DOUGLAS-RACHFORD SPLITTING AND ADMM FOR NONCONVEX OPTIMIZATION: ACCELERATED AND NEWTON-TYPE ALGORITHMS

ANDREAS THEMELIS, LORENZO STELLA AND PANAGIOTIS PATRINOS

ABSTRACT. Although the performance of popular optimization algorithms such as Douglas-Rachford splitting (DRS) and 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.

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. [16, 5, 6, 15, 22, 20, 41] for DRS and [21, 17, 14, 13, 43, 41] for ADMM. The former algorithm addresses the following composite minimization problems

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

for some $\varphi_1, \varphi_2 : \mathbb{R}^p \to \overline{\mathbb{R}}$ ($\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ 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

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

Here, $\mathbf{prox}_h$ denotes the *proximal mapping* of function h; cf. Section 1.3.

The ADMM addresses optimization problems that can be formulated as

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

1991 *Mathematics Subject Classification.* 90C06, 90C25, 90C26, 49J52, 49J53.

Key words and phrases. Nonsmooth nonconvex optimization, Douglas-Rachford splitting, ADMM, quasi-Newton methods.

A. Themelis and P. Patrinos are with the Department of Electrical Engineering (ESAT-STADIUS) – KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium, and are supported by the Research Foundation Flanders (FWO) research projects G086518N and G086318N; 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 no 30468160 (SeLMA)

L. Stella is with Amazon Research, Charlottenstrasse 4, 10969, Berlin, Germany *(the work was done prior to joining Amazon)*

andreas.themelis@kuleuven.be, panos.patrinos@esat.kuleuven.be, lorenzostella@gmail.com.

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

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

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

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

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 [44, 41] (see Fact 2.5 for the details). In both algorithms the relaxation λ serves as an averaging factor. The core of the respective iterations, independent of λ, can instead be summarized in the following oracles:

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

and

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

One of the main advantages of DRS and ADMM is 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. Aware of these pros an cons, 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, yet maintaining the same operational simplicity and worst-case convergence properties of the original methods. This is done by leveraging on favorable properties of the Douglas-Rachford envelope [28, 41], a continuous, real-valued, exact penalty function for problem (1.1), in a similar fashion to the predecessors forward-backward-based PANOC [36] and alternating-minimization-based NAMA [35] algorithms.

1.1. **Contributions.** In contrast to an *open-loop* preconditioining approach, we propose two linesearch algorithms that integrate DRS and ADMM with an *on-line* scaling based on Newton-type update directions. These are the first that (1) are compatible with fully nonconvex problems, (2) maintain the 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 and Anderson acceleration are also very effective in practice.

1.2. **Paper organization.** After concluding this section with the notational convention adopted throughout the paper, in Section 2 we offer a list of the key properties of the DRS and ADMM steps. These constitute the building blocks of the proposed linesearch Algorithms 1 and 2, introduced and detailed in Section 3 and which allow us to integrate DRS and ADMM with arbitrary update directions while preserving the convergence properties; favorable update directions that can considerably improve convergence speed are also discussed. 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 sparse principal component analysis problems. Section 6 concludes the paper.

1.3. **Notation and known facts.** We denote as $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ the extended-real line. With 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 $\mathbf{dist}(x, E) = \mathbf{inf}_{z \in S} \|z - x\|$.

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 $\frac{\|s^{k+1} - s_\star\|}{\|s^k - s_\star\|} \to 0$ as $k \to \infty$.

A function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *proper* if $h \not\equiv \infty$, in which case its *domain* is defined as the set $\mathbf{dom}\, h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$, and is *lower semicontinuous (lsc)* if for any $\bar{x} \in \mathbb{R}^n$ it holds that $h(\bar{x}) \leq \mathbf{liminf}_{x\to\bar{x}} h(x)$. A point $x_\star \in \mathbf{dom}\, h$ 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*. For $\alpha \in \mathbb{R}$, $\mathbf{lev}_{\leq\alpha} h$ is the *α-level set* of h, i.e., $\mathbf{lev}_{\leq\alpha} h := \{x \in \mathbb{R}^n \mid h(x) \leq \alpha\}$. We say that h is *level bounded* if $\mathbf{lev}_{\leq\alpha} h$ is bounded for all $\alpha \in \mathbb{R}$, a condition which is equivalent to $\mathbf{liminf}_{\|x\|\to\infty} f(x) = \infty$. With h^* we indicate the convex conjugate of h, pointwise defined as $h^*(y) = \mathbf{sup}\{\langle y, \cdot\rangle - h\}$.

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

$$v \in \hat{\partial} h(\bar{x}) \quad \Leftrightarrow \quad \mathop{\mathbf{liminf}}_{\bar{x}\neq x\to\bar{x}} \frac{h(x) - h(\bar{x}) - \langle v, x - \bar{x}\rangle}{\|x - \bar{x}\|} \geq 0. \tag{1.6}$$

The (limiting) *subdifferential* of h is $\partial h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$, where $v \in \partial h(x)$ iff there exists a sequence $(x^k, v^k)_{k\in\mathbb{N}}$ with $v^k \in \hat{\partial} h(x^k)$ such that $(x^k, h(x^k), v^k) \to (x, h(x), v)$ as $k \to \infty$. A necessary condition for local minimality of x for h is $0 \in \partial h(x)$, see [33, Thm. 10.1].

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

$$\mathbf{prox}_{\gamma h}(x) := \mathop{\mathbf{arg\,min}}_{w\in\mathbb{R}^n} \left\{h(w) + \tfrac{1}{2\gamma}\|w - x\|^2\right\}. \tag{1.7}$$

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

$$h^\gamma(x) := \mathop{\mathbf{inf}}_{w\in\mathbb{R}^n} \left\{h(w) + \tfrac{1}{2\gamma}\|w - x\|^2\right\}. \tag{1.8}$$

The necessary optimality condition for the problem defining $\mathbf{prox}_{\gamma h}$ together with the calculus rule of [33, Ex. 8.8] imply

$$\tfrac{1}{\gamma}(x - \bar{x}) \in \hat{\partial} h(\bar{x}) \quad \forall \bar{x} \in \mathbf{prox}_{\gamma h}(x). \tag{1.9}$$

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

$$\|s^+ - s_\star\|^2 \leq \|s - s_\star\|^2 - c\|s - s^+\|^2 \tag{2.1}$$

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 likes 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 [41] 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 linesearch algorithms developed in Section 3. Finally, by means of a primal equivalence first noticed in [44] 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 [22, 20] and later [41] extended the analysis to the nonconvex case when the functions satisfy the following requirements.

Assumption I (Requirements for DRS: the smooth φ_1 case)**.** *In problem* (1.1)*,*

A1 $\varphi_1 : \mathbb{R}^p \to \mathbb{R}$ *has* L_{φ_1}*-Lipschitz continuous gradient;*

A2 $\varphi_2 : \mathbb{R}^p \to \overline{\mathbb{R}}$ *is lsc;*

A3 *problem* (1.1) *admits a solution:* $\operatorname{\mathbf{arg\,min}} \varphi \neq \emptyset$.

In the setting of Assumption I, [22, 20] pioneered the idea of employing an augmented Lagrangian function as Lyapunov potential for DRS iterations, namely

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

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 γ, and subsequential convergence of the algorithm to stationary points was thus inferred. The results have been tightened in [41], where $c = 1/\gamma$ is shown to be an optimal choice for the Lagrangian penalty parameter and the augmented Lagrangian is regarded as a function of the sole variable s, thus recovering the *Douglas-Rachford envelope* (DRE) of [28], namely

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

where u and v are the result of a DRS-update with stepsize γ starting at s (u and v are independent of the relaxation λ). As detailed in [41, Prop. 2.3 and Rem. 3.1], when Assumption I is satisfied and $\gamma < 1/L_{\varphi_1}$ one has that $\mathbf{prox}_{\gamma\varphi_1}$ is Lipschitz continuous and $\mathbf{prox}_{\gamma\varphi_2}$ is a well-defined set-valued mapping, in the sense that $\mathbf{prox}_{\gamma\varphi_2}(x)$ is nonempty for every $x \in \mathbb{R}^p$. In fact, in order for the DRE to be well defined it suffices that $\mathbf{prox}_{\gamma\varphi_1}$ is single valued, as the expression (2.2) can easily be seen to equal

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

Nevertheless, under Assumption I the DRE enjoys a close kinship with the cost function φ, as summarized next.

Fact 2.1 ([41, Prop. 3.2 and Thm. 3.4])**.** *Suppose that Assumption I holds. Then, for all $\gamma < 1/L_{\varphi_1}$ the DRE $\varphi_\gamma^{\mathrm{DR}}$ is real valued, locally Lipschitz, and satisfies*

(i) $\inf \varphi = \inf \varphi_\gamma^{\mathrm{DR}}$ *and* $\operatorname{arg\,min} \varphi = \operatorname{prox}_{\gamma\varphi_1}(\operatorname{arg\,min} \varphi_\gamma^{\mathrm{DR}})$*;*

(ii) $\varphi_\gamma^{\mathrm{DR}}$ is level bounded iff φ is level bounded.

In the same spirit of the preceding works [22, 20], the convergence analysis of nonconvex DRS in [41] revolves around the following result, which assesses that 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 γ as in [41, Rem. 4.2], which only discerns whether φ_1 is convex or not. Tight ranges are given in [41, Thm. 4.1], and require the knowledge of the *hypoconvexity* modulus of φ_1.

