SPLITTING ALGORITHMS FOR NONCONVEX OPTIMIZATION: UNIFIED ANALYSIS AND NEWTON-TYPE ACCELERATION

Themelis, Andreas
Faculty of Information Science and Electrical Engineering (ISEE), Kyushu University

Splitting algorithms for nonconvex optimization
unified analysis and Newton-type acceleration

Andreas Themelis

andreas.themelis@ees.kyushu-u.ac.jp

Kyushu University

Faculty of Information Science and Electrical Engineering (ISEE)
744 Motooka, Nishi-ku, 819-0395 Fukuoka, Japan

NPU Optimization Seminar

Northwestern Polytechnical University

October 21, 2022

Outline

Outline

Introduction

Convex splitting algorithms

- Optimality conditions as a **monotone** inclusion problem

$$\text{find } x \text{ such that } 0 \in M(x) \qquad \text{(P)}$$

  e.g. $\quad \text{minimize}\, \varphi \quad \leftrightarrow \quad \text{find } x \text{ such that } 0 \in \partial\varphi(x)$

- M is **split** into "simpler" **monotone** operators *e.g.* $M = \sum_i A_i$
- (P) solved through **fixed-point iterations**

$$x \mapsto x^+ = (1 - \lambda)x + \lambda \mathcal{T} x$$

  where $\mathcal{T}$ operates on each A_i **separately**
  - direct evaluations $x \mapsto A_i x$,
  - resolvent steps $x \mapsto \mathcal{J}_{\gamma A_i} x := (\mathrm{id} + \gamma A_i)^{-1} x$,
  - forward steps $x \mapsto x - \gamma A_i x$, etc...

- **fixed points** $x^+ = x$ are (related to) **solutions** of (P)

Some examples

Forward-backward splitting

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

$$x \mapsto \begin{cases} \bar{x} &= \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)) \\ x^+ &= (1-\lambda)x + \lambda \bar{x} \end{cases}$$

Douglas-Rachford splitting

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

$$s \mapsto \begin{cases} x &= \operatorname{prox}_{\gamma f}(s) \\ \bar{x} &= \operatorname{prox}_{\gamma g}(2x - s) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

ADMM

$$\text{minimize } f(x) + g(z) \text{ s.t. } Ax + Bz = b$$

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto \begin{cases} x^+ &= \arg\min \mathcal{L}_\beta(\,\cdot\,, z, y) \\ y^+ &= y + \beta(Ax^+ + Bz - b) \\ z^+ &= \arg\min \mathcal{L}_\beta(x^+, \cdot\,, y^+) \end{cases}$$

Davis-Yin splitting

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

$$s \mapsto \begin{cases} x &= \operatorname{prox}_{\gamma f}(s) \\ \bar{x} &= \operatorname{prox}_{\gamma g}(2x - s - \gamma \nabla h(x)) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

($\operatorname{prox}_{\gamma\varphi}$ is the resolvent of $\gamma\partial\varphi$: $\operatorname{prox}_{\gamma\varphi} = \arg\min\left\{\varphi + \frac{1}{2\gamma}\|\cdot - x\|^2\right\}$)

Introduction

Convex splitting algorithms

Simple, elegant, **unified** convergence analysis

- $\mathcal{T}$ is α-averaged $\exists \alpha \in (0, 1]$ $(\mathcal{T} = (1-\alpha)\mathrm{id} + \alpha\mathcal{S},\ \exists \mathcal{S}$ 1-Lipschitz)
- thus, whenever $0 < \lambda < 1/\alpha$

$$\operatorname{dist}(x^{k+1}, \operatorname{fix}\mathcal{T})^2 \leq \operatorname{dist}(x^k, \operatorname{fix}\mathcal{T})^2 - \lambda(1/\alpha - \lambda)\|x^{k+1} - x^k\|^2$$

- by *"telescoping"*, $\|x^{k+1} - x^k\| \to 0$

Proof.

$$\begin{aligned}
\lambda(1/\alpha - \lambda)\sum_{k=0}^{K}\|x^{k+1} - x^k\|^2 \leq\ & \operatorname{dist}(x^0, \operatorname{fix}\mathcal{T})^2 - \operatorname{dist}(x^1, \operatorname{fix}\mathcal{T})^2 \\
&+ \operatorname{dist}(x^1, \operatorname{fix}\mathcal{T})^2 - \operatorname{dist}(x^2, \operatorname{fix}\mathcal{T})^2 \\
&\cdots \\
&+ \operatorname{dist}(x^K, \operatorname{fix}\mathcal{T})^2 - \operatorname{dist}(x^{K+1}, \operatorname{fix}\mathcal{T})^2 \\
=\ & \operatorname{dist}(x^0, \operatorname{fix}\mathcal{T})^2 - \operatorname{dist}(x^{K+1}, \operatorname{fix}\mathcal{T})^2 \\
\leq\ & \operatorname{dist}(x^0, \operatorname{fix}\mathcal{T})^2 \underbrace{- 0}_{\text{since } \operatorname{dist}(\cdot, \operatorname{fix}\mathcal{T})^2 \geq 0}
\end{aligned}$$

- *"Fejér monotonicity"* ensures $x^k \to x_\star \in \operatorname{fix}\mathcal{T}$ (if $\operatorname{fix}\mathcal{T} \neq \emptyset$)

Outline

What if problem is nonconvex?

- no averagedness, no Fejér-monotonicity
- operators are **set-valued** $\mathcal{T} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$

$$\textit{e.g.} \quad \mathcal{J}_{\gamma\varphi}(x) \supseteq \operatorname{prox}_{\gamma\varphi}(x) := \operatorname{arg\,min}\left\{\varphi + \tfrac{1}{2\gamma}\|\cdot - x\|^2\right\}$$

- local **VS** global solutions: $x \in \mathcal{T}(x)$ is only **necessary** for optimality

---

Can we still have a **unified** convergence analysis?

Introduction

Nonconvexity?

Splitting algorithm

$$x^+ \in \mathcal{T}(x)$$

Remarks.

- **Lyapunov potential.** Any **lower bounded** $\mathcal{V} : \mathbb{R}^n \to \mathbb{R}$ such that

$$\mathcal{V}(x^+) \leq \mathcal{V}(x) - c\|x^+ - x\|^2$$

  can replace $\operatorname{dist}(\,\cdot\,, \operatorname{fix}\mathcal{T})^2$

- **Subsequential convergence.** Having $\|x^{k+1} - x^k\| \to 0$, then

$$x^{k_j} \to x_\star \quad \Rightarrow \quad x_\star \in \mathcal{T}(x_\star) \qquad (\mathcal{T} \text{ is } \textit{outer semicontinuous})$$

Splitting algorithm

$$x^+ \in \mathcal{T}(x)$$

Remarks.

- **Lyapunov potential.** Any **lower bounded** $\mathcal{V} : \mathbb{R}^n \to \mathbb{R}$ such that

$$\mathcal{V}(x^+) \leq \mathcal{V}(x) - c\|x^+ - x\|^2$$

  can replace $\operatorname{dist}(\,\cdot\,, \operatorname{fix}\mathcal{T})^2$

- **Subsequential convergence.** Having $\|x^{k+1} - x^k\| \to 0$, then

$$x^{k_j} \to x_\star \quad \Rightarrow \quad x_\star \in \mathcal{T}(x_\star) \qquad (\mathcal{T} \text{ is outer semicontinuous})$$

Theorem

Suppose $\mathcal{T}$ is osc and admits a Lyapunov potential $\mathcal{V}$. Then its fixed-point iterations are ***subsequentially convergent****.*

Outline

In this talk:

- comprehensive class of osc algorithms $\mathcal{T}$ that admit potential $\mathcal{V}$
  (thus subsequentially convergent)
- **Newton-type** variants inheriting
  - (subsequential) convergence
  - oracle complexity

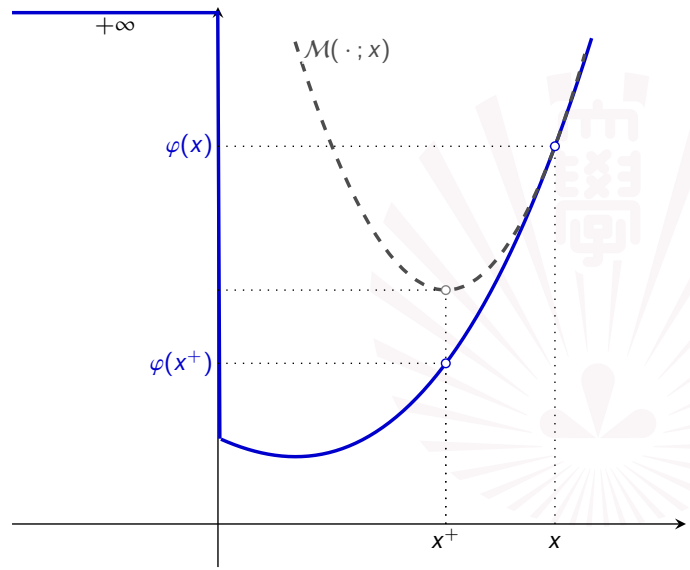
Outline

Outline

Problem. $\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \varphi(x)$ (P)

MM paradigm

1. Pick *majorizing model* $\mathcal{M}(w;x)$ i.e.
   - $\mathcal{M}(x;x) = \varphi(x)$
   - $\mathcal{M}(w;x) \geq \varphi(w)$
2. Let $x^+ \in \arg\min_w \mathcal{M}(w;x)$
3. Then, $\varphi(x^+) \leq \varphi(x)$



Example: proximal point
$\mathcal{M}^{\text{PP}}(w,x) = \varphi(w) + \frac{1}{2\gamma}\|w-x\|^2$

Why

- cope with nonconvex nonsmooth problems
- monotone algorithm

When

- $\mathcal{M}$ easy to minimize

Problem. $\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \varphi(x)$ (P)

Typical requirements

- φ and $\mathcal{M}(\,\cdot\,, x)$ directionally differentiable[1,3,4]
- φ continuous[2,3,4]
- $\operatorname{dom}\varphi$ convex[1,2,4]
- $\mathcal{M}(\,\cdot\,, x) - \varphi$ smooth[2]

[1] A. Beck and D. Pan. *Convergence of an inexact MM method for solving a class of composite optimization problems*, Springer, 2018

[2] J. Mairal. *Incremental MM optimization with application to large-scale machine learning*, SIOPT 25(2), 2015

[3] Y. Sun, P. Babu, and D. Palomar. *MM algorithms in signal processing, communications, and machine learning*, IEEE TSP 65(3), 2017

[4] G. Scutari and Y. Sun. *Parallel and distributed successive convex approximation methods for big-data optimization*, Springer, 2018

Algorithmic design

The majorization-minimization principle

Problem. $\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \varphi(x)$ (P)

Our working assumptions: $\mathcal{M}$ is a **proximal** majorizing model

- $\mathcal{M}$ is lsc
- $\varphi(w) + \frac{m_1}{2}\|w-x\|^2 \le \mathcal{M}(w,x) \le \varphi(w) + \frac{m_2}{2}\|w-x\|^2$

Example.
FBS is a proximal MM algorithm
(∇f L_f-Lipschitz, $\gamma < 1/L_f$)

$$\begin{aligned}\mathcal{T}(x) &= \operatorname{prox}_{\gamma g}(x-\gamma\nabla f(x))\\ &= \arg\min_w \mathcal{M}(w;x)\end{aligned}$$

Forward-backward splitting

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

$$x \mapsto \begin{cases}\bar{x} &= \operatorname{prox}_{\gamma g}(x-\gamma\nabla f(x))\\ x^+ &= (1-\lambda)x+\lambda\bar{x}\end{cases}$$

$$\mathcal{M}(w;x) := f(x)+\langle\nabla f(x), w-x\rangle + g(w) + \tfrac{1}{2\gamma}\|w-x\|^2 \qquad \begin{pmatrix} m_1 = 1-\gamma L_f \\ m_2 = 1+\gamma L_f\end{pmatrix}$$

Algorithmic design

The majorization-minimization principle

Given a proximal MM model $\mathcal{M}$, define $\mathcal{T}^{\mathcal{M}} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as

$$\mathcal{T}^{\mathcal{M}}(x) := \arg\min_{w} \mathcal{M}(w, x)$$

Theorem.

