九州大学学術情報リポジトリ
Kyushu University Institutional Repository

Douglas-Rachford Splitting and ADMM for Nonconvex Optimization: Tight Convergence Results

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

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

https://hdl.handle.net/2324/4377929

出版情報：SIAM Journal on Optimization. 30 (1), pp.149-181, 2020-01-09. Society for Industrial and Applied Mathematics
バージョン：
権利関係：

KYUSHU UNIVERSITY

DOUGLAS-RACHFORD SPLITTING AND ADMM FOR NONCONVEX OPTIMIZATION: TIGHT CONVERGENCE RESULTS*

ANDREAS THEMELIS† AND PANAGIOTIS PATRINOS†

Abstract. Although originally designed and analyzed for convex problems, the alternating direction method of multipliers (ADMM) and its close relatives, Douglas-Rachford splitting (DRS) and Peaceman-Rachford splitting (PRS), have been observed to perform remarkably well when applied to certain classes of structured nonconvex optimization problems. However, partial global convergence results in the nonconvex setting have only recently emerged. In this paper we show how the Douglas-Rachford envelope (DRE), introduced in 2014, can be employed to unify and considerably simplify the theory for devising global convergence guarantees for ADMM, DRS and PRS applied to nonconvex problems under less restrictive conditions, larger prox-stepsizes and over-relaxation parameters than previously known. In fact, our bounds are tight whenever the over-relaxation parameter ranges in $(0, 2]$. The analysis of ADMM uses a universal primal equivalence with DRS that generalizes the known duality of the algorithms.

Key words. Nonsmooth nonconvex optimization, Douglas-Rachford and Peaceman-Rachford splitting, ADMM.

AMS subject classifications. 90C06, 90C25, 90C26, 49J52, 49J53.

1. Introduction.

First introduced in [11] for finding numerical solutions of heat differential equations, the *Douglas-Rachford splitting* (DRS) is now a textbook algorithm in convex optimization or, more generally, in monotone inclusion problems. As the name suggests, DRS is a *splitting scheme*, meaning that it works on a problem decomposition by addressing each component separately, rather than operating on the whole problem which is typically too hard to be tackled directly. In optimization, the objective to be minimized is *split* as the sum of two functions, resulting in the following canonical framework addressed by DRS:

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

Here, $\varphi_1, \varphi_2 : \mathbb{R}^p \to \overline{\mathbb{R}}$ are proper, lower semicontinuous (lsc), extended-real-valued functions ($\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$ denotes the extended-real line). Starting from some $s \in \mathbb{R}^p$, one DR-iteration applied to (1.1) with *stepsize* $\gamma > 0$ and *relaxation* parameter $\lambda > 0$ amounts to

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

The case $\lambda = 1$ corresponds to the classical DRS, whereas for $\lambda = 2$ the scheme is also known as Peaceman-Rachford splitting (PRS). If s is a *fixed point* for the DR-iteration — that is, such that $s^+ = s$ — then it can be easily seen that u satisfies the first-order necessary condition for optimality in problem (1.1). When both φ_1 and φ_2 are convex functions, the condition is also sufficient and DRS iterations are known to converge for any $\gamma > 0$ and $\lambda \in (0, 2)$.

Closely related to DRS and possibly even more popular is the *alternating direction method of multipliers* (ADMM), first appeared in [17, 14], see also [16] for a recent historical overview. ADMM addresses linearly constrained optimization problems

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

*Submitted to the editors on January 5, 2018.

Funding: This work was 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).

†Department of Electrical Engineering (ESAT-STADIUS) – KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium (andreas.themelis@kuleuven.be, panos.patrinos@esat.kuleuven.be)

where $f : \mathbb{R}^m \to \overline{\mathbb{R}}$, $g : \mathbb{R}^n \to \overline{\mathbb{R}}$, $A \in \mathbb{R}^{p\times m}$, $B \in \mathbb{R}^{p\times n}$, and $b \in \mathbb{R}^p$. ADMM is an iterative scheme based on the following recursive steps

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

Here, $\beta > 0$ is a *penalty* parameter, $\lambda > 0$ is a possible *relaxation* parameter, and

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

is the β-augmented Lagrangian of (1.2) with $y \in \mathbb{R}^p$ as Lagrange equality multiplier. It is well known that for convex problems ADMM is simply DRS applied to a dual formulation [13], and its convergence properties for $\lambda = 1$ and arbitrary penalty parameters $\beta > 0$ are well documented in the literature, see e.g., [10]. Recently, DRS and ADMM have been observed to perform remarkably well when applied to certain classes of structured nonconvex optimization problems and partial or case-specific convergence results have also emerged.