Fact 2.2 (Sufficient decrease on the DRE [41, 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,*

$$\varphi_\gamma^{\mathrm{DR}}(s^+) \le \varphi_\gamma^{\mathrm{DR}}(s) - \tfrac{1}{\gamma} C(\gamma L_{\varphi_1}, \lambda)\|r\|^2, \tag{2.4}$$

where $r := u - v$ and

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

In particular, the constant C is strictly positive provided that

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

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 $u^k - v^k \to 0$ with $(u^k, v^k) \in \mathrm{DRS}_\gamma(s^k)$, any accumulation point $u^\star$ of $(v^k)_{k\in\mathbb{N}}$ satisfies the stationarity condition $0 \in \partial\varphi(u^\star)$. As detailed in the dedicated Section 3, 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 [41] 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 convex case.* As observed in [44], for convex problems DRS (or, equivalently, its sibling ADMM) is equivalent to itself applied to the dual formulation. To see this, let φ_1 and φ_2 be (proper, lsc) convex functions, and consider the following dual formulation of (1.1):[1]

$$\operatorname*{minimize}_{y\in\mathbb{R}^n} \psi(y) := \psi_1(y) + \psi_2(y), \tag{2.7}$$

where $\psi_1 := \varphi_1^*(-\,\cdot\,)$ and $\psi_2 = \varphi_2^*$. The Moreau identity [4, Thm. 14.3(ii)] yields

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

[1] Specifically, (2.7) is the dual of $\operatorname{minimize}_{x,z\in\mathbb{R}^n} \varphi_1(x) + \varphi_2(z)$ **subject to** $x - z = 0$.

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

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

leading to the following result.

Theorem 2.3 (Self-duality of DRS)**.** *Suppose that φ_1 and φ_2 are (proper, lsc) convex functions, and consider a DRS*-update $s_* \mapsto (u_*, v_*, s_*^+)$. Let $s := -\frac{s_*}{\gamma_*}$ and consider a DRS-update $s \mapsto (u, v, s^+)$ with stepsize $\gamma = 1/\gamma_*$ and relaxation $\lambda = \lambda_*$. Then, the variables are related as follows:*

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

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

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

$$\psi_1^{\gamma_*}(s_*) = (((\varphi_1 \circ (-\cdot))^*)^{1/\gamma}(-s/\gamma) = (\varphi_1^*)^{1/\gamma}(s/\gamma) = \tfrac{1}{2\gamma}\|s\|^2 - \varphi_1^\gamma(s),$$

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

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

as claimed. □

A consequence of the self-equivalence of DRS is that, for convex problems, it makes no difference if it is the primal formulation (1.1) or the dual (2.7) that satisfies the needed working assumptions. In fact, since the convex conjugate of a μ-strongly convex function is (convex and) $1/\mu$-Lipschitz differentiable, in parallel to Assumption I we can also include the following.

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

A1 *$\varphi_1 : \mathbb{R}^p \to \overline{\mathbb{R}}$ is proper, lsc, and μ_{φ_1}-strongly convex;*

A2 *$\varphi_2 : \mathbb{R}^p \to \overline{\mathbb{R}}$ is proper, lsc, and convex;*

A3 *a solution exists.*

When functions are convex and without necessarily either one being strongly so, plain DRS iterations are known to converge for any stepsize γ and relaxation $\lambda \in (0, 2)$. Strong convexity is instead crucial for the well definedness of Algorithm 1. This is similar to the setting of [28], where by viewing DRS as a scaled gradient descent algorithm on $\varphi^{\text{DR}}_\gamma$ it was possible to derive a Nesterov-type acceleration whenever φ_2 is convex and φ_1 is strongly convex and quadratic. The same arguments were then extended to ADMM in the follow-up work [29] 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,*

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

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

$$\operatorname*{\mathbf{minimize}}_{s \in \mathbb{R}^p} (Af)(s) + (Bg)(b - s), \tag{2.10}$$

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

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

It was shown in [44, Thm. 1] and later generalized in [41, 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 [41, 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* λ *and penalty* $\beta > 0$*. Let*

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

Then, the variables are related as follows:

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

Moreover,

(i) $-A^\top y^+ \in \hat{\partial} f(x^+)$*, and*

(ii) $\mathbf{dist}(-B^\top y^+, \hat{\partial} g(z^+)) \leq \beta\|B\|\|Ax^+ + Bz^+ - b\|$.

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),* $A \in \mathbb{R}^{p\times m}$*,* $B \in \mathbb{R}^{p\times n}$*,* $b \in \mathbb{R}^p$*,* $f : \mathbb{R}^m \to \mathbb{R}$ *and* $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ *are such that*

A1 *A is surjective (full row rank) and $(Af) : \mathbb{R}^p \to \mathbb{R}$ has $L_{(Af)}$-Lipschitz gradient;*

A2 *$(Bg) : \mathbb{R}^p \to \overline{\mathbb{R}}$ is lsc;*

A3 *a solution exists:* $\operatorname{\mathbf{arg\,min}} \Phi \neq \emptyset$*, 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 [41, 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 $\mu_f/\|A\|^2$.[2]

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

A1 *$f : \mathbb{R}^m \to \overline{\mathbb{R}}$ is lsc, proper, and μ_f-strongly convex;*

A2 *$(Bg) : \mathbb{R}^n \to \overline{\mathbb{R}}$ is lsc, proper, and convex (e.g. when g is proper, convex and coercive);*

A3 *a solution exists (equivalently,* $\mathbf{dom}\,\Phi$ *is nonempty).*

[2] In the limiting case $A = 0$, one has that $(Af) = f(0) + \delta_{\{0\}}$ is lsc and ∞-strongly convex for any f, and properness amounts to the condition $0 \in \mathbf{dom}\, f$.

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 λ 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 $\varphi_\gamma^{\mathrm{DR}}(s^+) = \mathscr{L}_\beta(x^+, z^+, y^+)$ for $s^+ = Ax^+ - y^+/\beta$ and $\gamma = 1/\beta$, with $\varphi_\gamma^{\mathrm{DR}}$ 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 $(x^-, y^-, z^-) \to (x, y, z) \to (x^+, z^+, y^+)$ with penalty β and relaxation λ. Let $r = Ax + Bz - b$ and C be as in (2.5).*

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

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

with the constant C being strictly positive provided that

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

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

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

and C is strictly positive.

3. The linesearch algorithms

As shown in [22, 20, 41], 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 3 (page 9), 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. Differently from the convex setting of [28], 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) [37, §4] already benchmarked with the PANOC [36] and NAMA [35] solvers, respectively based on the forward-backward splitting and the alternating minimization algorithm.

The chosen linesearch allows us to discard differentiability requirements and merely exploit 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

Algorithm 1. Linesearch DRS

REQUIRE

initial point $s^0 \in \mathbb{R}^p$; tolerance $\varepsilon > 0$; relaxation $\lambda \in (0,2)$; max # backtracks $i_{\max} \leq \infty$

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

1.0: $k = 0$; $(u^0, v^0) \in \mathrm{DRS}_\gamma(s^0)$; compute $\varphi_\gamma^{\mathrm{DR}}(s^0)$ using u^0 and v^0 as in (2.2)

1.1: $r^k = u^k - v^k$; **if** $\|r^k\| \leq \varepsilon$ **then return**; **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 $(u^{k+1}, v^{k+1}) \in \mathrm{DRS}_\gamma(s^{k+1})$ and use it to evaluate $\varphi_\gamma^{\mathrm{DR}}(s^{k+1})$ as in (2.2)

1.6: **if** $\pi\varphi_\gamma^{\mathrm{DR}}(s^{k+1}) \leq \pi\varphi_\gamma^{\mathrm{DR}}(s^k) - \frac{c}{\gamma}\|r^k\|^2$ **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 $(u^{k+1}, v^{k+1}) \in \mathrm{DRS}_\gamma(s^{k+1})$; $k \leftarrow k+1$ 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

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

UNDER ASSUMPTION II (set $\pi := 1$)	UNDER ASSUMPTION II* (set $\pi := -1$)
penalty β as in (2.11)	penalty $\beta < \mu_f/\|A\|^2$
constant $0 < c < C(L_{(Af)}/\beta, \lambda)$ as in (2.5)	constant $0 < c < C(\beta\|A\|^2/\mu_f, \lambda)$ as in (2.5)

2.0: $k = 0$; $r^{-1} = Ax^{-1} + Bz^{-1} - b$; $y^{-1/2} = y^{-1} - \beta(1-\lambda)r^{-1}$; $(x^0, y^0, z^0) \in \mathrm{ADMM}_\beta(y^{-1/2}, z^{-1})$

2.1: $r^k = Ax^k + Bz^k - b$; **if** $\|r^k\| \leq \varepsilon$ **then return**; **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 $(x^{k+1}, y^{k+1}, z^{k+1}) \in \mathrm{ADMM}_\beta(y^{k+1/2}, z^k)$ and evaluate $\mathscr{L}_\beta(x^{k+1}, y^{k+1}, z^{k+1})$

2.6: **if** $\pi\mathscr{L}_\beta(x^{k+1}, z^{k+1}, y^{k+1}) \leq \pi\mathscr{L}_\beta(x^k, z^k, y^k) - \beta c\|r^k\|^2$ **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

