Dual and primal-dual methods

Ye Shi

ShanghaiTech University

Outline

- Dual proximal gradient method
- Primal-dual proximal gradient method

Dual proximal gradient method

Constrained convex optimization

$$
\begin{array}{ll}
\text{minimize}_{\boldsymbol{x}} & f(\boldsymbol{x}) \\
\text{subject to} & \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b} \in \mathcal{C}
\end{array}
$$

where f is convex, and $\mathcal{C}$ is convex set

- projection onto such a feasible set could sometimes be highly nontrivial (even when projection onto $\mathcal{C}$ is easy)

Constrained convex optimization

$f(x) + h(x)$

More generally, consider

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) + h(\boldsymbol{A}\boldsymbol{x})$$

where f and h are convex

$z = Ax$

$$\left.\begin{array}{ll} \min & f(x) + h(z) \\ \text{s.t.} & z = Ax \end{array}\right\}$$

- computing the proximal operator w.r.t. $\tilde{h}(\boldsymbol{x}) := h(\boldsymbol{A}\boldsymbol{x})$ could be difficult (even when prox_h is inexpensive)

A possible route: dual formulation

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) + h(\boldsymbol{A}\boldsymbol{x})$$

$$\Updownarrow \quad \text{add auxiliary variable } \boldsymbol{z}$$

$$\begin{array}{rl} \text{minimize}_{\boldsymbol{x},\boldsymbol{z}} & f(\boldsymbol{x}) + h(\boldsymbol{z}) \\ \text{subject to} & \boldsymbol{A}\boldsymbol{x} = \boldsymbol{z} \end{array}$$

dual formulation:

$$\text{maximize}_{\boldsymbol{\lambda}} \quad \min_{\boldsymbol{x},\boldsymbol{z}} \underbrace{f(\boldsymbol{x}) + h(\boldsymbol{z}) + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{z} \rangle}_{=: \mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\lambda}) \text{ (Lagrangian)}} \checkmark$$

A possible route: dual formulation

$$f^*(x) = \sup_z \{ \langle x, z \rangle - f(z) \}$$

$$\min f(x) = -\max \{ -f(x) \}$$

$$\text{maximize}_{\boldsymbol{\lambda}} \quad \min_{\boldsymbol{x},\boldsymbol{z}} \ f(\boldsymbol{x}) + h(\boldsymbol{z}) + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{z} \rangle$$

$$\Updownarrow \quad \text{decouple } \boldsymbol{x} \text{ and } \boldsymbol{z}$$

$$\text{maximize}_{\boldsymbol{\lambda}} \quad \min_{\boldsymbol{x}} \left\{ \underbrace{\langle \boldsymbol{A}^\top \boldsymbol{\lambda}, \boldsymbol{x} \rangle + f(\boldsymbol{x})} \right\} + \min_{\boldsymbol{z}} \left\{ h(\boldsymbol{z}) \underline{- \langle \boldsymbol{\lambda}, \boldsymbol{z} \rangle} \right\}$$

$$\Updownarrow$$

$$\text{maximize}_{\boldsymbol{\lambda}} \quad - f^*(-\boldsymbol{A}^\top \boldsymbol{\lambda}) - h^*(\boldsymbol{\lambda})$$

where f^* (resp. h^*) is the Fenchel conjugate of f (resp. h)

Primal vs. dual problems

$$
\begin{array}{rll}
\text{(primal)} & \text{minimize}_{\boldsymbol{x}} & f(\boldsymbol{x}) + h(\underbrace{\boldsymbol{A}\boldsymbol{x}}) \\
\text{(dual)} & \text{minimize}_{\boldsymbol{\lambda}} & f^*(-\boldsymbol{A}^\top \boldsymbol{\lambda}) + h^*(\boldsymbol{\lambda})
\end{array}
$$

Dual formulation is useful if

- the proximal operator w.r.t. h is cheap (then we can use the Moreau decomposition $\text{prox}_{h^*}(\boldsymbol{x}) = \boldsymbol{x} - \text{prox}_h(\boldsymbol{x})$)
- f^* is smooth (or if f is strongly convex)

Dual proximal gradient methods

Apply proximal gradient methods to the dual problem:

Algorithm 9.1 Dual proximal gradient algorithm

1: **for** $t = 0, 1, \cdots$ **do**

2: $\quad \boldsymbol{\lambda}^{t+1} = \mathsf{prox}_{\underbrace{\eta_t h^*}_{\delta}}\Big(\boldsymbol{\lambda}^t + \eta_t \boldsymbol{A} \nabla f^*(- \boldsymbol{A}^\top \boldsymbol{\lambda}^t)\Big)$

- let $Q(\boldsymbol{\lambda}) := -f^*(-\boldsymbol{A}^\top \boldsymbol{\lambda}) - h^*(\boldsymbol{\lambda})$ and $Q^{\mathsf{opt}} = \max_{\boldsymbol{\lambda}} Q(\boldsymbol{\lambda})$, then

$$Q^{\mathsf{opt}} - Q(\boldsymbol{\lambda}^t) \lesssim \frac{1}{t} \tag{9.1}$$

Primal representation of dual proximal gradient methods

Algorithm 9.1 admits a more explicit primal representation

Algorithm 9.2 Dual proximal gradient algorithm (primal representation)