1.1. Contributions. Our contributions can be summarized as follows.

1) *New tight convergence results for nonconvex DRS*. We provide novel convergence results for DRS applied to nonconvex problems with one function being Lipschitz differentiable (Theorem 4.3). Differently from the results in the literature, we make no a priori assumption on the existence of accumulation points and we consider all relaxation parameters $\lambda \in (0, 4)$, as opposed to $\lambda \in \{1, 2\}$. Moreover, our results are tight for all $\lambda \in (0, 2]$ (Theorem 4.9). Figures 1.1a and 1.1b highlight the extent of the improvement with respect to the state of the art.
2) *Primal equivalence of DRS and ADMM*. We prove the equivalence of DRS and ADMM for arbitrary problems and relaxation parameters, so extending their well-known duality holding in the convex case and the recently observed primal equivalence when $\lambda = 1$.
3) *New convergence results for ADMM*. Thanks to the equivalence with DRS, not only do we provide new convergence results for the ADMM scheme, but we also offer an elegant unifying framework that greatly simplifies and generalizes the theory in the literature, is based on less restrictive assumptions, and provides explicit bounds for stepsizes and possible other coefficients. A comparison with the state of the art is shown in Figure 1.1c.
4) *A continuous and exact merit function for DRS and ADMM*. Our results are based on the *Douglas-Rachford Envelope* (DRE), first introduced in [31] for convex problems and here generalized. The DRE extends the known properties of the Moreau envelope and its connections to the proximal point algorithm to composite functions as in (1.1) and (1.2). In particular, we show that the DRE serves as an exact, continuous and real-valued (as opposed to extended-real-valued) merit function for the original problem, computable with quantities obtained in the iterations of DRS (or ADMM).

Finally, we propose out-of-the-box implementations of DRS and ADMM where no prior knowledge of quantities such as Lipschitz moduli is needed, as the stepsize γ and the penalty parameter β are adaptively tuned, and which preserve convergence guarantees of the original nonadpative algorithms.

1.2. Comparisons & related work. We now compare our results with a selection of recent related works which, to the best of our knowledge, represent the state of the art for generality and contributions.

1.2.1. ADMM. A primal equivalence of DRS and ADMM has been observed in [5, Rem. 3.14] when $A = -B = \mathrm{I}$ and $\lambda = 1$. In [36, Thm. 1] the equivalence is extended to arbitrary matrices; although limited to convex problems, the result is easily extendable. Our generalization to any relaxation parameter (and nonconvex problems) is largely based on this result and uses the same problem reformulation proposed therein. The relaxation considered in this paper corresponds to that introduced in [12]; it is worth mentioning that another type of relaxation has been proposed, corresponding to $\lambda = 1$ in (ADMM) but with a different steplength for the y-update: that is, with β replaced by $\theta\beta$ for some $\theta > 0$. The known convergence results for $\theta \in (0, \frac{1+\sqrt{5}}{2})$ in the convex case, see [15, §5], were recently extended to nonconvex problems and for $\theta \in (0, 2)$ in [18].

In [35] convergence of ADMM is studied for problems of the form

$$\underset{x=(x_0\dots x_t),z}{\textbf{minimize}}\; g(\boldsymbol{x}) + \textstyle\sum_{i=0}^{t} f_i(x_i) + h(z) \quad \textbf{subject to} \quad \boldsymbol{A}\boldsymbol{x} + Bz = 0.$$

Although addressing a more general class of problem than (1.2), when specialized to the standard two-function formulation analyzed in this paper it relies on numerous assumptions. These include Lipschitz continuous minimizers of all ADMM subproblems (in particular, uniqueness of their solution). For instance, the requirements rule out interesting cases involving discrete variables or rank constraints.

In [23] a class of nonconvex problems with more than two functions is presented and variants of ADMM with deterministic and random updates are discussed. The paper provides a nice theory and explicit bounds for the penalty parameter, which agree with ours in best- and worst-case scenerarios, but are more restrictive otherwise (cf. Figure 1.1c for a more detailed comparison). The main limitation of the proposed approach is that the theory only allows for functions either convex or smooth, differently from ours where the nonsmooth term can virtually be anything. Once again, many interesting applications are not covered.

The work [25] studies a proximal ADMM where a possible Bregman divergence term in the second block update is considered. By discarding the Bregman term so as to recover the original ADMM scheme, the same bound on the stepsize as in [23] is found. Another proximal variant is proposed in [18], under less restrictive assumptions related to the concept of smoothness relative to a matrix that we will introduce in Definition 5.12. When B is injective, the proximal term can be discarded and their method reduces to the classical ADMM.

The problem addressed in [19] is fully covered by our analysis, as they consider ADMM for (1.2) where f is L-Lipschitz continuously differentiable and A is the identity matrix. Their bound $\beta > 2L$ for the penalty parameter is more conservative than ours; in fact, the two coincide only in a worst-case scenario.

1.2.2. Douglas-Rachford splitting. Few exceptions apart [26, 24], advances in nonconvex DRS theory are problem specific and only provide local convergence results, at best. These mainly focus on feasibility problems, where the goal is to find points in the intersection of nonempty closed sets A and B subjected to some regularity conditions. This is done by applying DRS to the minimization of the sum of $\varphi_1 = \delta_A$ and $\varphi_2 = \delta_B$, where δ_C is the *indicator function* of a set C (see Subsection 2.1). The minimization subproblems in DRS then reduce to (set-valued) projections onto either sets, regardless of the stepsize parameter $\gamma > 0$. This is the case of [3], where A and B are finite unions of convex sets. Local linear convergence when A is affine, under some conditions on the (nonconvex) set B, are shown in [20, 21].