$(x^{k+1}, y^{k+1}, z^{k+1}) \in \mathrm{ADMM}_\beta(\bar{y}^{k+1/2}, z^k)$; $k \leftarrow k+1$ 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

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 β and relaxation λ, starting from (x^{-1}, y^{-1}, z^{-1}). For each $k \in \mathbb{N}$, let*

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

Then, $\varphi = \varphi_1 + \varphi_2$ with $\varphi_1 = (Af)$ and $\varphi_2 = (Bg)(b - \cdot)$ satisfies 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 λ, 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, $\varphi_\gamma^{\mathrm{DR}}(s^k) = \mathscr{L}_\beta(x^k, z^k, y^k)$ 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

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

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

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 initital 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

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

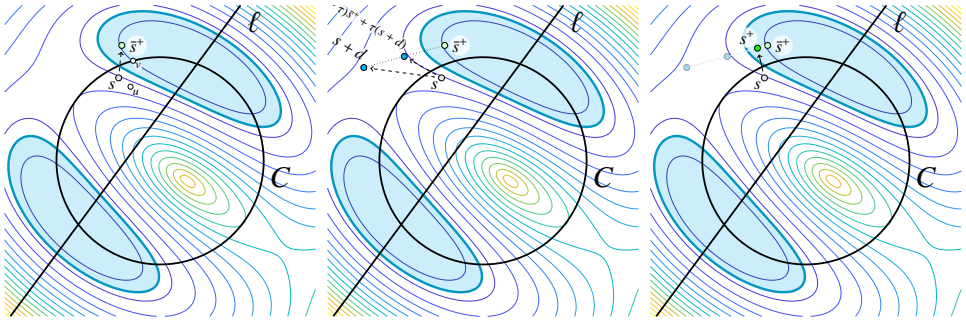
For $\tau \in \mathbb{R}$, let $\tilde{s}_\tau^{k+1} := (1 - \tau)\bar{s}^{k+1} + \tau(s^k + d^k)$ and $\tilde{y}_\tau^{k+1/2} := (1 - \tau)\bar{y}^{k+1/2} + \tau(y^k - \beta(r^k + d^k))$ be the candidate updates with stepsize τ at step 1.4 and step 2.4, respectively. We have

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

It then follows from (3.2) that $\varphi_\gamma^{\mathrm{DR}}(\tilde{s}_\tau^{k+1}) = \mathscr{L}_\beta(\tilde{x}_\tau^{k+1}, \tilde{z}_\tau^{k+1}, \tilde{y}_\tau^{k+1})$ for any $(\tilde{x}_\tau^{k+1}, \tilde{z}_\tau^{k+1}, \tilde{y}_\tau^{k+1}) \in \mathrm{ADMM}_\beta(\tilde{y}_\tau^{k+1/2}, z^k)$. 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. □



(a) Nominal DRS iterates ***(b)*** Failed linesearch trials ***(c)*** Stepsize accepted

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 $\frac{1}{\gamma}C\|\frac{v-u}{\gamma}\|^2$. Since $\varphi_\gamma^{\mathrm{DR}}$ is continuous and $c \lneqq C$, all points close enough to $\bar{s}^+$ belong to the sublevel set $[\varphi_\gamma^{\mathrm{DR}} \le \varphi_\gamma^{\mathrm{DR}}(s) - c\|u - v\|^2]$ (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 τ.*

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* [24], see also [19, §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 γ 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 $\varphi_\gamma^{\mathrm{DR}}$ (cf. Fact 2.2), it holds that

$$\varphi_\gamma^{\mathrm{DR}}(\bar{s}^+) \le \varphi_\gamma^{\mathrm{DR}}(s) - \tfrac{1}{\gamma}C(\gamma L_{\varphi_1}, \lambda)\|r\|^2$$

where C is as in (2.5) and $r = u - v$. However, nothing can be guaranteed as to whether $\varphi_\gamma^{\mathrm{DR}}(s + d)$ is also (sufficiently) smaller than $\varphi_\gamma^{\mathrm{DR}}(s)$ 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 $\varphi_\gamma^{\mathrm{DR}}$ (which is continuous but not necessarily differentiable, on top of the fact that the direction d is even arbitrary). Nevertheless, for any $c \lneqq C$, not only does $\bar{s}^+$ satisfy the sufficent decrease with constant c, but due to continuity of $\varphi_\gamma^{\mathrm{DR}}$ so do all the points around: loosely speaking,

$$\varphi_\gamma^{\mathrm{DR}}(s') \le \varphi_\gamma^{\mathrm{DR}}(s) - \tfrac{c}{\gamma}\|r\|^2 \quad \text{for all } s' \textit{ close to } \bar{s}^+. \tag{3.4}$$

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 ℓ and a circumference C, cast in the form (1.1) as

$$\operatorname*{\mathbf{minimize}}_{s\in\mathbb{R}^2} \varphi(s) \coloneqq \underbrace{\tfrac{1}{2}\operatorname{\mathbf{dist}}^2(s,\ell)}_{\varphi_1(s)} + \underbrace{\delta_C(s)}_{\varphi_2(s)}.$$

The contour levels correspond to those of $\varphi_\gamma^{\mathrm{DR}}$; notice that, since $\operatorname{\mathbf{prox}}_{\gamma\varphi_1} \equiv \mathrm{id}$ on ℓ and that $\operatorname{\mathbf{arg\,min}}\varphi \subset \ell$, for this problem it holds that $\operatorname{\mathbf{arg\,min}}\varphi = \operatorname{\mathbf{arg\,min}}\varphi_\gamma^{\mathrm{DR}}$, 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 $\varphi_\gamma^{\mathrm{DR}}$ or $\mathscr{L}_\beta$, 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.

Whenever function φ_1 is quadratic (not necessarily convex), however, the linearity of the u-update can conveniently be exploited to save computations in the linesearch. In fact, if φ_1 is quadratic, then $\operatorname{\mathbf{prox}}_{\gamma\varphi_1}$ is affine and the u-update at step 1.4 can be expanded to

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

Then, for $\tau_k = 1/2$ instead of computing directly u^{k+1} one can evaluate $\operatorname{\mathbf{prox}}_{\gamma\varphi_1}(\bar{s}^{k+1})$ and obtain u^{k+1} by linear combination, and similarly all subsequent trials for smaller values of τ_k will not require any additional evaluation of $\operatorname{\mathbf{prox}}_{\gamma\varphi_1}$. That is, in case φ_1 is quadratic, the number of calls to $\operatorname{\mathbf{prox}}_{\gamma\varphi_1}$ at iteration k will be $\operatorname{\mathbf{min}}\{1+i_k, 2\} \le 2$, where i_k is the number of failures of the linesearch condition at step 1.6, regardless of whether or not a maximum number of backtrackings $i_{\max}$ is set. This is particularly appealing when, additionally, evaluating $\operatorname{\mathbf{prox}}_{\gamma\varphi_2}$ is cheap (projections on simple sets, thresholding, ...), in which case each iteration is, at most, roughly twice as expensive as one of DRS. Such an example is discussed in Section 5; similar arguments also apply to the ADMM-Algorithm 2.

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 last part of the section provides an overview on some convenient choices of update directions d^k. Specifically, we propose a generalization of Nesterov extrapolation that is suited for nonconvex problems (although merely heuristical, without optimal convergence guarantees), and then discuss three popular quasi-Newton schemes. 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 [38, §15.6], useful for deriving directions based on *linear Newton approximation schemes* [11, §7.5.1].

3.3.1. *Nesterov acceleration.* It was shown in [28] that when φ_2 is convex and φ_1 is convex quadratic, then $\varphi_\gamma^{\mathrm{DR}}$ 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 [25] to DRS. By using duality arguments, this fact was extended in [29] where, under due assumptions, an *accelerated* ADMM scheme is proposed.

Although not supported by the theory, extensive numerical evidence suggests that such extrapolations perform quite effectively regardless of what function φ_2 or g in problems (1.1) and (1.2) are, while convexity of φ_1 and f seems instead to play an important role.

The main limitation in a direct employment of the acceleration is that convergence itself is not guaranteed. However, thanks to the arbitrarity 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, start with $d^0 = 0$, and for $k \geq 1$ select

$$d^k := \tfrac{k-1}{k+2}(\bar{s}^{k+1} - \bar{s}^k) - \lambda r^k$$

♠ Fast ADMM: in Algorithm 2, start with $d^0 = 0$, and for $k \geq 1$ select

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

These extensions differ from the approach proposed in [23] 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* [33, §13.F], eventually the updates $s \mapsto s^+$ are uniquely determined, that is, the inclusion $v \in \mathbf{prox}_{\gamma\varphi_2}(2u - s)$ 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 [42] 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 ($R_\gamma^{\mathrm{DR}}(s) = u - v$ for DRS and $R_\beta^{\mathrm{ADMM}}(y^{-/2}, z^{-1}) = Ax + Bz - b$ 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 the difference of the fixed-point residuals. Namely,

♠ quasi-Newton DRS: in Algorithm 1 use

$$d^k = -H_k r^k \quad \text{and} \quad \begin{cases} p^k = d^k \\ q^k = r_0^{k+1} - r^k \end{cases} \tag{3.5a}$$

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

♠ quasi-Newton ADMM: start with $H_0 = \mu \mathrm{I}$ for some $\mu > 0$, and in Algorithm 2 use

$$d^k = -H_k r^k \quad \text{and} \quad \begin{cases} p^k = d^k \\ q^k = r_0^{k+1} - r^k \end{cases} \tag{3.5b}$$

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

As it will be clear in the proof of Theorem 4.8, this particular choice of p_k and q_k rather than the conventional $p_k = s^{k+1} - s^k$ and $q_k = r^{k+1} - r^k$ is suited for 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 [32]. The limited-memory variant L-BFGS [26, 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 is *similar* to a symmetric matrix.

