Optimization for Data Science

Lec 11: Splitting

Yaoliang Yu

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS
DAVID R. CHERITON SCHOOL OF COMPUTER SCIENCE

Problem

Find zero of a maximal monotone map $\mathsf{T} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$:

$$\text{find } \mathbf{z} \ \text{ s.t. } \ \mathbf{0} \in \mathsf{T}\mathbf{z}, \quad \text{where} \quad \mathsf{T} = \mathsf{A} + \mathsf{B}$$

- $\mathsf{A}, \mathsf{B} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ can both be multi-valued
- Allow a unified treatment of many algorithms
- Dual problem:

$$\text{find } \mathbf{z}^* \ \text{ s.t. } \ \mathbf{0} \in \mathsf{T}^*\mathbf{z}^*, \quad \text{where} \quad \mathsf{T}^* := [-\mathsf{A}^{-1} \circ (-\mathrm{Id}) + \mathsf{B}^{-1}]$$

Some Definitions

- A multi-valued map $\mathsf{T} : \mathbb{R}^d \to \mathbb{R}^d$ is monotone iff for all $\mathbf{w}, \mathbf{z} \in \operatorname{dom} \mathsf{T}$, and all $\mathbf{w}^* \in \mathsf{T}\mathbf{w}, \mathbf{z}^* \in \mathsf{T}\mathbf{z}$,

$$\langle \mathbf{w} - \mathbf{z}, \mathbf{w}^* - \mathbf{z}^* \rangle \geq 0$$

- A monotone map is maximal if its graph is not contained in that of another monotone map:

$$\operatorname{gph} \mathsf{T} := \{(\mathbf{z}, \mathbf{z}^*) : \mathbf{z}^* \in \mathsf{T}\mathbf{z}\}$$

- Inverse: $\mathbf{z}^* \in \mathsf{T}\mathbf{z} \iff \mathbf{z} \in \mathsf{T}^{-1}\mathbf{z}^*$
- Resolvant: $J_{\mathsf{T}}^{\eta} := (\mathrm{Id} + \eta \mathsf{T})^{-1}$, i.e. $\mathbf{z}^* \in J_{\mathsf{T}}^{\eta}\mathbf{z} \iff \mathbf{z} \in \mathbf{z}^* + \eta \mathsf{T}\mathbf{z}^*$
- Sum: $(\mathsf{A} + \mathsf{B})(\mathbf{z}) = \{\mathbf{u}^* + \mathbf{v}^* : \mathbf{u}^* \in \mathsf{A}\mathbf{z}, \mathbf{v}^* \in \mathsf{B}\mathbf{z}\}$

Example: Convex minimization

Subdifferential ∂f of a convex function is maximal monotone.

$\mathbf{0} \in \partial f(\mathbf{z}) \iff \mathbf{z} \in \operatorname{argmin} f$.

$J^{\eta}_{\partial f}(\mathbf{z}) = \mathrm{P}^{\eta}_{f}(\mathbf{z}) = \operatorname{argmin}_{\mathbf{w}} \frac{1}{2\eta}\|\mathbf{w} - \mathbf{z}\|_2^2 + f(\mathbf{w})$

Example: Convex composite minimization

For convex f and g, let $\mathsf{A} = \partial f$, $\mathsf{B} = \partial g$ and $\mathsf{T} = \mathsf{A} + \mathsf{B} = \partial f + \partial g$.

$\mathbf{0} \in \mathsf{T}\mathbf{z} \iff \mathbf{z} \in \operatorname{argmin} f + g$

Example: Convex-concave minimax

For $f(\mathbf{x},\mathbf{y})$ convex in $\mathbf{x}$ and concave in $\mathbf{y}$, $\mathsf{T} = (\partial_{\mathbf{x}} f, \partial_{\mathbf{y}}\text{-}f)$ is maximal monotone.

$\mathbf{0} \in \mathsf{T}(\mathbf{x},\mathbf{y}) \iff (\mathbf{x},\mathbf{y}) \in \operatorname{argmin}\operatorname{argmax} f$, i.e. a Nash equilibrium.

- $\mathbf{0} \in \partial_{\mathbf{x}} f(\mathbf{x},\mathbf{y}) \iff$ fixing $\mathbf{y}$, $\mathbf{x} \in \operatorname{argmin} f(\cdot,\mathbf{y})$
- $\mathbf{0} \in \partial_{\mathbf{y}}\text{-}f(\mathbf{x},\mathbf{y}) \iff$ fixing $\mathbf{x}$, $\mathbf{y} \in \operatorname{argmax} f(\mathbf{x},\cdot)$

Example: Constrained convex-concave minimax

For $f(\mathbf{x},\mathbf{y})$ convex in $\mathbf{x}$ and concave in $\mathbf{y}$, $g(\mathbf{x})$ convex and $h(\mathbf{y})$ concave, let $\mathsf{A} = (\partial_{\mathbf{x}} f, \partial_{\mathbf{y}}\text{-}f)$ and $\mathsf{B} = (\partial_{\mathbf{x}} g, \partial_{\mathbf{y}}\text{-}h)$.