(i). **Regularity.** $\mathcal{T}^{\mathcal{M}}$ is osc & nonempty valued

(ii). **Cost decrease.** $\varphi(x^+) \leq \varphi(x) - \frac{m_1}{2}\|x - x^+\|^2$ for all $x^+ \in \mathcal{T}^{\mathcal{M}}(x)$

(iii). **Stationarity.** $x \in \mathcal{T}^{\mathcal{M}}(x) \Rightarrow 0 \in \hat{\partial}\varphi(x)$

Corollary. Consider $x^{k+1} \in \mathcal{T}^{\mathcal{M}}(x^k)$, $k \in \mathbb{N}$. Then,

(i). $x^{k+1} - x^k \to 0$

(ii). accumulation points are stationary for φ

Algorithmic design

The majorization-minimization principle

Given a proximal MM model $\mathcal{M}$, define $\mathcal{T}^{\mathcal{M}} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ as

$$\mathcal{T}^{\mathcal{M}}(x) := \arg\min_w \mathcal{M}(w, x)$$

Theorem.

(i). **Regularity.** $\mathcal{T}^{\mathcal{M}}$ is osc & nonempty valued

(ii). **Cost decrease.** $\varphi(x^+) \leq \varphi(x) - \frac{m_1}{2}\|x - x^+\|^2$ for all $x^+ \in \mathcal{T}^{\mathcal{M}}(x)$

(iii). **Stationarity.** $x \in \mathcal{T}^{\mathcal{M}}(x) \Rightarrow 0 \in \hat{\partial}\varphi(x)$

Corollary. Consider $x^{k+1} \in \mathcal{T}^{\mathcal{M}}(x^k)$, $k \in \mathbb{N}$. Then,

(i). $x^{k+1} - x^k \to 0$

(ii). accumulation points are stationary for φ

Issues:

- Can't cope with relaxation $(1 - \lambda)x + \lambda x^+$
- Limited to *"pure"* (proximal) MM algorithms (e.g. only monotone algorithms)

Outline

A glance to known splitting schemes

Douglas-Rachford splitting

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

$$s \mapsto \begin{cases} x &= \operatorname{prox}_{\gamma f}(s) \\ \bar{x} &= \operatorname{prox}_{\gamma g}(2x - s) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

Davis-Yin splitting

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

$$s \mapsto \begin{cases} x &= \operatorname{prox}_{\gamma f}(s) \\ \bar{x} &= \operatorname{prox}_{\gamma g}(2x - s - \gamma \nabla h(x)) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

A glance to known splitting schemes

Suppose ∇f Lipschitz, everything nonconvex[1,2,3]

- $x = \mathrm{prox}_{\gamma f}(s) \Leftrightarrow s = x + \gamma \nabla f(x)$
- $\mathrm{prox}_{\gamma f}$ Lipschitz and strongly montone

Douglas-Rachford splitting

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

$$s \mapsto \begin{cases} x &= \mathrm{prox}_{\gamma f}(s) \\ \bar{x} &= \mathrm{prox}_{\gamma g}(2x - s) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

Davis-Yin splitting

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

$$s \mapsto \begin{cases} x &= \mathrm{prox}_{\gamma f}(s) \\ \bar{x} &= \mathrm{prox}_{\gamma g}(2x - s - \gamma \nabla h(x)) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

[1] Li G. and Pong TK. *Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems*, MathProg 159(2), 2015

[2] Li G., Liu T. and Pong TK. *Peaceman-Rachford splitting for a class of nonconvex optimization problems*, COAP 68(2), 2017

[3] AT and P. Patrinos. *Douglas-Rachford splitting and ADMM for nonconvex optimization: tight convergence results*, SIOPT 30(1), 2020

A glance to known splitting schemes

Suppose ∇f Lipschitz, everything nonconvex[1,2,3]

- $x = \operatorname{prox}_{\gamma f}(s) \Leftrightarrow s = x + \gamma \nabla f(x)$
- $\operatorname{prox}_{\gamma f}$ Lipschitz and strongly montone

Douglas-Rachford splitting

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

$$s \mapsto \begin{cases} x &= \operatorname{prox}_{\gamma f}(s) \\ \bar{x} &\in \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

Davis-Yin splitting

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

$$s \mapsto \begin{cases} x &= \operatorname{prox}_{\gamma f}(s) \\ \bar{x} &\in \operatorname{prox}_{\gamma g}(x - \gamma \nabla (f+h)(x)) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

[1] Li G. and Pong TK. *Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems*, MathProg 159(2), 2015

[2] Li G., Liu T. and Pong TK. *Peaceman-Rachford splitting for a class of nonconvex optimization problems*, COAP 68(2), 2017

[3] AT and P. Patrinos. *Douglas-Rachford splitting and ADMM for nonconvex optimization: tight convergence results*, SIOPT 30(1), 2020

Algorithmic design

Generalized proximal MM algorithms

A glance to known splitting schemes...

Suppose ∇f Lipschitz, everything nonconvex

- $x = \operatorname{prox}_{\gamma f}(s) \Leftrightarrow s = x + \gamma \nabla f(x)$
- $\operatorname{prox}_{\gamma f}$ Lipschitz and strongly montone

Douglas-Rachford splitting

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

$$s \mapsto \begin{cases} x &= \operatorname{prox}_{\gamma f}(s) \\ \bar{x} &\in \operatorname{prox}_{\gamma g}(x - \gamma \nabla f(x)) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

Davis-Yin splitting

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

$$s \mapsto \begin{cases} x &= \operatorname{prox}_{\gamma f}(s) \\ \bar{x} &\in \operatorname{prox}_{\gamma g}(x - \gamma \nabla (f+h)(x)) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases}$$

...leads to Generalized Proximal MM schemes

$$s \mapsto \begin{cases} x &= G(s) \\ \bar{x} &\in \mathcal{T}(x) \\ s^+ &= s + \lambda(\bar{x} - x) \end{cases} \quad \text{with}$$

- G L_G-Lipschitz and μ_G-str. monotone
- $\mathcal{T} = \mathcal{T}^{\mathcal{M}} := \arg\min_w \mathcal{M}(w; \cdot)$

$\mathcal{A} \sim (\mathcal{T}, G)$ **is a GPMM algorithm**

Outline

Outline

Convergence analysis

Proximal envelopes

To a **GPMM scheme** $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

we associate a **proximal envelope function**

$$\varphi^{\mathcal{A}}(s) := \min_w \mathcal{M}(w; x) \quad \text{where } x = G(s)$$

Convergence analysis

Proximal envelopes

To a **GPMM scheme** $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

we associate a **proximal envelope function**

$$\varphi^{\mathcal{A}}(s) := \min_w \mathcal{M}(w; x) \quad \text{where } x = G(s)$$

Reduces to

- Moreau envelope for proximal point algorithm
- **F**orward-**B**ackward **E**nvelope (**FBE**) for FBS[1,2]
- **D**ouglas-**R**achford **E**nvelope (**DRE**) for DRS[3,4]
- **D**avis-**Y**in **E**nvelope (**DYE**) for DYS[5]

Basic properties

1. $\inf \varphi = \inf \varphi^{\mathcal{A}}$, $\arg\min \varphi = G(\arg\min \varphi^{\mathcal{A}})$
2. $\varphi^{\mathcal{A}}$ real valued, lsc (in fact, **continuous!**)
3. φ level bounded iff $\varphi^{\mathcal{A}}$ level bounded

since $x \mapsto \mathcal{M}(w; x)$ continuous $\forall w$

[1] P. Patrinos and A. Bemporad. *Proximal Newton methods for convex composite optimization*, IEEE CDC, 2013

[2] AT, L. Stella and P. Patrinos. *Forward-backward envelope for the sum of two nonconvex functions*, SIOPT 28(3), 2018

[3] P. Patrinos, L. Stella and A. Bemporad. *Douglas-Rachford splitting: Complexity estimates and accelerated variants*, IEEE CDC, 2014

[4] AT and P. Patrinos. *Douglas-Rachford splitting and ADMM for nonconvex optimization: tight convergence results*, SIOPT 30(1), 2020

[5] Liu Y. and Yin W. *An envelope for Davis-Yin splitting and strict saddle-point avoidance*, JOTA 181(2), 2019

Convergence analysis

Proximal envelopes

To a **GPMM scheme** $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

we associate a **proximal envelope function**

$$\varphi^{\mathcal{A}}(s) := \min_w \mathcal{M}(w; x) \quad \text{where } x = G(s)$$

Key property: "sufficient decrease"

- $\rho_{\mathcal{M}} := m_1/m_2 \in (0, 1]$ recall: $\varphi(w) + \frac{m_1}{2}\|w - x\|^2 \le \mathcal{M}(w, x) \le \varphi(w) + \frac{m_2}{2}\|w - x\|^2$
- $\rho_G := \mu_G/L_G \in (0, 1]$ G μ_G-strongly monotone, L_G-Lipschitz

Then

$$\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - \frac{m_2}{2}\left(\frac{2\mu_G}{\lambda} - \frac{1 - \rho_{\mathcal{M}}}{\lambda^2} - L_G^2\right)\|s - s^+\|^2$$

Convergence analysis

Proximal envelopes

To a **GPMM scheme** $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

we associate a **proximal envelope function**

$$\varphi^{\mathcal{A}}(s) := \min_w \mathcal{M}(w; x) \quad \text{where } x = G(s)$$

Key property: "sufficient decrease"

- $\rho_{\mathcal{M}} := m_1/m_2 \in (0, 1]$ recall: $\varphi(w) + \frac{m_1}{2}\|w - x\|^2 \le \mathcal{M}(w, x) \le \varphi(w) + \frac{m_2}{2}\|w - x\|^2$
- $\rho_G := \mu_G/L_G \in (0, 1]$ G μ_G-strongly monotone, L_G-Lipschitz
- $\Delta := \rho_G^2 + \rho_{\mathcal{M}} - 1$

Then for $\Delta > 0$ and $\frac{\rho_G - \sqrt{\Delta}}{L_G} < \lambda < \frac{\rho_G + \sqrt{\Delta}}{L_G}$

$$\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - \frac{m_2}{2}\underbrace{\left(\frac{2\mu_G}{\lambda} - \frac{1 - \rho_{\mathcal{M}}}{\lambda^2} - L_G^2\right)}_{>0}\|s - s^+\|^2$$

Convergence analysis

Proximal envelopes

To a **GPMM scheme** $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

we associate a **proximal envelope function**

$$\varphi^{\mathcal{A}}(s) := \min_w \mathcal{M}(w; x) \quad \text{where } x = G(s)$$

Key property: "sufficient decrease"

- $\rho_{\mathcal{M}} := {}^{m_1}/_{m_2} \in (0, 1]$
- $\rho_G := {}^{\mu_G}/_{L_G} \in (0, 1]$
- $\Delta := \rho_G^2 + \rho_{\mathcal{M}} - 1$

Theorem. If $\Delta > 0$, GPMM scheme is convergent for some range of λ:

