Accelerated Douglas-Rachford splitting and ADMM for structured nonconvex optimization

Panos Patrinos
KU Leuven (ESAT-STADIUS)
joint work with Andreas Themelis and Lorenzo Stella

CMO-BIRS Workshop
Splitting Algorithms, Modern Operator Theory and Applications
Oaxaca, Mexico

September 18, 2017

A. Themelis, L. Stella and P. Patrinos
Douglas–Rachford splitting and ADMM for nonconvex optimization: new convergence results and accelerated versions
https://arxiv.org/abs/1709.05747

Structured nonconvex optimization

composite problem

$$\textbf{minimize} \ \ \varphi_1(s) + \varphi_2(s)$$

separable problem

$$\begin{array}{ll} \textbf{minimize} & f(x) + g(z) \\ \textbf{subject to} & Ax + Bz = b \end{array}$$

- templates for large-scale structured optimization
- φ_1, φ_2, f, g can be nonsmooth
- numerous applications
  - machine learning
  - statistics
  - signal/image processing,
  - control...
- traditional algorithms usually do not apply

Structured nonconvex optimization

composite problem

$$\textbf{minimize}\ \varphi_1(s) + \varphi_2(s)$$

separable problem

$$\begin{aligned} \textbf{minimize}\quad & f(x) + g(z) \\ \textbf{subject to}\quad & Ax + Bz = b \end{aligned}$$

- resurgence of proximal algorithms (or operator splitting methods)
- reduce complex problem into a series of simpler subproblems
- perhaps most popular proximal algorithms

Douglas-Rachford Splitting (DRS)
Alternating Direction Method of Multipliers (ADMM)

- elegant, complete theory for **convex problems**
  (monotone operators, fixed-point iterations, Fejér sequences…[1])

[1] Bauschke H.H. and Combettes P.L. **Convex Analysis and Monotone Operator Theory in Hilbert Spaces.** Springer 2011

Contribution

composite problem

$$\textbf{minimize} \quad \varphi_1(s) + \varphi_2(s)$$

separable problem

$$\begin{aligned} \textbf{minimize} &\quad f(x) + g(z) \\ \textbf{subject to} &\quad Ax + Bz = b \end{aligned}$$

DRS & ADMM

- being fixed point iterations, DRS & ADMM can be **agonizingly slow**
- **nonconvex problems:** incomplete theory, results empirical or local[1,2]
- global results have recently emerged (see next slides)

this talk

- global convergence theory for **nonconvex problems** based on the

Douglas-Rachford Envelope (DRE)

- more importantly, **new, robust, faster algorithms**

[1] R. Hesse and R. Luke **Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems**. SIAM Opt. 23(4) 2013

[2] F. Artacho, J. Borwein and M. Tam **Recent Results on Douglas–Rachford Methods for Combinatorial Optimization Problems**. JOTA 163(1) 2014

Many applications...

- **ADMM:** amenable for **distributed** formulations (via **consensus**)
- **Nonconvex problems:** no need for convex relaxation
  rank constraints, 0/Schatten-norms, (mixed-) integer programming

Some examples:

- hybrid system MPC[1]
- distributed sparse principal component analysis (SPCA)[2]
- dictionary learning[3]
- background-foreground extraction[4,5]
- sparse representations (signal processing)[6]

---

[1] Takapoui R., Moehle N., Boyd S. and Bemporad A. **A simple effective heuristic for embedded mixed-integer quadratic programming**. IEEE ACC 2016

[2] Hajinezhad D. and Hong M. **Nonconvex ADMM for distributed sparse principal component analysis**. GlobalSIP 2015

[3] Wai H. T., Chang T. H. and Scaglione A. **A consensus-based decentralized algorithm for non-convex optimization with application to dictionary learning**. ICASSP 2015

[4] Chartrand R. **Nonconvex splitting for regularized low-rank + sparse decomposition**. IEEE TSP 2012

[5] Yang L., Pong T. K. and Chen X. **ADMM for a class of nonconvex and nonsmooth problems with applications to background/-foreground extraction**. SIAM 2017

[6] Chartrand R. and Wohlberg B. **A nonconvex ADMM algorithm for group sparsity with sparse groups**. ICASSP 2013

DRS for nonconvex problems

to solve

$$\textbf{minimize} \quad \varphi_1(s) + \varphi_2(s)$$

starting from $s \in \mathbb{R}^n$, iterate

$$\begin{aligned} u &= \mathbf{prox}_{\gamma\varphi_1}(s) \\ v &\in \mathbf{prox}_{\gamma\varphi_2}(2u - s) \\ s^+ &= s + \lambda(v - u) \end{aligned}$$

standing assumptions

1. φ_1 and φ_2 are *prox-friendly*, however **both can be nonconvex**
2. $\mathbf{dom}\,\varphi_1$ is affine and $\nabla\varphi_1$ is Lipschitz on $\mathbf{dom}\,\varphi_1$
3. $\varphi_2 + \frac{1}{2\gamma}\|\cdot\|^2$ is bounded below for some $\gamma > 0$ (**prox-bounded**)
4. $\mathbf{dom}\,\varphi_2 \subseteq \mathbf{dom}\,\varphi_1$

Structured Optimization

Tools: proximal map

Only **proximal operations** on φ_1 and φ_2:

$$\mathbf{prox}_{\gamma h}(s) = \mathop{\mathbf{argmin}}_{w} \left\{ h(w) + \tfrac{1}{2\gamma}\|w - s\|^2 \right\}, \qquad \gamma > 0$$

- a *generalized projection*: for $h = \delta_C$, $\mathbf{prox}_{\gamma h} = \mathbf{\Pi}_C$

Properties

- well defined for small γ
- Lipschitz for φ_1 (for small γ), but **set-valued** for φ_2
- *"prox-friendly"* (easily proximable) in many useful applications
- the value function is the **Moreau envelope**

$$h^\gamma(s) := \min_{w} \left\{ h(w) + \tfrac{1}{2\gamma}\|w - s\|^2 \right\}$$