$\mathbf{0} \in \mathsf{T}(\mathbf{x},\mathbf{y}) \iff (\mathbf{x},\mathbf{y}) \in \operatorname{argmin}_{\mathbf{x}}\operatorname{argmax}_{\mathbf{y}} f(\mathbf{x},\mathbf{y}) + g(\mathbf{x}) + h(\mathbf{y})$, i.e. a Nash equilibrium.

A Product Space Trick

$$\min_{\mathbf{w}} \sum_{i=1}^{m} f_i(\mathbf{w})$$

- m users, each having an objective f_i to optimize
- Users share the same model $\mathbf{w}$
- Introduce local copies $\mathbf{w}_1, \ldots, \mathbf{w}_m$, and add consensus constraint:

$$\min_{\mathbf{w}_1,\ldots,\mathbf{w}_m} \sum_{i=1}^{m} f_i(\mathbf{w}_i) + \iota_H(\mathbf{w}_1, \ldots, \mathbf{w}_m)$$

  - $H := \{(\mathbf{w}_1, \ldots, \mathbf{w}_m) : \mathbf{w}_1 = \mathbf{w}_2 = \cdots = \mathbf{w}_m\}$
- $J_{\partial \iota_H}(\mathbf{w}_1, \cdots, \mathbf{w}_m) = \mathrm{P}_H(\mathbf{w}_1, \ldots, \mathbf{w}_m) = (\bar{\mathbf{w}}, \bar{\mathbf{w}}, \cdots, \bar{\mathbf{w}})$, where $\bar{\mathbf{w}} := \frac{1}{m}\sum_{i=1}^{m} \mathbf{w}_i$

Algorithm 1: Forward-Backward Splitting

Input: $\mathbf{w}_0 \in \operatorname{dom} \mathsf{A}$

for $t = 0, 1, 2, \ldots$ **do**

  choose any $\mathbf{a}_t^* \in \mathsf{A}\mathbf{w}_t$

  $\mathbf{w}_{t+1} \leftarrow J_{\mathsf{B}}^{\eta_t}(\mathbf{w}_t - \eta_t \mathbf{a}_t^*)$ // $\eta_t \geq 0$ is the step size

  $\mathbf{z}_t \leftarrow \sum_{k=0}^{t} \bar{\eta}_{t,k} \mathbf{w}_k$ // ergodic averaging, $\bar{\eta}_{t,k} := \eta_k / H_t$, $H_t := \sum_{k=0}^{t} \eta_k$

G. B. Passty. "Ergodic convergence to a zero of the sum of monotone operators in Hilbert space". *Journal of Mathematical Analysis and Applications*, vol. 72, no. 2 (1979), pp. 383–390, R. E. Bruck. "On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space". *Journal of Mathematical Analysis and Applications*, vol. 61, no. 1 (1977), pp. 159–164.

Theorem:

For any $(\mathbf{w}, \mathbf{w}^*) \in \operatorname{gph} \mathsf{T}$ and $\mathbf{b}^* \in \mathsf{B}\mathbf{w}$:

$$\langle \mathbf{z}_t - \mathbf{w}, \mathbf{w}^* \rangle \le \sum_{k=0}^{t} \bar{\eta}_{t,k} \langle \mathbf{w}_k - \mathbf{w}, \mathbf{a}_k^* + \mathbf{b}^* \rangle \le \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \sum_{k=0}^{t} \eta_k^2 \|\mathbf{a}_k^* + \mathbf{b}^*\|_2^2}{2H_t}.$$

- If $\sum_t \|\eta_t \mathbf{a}_t^* + \mathbf{b}^*\|_2^2 < \infty$ and $H_t \to \infty$, then either no solution, in which case $\|\mathbf{z}_t\| \to \infty$; or $\mathbf{z}_t$ converges to a solution

Gradient-Descent-Ascent (GDA)

For simplicity, suppose we can choose $\mathbf{b}^* = \mathbf{0}$:

$$\frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2 + \sum_{k=0}^t \|\eta_k \mathbf{a}_k^*\|_2^2}{2H_t} \geq \sum_{k=0}^t \bar{\eta}_{t,k} \langle \mathbf{w}_k - \mathbf{w}, \mathbf{a}_k^* \rangle$$

$$= \sum_{k=0}^t \bar{\eta}_{t,k} \left[\langle \mathbf{x}_k - \mathbf{x}, \partial_{\mathbf{x}} f(\mathbf{x}_k, \mathbf{y}_k) \rangle + \langle \mathbf{y}_k - \mathbf{y}, \partial_{\mathbf{y}} - f(\mathbf{x}_k, \mathbf{y}_k) \rangle \right]$$

$$\geq \sum_{k=0}^t \bar{\eta}_{t,k} \left[-f(\mathbf{x}, \mathbf{y}_k) + f(\mathbf{x}_k, \mathbf{y}) \right]$$

$$\geq -f(\mathbf{x}, \bar{\mathbf{y}}_t) + f(\bar{\mathbf{x}}_t, \mathbf{y}), \quad \text{where} \quad (\bar{\mathbf{x}}_t, \bar{\mathbf{y}}_t) := \sum_{k=0}^t \bar{\eta}_{t,k} \mathbf{w}_k$$

