

## Douglas-Rachford splitting and ADMM for nonconvex optimization: Accelerated and Newton-type linesearch algorithms

Themelis, Andreas

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

Stella, Lorenzo

Amazon Research

Patrinos, Panagiotis

Department of Electrical Engineering (ESAT-STADIUS), KU Leuven

<https://hdl.handle.net/2324/4785265>

---

出版情報 : Computational Optimization and Applications. 82 (2), pp.395-440, 2022-05-11.  
Springer Nature

バージョン :

権利関係 :



# Douglas-Rachford splitting and ADMM for nonconvex optimization

## Accelerated and Newton-type linesearch algorithms

Andreas Themelis · Lorenzo Stella · Panagiotis Patrinos

Received: date / Accepted: date

**Abstract** Although the performance of popular optimization algorithms such as the Douglas-Rachford splitting (DRS) and the ADMM is satisfactory in small and well-scaled problems, ill conditioning and problem size pose a severe obstacle to their reliable employment. Expanding on recent convergence results for DRS and ADMM applied to nonconvex problems, we propose two linesearch algorithms to enhance and robustify these methods by means of quasi-Newton directions. The proposed algorithms are suited for nonconvex problems, require the same black-box oracle of DRS and ADMM, and maintain their (subsequential) convergence properties. Numerical evidence shows that the employment of L-BFGS in the proposed framework greatly improves convergence of DRS and ADMM, making them robust to ill conditioning. Under regularity and nondegeneracy assumptions at the limit point, superlinear convergence is shown when quasi-Newton Broyden directions are adopted.

**Keywords** Nonsmooth nonconvex optimization, Douglas-Rachford splitting, ADMM, quasi-Newton methods

---

A. Themelis is supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI grant JP21K17710.

L. Stella's work was done prior to joining Amazon.

P. Patrinos is supported by the Research Foundation Flanders (FWO) research projects G0A0920N, G086518N, G086318N, and G081222N; Research Council KU Leuven C1 project No. C14/18/068; Fonds de la Recherche Scientifique – FNRS and the Fonds Wetenschappelijk Onderzoek – Vlaanderen under EOS project 30468160 (SeLMA); European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 953348.

---

A. Themelis

Kyushu University, Faculty of Information Science and Electrical Engineering (ISEE), 744 Motooka Nishiku, 819-0395 Fukuoka, Japan. Tel.: +81/092-802-3718. E-mail: andreas.themelis@ees.kyushu-u.ac.jp

L. Stella

Amazon Research, Charlottenstrasse 4, 10969 Berlin, Germany. E-mail: lorenzostella@gmail.com

P. Patrinos

KU Leuven, Department of Electrical Engineering (ESAT-STADIUS), Kasteelpark Arenberg 10, 3001 Leuven, Belgium. Tel.: +32/016-374445. E-mail: panos.patrinos@esat.kuleuven.be

**Mathematics Subject Classification (2010)** 90C06, 90C25, 90C26, 49J52, 49J53.

## 1 Introduction

Due to their simplicity and versatility, the *Douglas-Rachford splitting* (DRS) and the *alternating direction method of multipliers* (ADMM) have gained much popularity in the last decades. Although originally designed for convex problems, their generalizations and extensions to nonconvex problems have recently attracted much attention, see e.g. [22, 5, 6, 23, 30, 28, 51] for DRS and [29, 24, 21, 27, 20, 53, 51] for ADMM. The former algorithm addresses the following composite minimization problems

$$\underset{s \in \mathbb{R}^p}{\text{minimize}} \varphi(s) \equiv \varphi_1(s) + \varphi_2(s), \quad (1.1)$$

for some  $\varphi_1, \varphi_2 : \mathbb{R}^p \rightarrow \overline{\mathbb{R}}$  ( $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$  denotes the extended-real line). Starting from some  $s \in \mathbb{R}^p$ , one iteration of DRS applied to (1.1) with *stepsize*  $\gamma > 0$  and *relaxation*  $\lambda > 0$  amounts to

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

Here,  $\text{prox}_h$  denotes the *proximal mapping* of function  $h$ ; cf. Section 1.5.

The ADMM addresses optimization problems that can be formulated as

$$\underset{(x,z) \in \mathbb{R}^m \times \mathbb{R}^n}{\text{minimize}} f(x) + g(z) \quad \text{subject to } Ax + Bz = b, \quad (1.2)$$

for some  $f : \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$ ,  $g : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ ,  $A \in \mathbb{R}^{p \times m}$ ,  $B \in \mathbb{R}^{p \times n}$  and  $b \in \mathbb{R}^p$ . Equivalently, problem (1.2) amounts to the minimization of  $\Phi : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  given by

$$\Phi(x, z) := \begin{cases} f(x) + g(z) & \text{if } Ax + Bz = b, \\ \infty & \text{otherwise.} \end{cases}$$

Starting from a triplet  $(x, y, z) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n$ , one iteration of ADMM applied to (1.2) with *penalty*  $\beta > 0$  and *relaxation*  $\lambda > 0$  amounts to

$$\begin{cases} y^{+1/2} = y - \beta(1 - \lambda)(Ax + Bz - b) \\ x^+ \in \arg \min \mathcal{L}_\beta(\cdot, y^{+1/2}, z) \\ y^+ = y^{+1/2} + \beta(Ax^+ + Bz - b) \\ z^+ \in \arg \min \mathcal{L}_\beta(x^+, \cdot, y^+), \end{cases} \quad (\text{ADMM})$$

where  $\mathcal{L}_\beta$  is the  $\beta$ -augmented Lagrangian of (1.2), namely

$$\mathcal{L}_\beta(x, z, y) := f(x) + g(z) + \langle y, Ax + Bz - b \rangle + \frac{\beta}{2} \|Ax + Bz - b\|^2. \quad (1.3)$$

Although apparently more general, ADMM is known to be equivalent to DRS applied to the Fenchel dual of (1.1) when the problem is convex; the identity of the

two algorithms has been recently shown to hold in general through a simple change of variable [55, 51] (see Fact 2.5 for the details).

In both algorithms, the relaxation  $\lambda$  serves as an averaging factor, and  $\lambda = 1$  corresponds to the standard form in which the methods are most known. In the particular case of ADMM, when  $\lambda = 1$  one has  $y^{+1/2} = y$ ; the intermediate variable  $y^{+1/2}$  can thus be removed from the formulation in this case, and one recovers the familiar 3-step update of the algorithm. The half-update  $y^{+1/2}$  for the general case is introduced for simplification purposes; other sources, including the pioneering work [14] (see Theorem 8 therein), avoid this by replacing  $y^{+1/2}$  with  $y$  and all occurrences of  $Ax^+$  in the  $z$ - and  $y$ -updates with  $\lambda Ax^+ - (1 - \lambda)(Bz - b)$ . Independently of  $\lambda$ , the core of the respective iterations can be summarized in the following familiar oracles:

$$\text{DRS}_\gamma(\tilde{s}) = \left\{ (u, v) \in \mathbb{R}^p \times \mathbb{R}^p \mid \begin{array}{l} u \in \text{prox}_{\gamma\varphi_1}(\tilde{s}) \\ v \in \text{prox}_{\gamma\varphi_2}(2u - \tilde{s}) \end{array} \right\} \quad (1.4)$$

and

$$\text{ADMM}_\beta(\tilde{y}, \tilde{z}) = \left\{ (x, y, z) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n \mid \begin{array}{l} x \in \arg \min \mathcal{L}_\beta(\cdot, \tilde{y}, \tilde{z}) \\ y = \tilde{y} + \beta(Ax + B\tilde{z} - b) \\ z \in \arg \min \mathcal{L}_\beta(x, y, \cdot) \end{array} \right\}. \quad (1.5)$$

One of the main advantages of DRS and ADMM lies in their “splitting” nature, in the sense that they exploit the additive structure of the respective problems (1.1) and (1.2) by performing operations involving only either one component. For this reason, these methods typically involve simple operations and are thus amenable to address large-scale problems. On the other hand, although every iteration is relatively cheap, whenever the problem is not well scaled convergence up to a satisfactory tolerance may require prohibitively many iterations.

## 1.1 Contributions

Aware of these pros and cons and in contrast to an *open-loop* preconditioning approach, in this paper we propose linesearch variants that allow us to integrate DRS and ADMM with fast update directions, stemming for instance from quasi-Newton schemes. To the best of our knowledge, the resulting algorithms are the first that (1) are compatible with *fully nonconvex* problems, (2) maintain the *same complexity* of the original DRS and ADMM iterations, and (3) preserve their *global (subsequential) convergence guarantees*. Moreover, under regularity and nondegeneracy assumptions at the limit point, they converge superlinearly when a modified Broyden’s scheme is used to compute the update directions. Extensive numerical simulations show that limited-memory quasi-Newton methods such as L-BFGS are also very effective in practice.

## 1.2 Paper organization

A preliminary discussion on the methodology is given in the following subsection, where the proposed linesearch Algorithms 1 and 2 are also presented. The section

will then conclude with a list of notational conventions adopted throughout the paper. [Section 2](#) recaps the key properties of the [DRS](#) and [ADMM](#) oracles, which constitute the building blocks of the proposed algorithms. Once these preliminaries are dealt with, [Section 3](#) offers a thorough overview of the proposed algorithms, together with a list of possible choices for the update directions that can considerably improve convergence speed. Parameter-free variants of the algorithms for out-of-the-box implementations are also provided in the concluding subsection. [Section 4](#) contains the convergence results of the two algorithms, with some auxiliary material deferred to [Appendix A](#). In [Section 5](#) we provide numerical evidence in support of the efficacy of the proposed algorithms with simulations on nonconvex sparse linear regression and sparse principal component analysis problems, and on a strongly convex model predictive control problem. [Section 6](#) concludes the paper.

### 1.3 Methodology overview and proposed linesearch algorithms

Although a complete and rigorous discussion will be given in the dedicated [Section 3](#) after the needed preliminaries have been dealt with, the overall methodology is quite

#### Algorithm 1. Linesearch DRS

REQUIRE  $s^0 \in \mathbb{R}^p$ ; tol.  $\varepsilon > 0$ ; relax.  $\lambda \in (0, 2)$ ; max # backtracks  $i_{\max} \leq \infty$

UNDER <a href="#">ASSUMPTION I</a> (set $\pi := 1$ ) stepsize $\gamma$ as in (2.6) constant $0 < c < C(\gamma L_{\varphi_1}, \lambda)$ as in (2.5)	UNDER <a href="#">ASSUMPTION I*</a> (set $\pi := -1$ ) stepsize $\gamma > 1/\mu_{\varphi_1}$ constant $0 < c < C(1/\gamma\mu_{\varphi_1}, \lambda)$ as in (2.5)
--	---

INITIALIZE  $k = 0$ ; compute  $(u^0, v^0) \in \text{DRS}_\gamma(s^0)$  and evaluate  $\varphi_\gamma^{\text{DR}}(s^0)$  as in (2.2)

1.1:  $r^k = u^k - v^k$ ; IF  $\frac{1}{\gamma}\|r^k\| \leq \varepsilon$  THEN RETURN  $(s^k, u^k, v^k)$ ; END IF

1.2: Set  $\bar{s}^{k+1} = s^k - \lambda r^k$  ▷Nominal DRS step

1.3: Select an update direction  $d^k \in \mathbb{R}^p$  and set  $\tau_k = 1$  and  $i_k = 0$

1.4: Define the candidate update as  $s^{k+1} = (1 - \tau_k)\bar{s}^{k+1} + \tau_k(s^k + d^k)$

1.5: Compute  $(u^{k+1}, v^{k+1}) \in \text{DRS}_\gamma(s^{k+1})$  and use it to evaluate  $\varphi_\gamma^{\text{DR}}(s^{k+1})$  as in (2.2)

1.6: IF  $\pi\varphi_\gamma^{\text{DR}}(s^{k+1}) \leq \pi\varphi_\gamma^{\text{DR}}(s^k) - \frac{c}{\gamma}\|r^k\|^2$  THEN ▷Update accepted  
      $k \leftarrow k + 1$  and go to [step 1.1](#)

1.7: ELSE IF  $i_k = i_{\max}$  THEN ▷Max #backtrackings: do nominal DRS step  
     Set  $s^{k+1} = \bar{s}^{k+1}$  and  $(u^{k+1}, v^{k+1}) \in \text{DRS}_\gamma(s^{k+1})$   
      $k \leftarrow k + 1$  and go to [step 1.1](#)

1.8: ELSE ▷Backtrack and retry  
      $\tau_k \leftarrow \tau_k/2$ ,  $i_k \leftarrow i_k + 1$ , and restart from [step 1.4](#)

**Algorithm 2.** Linesearch ADMM

REQUIRE  $(x^{-1}, y^{-1}, z^{-1}) \in \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^m$ ;  $\varepsilon > 0$ ;  $\lambda \in (0, 2)$ ;  $i_{\max} \leq \infty$

UNDER <a href="#">ASSUMPTION II</a> (set $\pi := 1$ ) penalty $\beta$ as in (2.11) constant $0 < c < C^{(L_{(A,F)}/\beta, \lambda)}$ as in (2.5)	UNDER <a href="#">ASSUMPTION II*</a> (set $\pi := -1$ ) penalty $\beta < \mu_f/\ A\ ^2$ constant $0 < c < C^{(\beta\ A\ ^2/\mu_f, \lambda)}$ as in (2.5)
--	--

INITIALIZE  $k = 0$ ;  $r^{-1} = Ax^{-1} + Bz^{-1} - b$ ;  $y^{-1/2} = y^{-1} - \beta(1 - \lambda)r^{-1}$ ;  
 $(x^0, y^0, z^0) \in \text{ADMM}_\beta(y^{-1/2}, z^{-1})$  and evaluate  $\mathcal{L}_\beta(x^0, y^0, z^0)$

---

2.1:  $r^k = Ax^k + Bz^k - b$ ; IF  $\beta\|r^k\| \leq \varepsilon$  THEN RETURN  $(x^k, y^k, z^k)$ ; END IF

2.2: Set  $\bar{y}^{k+1/2} = y^k - \beta(1 - \lambda)r^k$

2.3: Select an update direction  $d^k \in \mathbb{R}^p$  and set  $\tau_k = 1$  and  $i_k = 0$

2.4: Define the candidate update  $y^{k+1/2} = (1 - \tau_k)\bar{y}^{k+1/2} + \tau_k(y^k - \beta(r^k + d^k))$

2.5: Compute  $(x^{k+1}, y^{k+1}, z^{k+1}) \in \text{ADMM}_\beta(y^{k+1/2}, z^k)$  and evaluate  $\mathcal{L}_\beta(x^{k+1}, y^{k+1}, z^{k+1})$

2.6: IF  $\pi\mathcal{L}_\beta(x^{k+1}, z^{k+1}, y^{k+1}) \leq \pi\mathcal{L}_\beta(x^k, z^k, y^k) - \beta c\|r^k\|^2$  THEN ►Update accepted  
 $k \leftarrow k + 1$  and go to [step 2.1](#)

2.7: ELSE IF  $i_k = i_{\max}$  THEN ►Max #backtrackings: do nominal ADMM step  
 $(x^{k+1}, y^{k+1}, z^{k+1}) \in \text{ADMM}_\beta(\bar{y}^{k+1/2}, z^k)$   
 $k \leftarrow k + 1$  and go to [step 2.1](#)

2.8: ELSE ►Backtrack and retry  
 $\tau_k \leftarrow \tau_k/2$ ,  $i_k \leftarrow i_k + 1$ , and restart from [step 2.4](#)

simple and can be informally summarized in few sentences. The proposed algorithms leverage on favorable properties of the Douglas-Rachford envelope (DRE) [36, 51], a continuous, real-valued, exact penalty function for problem (1.1). Given a stepsize  $\gamma > 0$ , the DRE associated to problem (1.1) at a point  $s \in \mathbb{R}^p$  is given by

$$\varphi_\gamma^{\text{DR}}(s) := \varphi_1(u) + \varphi_2(v) + \frac{1}{\gamma}\langle s - u, v - u \rangle + \frac{1}{2\gamma}\|v - u\|^2,$$

where  $(u, v) \in \text{DRS}_\gamma(s)$  is the result of a [DRS](#)-update with stepsize  $\gamma$  at  $s$ . Under the assumptions on the problem dealt in this paper (see [Section 2](#) for the details), the right-hand side in the above equation is the same for any  $(u, v) \in \text{DRS}_\gamma(s)$ , making  $\varphi_\gamma^{\text{DR}}$  a well-defined function of variable  $s$ . In fact, the DRE turns out to be continuous and such that

$$\varphi_\gamma^{\text{DR}}(\bar{s}^+) \leq \varphi_\gamma^{\text{DR}}(s) - \frac{c}{\gamma}\|v - u\|^2 \quad (1.6)$$

holds for any  $u, v, \bar{s}^+$  resulting from a [DRS](#) update at  $s$ , provided that the stepsize  $\gamma$  and the relaxation  $\lambda$  are selected as instructed in [Algorithm 1](#)'s initialization. By replacing  $C$  with a smaller value  $c < C$ , thanks to the aforementioned continuity of  $\varphi_\gamma^{\text{DR}}$

not only will this inequality hold for  $\bar{s}^+$ , but also for all the points in a neighborhood. This enables the flexibility to choose an arbitrary update direction  $d$ , stemming for instance from a quasi-Newton method of choice, and backtrack the sought update  $s + d$  towards  $\bar{s}^+$  until it enters said neighborhood.

Skimming through the steps of [Algorithm 1](#), the decrease constant  $c < C$  is required at initialization time, and computed based on the problem assumptions. The backtracking instead occurs at [step 1.4](#), where by halvening  $\tau_k$  enough times the average between the standard DRS-update  $\bar{s}^{k+1}$  and the sought custom update  $s^k + d^k$  will be close enough to  $\bar{s}^{k+1}$ , causing the inequality in [step 1.6](#) to be satisfied. Surprisingly, the analysis of [Algorithm 1](#) will be enough to cover the ADMM counterpart of [Algorithm 2](#), for the steps in both methods will be shown to coincide.

#### 1.4 Connections with previous work

This work is based on the theoretical analysis of the original DRS and ADMM algorithms in the nonconvex setting developed in the paper [51] by some of the authors. In fact, part of the content originally appeared in a preprint (non peer-reviewed) version, and was then removed because of page limitation and lack of focus. The current version benefits from a thorough polishing and many additions, which include finite-termination and superlinear convergence analysis, strongly convex case through self-duality arguments, simplified assumptions, and numerical simulations. An overview of the results in (the latest, published, version of) [51] needed in this paper is given in [Section 2](#), which is dedicated to the properties of DRS and ADMM oracles and the relations existing among the two.

The proposed linesearch follows the same rationale of the predecessors PANOC [47] and NAMA [46] algorithms, which instead of DRS and ADMM oracles were based on proximal gradient and alternating minimization steps, respectively. Their rationale hinges on the same mechanisms explained in [Section 1.3](#), namely the identification of a continuous “envelope” serving as Lyapunov function for the nominal iterations, i.e., ensuring an inequality in the likes of (1.6), and a backtracking to bias the desired update close enough to the nominal algorithmic step in such a way to enforce a decrease condition on the envelope, as done here in [step 1.6](#). An initial attempt to encompass these methods under the same lens is given in the doctoral dissertation [48] with the *Continuous-Lyapunov Descent* framework (CLyD), which however is not based on peer-reviewed material and offers limited theoretical results.

Related, but substantially different, is the minFBE algorithm of [45], based on forward-backward iterations. Under a convexity assumption on the nonsmooth term, the nonsmooth minimization problem can be cast as the minimization of the continuously differentiable “forward-backward envelope” function (FBE), where classical smooth minimization techniques (with standard linesearch strategies) can directly be applied. Apart from the additional convexity assumption, not required in other methods, this early work suffers from a complication of the algorithmic oracle required for the evaluation of the gradient of the FBE, which entails Hessian evaluations of the smooth function (and, as a byproduct, an additional twice differentiability assumption).

### 1.5 Notation and known facts

We denote as  $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$  the extended-real line. With  $\text{id}$  we indicate the identity function  $x \mapsto x$  defined on a suitable space, and with  $I$  the identity matrix of suitable size. The distance of a point  $x \in \mathbb{R}^n$  to a nonempty set  $E \subseteq \mathbb{R}^n$  is given by  $\text{dist}(x, E) = \inf_{z \in E} \|z - x\|$ . The relative interior of  $E$ , denoted  $\text{relint } E$ , is the interior of  $E$  relative to the smallest affine space containing  $E$ .

For a sequence  $(x^k)_{k \in \mathbb{N}}$  we write  $(x^k)_{k \in \mathbb{N}} \subset E$  to indicate that  $x^k \in E$  for all  $k \in \mathbb{N}$ . We say that  $(x^k)_{k \in \mathbb{N}} \subset \mathbb{R}^n$  converges at  $R$ -linear rate (to a point  $x_\star$ ) if there exists  $c > 0$  and  $\rho \in (0, 1)$  such that  $\|x^k - x_\star\| \leq c\rho^k$  holds for every  $k$ , and at superlinear rate (to  $x_\star$ ) if either  $s^k = s_\star$  for some  $k$  or  $\frac{\|s^{k+1} - s_\star\|}{\|s^k - s_\star\|} \rightarrow 0$  as  $k \rightarrow \infty$ .