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

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

where

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

with the convention that $\mathbf{sgn}\,0 = 1$. The original Broyden formula [9] corresponds to $\vartheta_k \equiv 1$, while ϑ_k as in (3.6b) ensures nonsingularity of all matrices H_k [31].

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

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

where the columns of matrix $\mathcal{P}_k$ are the last vectors $p_{k-M}, \cdots, p_{k-1}$ and those of Q_k are the last vectors $q_{k-M}, \cdots, q_{k-1}$, with $M = \mathbf{min}\,\{k, m\}$. If Q_k is not full-column rank, for $x \in \mathbb{R}^M$ the product $(Q_k^\top Q_k)^{-1} x$ 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 [12] of *(inverse) multi-secant* update: H_k is the matrix closest to the identity (in the Frobenius norm) among those satisfying $H_k Q_k = \mathcal{P}_k$.

3.4. **Adaptive variants.** One drawback of Algorithms 1 and 2 is that both the stepsize γ in the former and the penalty β 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 [41, §4.1 and §5.3], where γ and β 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 γ according to an estimate of L_{φ_1} or μ_{φ_1}, and simply add the following routine after step 1.2:

1.2a: Evaluate $\varphi_\gamma^{\text{DR}}(\bar s^{k+1})$ using $(\bar u^{k+1}, \bar v^{k+1}) \in \mathrm{DRS}_\gamma(\bar s^{k+1})$ as in (2.2)

1.2b: **if** $\pi\varphi_\gamma^{\text{DR}}(\bar s^{k+1}) \geq \pi\varphi_\gamma^{\text{DR}}(s^k) - \frac{c}{\gamma}\|r^k\|^2$ **or** $\pi\varphi_\gamma^{\text{DR}}(\bar s^{k+1}) < \pi\varphi_{\text{LB}}$ **then**

$\gamma \leftarrow 2^{-\pi}\gamma$, $(u^k, v^k) \in \mathrm{DRS}_\gamma(s^k)$, 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 φ_{LB}, to be set offline equal to a known quantity that lower bounds $\mathbf{inf}\,\varphi$, 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 [41, §4.1]. As documented in the reference, this backtracking on γ can happen only a finite number of times, as eventually the condition at step 1.2b is never satisfied.

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 β according to an estimate of $L_{(Af)}$ or $\mu_{(Af)}$, setting offline a lower bound $\Phi_{\text{LB}} \leq \mathbf{inf}\,\{f(x)+g(z) \mid Ax+Bz=b\}$ (if known), and adding the following check after step 2.2:

2.2a: Evaluate $\mathscr{L}_\beta(\bar{x}^{k+1},\bar{z}^{k+1},\bar{y}^{k+1})$ 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** $\pi\mathscr{L}_\beta(\bar{x}^{k+1},\bar{z}^{k+1},\bar{y}^{k+1}) \geq \pi\mathscr{L}_\beta(x^k,z^k,y^k) - \beta c\|r^k\|^2$ **or** $\pi\mathscr{L}_\beta(\bar{x}^{k+1},\bar{z}^{k+1},\bar{y}^{k+1}) < \pi\Phi_{\text{LB}}$ **then** $\beta \leftarrow 2^\pi\beta$, $(x^k,z^k,y^k) \in \text{ADMM}_\beta(y^{k-1/2},z^{k-1})$, recompute $\mathscr{L}_\beta(x^k,z^k,y^k)$ 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 $K \leq \frac{\gamma|\varphi_\gamma^{\text{DR}}(s^0) - \mathbf{min}\,\varphi|}{c\varepsilon^2}$ iterations;

(iii) the last iterate of the algorithm yields

- *a point $z := v^K$ satisfying $\mathbf{dist}(0,\hat{\partial}\varphi(z)) \leq 2\varepsilon/\gamma$ if Assumption I holds,*
- *a triplet $(x,y,z) := (u^K, \gamma^{-1}(u^K - s^K), v^K)$ satisfying the approximate KKT*

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

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

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

Since $c < C$ by construction, one has

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

Continuity of $\varphi_\gamma^{\text{DR}}$ 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 ϵ-close to $\varphi_\gamma^{\text{DR}}(s^k)$. Since $(1-\tau_k)\bar{s}^{k+1} + \tau_k(s^k + d^k) \to \bar{s}^{k+1}$ as $\tau_k \to 0$, by halvening enough times τ_k at step 1.8,

- either the candidate point s^{k+1} is eventually ϵ-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

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

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

$$\pi\varphi_\gamma^{\mathrm{DR}}(s^{K+1}) \le \pi\varphi_\gamma^{\mathrm{DR}}(s^0) - \tfrac{c}{\gamma}\textstyle\sum_{k=0}^{K}\|r^k\|^2.$$

Since $\pi\,\mathbf{min}\,\varphi = \mathbf{min}\,\pi\varphi_\gamma^{\mathrm{DR}}$ and $\|r^k\| \ge \varepsilon$, we conclude that

$$\pi\varphi_\gamma^{\mathrm{DR}}(s^0) - \pi\,\mathbf{min}\,\varphi \ge \tfrac{c}{\gamma}\textstyle\sum_{k=0}^{K}\|r^k\|^2 \ge \frac{c(K+1)\varepsilon^2}{\gamma}, \tag{4.2}$$

resulting in the claimed bound on K. Finally, the bound on $\mathbf{dist}(0, \hat\partial\varphi(z))$ follows from [41, Thm. 4.3(ii)].

Suppose now that Assumption I* holds. Then, since $\pi = -1$, it follows from Theorem 2.3 that $\pi\varphi_\gamma^{\mathrm{DR}}(s) = \psi_{\gamma_*}^{\mathrm{DR}}(s_*)$, 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 (1.9) it follows that $-y \in \partial\varphi_1(x)$ and that $y + \frac{1}{\gamma}r^K \in \partial\varphi_2(z)$, hence that $\mathbf{dist}(y, \partial\varphi_2(z)) \le \varepsilon$, owing to the termination criterion $\|r^K\| \le \varepsilon$. Finally, $\|x - z\| = \|r^K\| \le \varepsilon$, completing the proof. □

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 $K \le \frac{|\mathscr{L}_\beta(x^0,z^0,y^0) - \mathbf{min}\,\Phi|}{c\beta\varepsilon^2}$ iterations and yields a triplet $(x, y, z) := (x^K, y^K, z^K)$ satisfying the approximate KKT conditions

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

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 $\mathbf{dist}(-B^\top y, \hat\partial g(z)) \le \|B\|\beta\varepsilon$ 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 β. □

4.1. **Subsequential convergence.** The remainder of the section is dedicated to asymptotic analysis where the tolerance is set to $\varepsilon = 0$ and the algorithms may thus run infinitely many iterations. In this first subsection, without imposing additional requirements other than one among Assumptions I, I*, II and II*, we show that every limit point of the generated sequences are stationary, and also give a sufficient condition ensuring boundedness. 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 φ and on which φ and $\varphi_\gamma^{\mathrm{DR}}$ have the same (constant) value, this being the limit of the monotonic sequence $(\varphi_\gamma^{\mathrm{DR}}(s^k))_{k\in\mathbb{N}}$;

(ii) if φ is level bounded, then $(s^k, u^k, v^k)_{k\in\mathbb{N}}$ remains bounded.

UNDER ASSUMPTION I*:

(iii) $\varphi_\gamma^{\mathrm{DR}}(s^k) \nearrow \mathbf{min}\,\varphi$ as $k \to \infty$, $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ converge to the unique (global) minimizer $x_\star$ of φ;

(iv) if the dual cost ψ as in (2.7) is level bounded, then $(s^k)_{k\in\mathbb{N}}$ remains 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 $K \to \infty$ in (4.2). In turn, this implies that $u^k - v^k \to 0$ as $k \to \infty$. We now consider the two separate cases.