- FP residual $s^k - s^{k+1}$ vanishes
- cluster points $x = \bar{x}$ with $0 \in \hat{\partial}\varphi(x)$
- sequence bounded if φ level bounded

Then for $\Delta > 0$ and $\frac{\rho_G - \sqrt{\Delta}}{L_G} < \lambda < \frac{\rho_G + \sqrt{\Delta}}{L_G}$

$$\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - \frac{m_2}{2}\underbrace{\left(\frac{2\mu_G}{\lambda} - \frac{1 - \rho_{\mathcal{M}}}{\lambda^2} - L_G^2\right)}_{>0}\|s - s^+\|^2$$

Convergence analysis

Proof of sufficient decrease

GPMM scheme $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

with

1. Proximal majorizing model $\mathcal{M}$
   - 1a. $\mathcal{M}(w; x) \geq \varphi(w) + \frac{m_1}{2}\|w - x\|^2$
   - 1b. $\mathcal{M}(w; x) \leq \varphi(w) + \frac{m_2}{2}\|w - x\|^2$
2. Transient mapping G
   - 2a. L_G-Lipschitz
   - 2b. μ_G-strongly monotone

$$\varphi^{\mathcal{A}}(s^+) \overset{(def)}{=} \min_w \mathcal{M}(w; x^+) \leq \mathcal{M}(\bar{x}; x^+)$$

Convergence analysis

Proof of sufficient decrease

GPMM scheme $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

with

1. Proximal majorizing model $\mathcal{M}$
   - 1a. $\mathcal{M}(w; x) \geq \varphi(w) + \frac{m_1}{2}\|w - x\|^2$
   - 1b. $\mathcal{M}(w; x) \leq \varphi(w) + \frac{m_2}{2}\|w - x\|^2$
2. Transient mapping G
   - 2a. L_G-Lipschitz
   - 2b. μ_G-strongly monotone

$$\varphi^{\mathcal{A}}(s^+) \overset{(def)}{=} \min_w \mathcal{M}(w; x^+) \leq \mathcal{M}(\bar{x}; x^+)$$

1b. $$\leq \varphi(\bar{x}) + \frac{m_2}{2}\|x^+ - \bar{x}\|^2$$

Convergence analysis

Proof of sufficient decrease

GPMM scheme $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

with

1. Proximal majorizing model $\mathcal{M}$
   - 1a. $\mathcal{M}(w; x) \geq \varphi(w) + \frac{m_1}{2}\|w - x\|^2$
   - 1b. $\mathcal{M}(w; x) \leq \varphi(w) + \frac{m_2}{2}\|w - x\|^2$
2. Transient mapping G
   - 2a. L_G-Lipschitz
   - 2b. μ_G-strongly monotone

$$\varphi^{\mathcal{A}}(s^+) \overset{(def)}{=} \min_w \mathcal{M}(w; x^+) \leq \mathcal{M}(\bar{x}; x^+)$$

- 1b. $\leq \varphi(\bar{x}) + \frac{m_2}{2}\|x^+ - \bar{x}\|^2$
- 1a. $\leq \mathcal{M}(\bar{x}; x) - \frac{m_1}{2}\|x - \bar{x}\|^2 + \frac{m_2}{2}\|x^+ - \bar{x}\|^2$

Convergence analysis

Proof of sufficient decrease

GPMM scheme $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

with

1. Proximal majorizing model $\mathcal{M}$
   - 1a. $\mathcal{M}(w; x) \geq \varphi(w) + \frac{m_1}{2}\|w - x\|^2$
   - 1b. $\mathcal{M}(w; x) \leq \varphi(w) + \frac{m_2}{2}\|w - x\|^2$
2. Transient mapping G
   - 2a. L_G-Lipschitz
   - 2b. μ_G-strongly monotone

$$\varphi^{\mathcal{A}}(s^+) \overset{(def)}{=} \min_w \mathcal{M}(w; x^+) \leq \mathcal{M}(\bar{x}; x^+)$$

1b. $\leq \varphi(\bar{x}) + \frac{m_2}{2}\|x^+ - \bar{x}\|^2$

1a. $\leq \mathcal{M}(\bar{x}; x) - \frac{m_1}{2}\|x - \bar{x}\|^2 + \frac{m_2}{2}\|x^+ - \bar{x}\|^2$

$= \varphi^{\mathcal{A}}(s) - \frac{m_1}{2}\|x - \bar{x}\|^2 + \frac{m_2}{2}\|x^+ - x\|^2 + \frac{m_2}{2}\|x - \bar{x}\|^2 - m_2\left\langle x - x^+, x - \bar{x}\right\rangle$

Convergence analysis

Proof of sufficient decrease

GPMM scheme $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

with

1. Proximal majorizing model $\mathcal{M}$
   - 1a. $\mathcal{M}(w; x) \geq \varphi(w) + \frac{m_1}{2}\|w - x\|^2$
   - 1b. $\mathcal{M}(w; x) \leq \varphi(w) + \frac{m_2}{2}\|w - x\|^2$
2. Transient mapping G
   - 2a. L_G-Lipschitz
   - 2b. μ_G-strongly monotone

$$\begin{aligned}
\varphi^{\mathcal{A}}(s^+) &\overset{(def)}{=} \min_w \mathcal{M}(w; x^+) \leq \mathcal{M}(\bar{x}; x^+) \\
&\text{1b.} \quad \leq \varphi(\bar{x}) + \tfrac{m_2}{2}\|x^+ - \bar{x}\|^2 \\
&\text{1a.} \quad \leq \mathcal{M}(\bar{x}; x) - \tfrac{m_1}{2}\|x - \bar{x}\|^2 + \tfrac{m_2}{2}\|x^+ - \bar{x}\|^2 \\
&= \varphi^{\mathcal{A}}(s) - \tfrac{m_1}{2}\|x - \bar{x}\|^2 + \tfrac{m_2}{2}\|x^+ - x\|^2 + \tfrac{m_2}{2}\|x - \bar{x}\|^2 - m_2\left\langle x - x^+, x - \bar{x}\right\rangle \\
&\overset{\bar{x}-x = \frac{s^+ - s}{\lambda}}{=} \varphi^{\mathcal{A}}(s) + \tfrac{m_2 - m_1}{2\lambda^2}\|s - s^+\|^2 + \tfrac{m_2}{2}\|x^+ - x\|^2 - \tfrac{m_2}{\lambda}\left\langle x - x^+, s - s^+\right\rangle
\end{aligned}$$

Convergence analysis

Proof of sufficient decrease

GPMM scheme $\mathcal{A} = (\mathcal{M}, G)$

$$s \mapsto \begin{cases} x = G(s) \\ \bar{x} \in \mathcal{T}^{\mathcal{M}}(x) = \arg\min_w \mathcal{M}(w; x) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

with

1. Proximal majorizing model $\mathcal{M}$
   - 1a. $\mathcal{M}(w; x) \geq \varphi(w) + \frac{m_1}{2}\|w - x\|^2$
   - 1b. $\mathcal{M}(w; x) \leq \varphi(w) + \frac{m_2}{2}\|w - x\|^2$
2. Transient mapping G
   - 2a. L_G-Lipschitz
   - 2b. μ_G-strongly monotone

$$\begin{aligned}
\varphi^{\mathcal{A}}(s^+) &\overset{(def)}{=} \min_w \mathcal{M}(w; x^+) \leq \mathcal{M}(\bar{x}; x^+) \\
&\overset{1b.}{\leq} \varphi(\bar{x}) + \tfrac{m_2}{2}\|x^+ - \bar{x}\|^2 \\
&\overset{1a.}{\leq} \mathcal{M}(\bar{x}; x) - \tfrac{m_1}{2}\|x - \bar{x}\|^2 + \tfrac{m_2}{2}\|x^+ - \bar{x}\|^2 \\
&= \varphi^{\mathcal{A}}(s) - \tfrac{m_1}{2}\|x - \bar{x}\|^2 + \tfrac{m_2}{2}\|x^+ - x\|^2 + \tfrac{m_2}{2}\|x - \bar{x}\|^2 - m_2\left\langle x - x^+, x - \bar{x}\right\rangle \\
&\overset{\bar{x}-x = \frac{s^+ - s}{\lambda}}{=} \varphi^{\mathcal{A}}(s) + \tfrac{m_2 - m_1}{2\lambda^2}\|s - s^+\|^2 + \tfrac{m_2}{2}\|x^+ - x\|^2 - \tfrac{m_2}{\lambda}\left\langle x - x^+, s - s^+\right\rangle \\
&\overset{2.}{\leq} \varphi^{\mathcal{A}}(s) - \left(\tfrac{m_2\mu_G}{\lambda} - \tfrac{m_2 - m_1}{2\lambda^2} - \tfrac{m_2 L_G^2}{2}\right)\|s - s^+\|^2
\end{aligned}$$

Outline

Convergence analysis

Notable examples

Algorithm	Problem	Assumptions	GPMM components
Proximal gradient	minimize $f + g$	∇f Lipschitz	$G = \mathrm{id}$ $\mathcal{T}^{\mathcal{M}} = \mathrm{prox}_{\gamma g}(\mathrm{id} - \gamma \nabla f)$
Douglas-Rachford		∇f Lipschitz	$G = \mathrm{prox}_{\gamma f}$ $\mathcal{T}^{\mathcal{M}} = \mathrm{prox}_{\gamma g}(\mathrm{id} - \gamma \nabla f)$
Davis-Yin	minimize $f + g + h$	∇f Lipschitz ∇h Lipschitz	$G = \mathrm{prox}_{\gamma f}$ $\mathcal{T}^{\mathcal{M}} = \mathrm{prox}_{\gamma g}(\mathrm{id} - \gamma \nabla (f + h))$
ADMM	minimize $f(x) + g(z)$ **s.t.** $Ax + Bz = b$	equivalent to DRS after change of variable	

Convergence analysis

Notable examples

Algorithm	Problem	Assumptions	GPMM components
Proximal gradient	minimize $f + g$	∇f Lipschitz	$G = \mathrm{id}$ $\mathcal{T}^{\mathcal{M}} = \mathrm{prox}_{\gamma g}(\mathrm{id} - \gamma \nabla f)$
Douglas-Rachford		∇f Lipschitz	$G = \mathrm{prox}_{\gamma f}$ $\mathcal{T}^{\mathcal{M}} = \mathrm{prox}_{\gamma g}(\mathrm{id} - \gamma \nabla f)$
Davis-Yin	minimize $f + g + h$	∇f Lipschitz ∇h Lipschitz	$G = \mathrm{prox}_{\gamma f}$ $\mathcal{T}^{\mathcal{M}} = \mathrm{prox}_{\gamma g}(\mathrm{id} - \gamma \nabla (f + h))$
ADMM	minimize $f(x) + g(z)$ **s.t.** $Ax + Bz = b$	equivalent to DRS after change of variable	
Proximal ADMM*		equivalent to **ADMM** with slack variables	
Chambolle-Pock*	minimize $f + g \circ A$	special case of **proximal ADMM**	

* results **do not** directly apply (some assumptions are not met) (work in progress. . .)