---

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

$\quad \boldsymbol{x}^t = \arg\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) + \langle \boldsymbol{A}^\top \boldsymbol{\lambda}^t, \boldsymbol{x} \rangle \right\}$

$\quad \boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \eta_t \boldsymbol{A}\boldsymbol{x}^t - \eta_t \mathsf{prox}_{\eta_t^{-1} h}\left(\eta_t^{-1} \boldsymbol{\lambda}^t + \boldsymbol{A}\boldsymbol{x}^t\right)$

$\underbrace{\partial f(x) + A^{T} \lambda^{t} \in 0}$

$\Downarrow$

$-A^{T}\lambda^{t} \in \partial f(x^{t})$

- $\{\boldsymbol{x}^t\}$ is a primal sequence, which is nonetheless *not always feasible*

$$\min_{\lambda} \underbrace{f^*(-A^T\lambda)}_{\Delta} + h^*(\lambda)$$

Proximal Gradient Method

$$\min_{\lambda} f^*(-\cancel{A^T\lambda^t}) + \langle -A\cdot\nabla f^*(-A^T\lambda^t), \lambda-\lambda^t\rangle + \frac{1}{2\eta_t}\|\lambda-\lambda^t\|^2 + h^*(\lambda)$$

$$\Leftrightarrow \min_{\lambda} \frac{1}{2}\|\lambda-(\lambda^t+\eta_t A\cdot\nabla f^*(-A^T\lambda^t))\|^2 + \eta_t h^*(\lambda)$$

$$= \underbrace{\boxed{\operatorname{prox}_{\eta_t h^*}(\lambda^t+\eta_t A\cdot\nabla f^*(-A^T\lambda^t))}}_{\Delta}$$

Extended Mor D

$$\underbrace{x = \operatorname{prox}_{\lambda f}(x) + \lambda \operatorname{prox}_{\lambda^{-1} f^*}(x/\lambda)}$$ ✓

Justification of the primal representation

By definition of $\boldsymbol{x}^t$,

$$-\boldsymbol{A}^\top \boldsymbol{\lambda}^t \in \partial f(\boldsymbol{x}^t)$$

This together with the conjugate subgradient theorem and the smoothness of f^* yields

$$\boldsymbol{x}^t = \nabla f^*(-\boldsymbol{A}^\top \boldsymbol{\lambda}^t)$$

Therefore, the dual proximal gradient update rule can be rewritten as

$$\boldsymbol{\lambda}^{t+1} = \mathsf{prox}_{\eta_t h^*}(\boldsymbol{\lambda}^t + \eta_t \boldsymbol{A}\boldsymbol{x}^t) \tag{9.2}$$

Justification of primal representation (cont.)

Moreover, from the extended Moreau decomposition, we know

$$\mathsf{prox}_{\eta_t h^*}(\boldsymbol{\lambda}^t + \eta_t \boldsymbol{A}\boldsymbol{x}^t) = \boldsymbol{\lambda}^t + \eta_t \boldsymbol{A}\boldsymbol{x}^t - \eta_t \mathsf{prox}_{\eta_t^{-1} h}(\eta_t^{-1}\boldsymbol{\lambda}^t + \boldsymbol{A}\boldsymbol{x}^t)$$

$$\Longrightarrow \qquad \boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \eta_t \boldsymbol{A}\boldsymbol{x}^t - \eta_t \mathsf{prox}_{\eta_t^{-1} h}(\eta_t^{-1}\boldsymbol{\lambda}^t + \boldsymbol{A}\boldsymbol{x}^t)$$

Accuracy of the primal sequence

One can control the primal accuracy via the dual accuracy:

Lemma 9.1

Let $\boldsymbol{x}_{\boldsymbol{\lambda}} := \arg\min_{\boldsymbol{x}} \left\{ f(\boldsymbol{x}) + \langle \boldsymbol{A}^\top \boldsymbol{\lambda}, \boldsymbol{x} \rangle \right\}$. *Suppose* f *is* μ*-strongly convex. Then*

$$\|\boldsymbol{x}^* - \boldsymbol{x}_{\boldsymbol{\lambda}}\|_2^2 \leq \frac{2\big(Q^{\mathsf{opt}} - Q(\boldsymbol{\lambda})\big)}{\mu}$$

- **consequence:** $\|\boldsymbol{x}^* - \boldsymbol{x}^t\|_2^2 \lesssim 1/t$ (using (9.1))

Proof of Lemma 9.1

Recall that Lagrangian is given by

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{z},\boldsymbol{\lambda}) := \underbrace{f(\boldsymbol{x}) + \langle \boldsymbol{A}^\top\boldsymbol{\lambda}, \boldsymbol{x}\rangle}_{=:\tilde{f}(\boldsymbol{x},\boldsymbol{\lambda})} + \underbrace{h(\boldsymbol{z}) - \langle \boldsymbol{\lambda}, \boldsymbol{z}\rangle}_{=:\tilde{h}(\boldsymbol{z},\boldsymbol{\lambda})} \quad \checkmark$$