♠ Assumption I. Having shown the vanishing of the residual, the proof is identical to that of [41, 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 $\psi^{\mathrm{DR}}_{\gamma_*}(s^k_*) = -\varphi^{\mathrm{DR}}_{\gamma}(s^k)$ is decreasing and bounded below by $\mathbf{inf}\,\psi = -\mathbf{min}\,\varphi$ (cf. (A.2)). In particular, $\psi(v^k_*) \le \psi^{\mathrm{DR}}_{\gamma_*}(s^k_*)$ is also bounded (the inequality following from Fact A.1), and since $r^k = u^k - v^k = \frac{u^k_* - v^k_*}{\gamma_*} \to 0$, from Lemma A.4 we conclude that $\psi(v^k_*) \to \mathbf{inf}\,\psi$. From Lemma A.2 we have that $\psi^{\mathrm{DR}}_{\gamma_*}(s^k_*) \le \psi(v^k_*) + \frac{1+\gamma L_{\psi^\perp}}{2\gamma_*}\|r^k_*\|^2$, implying that $\psi^{\mathrm{DR}}_{\gamma_*}(s^*_k) \to \mathbf{inf}\,\psi$, and from (A.3) we conclude that $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ converge to the unique minimizer of the primal function φ, hence assertion 4.3*(iii)*. Finally, since $(s^k_*)_{k\in\mathbb{N}}$ is contained in the sublevel set $\left\{x \mid \psi^{\mathrm{DR}}_{\gamma_*}(x) \le \varphi^{\mathrm{DR}}_{\gamma}(s^0_*)\right\}$, if ψ is coercive (and thus so is $\psi^{\mathrm{DR}}_{\gamma_*}$ owing to Fact 2.1*(ii)*) the sequence $(s^k_*)_{k\in\mathbb{N}}$ is bounded. Boundedness of $(s^k)_{k\in\mathbb{N}}$ then follows from the identity $s^k = -s^k_*/\gamma_*$. □

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

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

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

(ii) the sequence $(Ax^k, y^k, Bz^k)_{k\in\mathbb{N}}$ is bounded provided that the cost function Φ 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) $(\mathscr{L}_\beta(x^k, z^k, y^k))_{k\in\mathbb{N}} \nearrow \mathbf{min}\,\Phi$ as $k \to \infty$, $(Ax^k)_{k\in\mathbb{N}}$ and $(Bz^k)_{k\in\mathbb{N}}$ are convergent, and all 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 the dual cost

$$\Psi(y) := f^*(-A^\top y) + g^*(-B^\top y) + \langle b, y\rangle$$

is coercive, 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 [41, Thm. 5.6].

♠ Assumption II*. It follows from Theorem 4.3*(iii)* that $\varphi^{\mathrm{DR}}_{\gamma}(s^k) = \mathscr{L}_\beta(x^k, z^k, y^k) \nearrow \mathbf{inf}\,\Phi$, and that $(Ax^k = u^k)_{k\in\mathbb{N}}$ and $(b - Bz^k = v^k)_{k\in\mathbb{N}}$ are convergent. Suppose now that a subsequence $(x^k, y^k, z^k)_{k\in K}$ converges to a triplet (x, y, z). Since $r^k \to 0$, necessarily $Ax + Bz = b$. Moreover,

$$\mathbf{inf}\,\Phi = \lim_{K\ni k\to\infty} \mathscr{L}_\beta(x^k, z^k, y^k) = \lim_{K\ni k\to\infty} f(x^k) + g(z^k) \ge f(x) + g(z) = \Phi(x, z) \ge \mathbf{inf}\,\Phi,$$

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 $f(x^k) + g(z^k) \to f(x) + g(z)$ implies that $f(x^k) \to f(x)$ and $g(z^k) \to g(z)$. Therefore, the attentive outer semicontinuity of the subdifferential [33, Prop. 8.7] together with Facts 2.5*(i)* and 2.5*(ii)* implies the sought inclusions $-A^\top y \in \partial f(x)$ and $-B^\top y \in \partial g(z)$, which completes the proof of assertion 4.4*(iii)*. Finally, assertion 4.4*(iv)* is a direct consequence of Theorem 4.3*(iv)*, after observing that

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

for φ_1 and φ_2 as in (2.10), see [4, Prop.s 13.23(iii) and 13.24(iv)]. □

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 [11, §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

$$\lim_{k\to\infty} \frac{\|s^k + d^k - s_\star\|}{\|s^k - s_\star\|} = 0. \tag{4.4}$$

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

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

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

where the first inequality follows from $(1+\gamma L_{\varphi_1})$-strong monotonicity of $\mathbf{prox}_{\gamma\varphi_1}$ [41, Prop. 2.3(ii)] and the second one uses the Young's inequality. As ensured by Theorem 4.3*(i)*, $u^k - v^k \to 0$ and $\varphi(u_\star) = \varphi_\star$. Therefore, $v^k \to u_\star$ 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 $\varepsilon > 0$ and $k \geq K$ we thus have

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

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

Theorem 4.6 (Acceptance of the unit stepsize)**.** *Suppose that Assumption I 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 φ 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. Let $s_\star := u_\star - \gamma\nabla\varphi_1(u_\star)$ be the limit point of $(s^k)_{k\in\mathbb{N}}$. In light of Lemma 4.5, by possibly discarding the first iterates we may assume that

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

for some $\delta > 0$. Combined with Lemma A.2, we obtain that

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

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

$$\begin{aligned}
\varphi_\gamma^{\mathrm{DR}}(s^k + d^k) - \varphi_\gamma^{\mathrm{DR}}(s^k) &= -(1-\varepsilon_k)(\varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi_\gamma^{\mathrm{DR}}(s_\star)) \\
&\leq -(1-\varepsilon_k)(\varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi_\gamma^{\mathrm{DR}}(\bar{s}^k)) \\
&\overset{(2.5)}{\leq} -(1-\varepsilon_k)\tfrac{c}{\gamma}\|r^k\|^2.
\end{aligned}$$

Since $\varepsilon_k \to 0$ and $c < C$ as required in Algorithm 1, eventually $1 - \varepsilon_k \geq \frac{c}{C}$, resulting in $\varphi_\gamma^{\mathrm{DR}}(s^k + d^k) \leq \varphi_\gamma^{\mathrm{DR}}(s^k) - \frac{c}{\gamma}\|r^k\|^2$, proving that $\tau_k = 1$ passes the condition at step 1.6. □

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

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

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

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

is the residual of *forward-backward splitting* (FBS), see e.g., [2, 8, 42]. By combining [33, Prop. 13.24 and Ex. 13.35] and [30, Thm. 4.4(c)-(f) and Cor. 4.7], it follows that $\mathbf{prox}_{\gamma\varphi_1}(s)$ is continuously differentiable around $s_\star$ provided that φ_1 is twice continuously differentiable around $s_\star$. Thus, under this assumption from the chain rule of differentiation we conclude that R_γ^{DR} is (strictly) differentiable at $s_\star$ provided that R_γ^{FB} is (strictly) differentiable at $u_\star = \mathbf{prox}_{\gamma\varphi_1}(s_\star)$. Sufficient conditions for this latter property to hold are documented in [42, Thm. 4.10], namely twice (Lipschitz-) continuous differentiability of φ_1 around $u_\star$, and prox-regularity and (strict) twice epi-differentiability of φ_2 at $u_\star$ for $-\nabla\varphi_1(u_\star)$ [33, §13.B and 13.F].

Theorem 4.7 (Dennis-Moré criterion for superlinear directions)**.** *Suppose that Assumption I 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_γ^{DR} is strictly differentiable and with nonsingular Jacobian $JR_\gamma^{\mathrm{DR}}(s_\star)$. If the Dennis-Moré condition*

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

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

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

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

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

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

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

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_γ^{DR}, a condition that entails classical differentiability at the limit point but not necessarily around it, see [18].

Theorem 4.8 (Superlinear convergence with Broyden directions)**.** *Suppose that Assumption I 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 φ at which R_γ^{DR} is Lipschitz-semidifferentiable and has a nonsingular Jacobian $JR_\gamma^{\mathrm{DR}}(s_\star)$. 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 $G_\star := JR_\gamma^{\mathrm{DR}}(s_\star)$. Since R_γ^{DR} is differentiable at $s_\star$ with nonsingular Jacobian and since $s^k \to s_\star$, for k large enough $R_\gamma^{\mathrm{DR}}(s^k)$ is single valued and satisfies $\|R_\gamma^{\mathrm{DR}}(s^k)\| \geq \alpha\|s^k-s_\star\|$ for some $\alpha > 0$. Moreover, it follows from [18, Lem. 2.2] that an $L > 0$ exists such that

$$\begin{aligned}\frac{\|q^k - G_\star p^k\|}{\|p^k\|} = \frac{\|R_\gamma^{\mathrm{DR}}(s^k+d^k) - R_\gamma^{\mathrm{DR}}(s^k) - G_\star d^k\|}{\|d^k\|} &\leq L\,\mathbf{max}\left\{\|s^k+d^k-s_\star\|, \|s^k-s_\star\|\right\} \\ &\leq L(\|s^k-s_\star\| + \|d^k\|) \end{aligned} \tag{4.7}$$

for k large enough. Since $\mathbf{prox}_{\gamma\varphi_1}$ is $\frac{1}{1-\gamma L_{\varphi_1}}$-Lipschitz continuous [41, Prop. 2.3(ii)], denoting $u_\star := \mathbf{prox}_{\gamma\varphi_1}(s_\star)$ we have that

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

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

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

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_γ^{DR} is differentiable at $s_\star$, there exists $\alpha'' > 0$ such that $\|R_\gamma^{\mathrm{DR}}(s^k)\| = \|R_\gamma^{\mathrm{DR}}(s^k) - R_\gamma^{\mathrm{DR}}(s_\star)\| \leq \alpha''\|s^k-s_\star\|$. Therefore, also $(\|R_\gamma^{\mathrm{DR}}(s^k)\|)_{k\in\mathbb{N}}$ 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 $d^k = -H_k R_\gamma^{\mathrm{DR}}(s^k)$. 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 [40, Thm. VI.8], after observing that $\frac{\|R_\gamma^{\mathrm{DR}}(s^k)+G_\star d^k\|}{\|d^k\|} = \frac{\|(H_k^{-1}-G_\star)p^k\|}{\|p^k\|}$. 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. □

5. Simulations

In this section we show the effectiveness of the proposed Algorithm 1 and Algorithm 2, for different choices of the linesearch direction, compared to the standard DRS and ADMM. The implementations in Julia of all the algorithms are available online as part of the *ProximalAlgorithms.jl* package.[3] All experiments were run using Julia 1.4.1.

