CHAPTER 4. PROXIMAL GRADIENT DESCENT

Hao Yuan **Kun Yuan**

October 10, 2023

1 Problem formulation

This chapter considers the following problem

$$\min_{x \in \mathbb{R}^d} \quad \psi(x) = f(x) + h(x), \tag{1}$$

where $f(x)$ is a differentiable function and $h(x)$ is a convex function(maybe not differentiable).

Notation. We introduce the following notations:

- Let $x^\star := \arg\min_{x \in \mathbb{R}^d}\{\psi(x)\}$ be the optimial solution to problem (1).
- Let $\psi^\star := \min_{x \in \mathbb{R}^d}\{\psi(x)\}$ be the optimal function value.

2 Subgradients

2.1 Definition

Definition 2.1. Let $f : \mathbf{dom} f \to \mathbb{R}$. Then $g \in \mathbb{R}^d$ is a subgradient of f at $x \in \mathbf{dom} f$ if

$$f(y) \geq f(x) + g^\top (y - x), \quad \forall y \in \mathbf{dom} f.$$

The set of subgradients of f at x is called the subdifferential of x and is denoted by $\partial f(x)$.

2.2 Examples

- ℓ_1-norm: $\forall x \in \mathbb{R}^d$, $h(x) = \|x\|_1$, $\partial h(0) = \{g \in \mathbb{R}^d \mid |g_i| \le 1, i = 1, \ldots, d\}$.
- ℓ_2-norm: $\forall x \in \mathbb{R}^d$, $h(x) = \|x\|_2$, $\partial h(0) = \{g \in \mathbb{R}^d \mid \|g\|_2 \le 1\}$.

2.3 Optimality Conditions

Theorem 2.2. We suppose ψ is a convex and proper function, then $x^\star$ is a global minimum of problem (1) if and only if

$$0 \in \partial \psi(x^\star).$$

Proof. $\Leftarrow$:
Because $x^\star$ is a global minimum, we have

$$\psi(y) \ge \psi(x^\star) = \psi(x^\star) + 0^\top (y - x^\star), \quad \forall y \in \mathbb{R}^d,$$

then $0 \in \partial \psi(x^\star)$.

$\Rightarrow$:
Because $0 \in \partial \psi(x^\star)$, we have

$$\psi(y) \ge \psi(x^\star) + 0^\top (y - x^\star) = \psi(x^\star), \quad \forall y \in \mathbb{R}^d,$$

then $x^\star$ is a global minimum.

□

3 Proximity Operator

3.1 Definition

Definition 3.1. For a convex function h, its proximity operator is defined as

$$\operatorname{prox}_h(x) = \mathop{\arg\min}_{u \in \mathbf{dom} h} \left\{ h(u) + \frac{1}{2}\|u - x\|^2 \right\}.$$

Theorem 3.2. If h is a closed convex proper function, then for any $x \in \mathbb{R}^d$, $\operatorname{prox}_h(x)$ exists and is unique.

Proof. The proof can be found in [1] or `http://faculty.bicmr.pku.edu.cn/~wenzw/optbook/opt1.pdf`. □

Using Theorem 2.2, we can derive the following theorem.

Theorem 3.3. If h is a closed convex proper function, then

$$u = \mathrm{prox}_h(x) \iff x - u \in \partial h(u).$$

3.2 Examples

Using Theorem 3.3 and the results in Sec 2.2, we can derive the following results. In the following examples, $t > 0$ is a constant.

- ℓ_1-norm: $\forall x \in \mathbb{R}^d$, $h(x) = \|x\|_1$, $\Big(\mathrm{prox}_{th}(x)\Big)_i = \mathrm{sign}(x_i)\max\{|x_i| - t, 0\}$, $i = 1, \ldots, d$.
- ℓ_2-norm: $\forall x \in \mathbb{R}^d$, $h(x) = \|x\|_2$,

$$\mathrm{prox}_{th}(x) = \begin{cases} (1 - \dfrac{t}{\|x\|_2})x, & \|x\|_2 \geq t, \\ 0, & \text{otherwise.} \end{cases}$$

- projection: Let C be a closed convex set, the indicator function of C is defined as

$$I_C(x) = \begin{cases} 0, & x \in C, \\ +\infty, & \text{otherwise.} \end{cases}$$

  Then

$$\begin{aligned} \mathrm{prox}_{I_C}(x) &= \arg\min_u \left\{ I_C(u) + \frac{1}{2}\|u - x\|^2 \right\} \\ &= \arg\min_{u \in C} \|u - x\|^2 \\ &= \mathcal{P}_C(x) \end{aligned}$$

More examples can be found in `http://proximity-operator.net/index.html`.

3.3 Transformation rules

- $h(x) = g(\lambda x + a)$, $\lambda \neq 0$, $\mathrm{prox}_h(x) = \frac{1}{\lambda}(\mathrm{prox}_{\lambda^2 g}(\lambda x + a) - a)$;
- $h(x) = \lambda g(\frac{x}{\lambda})$, $\lambda > 0$, $\mathrm{prox}_h(x) = \lambda \mathrm{prox}_{\lambda^{-1} g}(\frac{x}{\lambda})$;
- $h(x) = g(x) + a^\top x$, $\mathrm{prox}_h(x) = \mathrm{prox}_g(x - a)$;
- $h(x) = g(x) + \frac{u}{2}\|x - a\|^2$, $\mathrm{prox}_h(x) = \mathrm{prox}_{\theta g}(\theta x + (1 - \theta)a)$, where $\theta = \frac{1}{1+u}$.