A function  $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  is *proper* if  $h \not\equiv \infty$ , in which case its *domain* is defined as the set  $\text{dom } h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$ , and is *lower semicontinuous* (*lsc*) if for any  $\bar{x} \in \mathbb{R}^n$  it holds that  $h(\bar{x}) \leq \liminf_{x \rightarrow \bar{x}} h(x)$ . A point  $x_\star \in \text{dom } h$  is a *local minimum* for  $h$  if  $h(x) \geq h(x_\star)$  holds for all  $x$  in a neighborhood of  $x_\star$ . If the inequality can be strengthened to  $h(x) \geq h(x_\star) + \frac{\mu}{2}\|x - x_\star\|^2$  for some  $\mu > 0$ , then  $x_\star$  is a *strong local minimum*. We say that  $h$  is *level bounded* if the sublevel sets  $\{x \in \mathbb{R}^n \mid h(x) \leq \alpha\}$  are bounded for all  $\alpha \in \mathbb{R}$ , a condition which is equivalent to  $\liminf_{\|x\| \rightarrow \infty} f(x) = \infty$ .

We denote by  $\hat{\partial}h$  the *regular subdifferential* of  $h$ , where

$$v \in \hat{\partial}h(\bar{x}) \iff \liminf_{\bar{x} \neq x \rightarrow \bar{x}} \frac{h(x) - h(\bar{x}) - \langle v, x - \bar{x} \rangle}{\|x - \bar{x}\|} \geq 0. \quad (1.7)$$

The (limiting) *subdifferential* of  $h$  is  $\partial h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ , where  $v \in \partial h(x)$  iff  $x \in \text{dom } h$  and there exists a sequence  $(x^k, v^k)_{k \in \mathbb{N}}$  with  $v^k \in \hat{\partial}h(x^k)$  such that  $(x^k, h(x^k), v^k) \rightarrow (x, h(x), v)$  as  $k \rightarrow \infty$ . A necessary condition for local minimality of  $x$  for  $h$  is  $0 \in \hat{\partial}h(x)$ , see [43, Thm. 10.1].

The *proximal mapping* of  $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  with parameter  $\gamma > 0$  is the set-valued mapping defined as

$$\text{prox}_{\gamma h}(x) := \arg \min_{w \in \mathbb{R}^n} \left\{ h(w) + \frac{1}{2\gamma} \|w - x\|^2 \right\}. \quad (1.8)$$

The value function of the corresponding minimization problem, namely the *Moreau envelope* with stepsize  $\gamma$ , is denoted as

$$h^\gamma(x) := \inf_{w \in \mathbb{R}^n} \left\{ h(w) + \frac{1}{2\gamma} \|w - x\|^2 \right\}. \quad (1.9)$$

The necessary optimality condition for the problem defining  $\text{prox}_{\gamma h}$  together with the calculus rule of [43, Ex. 8.8] implies

$$\frac{1}{\gamma}(x - \bar{x}) \in \hat{\partial}h(\bar{x}) \quad \forall \bar{x} \in \text{prox}_{\gamma h}(x). \quad (1.10)$$

With  $h^*$  we indicate the convex conjugate of function  $h$ , pointwise defined as  $h^*(y) = \sup \{\langle y, \cdot \rangle - h\}$ . We say that  $h$  is *strongly convex* if there exists a constant  $\mu_h > 0$  such that  $h - \frac{\mu_h}{2}\|\cdot\|^2$  is convex, in which case we may say that  $h$  is  $\mu_h$ -*strongly convex* to make the constant  $\mu_h$  explicit.



## 2 The building blocks: DRS and ADMM

**DRS** and **ADMM** are nowadays considered textbook algorithms of the realm of convex optimization, and their properties are well documented in the literature. For instance, it is common knowledge that both algorithms are mutually equivalent when applied to the respective dual formulations, and that convergence is guaranteed for arbitrary stepsize and penalty parameters under minimal assumptions. These algorithms are in fact well understood through an elegant and powerful link with monotone operator theory, a connection in which convexity plays an indispensable role and which can explain convergence through a Fejér-type monotonicity of the generated sequences. This property entails the existence of a constant  $c > 0$  such that

$$\|s^+ - s_\star\|^2 \leq \|s - s_\star\|^2 - c\|s - s^+\|^2 \quad (2.1)$$

holds for every solution  $s_\star$ ; by telescoping the inequality and with no information about the whereabouts of any solution  $s_\star$  required, from the lower boundedness of the (squared) norm it is immediate to deduce that the *residual*  $\|s^+ - s\|$  vanishes.

Under some smoothness assumption, a new descent condition in the form of (2.1) was shown to hold even for nonconvex problems, with the squared distance  $\|\cdot - s_\star\|^2$  being replaced by another lower bounded function; upon adopting the same telescoping arguments, this led to new convergence results in the absence of convexity. In this paper we show that the new descent condition also leads to linesearch extensions of **DRS** and **ADMM** that preserve the same convergence properties and oracle complexity. In this section we present all the preliminary material that is needed for their development. We begin with **DRS**, first by offering a brief recap of the key inequalities of the nonconvex analysis developed in [51] and then by showing, through duality arguments, that the needed smoothness requirement can be replaced by strong convexity. Although convergence of **DRS** is well known in the latter case, it allows us to generalize the standing assumptions of the proposed linesearch algorithms. Finally, by means of a primal equivalence first noticed in [55] that identifies **DRS** and **ADMM**, we will obtain a similar analysis for the latter algorithm.

### 2.1 Douglas-Rachford splitting

#### 2.1.1 The nonconvex case

For convex problems, both **DRS** and **ADMM** are well known to converge for arbitrary stepsize and penalty parameters under minimal assumptions. In the nonconvex setting, the works [30, 28] and later [51] extended the analysis to the nonconvex case when the functions satisfy the following requirements.

**Assumption I** (Requirements for **DRS**: the smooth  $\varphi_1$  case). *In problem (1.1):*

A1  $\varphi_1 : \mathbb{R}^p \rightarrow \mathbb{R}$  has  $L_{\varphi_1}$ -Lipschitz continuous gradient.

A2  $\varphi_2 : \mathbb{R}^p \rightarrow \overline{\mathbb{R}}$  is lsc.

A3 Problem (1.1) admits a solution:  $\arg \min \varphi \neq \emptyset$ .

In the setting of [Assumption I](#), [30,28] pioneered the idea of employing an augmented Lagrangian function as Lyapunov potential for [DRS](#) iterations, namely

$$\mathcal{L}_c(u, v, y) := \varphi_1(u) + \varphi_2(v) + \langle y, v - u \rangle + \frac{c}{2} \|v - u\|^2$$

for some  $c \in \mathbb{R}$  (not necessarily positive). It was shown that  $\mathcal{L}_c(u, v, \gamma^{-1}(s - u))$  decreases along the iterates generated by [DRS](#) for sufficiently small  $\gamma$ , and subsequential convergence of the algorithm to stationary points was thus inferred. The results have been improved in [51], where considering  $c = 1/\gamma$  is shown to lead to a tight convergence analysis that cannot be further improved without restricting the problem assumptions. Moreover, the augmented Lagrangian is regarded as a function of the sole variable  $s$ , and thus coincides with the *Douglas-Rachford envelope* (DRE) of [36], namely

$$\varphi_\gamma^{\text{DR}}(s) := \varphi_1(u) + \varphi_2(v) + \frac{1}{\gamma} \langle s - u, v - u \rangle + \frac{1}{2\gamma} \|v - u\|^2, \quad (2.2)$$

where  $u$  and  $v$  are the result of a [DRS](#)-update with stepsize  $\gamma$  starting at  $s$  ( $u$  and  $v$  are independent of the relaxation  $\lambda$ ). As detailed in [51, Prop. 2.3 and Rem. 3.1], when [Assumption I](#) is satisfied and  $\gamma < 1/L_{\varphi_1}$  one has that  $\text{prox}_{\gamma\varphi_1}$  is Lipschitz continuous and  $\text{prox}_{\gamma\varphi_2}$  is a well-defined set-valued mapping, in the sense that  $\text{prox}_{\gamma\varphi_2}(x)$  is nonempty for every  $x \in \mathbb{R}^p$ . In fact, in order for the DRE to be well defined it suffices that  $\text{prox}_{\gamma\varphi_1}$  is single valued, as the expression (2.2) can easily be seen to equal

$$\varphi_\gamma^{\text{DR}}(s) = \varphi_1^\gamma(s) - \frac{1}{\gamma} \|s - u\|^2 + \varphi_2^\gamma(2u - s), \quad \text{with } u = \text{prox}_{\gamma\varphi_1}(s). \quad (2.3)$$

Nevertheless, under [Assumption I](#) the DRE enjoys a close kinship with the cost function  $\varphi$ , as summarized next.

**Fact 2.1** ([51, Prop. 3.2 and Thm. 3.4]). *Suppose that [Assumption I](#) holds. Then, for all  $\gamma < 1/L_{\varphi_1}$  the DRE  $\varphi_\gamma^{\text{DR}}$  is real valued, locally Lipschitz, and satisfies the following:*

- (i)  $\inf \varphi = \inf \varphi_\gamma^{\text{DR}}$  and  $\arg \min \varphi = \text{prox}_{\gamma\varphi_1}(\arg \min \varphi_\gamma^{\text{DR}})$ .
- (ii)  $\varphi_\gamma^{\text{DR}}$  is level bounded iff  $\varphi$  is level bounded.

In the same spirit as the preceding works [30,28], the convergence analysis of nonconvex [DRS](#) in [51] revolves around the following result, which assesses that the DRE decreases along the iterations by a quantity which is proportional to the fixed-point residual  $\|s^k - s^{k+1}\|^2$ . For simplicity of exposition, we use the simplified bounds on the stepsize  $\gamma$  as in [51, Rem. 4.2], which only discern whether  $\varphi_1$  is convex or not. Tight ranges are given in [51, Thm. 4.1], and require the knowledge of the *hypoconvexity* modulus of  $\varphi_1$ .

**Fact 2.2** (Sufficient decrease on the DRE [51, Thm. 4.1]). *Suppose that [Assumption I](#) is satisfied, and consider one [DRS](#) update  $s \mapsto (u, v, s^+)$  for some stepsize  $\gamma < 1/L_{\varphi_1}$  and relaxation  $\lambda \in (0, 2)$ . Then,*

$$\varphi_\gamma^{\text{DR}}(s^+) \leq \varphi_\gamma^{\text{DR}}(s) - \frac{1}{\gamma} C(\gamma L_{\varphi_1}, \lambda) \|r\|^2, \quad (2.4)$$

where  $r := u - v$  and

$$C(\alpha, \lambda) := \frac{\lambda}{(1+\alpha)^2} \left( \frac{2-\lambda}{2} - \alpha \cdot \begin{cases} \max\{\alpha - \lambda/2, 0\} & \text{if } \varphi_1 \text{ is convex,} \\ 1 & \text{otherwise} \end{cases} \right). \quad (2.5)$$

In particular, the constant  $C$  is strictly positive provided that

$$\gamma < \begin{cases} \frac{1}{L_{\varphi_1}} & \text{if } \varphi_1 \text{ is convex,} \\ \frac{2-\lambda}{2L_{\varphi_1}} & \text{otherwise.} \end{cases} \quad (2.6)$$

Combined with the lower boundedness of the DRE (Fact 2.1), the vanishing of  $s^k - s^{k+1} = \lambda(u^k - v^k)$  for the nominal DRS algorithm was readily deduced. This is enough to guarantee *subsequential convergence* of DRS to stationary points, in the sense that whenever a sequence  $(s^k)_{k \in \mathbb{N}}$  satisfies  $u^k - v^k \rightarrow 0$  with  $(u^k, v^k) \in \text{DRS}_\gamma(s^k)$ , any accumulation point  $u^*$  of  $(v^k)_{k \in \mathbb{N}}$  satisfies the stationarity condition  $0 \in \partial\varphi(u^*)$ . As detailed in the dedicated Section 4, the analysis of the here proposed DRS-based Algorithm 1 follows the same line of proof, revolving around the *decrease condition* in Fact 2.2 and the *continuity* of the DRE in Fact 2.1. Before detailing the arguments, in the remaining subsections we set the ground for extending the theory beyond Assumption I. First, with duality arguments we replace the smoothness condition with a strong convexity requirement; then, by means of the same change of variable adopted in [51] we express ADMM operations (1.5) in terms of the DRS oracle (1.4), thus obtaining the ADMM-based Algorithm 2 as a simple byproduct.

### 2.1.2 The strongly convex case

As observed in [55], for convex problems DRS (or, equivalently, its sibling ADMM) is equivalent to itself applied to the dual formulation. To see this, let  $\varphi_1$  and  $\varphi_2$  be (proper, lsc) convex functions, and consider the following dual formulation of (1.1):<sup>1</sup>

$$\underset{y \in \mathbb{R}^n}{\text{minimize}} \psi(y) := \psi_1(y) + \psi_2(y), \quad (2.7)$$

with  $\psi_2 = \varphi_2^*$  and  $\psi_1 := (\mathcal{R}\varphi_1)^* = \mathcal{R}\varphi_1^*$  (see [4, Prop. 13.23(iv)] for the last equality), where  $\mathcal{R}$  denotes the “mirroring”  $\mathcal{R}\varphi_1 := \varphi_1(-\cdot)$ . The Moreau identity [4, Thm. 14.3(ii)] yields

$$\text{prox}_{\delta\psi_1}(y) = y + \delta \text{prox}_{\varphi_1/\delta}(-y/\delta) \quad \text{and} \quad \text{prox}_{\delta\psi_2}(y) = y - \delta \text{prox}_{\varphi_2/\delta}(y/\delta). \quad (2.8)$$

Therefore, DRS applied to (2.7) produces the following triplet:

$$s_* \mapsto \begin{cases} u_* = \text{prox}_{\gamma_*\psi_1}(s_*) \\ v_* = \text{prox}_{\gamma_*\psi_2}(2u_* - s_*) \\ s_*^+ = s_* + \lambda_*(v_* - u_*), \end{cases} \Leftrightarrow \begin{cases} u_* = s_* + \gamma_* \text{prox}_{\varphi_1/\gamma_*}(-s_*/\gamma_*) \\ v_* = 2u_* - s_* - \gamma_* \text{prox}_{\varphi_2/\gamma_*}((2u_* - s_*)/\gamma_*) \\ s_*^+ = s_* + \lambda_*(v_* - u_*), \end{cases} \quad (\text{DRS}^*)$$

leading to the following result.

**Theorem 2.3** (Self-duality of DRS). *Suppose that  $\varphi_1$  and  $\varphi_2$  are (proper, lsc) convex functions, and consider a  $\text{DRS}^*$ -update  $s_* \mapsto (u_*, v_*, s_*^+)$ . Let  $s := -\frac{s_*}{\gamma_*}$  and consider*

<sup>1</sup> Specifically, (2.7) is the dual of  $\underset{x, z \in \mathbb{R}^n}{\text{minimize}} \varphi_1(x) + \varphi_2(z)$  subject to  $x - z = 0$ .

a **DRS**-update  $s \mapsto (u, v, s^+)$  with stepsize  $\gamma = 1/\gamma_*$  and relaxation  $\lambda = \lambda_*$ . Then, the variables are related as follows:

$$\begin{cases} \lambda = \lambda_* \\ \gamma = \frac{1}{\gamma_*} \\ s = -\frac{s_*}{\gamma_*} \\ u = \frac{\gamma_*}{u_* - s_*} \\ v = \frac{\gamma_*}{2u_* - s_* - v_*} \end{cases} \quad \text{or, equivalently,} \quad \begin{cases} \lambda_* = \lambda \\ \gamma_* = \frac{1}{\gamma} \\ s_* = -\frac{s}{\gamma} \\ u_* = \frac{u - s}{\gamma} \\ v_* = \frac{2u - s - v}{\gamma} \end{cases} \quad (2.9)$$

Moreover,  $\psi_{\gamma_*}^{\text{DR}}(s_*) = -\varphi_{\gamma_*}^{\text{DR}}(s)$ , where  $\psi_{\gamma_*}^{\text{DR}}$  is the DRE with stepsize  $\gamma_*$  associated to the dual problem (2.7).

*Proof.* The identities in (2.9) follow by a direct application of those in (DRS\*). Next,

$$\psi_1^{\gamma_*}(s_*) = ((\mathcal{R}\varphi_1)^*)^{1/\gamma}(-s/\gamma) = (\varphi_1^*)^{1/\gamma}(s/\gamma) = \frac{1}{2\gamma}\|s\|^2 - \varphi_1^\gamma(s),$$

and similarly  $\psi_2^{\gamma_*}(2u_* - s_*) = \frac{1}{2\gamma}\|2u - s\|^2 - \varphi_2^\gamma(2u - s)$ . Therefore, from (2.3) we have

$$\begin{aligned} \psi_{\gamma_*}^{\text{DR}}(s_*) &= \psi_1^{\gamma_*}(s_*) - \frac{1}{\gamma_*}\|u_* - s_*\|^2 + \psi_2^{\gamma_*}(2u_* - s_*) \\ &= \underbrace{\frac{1}{2\gamma}\|s\|^2 - \varphi_1^\gamma(s)}_{(2.3)} - \frac{1}{\gamma}\|u\|^2 + \underbrace{\frac{1}{2\gamma}\|2u - s\|^2 - \varphi_2^\gamma(2u - s)}_{(2.3)} \\ &= -\varphi_1^\gamma(s) + \frac{1}{\gamma}\|u - s\|^2 - \varphi_2^\gamma(2u - s) \\ &= -\varphi_\gamma^{\text{DR}}(s) \end{aligned}$$

as claimed.  $\square$

The conjugate  $h^*$  of a proper convex lsc function  $h$  is also proper convex and lsc; when  $h$  is  $\mu_h$ -strongly convex, then  $h^*$  has also  $1/\mu_h$ -Lipschitz-continuous gradient. This means that whenever  $\varphi_2$  is proper lsc and convex and  $\varphi_1$  is additionally strongly convex, the theory presented in the previous subsection applies to the dual formulation (2.7), provided dual solutions exist. As shown in [4, Cor. 27.6.(i)a], this is guaranteed whenever a constraint qualification on the domains of the two functions holds,<sup>2</sup> which leads to the following convex, but fully nonsmooth, dual version of Assumption I.

**Assumption I\*** (Requirements for **DRS**: the strongly convex case). *In problem (1.1) the following hold:*

A1  $\varphi_1 : \mathbb{R}^p \rightarrow \overline{\mathbb{R}}$  is proper, lsc, and  $\mu_{\varphi_1}$ -strongly convex.

A2  $\varphi_2 : \mathbb{R}^p \rightarrow \overline{\mathbb{R}}$  is proper, lsc, and convex.

A3  $0 \in \text{relint}(\text{dom } \varphi_1 - \text{dom } \varphi_2)$ .

<sup>2</sup> Owing to strong convexity, the additional requirement of nonemptiness of  $\arg \min \varphi_1 + \varphi_2$  in the cited reference would be trivially satisfied.

When the functions are convex and without necessarily either one being strongly so, plain **DRS** iterations are known to converge for any stepsize  $\gamma$  and relaxation  $\lambda \in (0, 2)$  provided that  $\partial\varphi_1 + \partial\varphi_2$  attains a zero [4, Cor. 28.3]. **Assumption I\*.A3** is a mild and easily verifiable domain qualification ensuring this latter property, and is therefore quite standard. What is instead crucial in our setting is *strong* convexity, as it guarantees well definedness of **Algorithm 1** even in the absence of smoothness. The duality pairing strong convexity and smoothness was also exploited in [36], where the DRE was first introduced and **DRS** was shown to be a scaled gradient descent algorithm on it. This observation led to the development of a Nesterov-type acceleration whenever  $\varphi_2$  is convex and  $\varphi_1$  is strongly convex and quadratic. The analysis was then extended to **ADMM** in the follow-up work [37] by means of duality arguments.

A consequence of **Theorem 2.3** is the following dual version of **Fact 2.2**.

**Corollary 2.4.** *Suppose that **Assumption I\*** holds and consider one (**DRS**) update  $s \mapsto (u, v, s^+)$  applied to the primal formulation (1.1) with stepsize  $\gamma > 1/\mu_{\varphi_1}$  and relaxation  $\lambda \in (0, 2)$ . Then,*

$$\varphi_{\gamma}^{\text{DR}}(s^+) \geq \varphi_{\gamma}^{\text{DR}}(s) + \frac{1}{\gamma} C(1/\gamma\mu_{\varphi_1}, \lambda) \|r\|^2$$

where  $r := u - v$  and  $C$  is a strictly positive constant defined as in (2.5).

## 2.2 Alternating direction method of multipliers

Although apparently more general, it is well known that for convex problems **ADMM** coincides with **DRS** applied to the dual formulation, and vice versa. More generally, problem (1.2) can be reduced to a **DRS**-compatible form as

$$\underset{s \in \mathbb{R}^p}{\text{minimize}} (Af)(s) + (Bg)(b - s), \quad (2.10)$$

where for  $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$  and  $C \in \mathbb{R}^{p \times n}$  we indicate with  $(Ch)$  the *epicomposition*

$$(Ch) : \mathbb{R}^p \rightarrow \overline{\mathbb{R}}, \quad \text{defined as} \quad (Ch)(s) := \inf \{h(x) \mid Cx = s\}.$$

It was shown in [55, Thm. 1] and later generalized in [51, Thm. 5.5] that one iteration of **ADMM** applied to (1.2) is equivalent to one step of **DRS** applied to this new reformulation with stepsize  $\gamma = 1/\beta$ , as stated next.