[3] https://github.com/kul-forbes/ProximalAlgorithms.jl

5.1. **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:

$$\underset{x\in\mathbb{R}^n}{\textbf{maximize}}\ \tfrac{1}{m}\|Wx\|^2 \quad \textbf{subject to}\ \|x\| = 1,\ \|x\|_0 \le k, \tag{5.1}$$

where the ℓ_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, (5.1) amounts to a variance maximization problem. The formulation (5.1) was first introduced in [10], where the authors propose solving an SDP relaxation. Here, we consider tackling directly the nonconvex problem (5.1) instead.

Denoting the set of feasible points by

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

problem (5.1) takes the form (1.1) once we set $\varphi_1(x) = -\frac{1}{m}\|Wx\|^2$ and $\varphi_2(x) = \delta_{\mathcal{S}}(x)$. So formulated, the problem complies with Assumption I, therefore DRS can be readily applied. Note that, in this case,

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

where last equality uses the Woodbury identity. Whenever $\gamma < L_{\varphi_1}^{-1} = \frac{1}{m}\|W\|^2$, evaluating $\mathbf{prox}_{\gamma\varphi_1}$ amounts to solving square, positive definite linear system of dimension 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. On the other hand, evaluating $\mathbf{prox}_{\gamma\varphi_2} = \mathbf{\Pi}_{\mathcal{S}}$ amounts to setting to zero the $n-k$ smallest coefficients of x (in magnitude), and normalizing the resulting vector to project it on the ℓ_2-sphere: this has therefore a negligible cost.

Figure 2 shows the results when the algorithm is applied using a small subset of the *20newsgroup* dataset,[4] which only retains 100 features from the original dataset: $n = 16242$ and $m = 100$. This is the same dataset that was used in the experiments in [10]. 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 $\lambda = 1$, $\gamma = \frac{0.95}{2}L_{\varphi_1}^{-1}$, $c = \frac{1}{2}C(\gamma L_{\varphi_1}, \lambda)$, cf. (2.5), 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 (see remarks in Section 3.3.1).

5.2. **Sparse PCA: consensus formulation.** However, 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* $N \ge 1$, 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 $\boldsymbol{x} \in \mathbb{R}^{nN}$), and solve

$$\underset{x\in\mathbb{R}^{Nn}, z\in\mathbb{R}^n}{\textbf{minimize}} \sum_{i=1}^N -\tfrac{1}{m}\|W_i x_i\|^2 \quad \textbf{subject to}\ \|z\| = 1,\ \|z\|_0 \le k,\ x_i = z,\ i = 1\dots N.$$

[4] The dataset is available at: `https://cs.nyu.edu/~roweis/data.html`.

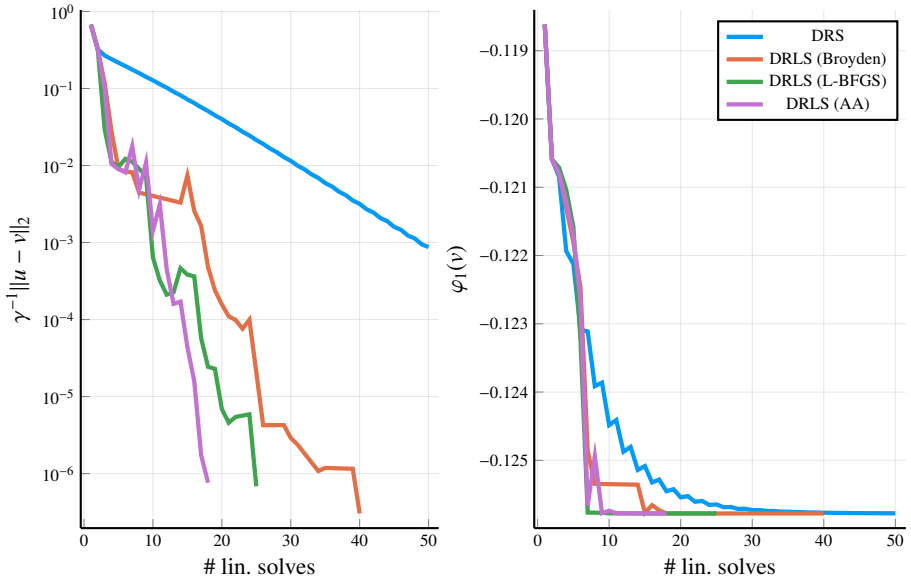


Figure 2. *Comparison between DRS and the linesearch variant Algorithm 2, using the modified Broyden, L-BFGS, and Anderson acceleration directions, when applied to the sparse PCA problem (5.1). The dataset used here is a small subset of the* 20newsgroup *dataset, with 100 features only. We used a memory parameter of 5 for L-BFGS and Anderson acceleration.*

This problem is equivalent to (5.1), and can be expressed in ADMM form (1.2) as

$$\underset{\boldsymbol{x}\in\mathbb{R}^{Nn},\, z\in\mathbb{R}^n}{\textbf{minimize}} \underbrace{\sum_{i=1}^{N} -\tfrac{1}{m}\|W_i x_i\|^2}_{f(\boldsymbol{x})} + \underbrace{\delta_{\mathcal{S}}(z)}_{g(z)} \quad \textbf{subject to } \boldsymbol{x} = \begin{pmatrix} \mathrm{I} \\ \vdots \\ \mathrm{I} \end{pmatrix} z. \tag{5.2}$$

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 $L_{(Af)} = \frac{1}{m}\,\mathbf{max}_{i=1\ldots N}\,\|W_i\|^2 \le \frac{1}{m}\|W\|^2$ and A has clearly full row rank.

The z-update as prescribed by ADMM comes at negligible cost, since

$$\underset{z\in\mathbb{R}^n}{\arg\min}\left\{\delta_{\mathcal{S}}(z) + \tfrac{\beta}{2}\|Bz-\boldsymbol{x}\|^2\right\} = \boldsymbol{\Pi}_{\mathcal{S}}(\tfrac{1}{N}\textstyle\sum_{i=1}^N x_i) \quad \forall \boldsymbol{x}\in\mathbb{R}^{nN},$$

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

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

for $i = 1\ldots N$, where the second equality uses the Woodbury identity. The Cholesky factors of the $m_i\times m_i$ matrices $m\beta\mathrm{I} - W_i W_i^\top$, $i = 1,\ldots,N$, can be computed once offline to efficiently solve the linear systems at each z-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 3. 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. Notice that f is quadratic, hence linearity can be exploited in the linesearch as discussed in Section 3.2.

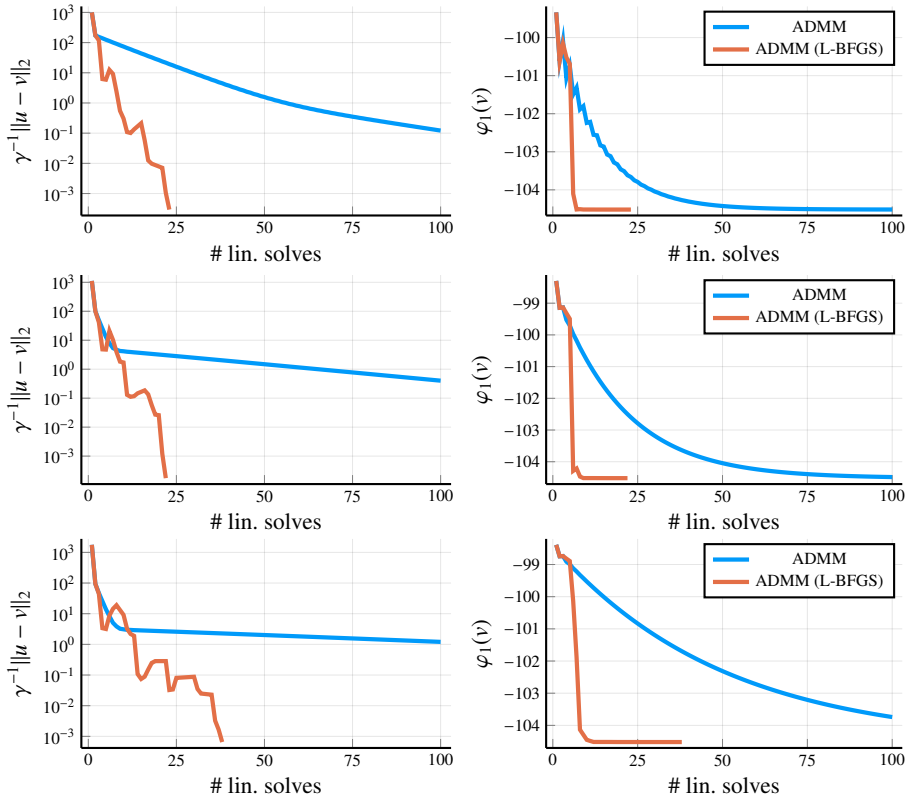


Figure 3. *Comparison between ADMM (in blue) and the L-BFGS enhancement (in red) for the consensus sparse PCA problem (5.2), using the full* 20newsgroup *dataset, for different number of agents* $N = 10, 20, 50$. *On the x-axis, the number of linear systems solved (needed for the* $\boldsymbol{x}$*-update), which in the case of plain ADMM coincides with the number of iterations. This is the unique expensive operation, as the x-update is negligible. On the left is the ADMM residual, on the right is the value of the cost function. Apparently, ADMM is severly affected by N, whereas using L-BFGS directions in Algorithm 2 consistently results in faster convergence.*

Figure 3 shows the result of running the proposed algorithm, with L-BFGS directions, to the full *20newsgroup* dataset:[5] this consists of the frequencies of $n = 26214$ words in $m = 18846$ documents. 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 (5.2). 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). In this case, we only considered L-BFGS directions with memory 5 since their computation scales better with the problem dimension. Both algorithms were started at the same initial iterates $y^0 = 0$ and $z^0 = (1/n, \ldots, 1/n) \in \mathbb{R}^n$. 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.