Convergence analysis

Notable examples – Douglas-Rachford splitting

DRS as a GPMM scheme

$$\text{minimize}\, \varphi = f + g$$

$$\begin{cases} x = \operatorname{prox}_{\gamma f}(s) \\ \bar{x} \in \operatorname{prox}_{\gamma g}(2x - s) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

∇f L_f-Lipschitz, σ_f-hypomonotone $(-L_f \le \sigma_f \le L_f)$

DRS with stepsize γ is $\mathcal{A} = (\mathcal{M}, G)$ with

$$\mathcal{M}(\,\cdot\,; x) = f(x) + \langle \nabla f(x), \,\cdot\, - x\rangle + g + \tfrac{1}{2\gamma}\|\,\cdot\, - x\|^2 \qquad G(s) = \operatorname{prox}_{\gamma f(s)}$$

- $m_1 = \frac{1-\gamma L_f}{\gamma}$
- $m_2 = \frac{1-\gamma \sigma_f}{\gamma}$
- $\rho_{\mathcal{M}} = \frac{1-\gamma L_f}{1-\gamma \sigma_f}$

- $\mu_G = \frac{1}{1+\gamma L_f}$
- $L_G = \frac{1}{1+\gamma \sigma_f}$
- $\rho_G = \frac{1+\gamma \sigma_f}{1+\gamma L_f}$

DRS with stepsize $\gamma \in (0, 1/L_f)$ *"converges"* for any λ with

$$\frac{\rho_G - \sqrt{\Delta}}{L_G} < \lambda < \frac{\rho_G + \sqrt{\Delta}}{L_G} \qquad (\Delta := \rho_G^2 + \rho_{\mathcal{M}} - 1)$$

Convergence analysis

Notable examples – Douglas-Rachford splitting

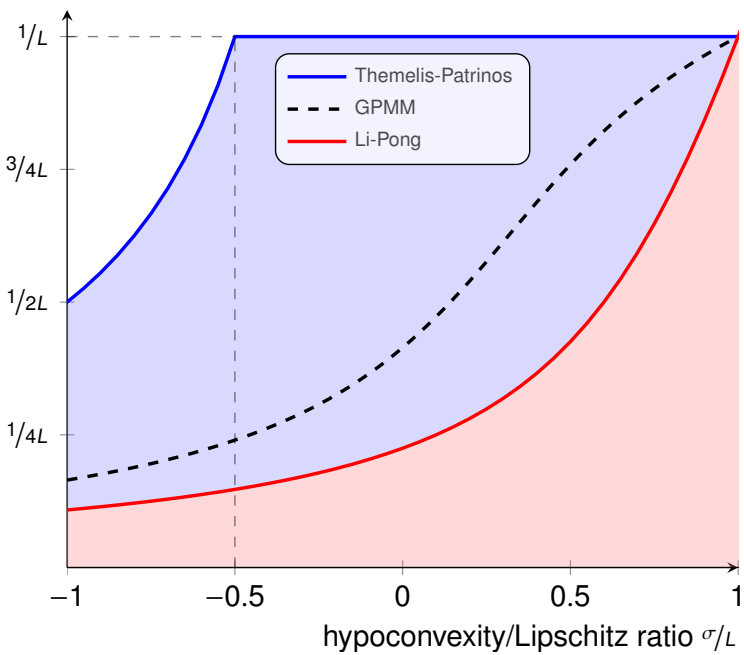
DRS as a GPMM scheme

$$\text{minimize}\, \varphi = f + g$$

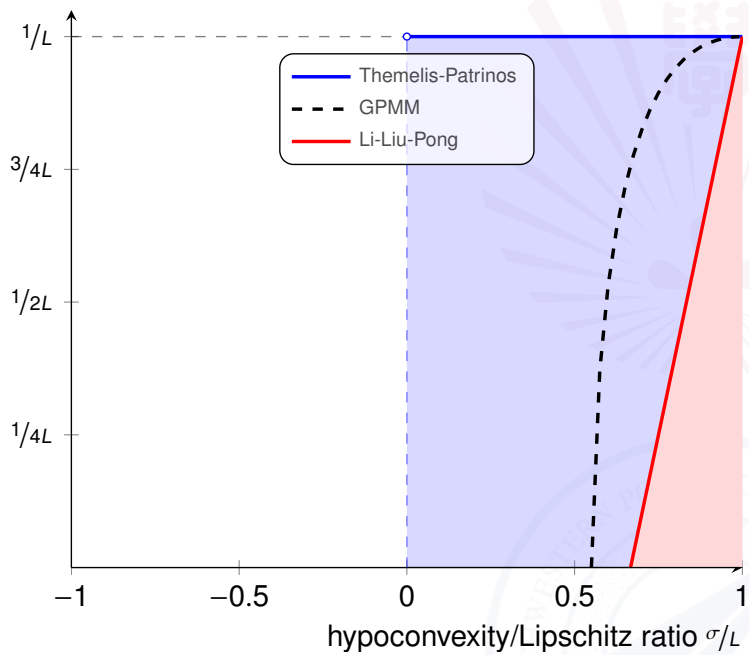
$$\begin{cases} x = \operatorname{prox}_{\gamma f}(s) \\ \bar{x} \in \operatorname{prox}_{\gamma g}(2x - s) \\ s^+ = s + \lambda(\bar{x} - x) \end{cases}$$

∇f L_f-Lipschitz, σ_f-hypomonotone $(-L_f \le \sigma_f \le L_f)$

Range of γ in DRS ($\lambda = 1$)



Range of γ in PRS ($\lambda = 2$)



Li G. and Pong TK. *Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems*, MathProg 159(2), 2015

Li G., Liu T. and Pong TK. *Peaceman-Rachford splitting for a class of nonconvex optimization problems*, COAP 68(2), 2017

AT and P. Patrinos. *Douglas-Rachford splitting and ADMM for nonconvex optimization: tight convergence results*, SIOPT 30(1), 2020

Convergence analysis

Notable examples – ADMM

ADMM as a transformed GPMM scheme

$$\begin{aligned} &\text{minimize } f(x) + g(z) \\ &\text{subject to } Ax + Bz = b \end{aligned} \qquad \begin{cases} x^+ \in \arg\min \mathcal{L}_\beta(\,\cdot\,, z, y) \\ y^+ = y + \beta(Ax + Bz - b) \\ z^+ \in \arg\min \mathcal{L}_\beta(x^+, \,\cdot\,, y^+) \end{cases}$$

Theorem (Yan & Yin, 2016 — Themelis & Patrinos, 2017)

ADMM is equivalent to DRS (with stepsize $\gamma = 1/\beta$) applied to

$$\operatorname{minimize}_s (Af)(s) + (Bg)(b - s)$$

where

$$(Ch)(s) := \inf_x \{h(x) \mid Cx = s\}$$

Convergence results for DRS readily translate!

- A surjective
- $\nabla(Af)$ $L_{(Af)}$-Lipschitz
- (Bg) lsc

Yan M. and Yin W. *Self equivalence of the alternating direction method of multipliers*, Springer, 2016

AT and P. Patrinos. *Douglas-Rachford splitting and ADMM for nonconvex optimization: tight convergence results*, SIOPT 30(1), 2020

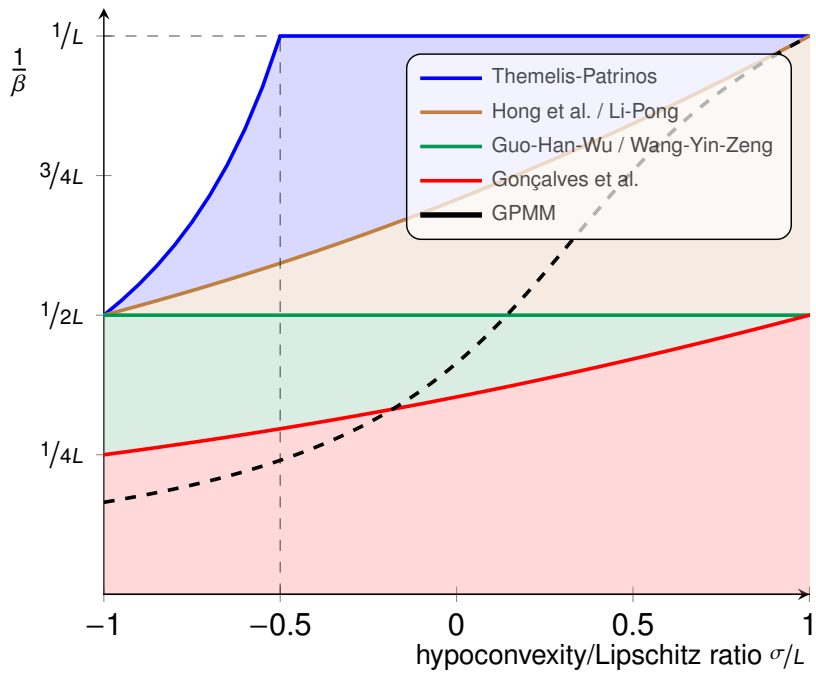
Convergence analysis

Notable examples – ADMM

ADMM as a transformed GPMM scheme

$$\text{minimize } f(x) + g(z) \quad \text{subject to } Ax + Bz = b$$

$$\begin{cases} x^+ \in \arg\min \mathcal{L}_\beta(\,\cdot\,, z, y) \\ y^+ = y + \beta(Ax + Bz - b) \\ z^+ \in \arg\min \mathcal{L}_\beta(x^+, \,\cdot\,, y^+) \end{cases}$$



AT and P. Patrinos. *Douglas-Rachford splitting and ADMM for nonconvex optimization: tight convergence results*, SIOPT 30(1), 2020

Hong M., Luo Z. and M. Razaviyayn. *Convergence Analysis of ADMM for a Family of Nonconvex Problems*, SIOPT 26(1), 2016

Li G. and Pong TK. *Global convergence of splitting methods for nonconvex composite optimization*, SIOPT 25(4), 2015

Wang Y., Yin W. and Zeng J. *Global convergence of ADMM in nonconvex nonsmooth optimization*, Jour. Sci. Comput. 78(29), 2019

M. Gonçalves, J. Melo and R. Monteiro. *Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems*, arXiv:1702.01850

Convergence analysis

Notable examples – ADMM

Ours[1]	Hong et al.[2]	Li and Pong[3]	Wang et al.[4]	Gonçalves et al.[5]
	g cvx or smooth			
f "A-smooth"	∇f Lipsch.	∇f Lipsch. $f \in C^2$	∇f Lipsch.	$\Pi_{A^\top} \nabla f$ Lipsch. f lower-C^2
$\mathcal{L}_\beta$ lev. bounded in x	$A = I$	$A = I$	$x(s)$ Lipsch.	A invertible
$z(s)$ loc. bound.	B full col. rank	B full row rank	$z(s)$ Lipsch.	