For any $\boldsymbol{\lambda}$, define $\boldsymbol{x}_{\boldsymbol{\lambda}} := \arg\min_{\boldsymbol{x}} \tilde{f}(\boldsymbol{x},\boldsymbol{\lambda})$ and $\boldsymbol{z}_{\boldsymbol{\lambda}} := \arg\min_{\boldsymbol{z}} \tilde{h}(\boldsymbol{z},\boldsymbol{\lambda})$ (non-rigorous). Then by strong convexity,

$\nabla\tilde{f}(x_\lambda,\lambda)=0$

$$\mathcal{L}(\boldsymbol{x}^*,\boldsymbol{z}^*,\boldsymbol{\lambda}) - \mathcal{L}(\boldsymbol{x}_{\boldsymbol{\lambda}},\boldsymbol{z}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}) \geq \tilde{f}(\boldsymbol{x}^*,\boldsymbol{\lambda}) - \tilde{f}(\boldsymbol{x}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}) \geq \frac{1}{2}\mu\|\boldsymbol{x}^* - \boldsymbol{x}_{\boldsymbol{\lambda}}\|_2^2$$

In addition, since $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{z}^*$, one has

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}^*,\boldsymbol{z}^*,\boldsymbol{\lambda}) &= f(\boldsymbol{x}^*) + h(\boldsymbol{z}^*) + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x}^* - \boldsymbol{z}^*\rangle = f(\boldsymbol{x}^*) + h(\boldsymbol{A}\boldsymbol{x}^*)\\ &= F^{\mathsf{opt}} \overset{\text{duality}}{=} Q^{\mathsf{opt}}\end{aligned}$$

This combined with $\mathcal{L}(\boldsymbol{x}_{\boldsymbol{\lambda}},\boldsymbol{z}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}) = Q(\boldsymbol{\lambda})$ gives

$$Q^{\mathsf{opt}} - Q(\boldsymbol{\lambda}) \geq \frac{1}{2}\mu\|\boldsymbol{x}^* - \boldsymbol{x}_{\boldsymbol{\lambda}}\|_2^2$$

as claimed

Accelerated dual proximal gradient methods

One can apply FISTA to dual problem to improve convergence:

Algorithm 9.3 Accelerated dual proximal gradient algorithm

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

$\quad \boldsymbol{\lambda}^{t+1} = \mathsf{prox}_{\eta_t h^*}\Big(\boldsymbol{w}^t + \eta_t \boldsymbol{A} \nabla f^*(-\boldsymbol{A}^\top \boldsymbol{w}^t)\Big)$

$\quad \theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$

$\quad \boldsymbol{w}^{t+1} = \boldsymbol{\lambda}^{t+1} + \frac{\theta_t - 1}{\theta_{t+1}}(\boldsymbol{\lambda}^{t+1} - \boldsymbol{\lambda}^t)$

- apply FISTA theory and Lemma 9.1 to get

$$Q^{\mathsf{opt}} - Q(\boldsymbol{\lambda}^t) \lesssim \frac{1}{t^2} \quad \text{and} \quad \|\boldsymbol{x}^* - \boldsymbol{x}^t\|_2^2 \lesssim \frac{1}{t^2}$$

Primal representation of accelerated dual proximal gradient methods

Algorithm 9.3 admits more explicit primal representation

Algorithm 9.4 Accelerated dual proximal gradient algorithm (primal representation)

1: **for** $t = 0, 1, \cdots$ **do**

2: $\quad \boldsymbol{x}^t = \arg\min_{\boldsymbol{x}} \; f(\boldsymbol{x}) + \langle \boldsymbol{A}^\top \boldsymbol{w}^t, \boldsymbol{x} \rangle$

3: $\quad \boldsymbol{\lambda}^{t+1} = \boldsymbol{w}^t + \eta_t \boldsymbol{A}\boldsymbol{x}^t - \eta_t \mathsf{prox}_{\eta_t^{-1} h}(\eta_t^{-1} \boldsymbol{w}^t + \boldsymbol{A}\boldsymbol{x}^t)$

4: $\quad \theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$

5: $\quad \boldsymbol{w}^{t+1} = \boldsymbol{\lambda}^{t+1} + \frac{\theta_t - 1}{\theta_{t+1}} (\boldsymbol{\lambda}^{t+1} - \boldsymbol{\lambda}^t)$

$$\sup_z \{ \langle x, z \rangle - f(z) \}$$

$$\langle \lambda, Ax \rangle \overset{?}{=} \langle A^T\lambda, x \rangle$$

$$= \mathrm{Tr}(\lambda^T \cdot Ax) = \mathrm{Tr}(\lambda^T A \cdot x) = \langle A^T\lambda, x$$

Primal-dual proximal gradient method

Nonsmooth optimization

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) + h(\boldsymbol{A}\boldsymbol{x})$$

where f and h are closed and convex

- both f and h might be non-smooth
- both f and h might have inexpensive proximal operators

Primal-dual approaches?