To satisfy $\sum_t \|\eta_t \mathbf{a}_t^*\|_2^2 < \infty$ we may choose $\eta_t = \frac{1}{\sqrt{\|\mathbf{a}_t^*\|_2^2 + 1}} \frac{1}{(t+1)^p}, \quad p \in (\frac{1}{2}, 1]$

A. S. Nemirovskii and D. B. Judin. "Cesari convergence of the gradient method of approximating saddle points of convex-concave functions". *Soviet Mathematics Doklady*, vol. 19, no. 2 (1978), pp. 482–486.

Algorithm 2: Backward-Backward Splitting

Input: $\mathbf{w}_0 \in \mathrm{dom}\,\mathsf{A}$

for $t = 0, 1, 2, \ldots$ **do**

  $\mathbf{w}_{t+1} \leftarrow J_{\mathsf{B}}^{\eta_t} J_{\mathsf{A}}^{\eta_t} \mathbf{w}_t$ `//` $\eta_t \geq 0$ `is the step size`

  $\mathbf{z}_t \leftarrow \sum_{k=0}^{t} \bar{\eta}_{t,k} \mathbf{w}_k$ `// ergodic averaging,` $\bar{\eta}_{t,k} := \eta_k / H_t, \;\; H_t := \sum_{k=0}^{t} \eta_k$

Theorem:

If $\sum_t \eta_t = \infty$ and $\eta_t \to 0$, then either there is no solution, in which case $\|\mathbf{z}_t\| \to \infty$, or $\mathbf{z}_t$ converges to a solution.

G. B. Passty. "Ergodic convergence to a zero of the sum of monotone operators in Hilbert space". *Journal of Mathematical Analysis and Applications*, vol. 72, no. 2 (1979), pp. 383–390, P.-L. Lions. "Une methode iterative de resolution d'une inequation variationnelle". *Israel Journal of Mathematics*, vol. 31, no. 2 (1978), pp. 204–208.

Comparing FB with BB

$$\mathfrak{F} := J^\eta_{\mathsf{B}}(\mathrm{Id} - \eta \mathsf{A}) \;\; vs. \;\; \mathfrak{B} := J^\eta_{\mathsf{B}} J^\eta_{\mathsf{A}}, \quad \text{where} \quad J^\eta_{\mathsf{T}} := (\mathrm{Id} + \eta \mathsf{T})^{-1}$$

- Let us define ${}^\eta\mathsf{A} := \frac{\mathrm{Id} - J^\eta_{\mathsf{A}}}{\eta}$ such that ${}^\eta\mathsf{A} \to {}^0\mathsf{A} \subseteq \mathsf{A}$ if $\eta \to 0$
- Then, backward-backward on $\mathsf{A} + \mathsf{B}$ is forward-backward on ${}^\eta\mathsf{A} + \mathsf{B}$!
  - $\mathrm{Id} - \eta \cdot {}^\eta\mathsf{A} = J^\eta_{\mathsf{A}}$
- Consider $\mathsf{A} = \partial f$ for a convex function f:
  - $J^\eta_{\mathsf{A}} = \mathrm{P}^\eta_f$
  - ${}^\eta\mathsf{A} = \frac{\mathrm{Id} - \mathrm{P}^\eta_f}{\eta} = \nabla \mathrm{M}^\eta_f$
  - backward-backward on $f + g$ is the same as forward-backward on $\mathrm{M}^\eta_f + g$!
- For small η, $J^\eta_{\mathsf{A}} = (\mathrm{Id} + \eta \mathsf{A})^{-1} \approx \mathrm{Id} - \eta \mathsf{A}$ when A is linear

H. H. Bauschke, P. L. Combettes, and S. Reich. "The asymptotic behavior of the composition of two resolvents". *Nonlinear Analysis: Theory, Methods & Applications*, vol. 60, no. 2 (2005), pp. 283–301.

Algorithm 3: The Barycenter Method

Input: $\mathbf{w}_0$

for $t = 0, 1, 2, \ldots$ **do**

  $\mathbf{w}_{t+1} \leftarrow \mathrm{Avg}(\mathrm{P}_{i_1}, \ldots, \mathrm{P}_{i_{k(t)}})\mathbf{w}_t$

- Let $H_i := \{\mathbf{w} : \langle \mathbf{w}, \mathbf{a}_i \rangle = b_i\}$ and P_i be the orthogonal projection onto it
- Kaczmarz's method: $k(t) \equiv 1$
- Cimmino's method: $k(t) \equiv m$
- A randomized version of backward-backward!