- $x(s) := \arg\min_x \{f(x) \mid Ax = s\}$ $\quad z(s) := \arg\min_z \{g(z) \mid Bz = s\}$
- f is *"A-smooth"* if
  $\nabla f(u), \nabla f(v) \in \operatorname{range} A^\top \;\Rightarrow\; \left|\langle \nabla f(u) - \nabla f(v), u - v\rangle\right| \le L\|A(u-v)\|^2$
  Ex. $A = [1\ 0]$, $f(x,y) = \frac{1}{2}x^2y^2$ is A-smooth, but $\Pi_{A^\top}\nabla f$ is not Lipschitz

Notice that

- A full column rank $\Rightarrow$ $x(s)$ Lipschitz & $\mathcal{L}_\beta$ level bounded in x
- B full column rank $\Rightarrow$ $z(s)$ Lipschitz $\Rightarrow$ $z(s)$ locally bounded

Disclaimer: Only compared against similar settings

- Fixed stepsize, no extrapolation, no inexactness
- For recent developements in this direction, see e.g.
  J. Bai, M. Zhang and H. Zhang. *An inexact ADMM for separable nonconvex and nonsmooth optimization*, 2022

---

[1] AT and P. Patrinos. *Douglas-Rachford splitting and ADMM for nonconvex optimization: tight convergence results*, SIOPT 30(1), 2020

[2] Hong M., Luo Z. and M. Razaviyayn. *Convergence analysis of ADMM for a family of nonconvex problems*, SIOPT 26(1), 2016

[3] Li G. and Pong TK. *Global convergence of splitting methods for nonconvex composite optimization*, SIOPT 25(4), 2015

[4] Wang Y., Yin W. and Zeng J. *Global convergence of ADMM in nonconvex nonsmooth optimization*, Jour. Sci. Comput. 78(29), 2019

[5] M. Gonçalves, J. Melo and R. Monteiro. *Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems*, arXiv:1702.01850

Convergence analysis

Notable examples – Davis-Yin splitting

DYS as a GPMM scheme

$$\text{minimize}\, \varphi = f + g + h$$

$$\begin{cases} x = \operatorname{prox}_{\gamma f}(s) & \overbrace{}^{x - \gamma\nabla(f+h)(x)} \\ \bar{x} \in \operatorname{prox}_{\gamma g}(2x - s - \gamma\nabla h(x)) & \\ s^+ = s + \lambda(\bar{x} - x) & \end{cases}$$

∇f L_f-Lipschitz
∇h L_h-Lipschitz

The inner step is a FBS step relative to $F + g$, where $F := f + h$

$$\mathcal{M}(\,\cdot\,; x) = F(x) + \langle \nabla F(x), \cdot - x\rangle + g + \tfrac{1}{2\gamma}\|\cdot - x\|^2 \qquad G(s) = \operatorname{prox}_{\gamma f(s)}$$

- $m_1 = \frac{1-\gamma L_F}{\gamma}$
- $m_2 = \frac{1-\gamma\sigma_F}{\gamma}$
- $\rho_{\mathcal{M}} = \frac{1-\gamma L_F}{1-\gamma\sigma_F}$

- $\mu_G = \frac{1}{1+\gamma L_f}$
- $L_G = \frac{1}{1+\gamma\sigma_f}$
- $\rho_G = \frac{1+\gamma\sigma_f}{1+\gamma L_f}$

DYS with stepsize $\gamma \in (0, 1/L_F)$ *"converges"* for any λ with

$$\frac{\rho_G - \sqrt{\Delta}}{L_G} < \lambda < \frac{\rho_G + \sqrt{\Delta}}{L_G} \qquad (\Delta := \rho_G^2 + \rho_{\mathcal{M}} - 1)$$

Proximal ADMM as a transformed GPMM scheme

$$\begin{aligned} &\text{minimize } f(x) + g(z) \\ &\text{subject to } Ax + Bz = b \end{aligned} \qquad \begin{cases} x^+ \in \arg\min \mathcal{L}_\beta(\,\cdot\,, z, y) + \frac{\beta}{2}\|Q(\,\cdot\, - x)\|^2 \\ y^+ = y + \beta(Ax + Bz - b) \\ z^+ \in \arg\min \mathcal{L}_\beta(x^+, \,\cdot\,, y^+) + \frac{\beta}{2}\|R(\,\cdot\, - z)\|^2 \end{cases}$$

Theorem (Bertsekas, 2016 — Themelis & Patrinos, 2019)

Proximal ADMM is equivalent to ADMM applied to

$$\underset{(x,\xi),\,(z,\zeta)}{\text{minimize}}\ f(x) + g(z) \quad \text{subject to} \begin{cases} Ax + Bz = b \\ Qx - \zeta = 0 \\ \xi - Rz = 0 \end{cases} \tag{P'}$$

However, need extra work (matrix $\binom{A}{Q}$ is likely *not* surjective...)

DP. Bertsekas. *Nonlinear Programming*, Athena scientific (3rd ed.), 2016

Convergence analysis

Notable examples – Chambolle-Pock

Chambolle-Pock as a transformed GPMM scheme

$$\begin{aligned} &\text{minimize } f(x) + g(z) \\ &\text{subject to } Ax = z \end{aligned} \qquad \begin{cases} x^+ \in \operatorname{prox}_{\tau f}(x - \tau A^\top (y + \sigma(Ax - z))) \\ y^+ = y + \sigma(Ax^+ - z) \\ z^+ \in \operatorname{prox}_{g/\sigma}(Ax^+ + {}^{y^+}/_{\sigma}) \end{cases}$$

Theorem (Shefi & Teboulle, 2014 — Boţ & Nguyen, 2018)

Chambolle-Pock is equivalent to proximal ADMM with

- $\beta = \sigma$
- $R = 0$
- Q such that $Q^\top Q = \frac{1}{\sigma\tau}\mathrm{I} - A^\top A$.

However, again need extra work (same issues. . .)

R. Shefi and M. Teboulle. *Rate of convergence analysis of decomposition methods based on the proximal method of multipliers for convex minimization*, SIOPT 24(1), 2014

R. Boţ and D. Nguyen. *The proximal ADMM in the nonconvex setting: convergence analysis and rates*, Math. Op. Res. 45(2), 2020

Outline

Outline

“Acceleration”

Challenges of higher-order methods

Example: Newton’s method for minimizing $f \in C^2$

$$f(x^k + d) \approx \underbrace{f(x^k) + \left\langle \nabla f(x^k), d \right\rangle + \tfrac{1}{2}\left\langle d, \nabla^2 f(x^k) d \right\rangle}_{\text{minimize wrt } d}$$

- minimize wrt d: $d^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$
- Far from solutions, $\approx$ inaccurate

$$x^{k+1} = x^k + d^k \quad \textbf{divergent!}$$

- Need to approach solutions first $\Rightarrow$ **linesearch**

E.g. Armijo: choose $\tau_k \in \{1, 1/2, 1/4, \ldots\}$ s.t.

$$f(x^k + \tau_k d^k) \le f(x^k) + \delta \tau_k \left\langle \nabla f(x^k), d^k \right\rangle$$

Works if

- $\exists$ Taylor expansion
- d^k is of descent: $\left\langle \nabla f(x^k), d^k \right\rangle < 0$

  otherwise e.g. $d^k = -(\nabla^2 f(x^k) + \mu I)^{-1} \nabla f(x^k)$

Works well if

- $\tau_k = 1$ when close to solutions

"Acceleration"

Challenges of higher-order methods

Example: Newton's method for minimizing $f \in C^2$

$$f(x^k + d) \approx \underbrace{f(x^k) + \left\langle \nabla f(x^k), d \right\rangle + \tfrac{1}{2}\left\langle d, \nabla^2 f(x^k) d \right\rangle}_{\text{minimize wrt } d}$$

- minimize wrt d: $d^k = -\nabla^2 f(x^k)^{-1} \nabla f(x^k)$
- Far from solutions, $\approx$ inaccurate

$$x^{k+1} = x^k + d^k \quad \textbf{divergent!}$$

- Need to approach solutions first $\Rightarrow$ **linesearch**

E.g. Armijo: choose $\tau_k \in \{1, 1/2, 1/4, \ldots\}$ s.t.

$$f(x^k + \tau_k d^k) \le f(x^k) + \delta \tau_k \left\langle \nabla f(x^k), d^k \right\rangle$$

Works if

- $\exists$ Taylor expansion
- d^k is of descent: $\left\langle \nabla f(x^k), d^k \right\rangle < 0$

  otherwise e.g. $d^k = -(\nabla^2 f(x^k) + \mu I)^{-1} \nabla f(x^k)$

Works well if

- $\tau_k = 1$ when close to solutions

Challenge

Design **nonsmooth** linesearch

- no Taylor expansion
- no notion of descent
- $\tau_k = 1$ close to solutions

Outline

“Acceleration”

Continuous-Lyapunov Descent framework

Continuous-Lyapunov Descent framework

Require $\alpha \in (0,1)$, $s \in \mathbb{R}^n$

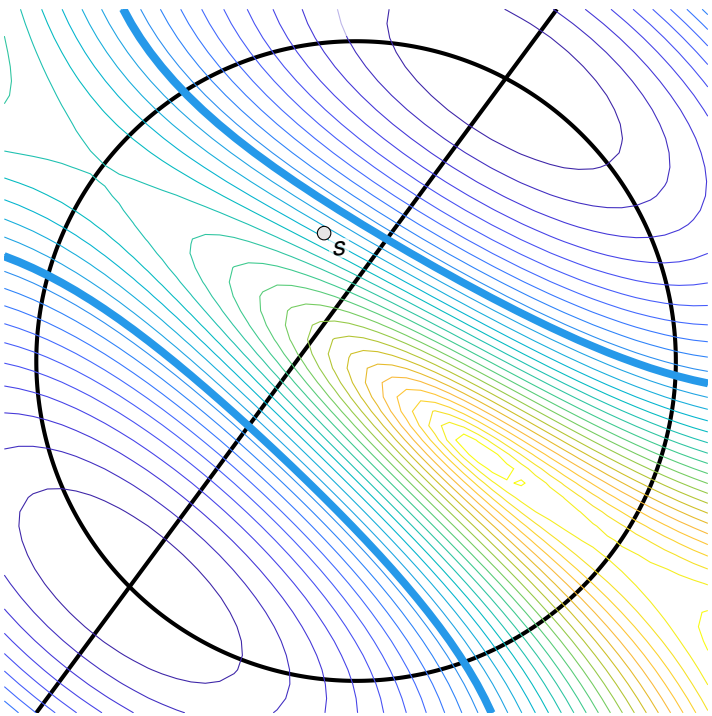
1: do one nominal FP iteration $s \mapsto s^+$ ▷ this ensures $\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - c\|s - s^+\|^2$

2: select $d \in \mathbb{R}^n$

3: backtrack $\tau \in (0,1]$ until $\tilde{s}^+ := (1-\tau)s^+ + \tau(s+d)$ satisfies

$$\varphi^{\mathcal{A}}(\tilde{s}^+) \le \varphi^{\mathcal{A}}(s) - \alpha c\|s - s^+\|^2 \qquad \text{(LS)}$$

4: $s \leftarrow \tilde{s}^+$ and go to step 1



This is how the NMPC solver **PANOC** works

L. Stella, AT, P. Sopasakis and P. Patrinos. *A simple and efficient algorithm for nonlinear model predictive control*, IEEE CDC, 2017

Alternating projections

(toy example)

Find intersection of line ℓ and sphere $\mathbb{S}$

Proximal gradient applied to

$$\text{minimize } \tfrac{1}{2}\operatorname{dist}^2(x,\ell) + \delta_{\mathbb{S}}(x)$$

“Acceleration”

Continuous-Lyapunov Descent framework

Continuous-Lyapunov Descent framework

Require $\alpha \in (0,1)$, $s \in \mathbb{R}^n$

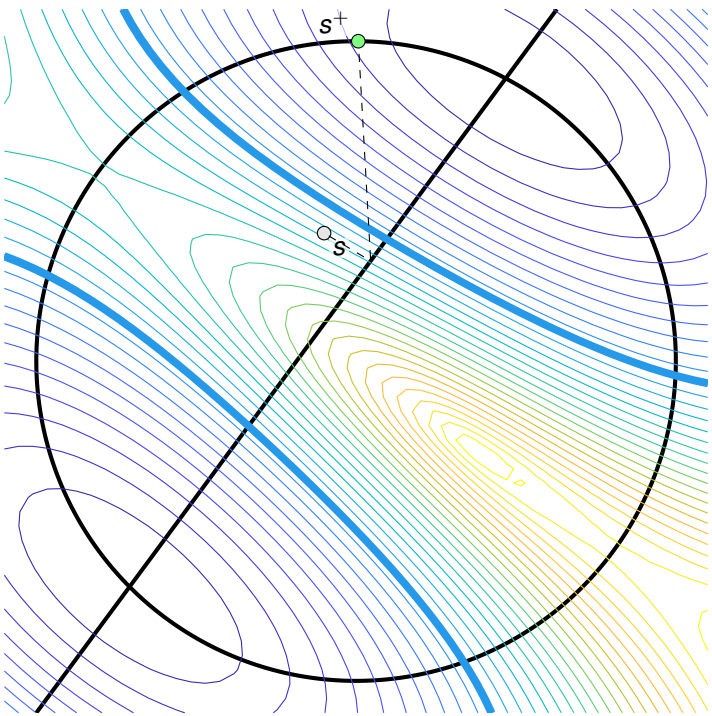
1: do one nominal FP iteration $s \mapsto s^+$ ▷ this ensures $\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - c\|s - s^+\|^2$

