Operator Splitting Methods

Goran Banjac

Large-Scale Convex Optimization
ETH Zurich

April 21, 2020

Monotone operators

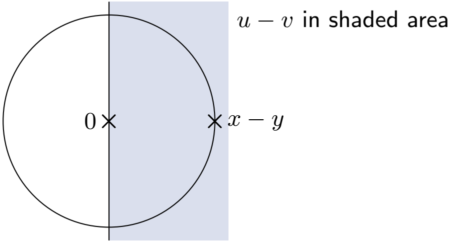
- the *graph* of an operator $A\colon \mathbb{R}^n \mapsto 2^{\mathbb{R}^n}$ is defined as

$$\operatorname{gph} A := \{(x,u) \mid u \in Ax\}$$

- operator A is *monotone* if

$$(u-v)^T(x-y) \geq 0$$

  for all $(x,u) \in \operatorname{gph} A$ and $(y,v) \in \operatorname{gph} A$



- A is maximally monotone if it is monotone and there exists no monotone operator B so that $\operatorname{gph} A \subset \operatorname{gph} B$

Lipschitz continuous operators

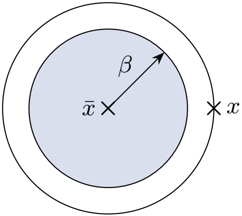
- let $\mathcal{D}$ be a subset of $\mathbb{R}^n$
- operator $T\colon \mathcal{D} \mapsto \mathbb{R}^n$ is β-Lipschitz continuous if

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

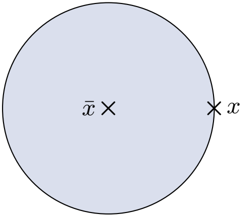
  holds for all $x, y \in \mathcal{D}$
- T is single-valued (show by letting $y = x$ and use contradiction)
- composition of Lipschitz continuous operators is Lipschitz continuous

$$T = T_1 \circ T_2 \quad \Longrightarrow \quad \beta = \beta_1 \beta_2$$

- graphical representation: $\bar{x} \in \operatorname{Fix} T$, Tx in shaded area



contractive: $\beta < 1$



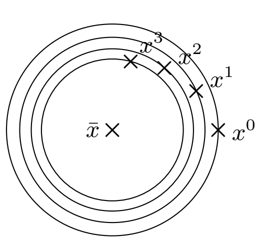
nonexpansive: $\beta = 1$

Iterating a nonexpansive operator

- contractive operators have unique fixed-points
- iteration $x^{k+1} = Tx^k$ converges linearly to the fixed-point $\bar{x}$

$$\|x^{k+1} - \bar{x}\| = \|Tx^k - \bar{x}\| \leq \beta\|x^k - \bar{x}\| \leq \ldots \leq \beta^{k+1}\|x^0 - \bar{x}\|$$

- a nonexpansive operator R need not have a fixed-point
- even if a fixed-point exists, iteration $x^{k+1} = Rx^k$ may not converge

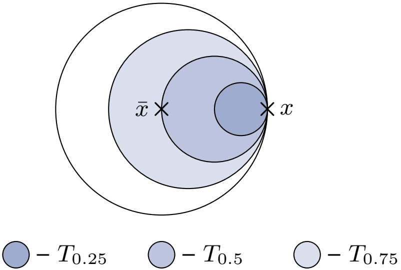


Averaged operators

- let $\alpha \in (0, 1)$ and $R\colon \mathcal{D} \mapsto \mathbb{R}^n$ be some nonexpansive operator
- operator $T\colon \mathcal{D} \mapsto \mathbb{R}^n$ is α-averaged if:

$$T = (1 - \alpha)\,\mathrm{Id} + \alpha R$$

- the fixed-points of T and R coincide
- composition of averaged operators is averaged
- if $\operatorname{Fix} T \neq \emptyset$, then iteration $x^{k+1} = Tx^k$ converges to some $\bar{x} \in \operatorname{Fix} T$
- $(x^{k+1} - x^k)$ always converges to some δx