S. Kaczmarz. "Angenäherte Auflösung von Systemen linearer Gleichunger". *Bulletin International de l'Académie Polonaise des Sciences et des Lettres*, vol. 35 (1937), pp. 355–357. "Approximate solution of systems of linear equations", English translation in International Journal of Control, 1993, vol. 57, no.6, pp. 1269–1271, G. Cimmino. "Calcolo Approssimato Per le Soluzioni dei Sistemi di Equazioni Lineari". *La Ricerca Scientifica*, vol. 9, no. 1 (1938), pp. 326–333.

$$\begin{cases} x + y = 0 \\ x - y = 0 \end{cases}$$

- $\mathrm{P}_1(x, y) = \left(\frac{x-y}{2}, \frac{y-x}{2}\right)$
- $\mathrm{P}_2(x, y) = \left(\frac{x+y}{2}, \frac{y+x}{2}\right)$
- $\mathrm{Avg}(\mathrm{P}_1, \mathrm{P}_2)(x, y) = \left(\frac{x}{2}, \frac{y}{2}\right) \to (0, 0)$

Algorithm 4: Extragradient for finding a zero of $\mathsf{T} = \mathsf{A} + \mathsf{B}$

Input: $\mathbf{w}_0 \in \operatorname{dom} \mathsf{A} \subseteq \operatorname{dom} \mathsf{B}$

for $t = 0, 1, 2, \ldots$ **do**

  choose step size $\eta_t > 0$

  $\tilde{\mathbf{w}}_t = J_{\mathsf{A}}^{\eta_t}(\mathbf{w}_t - \eta_t \mathsf{B} \mathbf{w}_t)$ `// peek`

  $\mathbf{w}_{t+1} = J_{\mathsf{A}}^{\eta_t}(\mathbf{w}_t - \eta_t \mathsf{B} \tilde{\mathbf{w}}_t)$ `// update with peeked information`

Theorem: Convergence of extra-gradient

Let $\mathsf{B} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be L-Lipschitz and $\operatorname{dom} \mathsf{A} \subseteq \operatorname{dom} \mathsf{B}$. Set $\eta_t \in [0, 1/\mathsf{L}]$. Then, for all $(\mathbf{w}, \mathbf{w}^*) \in \operatorname{gph} \mathsf{T}, \mathbf{a}^* \in \mathsf{A}\mathbf{w}$,

$$\langle \tilde{\mathbf{z}}_t - \mathbf{w}, \mathbf{w}^* \rangle \leq \sum_{k=0}^{t} \bar{\eta}_{t,k} \langle \tilde{\mathbf{w}}_k - \mathbf{w}, \mathsf{B}\tilde{\mathbf{w}}_k + \mathbf{a}^* \rangle \leq \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2}{2H_t}, \quad \text{where}$$

$$\tilde{\mathbf{z}}_t := \sum\nolimits_{k=0}^{t} \bar{\eta}_{t,k} \tilde{\mathbf{w}}_k, \quad \bar{\eta}_{t,k} := \eta_k / H_t, \qquad H_t := \sum\nolimits_{k=0}^{t} \eta_k.$$

G. M. Korpelevich. "The extragradient method for finding saddle points and other problems". *Ekonomika i matematicheskie metody*, vol. 12, no. 4 (1976), pp. 747–756.

- If $H_t \to \infty$, then either no solution and $\|\tilde{\mathbf{z}}_t\| \to \infty$, or $\tilde{\mathbf{z}}_t$ converges to a solution

- If $0 < \liminf_t \eta_t \leq \limsup_t \eta_t < 1/\mathsf{L}$ and assume a solution exists, then $\mathbf{w}_t - \tilde{\mathbf{w}}_t \to 0$ and $\tilde{\mathbf{w}}_t$ converges to a solution

- With $\eta_t \equiv \eta$ we obtain $O(1/t)$ rate of convergence for the averaged sequence $\tilde{\mathbf{z}}_t$

- Significantly faster than FB and BB (at best $O(1/\sqrt{t})$), under the additional assumption that B is Lipschitz continuous

- The direct sequence $\mathbf{w}_t$ converges at the slower $O(1/\sqrt{t})$ rate!

N. Golowich, S. Pattathil, C. Daskalakis, and A. Ozdaglar. "Last Iterate is Slower than Averaged Iterate in Smooth Convex-Concave Saddle Point Problems". In: *Proceedings of Thirty Third Conference on Learning Theory*. 2020, pp. 1758–1784.