2: select $d \in \mathbb{R}^n$

3: backtrack $\tau \in (0,1]$ until $\tilde{s}^+ := (1-\tau)s^+ + \tau(s+d)$ satisfies

$$\varphi^{\mathcal{A}}(\tilde{s}^+) \le \varphi^{\mathcal{A}}(s) - \alpha c\|s - s^+\|^2 \qquad \text{(LS)}$$

4: $s \leftarrow \tilde{s}^+$ and go to step 1



This is how the NMPC solver **PANOC** works

L. Stella, AT, P. Sopasakis and P. Patrinos. *A simple and efficient algorithm for nonlinear model predictive control*, IEEE CDC, 2017

Alternating projections

(toy example)

Find intersection of line ℓ and sphere $\mathbb{S}$

Proximal gradient applied to

$$\text{minimize } \tfrac{1}{2}\operatorname{dist}^2(x,\ell) + \delta_{\mathbb{S}}(x)$$

“Acceleration”

Continuous-Lyapunov Descent framework

Continuous-Lyapunov Descent framework

Require $\alpha \in (0,1)$, $s \in \mathbb{R}^n$

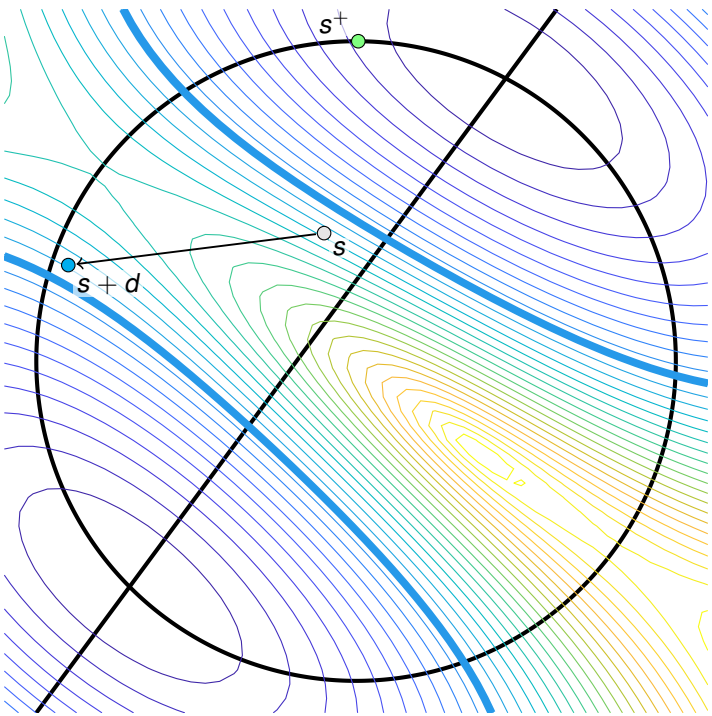
1: do one nominal FP iteration $s \mapsto s^+$ ▷ this ensures $\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - c\|s - s^+\|^2$

2: select $d \in \mathbb{R}^n$

3: backtrack $\tau \in (0,1]$ until $\tilde{s}^+ := (1-\tau)s^+ + \tau(s+d)$ satisfies

$$\varphi^{\mathcal{A}}(\tilde{s}^+) \le \varphi^{\mathcal{A}}(s) - \alpha c\|s - s^+\|^2 \qquad \text{(LS)}$$

4: $s \leftarrow \tilde{s}^+$ and go to step 1



This is how the NMPC solver **PANOC** works

L. Stella, AT, P. Sopasakis and P. Patrinos. *A simple and efficient algorithm for nonlinear model predictive control*, IEEE CDC, 2017

Alternating projections

(toy example)

Find intersection of line ℓ and sphere $\mathbb{S}$

Proximal gradient applied to

$$\text{minimize } \tfrac{1}{2}\,\text{dist}^2(x,\ell) + \delta_{\mathbb{S}}(x)$$

“Acceleration”

Continuous-Lyapunov Descent framework

Continuous-Lyapunov Descent framework

Require $\alpha \in (0,1)$, $s \in \mathbb{R}^n$

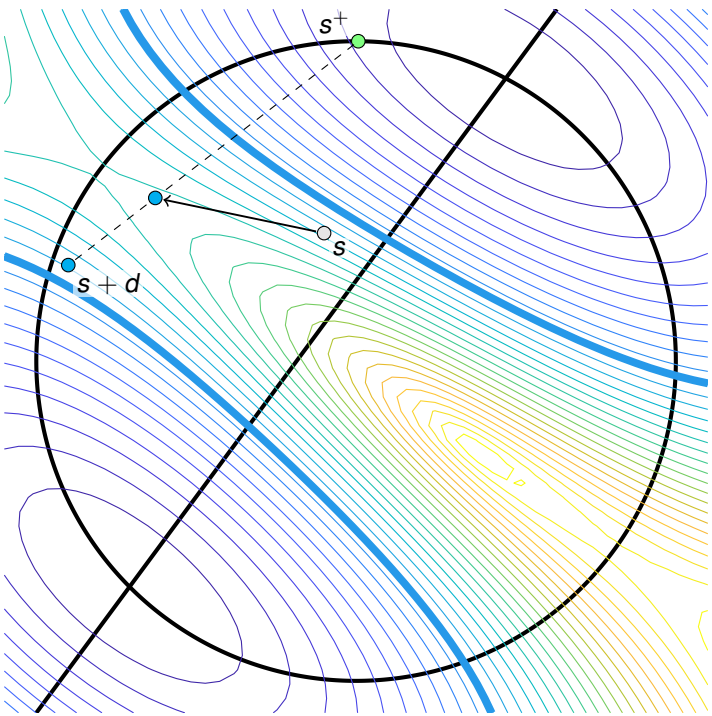
1: do one nominal FP iteration $s \mapsto s^+$ ▷ this ensures $\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - c\|s - s^+\|^2$

2: select $d \in \mathbb{R}^n$

3: backtrack $\tau \in (0,1]$ until $\tilde{s}^+ := (1-\tau)s^+ + \tau(s+d)$ satisfies

$$\varphi^{\mathcal{A}}(\tilde{s}^+) \le \varphi^{\mathcal{A}}(s) - \alpha c\|s - s^+\|^2 \qquad \text{(LS)}$$

4: $s \leftarrow \tilde{s}^+$ and go to step 1



This is how the NMPC solver **PANOC** works

L. Stella, AT, P. Sopasakis and P. Patrinos. *A simple and efficient algorithm for nonlinear model predictive control*, IEEE CDC, 2017

Alternating projections

(toy example)

Find intersection of line ℓ and sphere $\mathbb{S}$

Proximal gradient applied to

$$\text{minimize } \tfrac{1}{2}\operatorname{dist}^2(x,\ell) + \delta_{\mathbb{S}}(x)$$

“Acceleration”

Continuous-Lyapunov Descent framework

Continuous-Lyapunov Descent framework

Require $\alpha \in (0,1)$, $s \in \mathbb{R}^n$

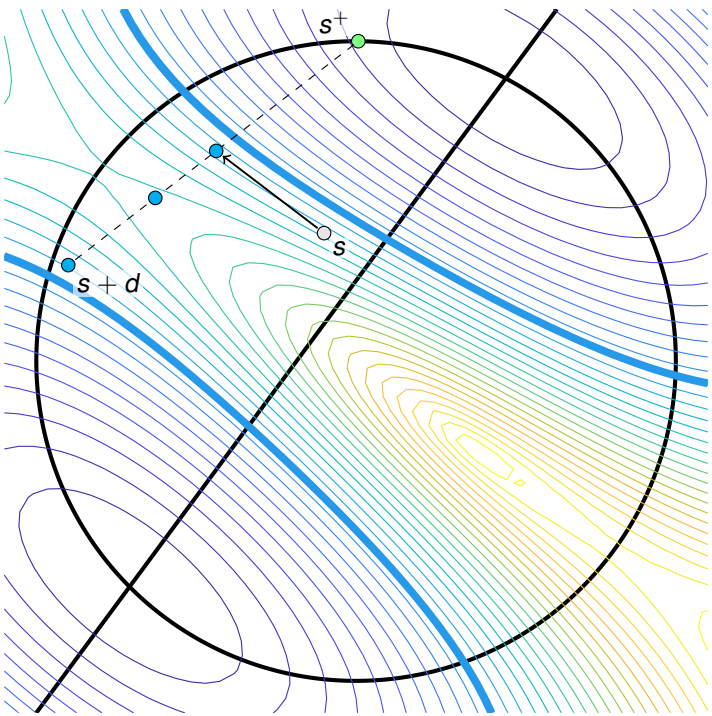
1: do one nominal FP iteration $s \mapsto s^+$ ▷ this ensures $\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - c\|s - s^+\|^2$

2: select $d \in \mathbb{R}^n$

3: backtrack $\tau \in (0,1]$ until $\tilde{s}^+ := (1-\tau)s^+ + \tau(s+d)$ satisfies

$$\varphi^{\mathcal{A}}(\tilde{s}^+) \le \varphi^{\mathcal{A}}(s) - \alpha c\|s - s^+\|^2 \qquad \text{(LS)}$$

4: $s \leftarrow \tilde{s}^+$ and go to step 1



This is how the NMPC solver **PANOC** works

L. Stella, AT, P. Sopasakis and P. Patrinos. *A simple and efficient algorithm for nonlinear model predictive control*, IEEE CDC, 2017

Alternating projections

(toy example)

Find intersection of line ℓ and sphere $\mathbb{S}$

Proximal gradient applied to

$$\text{minimize } \tfrac{1}{2}\operatorname{dist}^2(x,\ell) + \delta_{\mathbb{S}}(x)$$

“Acceleration”

Continuous-Lyapunov Descent framework

Continuous-Lyapunov Descent framework

Require $\alpha \in (0,1)$, $s \in \mathbb{R}^n$

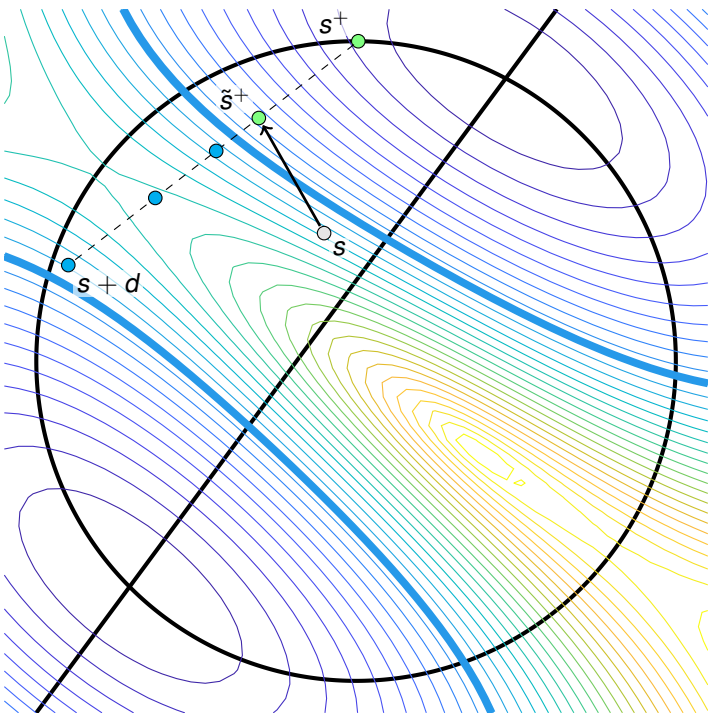
1: do one nominal FP iteration $s \mapsto s^+$ ▷ this ensures $\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - c\|s - s^+\|^2$