- h^γ is locally Lipschitz in general, even smooth for convex h

Douglas-Rachford Envelope

"Integrating" the fixed-point residual

$$\textbf{minimize}\ \ \varphi = \varphi_1 + \varphi_2 \qquad \begin{cases} u = \mathbf{prox}_{\gamma\varphi_1}(s) \\ v = \mathbf{prox}_{\gamma\varphi_2}(2u - s) \end{cases}$$

convex nonsmooth case ***with Douglas-Rachford***

- stationary points characterized by $u - v = 0$
- **Douglas-Rachford envelope** discovered for convex problems[1]

$$\boxed{\varphi_\gamma^{\mathrm{DR}}(s) := \varphi_1^\gamma(s) - \gamma\|\nabla\varphi_1^\gamma(s)\|^2 + \varphi_2^\gamma(s - 2\gamma\nabla\varphi_1^\gamma(s))}$$

  real-valued function with gradient *proportional* to the DR-residual (for $\varphi_1 \in C^2$, $\gamma < 1/L_{\varphi_1}$)

$$\nabla\varphi_\gamma^{\mathrm{DR}}(s) = M_\gamma(s)(u - v) \qquad M_\gamma(s) = I - 2\gamma\nabla^2\varphi_1^\gamma(s) \succ 0$$

- used to devise accelerated DRS (ADMM via dual[2])

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

[2] Pejcic I. and Jones C. **Accelerated ADMM based on accelerated Douglas-Rachford splitting**. ECC 2016

Douglas-Rachford Envelope

"Integrating" the fixed-point residual

$$\boxed{\varphi_\gamma^{\mathrm{DR}}(s) := \varphi_1^\gamma(s) - \gamma\|\nabla\varphi_1^\gamma(s)\|^2 + \varphi_2^\gamma(s - 2\gamma\nabla\varphi_1^\gamma(s))}$$

If

- $\varphi_1 : \mathbf{dom}\,\varphi_1 \to \mathbb{R}$ has L_{φ_1}-Lipschitz gradient
- $\mathbf{dom}\,\varphi_1$ is affine and contains $\mathbf{dom}\,\varphi_2$
- **no convexity assumptions!**

then for $\gamma < 1/L_{\varphi_1}$,

- $\mathbf{inf}\,\varphi = \mathbf{inf}\,\varphi_\gamma^{\mathrm{DR}}$
- $s \in \mathbf{argmin}\,\varphi_\gamma^{\mathrm{DR}} \iff \mathbf{prox}_{\gamma\varphi_1}(s) \in \mathbf{argmin}\,\varphi$

Minimizing φ is equivalent to minimizing $\varphi_\gamma^{\mathrm{DR}}$

Douglas-Rachford Envelope

"Integrating" the fixed-point residual

$$\boxed{\varphi_\gamma^{\mathrm{DR}}(s) := \varphi_1^\gamma(s) - \gamma\|\nabla\varphi_1^\gamma(s)\|^2 + \varphi_2^\gamma(s - 2\gamma\nabla\varphi_1^\gamma(s))}$$

If

- $\varphi_1 : \mathbf{dom}\,\varphi_1 \to \mathbb{R}$ has L_{φ_1}-Lipschitz gradient
- $\mathbf{dom}\,\varphi_1$ is affine and contains $\mathbf{dom}\,\varphi_2$
- **no convexity assumptions!**

then for $\gamma < 1/L_{\varphi_1}$,

- $\mathbf{inf}\,\varphi = \mathbf{inf}\,\varphi_\gamma^{\mathrm{DR}}$
- $s \in \mathbf{argmin}\,\varphi_\gamma^{\mathrm{DR}} \iff \mathbf{prox}_{\gamma\varphi_1}(s) \in \mathbf{argmin}\,\varphi$

Minimizing φ is equivalent to minimizing $\varphi_\gamma^{\mathrm{DR}}$

Notation: for $x \in \mathbf{dom}\,\varphi_1$, $\tilde{\nabla}\varphi_1(x)$ is the unique in $\mathbf{dom}\,\varphi_1^{\parallel}$ s.t.

$$\varphi_1(y) = \varphi_1(x) + \langle\tilde{\nabla}\varphi_1(x), y - x\rangle + o(\|y - x\|^2) \quad y \in \mathbf{dom}\,\varphi_1$$

Douglas-Rachford Envelope

DRE as an Augmented Lagrangian

- alternative expression

$$\varphi_\gamma^{\mathrm{DR}}(s) = \inf_{w\in\mathbb{R}^n} \left\{\varphi_1(u) + \varphi_2(w) + \langle\tilde\nabla\varphi_1(u), w-u\rangle + \tfrac{1}{2\gamma}\|w-u\|^2\right\}$$

  where $u = \mathbf{prox}_{\gamma\varphi_1}(s)$.
- minimum attained at $v \in \mathbf{prox}_{\gamma g}(2u - s)$:

$$\varphi_\gamma^{\mathrm{DR}}(s) = \varphi_1(u) + \varphi_2(v) + \langle\tilde\nabla\varphi_1(u), v-u\rangle + \tfrac{1}{2\gamma}\|v-u\|^2$$

- apparently,

$$\varphi_\gamma^{\mathrm{DR}}(s) = \mathcal{L}_\gamma(u,v,y) \quad \text{for} \quad y = -\tilde\nabla\varphi_1(u)$$

  where $\mathcal{L}_\gamma$ is the **augmented Lagrangian** relative to

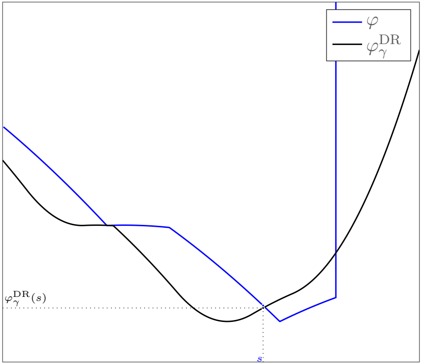
$$\textbf{minimize} \;\; \varphi_1(x) + \varphi_2(z) \quad \textbf{subject to} \quad x = z$$