Algorithm 5: Khobotov's linear search

$\eta_t \leftarrow 2\bar{\eta}$
$\mathbf{w}_t^* \leftarrow \mathsf{B}\mathbf{w}_t$
repeat
  $\eta_t \leftarrow \eta_t/2$
  $\tilde{\mathbf{w}}_t \leftarrow J_{\mathsf{A}}^{\eta_t}(\mathbf{w}_t - \eta_t \mathbf{w}_t^*)$
until $\eta_t \leq \gamma \frac{\|\mathbf{w}_t - \tilde{\mathbf{w}}_t\|_2}{\|\mathbf{w}_t^* - \mathsf{B}\tilde{\mathbf{w}}_t\|_2}$

- Inspecting the proof to see where L is used
- A constant number of line searches $\implies$ same convergence rate

E. N. Khobotov. "Modification of the extra-gradient method for solving variational inequalities and certain optimization problems". *USSR Computational Mathematics and Mathematical Physics*, vol. 27, no. 5 (1987), pp. 120–127.

Algorithm 6: Past extragradient for solving a smooth monotone VI

Input: $\mathbf{w}_0 = \tilde{\mathbf{w}}_{-1} \in C \subseteq \operatorname{dom} \mathsf{T}$

for $t = 0, 1, 2, \ldots$ **do**

  choose step size $\eta_t > 0$

  $\tilde{\mathbf{w}}_t = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T}\tilde{\mathbf{w}}_{t-1})$

  $\mathbf{w}_{t+1} = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T}\tilde{\mathbf{w}}_t)$

- Only requires 1 evaluation of the operator T
- But still 2 projections per step
- Compared to the extragradient, simply recycle the past evaluation $\mathsf{T}\tilde{\mathbf{w}}_{t-1}$ to replace $\mathsf{T}\mathbf{w}_t$, saving us 1 evaluation of T

L. D. Popov. "A modification of the Arrow-Hurwicz method for search of saddle points". *Mathematical notes of the Academy of Sciences of the USSR*, vol. 28, no. 5 (1980), pp. 845–848.

Algorithm 7: Tseng's modified forward-backward splitting

Input: $\mathbf{w}_0 \in C \subseteq \operatorname{dom} \mathsf{T}$

for $t = 0, 1, 2, \ldots$ **do**

  choose step size $\eta_t > 0$

  $\tilde{\mathbf{w}}_t = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T}\mathbf{w}_t)$

  $\mathbf{w}_{t+1} = \tilde{\mathbf{w}}_t - \eta_t(\mathsf{T}\tilde{\mathbf{w}}_t - \mathsf{T}\mathbf{w}_t)$

- Only requires 1 projection
- But still 2 evaluations of T per step
- This variant requires say $\operatorname{dom} \mathsf{T} = \mathbb{R}^d$

P. Tseng. "A Modified Forward-Backward Splitting Method for Maximal Monotone Mappings". *SIAM Journal on Control and Optimization*, vol. 38, no. 2 (2000), pp. 431–446.

Algorithm 8: Optimistic extragradient for solving a smooth monotone VI

Input: $\mathbf{w}_0 = \tilde{\mathbf{w}}_{-1} \in C \subseteq \operatorname{dom} \mathsf{T}$

for $t = 0, 1, 2, \ldots$ **do**

choose step size $\eta_t > 0$

$\tilde{\mathbf{w}}_t = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T}\tilde{\mathbf{w}}_{t-1})$

$\mathbf{w}_{t+1} = \tilde{\mathbf{w}}_t - \eta_t(\mathsf{T}\tilde{\mathbf{w}}_t - \mathsf{T}\tilde{\mathbf{w}}_{t-1})$

- Obviously, we can combine the previous two ideas
- Obtain a variant that only requires 1 projection and 1 evaluation of T per step!

C. Daskalakis, A. Ilyas, V. Syrgkanis, and H. Zeng. "Training GANs with optimism". In: *The 6th International Conference on Learning Representations*. 2018.

Algorithm 9: Reflected extragradient for solving a smooth monotone VI

Input: $\mathbf{w}_0 = \mathbf{w}_{-1} \in C \subseteq \operatorname{dom} \mathsf{T}$

for $t = 0, 1, 2, \ldots$ **do**

  choose step size $\eta_t > 0$

  $\tilde{\mathbf{w}}_t = 2\mathbf{w}_t - \mathbf{w}_{t-1}$

  $\mathbf{w}_{t+1} = \mathrm{P}_C(\mathbf{w}_t - \eta_t \mathsf{T}\tilde{\mathbf{w}}_t)$

- Uses reflection and also enjoyes 1 projection and 1 evaluation of T per step

- Requires say $\operatorname{dom} \mathsf{T} \supseteq 2C - C$

Y. Malitsky. "Projected Reflected Gradient Methods for Monotone Variational Inequalities". *SIAM Journal on Optimization*, vol. 25, no. 1 (2015), pp. 502–520.