Resolvent

- *resolvent* of a maximally monotone operator $A\colon \mathbb{R}^n \mapsto 2^{\mathbb{R}^n}$:

$$J_A = (\mathrm{Id} + A)^{-1}$$

- some important properties of resolvent J_A:
  - it full domain: $\operatorname{dom} J_A = \mathbb{R}^n$
  - it is single-valued
  - it is $\frac{1}{2}$-averaged
- $\operatorname{Fix} J_{\gamma A}$ coincides with the set of zeros of A:

$$\begin{aligned} 0 \in Ax \;&\Leftrightarrow\; x \in x + \gamma A x \\ &\Leftrightarrow\; x \in (\mathrm{Id} + \gamma A)x \\ &\Leftrightarrow\; x = (\mathrm{Id} + \gamma A)^{-1} x \\ &\Leftrightarrow\; x = J_{\gamma A} x \end{aligned}$$

- resolvent method: $x^{k+1} = J_{\gamma A} x^k$

Subdifferential and monotonicity

- assume $f\colon \mathbb{R}^n \mapsto \overline{\mathbb{R}}$ is a proper closed convex function
- then ∂f is maximally monotone
- let $A = \partial f$, then:

$$J_A x = \operatorname*{argmin}_y \{f(y) + \tfrac{1}{2}\|y - x\|_2^2\} =: \mathrm{prox}_f(x)$$

where prox_f is called the *proximal operator* of f

- proof: $z = \mathrm{prox}_f(x)$ if and only if

$$\begin{aligned} 0 \in \partial f(z) + z - x \;&\Leftrightarrow\; x \in \partial f(z) + z \\ &\Leftrightarrow\; x \in (\underline{\mathrm{Id} + \partial f) z} \\ &\Leftrightarrow\; z = (\mathrm{Id} + \partial f)^{-1} x = J\_A\,(x) \end{aligned}$$

- proximal operator can be seen as a generalization of projection:

$$\mathrm{prox}_{\mathcal{I}_\mathcal{C}}(x) = \operatorname*{argmin}_y \left\{\mathcal{I}_\mathcal{C}(y) + \tfrac{1}{2}\|y - x\|_2^2\right\} = \Pi_\mathcal{C}(x)$$

Proximal operator of separable functions

- consider a (block) separable function $g(x) = \sum_{i=1}^{n} g_i(x_i)$
- prox_f is (block) separable as well:

$$\begin{aligned}\mathrm{prox}_g(x) &= \underset{y}{\mathrm{argmin}} \left\{ g(y) + \tfrac{1}{2}\|y - x\|_2^2 \right\} \\ &= \underset{y}{\mathrm{argmin}} \left\{ \sum_{i=1}^{n} g_i(y_i) + \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i)^2 \right\} \\ &= \begin{bmatrix} \mathrm{argmin}_{x_1} \left\{ g_1(x_1) + \frac{1}{2}(y_1 - x_1)^2 \right\} \\ \vdots \\ \mathrm{argmin}_{x_n} \left\{ g_n(x_n) + \frac{1}{2}(y_n - x_n)^2 \right\} \end{bmatrix}\end{aligned}$$

- the proximal operator of $h = g \circ L$ (for an arbitrary matrix L) is

$$\mathrm{prox}_h(x) = \underset{y}{\mathrm{argmin}} \left\{ g(Ly) + \tfrac{1}{2}\|y - x\|_2^2 \right\}$$

- separability is lost in general

Moreau's identity

- proximal operators of f and f^* are related via the following identity:

$$\mathrm{prox}_f + \mathrm{prox}_{f^*} = \mathrm{Id}$$

- when f is scaled by $\gamma > 0$, we have

$$\mathrm{prox}_{\gamma f} + \mathrm{prox}_{(\gamma f)^*} = \mathrm{prox}_{\gamma f} + \gamma\, \mathrm{prox}_{\gamma^{-1} f^*} \circ \gamma^{-1}\, \mathrm{Id} = \mathrm{Id}$$