Although this particular application of DRS does not comply with our requirements, as φ_1 fails to be Lipschitz differentiable, however replacing δ_A with $\varphi_1 = \frac{1}{2}\mathbf{dist}_A^2$ yields an equivalent problem which fits into our framework when A is a convex set. In terms of DRS iterations, this simply amounts to replacing $\mathbf{\Pi}_A$, the projection onto set A, with a "relaxed"

version $\mathbf{\Pi}_{A,t} := (1-t)\mathrm{id} + t\,\mathbf{\Pi}_A$ for some $t \in (0,1)$. Then, it can be easily verified that for any $\alpha, \beta \in (0, +\infty]$ one DRS-step applied to

$$\underset{s\in\mathbb{R}^n}{\textbf{minimize}}\ \tfrac{\alpha}{2}\,\mathbf{dist}_A^2(s) + \tfrac{\beta}{2}\,\mathbf{dist}_B^2(s) \tag{1.4}$$

results in

$$s^+ \in (1 - \lambda/2)s + \lambda/2\,\mathbf{\Pi}_{B,q}\,\mathbf{\Pi}_{A,p}\,s \tag{1.5}$$

for $p = \frac{2\alpha\gamma}{1+\alpha\gamma}$ and $q = \frac{2\beta\gamma}{1+\beta\gamma}$. Notice that (1.5) is the $\lambda/2$-relaxation of the "method of alternating (p,q)-relaxed projections" ((p,q)-MARP) [6]. The (non-relaxed) (p,q)-MARP is recovered by setting $\lambda = 2$, that is, by applying PRS to (1.4). Local linear convergence of MARP was shown when A and B, both possibly nonconvex, satisfy some constraint qualifications, and also global convergence when some other requirements are met. When set A is convex, then $\frac{\alpha}{2}\,\mathbf{dist}_A^2$ is convex and α-Lipschitz differentiable; our theory then ensures convergence of the *fixed-point residual* and subsequential convergence of the iterations (1.5) for any $\lambda \in (0,2)$, $p \in (0,1)$ and $q \in (0,1]$, without any requirements on the (nonempty closed) set B. Here, $q = 1$ is obtained by replacing $\frac{\beta}{2}\,\mathbf{dist}_B^2$ with δ_B, which can be interpreted as the hard penalization obtained by letting $\beta = \infty$. Although the non-relaxed MARP is not covered due to the non-strong convexity of $\mathbf{dist}_A^2$, however λ can be set arbitrarily close to 2.

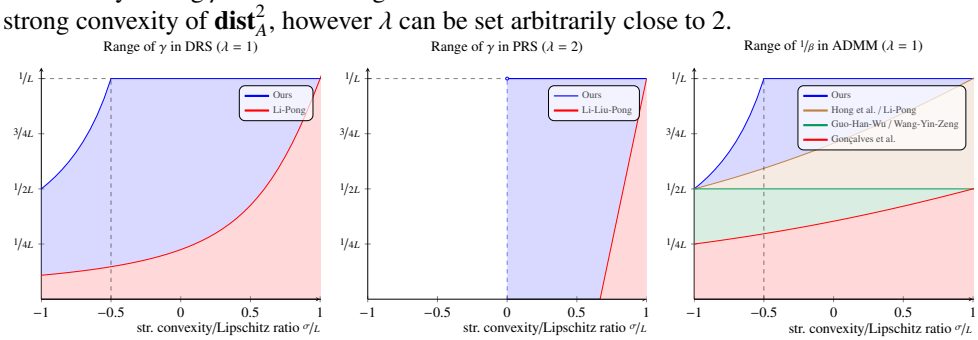


Figure 1.1: *Maximum stepsize γ ensuring convergence of DRS (Figure 1.1a) and PRS (Figure 1.1b), and maximum inverse of the penalty parameter $1/\beta$ in ADMM (Figure 1.1c); comparison between our bounds (blue plot) and [26] for DRS, [24] for PRS and [18, 19, 23, 25, 35] for ADMM. On the x-axis the ratio between hypoconvexity parameter σ and Lipschitz modulus L of the gradient of the smooth function. On the y-axis, the supremum of stepsize γ such that the algorithms converge. For ADMM the analysis is made for a common framework: 2-block ADMM with no Bregman or proximal terms, Lipschitz-differentiable f, A invertible and B identity; L and σ are relative to the transformed problem. Notice that, due to the proved analogy of DRS and ADMM, our bounds coincide in Figures 1.1a and 1.1c.*

The work [26] presents the first general analysis of global convergence of (non-relaxed) DRS for fully nonconvex problems where one function is Lipschitz differentiable. In [24] PRS is also considered under the additional requirement that the smooth function is strongly convex with strong-convexity/Lipschitz moduli ratio of at least $2/3$. For sufficiently small (explicitly computable) stepsizes one iteration of DRS or PRS yields a sufficient decrease on an augmented Lagrangian, and the generated sequences remain bounded when the cost function has bounded level sets.