Douglas-Rachford Envelope

A new tool for analyzing convergence

Key property: **sufficient decrease** after one DRS iteration

$$\begin{cases} u = \mathbf{prox}_{\gamma\varphi_1}(s) \\ v \in \mathbf{prox}_{\gamma\varphi_2}(2u - s) \\ s^+ = s + \lambda(v - u) \end{cases} \qquad \boxed{\varphi_\gamma^{\mathrm{DR}}(s^+) \leq \varphi_\gamma^{\mathrm{DR}}(s) - c\|u - v\|^2 \quad \exists c = c(\gamma, \lambda) > 0}$$



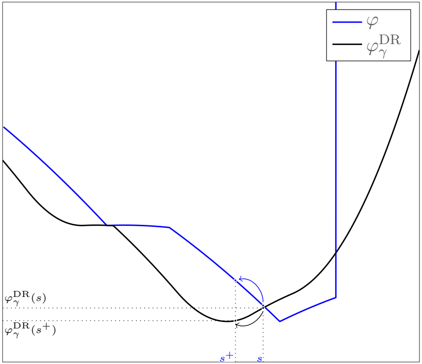
Douglas-Rachford Envelope

A new tool for analyzing convergence

Key property: **sufficient decrease** after one DRS iteration

$$\begin{cases} u &= \mathbf{prox}_{\gamma\varphi_1}(s) \\ v &\in \mathbf{prox}_{\gamma\varphi_2}(2u-s) \\ s^+ &= s + \lambda(v-u) \end{cases}$$

$$\boxed{\varphi_\gamma^{\mathrm{DR}}(s^+) \leq \varphi_\gamma^{\mathrm{DR}}(s) - c\|u-v\|^2 \quad \exists c = c(\gamma,\lambda) > 0}$$



Douglas-Rachford Envelope

A new tool for analyzing convergence

Key property: **sufficient decrease** after one DRS iteration

$$\begin{cases} u &= \mathbf{prox}_{\gamma\varphi_1}(s) \\ v &\in \mathbf{prox}_{\gamma\varphi_2}(2u - s) \\ s^+ &= s + \lambda(v - u) \end{cases} \quad \boxed{\varphi_\gamma^{\mathrm{DR}}(s^+) \leq \varphi_\gamma^{\mathrm{DR}}(s) - c\|u - v\|^2 \quad \exists c = c(\gamma, \lambda) > 0}$$

- nonconvex DRS studied only recently, using the DRE
- only $\lambda = 1$ (plain DRS) and $\lambda = 2$ (PRS) analyzed
- bounds on γ based on enforcing $c(\gamma, \lambda) > 0$

In this work,

- study extended to $\lambda \neq 1, 2$
- much less conservative upper bound on γ

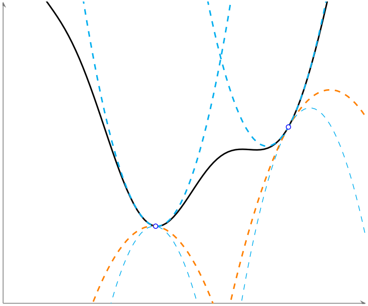
Douglas-Rachford Envelope

A new tool for analyzing convergence

Nicer results if we can improve the quadratic lower bound

$$\frac{\sigma_h}{2}\|x-y\|^2 \leq h(y) - h(x) - \langle \tilde{\nabla} h(x), y-x\rangle \leq \frac{L_h}{2}\|x-y\|^2$$

for some $\sigma_h \in [-L_h, L_h]$.



$h(x) = 4x^2 + \mathbf{sin}(5x)$ has
$L_h = 33$
$\sigma_h = -17$

key inequality: if $\sigma_h \leq 0$, for any $L \geq L_h$ with $L + \sigma_h > 0$

$$h(y) \geq h(x) + \langle \tilde{\nabla} h(x), y-x\rangle + \frac{\sigma_h L}{2(L+\sigma_h)}\|y-x\|^2 + \frac{1}{2(L+\sigma_h)}\|\tilde{\nabla} h(y) - \tilde{\nabla} h(x)\|^2$$

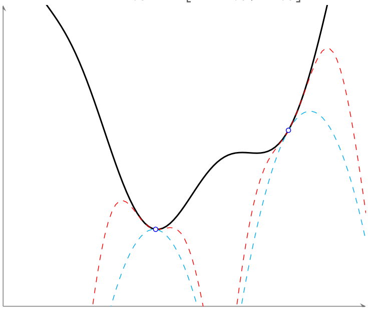
Douglas-Rachford Envelope

A new tool for analyzing convergence

Nicer results if we can improve the quadratic lower bound

$$\frac{\sigma_h}{2}\|x-y\|^2 \le h(y) - h(x) - \langle \tilde{\nabla} h(x), y - x\rangle \le \frac{L_h}{2}\|x-y\|^2$$

for some $\sigma_h \in [-L_h, L_h]$.



