6FMAI19 Nonlinear Optimization	Spring, 2022
Lecture #10 — 30/3, 2022	
Lecturer: Yura Malitsky	*Scribe: Daniel Arnström*

1 Proximal operators

We will now introduce the concept of *proximal operators*, which will motivate new methods and, in some sense, generalize previous discussions about, for example, projections.

Definition 1 (Proximal operator)**.** *For a convex function $f : \mathbb{R}^n \to R^n \cup \{\infty\}$ we define its proximal operator* $\operatorname{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ *through the rule*

$$\operatorname{prox}_f(x) \triangleq \arg\min_u \left\{ f(u) + \frac{1}{2}\|u - x\|^2 \right\}, \tag{1}$$

for any $x \in \mathbb{R}^n$.

The defining rule of prox_f in (1) is well-defined since f convex implies that $f(u)+\frac{1}{2}\|u-x\|^2$ is strongly convex, which ensures that $\operatorname{prox}_f(x)$ exists and is unique (i.e., is single-valued). Calling it a *proximal* operator originates from the term $\frac{1}{2}\|u - x\|_2^2$ forcing u to be in the *proximity* of x.

1.1 Canonical examples

To build some intuition for prox_f we give two canonical examples: First we consider the case when f is a quadratic, which highlights the regularizing properties of prox_f; then we consider the case when f is an indicator function, which highlights the projective properties of prox_f.

Example 1: Consider the convex quadratic function $f(x) = \frac{1}{2}\langle x, Qx\rangle + \langle b, x\rangle$, that is, Q is psd. In this case its proximal operator takes the closed form

$$\begin{aligned} \operatorname{prox}_f(x) &= \arg\min_u \left\{ \frac{1}{2}\langle u, Qu\rangle + \langle b, u\rangle + \frac{1}{2}\|u - x\|_2^2 \right\} \\ &= \arg\min_u \left\{ \frac{1}{2}\langle u, (Q + I)u\rangle + \langle b - x, u\rangle \right\} \\ &= (I + Q)^{-1}(x - b), \end{aligned} \tag{2}$$

where we have used that $\|u - x\|_2^2 = \|u\|_2^2 + \|x\|^2 - 2\langle x, u\rangle$ in the second equality, and that $(I + Q)^{-1}$ exists in the third equality since $Q \succeq 0 \implies (I + Q) \succ 0$.

Example 2: Consider the indicator function $f(x) = \delta_C(x) = \begin{cases} 0 & \text{if } x \in C \\ \infty & \text{if } x \notin C \end{cases}$, where C is a closed and convex set. In this case the proximal operator becomes a projection:

$$\begin{aligned} \operatorname{prox}_f(u) &= \arg\min_u \left\{ \delta_C(u) + \frac{1}{2}\|u - x\|_2^2 \right\} \\ &= \arg\min_{u \in C} \left\{ \frac{1}{2}\|u - x\|_2^2 \right\} \\ &= P_C(x). \end{aligned} \tag{3}$$

1.2 Equivalent characterizations

Theorem 2 (Equivalent charcterizations of prox_f). *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex and closed. Then the following are equivalent:*

(i) $u = \text{prox}_f(x)$

(ii) $x \in (I + \partial f)u$

(iii) $\langle u - x, y - u \rangle \geq f(u) - f(y) \quad \forall y$

Proof. $(i) \implies (ii)$: Using the definition of prox_f and Fermat's rule (i.e., that $0 \in \partial \phi(x)$ is a necessary and sufficient condition for x to be the minimizer to a convex function ϕ) yields

$$0 \in \partial(f(u) + \frac{1}{2}\|u - x\|_2^2) = \partial f(u) + u - x = (I + \partial f)u - x \Leftrightarrow x \in (I + \partial f)u. \tag{4}$$

$(ii) \Leftrightarrow (iii)$: Rewriting (ii) as $x - u \in \partial f$ and using the subgradient inequality yields

$$f(y) \geq f(u) + \langle x - u, y - u \rangle, \ \forall y \Leftrightarrow \langle u - x, y - u \rangle \geq f(u) - f(y), \ \forall y \tag{5}$$

□

Note that property (ii) can alternatively be written as $\text{prox}_f(x) = (I + \partial f)^{-1}x$, which is similar to the closed-form derived in Example 1 for a quadratic function.