**Fact 2.5** (Primal equivalence of **DRS** and **ADMM** [51, Thm. 5.5]). *Starting from a triplet  $(x, y, z) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n$ , consider an **ADMM**-update applied to problem (1.2) with relaxation  $\lambda$  and penalty  $\beta > 0$ . Let*

$$\begin{cases} s := Ax - y/\beta \\ u := Ax \\ v := b - Bz \end{cases} \quad \text{and, similarly,} \quad \begin{cases} s^+ := Ax^+ - y^+/\beta \\ u^+ := Ax^+ \\ v^+ := b - Bz^+. \end{cases}$$

Then, the variables are related as follows:

$$\begin{cases} s^+ = s + \lambda(v - u) \\ u^+ \in \text{prox}_{\gamma\varphi_1}(s^+) \\ v^+ \in \text{prox}_{\gamma\varphi_2}(2u^+ - s^+), \end{cases} \quad \text{where} \quad \begin{cases} \varphi_1 := (Af) \\ \varphi_2 := (Bg)(b - \cdot) \\ \gamma := 1/\beta. \end{cases}$$

Moreover,

- (i)  $-A^\top y^+ \in \hat{\partial} f(x^+)$ , and
- (ii)  $\text{dist}(-B^\top y^+, \hat{\partial} g(z^+)) \leq \beta \|B\| \|Ax^+ + Bz^+ - b\|$ .

This enabled the possibility to infer the convergence of nonconvex **ADMM** from the simpler analysis of that of **DRS**, when the reformulation (2.10) complies with the needed **DRS** requirements.

**Assumption II** (Requirements for **ADMM**: the smooth  $(Af)$  case). *In problem (1.2),  $A \in \mathbb{R}^{p \times m}$ ,  $B \in \mathbb{R}^{p \times n}$ ,  $b \in \mathbb{R}^p$ ,  $f : \mathbb{R}^m \rightarrow \mathbb{R}$  and  $g : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  are such that:*

- A1  *$A$  is surjective (full row rank) and  $(Af) : \mathbb{R}^p \rightarrow \mathbb{R}$  has  $L_{(Af)}$ -Lipschitz gradient.*
- A2  *$(Bg) : \mathbb{R}^p \rightarrow \bar{\mathbb{R}}$  is lsc.*
- A3 *a solution exists:  $\arg \min \Phi \neq \emptyset$ , where  $\Phi(x, z) = f(x) + g(z) + \delta_{Ax+Bz=b}(x, z)$ .*

Similarly, in parallel to **Assumption I\*** we may also consider the strongly convex case for **ADMM**. As shown in [51, Prop. 5.4], for any matrix  $A$  the function  $\varphi_1 := (Af)$  is strongly convex whenever so is  $f$ , in which case the strong convexity modulus is  $\mu_f/\|A\|^2$ .<sup>3</sup> Moreover, with  $\varphi_2 = (Bg)(b - \cdot)$  one has that  $\text{dom } \varphi_1 - \text{dom } \varphi_2 = A \text{ dom } f + B \text{ dom } g - b$ , owing to the fact that  $\text{dom } \varphi_1 = A \text{ dom } f$  and  $\text{dom } \varphi_2 = b - B \text{ dom } g$ , see [4, Prop. 12.36(i)]. Altogether, the following **ADMM** counterpart of **Assumption I\*** is obtained.

**Assumption II\*** (Requirements for **ADMM**: the strongly convex case). *In problem (1.2), the following hold:*

- A1  *$f : \mathbb{R}^m \rightarrow \bar{\mathbb{R}}$  is lsc, proper, and  $\mu_f$ -strongly convex.*
- A2  *$(Bg) : \mathbb{R}^n \rightarrow \bar{\mathbb{R}}$  is lsc, proper, and convex (e.g. when  $g$  is proper, convex and level bounded).*
- A3  *$b \in \text{relint}(A \text{ dom } f + B \text{ dom } g)$ .*

**Fact 2.6.** *Starting from a triplet  $(x, y, z) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n$ , consider an **ADMM**-update applied to problem (1.2) with relaxation  $\lambda$  and penalty  $\beta > 0$ . If*

- (A) *either **Assumption II** holds and  $\beta > L_{(Af)}$ ,*
- (B) *or **Assumption II\*** holds and  $\beta < \mu_f/\|A\|^2$ ,*

*then  $\varphi_\gamma^{\text{DR}}(s^+) = \mathcal{L}_\beta(x^+, z^+, y^+)$  for  $s^+ = Ax^+ - y^+/\beta$  and  $\gamma = 1/\beta$ , with  $\varphi_\gamma^{\text{DR}}$  being the DRE associated to the equivalent problem formulation (2.10).*

The following is a straightforward consequence of **Facts 2.2** and **2.6**.

**Corollary 2.7.** *Starting from a triplet  $(x^-, y^-, z^-) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n$ , consider the **ADMM** updates  $(x^-, y^-, z^-) \rightarrow (x, y, z) \rightarrow (x^+, z^+, y^+)$  with penalty  $\beta$  and relaxation  $\lambda$ . Let  $r = Ax + Bz - b$  and  $C$  be as in (2.5).*

- (i) *If **Assumption II** holds and  $\beta > L_{(Af)}$ , then*

$$\mathcal{L}_\beta(x^+, z^+, y^+) \leq \mathcal{L}_\beta(x, z, y) - \beta C(L_{(Af)}/\beta, \lambda) \|r\|^2,$$

*with the constant  $C$  being strictly positive provided that*

$$\beta > \begin{cases} L_{(Af)} & \text{if } f \text{ is convex,} \\ \frac{2L_{(Af)}}{2-\lambda} & \text{otherwise.} \end{cases} \quad (2.11)$$

<sup>3</sup> In the limiting case  $A = 0$ , one has that  $(Af) = f(0) + \delta_{\{0\}}$  is lsc and  $\infty$ -strongly convex for any  $f$ , and properness amounts to the condition  $0 \in \text{dom } f$ .

(ii) If *Assumption II\** holds and  $\beta < \mu_f/\|A\|^2$ , then

$$\mathcal{L}_\beta(x^+, z^+, y^+) \geq \mathcal{L}_\beta(x, z, y) + \beta C(\beta\|A\|^2/\mu_f, \lambda)\|r\|^2,$$

and  $C$  is strictly positive.

### 3 The linesearch algorithms

As shown in [30, 28, 51], both *DRS* and *ADMM* converge under mild assumptions that do not entail convexity of either functions. As a consequence, a wide range of nonsmooth and nonconvex problems can be addressed by iterating relatively simple operations. On the other hand, it is well known that even for convex problems the convergence can be prohibitively slow unless the problem is well scaled, which is rarely the case in practice. In contrast, fast local methods such as Newton-type exist that by exploiting higher-order information can suitably reshape the problem into a more convenient geometry. The major hindrance against their employment is their *local* nature, in the sense that convergence is guaranteed only if the starting point is already close enough to a solution, on top of some regularity criteria around such solution. For this reason, fast local methods are typically paired with a *linesearch* that globalizes convergence by ensuring a decrease condition on the cost or on a surrogate merit function.

The purpose of *Algorithms 1* and *2*, presented in *Section 1.3*, is exactly to complement the global (subsequential) convergence and operational simplicity of *DRS* and *ADMM* with the fast local convergence of Newton-type schemes by means of a tailored linesearch. It should be noted that the proposed method differs considerably from classical linesearch strategies, which can only cope with *directions of descent* and thus heavily hinge on differentiability requirements. Different from the convex setting of [36], where the envelope function is differentiable, in the setting dealt here it can only be guaranteed to be (locally Lipschitz) continuous (cf. *Fact 2.1*), and thus not suitable for a standard backtracking based, e.g., on the Armijo condition. For this reason, *Algorithms 1* and *2* here proposed adopt the novel linesearch protocol of the umbrella *Continuous-Lyapunov Descent method* (CLyD) [48, §4] already benchmarked with the PANOC [47] and NAMA [46] solvers, respectively based on the forward-backward splitting and the alternating minimization algorithm.

The chosen linesearch allows us to discard differentiability requirements as it merely exploits continuity of the DRE and the sufficient decrease property. Before elaborating on the details in the next subsection, we first prove that *Algorithms 1* and *2* are linked through the same equivalence relating the underlying *DRS* and *ADMM* oracles which, for exposition clarity, will be referred to as the *nominal steps*.

**Proposition 3.1** (Equivalence of *Algorithm 1* and *Algorithm 2*). *Suppose that Assumption II [resp. Assumption II\*] holds and consider the iterates generated by Algorithm 2 with penalty  $\beta$  and relaxation  $\lambda$ , starting from  $(x^{-1}, y^{-1}, z^{-1})$ . For each  $k \in \mathbb{N}$ , let*

$$\begin{cases} s^k := Ax^k - y^k/\beta = b - Bz^{k-1} - y^{k-1/2}/\beta \\ u^k := Ax^k \\ v^k := b - Bz^k. \end{cases} \quad (3.1)$$

Then,  $\varphi_1 := (Af)$ ,  $\varphi_2 := (Bg)(b - \cdot)$ , and  $\varphi := \varphi_1 + \varphi_2$  satisfy [Assumption I](#) [resp. [Assumption I\\*](#)], and  $(s^k, u^k, v^k)_{k \in \mathbb{N}}$  is a sequence generated by [Algorithm 1](#) with stepsize  $\gamma = 1/\beta$ , relaxation  $\lambda$ , starting from  $s^0 := b - Bz^{-1} - y^{-1/2}/\beta$  and with the same sufficient decrease constant  $c$  and choice of directions  $(d^k)_{k \in \mathbb{N}}$ . Moreover,  $\varphi_\gamma^{\text{DR}}(s^k) = \mathcal{L}_\beta(x^k, z^k, y^k)$  and  $Ax^k + Bz^k - b = u^k - v^k$  hold for every  $k$ , hence both sequences  $(r^k)_{k \in \mathbb{N}}$  and  $(\tau_k)_{k \in \mathbb{N}}$  coincide in the two algorithms at every iteration.

*Proof.* Clearly, [Assumption I](#) [resp. [Assumption I\\*](#)] is satisfied if [Assumption II](#) [resp. [Assumption II\\*](#)] holds. It follows from [Facts 2.5](#) and [2.6](#) that for any  $\tilde{y}, \tilde{z}$  such that  $s = -\tilde{y}/\beta - B\tilde{z} + b$  one has

$$\text{DRS}_\gamma(s) = \{(Ax, b - Bv) \mid (x, y, z) \in \text{ADMM}_\beta(\tilde{y}, \tilde{z})\}, \quad (3.2a)$$

$$\varphi_\gamma^{\text{DR}}(s) = \mathcal{L}_\beta(x, z, y) \quad \forall (x, y, z) \in \text{ADMM}_\beta(\tilde{y}, \tilde{z}). \quad (3.2b)$$

In particular, the entire claim will follow once we prove that  $s^k$  and  $(y^{k-1/2}, z^{k-1})$  are related as in [\(3.1\)](#) for every  $k$ . We proceed by induction. The case  $k = 0$  follows from the definition of the initial iterate  $s^0$  as in the statement. Suppose now that the identity holds up to iteration  $k$ , and let  $r^k = Ax^k + Bz^k - b$  be as in [Algorithm 2](#); then, observe that

$$\begin{aligned} \tilde{s}^{k+1} &= s^k + \lambda(v^k - u^k) \quad \overbrace{\quad}^{r^k} && \text{(step 1.2)} \\ &= Ax^k - y^k/\beta - \lambda(Ax^k + Bz^k - b) && \text{(induction)} \\ &= b - Bz^k - y^k/\beta + (1 - \lambda)r^k \\ &= b - Bz^k - \tilde{y}^{k+1/2}/\beta. && \text{(step 2.2)} \end{aligned} \quad (3.3)$$

For  $\tau \in \mathbb{R}$ , let  $\tilde{s}_\tau^{k+1} := (1 - \tau)\tilde{s}^{k+1} + \tau(s^k + d^k)$  and  $\tilde{y}_\tau^{k+1/2} := (1 - \tau)\tilde{y}^{k+1/2} + \tau(y^k - \beta(r^k + d^k))$  be the candidate updates with stepsize  $\tau$  at [step 1.4](#) and [step 2.4](#), respectively. We have

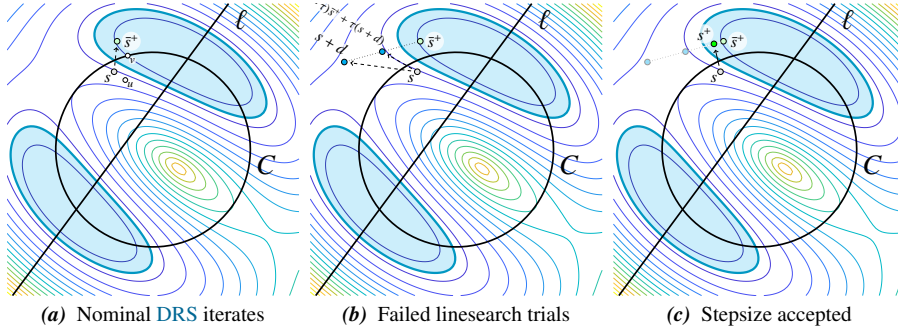
$$\begin{aligned} \tilde{s}_\tau^{k+1} &\stackrel{(3.3)}{=} (1 - \tau)\tilde{s}^{k+1} + \tau(s^k + d^k) \\ &= (1 - \tau)(b - \tilde{y}^{k+1/2}/\beta - Bz^k) + \tau \overbrace{(Ax^k - y^k/\beta + d^k)}^{s^k} && \text{(induction)} \\ &= b - Bz^k - \left[ (1 - \tau)\tilde{y}^{k+1/2}/\beta + \tau(y^k/\beta - Ax^k - Bz^k + b - d^k) \right] \\ &= b - Bz^k - \left[ (1 - \tau)\tilde{y}^{k+1/2}/\beta + \tau(y^k/\beta - r^k - d^k) \right] \\ &= b - Bz^k - \tilde{y}_\tau^{k+1/2}/\beta. \end{aligned}$$

It follows from [\(3.2\)](#) that  $\varphi_\gamma^{\text{DR}}(\tilde{s}_\tau^{k+1}) = \mathcal{L}_\beta(\tilde{x}_\tau^{k+1}, \tilde{z}_\tau^{k+1}, \tilde{y}_\tau^{k+1})$  for any  $(\tilde{x}_\tau^{k+1}, \tilde{z}_\tau^{k+1}, \tilde{y}_\tau^{k+1}) \in \text{ADMM}_\beta(\tilde{y}_\tau^{k+1/2}, z^k)$ . Combined with the fact that  $r^k = u^k - v^k$  holding by induction (that is,  $r^k$  is the same in both algorithms), we conclude that stepsize  $\tau = \tau_k$  is accepted at the  $k$ -th iteration of [Algorithm 2](#) iff so happens in [Algorithm 1](#). In particular, one has

$$s^{k+1} = \tilde{s}_{\tau_k}^{k+1} = b - Bz^k - \tilde{y}_{\tau_k}^{k+1/2} = b - Bz^k - y^{k+1/2},$$

which completes the induction argument.  $\square$





**Figure 1:** Main steps of [Algorithm 1](#). One call to the *DRS* oracle at  $s$  yields the pair  $(u, v)$  and the nominal *DRS* update  $\bar{s}^+ = s + \lambda(v - u)$ . On the DRE, this implies a decrease by at least  $\frac{1}{\gamma} C \|v - u\|^2$ . Since  $\varphi_\gamma^{\text{DR}}$  is continuous and  $c \leq C$ , all points close enough to  $\bar{s}^+$  belong to the sublevel set  $[\varphi_\gamma^{\text{DR}} \leq \varphi_\gamma^{\text{DR}}(s) - c\|u - v\|^2]$  (cyan-shaded region). Therefore, **for any direction**  $d$ , all points close to  $\bar{s}^+$  in the line segment  $[\bar{s}^+, s + d] = \{(1 - \tau)\bar{s}^+ + \tau(s + d) \mid \tau \in [0, 1]\}$  belong to this set, hence the linesearch is accepted for small enough  $\tau$ .

### 3.1 A continuity-based linesearch

We now discuss the core of [Algorithms 1](#) and [2](#), namely the linesearch strategy starting at [steps 1.6](#) and [2.6](#), respectively. Remarkably, not only is it flexible to *any* update direction  $d$  (and not only of descent type), but it will be shown in [Theorem 4.6](#) that it also accepts unit stepsize whenever  $d$  is “good”, in a sense that will be made precise in [Section 4.2](#). In other words, the proposed linesearch is robust against the *Maratos effect* [\[32\]](#), see also [\[26, §6.2\]](#), a pathology typical of linesearch methods in nonsmooth optimization such as sequential quadratic programming that inhibits the achievement of fast convergence rates.

Owing to the equivalence proven in [Proposition 3.1](#), we limit the preliminary discussion to the *DRS*-based [Algorithm 1](#), as the rationale of the *ADMM* counterpart uses the same arguments. Similarly, we may limit the discussion to the case in which [Assumption I](#) holds, as the complementary case of [Assumption I\\*](#) only differs from a change of sign in the DRE (cf. [Theorem 2.3](#)), covered by the initialization  $\pi = -1$ , and a choice of stepsize  $\gamma$  and decrease constant  $c$  consistent with [Corollary 2.4](#), whence the different initialization prescribed in [Algorithm 1](#).

Suppose that the current iterate is  $s \in \mathbb{R}^p$ , and let  $d \in \mathbb{R}^p$  be an arbitrary candidate update direction at  $s$ . What  $d$  is, and how it is retrieved is irrelevant at the moment and will be discussed in detail in the dedicated [Section 3.3](#); suffice it to say that the choice of an update direction  $d$  represents our degree of freedom for extending *DRS* while maintaining its (subsequential) convergence properties, and that “ideally” we would like to replace the *nominal DRS* update with the chosen  $s^+ = s + d$ , for we have reason to believe this choice will lead us closer to a solution.

Let the nominal update be  $\bar{s}^+ = s + \lambda(v - u)$  as in [step 1.2](#). Due to the sufficient decrease property on  $\varphi_\gamma^{\text{DR}}$  (cf. [Fact 2.2](#)), it holds that

$$\varphi_\gamma^{\text{DR}}(\bar{s}^+) \leq \varphi_\gamma^{\text{DR}}(s) - \frac{1}{\gamma} C(\gamma L_{\varphi_1}, \lambda) \|r\|^2$$

where  $C$  is as in (2.5) and  $r = u - v$ . However, nothing can be guaranteed as to whether  $\varphi_\gamma^{\text{DR}}(s + d)$  is also (sufficiently) smaller than  $\varphi_\gamma^{\text{DR}}(s)$  or not, nor can we hope to enforce the condition with a classical backtracking  $s + \tau d$  for small  $\tau > 0$ , as no notion of descent is known to  $\varphi_\gamma^{\text{DR}}$  (which is continuous but not necessarily differentiable, on top of the fact that the direction  $d$  is even arbitrary). Nevertheless, for any  $c \leq C$ , not only does  $\bar{s}^+$  satisfy the sufficient decrease with constant  $c$ , but due to continuity of  $\varphi_\gamma^{\text{DR}}$  so do all the points around: loosely speaking,

$$\varphi_\gamma^{\text{DR}}(s') \leq \varphi_\gamma^{\text{DR}}(s) - \frac{c}{\gamma} \|r\|^2 \quad \text{for all } s' \text{ close to } \bar{s}^+. \quad (3.4)$$

The idea is then to “push” the candidate update  $s + d$  towards the “safe” update  $\bar{s}^+$  until the *relaxed* decrease condition (3.4) holds. One way to do so is through a linesearch along the segment connecting the “ideal” update  $s + d$  and the “safe” nominal update  $\bar{s}^+$ , as done in step 1.4. The procedure is synopsized in Figure 1 for the toy problem of finding a point in the intersection of a line  $\ell$  and a circumference  $C$ , cast in the form (1.1) as

$$\underset{s \in \mathbb{R}^2}{\text{minimize}} \varphi(s) := \underbrace{\frac{1}{2} \text{dist}^2(s, \ell)}_{\varphi_1(s)} + \underbrace{\delta_C(s)}_{\varphi_2(s)}.$$

The contour levels correspond to those of  $\varphi_\gamma^{\text{DR}}$ ; notice that, since  $\text{prox}_{\gamma\varphi_1} \equiv \text{id}$  on  $\ell$  and that  $\arg \min \varphi \subset \ell$ , for this problem it holds that  $\arg \min \varphi = \arg \min \varphi_\gamma^{\text{DR}}$ , cf. Fact 2.1(i).

### 3.2 Iteration complexity

In both the proposed algorithms, a step of the nominal method is required for the evaluation of  $\varphi_\gamma^{\text{DR}}$  or  $\mathcal{L}_\beta$ , where with “nominal method” we indicate DRS for Algorithm 1 and ADMM for Algorithm 2. Therefore, the number of nominal steps performed at iteration  $k$  corresponds to the number of backtrackings  $i_k$ . In other words, the  $k$ -th iteration of the proposed algorithm is, in general, as expensive as  $i_k$  many nominal iterations. In order to bound complexity, a maximum number of backtrackings  $i_{\max}$  can be imposed, say  $i_{\max} = 5$ , so that whenever the linesearch condition fails  $i_{\max}$  many times, one can discard the direction  $d^k$  and proceed with a nominal update. Nevertheless, as explained in the following remark, whenever function  $\varphi_1$  is (generalized) quadratic, not necessarily convex, the linearity of the  $u$ -update can conveniently be exploited to save computations in the linesearch. This is particularly appealing when, additionally, evaluating  $\text{prox}_{\gamma\varphi_2}$  is cheap (projections onto simple sets, thresholding, ...), in which case each iteration is, at most, roughly twice as expensive as one of DRS. Similar arguments also apply to the ADMM-Algorithm 2, and examples of this kind will be given in the Simulations Section 5.