$h(x) = 4x^2 + \mathbf{sin}(5x)$ has
$L_h = 33$
$\sigma_h = -17$

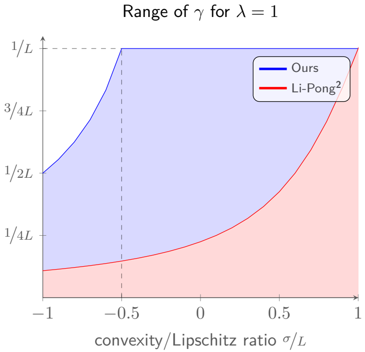
key inequality: if $\sigma_h \le 0$, for any $L \ge L_h$ with $L + \sigma_h > 0$

$$h(y) \ge h(x) + \langle \tilde{\nabla} h(x), y - x\rangle + \frac{\sigma_h L}{2(L+\sigma_h)}\|y-x\|^2 + \frac{1}{2(L+\sigma_h)}\|\tilde{\nabla} h(y) - \tilde{\nabla} h(x)\|^2$$

Douglas-Rachford Envelope

A new tool for analyzing convergence

- $\lambda = 1$: nonconvex DRS first studied by Li & Pong,[1] using the DRE



new bound much less conservative

- φ_2 plays **no role**
- $\sigma_{\varphi_1}/L_{\varphi_1} \in [-1, 1]$
- larger $\sigma_{\varphi_1}/L_{\varphi_1} \implies$ larger bound on γ
- φ_1 "mildly nonconvex":
  any $\gamma < 1/L_{\varphi_1}$ gives decrease
- can always use $\gamma < 1/(2L_{\varphi_1})$

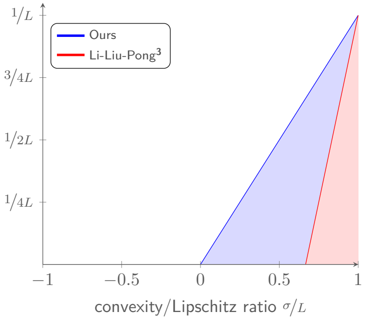
[1] Li G. and Pong T.K. **Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems**. Mathematical Programming 2016

Douglas-Rachford Envelope

A new tool for analyzing convergence

- $\lambda = 1$: nonconvex DRS first studied by Li & Pong,[1] using the DRE
- $\lambda = 2$: nonconvex PRS studied by Li, Liu & Pong,[2] using the DRE
  new bound much less conservative

Range of γ for $\lambda = 2$ (PRS)



- φ_2 plays **no role**
- can even choose $2 < \lambda < 4$!

[1] Li G. and Pong T.K. **Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems**. Mathematical Programming 2016

[2] Li G., Liu T. and Pong T.K. **Peaceman–Rachford splitting for a class of nonconvex optimization problems**. Computational Optimization and Applications 2017

Douglas-Rachford Envelope

Regularity

- if φ_1 is C^2 and φ_2 is convex, the DRE is C^1
- for nonconvex φ_1, φ_2, although not diff.ble, the DRE is locally Lipschitz

Furthermore, under mild conditions

- it is C^1 around minima
- and even twice diff.ble there!

The DRE leads to **novel fast DRS-based algorithms** for minimizing φ (this talk)

Douglas-Rachford Line-search Algorithm

A Lyapunov function for globalizing convergence

Choose λ, γ ensuring sufficient decrease, $0 < \sigma < c(\gamma, \lambda)$, and $s \in \mathbb{R}^n$

1: $u \leftarrow \mathbf{prox}_{\gamma\varphi_1}(s)$
2: $v \leftarrow \mathbf{prox}_{\gamma\varphi_2}(2u - s)$
3: Compute a direction $d \in \mathbf{dom}\,\varphi_1^{\|}$ and set $\tau \leftarrow 1$
4: $s^+ \leftarrow s + (1-\tau)\lambda(v-u) + \tau d$
5: **if** $\varphi_\gamma^{\mathrm{DR}}(s^+) \leq \varphi_\gamma^{\mathrm{DR}}(s) - \sigma\|v-u\|^2$ **then**
6: set $s \leftarrow s^+$ and go to step 1.
else
7: set $\tau \leftarrow \tau/2$ and go to step 4.

- step taken along convex combination of DR and custom directions
- continuity of φ_γ + suff. decrease of DR direction
  $\Rightarrow$ condition at step 5 passed for τ small enough

The DRE

- **globalizes convergence** for any d
- **favors fast directions**, thanks to local properties of the DRE

Douglas-Rachford Line-search Algorithm

A Lyapunov function for globalizing convergence

Convergence result

Suppose that the standing assumptions hold and γ, λ are s.t. $c(\gamma,\lambda) > 0$.

1. the sequence of DR-residuals $(\|v^k - u^k\|)_{k\in\mathbb{N}}$ is square-summable.
2. all cluster points of $(u^k)_{k\in\mathbb{N}}$, $(v^k)_{k\in\mathbb{N}}$ are stationary for φ

- result holds for *any* sequence of directions in $\mathbf{dom}\, f^{\|}$
- under extra mild assumptions (coercivity, KL property): convergence of entire sequence, linear convergence

Douglas-Rachford Line-search Algorithm

Examples of directions

$$s^+ = s + \underbrace{(1-\tau)\overbrace{\lambda(v-u)}^{\text{DR-residual}} + \tau \overbrace{d}^{\text{custom } d}}_{\text{convex combination}}$$

Key idea: d selected as *fast* direction for nonlinear equation

$$R_\gamma(s) = 0$$

where $R_\gamma(s) = v - u$ is the DR-residual.

- If d are "fast", eventually $\tau = 1$ when close to solution
- and algorithm reduces to the "fast" scheme $s^+ = s + d$.

Douglas-Rachford Line-search Algorithm

Examples of directions

$$s^+ = s + \underbrace{(1-\tau)\overbrace{\lambda(v-u)}^{\text{DR-residual}} + \tau \overbrace{d}^{\text{custom } d}}_{\text{convex combination}}$$

Possible choices:

- Newton-type directions

$$d = -HR_\gamma(s), \qquad H \text{ is } n \times n \text{ matrix}$$

  - quasi-Newton (BFGS, Broyden...): only linear algebra
  - limited-memory quasi-Newton (L-BFGS): **only scalar products**
- Nesterov-type acceleration (next slide): **negligible operations**

All such directions are **feasible**: $d \in \mathbf{dom}\, \varphi_1^{\|}$

Douglas-Rachford Line-search Algorithm

Examples of directions

$$s^+ = s + \underbrace{(1-\tau)\overbrace{\lambda(v-u)}^{\text{DR-residual}} + \tau \overbrace{d}^{\text{custom } d}}_{\text{convex combination}}$$

Nesterov-like acceleration:

$$d = \lambda(v-u) + \overbrace{\tfrac{k-1}{k+2}(w^+ - w)}^{\text{momentum term}} \quad \text{where } w^+ = s + \lambda(v-u)$$

- whenever $\tau = 1$ is accepted, iteration becomes Accelerated DRS[1]
- φ_1 convex quadratic, φ_2 convex $\implies$ $O(1/k^2)$ rate
- v and/or φ_2 nonconvex: no guarantee of acceleration
- but algorithm is **globally convergent**
- in practice, when φ_1 is not concave it seems we have acceleration

[1] Patrinos P., Stella L. and Bemporad A. **Douglas-Rachford splitting: Complexity estimates and accelerated variants**. 53rd IEEE CDC, 2014.

