Modern Optimization

Jean-Luc Bouchot

School of Mathematics and Statistics
Beijing Institute of Technology
jlbouchot@bit.edu.cn

Spring 2021

1. Proximal methods

Outline

1 Proximal methods

Definition 2.1 (Composite model)

Let $f(x) = g(x) + h(x)$ *where*

- g *is nice (i.e. for which the analysis from the previous sections carry over)*
- h *is simple – which we will describe later on*

This is called a **composite** *model.*

Example 2.1

Assume we are trying to solve the following constrained optimization problem

$$\min f_0(x)$$
$$\text{s.t. } x \in \Omega$$

where Ω is a convex body.
This can be rewritten in the form of a composite function with

- $g = f_0$
- $h = \chi_\Omega$ (which is 0 for points in Ω and ∞ elsewhere)

Example 2.2

Assume we are trying to solve the following constrained optimization problem

$$\min f_0(x)$$
$$\text{s.t. } Ax = 0$$

where $A \in \mathbb{R}^{m\times n}$.

This can be approximated via a composite function with

- $g = f_0$
- $h = \|Ax\|$

Remark 2.1

Note that if both functions g and h are differentiable, we're good to go!
The interesting part is if h is not differentiable (e.g. indicator function)

Remark 2.2

At each iterations, we will (try to) solve:

$$x^{k+1} := \operatorname{argmin} \left\{ \frac{1}{2\gamma} \|y - (x^k - \gamma \nabla g(x^k))\|^2 + h(y). \right\}$$

Definition 2.2

Let f be a function and $\gamma > 0$ a given parameter. We define the **proximal operator** *as*

$$\mathrm{prox}_{f,\gamma}(x) := \mathrm{argmin}\{f(y) + \frac{1}{2\gamma}\|y - x\|^2\}.$$

Example 2.3

Let C be a nonempty closed convex body and define

$$\chi_C(x) := \begin{cases} 0 & \text{if } x \in C, \\ \infty & \text{elsewhere.} \end{cases}$$

Its proximal operator is precisely the projection.

Definition 2.3

We define the **proximal gradient descent** *as the sequence of iterates*

$$x^{k+1} := \mathrm{prox}_{h,\gamma}(x^k - \gamma \nabla g(x^k))$$

with a certain starting point x^0.

Proposition 2.1 (Admitted)

Under some very general assumptions (e.g. f is proper closed and convex or f is proper closed and coercive) the proximal operator admits a unique valued and is defined.

Example 2.4

Let $A \in \mathbb{R}^{d\times d}$ be symmetric positive definite, $b \in \mathbb{R}^d$ be a constant vector and $c \in \mathbb{R}$ a scalar and define $f(x) = \frac{1}{2}x^T Ax + b^T x + c$. Then, for $\gamma > 0$,

$$\operatorname{prox}_{f,\gamma}(x) = \left(A + \frac{1}{\gamma}I\right)^{-1}(\frac{1}{\gamma}x - b).$$

Remark 2.3

The proximal gradient descent algorithm is a generalization of both the gradient descent and the projected gradient descent.

Definition 2.4

Let $f = g + h$ with g convex differentiable (and smooth) and h simple. We define its **generalized gradient** *as the operator*

$$G_{h,\gamma}(x) := \frac{1}{\gamma}\left(x - \mathrm{prox}_{h,\gamma}\left(x - \gamma \nabla g(x)\right)\right).$$

Proposition 2.2

The proximal gradient descent can also be written as a generalized gradient descent

$$x^{k+1} = x^k - \gamma G_{h,\gamma}(x^k).$$

Theorem 2.1

Let $f = g + h$ be a composite function such that g is convex (proper closed) and L-smooth and h is convex (and proper closed). Let $\{x^k\}$ be the sequence of iterates generated by the proximal gradient descent algorithm with stepsize $\gamma = 1/L$ and starting at $x^0 \in \mathbb{R}^d$. Assume moreover that the function f admits a minimum point x^*. Then for any $K \geq 1$ it holds

$$f(x^K) - f(x^*) \leq \frac{L}{2K} \|x^0 - x^*\|^2$$

Lemma 1

Let f be a (proper closed) convex function and $\gamma > 0$. For any x in the domain

$$u = \operatorname{prox}_{f,\gamma}(x) \Rightarrow \frac{1}{\gamma}\langle x-u, y-u\rangle \leq f(y) - f(u), \quad \forall y.$$