**Remark 3.2** (Exploiting linearity of  $\text{prox}_{\gamma\varphi_1}$ ). If function  $\varphi_1$  is quadratic (or *generalized* quadratic, i.e., quadratic plus the indicator of an affine subspace), then its proximal mapping is affine and the  $u$ -update at step 1.5 can be expanded to

$$u^{k+1} = \text{prox}_{\gamma\varphi_1}[(1 - \tau_k)\bar{s}^{k+1} + \tau_k(s^k + d^k)] = (1 - \tau_k) \text{prox}_{\gamma\varphi_1}(\bar{s}^{k+1}) + \tau_k \text{prox}_{\gamma\varphi_1}(s^k + d^k).$$

Then, for  $\tau_k = 1/2$  instead of computing directly  $u^{k+1}$  one can evaluate  $\text{prox}_{\gamma\varphi_1}(\bar{s}^{k+1})$  and obtain  $u^{k+1}$  by linear combination, and similarly all subsequent trials for smaller values of  $\tau_k$  will not require any additional evaluation of  $\text{prox}_{\gamma\varphi_1}$ . In this case, the number of evaluations of  $\text{prox}_{\gamma\varphi_1}$  at each iteration of the algorithm is at most 2.

Similarly, at most two evaluations of  $\varphi_1$  (needed for computing the value of  $\varphi_\gamma^{\text{DR}}$ ) will be enough. In fact, denoting  $\bar{u}^{k+1} := \text{prox}_{\gamma\varphi_1}(\bar{s}^{k+1})$  and  $u_0^{k+1} := \text{prox}_{\gamma\varphi_1}(s^k + d^k)$ , one has that

$$\ell(\tau) := \varphi_1((1 - \tau)\bar{u}^{k+1} + \tau u_0^{k+1})$$

is a one-dimensional quadratic function, hence  $\ell(\tau) = a + b\tau + c\tau^2$  with  $a = \ell(0)$ ,  $b = \ell'(0)$ , and  $c = \ell(1) - a - b$ , i.e.,

$$\begin{cases} a = \varphi_1(\bar{u}^{k+1}) \\ b = \frac{1}{\gamma} \langle \bar{s}^{k+1} - \bar{u}^{k+1}, u_0^{k+1} - \bar{u}^{k+1} \rangle \\ c = \varphi_1(u_0^{k+1}) - a - b. \end{cases}$$

The expression of  $b$  uses the fact that  $\bar{s}^{k+1} - \bar{u}^{k+1} \in \gamma\partial\varphi_1(\bar{u}^{k+1})$ , cf. (1.10). Consequently, once  $\varphi_1(u_0^{k+1})$  and  $\varphi_1(\bar{u}^{k+1})$  are computed, all the needed values of  $\ell(\tau_k)$  can be retrieved at negligible cost.  $\triangle$

### 3.3 Choice of direction

Although the proposed algorithmic framework is robust to any choice of directions  $d^k$ , its efficacy is greatly affected by the specific selection. This subsection provides an overview on some convenient choices of update directions  $d^k$ . Specifically, we propose a generalization of Nesterov extrapolation [33] that is suited for nonconvex problems (although merely heuristical, without optimal convergence guarantees), and then discuss three popular Newton-type schemes: (modified) Broyden [10, 39], BFGS [11, 17, 19, 44], and Anderson acceleration [1]. Although true higher-order information, when available, can also be considered, we prefer to limit our overview to methods that preserve the simple oracle of the original DRS and ADMM. In the convex case, the interested reader can find an extensive collection of generalized Jacobians of proximal mappings in [49, §15.6], useful for deriving directions based on *linear Newton approximation schemes* [15, §7.5.1].

**Remark 3.3** (Recommended directions). While all of the directions choices outlined in this section enable the subsequential convergence properties of Algorithm 1 and Algorithm 2 (cf. Section 4), some choice may perform better than the others on specific problems. In our experience, and as confirmed by the numerical evidence in Section 5, Broyden and L-BFGS directions consistently provided fast convergence: the Broyden method is supported by the theory, in that superlinear convergence can be proved under mild assumptions (cf. Section 4.2), while L-BFGS clearly scales better with the problem dimension, given its limited-memory nature. On the other hand, Nesterov directions proved effective only in case  $\varphi_1$  (or  $f$ ) is convex, while Anderson acceleration directions only performed well in one example in our experiments. These are purely empirical observations, and a thorough understanding of when (and why) these directions perform well should be the subject of future investigation.  $\triangle$

### 3.3.1 Nesterov acceleration

It was shown in [36] that when  $\varphi_2$  is convex and  $\varphi_1$  is convex quadratic, then  $\varphi_\gamma^{\text{DR}}$  is convex and continuously differentiable for  $\gamma < 1/L_{\varphi_1}$ . This enabled the possibility, for this specific case, to extend the employment of the optimal Nesterov acceleration techniques [33] to DRS. By using duality arguments, this fact was extended in [37] where an *accelerated ADMM* scheme is proposed for problems in the form (1.2) with  $B = I$ ,  $f$  strongly convex quadratic, and  $g$  convex.

Although not supported by the theory, extensive numerical evidence suggests that such extrapolations perform quite effectively regardless of what function  $\varphi_2$  or  $g$  in problems (1.1) and (1.2) are, while convexity of  $\varphi_1$  and  $f$  seems instead to play an important role, as mentioned in Remark 3.3. The main limitation in a direct employment of the acceleration is that convergence itself is not guaranteed. However, thanks to the arbitrariness of  $d^k$  in Theorems 4.3 and 4.4, we can enforce these updates into Algorithms 1 and 2 and thus obtain a (damped, monotone) extrapolation that is guaranteed to be globally (subsequentially) convergent.

♣ Fast DRS: in Algorithm 1, select

$$d^k := \frac{k}{k+3}(\bar{s}^{k+1} - \bar{s}^k) - \lambda r^k$$

♣ Fast ADMM: in Algorithm 2, select

$$d^k := -\lambda r^k - \frac{k}{k+3}(Bz^k - Bz^{k-1} + (\bar{y}^{k+1/2} - \bar{y}^{k-1/2})/\beta).$$

These extensions differ from the approach proposed in [31] for the proximal gradient method, as we do not discard the candidate fast direction when sufficient decrease is not satisfied but rather dampen it with a backtracking.

### 3.3.2 Quasi-Newton methods

The termination criterion for both Algorithms 1 and 2 is based on (the norm of) the fixed-point residual of the underlying splitting schemes, namely  $u - v = \frac{1}{\lambda}(\bar{s} - \bar{s}^+)$  for DRS, which corresponds to  $Ax + Bz - b$  in ADMM. Under some assumptions such as *prox-regularity* [43, §13.F], eventually the updates  $s \mapsto s^+$  are uniquely determined, that is, the inclusion  $v \in \text{prox}_{\gamma\varphi_2}(2u - s)$  in DRS becomes an equality. It then turns out that one ends up solving a system of nonlinear equations, namely finding  $s$  such that  $s - s^+ = 0$  in DRS, and similarly for the residual  $Ax + Bz - b = 0$  in ADMM.

Parallel to what done in [52] for (nonconvex) forward-backward splitting, we may then consider directions stemming from fast methods for nonlinear equations, namely  $d^k = -H_k R$ , where  $R$  is the fixed-point residual map ( $R_\gamma^{\text{DR}}(s) = u - v$  for DRS and  $R_\beta^{\text{ADMM}}(y^{1/2}, z^{-1}) = Ax + Bz - b$  for ADMM), and  $H_k$  is some approximation to the inverse of its (generalized) Jacobian. To maintain the simplicity of the original DRS and ADMM, this can be done efficiently by means of quasi-Newton methods, which starting from an invertible matrix  $H_0$  perform low-rank updates based on available quantities. Such quantities are pairs of vectors  $(p_k, q_k)$ , where  $p_k$  is the difference between consecutive iterates and  $q_k$  is the difference of the fixed-point residuals. Namely,

♣ quasi-Newton **DRS**: in [Algorithm 1](#) use

$$d^k = -H_k r^k \quad \text{and} \quad \begin{cases} p_k = d^k \\ q_k = r_0^{k+1} - r^k, \end{cases} \quad (3.5a)$$

where  $r_0^{k+1} = u_0^{k+1} - v_0^{k+1}$  with  $(u_0^{k+1}, v_0^{k+1}) \in \text{DRS}_\gamma(s^k + d^k)$  is the residual computed in the first linesearch trial.

♣ quasi-Newton **ADMM**: start with  $H_0 = \mu I$  for some  $\mu > 0$ , and in [Algorithm 2](#) use

$$d^k = -H_k r^k \quad \text{and} \quad \begin{cases} p_k = d^k \\ q_k = r_0^{k+1} - r^k, \end{cases} \quad (3.5b)$$

where  $r_0^{k+1}$  is the residual computed in the first linesearch trial, as in the **DRS** case.

In textbook quasi-Newton methods employed in smooth optimization, where the objective is to find a point where the gradient of the cost function is zero,  $q_k$  would amount to the difference of gradient values, see e.g., [34, pp. 24–26]. Here, where the objective is to find a zero of the residual operator,  $q_k$  is the difference of residual values. Nevertheless, a naive adaptation from the smooth case would suggest considering  $p_k = s^{k+1} - s^k$  and  $q_k = r^{k+1} - r^k$ , whereas here the values obtained from the first linesearch iteration trial are considered, regardless of whether stepsize  $\tau_k = 1$  is accepted or not. As will be clear in the superlinear convergence proof of [Theorem 4.8](#), the one considered here is a more educated adaptation that complies with the proposed innovative linesearch.

We will now list a few update rules for  $H_k$  based on the indicated pairs  $(p_k, q_k)$ .

**a. BFGS.** Start with  $H_0 > 0$  and update as follows:

$$H_{k+1} = H_k + \frac{\langle p_k, q_k \rangle + \langle H_k q_k, q_k \rangle}{(\langle p_k, q_k \rangle)^2} p_k p_k^\top - \frac{H_k q_k s_k^\top + s_k q_k^\top H_k}{\langle p_k, q_k \rangle}.$$

Whenever  $\langle p_k, q_k \rangle \leq 0$ , one can either set  $H_{k+1} = H_k$  or use a different vector  $p_k$  as proposed in [40]. The limited-memory variant L-BFGS [34, Alg. 7.4], which does not require storage of full matrices  $H_k$  or matrix-vector products but only storage of the last few pairs and scalar products, can conveniently be considered. Although very well performing in practice, to the best of our knowledge fast convergence of BFGS can only be shown when the Jacobian of  $R$  at the limit point is symmetric, which hardly ever holds in our framework. We suspect, however, that the well performance of BFGS derives from the observation that, when it exists, the Jacobian of  $R$  at a local minimum is diagonalizable and has all positive eigenvalues, and is thus similar to a symmetric positive semidefinite matrix (cf. (A.1a) and [52, Thm. 4.11]).

**b. Modified Broyden.** Fix  $\bar{\vartheta} \in (0, 1)$ , e.g.,  $\bar{\vartheta} = 0.2$ , an invertible matrix  $H_0$ , and update as follows:

$$H_{k+1} = H_k + \frac{p_k - H_k q_k}{\langle p_k, (1/\vartheta_k - 1)p_k + H_k q_k \rangle} p_k^\top H_k \quad (3.6a)$$

where

$$\vartheta_k := \begin{cases} 1 & \text{if } |\delta_k| \geq \bar{\vartheta} \\ \frac{1 - \text{sgn}(\delta_k)\bar{\vartheta}}{1 - \delta_k} & \text{if } |\delta_k| < \bar{\vartheta} \end{cases} \quad \text{and} \quad \delta_k := \frac{\langle H_k q_k, p_k \rangle}{\|p_k\|^2}, \quad (3.6b)$$

with the convention that  $\text{sgn } 0 = 1$ . The original Broyden formula [10] corresponds to  $\vartheta_k \equiv 1$ , while  $\vartheta_k$  as in (3.6b) ensures nonsingularity of all matrices  $H_k$  [39].

- c. **Anderson acceleration.** Fix a buffer size  $m \geq 1$  and start with  $H_0 = \mathbf{I}$ . For  $k \geq 1$ , let

$$H_k = \mathbf{I} + (\mathcal{P}_k - \mathbf{Q}_k)(\mathbf{Q}_k^\top \mathbf{Q}_k)^{-1} \mathbf{Q}_k^\top,$$

where the columns of matrix  $\mathcal{P}_k$  are the last vectors  $p_{k-M}, \dots, p_{k-1}$  and those of  $\mathbf{Q}_k$  are the last vectors  $q_{k-M}, \dots, q_{k-1}$ , with  $M = \min\{k, m\}$ . If  $\mathbf{Q}_k$  is not full-column rank, for  $x \in \mathbb{R}^M$  the product  $(\mathbf{Q}_k^\top \mathbf{Q}_k)^{-1}x$  is meant in a least-square sense. This is a limited-memory scheme, as it requires only the storage of few vectors and the solution of a small  $M \times M$  linear system. Anderson acceleration originated in [1]; here we use the interpretation well explained in [16] of (*inverse*) *multi-secant* update:  $H_k$  is the matrix closest to the identity (in the Frobenius norm) among those satisfying  $H_k \mathbf{Q}_k = \mathcal{P}_k$ .

### 3.4 Adaptive variants

One drawback of Algorithms 1 and 2 is that both the stepsize  $\gamma$  in the former and the penalty  $\beta$  in the latter have to be chosen offline based either on a Lipschitz constant or on a strong convexity modulus. In practice, the estimation of these quantities is often challenging and prone to yield very conservative approximations, dooming the algorithms to slow convergence in early iterations (that is, in the globalization stage when the effect of the fast local directions is not triggered yet). Moreover, even when such constants are known the algorithms may potentially work also with less conservative estimates which better reflect the local geometry.

In order to circumvent these issues and allow for out-of-the-box implementations, we may resort to the adaptive variants of the DRS and ADMM oracles as described in [51, §4.1 and §5.3], where  $\gamma$  and  $\beta$  are tuned online in such a way to ensure the needed sufficient decrease conditions and preserve convergence.

#### 3.4.1 Adaptive Algorithm 1

For the DRS-based Algorithm 1, it suffices to initialize  $\gamma$  according to an estimate of  $L_{\varphi_1}$  or  $\mu_{\varphi_1}$ , and simply add the following routine after step 1.2:

- 1.2a: Evaluate  $\varphi_\gamma^{\text{DR}}(\bar{s}^{k+1})$  using  $(\bar{u}^{k+1}, \bar{v}^{k+1}) \in \text{DRS}_\gamma(\bar{s}^{k+1})$  as in (2.2)  
 1.2b: IF  $\pi\varphi_\gamma^{\text{DR}}(\bar{s}^{k+1}) \geq \pi\varphi_\gamma^{\text{DR}}(s^k) - \frac{\varepsilon}{\gamma} \|r^k\|^2$  OR  $\pi\varphi_\gamma^{\text{DR}}(\bar{s}^{k+1}) < \pi\varphi_{\text{LB}}$  THEN  
      $\gamma \leftarrow 2^{-\pi}\gamma$ ,  $(u^k, v^k) \in \text{DRS}_\gamma(s^k)$ , recompute  $\varphi_\gamma^{\text{DR}}(s^k)$  and go to step 1.1.

Here, recall that  $\pi \in \{\pm 1\}$  is set at algorithm initialization according to whether Assumption I or Assumption I\* holds. The only new term is  $\varphi_{\text{LB}}$ , to be set offline equal to a known quantity that lower bounds  $\inf \varphi$ , which in practice is typically easily estimable. Its role is however a pure technicality that the not-too-fussy user can neglect; the interested reader can instead find the reasoning for the additional condition it enforces in [51, §4.1]. As documented in the reference, this backtracking on  $\gamma$  can happen only a finite number of times, as eventually the condition at step 1.2b

is never satisfied. In fact, under [Assumption I](#), by halvening  $\gamma$  enough yet finitely many times, the value will eventually fall under the (unknown) threshold prescribed by (2.6). Similarly, under [Assumption I\\*](#), by doubling  $\gamma$  finitely many times it will eventually exceed the (unknown) threshold  $1/\mu_{\varphi_1}$  prescribed in [Corollary 2.4](#). Either way, [Fact 2.2](#) and [Corollary 2.4](#) will guarantee that no more changes to  $\gamma$  will be triggered after that point.

### 3.4.2 Adaptive Algorithm 2

The same arguments as in the previous paragraph apply to the ADMM-based [Algorithm 2](#), in which case it suffices to initialize  $\beta$  according to an estimate of  $L_{(A_f)}$  or  $\mu_{(A_f)}$ , setting offline a lower bound  $\Phi_{\text{LB}} \leq \inf \{f(x) + g(z) \mid Ax + Bz = b\}$  (if known), and adding the following check after [step 2.2](#):

- 2.2a: Evaluate  $\mathcal{L}_\beta(\bar{x}^{k+1}, \bar{z}^{k+1}, \bar{y}^{k+1})$ , where  $(\bar{x}^{k+1}, \bar{z}^{k+1}, \bar{y}^{k+1}) \in \text{ADMM}_\beta(\bar{y}^{k+1/2}, z^k)$   
 2.2b: IF  $\pi \mathcal{L}_\beta(\bar{x}^{k+1}, \bar{z}^{k+1}, \bar{y}^{k+1}) \geq \pi \mathcal{L}_\beta(x^k, z^k, y^k) - \beta c \|r^k\|^2$  OR  $\pi \mathcal{L}_\beta(\bar{x}^{k+1}, \bar{z}^{k+1}, \bar{y}^{k+1}) < \pi \Phi_{\text{LB}}$  THEN  
 $\beta \leftarrow 2\pi\beta$ ,  $(x^k, z^k, y^k) \in \text{ADMM}_\beta(y^{k-1/2}, z^{k-1})$ , recompute  $\mathcal{L}_\beta(x^k, z^k, y^k)$  and go to [step 2.1](#).

## 4 Convergence results

This section is dedicated to the convergence properties of the proposed [Algorithms 1](#) and [2](#). We begin by addressing their well definedness, namely, that the iterations cannot get stuck in an infinite backtracking loop at [steps 1.8](#) and [2.8](#). We also show that for any strictly positive tolerance  $\varepsilon > 0$  the termination criterion is satisfied after finitely many iterations, and provide properties of the output quantities.

**Theorem 4.1** (Well definedness and finite termination of [Algorithm 1](#)). *Suppose that either [Assumption I](#) or [Assumption I\\*](#) is satisfied. Then, the following hold for the iterates generated by [Algorithm 1](#):*

- (i) *At every iteration the number of backtrackings at [step 1.8](#) is finite (regardless of whether  $i_{\max}$  is finite or not).*
- (ii) *The algorithm terminates in  $K \leq \frac{1}{c\gamma\varepsilon^2} |\varphi_\gamma^{\text{DR}}(s^0) - \min \varphi|$  iterations.*
- (iii) *The last iterate of the algorithm yields*
  - *a point  $z := v^K$  satisfying  $\text{dist}(0, \hat{\partial}\varphi(z)) \leq 2\varepsilon$  if [Assumption I](#) holds,*
  - *a triplet  $(x, y, z) := (u^K, \gamma^{-1}(u^K - s^K), v^K)$  satisfying the approximate KKT*

$$-y \in \partial\varphi_1(x), \quad \text{dist}(y, \partial\varphi_2(z)) \leq \varepsilon, \quad \|x - z\| \leq \gamma\varepsilon,$$

*if [Assumption I\\*](#) holds.*

*Proof.* We first consider the case in which [Assumption I](#) holds, so that  $\pi = 1$  and consequently  $\pi\varphi_\gamma^{\text{DR}} = \varphi_\gamma^{\text{DR}}$ .

♠ **4.1(i)** Testing the condition at [step 1.6](#) assumes  $\|r^k\| > 0$ , for otherwise the entire algorithm would have stopped at [step 1.1](#). Let  $C = C(\gamma L_{\varphi_1}, \lambda)$  be as in (2.5), so that the nominal DRS-update  $\bar{s}^{k+1}$  satisfies

$$\pi\varphi_\gamma^{\text{DR}}(\bar{s}^{k+1}) \leq \pi\varphi_\gamma^{\text{DR}}(s^k) - \frac{C}{\gamma} \|r^k\|^2,$$



as shown in [Fact 2.2](#). Since  $c < C$  by algorithm initialization, one has

$$\pi\varphi_\gamma^{\text{DR}}(s^k) - \frac{c}{\gamma}\|r^k\|^2 > \pi\varphi_\gamma^{\text{DR}}(s^k) - \frac{C}{\gamma}\|r^k\|^2.$$

Continuity of  $\varphi_\gamma^{\text{DR}}$  (cf. [Fact 2.1](#)) at  $\bar{s}^{k+1}$  thus entails the existence of  $\epsilon > 0$  such that  $\pi\varphi_\gamma^{\text{DR}}(s) \leq \pi\varphi_\gamma^{\text{DR}}(s^k) - \frac{c}{\gamma}\|r^k\|^2$  holds for every point  $s$   $\epsilon$ -close to  $\varphi_\gamma^{\text{DR}}(s^k)$ . Since  $(1 - \tau_k)\bar{s}^{k+1} + \tau_k(s^k + d^k) \rightarrow \bar{s}^{k+1}$  as  $\tau_k \rightarrow 0$ , by halvening enough times  $\tau_k$  at [step 1.8](#),

- either the candidate point  $s^{k+1}$  is eventually  $\epsilon$ -close to  $\bar{s}^{k+1}$  so that the needed condition at [step 1.6](#) holds,
- or the maximum number of backtrackings  $i_{\max}$  is reached (provided that  $i_{\max}$  is finite), in which case  $s^{k+1} = \bar{s}^{k+1}$ .