- when f is composed with L, we have

$$\mathrm{prox}_{\gamma(f \circ L)}(x) = x - \gamma L^T \mu^\star$$

where

$$\mu^\star \in \operatorname*{argmin}_{\mu} \left\{ f^*(\mu) + \tfrac{\gamma}{2} \|L^T \mu - \gamma^{-1} x\|_2^2 \right\}$$

(assuming the argmin is nonempty)

Monotone inclusion problems

- suppose A and B are maximally monotone operators
- we want to find x that solves the inclusion:

$$0 \in Ax + Bx$$

- there exist methods based on evaluating A, B, and their resolvents
- these methods can be used to solve

$$0 \in \partial f(x) + \partial g(x)$$

Forward-backward splitting

- suppose A and B are maximally monotone operators
- for any $\gamma > 0$, we have

$$\begin{aligned} 0 \in Ax + Bx &\Leftrightarrow -\gamma Bx \in \gamma Ax \\ &\Leftrightarrow (\mathrm{Id} - \gamma B)x \in (\mathrm{Id} + \gamma A)x \\ &\Leftrightarrow J_{\gamma A}(\mathrm{Id} - \gamma B)x = x \end{aligned}$$

- forward-backward splitting: $x^{k+1} = J_{\gamma A}(\mathrm{Id} - \gamma B)x^k$
- if $(\mathrm{Id} - \gamma B)$ is averaged and a fixed-point of the forward-backward operator exists, then the iteration converges

Proximal gradient method

- consider the composite minimization problem

$$\text{minimize} \quad f(x) + g(x)$$

  where f is β-smooth convex and g proper closed convex
- under suitable constraint qualification, it is equivalent to

$$0 \in \nabla f(x) + \partial g(x)$$

- FB splitting reduces to the *proximal gradient method*:

$$x^{k+1} = J_{\gamma \partial g}(\mathrm{Id} - \gamma \nabla f) x^k = \mathrm{prox}_{\gamma g}(\mathrm{Id} - \gamma \nabla f) x^k$$

- for $\gamma \in (0, \frac{2}{\beta})$, $(\mathrm{Id} - \gamma \nabla f)$ is $\frac{\gamma\beta}{2}$-averaged
- hence, the PG method converges to a fixed-point (provided it exists)
- if f is in addition strongly convex, then $(\mathrm{Id} - \gamma \nabla f)$ is contractive

Problems with composition

- consider the more general problem

$$\text{minimize} \quad f(x) + g(Lx)$$

  where f is β-smooth convex, g proper closed convex, L a matrix
- applying PG method gives:

$$x^{k+1} = \mathrm{prox}_{\gamma(g \circ L)}(\mathrm{Id} - \gamma \nabla f) x^k$$

- $\mathrm{prox}_{\gamma(g \circ L)}$ is often expensive to evaluate

Problems with composition

- consider the more general problem

$$\text{minimize} \quad f(x) + g(Lx)$$

  where f is β-smooth convex, g proper closed convex, L a matrix
- applying PG method gives:

$$x^{k+1} = \mathrm{prox}_{\gamma(g \circ L)}(\mathrm{Id} - \gamma \nabla f) x^k$$

- $\mathrm{prox}_{\gamma(g \circ L)}$ is often expensive to evaluate
- formulate dual problem:

$$\text{minimize} \quad f^*(-L^T \mu) + g^*(\mu)$$

- if f is σ-strongly convex, then $f^* \circ (-L^T)$ is $\frac{\|L\|_2^2}{\sigma}$-smooth