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) + h(\boldsymbol{A}\boldsymbol{x})$$

So far we have discussed proximal methods (resp. dual proximal methods), which essentially updates only primal (resp. dual) variables

Question: can we update both primal and dual variables simultaneously and take advantage of both prox_f and prox_h?

A saddle-point formulation

To this end, we first derive a saddle-point formulation that includes both primal and dual variables

$$\text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) + h(\boldsymbol{A}\boldsymbol{x})$$

$f^*(-A^T\lambda)$

$$\Updownarrow \quad \text{add an auxiliary variable } \boldsymbol{z}$$

$$\text{minimize}_{\boldsymbol{x},\boldsymbol{z}} \quad f(\boldsymbol{x}) + h(\boldsymbol{z}) \quad \text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}$$

$$\Updownarrow$$

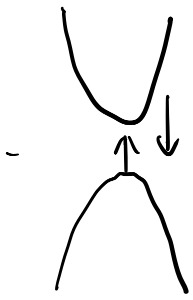
$$\text{maximize}_{\boldsymbol{\lambda}} \ \min_{\boldsymbol{x},\boldsymbol{z}} \ f(\boldsymbol{x}) + h(\boldsymbol{z}) + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} - \boldsymbol{z} \rangle$$

$$\Updownarrow$$

$$\text{maximize}_{\boldsymbol{\lambda}} \ \min_{\boldsymbol{x}} \ f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} \rangle - h^*(\boldsymbol{\lambda})$$

$$\Updownarrow$$

$$\text{minimize}_{\boldsymbol{x}} \ \max_{\boldsymbol{\lambda}} \ f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} \rangle - h^*(\boldsymbol{\lambda}) \quad \text{(saddle-point problem)}$$



A saddle-point formulation

$$\text{minimize}_{\boldsymbol{x}} \ \max_{\boldsymbol{\lambda}} \ \underbrace{f(\boldsymbol{x})} + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} \rangle - h^*(\boldsymbol{\lambda}) \qquad (9.3)$$

- one can then consider updating the primal variable $\boldsymbol{x}$ and the dual variable $\boldsymbol{\lambda}$ simultaneously
- we'll first examine the optimality condition for (9.3), which in turn gives ideas about how to jointly update primal and dual variables

Optimality condition

$$\text{minimize}_{\boldsymbol{x}} \ \max_{\boldsymbol{\lambda}} \ f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} \rangle - h^*(\boldsymbol{\lambda})$$

$$Tr(\lambda^T A x) = Tr(x \lambda^T A)$$
$$= Tr(\lambda^T A x)$$
$$= \langle A^T \lambda, x \rangle$$
$$= \langle x, A^T \lambda \rangle$$

optimality condition:

$$\begin{cases} \mathbf{0} \in & \partial f(\boldsymbol{x}) + \boldsymbol{A}^\top \boldsymbol{\lambda} \\ \mathbf{0} \in & -\boldsymbol{A}\boldsymbol{x} + \partial h^*(\boldsymbol{\lambda}) \end{cases}$$

$$\Longleftrightarrow \quad \mathbf{0} \in \begin{bmatrix} & \boldsymbol{A}^\top \\ -\boldsymbol{A} & \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \end{bmatrix} + \begin{bmatrix} \partial f(\boldsymbol{x}) \\ \partial h^*(\boldsymbol{\lambda}) \end{bmatrix} =: \mathcal{F}(\boldsymbol{x}, \boldsymbol{\lambda}) \quad (9.4)$$

primal + dual

key idea: iteratively update $(\boldsymbol{x}, \boldsymbol{\lambda})$ to reach a point obeying $\mathbf{0} \in \mathcal{F}(\boldsymbol{x}, \boldsymbol{\lambda})$

How to solve $\mathbf{0} \in \mathcal{F}(\boldsymbol{x})$ in general?

In general, finding solution to

$$\underbrace{\mathbf{0} \in \mathcal{F}(\boldsymbol{x})}_{\text{called "monotone inclusion problem" if } \mathcal{F} \text{ is maximal monotone}}$$

$$\Longleftrightarrow \qquad \boldsymbol{x} \in (\mathcal{I} + \mathcal{F})(\boldsymbol{x})$$

$x \in x + \mathcal{F}(x)$

is equivalent to finding fixed points of $\underbrace{(\mathcal{I} + \eta\mathcal{F})^{-1}}_{\text{resolvent of } \mathcal{F}}$, i.e. solutions to

$$\boldsymbol{x} = (\mathcal{I} + \eta\mathcal{F})^{-1}(\boldsymbol{x})$$

$x_{t+1} = x_t + \mathcal{F}(x_t)$

This suggests a natural fixed-point iteration / resolvent iteration:

$$\boldsymbol{x}^{t+1} = (\mathcal{I} + \eta\mathcal{F})^{-1}(\boldsymbol{x}^t), \qquad t = 0, 1, \cdots$$

$$x^{t+1} = x^t - \eta_t \nabla f(x^t) \qquad \text{GD}.$$

$$\vdots$$

$$\boxed{x^* = x^* - \eta_t \nabla f(x^*)} \qquad \text{fix-point}$$

$$x = T(x) = \underbrace{(I - \eta_t \nabla f)}_{\text{Operator}}(x)$$

$$\operatorname{prox}_f(x) = \operatorname*{argmin}_z \frac{1}{2}\|z-x\|^2 + f(z)$$