Douglas-Rachford Line-search Algorithm

Superlinear convergence

Superlinear convergence result

Suppose that the basic assumptions hold and that

1. $(u^k)_{k\in\mathbb{N}}$ converges to a strong local minimum $u^\star$ of φ
2. φ_1 is C^2 around $u^\star$
3. φ_2 is prox-regular at $u^\star$ for $-\tilde{\nabla}\varphi_1(u^\star)$,
   and has generalized quadratic second-order epiderivative.

If the directions satisfy the Dennis-Moré condition (*e.g.*, Broyden)

$$\lim_{k\to\infty} \frac{v^k - u^k + JR_\gamma(s_\star)d^k}{\|d^k\|} = 0,$$

$s_\star$ being the limit point of s^k, then

- unit stepsize $\tau_k = 1$ is eventually always accepted, and
- the sequence $(s^k)_{k\in\mathbb{N}}$ converges **superlinearly** to $s^\star$.

Separable problems

- ADMM first interpreted DRS **on the dual** (Eckstein & Bertsekas)
- **No convexity**: we interpret ADMM as DRS **on the primal**

$$\begin{array}{rl} \textbf{minimize} & f(x) + g(z) \\ \textbf{subject to} & Ax + Bz = b \end{array}$$

- rewrite as

$$\begin{array}{rl} \underset{x,z,s}{\textbf{minimize}} & f(x) + g(z) \\ \textbf{subject to} & Ax = b - s, Bz = s \end{array}$$

- minimizing first with respect to x, z

$$\underset{s}{\textbf{minimize}} \ (Af)(b - s) + (Bg)(s)$$

where

$$(Lh)(s) = \inf_x \{h(x) \mid Lx = s\}$$

is the **image function**

ADMM & DRS

separable problem

$$\begin{aligned} \textbf{minimize} \quad & f(x) + g(z) \\ \textbf{subject to} \quad & Ax + Bz = b \end{aligned}$$

image formulation

$$\underset{s}{\textbf{minimize}} \quad \underbrace{(Bg)(s)}_{\varphi_1(s)} + \underbrace{(Af)(b-s)}_{\varphi_2(s)}$$

- apply DRS to equivalent image formulation

$$\text{(update order shifted)} \quad \begin{cases} v^+ \in \mathbf{prox}_{\gamma\varphi_2}(2u - s) \\ s^+ = s + v^+ - u \\ u^+ = \mathbf{prox}_{\gamma\varphi_1}(s^+) \end{cases}$$

- use proximal calculus rules

$$v^+ = b - Ax^+ \quad \text{where} \quad x^+ \in \mathbf{argmin}_x \left\{ f(x) + \tfrac{1}{2\gamma}\|Ax - b + s\|^2 \right\}$$

$$u^+ = Bz^+ \quad \text{where} \quad z^+ \in \mathbf{argmin}_z \left\{ g(z) + \tfrac{1}{2\gamma}\|Bz - s\|^2 \right\}$$

- introduce

$$y = -\tilde{\nabla}\varphi_1(v) = \gamma^{-1}(Bz - s)$$

and eliminate s...

ADMM & DRS

separable problem

$$\begin{array}{ll} \textbf{minimize} & f(x) + g(z) \\ \textbf{subject to} & Ax + Bz = b \end{array}$$

image formulation

$$\underset{s}{\textbf{minimize}} \quad \underbrace{(Bg)(s)}_{\varphi_1(s)} + \underbrace{(Af)(b-s)}_{\varphi_2(s)}$$

- ... to arrive at ADMM

$$\begin{cases} x^+ = \textbf{argmin}_x \mathcal{L}_\beta(x, z, y) \\ z^+ = \textbf{argmin}_z \mathcal{L}_\beta(x^+, z, y) \\ y^+ = y + \beta(Ax^+ + Bz^+ - b) \end{cases}$$

- where $\beta = 1/\gamma$ and

$$\mathcal{L}_\beta(x, z, y) = f(x) + g(z) + \langle y, Ax + Bz - b \rangle + \frac{\beta}{2} \|Ax + Bz - b\|^2$$

is the augmented Lagrangian

ADMM & DRS

separable problem

$$\begin{aligned} \textbf{minimize} \quad & f(x) + g(z) \\ \textbf{subject to} \quad & Ax + Bz = b \end{aligned}$$

image formulation

$$\underset{s}{\textbf{minimize}} \quad \underbrace{(Bg)(s)}_{\varphi_1(s)} + \underbrace{(Af)(b - s)}_{\varphi_2(s)}$$

- equivalence between DRE and augmented Lagrangian

$$\varphi^{\mathrm{DR}}_{1/\beta}(s) = \mathcal{L}_\beta(x, z, y) \quad \text{for} \begin{cases} x \in \textbf{argmin}_x \left\{ f(x) + \frac{\beta}{2}\|Ax + s - b\|^2 \right\} \\ y = \beta(Bz - s) \\ z \in \textbf{argmin}_z \, \mathcal{L}_\beta(x, z, y) \end{cases}$$