2 The proximal point algorithm

Proximal operators can be used to derive a simple algorithm for solving $\min_x f(x)$:

Algorithm 1 The proximal point algorithm (PPA)

Input: x_0, rule for selecting $\alpha_k > 0$
Output: $\approx x^*$

```
k ← 0
repeat
    x_{k+1} ← prox_{α_k f}(x_k).
    k ← k + 1
until termination criterion satisfied
```

Note that although Algorithm 1 is simple to formulate, prox_f is generally difficult to evaluate for an arbitrary f.

If f is differentiable, it follows from (ii) in Theorem 2 that an iteration in Algorithm 1 takes the form

$$x_{k+1} = x_k - \alpha_k \nabla f(x_{k+1}). \tag{6}$$

This is reminiscent of an iteration in gradient descent, except that the gradient is evaluated in x_{k+1} rather than in x_k, making (6) an implicit rule. For those familiar with integration of differential equations, this is analogous to the forward Euler method (GD) vs the backward Euler method (PPA).

2.1 Convergence

Before deriving the convergence rate of Algorithm 1, we show that it is a descent method.

Lemma 3 (Descent in PPA)**.** *If $\alpha > 0$ in Algorithm 1, the iterates are monotonically decreasing w.r.t. to f. That is, $f(x_{k+1}) \leq f(x_k)$.*

Proof. By letting $x_{k+1} = \text{prox}_{\alpha_k f}(x_k)$ and $y = x_k$ in the inequality (iii) in Theorem 2 we get

$$\begin{aligned} \langle x_{k+1} - x_k, x_k - x_{k+1} \rangle \geq \alpha_k (f(x_{k+1}) - f(x_k)) \quad &\Leftrightarrow \\ -\|x_{k+1} - x_k\|_2^2 \geq \alpha_k (f(x_{k+1}) - f(x_k)) \quad &\Leftrightarrow \\ f(x_{k+1}) - f(x_k) \leq \frac{\|x_{k+1} - x_k\|_2^2}{\alpha_k} \leq 0, & \end{aligned} \tag{7}$$

where we, in fact, have strict descent if $x_{k+1} \neq x_k$, i.e., if x_k is not a fixed-point to $\text{prox}_{\alpha_k f}$. □

The strict descent in PPA as long as x_k is not a fixed-point motivates the common termination rule of terminating when $x_{k+1} \approx x_k$.

We are now ready to derive the converge rate for Algorithm 1.

Theorem 4 (Convergence of PPA)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and closed and denote $x^* \in \arg\min_x f(x)$ and $f_* = f(x^*)$. Then the iterates in Algorithm 1 satisfy*

$$f(x_{k+1}) - f_* \leq \frac{\|x_0 - x^*\|_2^2}{2\sum_{i=0}^k \alpha_i}. \tag{8}$$

Proof. By letting $x_{k+1} = \text{prox}_{\alpha_k f}(x_k)$ and $y = x^*$ in inequality (iii) in Theorem 2 we get

$$\langle x_{k+1} - x_k, x^* - x_{k+1} \rangle \geq \alpha_k (f(x_{k+1}) - f_*) \tag{9}$$

Using the identity $\langle a, b \rangle = \frac{1}{2}\|a+b\|_2^2 - \frac{1}{2}\|a\|_2^2 - \frac{1}{2}\|b\|_2^2$ and reordering terms yield

$$\frac{1}{2}\|x_{k+1} - x^*\|_2^2 + \alpha_k (f(x_{k+1}) - f_*) + \frac{1}{2}\|x_{k+1} - x_k\|_2^2 \leq \frac{1}{2}\|x_k - x^*\|_2^2. \tag{10}$$