Either way, only a finite number of backtrackings is performed, and the inequality

$$\pi\varphi_\gamma^{\text{DR}}(s^{k+1}) \leq \pi\varphi_\gamma^{\text{DR}}(s^k) - \frac{c}{\gamma}\|r^k\|^2 \quad (4.1)$$

holds for every  $k \in \mathbb{N}$ .

♣ [4.1\(ii\)](#) By combining the decrease condition (4.1) with the termination criterion at [step 1.1](#), we have that  $K$  iterations of the algorithm result in a decrease of the DRE by at least

$$\pi\varphi_\gamma^{\text{DR}}(s^{K+1}) \leq \pi\varphi_\gamma^{\text{DR}}(s^0) - \frac{c}{\gamma} \sum_{k=0}^K \|r^k\|^2.$$

Since  $\pi \min \varphi = \min \pi\varphi_\gamma^{\text{DR}}$  (cf. [Fact 2.1\(i\)](#)) and  $\|r^k\| \geq \gamma\epsilon$  for  $k < K$ , we conclude that

$$\pi\varphi_\gamma^{\text{DR}}(s^0) - \pi \min \varphi \geq \frac{c}{\gamma} \sum_{k=0}^{K-1} \|r^k\|^2 \geq cK\gamma\epsilon^2, \quad (4.2)$$

resulting in the claimed bound on  $K$ . Finally, the bound on  $\text{dist}(0, \hat{\partial}\varphi(z))$  follows from [\[51, Thm. 4.3\(ii\)\]](#).

Suppose now that [Assumption I\\*](#) holds. Then, since  $\pi = -1$ , it follows from [Theorem 2.3](#) that  $\pi\varphi_\gamma^{\text{DR}}(s) = \psi_{\gamma s_*}^{\text{DR}}(s_*)$ , where  $\gamma_* = 1/\gamma$  and  $s_* = -s/\gamma$ . Moreover, one has that  $u - v = \frac{u_* - v_*}{\gamma_*}$ , where  $u_*$  and  $v_*$  are as in [Theorem 2.3](#). The previous arguments can thus be replicated in terms of the dual formulation (2.7) to prove assertion [4.1\(i\)](#) and the bound on the number of iterations  $K$ .

To conclude, let the triplet  $(x, y, z) = (u^K, \frac{u^K - s^K}{\gamma}, v^K)$  be as in the statement. From (1.10) one has that  $-y \in \partial\varphi_1(x)$  and that  $y + \frac{1}{\gamma}r^K \in \partial\varphi_2(z)$ , hence that  $\text{dist}(y, \partial\varphi_2(z)) \leq \epsilon$ , owing to the termination criterion  $\|r^K\| \leq \gamma\epsilon$ . Finally,  $\|x - z\| = \|r^K\| \leq \epsilon$ , completing the proof.  $\square$

**Theorem 4.2** (Well definedness and finite termination of [Algorithm 2](#)). *Suppose that either [Assumption II](#) or [Assumption II\\*](#) is satisfied. Then, the following hold for the iterates generated by [Algorithm 2](#):*

- (i) *At every iteration the number of backtrackings at [step 2.8](#) is finite (regardless of whether  $i_{\max}$  is finite or not).*
- (ii) *The algorithm terminates in  $K \leq \frac{\beta}{c\epsilon^2} |\mathcal{L}_\beta(x^0, z^0, y^0) - \min \Phi|$  iterations and yields a triplet  $(x, y, z) := (x^K, y^K, z^K)$  satisfying the approximate KKT conditions*

$$\|Ax + Bz - b\| \leq \epsilon/\beta, \quad -A^\top y \in \hat{\partial}f(x), \quad \text{dist}(-B^\top y, \hat{\partial}g(z)) \leq \|B\|\epsilon.$$



*Proof.* Assertion 4.2(i) and the bound on the number of iterations  $K$  follow from Theorem 4.1, owing to the equivalence of Algorithms 1 and 2 stated in Proposition 3.1. In turn, the conditions  $-A^\top y \in \hat{\partial}f(x)$  and  $\text{dist}(-B^\top y, \hat{\partial}g(z)) \leq \|B\|\varepsilon$  follow from Facts 2.5(i) and 2.5(ii), since for every  $k$  and independently of the choice of the direction  $d^k$  the triplet  $(x^k, y^k, z^k)$  is the result of an ADMM-step with penalty  $\beta$ .  $\square$

#### 4.1 Subsequential convergence

The remainder of the section will focus on more theoretical aspects of the algorithms, and specifically on asymptotic behaviors. To this end, an idealistic tolerance  $\varepsilon = 0$  will be considered, so that the algorithms may run infinitely many iterations. In this first subsection, without imposing additional requirements other than either one among Assumptions I, I\*, II, or II\*, we show that every limit point of the generated sequences is stationary, and also give a sufficient condition ensuring boundedness. Under additional assumptions, we will later demonstrate and rigorously define the speed-up triggered by suitable choices of directions, which will ultimately be backed up by numerical evidence in Section 5. Some results are based on auxiliary material presented in Appendix A.

**Theorem 4.3** (Subsequential convergence of Algorithm 1). *Suppose that either Assumption I or Assumption I\* is satisfied, and consider the (possibly infinite) iterates generated by Algorithm 1 with tolerance  $\varepsilon = 0$ . Then, the sequence of squared residuals  $(\|r^k\|^2)_{k \in \mathbb{N}}$  has finite sum. Moreover,*

UNDER ASSUMPTION I:

- (i) *The sequences  $(u^k)_{k \in \mathbb{N}}$  and  $(v^k)_{k \in \mathbb{N}}$  have the same cluster points, all of which are stationary for  $\varphi$  and on which  $\varphi$  and  $\varphi_\gamma^{\text{DR}}$  have the same (constant) value, this being the limit of the monotonic sequence  $(\varphi_\gamma^{\text{DR}}(s^k))_{k \in \mathbb{N}}$ .*
- (ii) *If  $\varphi$  is level bounded, then  $(s^k, u^k, v^k)_{k \in \mathbb{N}}$  remains bounded.*

UNDER ASSUMPTION I\*:

- (iii)  *$\varphi_\gamma^{\text{DR}}(s^k) \nearrow \min \varphi$  as  $k \rightarrow \infty$ , and  $(u^k)_{k \in \mathbb{N}}$  and  $(v^k)_{k \in \mathbb{N}}$  converge to the unique (global) minimizer  $x_\star$  of  $\varphi$ .*
- (iv) *If  $0 \in \text{int}(\text{dom } \varphi_1 - \text{dom } \varphi_2)$  (equivalently, if the dual cost  $\psi$  as in (2.7) is level bounded), then  $(s^k)_{k \in \mathbb{N}}$  is bounded.*

*Proof.* To rule out trivialities we may assume that the stopping criterion  $\|r^k\| = 0$  is never reached, hence that the algorithm generates infinitely many iterates. That the squared residuals  $(\|r^k\|^2)_{k \in \mathbb{N}}$  have finite sum follows by letting  $K \rightarrow \infty$  in (4.2). In turn, this implies that  $u^k - v^k \rightarrow 0$  as  $k \rightarrow \infty$ . We now consider the two separate cases.

♠ **ASSUMPTION I.** Having shown the vanishing of the residual, the proof is identical to that of [51, Thm. 4.3(ii)-(iii)].

♠ **ASSUMPTION I\*.** In this case, the dual formulation (2.7) satisfies Assumption I. In particular, adopting the  $*$ -notation of Theorem 2.3 we have that  $\psi_{\gamma_*}^{\text{DR}}(s^k) = -\varphi_\gamma^{\text{DR}}(s^k)$  is decreasing and bounded below by  $\inf \psi = -\min \varphi$  (cf. (A.2)), and  $u_*^k - v_*^k = \frac{v^k - u^k}{\gamma} \rightarrow 0$ . Note that inequality (A.3) guarantees that  $(u^k)_{k \in \mathbb{N}}$  and  $(v^k)_{k \in \mathbb{N}}$  are bounded; in what

follows, we will make use of [Assumption I\\*.A3](#) to show that they actually converge to the global minimizer of  $\varphi$ . To this end, it will suffice to show that  $\psi(v_*^k) \rightarrow \inf \psi$ , since

$$\psi_{\gamma_*}^{\text{DR}}(s_*^k) - \inf \psi \leq \psi(v_*^k) + \frac{1+\gamma_* L_{\psi_1}}{2\gamma_*} \|u_*^k - v_*^k\|^2 - \inf \psi = \psi(v_*^k) + \frac{1+\mu_{\varphi_1} \gamma}{2\mu_{\varphi_1}} \|u_*^k - v_*^k\|^2 - \inf \psi$$

and  $\|u_*^k - v_*^k\| \rightarrow 0$ , where the inequality uses  $\bar{u} = v_*^k$  in [Lemma A.2](#) applied to the dual formulation (2.7). We start by observing that (1.10) implies that  $\frac{2u_*^k - s_*^k - v_*^k}{\gamma_*} \in \partial\psi_2(v_*^k)$  and  $\frac{s_*^k - u_*^k}{\gamma_*} = \nabla\psi_1(u_*^k)$ , which together result in  $\partial\psi(v_*^k) \ni \frac{u_*^k - v_*^k}{\gamma_*} - (\nabla\psi_1(u_*^k) - \nabla\psi_1(v_*^k))$ , owing to the calculus rule of [42, Thm. 23.8]. Therefore,  $\text{dist}(0, \partial\psi(v_*^k)) \leq \|u_*^k - v_*^k\| \rightarrow 0$ . Adopting the terminology of [3, Def. 4.2.1], this means that  $(v_*^k)_{k \in \mathbb{N}}$  is a *stationary sequence* for  $\psi$ . We will now show that  $\psi$  is *asymptotically well behaved*, so that stationarity of  $(v_*^k)_{k \in \mathbb{N}}$  guarantees that  $\psi(v_*^k) \rightarrow \inf \psi$ , cf. [3, Def. 4.2.2]. To this end, observe that  $\psi^* = ((\mathcal{R}\varphi_1)^* + \varphi_2^*)^* = (\mathcal{R}\varphi_1) \square g$ , see [4, Prop. 13.24(i) and Thm. 13.37], where  $\square$  denotes the infimal convolution operator. It then follows from [4, Prop. 12.6(ii)] that  $\text{dom } \psi^* = \text{dom}(\mathcal{R}\varphi_1) + \text{dom } \varphi_2 = \text{dom } \varphi_2 - \text{dom } \varphi_1$ . The validity of [Assumption I\\*.A3](#) then ensures through [3, Thm. 3.2.1(a)-(d) and Cor. 4.2.1(a)] the sought asymptotic well behavior of  $\psi$ , hence that  $\psi(v_*^k) \rightarrow \inf \psi = -\min \varphi$ . This concludes the proof of assertion 4.3(iii).

Finally, the equivalence of the condition  $0 \in \text{int}(\text{dom } \varphi_1 - \text{dom } \varphi_2)$  and level boundedness of the dual cost  $\psi$  is shown in [3, Thm. 5.2.1(a)]. Assertion 4.3(ii) applied to the dual formulation then ensures boundedness of  $(s_*^k)_{k \in \mathbb{N}}$ , and that of  $(s^k)_{k \in \mathbb{N}}$  then follows from the identity  $s_k = -s_*^k/\gamma_*$ .  $\square$

**Theorem 4.4** (Subsequential convergence of [Algorithm 2](#)). *Suppose that either [Assumption II](#) or [Assumption II\\*](#) is satisfied, and consider the (possibly infinite) iterates generated by [Algorithm 2](#) with tolerance  $\varepsilon = 0$ . Then, the sequence of squared residuals  $(\|r^k\|^2)_{k \in \mathbb{N}}$  has finite sum. Moreover,*

UNDER [ASSUMPTION II](#):

- (i) All cluster points  $(x, y, z)$  of  $(x^k, y^k, z^k)_{k \in \mathbb{N}}$  satisfy the KKT conditions

$$-A^\top y \in \partial f(x), \quad -B^\top y \in \partial g(z), \quad Ax + Bz = b, \quad (4.3)$$

and attain the same (finite) cost  $f(x) + g(z)$ , this being the limit of the sequence  $(\mathcal{L}_\beta(x^k, z^k, y^k))_{k \in \mathbb{N}}$ .

- (ii) The sequence  $(Ax^k, y^k, Bz^k)_{k \in \mathbb{N}}$  is bounded provided that the cost function  $\Phi$  is level bounded. If, additionally,  $f \in C^{1,1}(\mathbb{R}^m)$ , then the sequence  $(x^k, y^k, z^k)_{k \in \mathbb{N}}$  is bounded.

UNDER [ASSUMPTION II\\*](#):

- (iii)  $(\mathcal{L}_\beta(x^k, z^k, y^k))_{k \in \mathbb{N}} \nearrow \min \Phi$  as  $k \rightarrow \infty$ ,  $(Ax^k)_{k \in \mathbb{N}}$  and  $(Bz^k)_{k \in \mathbb{N}}$  are convergent, and all the cluster points  $(x, y, z)$  of  $(x^k, y^k, z^k)_{k \in \mathbb{N}}$  are KKT-optimal triplets (satisfying (4.3)).
- (iv) If  $b \in \text{int}(A \text{ dom } f + B \text{ dom } g)$  (equivalently, if the dual cost  $\Psi(y) := f^*(-A^\top y) + g^*(-B^\top y) + \langle b, y \rangle$  is level bounded), then  $(x^k, y^k, z^k)_{k \in \mathbb{N}}$  remains bounded.

*Proof.* We shall again invoke the equivalence of Algorithms 1 and 2 stated in Proposition 3.1, and import the same notation for convenience. Note that no clash occurs in using  $r^k$  to express the residual in both algorithms, having  $r^k = u^k - v^k = Ax^k + Bz^k - b$  through the relations of Proposition 3.1. In particular, that the squared residuals  $(\|r^k\|^2)_{k \in \mathbb{N}}$  have finite sum is a direct consequence of Theorem 4.3. We now consider the two separate cases.

♣ **ASSUMPTION II.** Having shown the vanishing of the residual, the proof is identical to that of [51, Thm. 5.6].

♣ **ASSUMPTION II\*.** Notice that  $\text{dom } \varphi_1 - \text{dom } \varphi_2 = A \text{ dom } f + B \text{ dom } g - b$  for  $\varphi_1$  and  $\varphi_2$  as in (2.10), owing to the fact that  $\text{dom}(Af) = A \text{ dom } f$  and  $\text{dom}(Bg)(b - \cdot) = b - B \text{ dom } g$ , see [4, Prop. 12.36(i)]. Notice further that

$$\varphi_1^* = f(-A^\top \cdot) \quad \text{and} \quad \varphi_2^* = g(-B^\top \cdot) + \langle b, \cdot \rangle,$$

see [4, Props. 13.23(iii) and 13.24(iv)]. Therefore, it suffices to show that any limit triplet  $(x, y, z)$  is KKT-optimal, for all other claims are direct translations of Theorem 4.3(iii). To this end, suppose that a subsequence  $(x^k, y^k, z^k)_{k \in K}$  converges to  $(x, y, z)$ . Since  $r^k \rightarrow 0$ , necessarily  $Ax + Bz = b$ . Moreover,

$$\inf \Phi = \lim_{K \ni k \rightarrow \infty} \mathcal{L}_\beta(x^k, z^k, y^k) = \lim_{K \ni k \rightarrow \infty} f(x^k) + g(z^k) \geq f(x) + g(z) = \Phi(x, z) \geq \inf \Phi,$$

where the second equality follows from the fact that  $(y^k)_{k \in K}$  is bounded (since it converges), the first inequality from lsc of  $f$  and  $g$ , and the last equality from the fact that  $Ax + Bz - b = 0$ . In fact, notice that having  $f(x^k) + g(z^k) \rightarrow f(x) + g(z)$  implies that  $f(x^k) \rightarrow f(x)$  and  $g(z^k) \rightarrow g(z)$ . Therefore, the continuity of the convex subdifferential [42, Thm. 24.4] together with Facts 2.5(i) and 2.5(ii) implies the sought inclusions  $-A^\top y \in \partial f(x)$  and  $-B^\top y \in \partial g(z)$ , yielding the claimed KKT-optimality.  $\square$

## 4.2 Superlinear convergence

We now provide sufficient conditions that ensure superlinear convergence of the proposed algorithms. For the sake of simplicity we limit the analysis to Algorithm 1 under Assumption I; the other cases can be inferred through the equivalence between Algorithm 1 and Algorithm 2 and the self-duality of DRS proven in Theorem 2.3 and Proposition 3.1.

As a measure of “quality” of the oracle producing the update directions, borrowing the terminology of [15, §7.5], we say that  $(d^k)_{k \in \mathbb{N}}$  is a sequence of *superlinear directions* (relative to a sequence  $(s^k)_{k \in \mathbb{N}}$  converging to  $s_\star$ ) if

$$\lim_{k \rightarrow \infty} \frac{\|s^k + d^k - s_\star\|}{\|s^k - s_\star\|} = 0. \quad (4.4)$$

The next theorem shows that whenever the algorithm converges to a strong local minimum and the directions comply with the qualitative criterion (4.4), the linesearch condition at step 1.6 is eventually always passed at the first trial, and the iterates reduce to  $s^{k+1} = s^k + d^k$  and converge superlinearly. To do so, we will use the following lemma showing how strong local minimality on the original cost reflects on the DRE.

**Lemma 4.5.** *Suppose that [Assumption 1](#) holds, and consider the iterates generated by [Algorithm 1](#) with  $\varepsilon = 0$ , and let  $\varphi_\star$  be the limit of the sequence  $(\varphi_\gamma^{\text{DR}}(s^k))_{k \in \mathbb{N}}$ . Suppose that  $(u^k)_{k \in \mathbb{N}}$  converges to a strong local minimum  $u_\star$  of  $\varphi$ . Then, there exists  $\delta > 0$  such that  $\varphi_\gamma^{\text{DR}}(s^k) - \varphi_\star \geq \frac{\delta}{2} \|s^k - s_\star\|^2$  holds for every  $k$  large enough, where  $s_\star := u_\star + \gamma \nabla \varphi_1(u_\star)$ .*

*Proof.* We begin by observing that for every  $k \in \mathbb{N}$  and  $\epsilon > 0$  it holds that

$$\frac{1}{(1+\gamma L_{\varphi_1})^2} \|s^k - s_\star\|^2 \leq \|u^k - u_\star\|^2 \leq \frac{\epsilon+1}{\epsilon} \|u^k - v^k\|^2 + (1+\epsilon) \|v^k - u_\star\|^2, \quad (4.5)$$

where the first inequality follows from  $\frac{1}{1+\gamma L_{\varphi_1}}$ -strong monotonicity of  $\text{prox}_{\gamma \varphi_1}$  [[51](#), Prop. 2.3(ii)] and the fact that  $u_\star = \text{prox}_{\gamma \varphi_1}(s_\star)$  as it follows from [\(1.10\)](#), and the second one uses Young's inequality. As ensured by [Theorem 4.3\(i\)](#),  $u^k - v^k \rightarrow 0$  and  $\varphi(u_\star) = \varphi_\star$ . Therefore,  $v^k \rightarrow u_\star$  and because of strong local minimality there exists  $\mu > 0$  and  $K \in \mathbb{N}$  such that  $\varphi(v^k) - \varphi_\star \geq \frac{\mu}{2} \|v^k - u_\star\|^2$  for all  $k \geq K$ . For all  $\epsilon > 0$  and  $k \geq K$  we thus have

$$\begin{aligned} \varphi_\gamma^{\text{DR}}(s^k) &\stackrel{\text{A.1}}{\geq} \varphi(v^k) + \frac{1-\gamma L_{\varphi_1}}{2\gamma} \|v^k - u^k\|^2 \\ &\geq \varphi_\star + \frac{\mu}{2} \|v^k - u_\star\|^2 + \frac{1-\gamma L_{\varphi_1}}{2\gamma} \|v^k - u^k\|^2 \\ &\stackrel{(4.5)}{\geq} \varphi_\star + \frac{\mu}{2(1+\epsilon)(1+\gamma L_{\varphi_1})^2} \|s^k - s_\star\|^2 + \left( \frac{1-\gamma L_{\varphi_1}}{2\gamma} - \frac{\mu}{2\epsilon} \right) \|v^k - u^k\|^2 \\ &= \varphi_\star + \frac{1-\gamma L_{\varphi_1}}{(1-\gamma(L_{\varphi_1}-\mu))(1+\gamma L_{\varphi_1})^2} \frac{\mu}{2} \|s^k - s_\star\|^2, \end{aligned}$$

where the last equality uses  $\epsilon = \frac{\gamma\mu}{1-\gamma L_{\varphi_1}}$ .  $\square$

**Theorem 4.6** (Acceptance of the unit stepsize). *Suppose that [Assumption 1](#) holds, and consider the iterates generated by [Algorithm 1](#). Suppose that  $(u^k)_{k \in \mathbb{N}}$  converges to a strong local minimum  $u_\star$  of  $\varphi$  and that  $(d^k)_{k \in \mathbb{N}}$  are superlinear directions as in [\(4.4\)](#). Then, eventually unit stepsize  $\tau_k = 1$  is always accepted, hence the iterates reduce to  $s^{k+1} = s^k + d^k$  and converge superlinearly.*