Other than completing the analysis to all relaxation parameters $\lambda \in (0,4)$, as opposed to $\lambda \in \{1,2\}$, we improve their results by showing convergence for a considerably larger range of stepsizes and, in the case of PRS, with no restriction on the strong convexity modulus of the smooth function. We also show that our bounds are optimal whenever $\lambda \in (0,2]$. The

extent of the improvement is evident in the comparisons outlined in Figure 1.1. Thanks to the lower boundedness of the DRE, as opposed to the lower unbounded augmented Lagrangian, we show that the vanishing of the fixed-point residual occurs without coercivity assumptions.

1.3. Organization of the paper. The paper is organized as follows. Section 2 introduces some notation and offers a brief recap of the needed theory; the proof of the results therein is deferred to the dedicated Appendix A. In Section 3, after formally stating the needed assumptions for the DRS problem formulation (1.1) we introduce the DRE and analyze in detail its key properties. Based on these properties, in Section 4 we prove convergence results of DRS and show the tightness of our findings by means of suitable counterexamples. In Section 5 we deal with ADMM and show its equivalence with DRS; based on this, convergence results for ADMM are derived from the ones already proven for DRS. Section 6 concludes the paper.

2. Background.

2.1. Notation. The extended-real line is $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty\}$. The positive and negative parts of $r \in \mathbb{R}$ are defined respectively as $[r]_+ := \mathbf{max}\{0, r\}$ and $[r]_- := \mathbf{max}\{0, -r\}$, so that $r = [r]_+ - [r]_-$. We adopt the convention that $1/0 = \infty$.

The open and closed balls centered in x and with radius r are denoted by $\mathbf{B}(x; r)$ and $\overline{\mathbf{B}}(x; r)$, respectively. With id we indicate the identity function $x \mapsto x$ defined on a suitable space, and with I the identity matrix of suitable size. For a nonzero matrix $M \in \mathbb{R}^{p\times n}$ we let $\sigma_+(M)$ denote its smallest nonzero singular value.

For a set E and 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 $(x^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ is *summable* if $\sum_{k\in\mathbb{N}} \|x^k\|$ is finite, and *square summable* if $(\|x^k\|^2)_{k\in\mathbb{N}}$ is summable.

We use the notation $H : \mathbb{R}^n \rightrightarrows \mathbb{R}^m$ to indicate a point-to-set mapping $H : \mathbb{R}^n \to \mathcal{P}(\mathbb{R}^m)$, where $\mathcal{P}(\mathbb{R}^m)$ is the power set of $\mathbb{R}^m$ (the set of all subsets of $\mathbb{R}^m$). The *graph* of H is the set $\mathbf{gph}\, H := \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mid y \in H(x)\}$.

The *domain* of an extended-real-valued function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is the set $\mathbf{dom}\, h := \{x \in \mathbb{R}^n \mid h(x) < \infty\}$, while its *epigraph* is the set $\mathbf{epi}\, h := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} \mid h(x) \le \alpha\}$. h is said to be *proper* if $\mathbf{dom}\, h \neq \emptyset$, and *lower semicontinuous (lsc)* if $\mathbf{epi}\, h$ is a closed subset of $\mathbb{R}^{n+1}$. For $\alpha \in \mathbb{R}$, $\mathbf{lev}_{\le\alpha}\, h$ is the *α-level set* of h, i.e., $\mathbf{lev}_{\le\alpha}\, h := \{x \in \mathbb{R}^n \mid h(x) \le \alpha\}$. We say that h is *level bounded* if $\mathbf{lev}_{\le\alpha}\, h$ is bounded for all $\alpha \in \mathbb{R}$. We denote by $\hat{\partial} h : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ the *regular subdifferential* of h, where

$$v \in \hat{\partial} h(\bar{x}) \quad \Leftrightarrow \quad \liminf_{\substack{x\to\bar{x}\\ x\neq\bar{x}}} \frac{h(x) - h(\bar{x}) - \langle v, x - \bar{x}\rangle}{\|x - \bar{x}\|} \ge 0. \tag{2.1}$$

A necessary condition for local minimality of x for h is $0 \in \hat{\partial} h(x)$, see [32, Thm. 10.1]. 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 $(x^k, v^k)_{k\in\mathbb{N}} \subseteq \mathbf{gph}\, \hat{\partial} h$ such that $(x^k, h(x^k), v^k) \to (x, h(x), v)$ as $k \to \infty$. The set of *horizon subgradients* of h at x is $\partial^\infty h(x)$, defined as $\partial h(x)$ except that $v^k \to v$ is meant in the "cosmic" sense, namely $\lambda_k v^k \to v$ for some $\lambda_k \searrow 0$.