$$z - x + \partial f(z) \ni 0$$

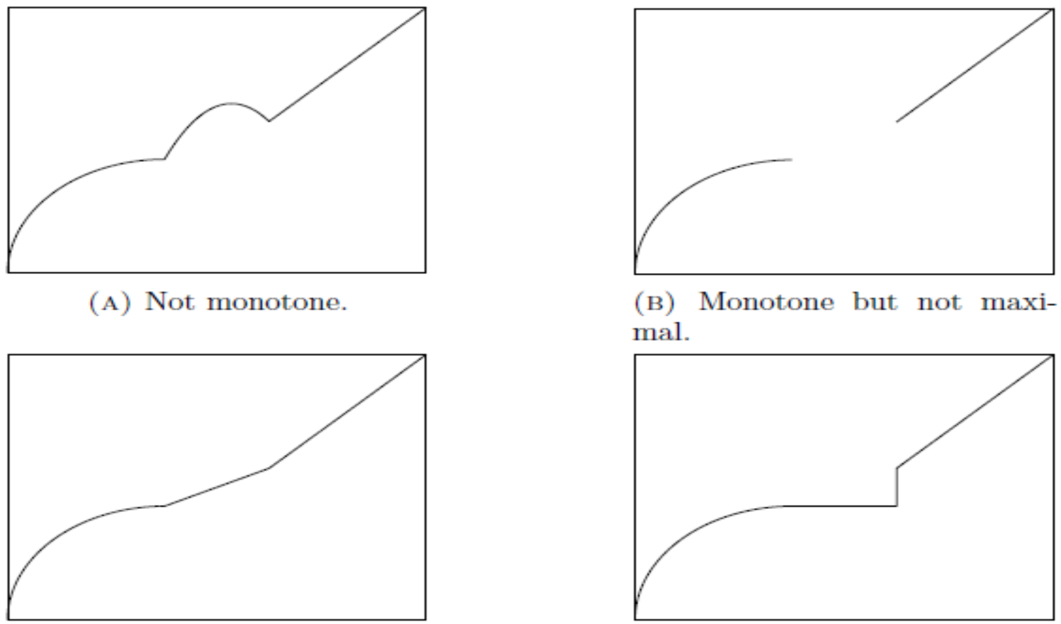
$$\Rightarrow \quad x \in z + \partial f(z)$$

$$:= (I + \partial f)(z)$$

$$z = (I + \partial f)^{-1} x$$

Aside: monotone operators

— *Ryu, Boyd '16*



(A) Not monotone.

(B) Monotone but not maximal.

(C) Maximal monotone function.

(D) Maximal monotone but not a function.

- a relation $\mathcal{F}$ is called *monotone* if

$$\langle \boldsymbol{u} - \boldsymbol{v}, \boldsymbol{x} - \boldsymbol{y} \rangle \geq 0, \quad \forall (\boldsymbol{x}, \underbrace{\boldsymbol{u}}_{T(x)}), (\boldsymbol{y}, \underbrace{\boldsymbol{v}}_{T(y)}) \in \mathcal{F}$$

- relation $\mathcal{F}$ is called *maximal monotone* if there is no monotone operator that contains it

$$\langle T(x) - T(y), x - y \rangle \geq 0, \quad \checkmark$$

$$\underset{\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim\sim}{\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0} \quad \Delta$$

Proximal point method

$$\boldsymbol{x}^{t+1} = (\mathcal{I} + \eta_t \mathcal{F})^{-1}(\boldsymbol{x}^t), \qquad t = 0, 1, \cdots$$

If $\mathcal{F} = \partial f$ for some convex function f, then this proximal point method becomes

$$\boldsymbol{x}^{t+1} = \mathsf{prox}_{\eta_t f}(\boldsymbol{x}^t), \qquad t = 0, 1, \cdots$$

- useful when $\mathsf{prox}_{\eta_t f}$ is cheap

Back to primal-dual approaches

Recall that we want to solve

$$\mathbf{0} \in \underbrace{\begin{bmatrix} & \boldsymbol{A}^\top \\ -\boldsymbol{A} & \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \end{bmatrix}}_{A(x,\lambda)} + \underbrace{\begin{bmatrix} \partial f(\boldsymbol{x}) \\ \partial h^*(\boldsymbol{\lambda}) \end{bmatrix}}_{B(x,\lambda)} =: \mathcal{F}(\boldsymbol{x}, \boldsymbol{\lambda}) \quad \checkmark$$

the issue of proximal point methods: computing $(\mathcal{I} + \eta\mathcal{F})^{-1}$ is in general difficult

$$x = (I + \eta A)^{-1}(z)$$

$$\begin{bmatrix} x \\ \lambda \end{bmatrix} + \eta A(x,\lambda) = z$$

$$Ax = b$$

$$\downarrow$$

Back to primal-dual approaches

observation: practically we may often consider splitting $\mathcal{F}$ into two operators

$$\mathbf{0} \in \mathcal{A}(\boldsymbol{x}, \boldsymbol{\lambda}) + \mathcal{B}(\boldsymbol{x}, \boldsymbol{\lambda})$$

$$\text{with} \quad \mathcal{A}(\boldsymbol{x}, \boldsymbol{\lambda}) = \begin{bmatrix} & \boldsymbol{A} \\ -\boldsymbol{A}^\top & \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \end{bmatrix}, \quad \mathcal{B}(\boldsymbol{x}, \boldsymbol{\lambda}) = \begin{bmatrix} \partial f(\boldsymbol{x}) \\ \partial h^*(\boldsymbol{\lambda}) \end{bmatrix} \tag{9.5}$$