Reflectors

- $R_{\mathsf{T}}^{\eta} := 2J_{\mathsf{T}}^{\eta} - \mathrm{Id} \subseteq (\mathrm{Id} - \eta\mathsf{T})(\mathrm{Id} + \eta\mathsf{T})^{-1}$
  - a backward step, followed by a forward step
- $(\mathrm{Id} - \eta\mathsf{T})(\mathrm{Id} + \eta\mathsf{T})^{-1} \neq (\mathrm{Id} + \eta\mathsf{T})^{-1}(\mathrm{Id} - \eta\mathsf{T}) = (\mathrm{Id} + \eta\mathsf{T})^{-1}(\mathrm{Id} - \eta\mathsf{T})(\mathrm{Id} + \eta\mathsf{T})^{-1}(\mathrm{Id} + \eta\mathsf{T})$
- $\mathbf{w} \in R_{\mathsf{T}}^{\eta}\mathbf{w} \iff \mathbf{w} \in J_{\mathsf{T}}^{\eta}(\mathbf{w}) \iff \mathbf{0} \in \mathsf{T}\mathbf{w}$

Algorithm 10: A general splitting algorithm based on reflectors

Input: $\mathbf{w}_0$

for $t = 0, 1, \ldots$ **do**

  choose step size $\eta_t \geq 0$ and relaxation size $\gamma_t \in [0, 1]$

  $\mathbf{w}_{t+1} = (1 - \gamma_t)\mathbf{w}_t + \gamma_t R_{\mathsf{B}}^{\eta_t} R_{\mathsf{A}}^{\eta_t}(\mathbf{w}_t) + \boldsymbol{\epsilon}_t$ `// allow error` $\boldsymbol{\epsilon}_t$

return $\mathbf{z} \in J_{\mathsf{A}}^{\eta}\mathbf{w}$ `// assuming the for-loop returns` $\mathbf{w}$ `and` $\eta_t \equiv \eta$

- $\gamma_t \equiv 1$: Peaceman-Rachford (PR) splitting
- $\gamma_t \equiv \frac{1}{2}$: Douglas-Rachford (DR) splitting
- $\gamma_t \in [\frac{1}{2}, 1]$: over-relaxation; $\gamma_t \in [0, \frac{1}{2}]$: under-relaxation

J. Eckstein and D. P. Bertsekas. "On the Douglas—Rachford splitting method and the proximal point algorithm for maximal monotone operators". *Mathematical Programming*, vol. 55 (1992), pp. 293–318.

General Convergence

Let $\mathsf{A}, \mathsf{B} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be maximal monotone and $\mathsf{T} = \mathsf{A} + \mathsf{B}$. Suppose $\sum_t \gamma_t(1-\gamma_t) = \infty$, $\eta_t \equiv \eta > 0$, and $\sum_t \|\boldsymbol{\epsilon}_t\|_2 < \infty$.

- If $\mathsf{T}^{-1}\mathbf{0} = (\mathsf{A}+\mathsf{B})^{-1}\mathbf{0} = \emptyset$, then $\{\mathbf{w}_t\}$ is unbounded.
- If $\mathsf{T}^{-1}\mathbf{0} \neq \emptyset$, then $\mathbf{w}_t \rightharpoonup \mathbf{w}_\infty = R^\eta_{\mathsf{B}} R^\eta_{\mathsf{A}} \mathbf{w}_\infty$, $\mathbf{z}_t := J^\eta_{\mathsf{A}} \mathbf{w}_t \rightharpoonup J^\eta_{\mathsf{A}} \mathbf{w}_\infty \in \mathsf{T}^{-1}\mathbf{0}$ and $\mathbf{z}^*_t := {}^\eta\mathsf{A}\mathbf{w}_t \rightharpoonup {}^\eta\mathsf{A}\mathbf{w}_\infty \in \mathsf{T}^{*-1}\mathbf{0}$, where recall that $\mathsf{T}^* := -\mathsf{B}^{-1} \circ (-\mathrm{Id}) + \mathsf{A}^{-1}$ and ${}^\eta\mathsf{A} := (\mathrm{Id} - J^\eta_{\mathsf{A}})/\eta$.

B. F. Svaiter. "On Weak Convergence of the Douglas-Rachford Method". *SIAM Journal on Control and Optimization*, vol. 49, no. 1 (2011), pp. 280–287.

Convergence of PR

Let $\mathsf{A}, \mathsf{B} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ be maximal monotone and $\mathsf{T} = \mathsf{A} + \mathsf{B}$. Consider PR with $\gamma_t \equiv 1$, $\eta_t \equiv \eta > 0$, and $\sum_t \|\boldsymbol{\epsilon}_t\|_2 < \infty$. Assume A is strictly monotone:

- There exists at most one zero $\mathbf{z}$ of T
- If $\mathbf{z} = \mathsf{T}^{-1}\mathbf{0} = (\mathsf{A} + \mathsf{B})^{-1}\mathbf{0}$ exists, then $\mathbf{z}_t := J_{\mathsf{A}}^{\eta}\mathbf{w}_t \rightharpoonup \mathbf{z}$
- If $\mathsf{T}^{-1}\mathbf{0} = (\mathsf{A} + \mathsf{B})^{-1}\mathbf{0} = \emptyset$, then $\{\mathbf{w}_t\}$ is unbounded

P.-L. Lions and B. Mercier. "Splitting Algorithms for the Sum of Two Nonlinear Operators". *SIAM Journal on Numerical Analysis*, vol. 16, no. 6 (1979), pp. 964–979.