2.2. Smoothness and hypoconvexity. The class of functions $h : \mathbb{R}^n \to \mathbb{R}$ that are k times continuously differentiable is denoted as $\mathrm{C}^k(\mathbb{R}^n)$. We write $h \in \mathrm{C}^{1,1}(\mathbb{R}^n)$ to indicate that $h \in \mathrm{C}^1(\mathbb{R}^n)$ and that ∇h is Lipschitz continuous with modulus L_h. To simplify the terminology, we will say that such an h is *L_h-smooth*. It follows from [7, Prop. A.24] that if h is L_h-smooth, then $|h(y) - h(x) - \langle \nabla h(x), y - x\rangle| \le \frac{L_h}{2}\|y - x\|^2$ for all $x, y \in \mathbb{R}^n$. In particular, there exists $\sigma_h \in [-L_h, L_h]$ such that h is *σ_h-hypoconvex*, in the sense that $h - \frac{\sigma_h}{2}\|\cdot\|^2$ is a convex function. Thus, every L_h-smooth and σ_h-hypoconvex function h satisfies

$$\tfrac{\sigma_h}{2}\|y - x\|^2 \le h(y) - h(x) - \langle \nabla h(x), y - x\rangle \le \tfrac{L_h}{2}\|y - x\|^2 \quad \forall x, y \in \mathbb{R}^n. \tag{2.2}$$

By applying [29, Thm. 2.1.5] to the (convex) function $\psi = h - \frac{\sigma_h}{2}\|\cdot\|^2$ we obtain that this is equivalent to having

$$\sigma_h\|y-x\|^2 \le \langle \nabla h(y) - \nabla h(x), y-x\rangle \le L_h\|y-x\|^2 \quad \forall x, y \in \mathbb{R}^n. \tag{2.3}$$

Note that σ_h-hypoconvexity generalizes the notion of (strong) convexity by allowing negative strong convexity moduli. In fact, if $\sigma_h = 0$ then σ_h-hypoconvexity reduces to convexity, while for $\sigma_h > 0$ it denotes σ_h-strong convexity.

Lemma 2.1 (Subdifferential characterization of smoothness). *Let $h : \mathbb{R}^n \to \mathbb{R}$ be such that $\partial h(x) \neq \emptyset$ for all $x \in \mathbb{R}^n$, and suppose that there exist $L \ge 0$ and $\sigma \in [-L, L]$ such that*

$$\sigma\|x_1 - x_2\|^2 \le \langle v_1 - v_2, x_1 - x_2\rangle \le L\|x_1 - x_2\|^2 \quad \forall x_i \in \mathbb{R}^n,\ v_i \in \partial h(x_i),\ i = 1, 2. \tag{2.4}$$

Then, $h \in \mathrm{C}^{1,1}(\mathbb{R}^n)$ is L-smooth and σ-hypoconvex.

Proof. See Appendix A. □

Theorem 2.2 (Lower bounds for smooth functions). *Let $h \in C^{1,1}(\mathbb{R}^n)$ be L_h-smooth and σ_h-hypoconvex. Then, for all $x, y \in \mathbb{R}^n$ it holds that*

$$h(y) \ge h(x) + \langle \nabla h(x), y - x\rangle + \rho(y, x),$$

where

(i) either $\rho(y, x) = \frac{\sigma_h}{2}\|y - x\|^2$,

(ii) or $\rho(y, x) = \frac{\sigma_h L_h}{2(L_h+\sigma_h)}\|y - x\|^2 + \frac{1}{2(L_h+\sigma_h)}\|\nabla h(y) - \nabla h(x)\|^2$, *provided that* $-L_h < \sigma_h \le 0$.

Clearly, all inequalities remain valid if L_h is replaced with any $L \ge L_h$ and σ_h with any $\sigma \in [-L, \sigma_h]$.

Proof. See Appendix A. □

2.3. Proximal mapping. The *proximal mapping* of $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ with parameter $\gamma > 0$ is $\mathbf{prox}_{\gamma h} : \mathbb{R}^n \rightrightarrows \mathbf{dom}\, h$ defined as

$$\mathbf{prox}_{\gamma h}(x) := \operatorname*{\mathbf{arg\,min}}_{w\in\mathbb{R}^n} \left\{h(w) + \tfrac{1}{2\gamma}\|w - x\|^2\right\}. \tag{2.5}$$

We say that a function h is *prox-bounded* if $h + \frac{1}{2\gamma}\|\cdot\|^2$ is lower bounded for some $\gamma > 0$. The supremum of all such γ is the *threshold of prox-boundedness of h*, denoted as γ_h. If h is proper and lsc, $\mathbf{prox}_{\gamma h}$ is nonempty- and compact-valued over $\mathbb{R}^n$ for $\gamma \in (0, \gamma_h)$ [32, Thm. 1.25]. The value function of the minimization problem defining the proximal mapping, namely the *Moreau envelope* with stepsize $\gamma \in (0, \gamma_h)$, denoted by $h^\gamma : \mathbb{R}^n \to \mathbb{R}$ and defined as

$$h^\gamma(x) := \inf_{w\in\mathbb{R}^n} \left\{h(w) + \tfrac{1}{2\gamma}\|w - x\|^2\right\}, \tag{2.6}$$

is finite and strictly continuous [32, Thm. 1.25 and Ex. 10.32]. The necessary optimality conditions of the problem defining $\mathbf{prox}_{\gamma g}$ together with [32, Thm. 10.1 and Ex. 8.8] imply

$$\tfrac{1}{\gamma}(x - \bar{x}) \in \hat{\partial} h(\bar{x}) \quad \forall \bar{x} \in \mathbf{prox}_{\gamma h}(x). \tag{2.7}$$

When $h \in \mathrm{C}^{1,1}(\mathbb{R}^n)$, its proximal mapping and Moreau envelope enjoy many favorable properties which we summarize next.