*Proof.* In light of [Lemma 4.5](#), by possibly discarding the first iterates we may assume that

$$\varphi_\gamma^{\text{DR}}(s^k) - \varphi_\gamma^{\text{DR}}(s_\star) \geq \frac{\delta}{2} \|s^k - s_\star\|^2 \quad \text{for all } k\text{'s}$$

for some  $\delta > 0$ , where  $s_\star := u_\star + \gamma \nabla \varphi_1(u_\star)$  is the limit point of  $(s^k)_{k \in \mathbb{N}}$ . Combined with [Lemma A.2](#), we obtain that

$$\epsilon_k := \frac{\varphi_\gamma^{\text{DR}}(s^k + d^k) - \varphi_\gamma^{\text{DR}}(s_\star)}{\varphi_\gamma^{\text{DR}}(s^k) - \varphi_\gamma^{\text{DR}}(s_\star)} \leq \frac{1 + \gamma L_f}{\gamma \delta} \frac{\|s^k + d^k - s_\star\|^2}{\|s^k - s_\star\|^2} \rightarrow 0.$$

Since  $u^k - v^k \rightarrow 0$ , we have that  $\bar{s}^k \rightarrow s_\star$ . Therefore, eventually  $\varphi_\gamma^{\text{DR}}(\bar{s}^k) \geq \varphi_\gamma^{\text{DR}}(s_\star)$  and  $\epsilon_k \leq 1$ . Then, denoting  $C = C(\gamma L_{\varphi_1}, \lambda)$  as in [\(2.5\)](#) we have

$$\begin{aligned} \varphi_\gamma^{\text{DR}}(s^k + d^k) - \varphi_\gamma^{\text{DR}}(s^k) &= -(1 - \epsilon_k)(\varphi_\gamma^{\text{DR}}(s^k) - \varphi_\gamma^{\text{DR}}(s_\star)) \\ &\leq -(1 - \epsilon_k)(\varphi_\gamma^{\text{DR}}(s^k) - \varphi_\gamma^{\text{DR}}(\bar{s}^k)) \\ &\stackrel{(2.5)}{\leq} -(1 - \epsilon_k) \frac{C}{\gamma} \|r^k\|^2. \end{aligned}$$

Since  $\epsilon_k \rightarrow 0$  and  $c < C$  as required in [Algorithm 1](#), eventually  $1 - \epsilon_k \geq \frac{c}{C}$ , resulting in  $\varphi_\gamma^{\text{DR}}(s^k + d^k) \leq \varphi_\gamma^{\text{DR}}(s^k) - \frac{c}{\gamma} \|d^k\|^2$ , proving that  $\tau_k = 1$  passes the condition at [step 1.6](#).  $\square$

Although the [DRS](#)-update  $s \mapsto s^+$  is not uniquely determined owing to the multi-valuedness of  $\text{prox}_{\gamma\varphi_2}$ , under some regularity assumptions not only is it single valued, but even differentiable, when close to solutions. To see this, observe that

$$R_\gamma^{\text{DR}} = R_\gamma^{\text{FB}} \circ \text{prox}_{\gamma\varphi_1},$$

where  $R_\gamma^{\text{DR}} := \text{prox}_{\gamma\varphi_1} - \text{prox}_{\gamma\varphi_2}(2 \text{prox}_{\gamma\varphi_1} - \text{id})$  is the Douglas-Rachford residual and

$$R_\gamma^{\text{FB}}(u) := u - \text{prox}_{\gamma\varphi_2}(u - \gamma \nabla \varphi_1(u))$$

is the residual of *forward-backward splitting* (FBS), see e.g., [\[2, 9, 52\]](#). By combining [\[43, Prop. 13.24 and Ex. 13.35\]](#) and [\[38, Thm. 4.4\(c\)-\(f\) and Cor. 4.7\]](#), it follows that  $\text{prox}_{\gamma\varphi_1}(s)$  is continuously differentiable around  $s_\star$  provided that  $\varphi_1$  is twice continuously differentiable around  $s_\star$ . Thus, under this assumption, from the chain rule of differentiation we conclude that  $R_\gamma^{\text{DR}}$  is (strictly) differentiable at  $s_\star$  provided that  $R_\gamma^{\text{FB}}$  is (strictly) differentiable at  $u_\star = \text{prox}_{\gamma\varphi_1}(s_\star)$ . Sufficient conditions for this latter property to hold are documented in [\[52, Thm. 4.10\]](#), namely twice (Lipschitz-) continuous differentiability of  $\varphi_1$  around  $u_\star$ , and prox-regularity and (strict) twice epi-differentiability of  $\varphi_2$  at  $u_\star$  for  $-\nabla \varphi_1(u_\star)$  [\[43, §13.B and 13.F\]](#).

**Theorem 4.7** (Dennis-Moré criterion for superlinear directions). *Suppose that [Assumption 1](#) holds, and consider the iterates generated by [Algorithm 1](#). Suppose that  $(s^k)_{k \in \mathbb{N}}$  converges to a point  $s_\star$  at which  $R_\gamma^{\text{DR}}$  is strictly differentiable and with nonsingular Jacobian  $JR_\gamma^{\text{DR}}(s_\star)$ . If the Dennis-Moré condition*

$$\lim_{k \rightarrow \infty} \frac{\|R_\gamma^{\text{DR}}(s^k) + JR_\gamma^{\text{DR}}(s_\star)d^k\|}{\|d^k\|} = 0 \quad (4.6)$$

*holds, then  $(d^k)_{k \in \mathbb{N}}$  are superlinear directions and the claims of [Theorem 4.6](#) hold.*

*Proof.* Since  $JR_\gamma^{\text{DR}}$  exists and is nonsingular at  $s_\star$ , for  $k$  large enough  $R_\gamma^{\text{DR}}(s^k)$  is single valued and satisfies  $\|R_\gamma^{\text{DR}}(s^k)\| \geq \alpha \|s^k - s_\star\|$  for some  $\alpha > 0$ . Due to strict differentiability,

$$\lim_{k \rightarrow \infty} \frac{\|R_\gamma^{\text{DR}}(s^k + d^k) - R_\gamma^{\text{DR}}(s^k) - JR_\gamma^{\text{DR}}(s_\star)d^k\|}{\|d^k\|} = 0,$$

and from the Dennis-Moré condition (4.6) it then follows that  $\frac{\|R_\gamma^{\text{DR}}(s^k + d^k)\|}{\|d^k\|} \rightarrow 0$ . Since  $\|R_\gamma^{\text{DR}}(s^k + d^k)\| \geq \alpha \|s^k + d^k - s_\star\|$ , we conclude that  $\frac{\|s^k + d^k - s_\star\|}{\|d^k\|} \rightarrow 0$  too. Therefore,

$$\frac{\|s^k + d^k - s_\star\|}{\|s^k - s_\star\|} \leq \frac{\|s^k + d^k - s_\star\|}{\|d^k\| \left|1 - \frac{\|s^k + d^k - s_\star\|}{\|d^k\|}\right|} \rightarrow 0,$$

proving that  $(d^k)_{k \in \mathbb{N}}$  are superlinear directions with respect to  $(s^k)_{k \in \mathbb{N}}$ .  $\square$

We conclude by showing that the modified Broyden scheme described in §3.3.2b enables superlinear rates when some regularity requirements are met at the limit point. These include *Lipschitz semidifferentiability* of the residual  $R_\gamma^{\text{DR}}$ , a condition that entails classical differentiability at the limit point but not necessarily around it, see [25].

**Theorem 4.8** (Superlinear convergence with Broyden directions). *Suppose that Assumption 1 holds, and consider the iterates generated by Algorithm 1 with directions being selected with the modified Broyden method (§3.3.2b). Suppose that the sequence of Broyden matrices  $(H_k)_{k \in \mathbb{N}}$  is bounded, and that  $(s^k)_{k \in \mathbb{N}}$  converges to a strong local minimum  $s_\star$  of  $\varphi$  at which  $R_\gamma^{\text{DR}}$  is Lipschitz-semidifferentiable and has a nonsingular Jacobian  $JR_\gamma^{\text{DR}}(s_\star)$ . Then, the Dennis-Moré condition (4.6) holds, and in particular the unit stepsize  $\tau_k = 1$  is eventually always accepted and  $(s^k)_{k \in \mathbb{N}}$  converges superlinearly.*

*Proof.* Let  $p_k$  and  $q_k$  be as in (3.5a),  $H_k$  be as in (3.6), and denote  $G_\star := JR_\gamma^{\text{DR}}(s_\star)$ . Since  $R_\gamma^{\text{DR}}$  is differentiable at  $s_\star$  with nonsingular Jacobian and since  $s^k \rightarrow s_\star$ , for  $k$  large enough  $R_\gamma^{\text{DR}}(s^k)$  is single valued and satisfies  $\|R_\gamma^{\text{DR}}(s^k)\| \geq \alpha \|s^k - s_\star\|$  for some  $\alpha > 0$ . Moreover, it follows from [25, Lem. 2.2] that an  $L > 0$  exists such that

$$\begin{aligned} \frac{\|q_k - G_\star p_k\|}{\|p_k\|} &= \frac{\|R_\gamma^{\text{DR}}(s^k + d^k) - R_\gamma^{\text{DR}}(s^k) - G_\star d^k\|}{\|d^k\|} \leq L \max\{\|s^k + d^k - s_\star\|, \|s^k - s_\star\|\} \\ &\leq L(\|s^k - s_\star\| + \|d^k\|) \end{aligned} \quad (4.7)$$

for  $k$  large enough. Since  $\text{prox}_{\gamma\varphi_1}$  is  $\frac{1}{1-\gamma L_{\varphi_1}}$ -Lipschitz continuous [51, Prop. 2.3(ii)], denoting  $u_\star := \text{prox}_{\gamma\varphi_1}(s_\star)$  we have that

$$\|R_\gamma^{\text{DR}}(s^k)\|^2 \geq \alpha^2 \|s^k - s_\star\|^2 \geq \alpha^2 (1 - \gamma L_{\varphi_1})^2 \|u^k - u_\star\|^2 \stackrel{\text{A.2}}{\geq} \alpha' (\varphi_\gamma^{\text{DR}}(s^k) - \varphi_\star)$$

for some  $\alpha' > 0$ . Therefore,

$$\varphi_\gamma^{\text{DR}}(s^{k+1}) - \varphi_\star \leq \varphi_\gamma^{\text{DR}}(s^k) - \varphi_\star - c \|R_\gamma^{\text{DR}}(s^k)\|^2 \leq (1 - c\alpha') (\varphi_\gamma^{\text{DR}}(s^k) - \varphi_\star),$$

and invoking Lemma 4.5 we conclude that  $(\|s^k - s_\star\|)_{k \in \mathbb{N}}$  converges  $R$ -linearly and thus has finite sum. Since  $R_\gamma^{\text{DR}}$  is differentiable at  $s_\star$ , there exists  $\alpha'' > 0$  such that  $\|R_\gamma^{\text{DR}}(s^k)\| = \|R_\gamma^{\text{DR}}(s^k) - R_\gamma^{\text{DR}}(s_\star)\| \leq \alpha'' \|s^k - s_\star\|$ . Therefore, also  $(\|R_\gamma^{\text{DR}}(s^k)\|)_{k \in \mathbb{N}}$  has finite sum, and in turn so does  $(\|d^k\|)_{k \in \mathbb{N}}$  owing to boundedness of  $(H_k)_{k \in \mathbb{N}}$  and the fact that  $d^k = -H_k R_\gamma^{\text{DR}}(s^k)$ . From (4.7) we conclude that  $(\frac{\|q_k - G_\star p_k\|}{\|p_k\|})_{k \in \mathbb{N}}$  has finite sum as well, and the claimed Dennis-Moré condition follows by verbatim importing the conclusions of the proof of [50, Thm. VI.8], after observing that  $\frac{\|R_\gamma^{\text{DR}}(s^k) + G_\star d^k\|}{\|d^k\|} = \frac{\|(H_k^{-1} - G_\star)p_k\|}{\|p_k\|}$ . Finally, the acceptance of the unit stepsize and superlinear convergence of  $(s^k)_{k \in \mathbb{N}}$  follow from Theorems 4.6 and 4.7.  $\square$

## 5 Simulations

In this section we show the effectiveness of the proposed [Algorithms 1](#) and [2](#), for different choices of the linesearch direction, compared to the standard [DRS](#) and [ADMM](#). In all problems, the (generalized) quadratic structure of one component of the cost functions is exploited as described in [Remark 3.2](#), thus resulting in at most two evaluations of  $\text{prox}_{\gamma\varphi_1}$  for every iteration of [Algorithm 1](#), or  $x$ -minimizations for every iteration of [Algorithm 2](#). The implementations in Julia of all the algorithms are available online as part of the *ProximalAlgorithms.jl* package.<sup>4</sup> All experiments were run using Julia 1.6.3.

### 5.1 Nonconvex sparse least squares

To find a sparse, least-squares solution  $x \in \mathbb{R}^n$  to a linear system  $Ax = b$ , we consider the formulation

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|Ax - b\|^2 + r \|x\|_{1/2}^{1/2}, \quad (5.1)$$

where  $A \in \mathbb{R}^{m \times n}$ ,  $\|x\|_{1/2}^{1/2} = \sum_i \sqrt{|x_i|}$  is the square root of the  $\ell_{1/2}$  quasi-norm, and  $r > 0$  is a regularization parameter. The  $\ell_{1/2}$  penalty term has favorable properties compared to the popular  $\ell_1$  regularization, as thoroughly documented in [\[54, Sec. II\]](#); yet its nonconvexity makes problem (5.1) more challenging to solve. As derived in [\[54\]](#), the proximal operator for the regularization term can be computed in closed form as follows

$$\left[ \text{prox}_{\gamma\|\cdot\|_{1/2}^{1/2}}(x) \right]_i = \begin{cases} \frac{1}{2} \left( 1 + \cos \frac{2}{3} \left( \pi - \arccos \frac{\gamma}{4} \left( \frac{x_i}{3} \right)^{-3/2} \right) \right) x_i & \text{if } |x_i| > \frac{3}{2} \gamma^{2/3}, \\ 0 & \text{otherwise.} \end{cases} \quad (5.2)$$

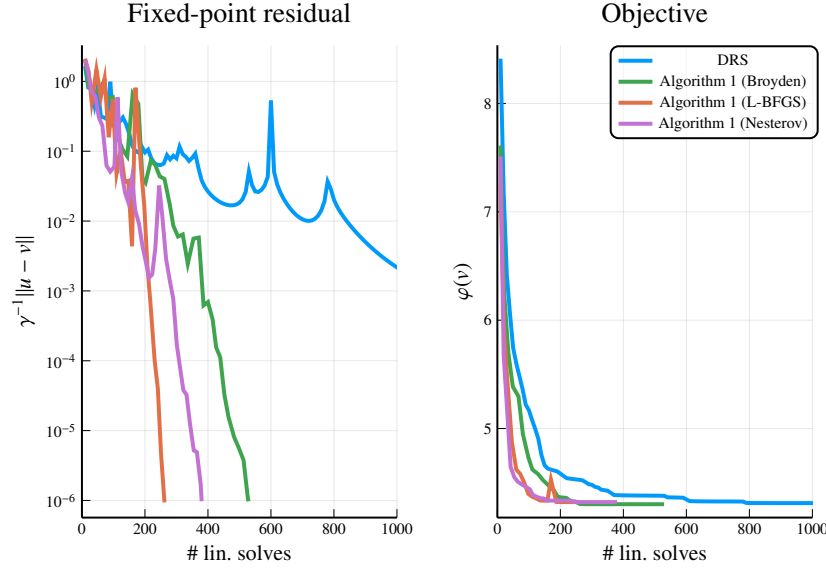
For the least squares term,  $\text{prox}_{\gamma/2\|A \cdot - b\|^2}(x) = (A^T A + \gamma^{-1} I)^{-1} (A^T b + \gamma^{-1} x)$  involves solving a positive definite linear system, and is therefore the most computationally demanding operation.

We generated random instances of problem (5.1) similarly to the setup of [\[13, Sec. 8.2\]](#): matrix  $A$  has i.i.d. Gaussian entries with variance  $1/m$ , while  $b = A\hat{x}$  for a random vector  $\hat{x}$  with  $k$  nonzero coefficients. In the [DRS-Algorithm 1](#), we used  $\lambda = 1$ ,  $\gamma = 0.95 \cdot L_{\varphi_1}^{-1}$ ,  $c = \frac{1}{2} C(\gamma L_{\varphi_1}, \lambda)$ , cf. (2.5) and (2.6).

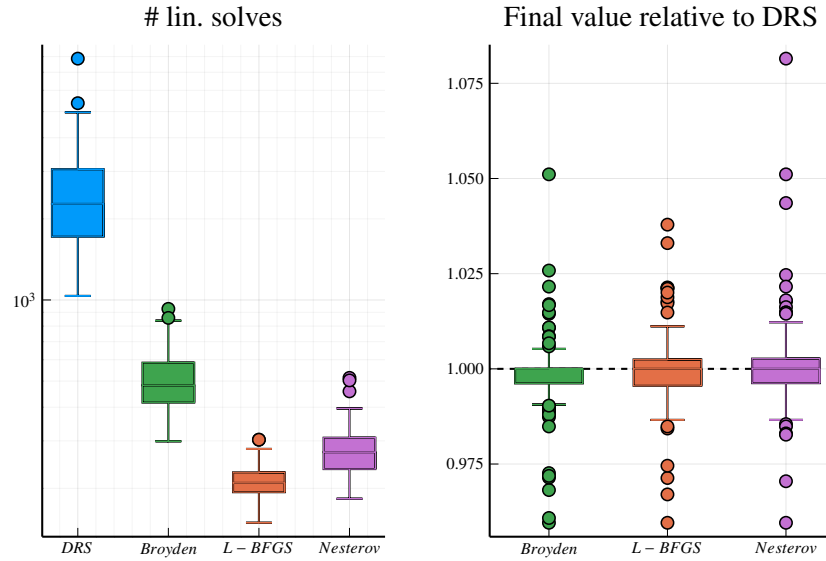
The convergence of the proposed [Algorithm 1](#) compared to [DRS](#) is exemplified in [Figure 2a](#), where the algorithms were applied on a randomly generated problem instance with  $n = 500$ ,  $m = 100$ , and  $k = 50$ . For this experiment, in [Algorithm 1](#) we used Broyden, L-BFGS (with memory 5) and Nesterov directions, cf. [Section 3.3](#): these choices of directions significantly accelerate convergence, compared to [DRS](#). Using Anderson acceleration directions, in this case, did not perform well. In [Figure 2b](#), the result of running the algorithms on 100 randomly generated problems is displayed: there, the distribution of the number of linear solves (i.e. evaluations of  $\text{prox}_{\gamma/2\|A \cdot - b\|^2}$ ) over all problems is shown for each algorithm, as well as the distribution of the best objective values reached relative to the one obtained by [DRS](#). It is

<sup>4</sup> <https://github.com/JuliaFirstOrder/ProximalAlgorithms.jl>

**Figure 2:** §5.1: nonconvex sparse least squares problem (5.1). Comparison between *DRS* and the line-search variant *Algorithm 1* using modified Broyden, *L-BFGS*, and Nesterov acceleration directions.



(a) Comparison of the methods on one randomly generated problem instance: matrix  $A \in \mathbb{R}^{100 \times 500}$  has random coefficients with variance 0.01, while  $b$  was generated from a known sparse  $\hat{x}$  with 50 nonzero coefficients, and  $r = 0.1$ . On the x-axis, the number of linear systems solved (needed for the  $u$ -update): in the case of *DRS*, this coincides with the number of iterations, while for *Algorithm 1* it accounts for all operations performed in the linesearch.



(b) Comparison of the methods on 100 randomly generated problems. Left: box plot of the number of operations to reach a fixed-point residual of norm  $10^{-6}$ . Right: box plot of the objective value achieved, relative to the one obtained by *DRS*. The boxes represent the interquartile range (the range between the P25 and P75 percentiles), with the median highlighted; the bars extend below to  $P25 - 1.5(P75 - P25)$ , and above to  $P75 + 1.5(P75 - P25)$ ; all observations outside of this range are displayed as stand-alone points.



apparent how [Algorithm 1](#) converges to critical points in a fraction of the operations required by [DRS](#).

## 5.2 Sparse PCA

Given a dataset of points in  $\mathbb{R}^n$ , the goal of *sparse principal component analysis* (SPCA) is to explain as much variability in the data as possible by using only  $k \ll n$  variables. Let the data matrix be  $W \in \mathbb{R}^{m \times n}$  (this can be assumed to be centered, i.e., with zero-mean columns), then the problem can be formulated as follows:

$$\underset{x \in \mathbb{R}^n}{\text{maximize}} \quad \frac{1}{2m} \|Wx\|^2 \quad \text{subject to } \|x\| = 1, \|x\|_0 \leq k, \quad (5.3)$$

where the  $\ell_0$ -quasi-norm  $\|x\|_0$  denotes the number of nonzero elements of vector  $x$ . Being  $\frac{1}{m} W^\top W \in \mathbb{R}^{n \times n}$  the covariance matrix of  $W$ , (5.3) amounts to a variance maximization problem. The formulation (5.3) was first introduced in [12], where the authors propose solving an SDP relaxation. Here, we consider tackling directly the non-convex problem (5.3) instead.