2: select $d \in \mathbb{R}^n$

3: backtrack $\tau \in (0,1]$ until $\tilde{s}^+ := (1-\tau)s^+ + \tau(s+d)$ satisfies

$$\varphi^{\mathcal{A}}(\tilde{s}^+) \le \varphi^{\mathcal{A}}(s) - \alpha c\|s - s^+\|^2 \qquad \text{(LS)}$$

4: $s \leftarrow \tilde{s}^+$ and go to step 1



This is how the NMPC solver **PANOC** works

L. Stella, AT, P. Sopasakis and P. Patrinos. *A simple and efficient algorithm for nonlinear model predictive control*, IEEE CDC, 2017

Alternating projections

(toy example)

Find intersection of line ℓ and sphere $\mathbb{S}$

Proximal gradient applied to

$$\text{minimize } \tfrac{1}{2}\,\text{dist}^2(x,\ell) + \delta_{\mathbb{S}}(x)$$

“Acceleration”

Continuous-Lyapunov Descent framework

Continuous-Lyapunov Descent framework

Require $\alpha \in (0,1)$, $s \in \mathbb{R}^n$

1: do one nominal FP iteration $s \mapsto s^+$ ▷ this ensures $\varphi^{\mathcal{A}}(s^+) \le \varphi^{\mathcal{A}}(s) - c\|s - s^+\|^2$

2: select $d \in \mathbb{R}^n$

3: backtrack $\tau \in (0,1]$ until $\tilde{s}^+ := (1-\tau)s^+ + \tau(s+d)$ satisfies

$$\varphi^{\mathcal{A}}(\tilde{s}^+) \le \varphi^{\mathcal{A}}(s) - \alpha c\|s - s^+\|^2 \qquad \text{(LS)}$$

4: $s \leftarrow \tilde{s}^+$ and go to step 1

Features.

1. **CLyD** well defined for **any choice** of d (since $\varphi^{\mathcal{A}}$ is continuous)
2. **Same convergence** of FP iterations!
3. **Same computations** as in FP iterations! (no geometry change)
4. **No Maratos effect:** eventually $\tau = 1$ if directions are “good”
   - **Superlinear** rates with Broyden directions (under assumptions)
   - L-BFGS very effective in practice

This is how the NMPC solver **PANOC** works

L. Stella, AT, P. Sopasakis and P. Patrinos. *A simple and efficient algorithm for nonlinear model predictive control*, IEEE CDC, 2017

Outline

Sparse principal component analysis

$$\text{maximize } \langle z, \Sigma z \rangle \quad \text{subject to } \|z\|_2 = 1,\ \|z\|_0 \le k$$

- $\Sigma = A^\top A$ covariance matrix of data matrix $A \in \mathbb{R}^{m \times n}$
- explain as much variability in data by using only $k \ll n$ variables

Centralized SPCA formulation

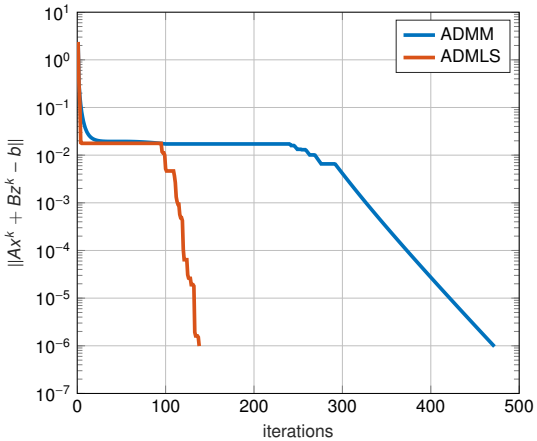
$$\begin{aligned} &\text{minimize} && -\|Az\|_2^2 && \text{concave cost} \\ &\text{subject to } \|z\|_2 = 1, && \|z\|_0 \le k && \text{nonconvex constraints} \end{aligned}$$

Distributed SPCA formulation: introduce copies $x_1, \ldots, x_N$ of z

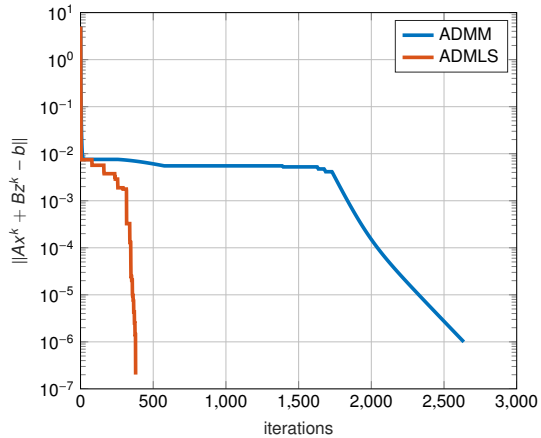
$$\text{minimize } \sum_{i=1}^{N} \underbrace{-\|A_i x_i\|_2^2}_{f_i(x_i)} + g(z) \quad \text{subject to } x_i = z$$

- data distributed across different agents/workers or A is huge
- each term $\frac{1}{2}\|A_i x_i\|^2$ prox-ed **separately**
- no exchange of data A_i, only variables

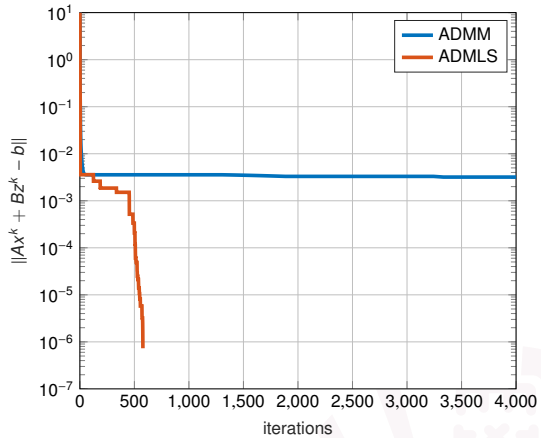
Consensus SPCA: example



$N = 5$



$N = 25$



$N = 100$

- $n = 100,000$ features, $m = 50,000$ data points
- rows are split into N subsets

Pre-processing: Computing prox of $-\|A_i x_i\|^2$ requires factoring (once)

$$I - \gamma A_i A_i^\top \in \mathbb{R}^{m_i \times m_i}$$

- $N = 1$ workers $\Rightarrow m_1 = m = 50,000$, $>$ 1 hour
- $N = 5$ workers $\Rightarrow m_i = 10,000$, $\approx$ 7 seconds
- $N = 50$ workers $\Rightarrow m_i = 1,000$, $\approx$ 0.03 seconds

---

AT, L. Stella and P. Patrinos. *Douglas-Rachford splitting and ADMM for nonconvex optimization: Accelerated and Newton-type algorithms*, COAP 82, 2022

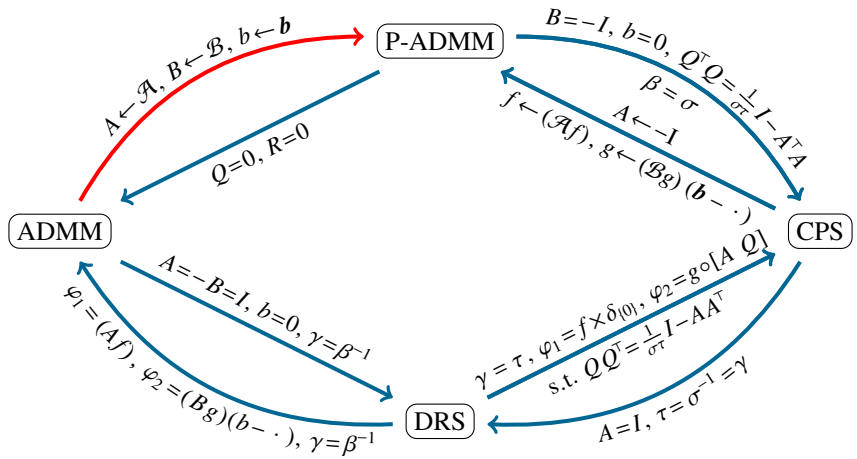
Outline

Conclusions

In this talk

Nonconvex splitting algorithms: **unified** interpretation as **GPMM** schemes

- Proximal envelopes: continuous Lyapunov functions
- *"Acceleration"* with **CLyD** (nonsmooth linesearch)
- Equivalences DRS ⇔ ADMM ⇔ proximal ADMM ⇔ Chambolle-Pock



Extensions

- Bregman metrics (relative smoothness)[1,2]
- Block-coordinate updates [2]
- **Locally** Lipschitz ∇f [3,4]

---

[1] M. Ahookhosh, AT, and P. Patrinos. *A Bregman forward-backward linesearch algorithm for nonconvex composite optimization: Superlinear convergence to nonisolated local minima*, SIOPT 31(1), 2021

[2] P. Latafat, AT, M. Ahookhosh, and P. Patrinos. *Bregman Finito/MISO for nonconvex regularized finite sum minimization without Lipschitz gradient continuity*, SIOPT 32(3), 2022

[3] A. De Marchi and AT. *Proximal gradient algorithms under local Lipschitz gradient continuity: A convergence and robustness analysis of PANOC*, JOTA 194, 2022

[4] A. De Marchi and AT. *An interior proximal gradient method for nonconvex optimization*, arXiv:2208.00799

Appendix

Universal equivalence of ADMM and proximal ADMM

$$\text{minimize } f(x)+g(z) \qquad \text{(P')}$$

$$\text{subject to } \begin{cases} Ax+Bz=b \\ Qx-\zeta=0 \\ \xi-Rz=0 \end{cases}$$

$$\tilde{\mathcal{L}}_\beta := \mathcal{L}_\beta(x,z,y)+\langle u, Qx-\zeta\rangle+\langle v,\xi-Rz\rangle + \tfrac{\beta}{2}\|Qx-\zeta\|^2+\tfrac{\beta}{2}\|\xi-Rz\|^2$$

ADMM on (P')

$$\begin{cases} \begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta(\,\cdot\,,\begin{pmatrix} z \\ \zeta \end{pmatrix},\begin{pmatrix} y \\ u \\ v \end{pmatrix}) & \bullet \\ \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix} = \begin{pmatrix} y \\ u \\ v \end{pmatrix} + \beta \begin{pmatrix} Ax^+ + Bz - b \\ Qx^+ - \zeta \\ \xi^+ - Rz \end{pmatrix} & \bullet \\ \begin{pmatrix} z^+ \\ \zeta^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta(\begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix},\,\cdot\,,\begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix}) & \end{cases}$$

Proof ($\lambda = 1$ for simplicity).