Proposition 2.3 (Proximal properties of smooth functions). *Let $h \in \mathrm{C}^{1,1}(\mathbf{dom}\, h)$ be L_h-smooth, hence σ_h-hypoconvex for some $\sigma_h \in [-L_h, L_h]$. Then, h is prox-bounded with $\gamma_h \ge 1/[\sigma_h]_-$ and for all $\gamma < 1/[\sigma_h]_-$ the following hold:*

(i) $\mathbf{prox}_{\gamma h}$ *is single valued, and for all* $s \in \mathbb{R}^n$ *it holds that* $u = \mathbf{prox}_{\gamma h}(s)$ *iff* $s = u + \gamma\nabla h(u)$.

(ii) $\mathbf{prox}_{\gamma h}$ *is* $(\frac{1}{1+\gamma L_h})$*-strongly monotone and* $(1+\gamma\sigma_h)$*-cocoercive, in the sense that*

$$\langle u-u', s-s'\rangle \geq \tfrac{1}{1+\gamma L_h}\|s-s'\|^2 \quad \textit{and} \quad \langle u-u', s-s'\rangle \geq (1+\gamma\sigma_h)\|u-u'\|^2$$

for all $s, s' \in \mathbb{R}^n$*, where* $u = \mathbf{prox}_{\gamma h}(s)$ *and* $u' = \mathbf{prox}_{\gamma h}(s')$*. In particular,*

$$\tfrac{1}{1+\gamma L_h}\|s-s'\| \leq \|u-u'\| \leq \tfrac{1}{1+\gamma\sigma_h}\|s-s'\|. \tag{2.8}$$

Thus, $\mathbf{prox}_{\gamma h}$ *is a* $\frac{1}{1+\gamma\sigma_h}$*-Lipschitz and invertible mapping, and its inverse* $\mathrm{id} + \gamma\nabla h$ *is* $(1+\gamma L_h)$*-Lipschitz continuous.*

(iii) $h^\gamma \in \mathrm{C}^{1,1}(\mathbb{R}^n)$ *is* L_{h^γ}*-smooth and* σ_{h^γ}*-hypoconvex, with* $L_{h^\gamma} = \mathbf{max}\left\{\frac{L_h}{1+\gamma L_h}, \frac{[\sigma_h]_-}{1+\gamma\sigma_h}\right\}$ *and* $\sigma_{h^\gamma} = \frac{\sigma_h}{1+\gamma\sigma_h}$*. Moreover,* $\nabla h^\gamma(s) = \frac{1}{\gamma}(s - \mathbf{prox}_{\gamma h}(s))$ *and* $\nabla h(\mathbf{prox}_{\gamma h}(s)) = \frac{1}{\gamma}(s - \mathbf{prox}_{\gamma h}(s))$.

Proof. See Appendix A. □

3. Douglas-Rachford envelope. We now list the blanket assumptions for the functions in problem (1.1).

Assumption I (Requirements for the DRS formulation (1.1)). *The following hold*

(i) $\varphi_1 \in \mathrm{C}^{1,1}(\mathbb{R}^n)$ *is* L_{φ_1}*-smooth, hence* σ_{φ_1}*-hypoconvex for some* $\sigma_{\varphi_1} \in [-L_{\varphi_1}, L_{\varphi_1}]$.

(ii) φ_2 *is proper and lsc.*

(iii) Problem (1.1) *has a solution, that is,* $\mathbf{arg\,min}\,\varphi \neq \emptyset$.

Remark 3.1 (Feasible stepsizes for DRS). Under Assumption I, both φ_1 and φ_2 are prox-bounded with threshold at least $1/L_{\varphi_1}$, and in particular DRS iterations are well defined for all $\gamma \in (0, 1/L_{\varphi_1})$. That $\gamma_{\varphi_1} \geq 1/L_{\varphi_1}$ follows from Proposition 2.3, having $1/[\sigma_{\varphi_1}]_- \geq 1/L_{\varphi_1}$. As for φ_2, for all $s \in \mathbb{R}^p$ it holds that

$$\mathbf{inf}\,\varphi \leq \varphi_1(s) + \varphi_2(s) \overset{(2.2)}{\leq} \varphi_1(0) + \langle\nabla\varphi_1(0), s\rangle + \tfrac{L_{\varphi_1}}{2}\|s\|^2 + \varphi_2(s),$$

hence, for all $\gamma < 1/L_{\varphi_1}$ the function $s \mapsto \varphi_2(s) + \frac{1}{2\gamma}\|s\|^2$ is lower bounded. □