Denoting the set of feasible points by

$$\mathcal{S} := \{x \in \mathbb{R}^n \mid \|x\| = 1, \|x\|_0 \leq k\},$$

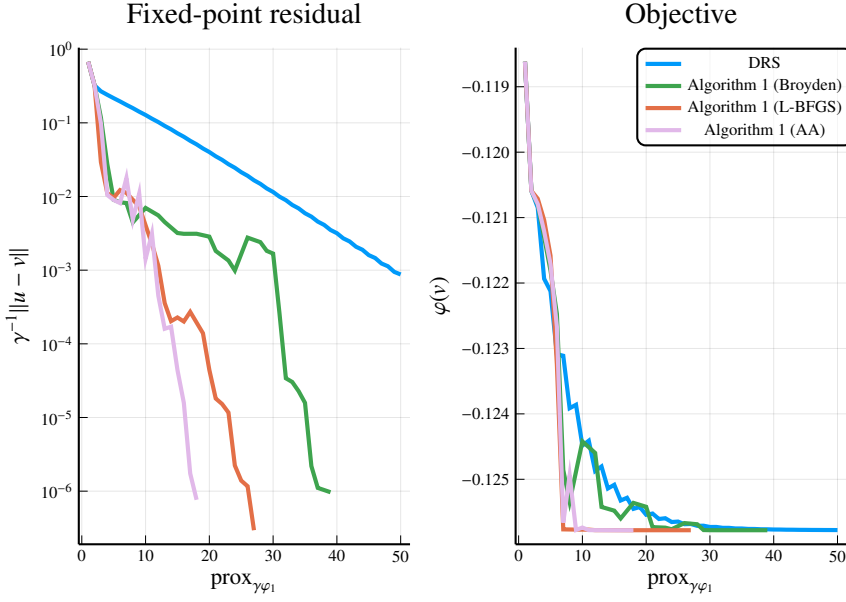
problem (5.3) takes the form (1.1) once we set  $\varphi_1(x) = -\frac{1}{2m} \|Wx\|^2$  and  $\varphi_2(x) = \delta_{\mathcal{S}}(x)$  (the *indicator function* of  $\mathcal{S}$ , namely  $\delta_{\mathcal{S}}(x) = 0$  if  $x \in \mathcal{S}$  and  $\delta_{\mathcal{S}}(x) = \infty$  otherwise). So formulated, the problem complies with [Assumption I](#), therefore [DRS](#) can be readily applied. Note that, in this case,

$$\begin{aligned} \text{prox}_{\gamma\varphi_1}(x) &= \arg \min_z \left\{ -\frac{1}{2m} \|Wz\|^2 + \frac{1}{2\gamma} \|z - x\|^2 \right\} = \left( \mathbf{I} - \frac{\gamma}{m} W^\top W \right)^{-1} x \\ &= x - W^\top \left( \frac{m}{\gamma} \mathbf{I} - WW^\top \right)^{-1} Wx, \end{aligned}$$

where the last equality uses the Woodbury identity. Whenever  $\gamma < L_{\varphi_1}^{-1} = \|W\|^2/m$ , evaluating  $\text{prox}_{\gamma\varphi_1}$  amounts to solving a square positive definite linear system of dimension either  $n$  or  $m$ , depending on which one is smaller: this can be done by computing the Cholesky factor offline once, and caching the factorization throughout the iterations. Recall that, since  $\varphi_1$  is quadratic, as illustrated in [Remark 3.2](#) no more than two evaluations of its proximal mapping will be necessary at every iteration. In addition, evaluating  $\text{prox}_{\gamma\varphi_2} = \Pi_{\mathcal{S}}$  (the set-valued projection onto  $\mathcal{S}$ ) amounts to setting to zero the  $n - k$  smallest coefficients of  $x$  (in magnitude), and normalizing the resulting vector to project it on the  $\ell_2$ -sphere: this has therefore a negligible cost.

[Figure 3](#) shows the results when the algorithm is applied using a small subset of the *20newsgroup* dataset,<sup>5</sup> which only retains 100 features from the original dataset:  $m = 16242$  and  $n = 100$ . This is the same dataset that was used in the experiments in [12]. Here, the initial point was chosen as the vector  $(1/n, \dots, 1/n) \in \mathbb{R}^n$ , which in our experiments gave consistently better results compared to random initialization, in terms of the objective value reached. In the [DRS-Algorithm 1](#), we used  $\lambda = 1$ ,

<sup>5</sup> <https://cs.nyu.edu/~roweis/data.html>



**Figure 3:** §5.2: sparse PCA problem (5.3) on a subset of the 20newsgroup dataset (100 features only). Comparison between *DRS* and the linesearch variant *Algorithm 1*, using modified Broyden, *L-BFGS*, and Anderson acceleration directions, when applied to the sparse PCA problem (5.3). On the x-axis, the number of linear systems solved (needed for the u-update): in the case of *DRS*, this coincides with the number of iterations, while for *Algorithm 1* it accounts for all operations performed in the linesearch. We used a memory parameter of 5 for *L-BFGS* and Anderson acceleration.

$\gamma = \frac{0.95}{2} L_{\varphi_1}^{-1}$ ,  $c = \frac{1}{2} C(\gamma L_{\varphi_1}, \lambda)$ , cf. (2.5)-(2.6), and the directions given by the modified Broyden, *L-BFGS* and Anderson acceleration (the latter two with memory 5). From this experiment, these directions in *Algorithm 1* greatly improve the convergence over *DRS*, both in terms of the fixed point residual norm as well as the objective value. We did not include results with Nesterov directions, as these did not perform well on this problem, a phenomenon that we believe is related to the nonconvexity of  $\varphi_1$  (cf. Remark 3.3).

### 5.3 Sparse PCA: consensus formulation

As the problem size grows, a big limitation is the need to store and operate with large matrices. To account for this issue, we consider the following *consensus formulation*: having fixed a number of *agents*  $N \geq 1$ , decompose matrix  $W$  into  $N$  row blocks  $W_1, \dots, W_N$ , so that  $W^\top = [W_1^\top \cdots W_N^\top]$  and  $\|Wx\|^2 = \sum_{i=1}^N \|W_i x\|^2$ , introduce  $N$  copies  $x_1, \dots, x_N$  of  $x$  (stacked in a vector  $\mathbf{x} \in \mathbb{R}^{nN}$ ), and solve

$$\underset{\mathbf{x} \in \mathbb{R}^{nN}, \mathbf{z} \in \mathbb{R}^n}{\text{minimize}} \sum_{i=1}^N -\frac{1}{2m} \|W_i x_i\|^2 \quad \text{subject to } \|z\| = 1, \|z\|_0 \leq k, x_i = z, i = 1 \dots N.$$

This problem is equivalent to (5.3), and can be expressed in ADMM form (1.2) as

$$\underset{\mathbf{x} \in \mathbb{R}^{nN}, \mathbf{z} \in \mathbb{R}^n}{\text{minimize}} \underbrace{\sum_{i=1}^N -\frac{1}{2m} \|W_i x_i\|^2}_{f(\mathbf{x})} + \underbrace{\delta_S(\mathbf{z})}_{g(\mathbf{z})} \quad \text{subject to} \quad \mathbf{x} - \begin{pmatrix} \mathbf{I} \\ \vdots \\ \mathbf{I} \end{pmatrix} \mathbf{z} = 0. \quad (5.4)$$

Apparently, the ADMM matrix  $A$  is the  $nN \times nN$  identity,  $B \in \mathbb{R}^{nN \times n}$  is the vertical stacking of  $N$  many  $n \times n$  negative identity matrices, and  $b$  is the zero  $\mathbb{R}^{nN}$  vector. Notice that Assumption II is satisfied, as  $(Af) = f$  has Lipschitz-continuous gradient with modulus  $L_{(Af)} = \frac{1}{m} \max_{i=1 \dots N} \|W_i\|^2 \leq \frac{1}{m} \|W\|^2$  and  $A$  has clearly full row rank.

The  $\mathbf{z}$ -update as prescribed by ADMM comes at negligible cost, since

$$\arg \min_{\mathbf{z} \in \mathbb{R}^n} \left\{ \delta_S(\mathbf{z}) + \frac{\beta}{2} \|\mathbf{x} + B\mathbf{z}\|^2 \right\} = \Pi_S\left(\frac{1}{N} \sum_{i=1}^N x_i\right) \quad \forall \mathbf{x} \in \mathbb{R}^{nN}.$$

The  $\mathbf{x}$ -update amounts to solving (in parallel)  $N$  (small) linear systems:

$$\arg \min_{x_i \in \mathbb{R}^n} \left\{ -\frac{1}{2m} \|W_i x_i\|^2 + \frac{\beta}{2} \|x_i - \mathbf{z}\|^2 \right\} = \left( \mathbf{I} - \frac{1}{m\beta} W_i^\top W_i \right)^{-1} \mathbf{z} = \mathbf{z} + W_i^\top (m\beta \mathbf{I} - W_i W_i^\top)^{-1} W_i \mathbf{z},$$

for  $i = 1 \dots N$ , where the second equality uses the Woodbury identity. The Cholesky factors of the  $m_i \times m_i$  matrices  $m\beta \mathbf{I} - W_i W_i^\top$ ,  $i = 1, \dots, N$ , can be computed once offline to efficiently solve the linear systems at each  $\mathbf{x}$ -update, resulting in  $O(\sum_{i=1}^N m_i^2)$  memory requirement, as opposed to  $O(N^2) = O(\sum_{i=1}^N m_i)^2$  (let alone the operational cost) needed for the original single-agent problem expression.

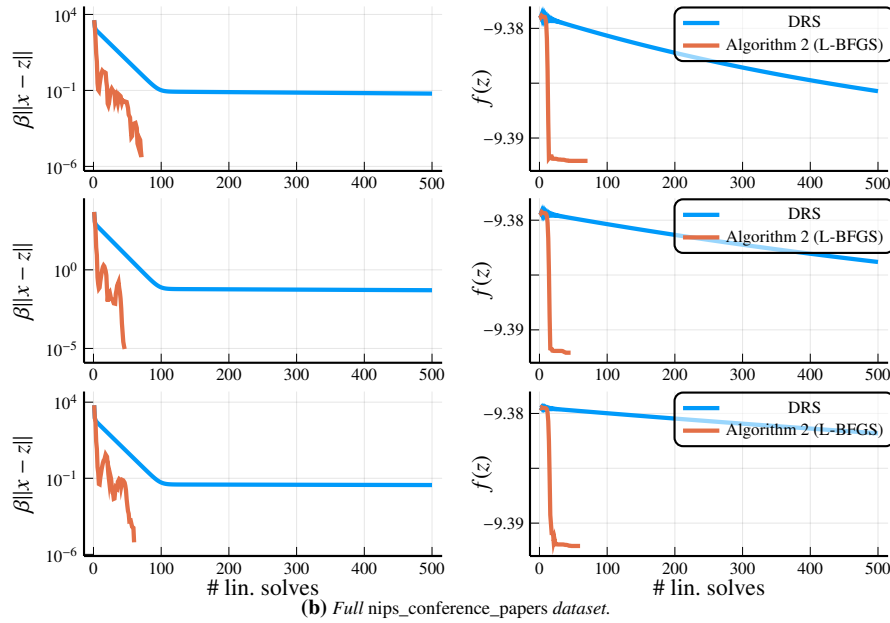
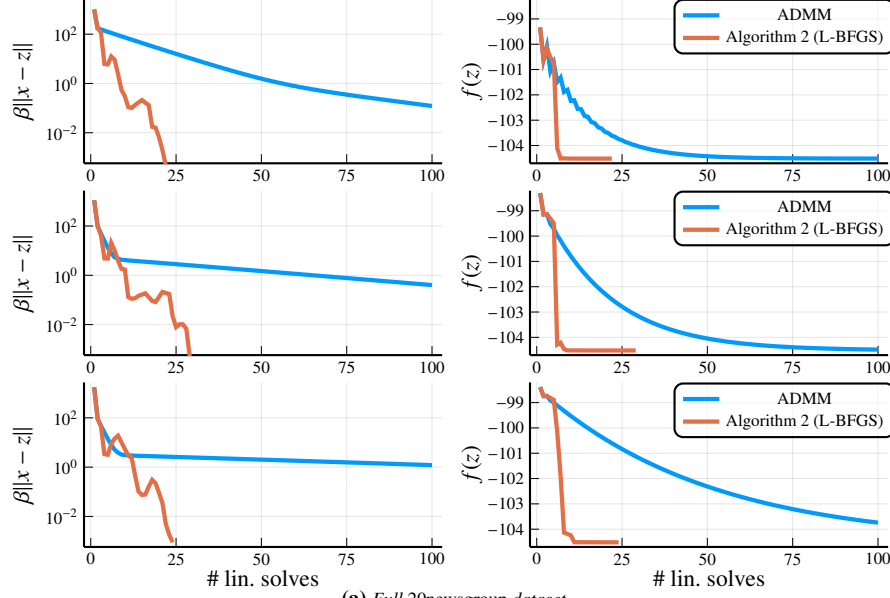
This consensus formulation, however, increases the problem size and thus the ill conditioning, and for moderate values of  $m$ ,  $n$  and  $N$  the convergence of plain ADMM is already prohibitively slow, cf. Figure 4. On the contrary, the adoption of L-BFGS directions in the ADMM-Algorithm 2 robustifies the performance at the negligible cost of few scalar products per iteration.

Figure 4 shows the result of running the proposed algorithm, with L-BFGS directions, to the full *20newsgroup*<sup>6</sup> and *nips\_conference\_papers*<sup>7</sup> datasets. The former consists of the frequencies of  $n = 26214$  words in  $m = 18846$  documents, and the latter contains  $n = 11463$  word counts from  $m = 5811$  NIPS conference papers. We split the data in  $N$  subsets (by row) of approximately equal size, for  $N \in \{10, 20, 50\}$ , to put the problem in the form (5.4). In each experiment we used  $\lambda = 1$ , the penalty parameter in both Algorithm 2 and the nominal ADMM was set to  $\beta = \frac{2}{0.95} L_{(Af)}$ , and  $c = \frac{1}{2} C(\beta^{-1} L_{(Af)}, \lambda)$ , cf. (2.5) and (2.11). In this case, we only considered L-BFGS directions with memory 5 since their computation scales better with the problem dimension (cf. Remark 3.3). Both algorithms were started at the same initial iterates  $\mathbf{y}^0 = 0$  and  $\mathbf{z}^0 = (1/n, \dots, 1/n) \in \mathbb{R}^n$ . Apparently, Algorithm 2 using L-BFGS converges faster than the nominal ADMM, and its convergence speed is significantly less susceptible to the number  $N$  of agents.

<sup>6</sup> <http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html>

<sup>7</sup> <http://archive.ics.uci.edu/ml/datasets/NIPS+Conference+Papers+1987-2015>

**Figure 4:** §5.3: consensus sparse PCA problem (5.4) on full datasets. Comparison between ADMM (blue) and the L-BFGS enhancement (red), for different number of agents  $N = 10, 20, 50$ . Left: ADMM residual; Right: cost. On the x-axis, the number of linear systems solved (needed for the  $x$ -update): in ADMM this coincides with the number of iterations, while in Algorithm 2 it accounts for each linesearch step. This is the only expensive operation, as the  $z$ -update is negligible. Apparently, ADMM is severely affected by  $N$ , whereas using L-BFGS directions in Algorithm 2 consistently results in faster convergence.



#### 5.4 Linear model predictive control (strongly convex)

To showcase the performance of the proposed method in the strongly convex case, we apply [Algorithm 1](#) to linear model predictive control (MPC) problems [18], i.e. finite-horizon, discrete-time, linear optimal control problems of the form

$$\begin{aligned}
 & \underset{\substack{u_0, \dots, u_{N-1} \in \mathbb{R}^{n_u} \\ x_1, \dots, x_N \in \mathbb{R}^{n_x}}}{\text{minimize}} \quad \sum_{i=0}^{N-1} \left( \|x_{i+1} - \bar{x}\|_Q^2 + \|u_i\|_R^2 + g(u_i, x_{i+1}) \right) \\
 & \text{subject to} \quad x_0 \in \mathbb{R}^{n_x} \text{ given} \\
 & \quad \quad \quad x_{i+1} = Ax_i + Bu_i, \quad i = 0, \dots, N-1.
 \end{aligned} \tag{5.5}$$

The decision variables are the system inputs  $u_i$  and states  $x_{i+1}$ ,  $i = 0, \dots, N-1$ . In the quadratic cost terms,  $Q > 0$  and  $R > 0$ . The objective is find the optimal sequence of inputs  $u_i$  that drive the system towards the reference state  $\bar{x}$ . The equality constraints enforce the linear dynamics, while  $g$  is a convex functions that models constraints on inputs and states.

Grouping all variables into  $z = (u_0, \dots, u_{N-1}, x_1, \dots, x_N) \in \mathbb{R}^{N(n_u+n_x)}$ , and denoting the set  $\mathcal{D}_{x_0} = \{z \in \mathbb{R}^{N(n_u+n_x)} \mid x_{i+1} = Ax_i + Bu_i, \quad i = 0, \dots, N-1\}$ , problem (5.5) can be solved with [Algorithm 1](#) by setting

$$\begin{aligned}
 \varphi_1(z) &= \sum_{i=0}^{N-1} \left( \|x_{i+1} - \bar{x}\|_Q^2 + \|u_i\|_R^2 + \delta_{\mathcal{D}_{x_0}}(s) \right), \\
 \varphi_2(z) &= \sum_{i=0}^{N-1} g(u_i, x_{i+1}).
 \end{aligned}$$

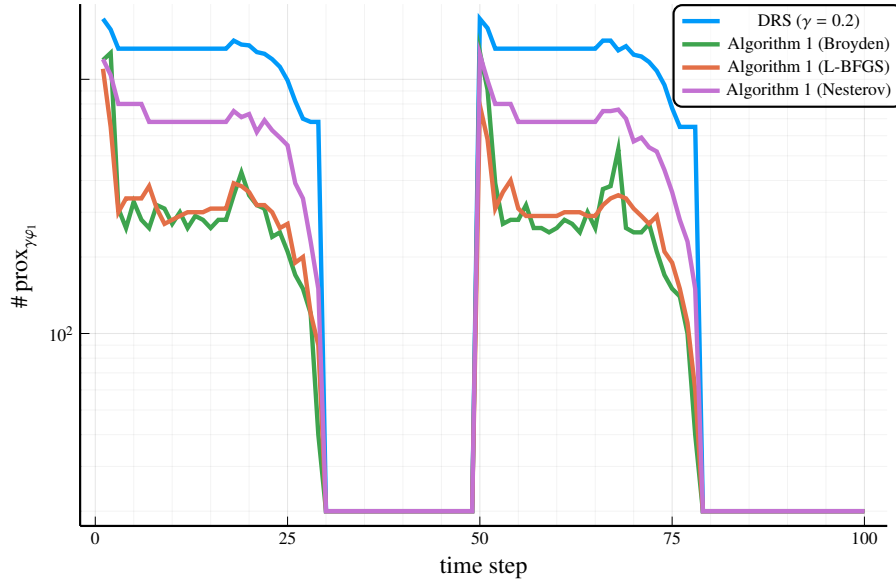
Both terms are nonsmooth in general; however,  $\varphi_1$  is strongly convex and computing  $\text{prox}_{\gamma\varphi_1}$  amounts to solving a strongly convex, equality-constrained quadratic program: due to the problem structure, this can be done efficiently via the Riccati equation, cf. [8, Sec. 1.9]. Therefore, [Algorithm 1](#) can be applied under [Assumption I\\*](#). Notice further that  $\mathcal{D}_{x_0}$  is an affine subspace, resulting in function  $\varphi_1$  being *generalized* quadratic; the linearity of  $\text{prox}_{\gamma\varphi_1}$  can be thus conveniently exploited in the linesearch as described in [Remark 3.2](#).