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

[5] Dataset available at `http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html`.

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.

Appendix A. Auxiliary results

This appendix contains some auxiliary results needed for the convergence analysis of Section 4. As shown in [41, Eq. (3.4)], the DRE can be expressed in terms of the forward-backward envelope $\varphi^{\mathrm{FB}}_\gamma$ [27, 34, 42] as

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

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

$$\varphi^{\mathrm{DR}}_\gamma(s) = \varphi_1(u) - \tfrac{\gamma}{2}\|\nabla\varphi_1(u)\|^2 + \varphi_2^\gamma(u - \gamma\nabla\varphi_1(u)). \tag{A.1b}$$

Fact A.1 ([41, Prop. 3.3])**.** *Suppose that Assumption I holds and let $\gamma < 1/L_{\varphi_1}$ be fixed. Then, for all $s \in \mathbb{R}^p$ and $(u, v) \in \mathrm{DRS}_\gamma(s)$ it holds that*

$$\varphi(v) + \tfrac{1-\gamma L_{\varphi_1}}{2\gamma}\|v-u\|^2 \le \varphi^{\mathrm{DR}}_\gamma(s) \le \varphi(u).$$

□

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

$$\varphi^{\mathrm{DR}}_\gamma(s) - \varphi(\bar{u}) \le \tfrac{1+\gamma L_{\varphi_1}}{2\gamma}\|\mathbf{prox}_{\gamma\varphi_1}(s) - \bar{u}\|^2.$$

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

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

where the second inequality uses the known quadratic upper bound [7, Prop. A.24] for functions with Lipschitz-continuous gradient. □

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

$$\inf\varphi = -\inf\psi = -\inf\psi^{\mathrm{DR}}_{\gamma_*} = \sup\varphi^{\mathrm{DR}}_\gamma \tag{A.2}$$

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

$$\tfrac{1}{2\gamma}\|x_\star - v\|^2 + \tfrac{\gamma\mu_{\varphi_1}-1}{2\gamma}\|x_\star - u\|^2 \le \inf\varphi - \varphi^{\mathrm{DR}}_\gamma(s) = \psi^{\mathrm{DR}}_{\gamma_*}(s_*) - \inf\psi \tag{A.3}$$

where $x_\star$ is the unique minimizer of φ and $(u, v) = \mathrm{DRS}_\gamma(s)$.

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 ψ_1 is $1/\mu_{\varphi_1}$-smooth, it follows from Fact 2.1*(i)* that $\inf\psi^{\mathrm{DR}}_{\gamma_*} = \inf\psi$ for every $\gamma_* < 1/\mu_{\varphi_1}$; combined with the identity $\varphi^{\mathrm{DR}}_\gamma(s) = -\psi^{\mathrm{DR}}_{\gamma_*}(-s/\gamma)$ holding for $\gamma_* = 1/\gamma$ (cf. Theorem 2.3), (A.2) is obtained. Let now $s \in \mathbb{R}^p$ be fixed and consider $(u, v) = \mathrm{DRS}_\gamma(s)$. From the inclusion $\frac{s-u}{\gamma} \in \partial\varphi_1(u)$ and strong convexity of φ_1 one has

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

where $x_\star = \arg\min\varphi$. Similarly, since $\frac{2u-s-v}{\gamma} \in \partial\varphi_2(v)$ and φ_2 is convex, one has

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

Summing the two inequalities yields

$$\begin{aligned}\inf\varphi &\geq \varphi_1(u)+\varphi_2(v)+\tfrac{1}{\gamma}\langle s-u, x_\star-u\rangle+\tfrac{1}{\gamma}\langle 2u-s-v, x_\star-v\rangle+\tfrac{\mu_{\varphi_1}}{2}\|x_\star-u\|^2\\ &= \varphi_1(u)+\varphi_2(v)+\tfrac{1}{\gamma}\langle s-u, v-u\rangle+\tfrac{1}{\gamma}\langle u-v, x_\star-v\rangle+\tfrac{\mu_{\varphi_1}}{2}\|x_\star-u\|^2\\ &= \varphi_1(u)+\varphi_2(v)+\tfrac{1}{\gamma}\langle s-u, v-u\rangle+\tfrac{1}{2\gamma}\|u-v\|^2+\tfrac{1}{2\gamma}\|x_\star-v\|^2+\tfrac{\gamma\mu_{\varphi_1}-1}{2\gamma}\|x_\star-u\|^2\\ &= \varphi^{\mathrm{DR}}_\gamma(s)+\tfrac{1}{2\gamma}\|x_\star-v\|^2+\tfrac{\gamma\mu_{\varphi_1}-1}{2\gamma}\|x_\star-u\|^2\\ &= -\psi^{\mathrm{DR}}_{\gamma_*}(s_*)+\tfrac{1}{2\gamma}\|x_\star-v\|^2+\tfrac{\gamma\mu_{\varphi_1}-1}{2\gamma}\|x_\star-u\|^2.\end{aligned}$$

The claim now follows from the identity $\psi^{\mathrm{DR}}_{\gamma_*}(s_*)=-\varphi^{\mathrm{DR}}_\gamma(s)$. □

Lemma A.4. *Additionally to Assumption I, suppose that φ_1 and φ_2 are convex functions. Let $(s^k)_{k\in\mathbb{N}}$ be a sequence such that $u^k-v^k\to 0$ as $k\to\infty$, where $(u^k,v^k)=\mathrm{DRS}_\gamma(s^k)$ with $\gamma\in(0,1/L_{\varphi_1})$. If $(\varphi(v^k))_{k\in\mathbb{N}}$ is bounded, then $\varphi(v^k)\to\inf\varphi$ as $k\to\infty$.*

Proof. Since $\gamma<1/L_{\varphi_1}=1/L_{-\varphi_1}$, the mapping $\mathbf{prox}_{-\gamma\varphi_1}$ is a Lipschitz homeomorphism on $\mathbb{R}^p$, characterized as $\tilde{u}=\mathbf{prox}_{-\gamma\varphi_1}(u)$ iff $u=\tilde{u}-\gamma\nabla\varphi_1(\tilde{u})$. Moreover, from (A.1b) we have

$$\begin{aligned}\varphi^{\mathrm{FB}}_\gamma\circ\mathbf{prox}_{-\gamma\varphi_1}(u)=\varphi^{\mathrm{FB}}_\gamma(\tilde{u})&=\varphi_1(\tilde{u})-\tfrac{\gamma}{2}\|\nabla\varphi_1(\tilde{u})\|^2+\varphi_2^\gamma(\tilde{u}-\gamma\nabla\varphi_1(\tilde{u}))\\ &=\varphi_1(\tilde{u})-\tfrac{1}{2\gamma}\|\tilde{u}-u\|^2+\varphi_2^\gamma(u)\\ &=\varphi_2^\gamma(u)-(-\varphi_1)^\gamma(u).\end{aligned}$$

This reparametrization of the forward-backward envelope has already been observed in [39, §III-A] and shown to be convex and Lipschitz-differentiable. For the sake of self-containedness we detail the proof. Since φ_2 is convex and $-\varphi_1$ is L_{φ_1}-smooth, the function

$$h(x):=\varphi^{\mathrm{FB}}_\gamma\circ\mathbf{prox}_{-\gamma\varphi_1}(x)=\varphi_2^\gamma(x)-(-\varphi_1)^\gamma(x)$$

is (Lipschitz) differentiable, with

$$\nabla h(x)=\tfrac{1}{\gamma}\Big(\mathbf{prox}_{-\gamma\varphi_1}(x)-\mathbf{prox}_{\gamma\varphi_2}(x)\Big)=\tfrac{1}{\gamma}\Big(\mathrm{id}-\mathbf{prox}_{\gamma\varphi_2}(\mathrm{id}-\gamma\nabla\varphi_1)\Big)\circ\mathbf{prox}_{-\gamma\varphi_1}(x).$$

