convexity:

$$g(\boldsymbol{y}) \ge g(\boldsymbol{x}) + \nabla g(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}).$$

- quadratic upper bound from Lipschitz property

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

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

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

A consequence of this is that the line search inequality

$$g(\boldsymbol{x} - tG_t(\boldsymbol{x})) \le g(\boldsymbol{x}) - t\nabla g(\boldsymbol{x})G_t(\boldsymbol{x}) + \frac{t}{2}\|G_t(\boldsymbol{x})\|^2 \tag{1}$$

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

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

$$f(\boldsymbol{x} - tG_t(\boldsymbol{x})) \le f(\boldsymbol{z}) + G_t(\boldsymbol{x})^T(\boldsymbol{x} - \boldsymbol{z}) - \frac{t}{2}\|G_t(\boldsymbol{x})\|^2. \tag{2}$$

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

$$f(\boldsymbol{x}^+) - f(\boldsymbol{x}^\star) \le \frac{1}{2t}\big(\|\boldsymbol{x} - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}^+ - \boldsymbol{x}^\star\|^2\big).$$

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

$$\begin{aligned}
\sum_{i=1}^k (f(\boldsymbol{x}^{(i)}) - f(\boldsymbol{x}^\star)) &\le \sum_{i=1}^k \frac{1}{2t}\big(\|\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^\star\|^2\big) \\
&\le \frac{1}{2t}\sum_{i=1}^k \big(\|\boldsymbol{x}^{(i-1)} - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}^{(i)} - \boldsymbol{x}^\star\|^2\big) \\
&= \frac{1}{2t}\big(\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|^2 - \|\boldsymbol{x}^{(k)} - \boldsymbol{x}^\star\|^2\big).
\end{aligned}$$

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

$$f(\boldsymbol{x}^{(k)}) - f^\star \le \frac{1}{2kt}\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^\star\|^2.$$

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

Analysis with line search. The derivation is quite similar:

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

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

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

□