$\begin{pmatrix} x \\ \xi \end{pmatrix}$- and v-updates

1. **a)** $x^+ \in \arg\min \mathcal{L}_\beta(\,\cdot\,,z,y)+\frac{\beta}{2}\|Q\,\cdot\,-\zeta+u/\beta\|^2$

   b) $\xi^+ = Rz - v/\beta \quad\Rightarrow\quad v^+ = 0 \quad\Rightarrow\quad v=0,\ \xi^+ = Rz$

Appendix

Universal equivalence of ADMM and proximal ADMM

$$\text{minimize } f(x)+g(z) \qquad \text{(P')}$$

$$\text{subject to } \begin{cases} Ax+Bz=b \\ Qx-\zeta=0 \\ \xi-Rz=0 \end{cases}$$

$$\tilde{\mathcal{L}}_\beta := \mathcal{L}_\beta(x,z,y)+\langle u, Qx-\zeta\rangle+\langle v,\xi-Rz\rangle + \tfrac{\beta}{2}\|Qx-\zeta\|^2+\tfrac{\beta}{2}\|\xi-Rz\|^2$$

ADMM on (P')

$$\begin{cases} \begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\cdot, \begin{pmatrix} z \\ \zeta \end{pmatrix}, \begin{pmatrix} y \\ u \\ v \end{pmatrix}\right) \\ \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix} = \begin{pmatrix} y \\ u \\ v \end{pmatrix} + \beta \begin{pmatrix} Ax^+ + Bz - b \\ Qx^+ - \zeta \\ \xi^+ - Rz \end{pmatrix} \\ \begin{pmatrix} z^+ \\ \zeta^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix}, \cdot, \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix}\right) \end{cases}$$

Proof ($\lambda = 1$ for simplicity).

$\begin{pmatrix} z \\ \zeta \end{pmatrix}$- and u-updates

1. **a)** $x^+ \in \arg\min \mathcal{L}_\beta(\cdot, z, y) + \frac{\beta}{2}\|Q\cdot - \zeta + u/\beta\|^2$

   b) $\xi^+ = Rz - v/\beta \quad \Rightarrow \quad v^+ = 0 \quad \Rightarrow \quad v = 0,\ \xi^+ = Rz$

2. **a)** $z^+ \in \arg\min \mathcal{L}_\beta(x^+, \cdot, y^+) + \frac{\beta}{2}\|\xi^+ - R\cdot + v^+/\beta\|^2$

   b) $\zeta^+ = Qx^+ + u^+/\beta \quad \Rightarrow \quad \zeta - u/\beta = Qx$

Appendix

Universal equivalence of ADMM and proximal ADMM

$$\text{minimize } f(x)+g(z) \qquad \text{(P')}$$

$$\text{subject to } \begin{cases} Ax+Bz=b \\ Qx-\zeta=0 \\ \xi-Rz=0 \end{cases}$$

$$\tilde{\mathcal{L}}_\beta := \mathcal{L}_\beta(x,z,y)+\langle u, Qx-\zeta\rangle+\langle v,\xi-Rz\rangle + \tfrac{\beta}{2}\|Qx-\zeta\|^2+\tfrac{\beta}{2}\|\xi-Rz\|^2$$

ADMM on (P')

$$\begin{cases} \begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\cdot, \begin{pmatrix} z \\ \zeta \end{pmatrix}, \begin{pmatrix} y \\ u \\ v \end{pmatrix}\right) \\ \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix} = \begin{pmatrix} y \\ u \\ v \end{pmatrix} + \beta \begin{pmatrix} Ax^+ + Bz - b \\ Qx^+ - \zeta \\ \xi^+ - Rz \end{pmatrix} \\ \begin{pmatrix} z^+ \\ \zeta^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix}, \cdot, \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix}\right) \end{cases}$$

Proof ($\lambda = 1$ for simplicity).

replace $\zeta - u/\beta$ from **(2b)** in **(1a)**

1. **a)** $x^+ \in \arg\min \mathcal{L}_\beta(\cdot, z, y) + \frac{\beta}{2}\|Q\cdot - \zeta + u/\beta\|^2$

   b) $\xi^+ = Rz - v/\beta \quad \Rightarrow \quad v^+ = 0 \quad \Rightarrow \quad v = 0,\ \xi^+ = Rz$

2. **a)** $z^+ \in \arg\min \mathcal{L}_\beta(x^+, \cdot, y^+) + \frac{\beta}{2}\|\xi^+ - R\cdot + v^+/\beta\|^2$

   b) $\zeta^+ = Qx^+ + u^+/\beta \quad \Rightarrow \quad \zeta - u/\beta = Qx$

3. $x^+ \in \arg\min \mathcal{L}_\beta(\cdot, z, y) + \frac{\beta}{2}\|Q(\cdot - x)\|^2$

Appendix

Universal equivalence of ADMM and proximal ADMM

$$\text{minimize } f(x)+g(z) \qquad \text{(P')}$$

$$\text{subject to } \begin{cases} Ax+Bz=b \\ Qx-\zeta=0 \\ \xi-Rz=0 \end{cases}$$

$$\tilde{\mathcal{L}}_\beta := \mathcal{L}_\beta(x,z,y)+\langle u, Qx-\zeta\rangle+\langle v,\xi-Rz\rangle + \tfrac{\beta}{2}\|Qx-\zeta\|^2+\tfrac{\beta}{2}\|\xi-Rz\|^2$$

ADMM on (P')

$$\begin{cases} \begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\cdot, \begin{pmatrix} z \\ \zeta \end{pmatrix}, \begin{pmatrix} y \\ u \\ v \end{pmatrix}\right) \\ \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix} = \begin{pmatrix} y \\ u \\ v \end{pmatrix} + \beta \begin{pmatrix} Ax^+ + Bz - b \\ Qx^+ - \zeta \\ \xi^+ - Rz \end{pmatrix} \\ \begin{pmatrix} z^+ \\ \zeta^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix}, \cdot, \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix}\right) \end{cases}$$

Proof ($\lambda = 1$ for simplicity).

y-update

1. **a)** $x^+ \in \arg\min \mathcal{L}_\beta(\cdot, z, y) + \frac{\beta}{2}\|Q\cdot - \zeta + u/\beta\|^2$

   b) $\xi^+ = Rz - v/\beta \quad \Rightarrow \quad v^+ = 0 \quad \Rightarrow \quad v = 0,\ \xi^+ = Rz$

2. **a)** $z^+ \in \arg\min \mathcal{L}_\beta(x^+, \cdot, y^+) + \frac{\beta}{2}\|\xi^+ - R\cdot + v^+/\beta\|^2$

   b) $\zeta^+ = Qx^+ + u^+/\beta \quad \Rightarrow \quad \zeta - u/\beta = Qx$

3. $x^+ \in \arg\min \mathcal{L}_\beta(\cdot, z, y) + \frac{\beta}{2}\|Q(\cdot - x)\|^2$

4. $y^+ = y + \beta(Ax^+ + Bz - b)$

Appendix

Universal equivalence of ADMM and proximal ADMM

$$\text{minimize } f(x)+g(z) \qquad \text{(P')}$$

$$\text{subject to } \begin{cases} Ax+Bz=b \\ Qx-\zeta=0 \\ \xi-Rz=0 \end{cases}$$

$$\tilde{\mathcal{L}}_\beta := \mathcal{L}_\beta(x,z,y)+\langle u, Qx-\zeta\rangle+\langle v,\xi-Rz\rangle + \tfrac{\beta}{2}\|Qx-\zeta\|^2+\tfrac{\beta}{2}\|\xi-Rz\|^2$$

ADMM on (P')

$$\begin{cases} \begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\cdot, \begin{pmatrix} z \\ \zeta \end{pmatrix}, \begin{pmatrix} y \\ u \\ v \end{pmatrix}\right) \\ \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix} = \begin{pmatrix} y \\ u \\ v \end{pmatrix} + \beta \begin{pmatrix} Ax^+ + Bz - b \\ Qx^+ - \zeta \\ \xi^+ - Rz \end{pmatrix} \\ \begin{pmatrix} z^+ \\ \zeta^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix}, \cdot, \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix}\right) \end{cases}$$

Proof ($\lambda = 1$ for simplicity).

replace ξ^+ and v^+ from **(1b)** in **(2a)**

1. **a)** $x^+ \in \arg\min \mathcal{L}_\beta(\cdot, z, y) + \frac{\beta}{2}\|Q\cdot - \zeta + u/\beta\|^2$

   b) $\xi^+ = Rz - v/\beta \quad \Rightarrow \quad v^+ = 0 \quad \Rightarrow \quad v = 0,\ \xi^+ = Rz$

2. **a)** $z^+ \in \arg\min \mathcal{L}_\beta(x^+, \cdot, y^+) + \frac{\beta}{2}\|\xi^+ - R\cdot + v^+/\beta\|^2$

   b) $\zeta^+ = Qx^+ + u^+/\beta \quad \Rightarrow \quad \zeta - u/\beta = Qx$

3. $x^+ \in \arg\min \mathcal{L}_\beta(\cdot, z, y) + \frac{\beta}{2}\|Q(\cdot - x)\|^2$

4. $y^+ = y + \beta(Ax^+ + Bz - b)$

5. $z^+ \in \arg\min \mathcal{L}_\beta(x^+, \cdot, y) + \frac{\beta}{2}\|R(\cdot - z)\|^2$

Appendix

Universal equivalence of ADMM and proximal ADMM

$$\text{minimize } f(x)+g(z) \qquad \text{(P')}$$
$$\text{subject to } \begin{cases} Ax+Bz=b \\ Qx-\zeta=0 \\ \xi-Rz=0 \end{cases}$$
$$\tilde{\mathcal{L}}_\beta := \mathcal{L}_\beta(x,z,y)+\langle u, Qx-\zeta\rangle+\langle v,\xi-Rz\rangle + \tfrac{\beta}{2}\|Qx-\zeta\|^2+\tfrac{\beta}{2}\|\xi-Rz\|^2$$

ADMM on (P')

$$\begin{cases} \begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\cdot, \begin{pmatrix} z \\ \zeta \end{pmatrix}, \begin{pmatrix} y \\ u \\ v \end{pmatrix}\right) \\ \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix} = \begin{pmatrix} y \\ u \\ v \end{pmatrix} + \beta \begin{pmatrix} Ax^+ + Bz - b \\ Qx^+ - \zeta \\ \xi^+ - Rz \end{pmatrix} \\ \begin{pmatrix} z^+ \\ \zeta^+ \end{pmatrix} \in \arg\min \tilde{\mathcal{L}}_\beta\left(\begin{pmatrix} x^+ \\ \xi^+ \end{pmatrix}, \cdot, \begin{pmatrix} y^+ \\ u^+ \\ v^+ \end{pmatrix}\right) \end{cases}$$

Proof ($\lambda = 1$ for simplicity).

This is (Q,R)-proximal ADMM!

1. **a)** $x^+ \in \arg\min \mathcal{L}_\beta(\cdot, z, y) + \frac{\beta}{2}\|Q\cdot - \zeta + u/\beta\|^2$
   b) $\xi^+ = Rz - v/\beta \quad\Rightarrow\quad v^+ = 0 \quad\Rightarrow\quad v = 0,\ \xi^+ = Rz$
2. **a)** $z^+ \in \arg\min \mathcal{L}_\beta(x^+, \cdot, y^+) + \frac{\beta}{2}\|\xi^+ - R\cdot + v^+/\beta\|^2$
   b) $\zeta^+ = Qx^+ + u^+/\beta \quad\Rightarrow\quad \zeta - u/\beta = Qx$
3. $x^+ \in \arg\min \mathcal{L}_\beta(\cdot, z, y) + \frac{\beta}{2}\|Q(\cdot - x)\|^2$
4. $y^+ = y + \beta(Ax^+ + Bz - b)$
5. $z^+ \in \arg\min \mathcal{L}_\beta(x^+, \cdot, y) + \frac{\beta}{2}\|R(\cdot - z)\|^2$

□