Since $\mathbf{prox}_{-\gamma\varphi_1}$ is invertible and $\inf\varphi^{\mathrm{FB}}_\gamma=\inf\varphi$ [42, Thm. 4.4], one has $\inf h=\inf\varphi$ as well. We now show that h is convex. Let $x_i\in\mathbb{R}^p$, $i=1,2$, be arbitrary, and let us denote $\tilde{x}_i:=\mathbf{prox}_{-\gamma\varphi_1}(x_i)$ so that $x_i=\tilde{x}_i-\gamma\nabla\varphi_1(\tilde{x}_i)$. We have

$$\begin{aligned}\langle\nabla h(x_1)-\nabla h(x_2),x_1-x_2\rangle&=\tfrac{1}{\gamma}\langle\tilde{x}_1-\tilde{x}_2,x_1-x_2\rangle-\tfrac{1}{\gamma}\langle\mathbf{prox}_{\gamma\varphi_2}(x_1)-\mathbf{prox}_{\gamma\varphi_2}(x_2),x_1-x_2\rangle\\ &=\tfrac{1}{\gamma}\langle(\tilde{x}_1-x_1)-(\tilde{x}_2-x_2),x_1-x_2\rangle\\ &\quad+\tfrac{1}{\gamma}\underbrace{\langle(\mathrm{id}-\mathbf{prox}_{\gamma\varphi_2})(x_1)-(\mathrm{id}-\mathbf{prox}_{\gamma\varphi_2})(x_2),x_1-x_2\rangle}_{\geq 0}\\ &\geq\langle\nabla\varphi_1(\tilde{x}_1)-\nabla\varphi_1(\tilde{x}_2),(\tilde{x}_1-\gamma\nabla\varphi_1(\tilde{x}_1))-(\tilde{x}_2-\gamma\nabla\varphi_1(\tilde{x}_2))\rangle\\ &\geq\Big(\tfrac{1}{L_{\varphi_1}}-\gamma\Big)\|\nabla\varphi_1(\tilde{x}_1)-\nabla\varphi_1(\tilde{x}_2)\|^2\geq 0,\end{aligned}$$

proving that h is convex. Here, the first inequality follows from monotonicity of $\mathrm{id}-\mathbf{prox}_{\gamma\varphi_2}$ [4, Prop. 12.28], and the second one from $\frac{1}{L_{\varphi_1}}$-cocoercivity of $\nabla\varphi_1$ [4, Cor. 18.17].

Let now $(s^k,u^k,v^k)_{k\in\mathbb{N}}$ be as in the statement and observe that

$$\nabla h(u^k-\gamma\nabla\varphi_1(u^k))=\tfrac{1}{\gamma}\Big(u^k-\mathbf{prox}_{\gamma\varphi_2}\big(u^k-\gamma\nabla\varphi_1(u^k)\big)\Big)=\tfrac{1}{\gamma}(u^k-v^k)\to 0\quad\text{as } k\to\infty.$$

Since h is convex with $\mathbf{dom}\,h=\mathbb{R}^p$, we infer from [3, Thm. 4.2.2] that $h(u^k-\gamma\nabla\varphi_1(u^k))\to\inf h=\inf\varphi$ as $k\to\infty$. Therefore,

$$\inf\varphi\leq\varphi(v^k)\overset{\text{A.1}}{\leq}\varphi^{\mathrm{DR}}_\gamma(s^k)=\varphi^{\mathrm{FB}}_\gamma(u^k)=h(u^k-\gamma\nabla\varphi_1(u^k))\to\inf\varphi,$$

hence $\varphi(v^k)\to\inf\varphi$ as claimed. □

REFERENCES

[1] Donald G. Anderson. Iterative procedures for nonlinear integral equations. *J. ACM*, 12(4):547–560, oct 1965.

[2] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. 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, Feb 2013.

[3] Alfred Auslender and Marc Teboulle. *Asymptotic Cones and Functions in Optimization and Variational Inequalities*. Springer Monographs in Mathematics. Springer New York, 2002.

[4] Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics. Springer, 2017.

[5] Heinz H. Bauschke and Dominikus Noll. On the local convergence of the Douglas-Rachford algorithm. *Archiv der Mathematik*, 102(6):589–600, Jun 2014.

[6] Heinz H. Bauschke, Hung M. Phan, and Xianfu Wang. The method of alternating relaxed projections for two nonconvex sets. *Vietnam Journal of Mathematics*, 42(4):421–450, Dec 2014.

[7] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2016.

[8] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal Alternating Linearized Minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494, 2014.

[9] Charles G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 19(92):577–593, 1965.

[10] Alexandre d'Aspremont, Laurent E Ghaoui, Michael I Jordan, and Gert R Lanckriet. A direct formulation for sparse PCA using semidefinite programming. In *Advances in neural information processing systems*, pages 41–48, 2005.

[11] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*, volume II. Springer, 2003.

[12] Haw-ren Fang and Yousef Saad. Two classes of multisecant methods for nonlinear acceleration. *Numerical Linear Algebra with Applications*, 16(3):197–221, 2009.

[13] Max L. N. Goncalves, Jefferson G. Melo, and Renato D. C. Monteiro. 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.

[14] Ke Guo, Deren Han, and Ting-Ting Wu. 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.

[15] Robert Hesse, Russel Luke, and Patrick Neumann. Alternating projections and Douglas-Rachford for sparse affine feasibility. *IEEE Transactions on Signal Processing*, 62(18):4868–4881, Sept 2014.

[16] Robert Hesse and Russell Luke. Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *SIAM Journal on Optimization*, 23(4):2397–2419, 2013.

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

[18] Chi-Ming Ip and Jerzy Kyparisis. Local convergence of quasi-Newton methods for B-differentiable equations. *Mathematical Programming*, 56(1-3):71–89, 1992.

[19] Alexey F. Izmailov and Mikhail V. Solodov. *Newton-type methods for optimization and variational problems*. Springer, 2014.

[20] Guoyin Li, Tianxiang Liu, and Ting Kei Pong. Peaceman-Rachford splitting for a class of nonconvex optimization problems. *Computational Optimization and Applications*, May 2017.

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

[22] Guoyin Li and Ting Kei Pong. Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical Programming*, 159(1):371–401, Sep 2016.

[23] Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 379–387. Curran Associates, Inc., 2015.

[24] Nicholas Maratos. *Exact penalty function algorithms for finite dimensional and control optimization problems*. PhD thesis, Imperial College London (University of London), 1978.

[25] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27:372–376, 1983.

[26] Jorge Nocedal and Stephen Wright. *Numerical Optimization*. Springer, New York, 2nd edition edition, August 2006.

[27] Panagiotis Patrinos and Alberto Bemporad. Proximal Newton methods for convex composite optimization. In *52nd IEEE Conference on Decision and Control*, pages 2358–2363, Dec 2013.

[28] Panagiotis Patrinos, Lorenzo Stella, and Alberto Bemporad. Douglas-Rachford splitting: Complexity estimates and accelerated variants. In *53rd IEEE Conference on Decision and Control*, pages 4234–4239, Dec 2014.

[29] Ivan Pejcic and Colin Jones. Accelerated ADMM based on accelerated Douglas-Rachford splitting. In *2016 European Control Conference (ECC)*, pages 1952–1957, June 2016.

[30] René A. Poliquin and R. Tyrrell Rockafellar. Generalized Hessian properties of regularized nonsmooth functions. *SIAM Journal on Optimization*, 6(4):1121–1137, 1996.

[31] Michael J.D. Powell. A hybrid method for nonlinear equations. *Numerical Methods for Nonlinear Algebraic Equations*, pages 87–144, 1970.

[32] Michael J.D. Powell. A fast algorithm for nonlinearly constrained optimization calculations. In G. A. Watson, editor, *Numerical Analysis*, pages 144–157, Berlin, Heidelberg, 1978. Springer Berlin Heidelberg.

[33] R. Tyrrell Rockafellar and Roger J.B. Wets. *Variational analysis*, volume 317. Springer, 2011.

[34] Lorenzo Stella, Andreas Themelis, and Panagiotis Patrinos. Forward-backward quasi-Newton methods for nonsmooth optimization problems. *Computational Optimization and Applications*, 67(3):443–487, Jul 2017.

[35] Lorenzo Stella, Andreas Themelis, and Panagiotis Patrinos. Newton-type alternating minimization algorithm for convex optimization. *IEEE Transactions on Automatic Control*, 2018.

[36] Lorenzo Stella, Andreas Themelis, Pantelis Sopasakis, and Panagiotis Patrinos. A simple and efficient algorithm for nonlinear model predictive control. *ArXiv e-prints*, September 2017.

[37] Andreas Themelis. *Proximal Algorithms for Structured Nonconvex Optimization*. PhD thesis, KU Leuven, 12 2018.

[38] Andreas Themelis, Masoud Ahookhosh, and Panagiotis Patrinos. On the acceleration of forward-backward splitting via an inexact Newton method. In Heinz H. Bauschke, Regina S. Burachik, and D. Russell Luke, editors, *Splitting Algorithms, Modern Operator Theory, and Applications*, pages 363–412. Springer International Publishing, Cham, 2019.

[39] Andreas Themelis, Ben Hermans, and Panagiotis Patrinos. A new envelope function for nonsmooth DC optimization. *arXiv e-prints*, 3 2020.

[40] Andreas Themelis and Panagiotis Patrinos. SuperMann: a superlinearly convergent algorithm for finding fixed points of nonexpansive operators. *IEEE Transactions on Automatic Control*, 64(12):4875–4890, 12 2019.

[41] Andreas Themelis and Panagiotis Patrinos. Douglas–Rachford splitting and ADMM for nonconvex optimization: Tight convergence results. *SIAM Journal on Optimization*, 30(1):149–181, 2020.

[42] Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. 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.

[43] Yu Wang, Wotao Yin, and Jinshan Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *Journal of Scientific Computing*, 78(1):29–63, 2019.

[44] Ming Yan and Wotao Yin. *Self Equivalence of the Alternating Direction Method of Multipliers*, pages 165–194. Springer International Publishing, Cham, 2016.