Since $\|x_{k+1} - x_k\|_2^2 \geq 0$ we get $\frac{1}{2}\|x_{k+1} - x^*\|_2^2 + \alpha_k (f(x_{k+1}) - f_*) \leq \frac{1}{2}\|x_k - x^*\|_2^2$, and telescoping this inequality from iteration 0 to iteration k gives

$$\sum_{i=0}^k \alpha_i (f(x_{i+1} - f_*)) \leq \frac{1}{2}\|x_0 - x^*\|_2^2. \tag{11}$$

Finally, from the descent property in Lemma 3 we have that $f(x_{k+1}) \leq f(x_{i+1})$ for all $i \leq k$. Hence, we have that $(f(x_{k+1}) - f_*) \sum_{i=0}^k \alpha_i \leq \sum_{i=0}^k \alpha_i (f(x_{i+1}) - f_*)$, which inserted into (11) yields

$$f(x_{k+1}) - f_* \leq \frac{\|x_0 - x^*\|_2^2}{2\sum_{i=0}^k \alpha_i}. \tag{12}$$

□

At a first glance, the result in Theorem 4 seem to imply that the rate of convergence can be arbitrary fast by letting $\alpha_k \to \infty$. Theoretically, this is correct, although closer inspection reveals practical limitations. By assuming that $\alpha_k > 0$ we get that x_{k+1} in PPA is given by

$$x_{k+1} = \text{prox}_{\alpha_k f}(x_k) = \arg\min_u \left\{ \alpha_k f(u) + \frac{1}{2}\|u - x_k\|_2^2 \right\} = \arg\min_u \left\{ f(u) + \frac{1}{2\alpha_k}\|u - x_k\|_2^2 \right\}$$

Hence, if $\alpha_k \to \infty$, evaluating $\text{prox}_{\alpha_k f}$ becomes equivalent to solving the original problem $\min_x f(x)$. There is, hence, a trade-off in the selection α_k: a larger α_k leads to faster convergence but harder inner subproblems (since they becomes less regularized).

2.2 The proximal gradient method

The PPA is seldom applied in practice directly since evaluating prox_f for an arbitrary f is difficult. A related method that more often finds practical application is the proximal gradient method (PGM), which works on problems where the objective function can be split into to two parts as

$$\min_x f(x) + g(x), \tag{13}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is closed, convex and smooth and $g : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is convex and "prox-friendly", in the sense that prox_g is easy to evaluate. The PGM is outlined below.

Algorithm 2 The proximal gradient method (PGM)

Input: x_0, rule for selecting $\alpha_k > 0$
Output: $\approx x^*$

1: $k \leftarrow 0$
2: **repeat**
3: $\quad x_{k+1} = \text{prox}_{\alpha_k g}(x_k - \alpha_k \nabla f(x_k))$
4: $\quad k \leftarrow k + 1$
5: **until** termination criterion satisfied

When g is an indicator function for a closed and convex set C, i.e., $g = \delta_C(x)$, Algorithm 2 simply becomes projected gradient descent. Another important example of when PGM is used is when a term containing $\|\cdot\|_1$ is added to the objective function to obtain sparse solutions:

Example 3: Consider the minimization problem (sometimes called a "Lasso-regularized" least-squares problem)

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \lambda\|x\|_1, \tag{14}$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $\lambda > 0$. By letting $f(x) \triangleq \frac{1}{2}\|Ax - b\|_2^2$ and $g(x) \triangleq \lambda\|x\|_1$, the proximal operator for g takes the closed-form

$$[\text{prox}_g(x)]_i = \text{sign}([x]_i) \max\{|[x]_i| - \lambda, 0\}, \tag{15}$$

where $[\cdot]_i$ denotes the ith component of a vector. Applying Algorithm 2 to solve (14), hence, results in an iteration consisting of a gradient step with f followed by a soft thresholding according to (15).