- $(\mathcal{I} + \eta\mathcal{A})^{-1}$ can be computed by solving linear systems
- $(\mathcal{I} + \eta\mathcal{B})^{-1}$ is easy if prox_f and prox_{h^*} are both inexpensive

solution: design update rules based on $(\mathcal{I} + \eta\mathcal{A})^{-1}$ and $(\mathcal{I} + \eta\mathcal{B})^{-1}$ instead of $(\mathcal{I} + \eta\mathcal{F})^{-1}$

Operator splitting via Cayley operators

We now introduce a principled approach based on operator splitting

$$\text{find } \boldsymbol{x} \quad \text{s.t. } \mathbf{0} \in \mathcal{F}(\boldsymbol{x}) = \underbrace{\mathcal{A}(\boldsymbol{x}) + \mathcal{B}(\boldsymbol{x})}_{\text{operator splitting}}$$

R_{A+B}

let $\mathcal{R}_{\mathcal{A}} := (\mathcal{I} + \eta\mathcal{A})^{-1}$ and $\mathcal{R}_{\mathcal{B}} := (\mathcal{I} + \eta\mathcal{B})^{-1}$ be the resolvents, and $\mathcal{C}_{\mathcal{A}} := 2\mathcal{R}_{\mathcal{A}} - \mathcal{I}$ and $\mathcal{C}_{\mathcal{B}} := 2\mathcal{R}_{\mathcal{B}} - \mathcal{I}$ be the Cayley operators

Lemma 9.2

$C_A(C_B(z)) = C_A((2R_B - I)(z))$

$= C_A(2R_B(z) - z)$

$F(x)$

$$\underbrace{\mathbf{0} \in \mathcal{A}(\boldsymbol{x}) + \mathcal{B}(\boldsymbol{x})}_{\boldsymbol{x} \in \mathcal{R}_{\mathcal{A}+\mathcal{B}}(\boldsymbol{x})} \iff \underbrace{\mathcal{C}_{\mathcal{A}}\mathcal{C}_{\mathcal{B}}(\boldsymbol{z}) = \boldsymbol{z} \text{ with } \boldsymbol{x} = \mathcal{R}_{\mathcal{B}}(\boldsymbol{z})}_{\textit{it comes down to finding fixed points of } \mathcal{C}_{\mathcal{A}}\mathcal{C}_{\mathcal{B}}} \tag{9.6}$$

Operator splitting via Cayley operators

$$\boldsymbol{x} \in \mathcal{R}_{\mathcal{A}+\mathcal{B}}(\boldsymbol{x}) \quad \Longleftrightarrow \quad \mathcal{C}_{\mathcal{A}}\mathcal{C}_{\mathcal{B}}(\boldsymbol{z}) = \boldsymbol{z}$$

- **advantage:** allows us to apply $\mathcal{C}_{\mathcal{A}}$ (resp. $\mathcal{R}_{\mathcal{A}}$) and $\mathcal{C}_{\mathcal{B}}$ (resp. $\mathcal{R}_{\mathcal{B}}$) sequentially (instead of computing $\mathcal{R}_{\mathcal{A}+\mathcal{B}}$ directly)

Proof of Lemma 9.2

$C_{\mathcal{B}}(z) = (2R_{\mathcal{B}} - I)(z)$
$= 2R_{\mathcal{B}}(z) - z$
$= \boxed{2x - z}$
$:= \tilde{z}$

$2x = z + \tilde{z}$

$C_{\mathcal{A}}C_{\mathcal{B}}(z) = C_{\mathcal{A}}(\tilde{z})$
$= 2R_{\mathcal{A}}(\tilde{z}) - \tilde{z}$
$:= \boxed{2\tilde{x} - \tilde{z}}$
$= z$

$2\tilde{x} = z + \tilde{z}$

$$\mathcal{C}_{\mathcal{A}}\mathcal{C}_{\mathcal{B}}(z) = z$$

$$\Longleftrightarrow$$

$$x = \boxed{\mathcal{R}_{\mathcal{B}}(z)} \tag{9.7a}$$
$$\tilde{z} = 2x - z \tag{9.7b}$$
$$\tilde{x} = \boxed{\mathcal{R}_{\mathcal{A}}(\tilde{z})} \;\checkmark \tag{9.7c}$$
$$z = 2\tilde{x} - \tilde{z} \tag{9.7d}$$

From (9.7b) and (9.7d), we see that $R_{\mathcal{B}}(z) = R_{\mathcal{A}}(2R_{\mathcal{B}}(z) - z)$

$$\tilde{x} = x$$

which together with (9.7d) gives $z = 2R_{\mathcal{A}}(2R_{\mathcal{B}}(z) - z)) - (2R_{\mathcal{B}}(z) - z)$

$$2x = z + \tilde{z} \tag{9.8}$$

Proof of Lemma 9.2 (cont.)