Non-convergence of PR

Consider the rotation in $\mathbb{R}^2$:

$$\mathsf{A} = \mathsf{B} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad \text{where} \quad (\mathsf{A}+\mathsf{B})\mathbf{z} = \mathbf{0} \iff \mathbf{z} = \mathbf{0}.$$

Since $\langle \mathbf{w}, \mathsf{A}\mathbf{w}\rangle = 0$ for all $\mathbf{w}$, we know A is maximal monotone. We have

$$J_{\mathsf{A}}^{\eta} = \frac{1}{(1+\eta)^2}\begin{bmatrix} 1 & \eta \\ -\eta & 1 \end{bmatrix}, \quad R_{\mathsf{A}}^{\eta} = \frac{1}{(1+\eta)^2}\begin{bmatrix} 1-\eta^2 & 2\eta \\ -2\eta & 1-\eta^2 \end{bmatrix}.$$

Since both J_{A}^{η} and R_{A}^{η} are rotations (i.e. $\det = 1$), with $\gamma_t \equiv 1$ and $\boldsymbol{\epsilon}_t \equiv \mathbf{0}$, $\|\mathbf{z}_t\|_2 \equiv \|\mathbf{w}_0\|_2$ hence may not converge to any point (hence also $\mathbf{0}$, the unique zero of $\mathsf{A}+\mathsf{B}$). Moreover, $\mathbf{w}_t$ may not converge (to any point) either.

We verify that A is not strictly monotone and convince yourself that $\mathbf{z}_t$ and $\mathbf{w}_t$ do converge if we choose say $\gamma_t \equiv \gamma \in (0,1)$.

Alternating Direction Method of Multipliers (ADMM)

Consider the generic minimization problem

$$\inf_{\mathbf{a}} \; g(L\mathbf{a}) + h(\mathbf{a}), \text{ or equivalently } \inf_{\mathbf{a},\mathbf{b}} \; g(\mathbf{b}) + h(\mathbf{a}), \quad \text{s.t.} \quad L\mathbf{a} = \mathbf{b},$$

and its Fenchel-Rockafellar dual

$$-\inf_{\boldsymbol{\mu}} \; h^*(-L^\top \boldsymbol{\mu}) + g^*(\boldsymbol{\mu}),$$

where $g : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ and $h : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$ are closed proper convex and $L : \mathbb{R}^d \to \mathbb{R}^p$ is linear.

Introducing the Lagrangian multiplier $\boldsymbol{\mu}$ we obtain the Lagrangian:

$$\inf_{\mathbf{a},\mathbf{b}} \sup_{\boldsymbol{\mu}} \ g(\mathbf{b}) + h(\mathbf{a}) + \langle L\mathbf{a} - \mathbf{b}; \boldsymbol{\mu}\rangle \quad = \quad \sup_{\boldsymbol{\mu}} \underbrace{\inf_{\mathbf{a},\mathbf{b}} \ g(\mathbf{b}) + h(\mathbf{a}) + \langle L\mathbf{a} - \mathbf{b}; \boldsymbol{\mu}\rangle}_{\underline{L}(\boldsymbol{\mu})}.$$

- May apply Uzawa's algorithm to maximize the dual function $\underline{L}(\boldsymbol{\mu})$
- Nonsmooth hence requires diminishing step sizes

Instead, consider the augmented Lagrangian, where the penalty parameter η need not increase to ∞:

$$\sup_{\boldsymbol{\mu}} \left[\inf_{\mathbf{a},\mathbf{b}} \ g(\mathbf{b}) + h(\mathbf{a}) + \langle L\mathbf{a} - \mathbf{b}; \boldsymbol{\mu}\rangle + \tfrac{\eta}{2}\|L\mathbf{a} - \mathbf{b}\|_2^2\right] \equiv \sup_{\boldsymbol{\mu}} \left[\sup_{\boldsymbol{\nu}} \underline{L}(\boldsymbol{\nu}) - \tfrac{1}{2\eta}\|\boldsymbol{\nu} - \boldsymbol{\mu}\|_2^2\right],$$

- smooth inner function
- May apply Uzawa's algorithm with constant step size η:

$$\boldsymbol{\mu}_{t+1} \leftarrow \boldsymbol{\mu}_t + \eta(L\mathbf{a}_{t+1} - \mathbf{b}_{t+1})$$