- sufficient decrease on DRE becomes (for simplicity, $\lambda = 1$)

$$\mathcal{L}_\beta(x^+, z^+, y^+) \le \mathcal{L}_\beta(x, z, y) - c\|Ax + Bz - b\|^2$$

$$\text{for ADMM updates} \begin{cases} x^+ = \textbf{argmin}_x \mathcal{L}_\beta(x, z, y) \\ z^+ = \textbf{argmin}_z \mathcal{L}_\beta(x^+, z, y) \\ y^+ = y + \beta(Ax^+ + Bz^+ - b) \end{cases}$$

ADMM-LS

Choose β large enough ensuring sufficient decrease, $0 < \sigma < c(\beta)$

1: Compute a direction $d \in B\,\mathbf{dom}\,g^{\|}$ and set $\tau \leftarrow 1$
2: $y^{+/2} \leftarrow y - \beta\tau(Ax + Bz - b + d)$
3: $z^{+} \leftarrow \mathbf{argmin}_z \mathcal{L}_\beta(x, z, y^{1/2})$
4: $y^{+} \leftarrow y^{+/2} + \beta(Ax + Bz^{+} - b)$
5: $x^{+} \leftarrow \mathbf{argmin}_x \mathcal{L}_\beta(x, z^{+}, y^{+})$
6: **if** $\mathcal{L}_\beta(x^{+}, z^{+}, y^{+}) \leq \mathcal{L}_\beta(x, z, y) - \sigma\|Ax + Bz - b\|^2$ **then**
7: set $x \leftarrow x^{+}, z \leftarrow z^{+}, y \leftarrow y^{+}$ and go to step 1.
else
8: set $\tau \leftarrow \tau/2$ and go to step 2.

- algorithm is DRLS applied to image formulation
- $\tau = 0 \implies$ only steps 3,4,5 needed: algorithm equivalent to ADMM (after update order shift)

ADMM

Convergence result

Suppose that

1. $B\,\mathbf{dom}\,g \supseteq b - A\,\mathbf{dom}\,f$
2. (Bg) is Lipschitz smooth on $B\,\mathbf{dom}\,g$ (see next slide)
3. ADMM subproblems level bounded wrt minimization variable
4. β is s.t. $c(\beta) > 0$ (always exists)

Then

1. square-summable ADMM-residuals $(\|Ax^k + Bz^k - b\|)_{k\in\mathbb{N}}$
2. all cluster points of $(x^k, z^k, y^k)_{k\in\mathbb{N}}$ satisfy KKT

$$0 \in \partial f(x^\star) + A^\top y^\star,\ 0 \in \partial f(z^\star) + B^\top y^\star,\ Ax^\star + Bz^\star = b$$

- much less restrictive than existing results (see next slides)

ADMM

Sufficient conditions for

$$\varphi_1(s) = \inf_z \{g(z) \mid Bz = s\}$$

to be Lipschitz smooth on its domain: g Lipschitz smooth and

- B full column rank: choose

$$\beta > 2L_{\varphi_1} \qquad \text{where} \qquad L_{\varphi_1} = \frac{L_g}{\lambda_{\min}(B^\top B)}$$

- g convex, B full row rank: choose

$$\beta > L_{\varphi_1} \qquad \text{where} \qquad L_{\varphi_1} = \frac{L_g}{\lambda_{\min}(BB^\top)}$$

- $z(s) = \mathbf{argmin}_z \{g(z) \mid Bz = s\}$ is Lipschitz on $B\,\mathbf{dom}\,g$[1]

[1] standing assumption in Wang, Yin, Zeng (2015), for both $z(s)$ and $x(s) = \mathbf{argmin}_x \{f(x) \mid Ax = b - s\}$

ADMM

Sufficient conditions for

$$\varphi_1(s) = \inf_z \{g(z) \mid Bz = s\}$$

to be Lipschitz smooth on its domain:
alternatively,

- g "B-smooth":

$$|\langle \tilde{\nabla} g(x) - \tilde{\nabla} g(y), x - y \rangle| \leq L_{g,B} \|B(x-y)\|^2$$

only for x, y **such that** $\tilde{\nabla} g(x), \tilde{\nabla} g(y) \in \mathbf{range}\, B^\top$

In any case, L_{φ_1} can be retrieved adaptively!

ADMM

Comparisons (bringing all under the same framework. . .)

Ours	Hong et al.[2]	Li and Pong[4]	Wang et al.[5]	Gonçalves et al.[6]
	f cvx or smooth			
g "B-smooth" $\mathbf{dom}\, g$ affine	∇g Lipsch.	∇g Lipsch. $g \in C^2$	∇g Lipsch.	$\Pi_{B^\top} \nabla g$ Lipsch. g lower-C^2
$x(s)$ loc. bound.	$A = I$	A full row rank	$x(s)$ Lipsch.	
$\mathcal{L}_\beta$ level bound. in z	B full col. rank	$B = I$	$z(s)$ Lipsch.	B full col. rank

$$x(s) = \mathbf{argmin}_x \{f(x) \mid Ax = s\} \quad \text{and} \quad z(s) = \mathbf{argmin}_z \{g(z) \mid Bz = s\}$$

Notice that

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

---

[3] M. Hong, Z. Luo and M. Razaviyayn **Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems** SIAM Opt. 26(1) 2016

[4] G. Li and T.K. Pong **Global Convergence of Splitting Methods for Nonconvex Composite Optimization**. SIAM Opt. 25(4) 2015

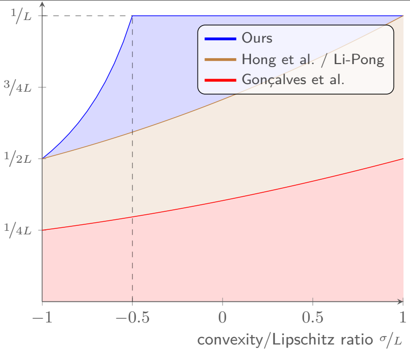
[5] Y. Wang, W. Yin and J. Zeng **Global Convergence of ADMM in Nonconvex Nonsmooth Optimization** arXiv:1511.06324 2015

[6] 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 2017

ADMM

Comparisons (bringing all under the same framework. . .)