$x = R_B(z) = (I+\eta B)^{-1}(z) \Rightarrow z \in x + \eta B(x)$ ✓

$x = R_A(\tilde{z}) \Rightarrow \tilde{z} \in x + \eta A(x)$ ✓

Recall that

$$\boldsymbol{z} \in \boldsymbol{x} + \eta\mathcal{B}(\boldsymbol{x}) \qquad \text{and} \qquad \tilde{\boldsymbol{z}} \in \boldsymbol{x} + \eta\mathcal{A}(\boldsymbol{x})$$

Adding these two facts and using (9.8), we get

$$2\boldsymbol{x} = \boldsymbol{z} + \tilde{\boldsymbol{z}} \in 2\boldsymbol{x} + \eta\mathcal{B}(\boldsymbol{x}) + \eta\mathcal{A}(\boldsymbol{x})$$

$$\Longleftrightarrow \qquad \boldsymbol{0} \in \mathcal{A}(\boldsymbol{x}) + \mathcal{B}(\boldsymbol{x})$$

Douglas-Rachford splitting

How to find points obeying $\boldsymbol{x} = \mathcal{C}_{\mathcal{A}}\mathcal{C}_{\mathcal{B}}(\boldsymbol{x})$?

- First attempt: fixed-point iteration

$$\boldsymbol{z}^{t+1} = \mathcal{C}_{\mathcal{A}}\mathcal{C}_{\mathcal{B}}(\boldsymbol{z}^{t})$$

  unfortunately, it may not converge in general

- **Douglas-Rachford splitting**: damped fixed-point iteration

$$\boldsymbol{z}^{t+1} = \frac{1}{2}(\mathcal{I} + \mathcal{C}_{\mathcal{A}}\mathcal{C}_{\mathcal{B}})(\boldsymbol{z}^{t})$$

  converges when a solution to $\mathbf{0} \in \mathcal{A}(\boldsymbol{x}) + \mathcal{B}(\boldsymbol{x})$ exists!

More explicit expression for D-R splitting

$z^t \to z^{t+1}$

Douglas-Rachford splitting update rule $z^{t+1} = \frac{1}{2}(\mathcal{I} + \mathcal{C}_{\mathcal{A}}\mathcal{C}_{\mathcal{B}})(z^t)$ is essentially:

$$\begin{aligned}
x^{t+\frac{1}{2}} &= \mathcal{R}_{\mathcal{B}}(z^t) \\
z^{t+\frac{1}{2}} &= 2x^{t+\frac{1}{2}} - z^t \\
x^{t+1} &= \mathcal{R}_{\mathcal{A}}(z^{t+\frac{1}{2}}) \\
z^{t+1} &= \frac{1}{2}(z^t + 2x^{t+1} - z^{t+\frac{1}{2}}) \\
&= z^t + x^{t+1} - x^{t+\frac{1}{2}}
\end{aligned}$$

$\underbrace{z^t - (2x^{t+\frac{1}{2}} - z^t)}_{= 2z^t - 2x^{t+\frac{1}{2}}}$

$C_A C_B (z^t)$

where $x^{t+\frac{1}{2}}$ and $z^{t+\frac{1}{2}}$ are auxiliary variables

More explicit expression for D-R splitting

or equivalently,

$$
\begin{aligned}
\boldsymbol{x}^{t+\frac{1}{2}} &= \mathcal{R}_{\mathcal{B}}(\boldsymbol{z}^t) \checkmark \\
\boldsymbol{x}^{t+1} &= \mathcal{R}_{\mathcal{A}}\big(\underbrace{2\boldsymbol{x}^{t+\frac{1}{2}} - \boldsymbol{z}^t}\big) \checkmark \\
\boldsymbol{z}^{t+1} &= \boldsymbol{z}^t + \underbrace{\boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+\frac{1}{2}}}_{\Delta}
\end{aligned}
$$

$$z^{t+1} = z^t + \underbrace{R_A(2x^{t+\frac{1}{2}} - z^t) - R_B(z^t)}_{\Delta} \checkmark$$

$$\min \quad f(x) + g(x)$$

f, g closed and convex

D-R Iteration:

Start at any y_0 and repeat for $t = 0, 1, \dots$

$$x^{t+1} = \text{prox}_f (y^t) \quad \checkmark$$

$$y^{t+1} = y^t + \text{prox}_g (2x^{t+1} - y^t) - x^{t+1} \quad \checkmark$$

x^t converges to a solution of

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

Douglas-Rachford primal-dual splitting

$$\text{minimize}_{\boldsymbol{x}} \ \max_{\boldsymbol{\lambda}} \ f(\boldsymbol{x}) + \langle \boldsymbol{\lambda}, \boldsymbol{A}\boldsymbol{x} \rangle - h^*(\boldsymbol{\lambda})$$

