Lecture 9.5: The Proximal Gradient Method*

Shi Pu

School of Data Science (SDS)
The Chinese University of Hong Kong, Shenzhen

Outline

Outline

1. The Composite Model
2. The Proximal Gradient Method
3. Convergence of The Proximal Gradient Method

The Composite Model

Consider the composite problem (P) given by

$$(P) \qquad \begin{aligned} \min \quad & F(\mathbf{x}) \equiv f(\mathbf{x}) + g(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{x} \in \mathbb{R}^n \end{aligned}$$

- f - continuously differentiable over $\mathbb{R}^n$, L-smooth. Not necessarily convex.
- g - convex. Not necessarily continuous or differentiable.

Examples

- Unconstrained smooth minimization: $g(\mathbf{x}) \equiv 0$.
- Convex constrained minimization: $g(\mathbf{x}) = \delta_C(\mathbf{x})$, where C is nonempty closed and convex.
- l_1-regularized minimization: $g(\mathbf{x}) = \lambda \|\mathbf{x}\|_1$ for some $\lambda > 0$.

Outline

- The general update step of the gradient projection method takes the form
$$\mathbf{x}_{k+1} = P_C(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k)),$$
where $t_k > 0$ is the stepsize at iteration k.
- The update step can also be written as

$$\begin{aligned}
\mathbf{x}_{k+1} &= \arg\min_{\mathbf{x}\in C}\{\|\mathbf{x} - (\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k))\|^2\} \\
&= \arg\min_{\mathbf{x}\in C}\{f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}_k\|^2\} \\
&= \arg\min_{\mathbf{x}\in \mathbb{R}^n}\{f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2t_k}\|\mathbf{x} - \mathbf{x}_k\|^2 + \delta_C(\mathbf{x})\}.
\end{aligned}$$

The Proximal Gradient Update

- For the general problem (P), replacing δ_C by g yields

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n}\{f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^\top(\mathbf{x}-\mathbf{x}_k) + \frac{1}{2t_k}\|\mathbf{x}-\mathbf{x}_k\|^2 + g(\mathbf{x})\}.$$

- After some simple algebraic manipulations and cancellation of constant terms, the update can be rewritten as

$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n}\{t_k g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}-(\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k))\|^2\}.$$

The Proximal Gradient Update

- Define the proximal mapping:

$$\mathrm{prox}_g(\mathbf{x}) = \arg\min_{\mathbf{y}\in\mathbb{R}^n}\{g(\mathbf{y}) + \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|^2\}.$$

- The proximal gradient update can be written as

$$\mathbf{x}_{k+1} = \mathrm{prox}_{t_k g}(\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)).$$

Examples of Proximal Mappings

- $$g_1(x) = \begin{cases} \mu x, & x \geq 0 \\ \infty, & x < 0 \end{cases}$$

Examples of Proximal Mappings

- $$g_1(x) = \begin{cases} \mu x, & x \geq 0 \\ \infty, & x < 0 \end{cases}$$

  $$\text{prox}_{g_1}(x) = [x - \mu]_+.$$

- $$g_2(x) = \lambda|x|.$$

- $$g_2(x) = \lambda |x|.$$

$$\mathrm{prox}_{g_2}(x) = \begin{cases} x+\lambda, & x < -\lambda \\ 0, & |x| \leq \lambda \\ x-\lambda, & x > \lambda \end{cases}$$

Soft-thresholding function.

Algorithm 1 The Proximal Gradient Method

1: **Initialization:** pick $\mathbf{x}_0 \in \mathbb{R}^n$.
2: **General step:**
3: **for** $k = 0, 1, 2, \cdots$ execute the following steps: **do**
4: pick $t_k > 0$
5: set $\mathbf{x}_{k+1} = \text{prox}_{t_k g}(\mathbf{x}_k - t_k \nabla f(\mathbf{x}_k))$.
6: **end for**

- There are several strategies for choosing the stepsizes t_k.
- When $f \in C_L^{1,1}$, we can choose t_k to be constant and equal to $\frac{1}{L}$.

Iterative Shrinkage-Thresholding Algorithm (ISTA)

- In the context of solving the l_1-norm regularized problem, the proximal gradient method is

$$\begin{aligned}\mathbf{x}_{k+1} =& \text{prox}_{t_k\lambda\|\cdot\|_1}(\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k)) \\ =& \arg\min_{\mathbf{x}\in\mathbb{R}^n}\{t_k\lambda\|\mathbf{x}\|_1 + \frac{1}{2}\|\mathbf{x} - (\mathbf{x}_k - t_k\nabla f(\mathbf{x}_k))\|^2\}.\end{aligned}$$

- The optimization problem on the right-hand side of the expression is separable and can be solved in closed form.

$$[x_{k+1}]_i = \begin{cases} [x_k - t_k \nabla f(\mathbf{x}_k)]_i + t_k\lambda & [x_k - t_k \nabla f(\mathbf{x}_k)]_i < -t_k\lambda \\ 0 & [x_k - t_k \nabla f(\mathbf{x}_k)]_i \in [-t_k\lambda, t_k\lambda] \\ [x_k - t_k \nabla f(\mathbf{x}_k)]_i - t_k\lambda & [x_k - t_k \nabla f(\mathbf{x}_k)]_i > t_k\lambda \end{cases}$$

- Define the gradient mapping as

$$G_L^{f,g}(\mathbf{x}) = L\left[\mathbf{x} - \mathrm{prox}_{\frac{1}{L}g}(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x}))\right],$$

  where $L > 0$.
- When $g \equiv 0$, $G_L^{f,g}(\mathbf{x}) = \nabla f(\mathbf{x})$.
- $G_L^{f,g}(\mathbf{x}) = \mathbf{0}$ if and only if $\mathbf{x}$ is a stationary point of (P). Hence we can consider $\|G_L^{f,g}(\mathbf{x})\|^2$ to be the optimality measure.
- The update of the proximal gradient method can be rewritten as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - t_k G_{\frac{1}{t_k}}^{f,g}(\mathbf{x}_k).$$

Convergence of the Proximal Gradient Method

Theorem

Let $\{\mathbf{x}_k\}$ be the sequence generated by the proximal gradient method for solving problem (P) with a constant stepsize defined by $t_k = \bar{t} \in (0, \frac{2}{L})$, where L is a Lipschitz constant of ∇f. Assume that f is bounded below. Then

1. *The sequence $\{F(\mathbf{x}_k)\}$ is nonincreasing.*
2. $G^{f,g}_{\frac{1}{\bar{t}}}(\mathbf{x}_k) \to 0$ *as* $k \to \infty$

- Any limit point of the sequence is a stationary point of the problem.
- Rate of convergence of gradient mapping norms can be derived (similar to GD for unconstrained problem).