Advanced Topics in Control 2020: Large-Scale Convex Optimization

Summary 7: Operator Splitting Methods

Goran Banjac, Mathias Hudoba de Badyn, Andrea Iannelli,
Angeliki Kamoutsi, Ilnura Usmanova

April 25, 2020

1 Proximal Operators

Proximal mapping. Given a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the **proximal mapping** is the operator given by

$$\mathbf{prox}_f(x) = \arg\min_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{1}{2}\|u - x\|_2^2 \right\} \tag{1}$$

Note that $\mathbf{prox}_{\gamma f}$ is the proximal operator $\mathbf{prox}$ of the scaled function $\gamma f(x)$:

$$\mathbf{prox}_{\gamma f}(x) = \arg\min_{u \in \mathbb{R}^n} \left\{ f(u) + \frac{1}{2\gamma}\|u - x\|_2^2 \right\} \tag{2}$$

You can see, that $\mathbf{prox}_f(x)$ is a mappping from $\mathbb{R}^n$ to $\mathbb{R}^n$, which might be a subset of $\mathbb{R}^n$, a singleton, or the empty set.

Theorem 1. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper closed and convex function. Then $\mathbf{prox}_f(x)$ is a* ***singleton*** *for any $x \in \mathbb{R}^n$.*

Examples:

- Constant: $f(x) \equiv c$

$$\mathbf{prox}_f(x) = \arg\min_u \left\{ c + \frac{1}{2}\|u - x\|^2 \right\} = x.$$

- Affine function $f(x) = \langle a, x \rangle + b$

$$\mathbf{prox}_{\gamma f}(x) = \arg\min_u \left\{ h(u) := \langle a, u \rangle + b + \frac{1}{2\gamma}\|u - x\|^2 \right\}$$

$$\nabla h(u^*) = a + \frac{1}{\gamma}(u^* - x) = 0$$

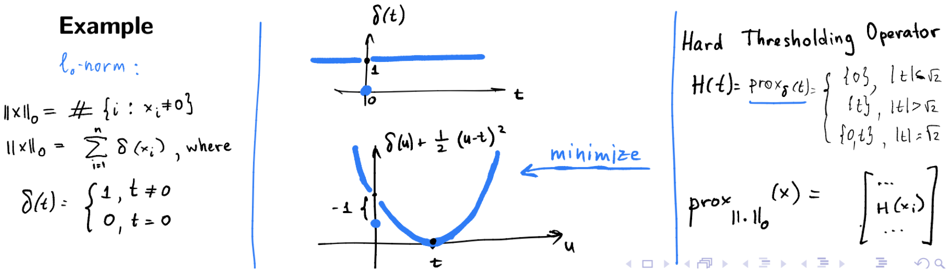
$$\mathbf{prox}_{\gamma f}(x) = u^* = x - \gamma a$$

1.1 Prox of separable functions

Theorem 2. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be block separable and be given by*

$$f(x) = \sum_{i=1}^{m} f_i(x_i), x_i \in \mathbb{R}^{n_i}, \sum_{i=1}^{m} n_i = n. \text{ Then, } \boldsymbol{prox}_f(x) = \begin{bmatrix} \boldsymbol{prox}_{f_1}(x_1) \\ \dots \\ \boldsymbol{prox}_{f_m}(x_m) \end{bmatrix}.$$

Examples: $\|\cdot\|_1 = \sum_{i=1}^{n} |x_i|$, $\|\cdot\|_p^p = \sum_{i=1}^{n} |x_i|^p$, $f(x) + g(z)$.



1.2 Prox of indicators: Orthogonal Projections

We denote the indicator function of the set C by

$$\mathcal{I}_C(x) = \begin{cases} 0, & x \in C \\ +\infty, & x \notin C \end{cases}. \tag{3}$$

Theorem 3. *Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be given by $g(x) = \mathcal{I}_C(x)$, where C is a nonempty set. Then, $\boldsymbol{prox}_g(x)$ is the orthogonal projection operator.*

$$\boldsymbol{prox}_g(x) = \arg\min_{u \in \mathbb{R}^n} \left\{ \mathcal{I}_C(u) + \frac{1}{2}\|u - x\|^2 \right\} = \arg\min_{u \in C} \|u - x\| = \boxed{\Pi_C(x)}.$$

1.3 Fixed point

The main link between Fixed Point theory and Proximal Operators:

Theorem 4. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper, closed, and convex. Then*

- *$u = \boldsymbol{prox}_{\gamma f}(x)$ if and only if $x - u \in \gamma \partial f(u)$.*
- *x is a minimizer of f if and only if $x = \boldsymbol{prox}_{\gamma f}(x)$.*

That is, finding a fixed point of $\mathbf{prox}_{\gamma f}(x)$ is a valid optimization algorithm.

this is very interesting theorem relating the proximal operators to the minimizers.