Starting from $s \in \mathbb{R}^p$, let (u, v) be generated by a DRS step under Assumption I. As first noted in [31], from the relation $s = u + \gamma\nabla\varphi_1(u)$ (see Proposition 2.3*(i)*) it follows that

$$v \in \mathbf{prox}_{\gamma\varphi_2}(u - \gamma\nabla\varphi_1(u)) \tag{3.1}$$

is the result of a *forward-backward step* at u, amounting to

$$v \in \underset{w\in\mathbb{R}^p}{\mathbf{arg\,min}}\Big\{\varphi_2(w) + \underbrace{\varphi_1(u) + \langle\nabla\varphi_1(u), w-u\rangle + \tfrac{1}{2\gamma}\|w-u\|^2}\Big\}, \tag{3.2}$$

see e.g., [9, 34] for an extensive discussion on nonconvex forward-backward splitting (FBS). This shows that v is the result of the minimization of a *majorization model* for the original function $\varphi = \varphi_1 + \varphi_2$, where the smooth function φ_1 is replaced by the quadratic upper bound emphasized by the under-bracket in (3.2). First introduced in [31] for convex problems, the *Douglas-Rachford envelope* (DRE) is the function $\varphi_\gamma^{\mathrm{DR}} : \mathbb{R}^p \to \mathbb{R}$ defined as

$$\varphi_\gamma^{\mathrm{DR}}(s) := \min_{w\in\mathbb{R}^p}\Big\{\varphi_2(w) + \varphi_1(u) + \langle\nabla\varphi_1(u), w-u\rangle + \tfrac{1}{2\gamma}\|w-u\|^2\Big\}, \tag{3.3}$$

where $u := \mathbf{prox}_{\gamma\varphi_1}(s)$. Namely, rather than the *minimizer* v, $\varphi_\gamma^{\mathrm{DR}}(s)$ is the *value* of the minimization problem (3.2) defining the v-update in (DRS). The expression (3.3) emphasizes the close connection that the DRE has with the *forward-backward envelope* (FBE) as in [34], here denoted $\varphi_\gamma^{\mathrm{FB}}$, namely

$$\varphi_\gamma^{\mathrm{DR}}(s) = \varphi_\gamma^{\mathrm{FB}}(u), \quad \text{where } u = \mathbf{prox}_{\gamma\varphi_1}(s). \tag{3.4}$$

The FBE is an exact penalty function for FBS, which was initially proposed for convex problems in [30] and later extended and further analyzed in [33, 34, 27]. In this section we will see that, under Assumption I, the DRE serves a similar role with respect to DRS which will be key for establishing (tight) convergence results in the nonconvex setting. Another useful intepretation of the DRE is obtained by plugging the minimizer $w = v$ in (3.3). This leads to

$$\varphi_\gamma^{\mathrm{DR}}(s) = \mathscr{L}_{1/\gamma}(u, v, \gamma^{-1}(u - s)), \tag{3.5}$$

where u and v come from the DRS iteration and

$$\mathscr{L}_\beta(x, z, y) := \varphi_1(x) + \varphi_2(z) + \langle y, x - z\rangle + \tfrac{\beta}{2}\|x - z\|^2 \tag{3.6}$$

is the β-augmented Lagrangian relative to the equivalent problem formulation

$$\underset{x,z\in\mathbb{R}^p}{\textbf{minimize}}\ \varphi_1(x) + \varphi_2(z) \quad \textbf{subject to}\ \ x - z = 0. \tag{3.7}$$

This expression also emphasizes that evaluating $\varphi_\gamma^{\mathrm{DR}}(s)$ requires the same operations as performing one DRS update $s \mapsto (u, v)$.

3.1. Properties. Building upon the connection with the FBE emphasized in (3.4), in this section we highlight some important properties enjoyed by the DRE. We start by observing that $\varphi_\gamma^{\mathrm{DR}}$ is a strictly continuous function for $\gamma < 1/L_{\varphi_1}$, owing to the fact that so is the FBE [34, Prop. 4.2], and that $\mathbf{prox}_{\gamma\varphi_1}$ is Lipschitz continuous as shown in Proposition 2.3*(ii)*.

Proposition 3.2 (Strict continuity). *Suppose that Assumption I is satisfied. For all $\gamma < 1/L_{\varphi_1}$ the DRE $\varphi_\gamma^{\mathrm{DR}}$ is a real-valued and strictly continuous function.*

Next, we investigate the fundamental connections relating the DRE $\varphi_\gamma^{\mathrm{DR}}$ and the cost function φ. We show, for γ small enough and up to an (invertible) change of variable, that infima and minimizers of the two functions coincide, as well as equivalence of level boundedness.

Proposition 3.3 (Sandwiching property). *Suppose that Assumption I is satisfied. Let $\gamma < 1/L_{\varphi_1}$ be fixed, and consider u, v generated by one DRS iteration starting from $s \in \mathbb{R}^p$. Then,*

(i) $\varphi_\gamma^{\mathrm{DR}}(s) \le \varphi(u)$.