Ours	Hong et al.[2]	Li and Pong[4]	Wang et al.[5]	Gonçalves et al.[6]



upper bound for $1/\beta$ (the higher the better)

- the nonsmooth function plays no role
- L is the Lipschitz constant in the DRS-equivalent problem $(L = L_{(Bg)})$
- ours is the **same bound as** $\gamma = 1/\beta$ **in DRS**

[3] M. Hong, Z. Luo and M. Razaviyayn **Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems** SIAM Opt. 26(1) 2016

[4] G. Li and T.K. Pong **Global Convergence of Splitting Methods for Nonconvex Composite Optimization**. SIAM Opt. 25(4) 2015

[5] Y. Wang, W. Yin and J. Zeng **Global Convergence of ADMM in Nonconvex Nonsmooth Optimization** arXiv:1511.06324 2015

[6] 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 2017

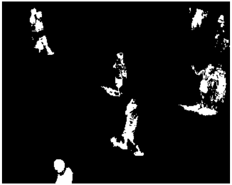
Matrix decomposition

Split a signal S into a **sparse** X and **low-rank** Y:

$$\begin{aligned} \textbf{minimize} \quad & \tfrac{1}{2}\|X + Y - S\|^2 + \lambda \|X\|_0 \\ \textbf{subject to} \quad & \textbf{rank}(Y) \le r \end{aligned}$$

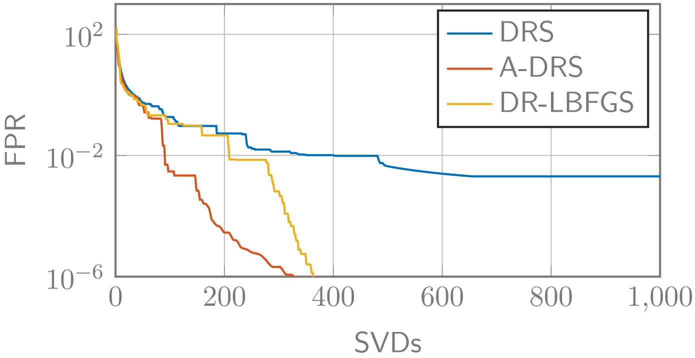
Example: separate foreground objects from background in a sequence of video frames

- S is a matrix where each column is a video frame
- the background is mainly constant over time $\Rightarrow$ Y **low rank**
- foreground moving objects $\Rightarrow$ X **sparse**



Examples

- S contains 100 frames from the *ShoppingMall* dataset
- $r = 1, \lambda = 5 \cdot 10^{-3}$, 8192000 variables



Cost achieved:
DRS $= 4.1330 \cdot 10^3$, A-DRS $= 4.1118 \cdot 10^3$, **DR-LBFGS =** $4.0556 \cdot 10^3$

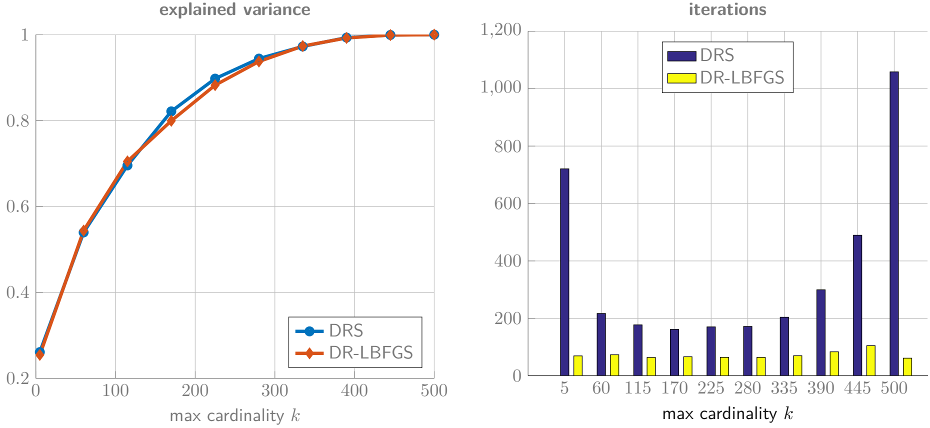
Sparse PCA

$$\begin{aligned} \textbf{maximize} \quad & \langle x, \Sigma x \rangle \\ \textbf{subject to} \quad & \|x\|_2 = 1, \qquad \|x\|_0 \le k \end{aligned}$$

- $\Sigma = A^\top A$ 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
- DRLS is readily applicable
- $f(x) = -\langle x, \Sigma x \rangle$ nonconvex (concave)
- g models intersection of unit ℓ_2 sphere with ℓ_0 ball (nonconvex)

Sparse PCA example

SPCA path



Consensus SPCA

centralized SPCA formulation

$$\begin{array}{rl} \textbf{minimize} & -\|Az\|_2^2 \\ \textbf{subject to} & \|z\|_2 = 1, \qquad \|z\|_0 \le k \end{array}$$

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

$$\begin{array}{rl} \textbf{minimize} & \sum_{i=1}^{N} \overbrace{-\|A_i x_i\|_2^2}^{f_i(x_i)} + g(z) \\ \textbf{subject to} & x_i = z \end{array}$$

the problem is in ADMM form

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

Consensus SPCA: example

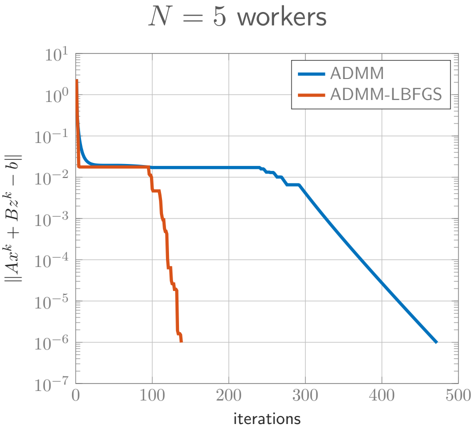
- each $A \in \mathbb{R}^{m \times n}$ sparse, randomly generated
- $n = 100,000$ features, $m = 50,000$ data points
- rows are split into N subsets