Take it as a convergent problem, move the x close to f, then we find the minimizer of f

1.4 Moreau's identity

The main link between **prox** operators and conjugate functions.

Theorem 5. *Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper, closed, and convex. Then for any $x \in \mathbb{R}^n$*

$$\boldsymbol{prox}_f(x) + \boldsymbol{prox}_{f^*}(x) = x.$$

Extended version, $\gamma > 0$:

$$\boldsymbol{prox}_{\gamma f}(x) + \boldsymbol{prox}_{\gamma^{-1} f^*}(x/\gamma) = x.$$

Proof: From the previous theorem it follows that $x - u \in \partial f(u)$, which is equivalent to $u \in \partial f^*(x - u)$ by the conjugate subgradient theorem. Using the previous theorem again, we conclude that $x - u = \mathbf{prox}_{f^*}(x)$.

x-(x-u)\in \partial f^*(x-u)

• **Conjugate Subgradient Theorem:** If f is closed proper convex, the following are equivalent for a pair of vectors (x, y):

(i) $x'y = f(x) + f^*(y)$.

(ii) $y \in \partial f(x)$.

(iii) $x \in \partial f^*(y)$.

2 Operators

Let T be an operator. Then, it is called:

- **Monotone operator** if $\langle Tx - Ty, x - y\rangle \geq 0$
- **Lipschitz continuous operator** if $\exists \beta > 0$ such that

$$\|Tx - Ty\| \leq \beta \|x - y\|$$

- **Contractive operator** if $\beta < 1$
  1) Always has a fixed point
  2) Simple iteration always converges
- **Nonexpancive operator** if $\beta \leq 1$
  1) Does not necessarily has a fixed point
  2) Even if it exists, no guarantee of convergence of the simple iteration

  without any averaged operation
- **Averaged operator** if $T = (1 - \alpha)I + \alpha R$ where R is nonexpancive.
  1) It guarantees the convergence of the fixed point simple iteration.
  2) Composition of averaged operators is averaged
  3) If R has a fixed point, then T has the same fixed point.

  if T has fixed point, then it converges, otherwise, consider T = 0.5I+0.5(I+2)=I+1

2.1 Resolvent

- **Resolvent** of an operator A is $J_{\gamma A} = (I + \gamma A)^{-1}$.
- If A is maximal monotone, then $J_{\gamma A}$ is
  - a singleton, is a function
  - full domain,

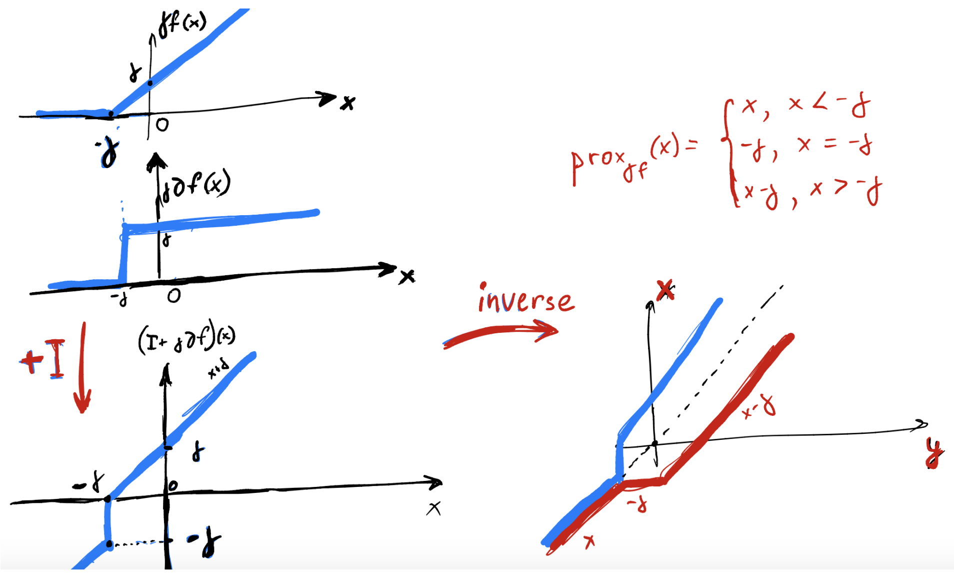
– $\frac{1}{2}$-averaged operator because reflected resolvent: R_A = 2J_A - I_d

- $\partial f(x)$ for the proper closed convex $f(x)$ is maximal monotone
- If $u = \mathbf{prox}_{\gamma f}(x)$ then $u \in J_{\gamma \partial f}(x) = (I + \gamma \partial f)^{-1}(x)$
Indeed, $u = \mathbf{prox}_{\gamma f}(x) \iff 0 \in \partial f(u) + u - x \iff x \in \partial f(u) + u \iff u \in (I + \partial f)^{-1}(x)$
- For the proper closed convex $f(x)$ we have $\mathbf{prox}_{\gamma f}(x) = (I + \gamma \partial f)^{-1}(x)$

Example: Derivation of $\mathbf{prox}_f(x)$ of one dimensional function:

$$f(x) = \max(0, 1 + x)$$

using graphical arguments:



Hence, simple iteration for finding a fixed point of $\mathbf{prox}_{\gamma f}(x)$ is a converging valid optimization algorithm.