Solving the dual

$$\text{minimize} \quad f^*(-L^T\mu) + g^*(\mu)$$

- applying PG method to the dual gives:

$$\begin{aligned}\mu^{k+1} &= \operatorname{prox}_{\gamma g^*}\left(\operatorname{Id} - \gamma\nabla(f^* \circ (-L^T))\right)\mu^k \\ &= \operatorname{prox}_{\gamma g^*}\left(\mu^k + \gamma L \nabla f^*(-L^T\mu^k)\right)\end{aligned}$$

- the method converges for $\gamma \in (0, \frac{2\sigma}{\|L\|_2^2})$

Solving the dual

$$\text{minimize} \quad f^*(-L^T\mu) + g^*(\mu)$$

- applying PG method to the dual gives:

$$\begin{aligned}\mu^{k+1} &= \operatorname{prox}_{\gamma g^*}\left(\operatorname{Id} - \gamma\nabla(f^* \circ (-L^T))\right)\mu^k \\ &= \operatorname{prox}_{\gamma g^*}\left(\mu^k + \gamma L \nabla f^*(-L^T\mu^k)\right)\end{aligned}$$

- the method converges for $\gamma \in (0, \frac{2\sigma}{\|L\|_2^2})$
- letting $x^k = \nabla f^*(-L^T\mu^k)$, we obtain

$$\begin{aligned}x^k &= \nabla f^*(-L^T\mu^k) \\ \mu^{k+1} &= \operatorname{prox}_{\gamma g^*}\left(\mu^k + \gamma L x^k\right)\end{aligned}$$

Recovering the primal

- since the dual PG method converges to a fixed-point $\bar{\mu}$, we have

$$\bar{x} = \nabla f^*(-L^T \bar{\mu})$$
$$\bar{\mu} = \operatorname{prox}_{\gamma g^*}(\bar{\mu} + \gamma L \bar{x})$$

- Fermat's rule gives

$$0 \in \partial g^*(\bar{\mu}) + \gamma^{-1}(\bar{\mu} - (\bar{\mu} + \gamma L \bar{x})) = \partial g^*(\bar{\mu}) - L\bar{x}$$

- recall that the optimality conditions can be written as

$$\begin{cases} x \in \partial f^*(-L^T \mu) \\ Lx \in \partial g^*(\mu) \end{cases}$$

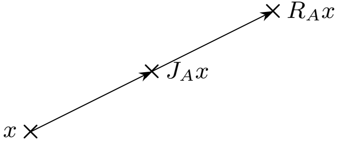
- therefore, the method outputs both primal and dual solutions

Reflected resolvent

- *reflected resolvent* of a maximally monotone operator $A\colon \mathbb{R}^n \mapsto 2^{\mathbb{R}^n}$:

$$R_A = 2J_A - \mathrm{Id}$$

- it gives the reflection point



- R_A is always nonexpansive
- if $A = \partial f$, then *reflected proximal operator* is

$$R_{\partial f} = 2\,\mathrm{prox}_f - \mathrm{Id} =: \mathrm{rprox}_f$$

- the following identity holds:

$$\begin{aligned} R_{\gamma A}(\mathrm{Id} + \gamma A) &= 2(\mathrm{Id} + \gamma A)^{-1}(\mathrm{Id} + \gamma A) - (\mathrm{Id} + \gamma A) \\ &= 2\,\mathrm{Id} - (\mathrm{Id} + \gamma A) \\ &= \mathrm{Id} - \gamma A \end{aligned}$$

Peaceman-Rachford splitting

- suppose A and B are maximally monotone operators
- then we have

$$
\begin{aligned}
0 \in Ax + Bx \;&\Leftrightarrow\; 0 \in (\mathrm{Id} + \gamma A)x - (\mathrm{Id} - \gamma B)x \\
&\Leftrightarrow\; 0 \in (\mathrm{Id} + \gamma A)x - R_{\gamma B}(\mathrm{Id} + \gamma B)x \\
&\Leftrightarrow\; 0 \in (\mathrm{Id} + \gamma A)x - R_{\gamma B}z, && z \in (\mathrm{Id} + \gamma B)x \\
&\Leftrightarrow\; R_{\gamma B}z \in (\mathrm{Id} + \gamma A)x, && z \in (\mathrm{Id} + \gamma B)x \\
&\Leftrightarrow\; J_{\gamma A}R_{\gamma B}z = J_{\gamma B}z, && x \in J_{\gamma B}z
\end{aligned}
$$