3.4 Non-expansive property

Theorem 3.4. If h is a closed convex proper function, then

$$\|\mathrm{prox}_h(x) - \mathrm{prox}_h(y)\| \leq \|x - y\|.$$

4 Proximal gradient descent

For optimization problem 1, given any arbitrary initialization variable $x_0 \in \mathbb{R}^d$, proximal gradient descent iterates as follows

$$y_{k+1} = x_k - \gamma \nabla f(x_k), \tag{2a}$$
$$x_{k+1} = \mathrm{prox}_{\gamma h}(y_{k+1}), \quad \forall\, k = 0, 1, 2, \cdots \tag{2b}$$

where γ is the learning rate.

5 Convergence analysis

5.1 Smooth and generally convex problem

Assumption 5.1. 1. f is convex and L-smooth on $\mathbb{R}^d$.

2. h is a closed convex proper function.

3. $\psi^\star$ exists, and $\psi(x^\star) = \psi^\star$.

Definition 5.2. Under Assumption 5.1, we define

$$G_t(x) = \frac{1}{t}(x - \mathrm{prox}_{th}(x - t\nabla f(x))).$$

And it holds that

$$x^{k+1} = \mathrm{prox}_{th}(x^k - t\nabla f(x^k)) = x^k - tG_t(x^k).$$

Lemma 5.3. Under Assumption 5.1 and using Definition 5.2, we have

$$G_t(x) - \nabla f(x) \in \partial h(x - tG_t(x)).$$

Theorem 5.4. Under Assumption 5.1, if the stepsize $t = \frac{1}{L}$, proximal gradient descent 2 with arbitrary x_0 satisfies

$$\psi(x^K) - \psi^\star \leq \frac{L}{2K}\|x_0 - x^\star\|^2.$$

Proof. Because f is L-smooth, we have

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d.$$

Let $y = x - tG_t(x)$, it holds that

$$f(x - tG_t(x)) \leq f(x) - t\langle \nabla f(x), G_t(x)\rangle + \frac{t^2 L}{2}\|G_t(x)\|^2.$$

Using $t = \frac{1}{L}$, we have

$$f(x - tG_t(x)) \leq f(x) - t\langle \nabla f(x), G_t(x) \rangle + \frac{t}{2}\|G_t(x)\|^2. \tag{3}$$

Because f, h are convex, using Lemma 5.3 for $\forall z \in \mathbb{R}^d$, we have

$$\begin{aligned}
&f(z) \geq f(x) + \langle \nabla f(x), z - x \rangle, \\
&h(z) \geq h(x - tG_t(x)) + \langle G_t(x) - \nabla f(x), z - x + tG_t(x) \rangle.
\end{aligned}$$

Rearranging the above two inequalities, we have

$$h(x - G_t(x)) \leq h(z) - \langle G_t(x) - \nabla f(x), z - x + tG_t(x) \rangle, \tag{4}$$
$$f(x) \leq f(z) - \langle \nabla f(x), z - x \rangle. \tag{5}$$

Combining (3),(4) and (5), it holds that

$$\psi(x - tG_t(x)) \leq \psi(z) + \langle G_t(x), x - z \rangle - \frac{t}{2}\|G_t(x)\|^2. \tag{6}$$

In (6), let $z = x^\star$, $\tilde{x} = x - tG_t(x)$, we have

$$\begin{aligned}
\psi(\tilde{x}) - \psi^\star &\leq \langle G_t(x), x - x^\star \rangle - \frac{t}{2}\|G_t(x)\|^2 \\
&= \frac{1}{2t}(\|x - x^\star\|^2 - \|x - x^\star - tG_t(x)\|^2) \\
&= \frac{1}{2t}(\|x - x^\star\|^2 - \|\tilde{x} - x^\star\|^2).
\end{aligned} \tag{7}$$

Let $x = x^{i-1}$, $\tilde{x} = x^i$, $i = 1, 2, \cdots, K$, in (7) respectively, and take the telescoping sum, we have

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

In (6), let $z = x$, we obtain

$$\psi(\tilde{x}) \leq \psi(x) - \frac{t}{2}\|G_t(x)\|^2 < \psi(x). \tag{8}$$

Then we have

$$\psi(x^K) - \psi^\star \leq \frac{1}{K}\sum_{i=1}^{K}(\psi(x^i) - \psi^\star) \leq \frac{1}{2Kt}\|x^0 - x^\star\|^2 = \frac{L}{2K}\|x^0 - x^\star\|^2. \tag{9}$$

$\square$

5.2 Smooth and strongly convex problem

Lemma 5.5. Under Assumption 5.1, if the stepsize $t = \frac{1}{L}$, proximal gradient descent 2 with arbitrary x_0 satisfies

$$x^\star = \mathrm{prox}_{th}(x^\star - t\nabla f(x^\star)).$$

Assumption 5.6.
1. f is μ-strongly convex and L-smooth on $\mathbb{R}^d$.
2. h is a closed convex proper function.
3. $\psi^\star$ exists, and $\psi(x^\star) = \psi^\star$.

Theorem 5.7. Under Assumption 5.6, if the stepsize $t = \frac{1}{L}$, proximal gradient descent 2 with arbitrary x_0 satisfies

$$\|x^K - x^\star\| \leq (1 - \frac{\mu}{L})^K \|x^0 - x^\star\|.$$

References

[1] H. Bauschke and P. Combettes, "Convex analysis and monotone operator theory in hilbert spaces, 2011," *CMS books in mathematics). DOI*, vol. 10, pp. 978–1.