As specific instance of the problem, we considered the AFTI-16 system [7]: this has  $n_u = 2$  inputs,  $n_x = 4$  states, and unstable dynamics. We imposed hard constraint on the input variables and soft constraints on the system states, with

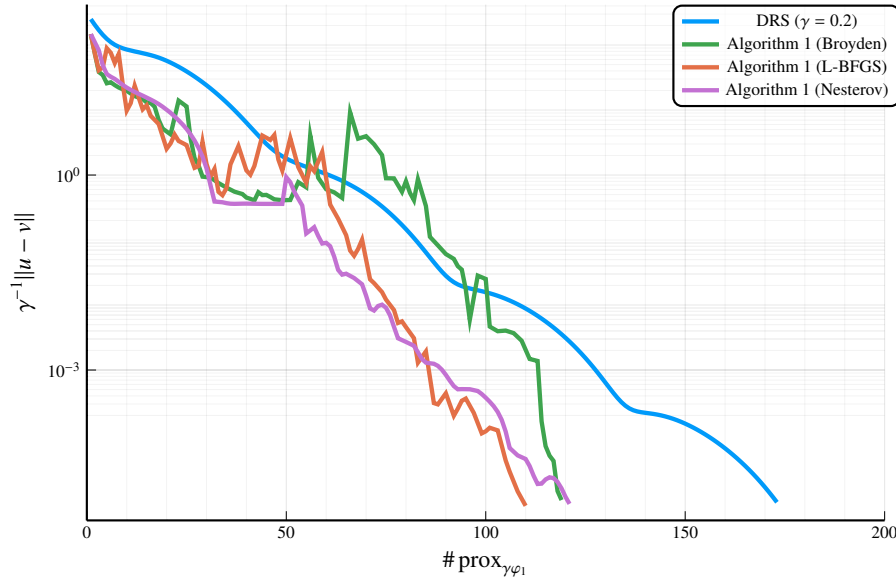
$$g(u_i, x_{i+1}) = \delta_{[-25, 25]}(u_i) + 10^6 \left[ \max \{0, |x_{i+1}^{(2)}| - 0.5\} + \max \{0, |x_{i+1}^{(4)}| - 100\} \right],$$

and set  $Q = \text{diag}(10^{-4}, 10^2, 10^{-3}, 10^2)$ ,  $R = \text{diag}(10^{-2}, 10^{-2})$ . The prediction horizon was set to  $N = 10$ , and the problem was scaled so as to have identity Hessian in the quadratic cost: this is known to improve significantly the convergence speed of [DRS](#) [41] and proved beneficial for [Algorithm 1](#) as well in our experiments. Therefore the scaled problem has  $\mu_{\varphi_1} = 1$ . In applying [Algorithm 1](#), we used  $\lambda = 1$ ,  $\gamma = (0.95 \cdot \mu_{\varphi_1})^{-1}$ ,  $c = \frac{1}{2}C((\gamma\mu_{\varphi_1})^{-1}, \lambda)$ , cf. [Corollary 2.4](#). For [DRS](#), since the problem is convex, it is known that any  $\gamma > 0$  is feasible. The choice of  $\gamma$  that performs best has

**Figure 5:** §5.4: linear MPC problem (5.5) for the AFTI-16 system. Comparison between *DRS* and the linesearch variant *Algorithm 1* using modified Broyden, L-BFGS, and Nesterov acceleration directions, to reach a tolerance of  $10^{-5}$  at each time step.



(a) Number of calls to  $\text{prox}_{\gamma\varphi_1}$  required by each method. After initially harder problems, the task becomes easier as the system reaches the imposed reference state; when the reference state changes at step 50, the problem becomes again harder to solve.



(b) Convergence of the fixed point residual for the different methods, on the initial problem of the simulation.

been made in hindsight; it is well known that the performance of **DRS** is sensitive to this choice, and we are not aware of any generically applicable rule. Therefore, after inspecting a grid of values for  $\gamma$ , we empirically found  $\gamma = 0.2$  to give the best performance with **DRS** on this specific problem, and used this value as baseline.

We simulated the system for 100 time steps, which correspond to 5 seconds when the original continuous-time dynamics is discretized with a step of 0.05 seconds. At each time step, problem (5.5) is solved with tolerance  $10^{-5}$ ; then, the first optimal input  $u_0$  is applied and the system evolves to the next time step, and the next problem is solved. On each problem, both **DRS** and **Algorithm 1** were warm-started by providing the final  $s$ -iterate to the previous problem as initial  $s$ -iterate: this proved beneficial for all algorithms. The initial system state is set to  $x_0 = (0, 0, 0, 0)$  at the beginning of the simulation; the reference state was set to  $\bar{x} = (0, 0, 0, 10)$  for the first 50 time steps (2.5 seconds), and to  $\bar{x} = (0, 0, 0, 0)$  for the remaining steps.

The performance of **Algorithm 1** is illustrated in **Figure 5a**, where the total number of calls to  $\text{prox}_{\gamma\varphi_1}$  at each time step is reported, for all considered algorithms. **Figure 5b** shows the convergence of the methods for the very first problem in the simulation. While using Anderson acceleration directions did not perform well in this example, it is clear that Broyden, L-BFGS, and Nesterov directions perform significantly better than vanilla **DRS**.

## 6 Conclusions

We proposed two linesearch algorithms that allow the employment of Newton-like update directions to enhance **DRS** and **ADMM**. The choice of quasi-Newton directions maintains the same low complexity as the original **DRS** and **ADMM** algorithms, as it prescribes only additional direct linear algebra. Simulations confirm that L-BFGS considerably robustifies the convergence, rendering these first-order algorithms extremely fast and unaffected by problem size and ill conditioning. The proposed algorithms are tuning-free and out-of-the-box, as the needed stepsizes and parameters can adaptively be retrieved without prior knowledge. Last but not least, they are suited for fully nonconvex problems, and maintain the same worst-case convergence properties of **DRS** and **ADMM**.

## A Auxiliary results

This appendix contains some auxiliary results needed for the convergence analysis of **Section 4**. As shown in [51, Eq. (3.4)], the DRE can be expressed in terms of the forward-backward envelope  $\varphi_\gamma^{\text{FB}}$  [35, 45, 52] as

$$\varphi_\gamma^{\text{DR}}(s) = \varphi_\gamma^{\text{FB}}(u) := \min_{w \in \mathbb{R}^p} \left\{ \varphi_1(u) + \varphi_2(w) + \langle \nabla \varphi_1(u), w - u \rangle + \frac{1}{2\gamma} \|w - u\|^2 \right\} \quad (\text{A.1a})$$

where  $u = \text{prox}_{\gamma\varphi_1}(s)$  and the minimum is attained at any  $v \in \text{prox}_{\gamma\varphi_2}(2u - s)$ . Equivalently,

$$\varphi_\gamma^{\text{DR}}(s) = \varphi_1(u) - \frac{\gamma}{2} \|\nabla \varphi_1(u)\|^2 + \varphi_2^\gamma(u - \gamma \nabla \varphi_1(u)). \quad (\text{A.1b})$$

**Fact A.1** ([51, Prop. 3.3]). *Suppose that **Assumption 1** holds and let  $\gamma < 1/L_{\varphi_1}$  be fixed. Then, for all  $s \in \mathbb{R}^p$  and  $(u, v) \in \text{DRS}_\gamma(s)$  it holds that*

$$\varphi(v) + \frac{1-\gamma L_{\varphi_1}}{2\gamma} \|v - u\|^2 \leq \varphi_\gamma^{\text{DR}}(s) \leq \varphi(u). \quad \square$$

**Lemma A.2.** Suppose that [Assumption I](#) holds. Then, for all  $s, \bar{u} \in \mathbb{R}^p$  and  $\gamma < 1/L_{\varphi_1}$

$$\varphi_\gamma^{\text{DR}}(s) - \varphi(\bar{u}) \leq \frac{1+\gamma L_{\varphi_1}}{2\gamma} \|\text{prox}_{\gamma\varphi_1}(s) - \bar{u}\|^2.$$

*Proof.* Let  $u := \text{prox}_{\gamma\varphi_1}(s)$  for brevity. By plugging  $w = \bar{u}$  into (A.1a) we obtain

$$\begin{aligned} \varphi_\gamma^{\text{DR}}(s) &\leq \varphi_2(\bar{u}) + \underbrace{\varphi_1(u) + \langle \nabla \varphi_1(u), \bar{u} - u \rangle}_{\varphi_1(\bar{u}) + \frac{L_{\varphi_1}}{2} \|\bar{u} - u\|^2} + \frac{1}{2\gamma} \|\bar{u} - u\|^2 \\ &\leq \varphi_2(\bar{u}) + \varphi_1(\bar{u}) + \frac{L_{\varphi_1}}{2} \|\bar{u} - u\|^2 + \frac{1}{2\gamma} \|\bar{u} - u\|^2, \end{aligned}$$

where the second inequality uses the known quadratic upper bound [8, Prop. A.24] for functions with Lipschitz-continuous gradient.  $\square$

**Lemma A.3.** Suppose that [Assumption I\\*](#) holds and let  $\gamma > 1/\mu_{\varphi_1}$  be fixed. Then,

$$\inf \varphi = -\inf \psi = -\inf \psi_{\gamma_*}^{\text{DR}} = \sup \varphi_\gamma^{\text{DR}} \quad (\text{A.2})$$

where  $\gamma_* = 1/\gamma$ . Moreover, for any  $s \in \mathbb{R}^p$  it holds that

$$\frac{1}{2\gamma} \|x_\star - v\|^2 + \frac{\gamma\mu_{\varphi_1}-1}{2\gamma} \|x_\star - u\|^2 \leq \inf \varphi - \varphi_\gamma^{\text{DR}}(s) = \psi_{\gamma_*}^{\text{DR}}(s) - \inf \psi \quad (\text{A.3})$$

where  $x_\star$  is the unique minimizer of  $\varphi$ ,  $s_* = -s/\gamma$ , and  $(u, v) = \text{DRS}_\gamma(s)$ .

*Proof.* Due to strong convexity, the set of primal solutions  $\arg \min \varphi$  is a singleton, ensuring strong duality  $\inf \varphi = -\inf \psi$  through [3, Thm. 5.2.1(b)-(c)]. Since  $\psi_1$  is  $1/\mu_{\varphi_1}$ -smooth, it follows from [Fact 2.1\(i\)](#) that  $\inf \psi_{\gamma_*}^{\text{DR}} = \inf \psi$  for every  $\gamma_* < \mu_{\varphi_1}$ ; combined with the identity  $\varphi_\gamma^{\text{DR}}(s) = -\psi_{\gamma_*}^{\text{DR}}(-s/\gamma)$  holding for  $\gamma_* = 1/\gamma$  (cf. [Theorem 2.3](#)), (A.2) is obtained. Let now  $s \in \mathbb{R}^p$  be fixed and consider  $(u, v) = \text{DRS}_\gamma(s)$ . From the inclusion  $\frac{s-u}{\gamma} \in \partial \varphi_1(u)$  (cf. (1.10)) and strong convexity of  $\varphi_1$  one has

$$\varphi_1(x_\star) \geq \varphi_1(u) + \frac{1}{\gamma} \langle s - u, x_\star - u \rangle + \frac{\mu_{\varphi_1}}{2} \|x_\star - u\|^2.$$

Similarly, since  $\frac{2u-s-v}{\gamma} \in \partial \varphi_2(v)$  and  $\varphi_2$  is convex, one has

$$\varphi_2(x_\star) \geq \varphi_2(v) + \frac{1}{\gamma} \langle 2u - s - v, x_\star - v \rangle.$$

Summing the two inequalities yields

$$\begin{aligned} \inf \varphi &\geq \varphi_1(u) + \varphi_2(v) + \frac{1}{\gamma} \langle s - u, x_\star - u \rangle + \frac{1}{\gamma} \langle 2u - s - v, x_\star - v \rangle + \frac{\mu_{\varphi_1}}{2} \|x_\star - u\|^2 \\ &= \varphi_1(u) + \varphi_2(v) + \frac{1}{\gamma} \langle s - u, v - u \rangle + \frac{1}{\gamma} \langle u - v, x_\star - v \rangle + \frac{\mu_{\varphi_1}}{2} \|x_\star - u\|^2 \\ &= \varphi_1(u) + \varphi_2(v) + \frac{1}{\gamma} \langle s - u, v - u \rangle + \frac{1}{2\gamma} \|u - v\|^2 + \frac{1}{2\gamma} \|x_\star - v\|^2 + \frac{\gamma\mu_{\varphi_1}-1}{2\gamma} \|x_\star - u\|^2 \\ &= \varphi_\gamma^{\text{DR}}(s) + \frac{1}{2\gamma} \|x_\star - v\|^2 + \frac{\gamma\mu_{\varphi_1}-1}{2\gamma} \|x_\star - u\|^2. \end{aligned}$$

The claim now follows from the identity  $\psi_{\gamma_*}^{\text{DR}}(s) = -\varphi_\gamma^{\text{DR}}(s)$  shown in [Theorem 2.3](#).  $\square$

## References

1. Anderson, D.G.: Iterative procedures for nonlinear integral equations. J. ACM **12**(4), 547–560 (1965). DOI 10.1145/321296.321305
2. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Mathematical Programming **137**(1), 91–129 (2013). DOI 10.1007/s10107-011-0484-9
3. Auslender, A., Teboulle, M.: Asymptotic Cones and Functions in Optimization and Variational Inequalities. Springer Monographs in Mathematics. Springer New York (2002)



4. Bauschke, H.H., Combettes, P.L.: *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics. Springer (2017). DOI 10.1007/978-3-319-48311-5
5. Bauschke, H.H., Noll, D.: On the local convergence of the Douglas-Rachford algorithm. *Archiv der Mathematik* **102**(6), 589–600 (2014). DOI 10.1007/s00013-014-0652-2
6. Bauschke, H.H., Phan, H.M., Wang, X.: The method of alternating relaxed projections for two nonconvex sets. *Vietnam Journal of Mathematics* **42**(4), 421–450 (2014). DOI 10.1007/s10013-013-0049-8
7. Bemporad, A., Casavola, A., Mosca, E.: Nonlinear control of constrained linear systems via predictive reference management. *IEEE Transactions on Automatic Control* **42**(3), 340–349 (1997). DOI 10.1109/9.557577
8. Bertsekas, D.P.: *Nonlinear Programming*, 2nd edition edn. Athena Scientific (1999)
9. Bolte, J., Sabach, S., Teboulle, M.: Proximal Alternating Linearized Minimization for nonconvex and nonsmooth problems. *Mathematical Programming* **146**(1–2), 459–494 (2014). DOI 10.1007/s10107-013-0701-9
10. Broyden, C.G.: A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation* **19**(92), 577–593 (1965)
11. Broyden, C.G.: The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics* **6**(1), 76–90 (1970). DOI 10.1093/imamat/6.1.76
12. d’Aspremont, A., Ghaoui, L.E., Jordan, M.I., Lanckriet, G.R.: A direct formulation for sparse PCA using semidefinite programming. In: *Advances in neural information processing systems*, pp. 41–48 (2005)
13. Daubechies, I., DeVore, R., Fornasier, M., Güntürk, C.S.: Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics* **63**(1), 1–38 (2010). DOI 10.1002/cpa.20303
14. Eckstein, J., Bertsekas, D.P.: On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming* **55**(1), 293–318 (1992). DOI 10.1007/BF01581204
15. Facchinei, F., Pang, J.S.: *Finite-dimensional variational inequalities and complementarity problems*, vol. II. Springer (2003)
16. Fang, H.r., Saad, Y.: Two classes of multisection methods for nonlinear acceleration. *Numerical Linear Algebra with Applications* **16**(3), 197–221 (2009). DOI 10.1002/nla.617
17. Fletcher, R.: A new approach to variable metric algorithms. *The Computer Journal* **13**(3), 317–322 (1970). DOI 10.1093/comjnl/13.3.317
18. García, C.E., Prett, D.M., Morari, M.: Model predictive control: Theory and practice – A survey. *Automatica* **25**(3), 335–348 (1989). DOI 10.1016/0005-1098(89)90002-2
19. Goldfarb, D.: A family of variable-metric methods derived by variational means. *Mathematics of computation* **24**(109), 23–26 (1970)
20. Goncalves, M.L.N., Melo, J.G., Monteiro, R.D.C.: Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems. *Pacific Journal of Optimization* **15**, 378–398 (2019)
21. Guo, K., Han, D., Wu, T.T.: Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints. *International Journal of Computer Mathematics* **94**(8), 1653–1669 (2017). DOI 10.1080/00207160.2016.1227432
22. Hesse, R., Luke, R.: Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *SIAM Journal on Optimization* **23**(4), 2397–2419 (2013). DOI 10.1137/120902653
23. Hesse, R., Luke, R., Neumann, P.: Alternating projections and Douglas-Rachford for sparse affine feasibility. *IEEE Transactions on Signal Processing* **62**(18), 4868–4881 (2014). DOI 10.1109/TSP.2014.2339801
24. Hong, M., Luo, Z.Q., Razaviyayn, M.: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization* **26**(1), 337–364 (2016). DOI 10.1137/140990309
25. Ip, C.M., Kyriasis, J.: Local convergence of quasi-Newton methods for B-differentiable equations. *Mathematical Programming* **56**(1-3), 71–89 (1992)
26. Izmailov, A.F., Solodov, M.V.: *Newton-type methods for optimization and variational problems*. Springer (2014)
27. Jiang, B., Lin, T., Ma, S., Zhang, S.: Structured nonconvex and nonsmooth optimization: algorithms and iteration complexity analysis. *Computational Optimization and Applications* **72**(1), 115–157 (2019)

28. Li, G., Liu, T., Pong, T.K.: Peaceman-Rachford splitting for a class of nonconvex optimization problems. *Computational Optimization and Applications* (2017). DOI 10.1007/s10589-017-9915-8
29. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization* **25**(4), 2434–2460 (2015). DOI 10.1137/140998135
30. Li, G., Pong, T.K.: Douglas-Rachford splitting for nonconvex optimization with application to non-convex feasibility problems. *Mathematical Programming* **159**(1), 371–401 (2016). DOI 10.1007/s10107-015-0963-5
31. Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, R. Garnett (eds.) *Advances in Neural Information Processing Systems* 28, pp. 379–387. Curran Associates, Inc. (2015)
32. Maratos, N.: Exact penalty function algorithms for finite dimensional and control optimization problems. Ph.D. thesis, Imperial College London (University of London) (1978)
33. Nesterov, Y.: A method of solving a convex programming problem with convergence rate  $o(1/k^2)$ . *Soviet Mathematics Doklady* **27**, 372–376 (1983)
34. Nocedal, J., Wright, S.: *Numerical Optimization*, 2nd edition edn. Springer, New York (2006)
35. Patrinos, P., Bemporad, A.: Proximal Newton methods for convex composite optimization. In: 52nd IEEE Conference on Decision and Control, pp. 2358–2363 (2013). DOI 10.1109/CDC.2013.6760233
36. Patrinos, P., Stella, L., Bemporad, A.: Douglas-Rachford splitting: Complexity estimates and accelerated variants. In: 53rd IEEE Conference on Decision and Control, pp. 4234–4239 (2014). DOI 10.1109/CDC.2014.7040049
37. Pejčic, I., Jones, C.: Accelerated ADMM based on accelerated Douglas-Rachford splitting. In: 2016 European Control Conference (ECC), pp. 1952–1957 (2016). DOI 10.1109/ECC.2016.7810577
38. Poliquin, R.A., Rockafellar, R.T.: Generalized Hessian properties of regularized nonsmooth functions. *SIAM Journal on Optimization* **6**(4), 1121–1137 (1996)
39. Powell, M.J.: A hybrid method for nonlinear equations. *Numerical Methods for Nonlinear Algebraic Equations* pp. 87–144 (1970)
40. Powell, M.J.: A fast algorithm for nonlinearly constrained optimization calculations. In: G.A. Watson (ed.) *Numerical Analysis*, pp. 144–157. Springer Berlin Heidelberg, Berlin, Heidelberg (1978)
41. Rey, F., Frick, D., Domahidi, A., Jerez, J., Morari, M., Lygeros, J.: ADMM prescaling for model predictive control. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 3662–3667. IEEE, Las Vegas, NV, USA (2016). DOI 10.1109/CDC.2016.7798820
42. Rockafellar, R.T.: *Convex analysis*, vol. 28. Princeton university press (1970)
43. Rockafellar, R.T., Wets, R.J.: *Variational analysis*, vol. 317. Springer (2011)
44. Shanno, D.F.: Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation* **24**(111), 647–656 (1970)
45. Stella, L., Themelis, A., Patrinos, P.: Forward-backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications* **67**(3), 443–487 (2017). DOI 10.1007/s10589-017-9912-y
46. Stella, L., Themelis, A., Patrinos, P.: Newton-type alternating minimization algorithm for convex optimization. *IEEE Transactions on Automatic Control* **64**(2), 697–711 (2019). DOI 10.1109/TAC.2018.2872203
47. Stella, L., Themelis, A., Sopasakis, P., Patrinos, P.: A simple and efficient algorithm for nonlinear model predictive control. In: 2017 IEEE 56th Annual Conference on Decision and Control (CDC), pp. 1939–1944 (2017). DOI 10.1109/CDC.2017.8263933
48. Themelis, A.: Proximal algorithms for structured nonconvex optimization. Ph.D. thesis, KU Leuven (2018)
49. Themelis, A., Ahookhosh, M., Patrinos, P.: On the acceleration of forward-backward splitting via an inexact Newton method. In: H.H. Bauschke, R.S. Burachik, D.R. Luke (eds.) *Splitting Algorithms, Modern Operator Theory, and Applications*, pp. 363–412. Springer International Publishing, Cham (2019). DOI 10.1007/978-3-030-25939-6\_15
50. Themelis, A., Patrinos, P.: SuperMann: a superlinearly convergent algorithm for finding fixed points of nonexpansive operators. *IEEE Transactions on Automatic Control* **64**(12), 4875–4890 (2019). DOI 10.1109/TAC.2019.2906393
51. Themelis, A., Patrinos, P.: Douglas–Rachford splitting and ADMM for nonconvex optimization: Tight convergence results. *SIAM Journal on Optimization* **30**(1), 149–181 (2020). DOI 10.1137/18M1163993
52. Themelis, A., Stella, L., Patrinos, P.: Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization* **28**(3), 2274–2303 (2018). DOI 10.1137/16M1080240

53. Wang, Y., Yin, W., Zeng, J.: Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing* **78**(1), 29–63 (2019)
54. Xu, Z., Chang, X., Xu, F., Zhang, H.:  $L_{1/2}$  regularization: A thresholding representation theory and a fast solver. *IEEE Transactions on Neural Networks and Learning Systems* **23**(7), 1013–1027 (2012). DOI 10.1109/TNNLS.2012.2197412
55. Yan, M., Yin, W.: Self Equivalence of the Alternating Direction Method of Multipliers, pp. 165–194. Springer International Publishing, Cham (2016). DOI 10.1007/978-3-319-41589-5\_5