- finally, this is equivalent to

$$R_{\gamma A}R_{\gamma B}z = 2J_{\gamma A}R_{\gamma B}z - R_{\gamma B}z = 2J_{\gamma B}z - R_{\gamma B}z = z$$

- in other words, $0 \in Ax + Bx$ if and only if

$$z = R_{\gamma A}R_{\gamma B}z, \quad x = J_{\gamma B}z$$

- Peaceman-Rachford splitting: $z^{k+1} = R_{\gamma A}R_{\gamma B}z^k$

Douglas-Rachford splitting

- iterating $R_{\gamma A} \circ R_{\gamma B}$ may not converge as it is nonexpansive in general
- we instead iterate the averaged map (with $\alpha \in (0,1)$):

$$z^{k+1} = \left((1-\alpha)\,\mathrm{Id} + \alpha R_{\gamma A} R_{\gamma B}\right) z^k$$

- provided that a fixed-point exists, the method converges for any $\gamma > 0$
- convergence rate depends on the value of γ
- the algorithm can be implemented as

$$\begin{aligned} x^k &= J_{\gamma B}(z^k) \\ y^k &= J_{\gamma A}(2x^k - z^k) \\ z^{k+1} &= z^k + 2\alpha(y^k - x^k) \end{aligned}$$

Douglas-Rachford for optimization

- consider the composite minimization problem

$$\text{minimize} \quad f(x) + g(x)$$

  where f and g are proper closed convex
- under suitable constraint qualification, it is equivalent to

$$0 \in \partial f(x) + \partial g(x)$$

- DR splitting can be implemented as

$$\begin{aligned} x^k &= \operatorname{prox}_{\gamma f}(z^k) \\ y^k &= \operatorname{prox}_{\gamma g}(2x^k - z^k) \\ z^{k+1} &= z^k + 2\alpha(y^k - x^k) \end{aligned}$$

- z^k converges to a fixed-point of $\operatorname{rprox}_{\gamma g} \circ \operatorname{rprox}_{\gamma f}$
- x^k converges to a solution of the optimization problem
- if f is strongly convex and β-smooth, then $\operatorname{rprox}_{\gamma f}$ is contractive

Optimality conditions

- since DR splitting converges to a fixed-point $\bar{z}$, we have:

$$\begin{aligned}\bar{x} &= \mathrm{prox}_{\gamma f}(\bar{z}) \\ \bar{y} &= \mathrm{prox}_{\gamma g}(2\bar{x} - \bar{z}) \\ \bar{z} &= \bar{z} + 2\alpha(\bar{y} - \bar{x})\end{aligned}$$

- Fermat's rule gives

$$\begin{aligned}0 &\in \gamma\partial f(\bar{x}) + \bar{x} - \bar{z} \\ 0 &\in \gamma\partial g(\bar{y}) + \bar{y} - 2\bar{x} + \bar{z} \\ 0 &= \bar{y} - \bar{x}\end{aligned}$$

- letting $\mu = \frac{1}{\gamma}(\bar{x} - \bar{z})$, we obtain

$$\begin{aligned}0 &\in \partial f(\bar{x}) + \mu \\ 0 &\in \partial g(\bar{y}) - \mu \\ 0 &= \bar{y} - \bar{x}\end{aligned}$$

- therefore, $\bar{x} = \bar{y}$ is primal and μ is dual solution

References

- these lecture notes are based to a large extent on the Large-Scale Convex Optimization course developed by Pontus Giselsson at Lund
- the original slides can be downloaded from

`https://archive.control.lth.se/ls-convex-2015/`