Theorem 2.2 (Prox of separable functions)

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ *is* **separable**:

$$f(x) = f(x_1, \cdots, x_d) = \sum_{i=1}^{d} f_i(x_i),$$

where all the f_i*'s a proper closed and convex univariate functions. Then*

$$\operatorname{prox}_{f,\gamma}(x) = \left(\operatorname{prox}_{f_i,\gamma}(x_i)\right)_{i=1}^{d}.$$

Remark 2.4

This previous result is a consequence of the more general result: if

$$f : \begin{cases} \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \cdots \times \mathbb{R}^{d_m} & \to & (-\infty, \infty] \\ (x_1, \cdots, x_m) & \mapsto & \sum_{i=1}^m f_i(x_i) \end{cases}$$

then

$$\mathrm{prox}_{f,\gamma}(x) = \mathrm{prox}_{f_1,\gamma}(x_1) \times \mathrm{prox}_{f_2,\gamma}(x_2) \times \cdots \times \mathrm{prox}_{f_m,\gamma}(x_m)$$

Example 2.5

Let $f(x) = \|x\|_1$. For any $\gamma > 0$, its proximal operator is given by

$$\mathrm{prox}_{f,\gamma}(x) = S_\gamma(x)$$

where S_γ denotes the soft-thresholding operator applied component-wise and defined as

$$S_\gamma(x)_i = \max\{|x_i| - \gamma, 0\}\,\mathrm{sign}(x_i), \quad \forall 1 \leq i \leq d.$$

Theorem 2.3 (Prox with scaling and translation)

Let g be a proper function. For any scaling parameter $\lambda \neq 0$ and translation $a \in \mathbb{R}^d$, define $f(x) = g(\lambda x + a)$. It follows that

$$\operatorname{prox}_{f,\gamma}(x) = \frac{1}{\lambda}\left(\operatorname{prox}_{\lambda^2 g,\gamma}(\lambda x + a) - a\right).$$

Theorem 2.4 (Proved as part of the convergence of the proximal gradient descent)

Let f be a (proper closed) convex function. Then for any $x, y \in \mathbb{R}^d$ the following statements are equivalent

1. $y = \text{prox}_{f,1}(x)$
2. $x - y \in \partial f(y)$ *(the subgradient – see exercises)*
3. $\langle x - y, z - y \rangle \leq f(z) - f(y)$ *for all* $z \in \mathbb{R}^d$.

Proposition 2.3

Let f be a (proper closed) convex function. Then x is a minimizer of f if and only if it is a fixed point of its proximal operator:

$$x = \operatorname{prox}_{f,1}(x).$$

Definition 2.5

The **fast proximal gradient method** *is defined by the following set of instructions:*

$$x^{k+1} = \operatorname{prox}_{h,\gamma}\left(y^k - \gamma \nabla g(y^k)\right)$$
$$y^{k+1} = x^{k+1} + \left(\frac{\lambda_k - 1}{\lambda_{k+1}}\right)\left(x^{k+1} - x^k\right)$$

where

$$\lambda_{k+1} = \frac{1 + \sqrt{1 + 4\lambda_k^2}}{2}$$

and with some $\gamma > 0$ *and*

$$\lambda_0 = 1$$
$$x^0 = y^0.$$

Theorem 2.5

Let $f = g + h$ be a composite function with g L-smooth convex, and h convex. Assume further that the set of minimizers F^* of the function f is not empty and let $x^* \in F^*$. Then

$$f(x^K) - f(x^*) \leq \frac{2L}{(K+1)^2} \|x^0 - x^*\|_2^2.$$