Applying Douglas-Rachford splitting to (9.5) yields

$$\boldsymbol{x}^{t+\frac{1}{2}} = \text{prox}_{\eta f}(\boldsymbol{p}^t) \quad \checkmark$$

$$\boldsymbol{\lambda}^{t+\frac{1}{2}} = \text{prox}_{\eta h^*}(\boldsymbol{q}^t) \quad \checkmark$$

$$\begin{bmatrix} \boldsymbol{x}^{t+1} \\ \boldsymbol{\lambda}^{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \eta \boldsymbol{A}^\top \\ -\eta \boldsymbol{A} & \boldsymbol{I} \end{bmatrix}^{-1} \begin{bmatrix} 2\boldsymbol{x}^{t+\frac{1}{2}} - \boldsymbol{p}^t \\ 2\boldsymbol{\lambda}^{t+\frac{1}{2}} - \boldsymbol{q}^t \end{bmatrix} \quad \checkmark$$

$$\boldsymbol{p}^{t+1} = \boldsymbol{p}^t + \boldsymbol{x}^{t+1} - \boldsymbol{x}^{t+\frac{1}{2}}$$

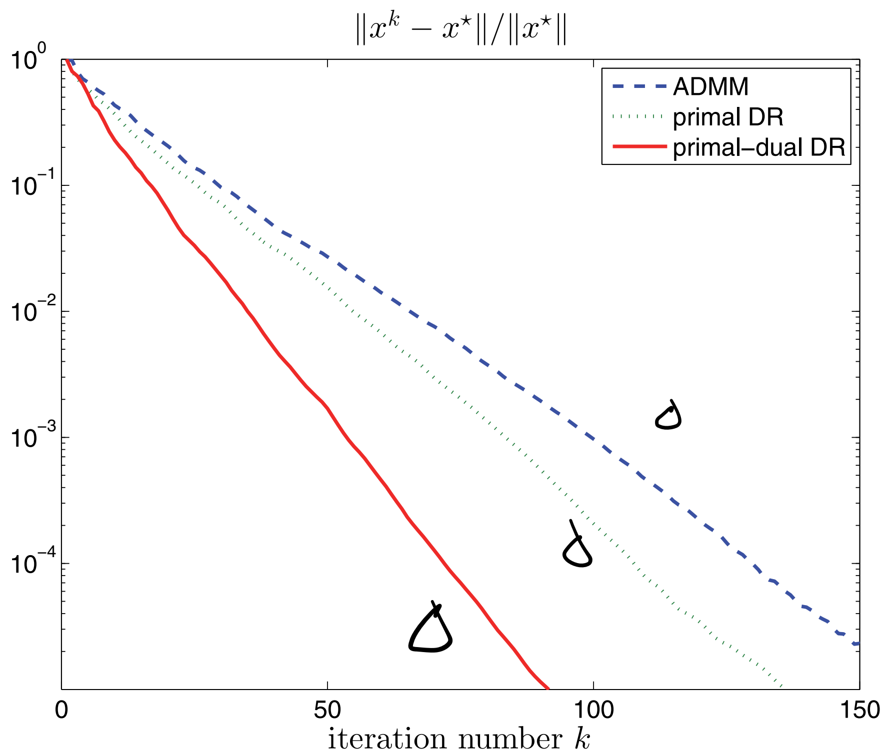
$$\boldsymbol{q}^{t+1} = \boldsymbol{q}^t + \boldsymbol{\lambda}^{t+1} - \boldsymbol{\lambda}^{t+\frac{1}{2}} \quad \checkmark$$

Example

$$\text{minimize}_{\boldsymbol{x}} \quad \|\boldsymbol{x}\|_2 + \gamma\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_1$$

$$\Longleftrightarrow \qquad \text{minimize}_{\boldsymbol{x}} \quad f(\boldsymbol{x}) + g(\boldsymbol{A}\boldsymbol{x})$$

with $f(\boldsymbol{x}) := \|\boldsymbol{x}\|_2$ and $g(\boldsymbol{y}) := \gamma\|\boldsymbol{y} - \boldsymbol{b}\|_1$



— *Connor, Vandenberghe '14*

Reference

[1] "*A First Course in Convex Optimization Theory*," E. Ryu, W. Yin.

[2] "*Optimization methods for large-scale systems, EE236C lecture notes*," L. Vandenberghe, UCLA.

[3] "*Convex optimization, EE364B lecture notes*," S. Boyd, Stanford.

[4] "*Mathematical optimization, MATH301 lecture notes*," E. Candes, Stanford.

[5] "*First-order methods in optimization*," A. Beck, Vol. 25, SIAM, 2017.

[6] "*Primal-dual decomposition by operator splitting and applications to image deblurring*," D. O' Connor, L. Vandenberghe, *SIAM Journal on Imaging Sciences*, 2014.

[7] "*On the numerical solution of heat conduction problems in two and three space variables*," J. Douglas, H. Rachford, *Transactions of the American mathematical Society*, 1956.

Reference

[8] "*A primer on monotone operator methods*," E. Ryu, S. Boyd, *Appl. Comput. Math.*, 2016.