2.2 Proximal Minimization (resolvent method)

$$x^{k+1} = \mathbf{prox}_{\gamma f}(x^k)$$

- The simplest proximal method
- Simple iteration for finding a fixed point of $\mathbf{prox}_f$
- Example: for quadratic fucntion $\frac{1}{2}x^T Ax - b^T x$ reduces to iterative refinement for $Ax = b$ (you can check this using your homework)

Splitting Operator Methods

$$\min f(x) + g(x). \tag{4}$$

- **Why Splitting?**: Makes use of the $f(x) + g(x)$ when $f(x)$ and $g(x)$ have useful structure separately.
- These algorithms minimize $f + g$ only using $\mathbf{prox}_f$ or $\mathbf{prox}_g$

1. **Proximal Gradient Method**

$$x^{k+1} = \mathbf{prox}_{\gamma^k g}(x^k - \gamma^k \nabla f(x^k))$$

If $g(x) = \mathcal{I}_C(x)$ is the indicator, turns to Projected Gradient Descent.

Interpretations:

(a) majorization-minimization for $f + g$:

$$x^{k+1} = \arg\min_{x \in \mathbb{R}^n} \left\{ f(x^k) + \langle \nabla f(x^k), x - x^k \rangle + \frac{1}{2\gamma}\|x - x^k\|^2 + \gamma g(x^k) \right\}$$

convex upper bound tight at previous iterate ($\gamma \in (0, 1/L)$).

(b) x^{k+1} is solution to

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|x - (x^k - \gamma \nabla f(x^k))\|^2 + \gamma g(x)$$

trade off between minimising g and being close to gradient step for f

(c) $0 \in \nabla f(x^*) + \partial g(x^*)$ if and only if

$$x^* = (I + \gamma \partial g)^{-1}(I - \gamma \nabla f)(x^*),$$

i.e., x^* is a fixed point of *forward-backward* operator.

Remark: There exisits the accelerated version of proximal gradient, called FISTA.

2. **Peaceman-Rachford splitting.** The solutions to $\min f(x) + g(x)$ are characterized by the fixed point

$$z = R_{\gamma \partial g} R_{\gamma \partial f} z = (2\mathbf{prox}_{\gamma g} - I)(2\mathbf{prox}_{\gamma f} - I)z, \tag{5}$$

$$x = \mathbf{prox}_{\gamma f}(z), \tag{6}$$

where $R_{\gamma \partial f} = 2\mathbf{prox}_{\gamma f} - I := r\mathbf{prox}_{\gamma f}(x)$ is the reflected proximal operator.
If we write the simple iteration for $R_{\gamma \partial g} R_{\gamma \partial f}$, we get the Peaceman-Rachford splitting algorithm:

$$z^{k+1} = R_{\gamma \partial g} R_{\gamma \partial f} z^k \tag{7}$$

$$x^k = \mathbf{prox}_{\gamma f}(z^k) \tag{8}$$

can be shown by definition. and use the
property of resolvent's monotone operator

- $R_{\gamma\partial g}, R_{\gamma\partial f}$ are nonexpancive, hence do not guarantee convergence.
- If f is strongly convex and β-smooth, then $r\mathbf{prox}_f$ is contractive

3. **Douglas-Rachford splitting.** Douglas-Rachford algorithm takes averaged map of $R_{\gamma\partial g}R_{\gamma\partial f}$ operator with $0 < \alpha < 1$:

$$z^{k+1} = ((1-\alpha)I + \alpha R_{\gamma\partial g}R_{\gamma\partial f})\, z^k \tag{9}$$
$$x^k = \mathbf{prox}_{\gamma f}(z^k) \tag{10}$$

or more explicitly

$$x^k = \mathbf{prox}_{\gamma f}(z^k) \tag{11}$$
$$y^k = \mathbf{prox}_{\gamma g}(2x^k - z^k) \tag{12}$$
$$z^{k+1} = z^k + 2\alpha(y^k - x^k) \tag{13}$$

Remark The above algorithms that can be applied for general operators A and B, not necessarily $\partial f, \partial g$