Computing $\mathbf{prox}$ of $-\|A_i x_i\|^2$ requires factoring (once)

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

- Cholesky factorization (*e.g.*, using `ldlchol`) $O(m_i^3)$
- $N = 50$ workers $\Rightarrow m_i = 1,000$, ≈ 0.03 seconds
- $N = 5$ workers $\Rightarrow m_i = 10,000$, ≈ 7 seconds
- $N = 1$ workers $\Rightarrow m_1 = m = 50,000$, > 1 hour

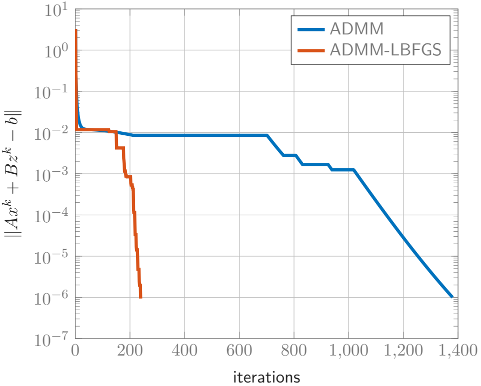
Consensus SPCA



	final $\langle z, \Sigma z \rangle$	iterations
ADMM	183	472
ADMM-LBFGS	185	138

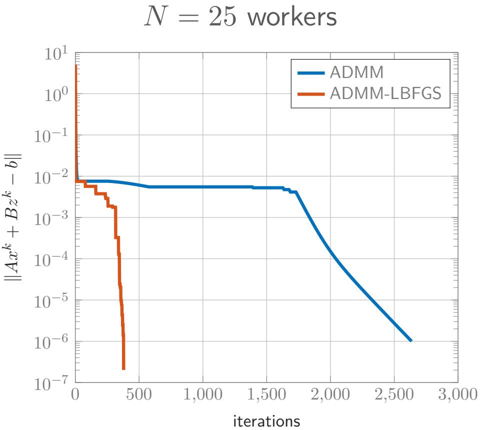
Consensus SPCA

$N = 10$ workers



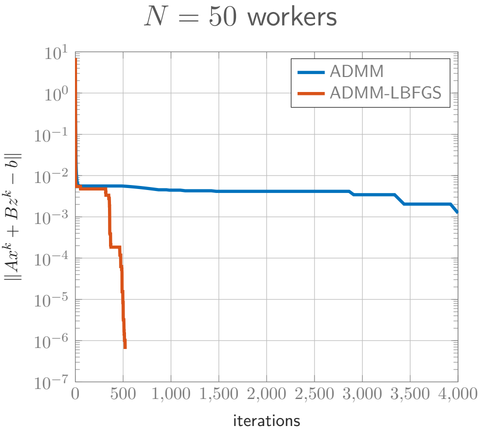
	final $\langle z, \Sigma z\rangle$	iterations
ADMM	181	1380
ADMM-LBFGS	187	239

Consensus SPCA



	final $\langle z, \Sigma z \rangle$	iterations
ADMM	169	2636
ADMM-LBFGS	180	379

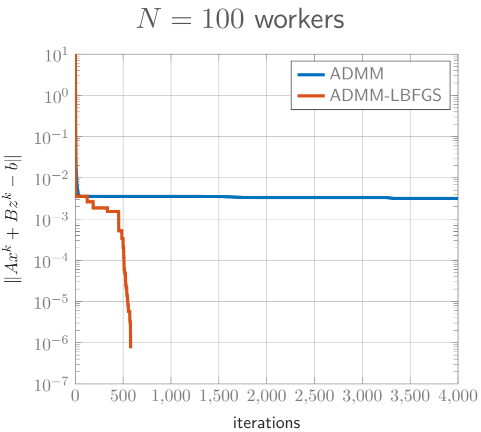
Consensus SPCA



	final $\langle z, \Sigma z \rangle$	iterations
ADMM	168	4000*
ADMM-LBFGS	175	521

*reached maximum number of iterations

Consensus SPCA



	final $\langle z, \Sigma z \rangle$	iterations
ADMM	95	4000*
ADMM-LBFGS	175	578

*reached maximum number of iterations

H.H. Bauschke and P.L. Combettes.
Convex Analysis and Monotone Operator Theory in Hilbert Spaces.
CMS Books in Mathematics. Springer, 2011.

M. L. N. Goncalves, J. G. Melo, and R. D. C. Monteiro.
Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems.
ArXiv e-prints, February 2017.

Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn.
Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems.
SIAM Journal on Optimization, 26(1):337–364, 2016.

G. Li, T. Liu, and T.K. Pong.
Peaceman–Rachford splitting for a class of nonconvex optimization problems.
Computational Optimization and Applications, pages 1–30, 2017.

G. Li and T.K. Pong.
Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems.
Mathematical Programming, 159(1):371–401, 2016.

Guoyin Li and Ting Kei Pong.
Global convergence of splitting methods for nonconvex composite optimization.
SIAM Journal on Optimization, 25(4):2434–2460, 2015.

P. Patrinos, L. Stella, and A. Bemporad.
Douglas–Rachford splitting: Complexity estimates and accelerated variants.
In *53rd IEEE Conference on Decision and Control*, pages 4234–4239, Dec 2014.

A. Themelis, L. Stella, and P. Patrinos.
Douglas–Rachford splitting and ADMM for nonconvex optimization: new convergence results and accelerated versions.
arXiv, 2017.

Y. Wang, W. Yin, and J. Zeng.
Global convergence of ADMM in nonconvex nonsmooth optimization.
ArXiv e-prints, November 2015.