- Solving $\mathbf{a}$ and $\mathbf{b}$ simultaneously in the augmented Lagrangian is challenging
- Fortunately, may apply just one step of alternating minimization to $\mathbf{a}$ and $\mathbf{b}$ sequentially:

$$\mathbf{a}_{t+1} \in \operatorname*{argmin}_{\mathbf{a}} h(\mathbf{a}) + \langle L\mathbf{a} - \mathbf{b}_t; \boldsymbol{\mu}_t \rangle + \tfrac{\eta}{2}\|L\mathbf{a} - \mathbf{b}_t\|_2^2 \equiv h(\mathbf{a}) + \tfrac{\eta}{2}\|L\mathbf{a} - \mathbf{b}_t + \boldsymbol{\mu}_t/\eta\|_2^2$$

$$\mathbf{b}_{t+1} = \operatorname*{argmin}_{\mathbf{b}} g(\mathbf{b}) + \langle L\mathbf{a}_{t+1} - \mathbf{b}; \boldsymbol{\mu}_t \rangle + \tfrac{\eta}{2}\|L\mathbf{a}_{t+1} - \mathbf{b}\|_2^2 \equiv g(\mathbf{b}) + \tfrac{\eta}{2}\|L\mathbf{a}_{t+1} - \mathbf{b} + \boldsymbol{\mu}_t/\eta\|_2^2$$

To understand the above updates, let us apply the Fenchel-Rockafellar duality again:

$$\mathbf{a}^*_{t+1} - \eta L\mathbf{a}_{t+1} = -\eta\mathbf{b}_t + \boldsymbol{\mu}_t, \quad \mathbf{a}^*_{t+1} = \operatorname*{argmin}_{\mathbf{a}^*} \ \tfrac{1}{2\eta}\|\mathbf{a}^* + \eta\mathbf{b}_t - \boldsymbol{\mu}_t\|_2^2 + h^*(-L^\top \mathbf{a}^*)$$

$$\mathbf{b}^*_{t+1} + \eta\mathbf{b}_{t+1} = \eta L\mathbf{a}_{t+1} + \boldsymbol{\mu}_t, \quad \mathbf{b}^*_{t+1} = \operatorname*{argmin}_{\mathbf{b}^*} \ \tfrac{1}{2\eta}\|\mathbf{b}^* - \eta L\mathbf{a}_{t+1} - \boldsymbol{\mu}_t\|_2^2 + g^*(\mathbf{b}^*)$$

It can be shown that $\boldsymbol{\mu}_t = \mathbf{b}^*_t$.

From the optimality conditions of $\mathbf{b}^*_{t+1}$ and $\mathbf{a}_{t+1}$ we verify that:

$$\begin{aligned} (\mathbf{b}^*_{t+1}, \mathbf{b}_{t+1}) \in \operatorname{gph} \partial g^* &\implies \mathbf{b}^*_{t+1} = J^\eta_{\partial g^*} \mathbf{w}_{t+1}, \quad \mathbf{w}_{t+1} := \eta \mathbf{b}_{t+1} + \mathbf{b}^*_{t+1}, \\ (\mathbf{a}^*_{t+1}, \mathbf{a}_{t+1}) \in \operatorname{gph}[\partial h^* \circ (-L^\top)] &\implies \mathbf{a}^*_{t+1} = J^\eta_{-L\circ\partial h^* \circ(-L^\top)}(-\eta L \mathbf{a}_{t+1} + \mathbf{a}^*_{t+1}) \\ (\text{ since } \boldsymbol{\mu}_t = \mathbf{b}^*_t) & \qquad\qquad = J^\eta_{-L\circ\partial h^* \circ(-L^\top)}(-\eta \mathbf{b}_t + \mathbf{b}^*_t) \end{aligned}$$

Therefore, we deduce that

$$\begin{aligned} \mathbf{w}_{t+1} &:= \eta \mathbf{b}_{t+1} + \mathbf{b}^*_{t+1} = \eta L \mathbf{a}_{t+1} + \boldsymbol{\mu}_t = \mathbf{a}^*_{t+1} + \eta \mathbf{b}_t = J^\eta_{-L\circ\partial h^* \circ(-L^\top)}(-\eta \mathbf{b}_t + \mathbf{b}^*_t) + \eta \mathbf{b}_t \\ &= J^\eta_{-L\circ\partial h^* \circ(-L^\top)}(2\mathbf{b}^*_t - \mathbf{w}_t) + \mathbf{w}_t - \mathbf{b}^*_t = \frac{\mathrm{Id} + R^\eta_{-L\circ\partial h^*\circ(-L^\top)} R^\eta_{\partial g^*}}{2} \mathbf{w}_t, \end{aligned}$$

Exactly the Douglas-Rachford algorithm applied to the dual problem, with the maximal monotone maps ∂h^* and $-L \circ \partial g^* \circ (-L^\top)$!

D. Gabay. "Applications of the Method of Multipliers to Variational Inequalities". In: *Augmented Lagrangian methods: Applications to the numerical solution of boundary-value problems.* Vol. 15. 9. 1983, pp. 299–331.