(ii) $\varphi(v) \le \varphi_\gamma^{\mathrm{DR}}(s) - \frac{1-\gamma L_{\varphi_1}}{2\gamma}\|u - v\|^2$.

Proof. 3.3*(i)* is easily inferred from definition (3.3) by considering $w = u$. Moreover, it follows from [34, Prop. 4.3] and the fact that $v \in \mathbf{prox}_{\gamma\varphi_2}(u - \gamma\nabla\varphi_1(u))$, cf. (3.1), that $\varphi(v) \le \varphi_\gamma^{\mathrm{FB}}(u) - \frac{1-\gamma L_{\varphi_1}}{2\gamma}\|u - v\|^2$. 3.3*(ii)* then follows from (3.4). ∎

Theorem 3.4 (Minimization and level-boundedness equivalence). *Suppose that Assumption I is satisfied. For any $\gamma < 1/L_{\varphi_1}$ the following hold:*

(i) $\mathbf{inf}\,\varphi = \mathbf{inf}\,\varphi_\gamma^{\mathrm{DR}}$.

(ii) $\mathbf{arg\,min}\,\varphi = \mathbf{prox}_{\gamma\varphi_1}(\mathbf{arg\,min}\,\varphi_\gamma^{\mathrm{DR}})$.

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

Proof. It follows from [34, Thm. 4.4] that the FBE satisfies $\mathbf{arg\,min}\,\varphi = \mathbf{arg\,min}\,\varphi_\gamma^{\mathrm{FB}}$ and $\mathbf{inf}\,\varphi = \mathbf{inf}\,\varphi_\gamma^{\mathrm{FB}}$. The similar properties 3.4*(i)* and 3.4*(ii)* of the DRE then follow from the identity $\varphi_\gamma^{\mathrm{DR}} = \varphi_\gamma^{\mathrm{FB}} \circ \mathbf{prox}_{\gamma\varphi_1}$, cf. (3.4), and the fact that $\mathbf{prox}_{\gamma\varphi_1}$ is invertible, as shown in Proposition 2.3.

We now show 3.4*(iii)*. Denote $\varphi_\star := \mathbf{inf}\,\varphi = \mathbf{inf}\,\varphi_\gamma^{\mathrm{DR}}$, which is finite by assumption.

♠ Suppose that $\varphi_\gamma^{\mathrm{DR}}$ is level bounded, and let $u \in \mathbf{lev}_{\le\alpha}\,\varphi$ for some $\alpha > \varphi_\star$. Then, $s := u + \gamma\nabla\varphi_1(u)$ is such that $\mathbf{prox}_{\gamma\varphi_1}(s) = u$, as shown in Proposition 2.3*(i)*. Thus, from Proposition 3.3 it follows that $s \in \mathbf{lev}_{\le\alpha}\,\varphi_\gamma^{\mathrm{DR}}$. In particular, $\mathbf{lev}_{\le\alpha}\,\varphi \subseteq [I + \gamma\nabla\varphi_1]^{-1}(\mathbf{lev}_{\le\alpha}\,\varphi_\gamma^{\mathrm{DR}})$, and since $\mathbf{prox}_{\gamma\varphi_1} = [\mathrm{I} + \gamma\nabla\varphi_1]^{-1}$ is Lipschitz continuous and $\mathbf{lev}_{\le\alpha}\,\varphi_\gamma^{\mathrm{DR}}$ is bounded by assumption, it follows that $\mathbf{lev}_{\le\alpha}\,\varphi$ is also bounded.

♠ Suppose now that $\varphi_\gamma^{\mathrm{DR}}$ is not level bounded. Then, there exists $\alpha > \varphi_\star$ together with a sequence $(s_k)_{k\in\mathbb{N}}$ satisfying $s_k \in \mathbf{lev}_{\leq\alpha}\,\varphi_\gamma^{\mathrm{DR}} \setminus \mathbf{B}(0;k)$ for all $k \in \mathbb{N}$. Let $u_k := \mathbf{prox}_{\gamma\varphi_1}(s_k)$, so that $s_k = u_k + \gamma\nabla\varphi_1(u_k)$ (Proposition 2.3*(i)*), and let $v_k \in \mathbf{prox}_{\gamma\varphi_2}(u_k - \gamma\nabla\varphi_1(u_k))$. From Proposition 3.3*(ii)* it then follows that $v_k \in \mathbf{lev}_{\leq\alpha}\,\varphi$, and that

$$\alpha - \varphi_\star \geq \varphi_\gamma^{\mathrm{DR}}(s_k) - \varphi_\star \geq \varphi_\gamma^{\mathrm{DR}}(s_k) - \varphi(v_k) \geq \tfrac{1-\gamma L_{\varphi_1}}{2\gamma}\|u_k - v_k\|^2.$$

Therefore, $\|u_k - v_k\|^2 \leq \frac{2\gamma(\alpha-\varphi_\star)}{1-\gamma L_{\varphi_1}}$ and

$$\begin{aligned}\|v_k\| &\geq \|u_k - u_0\| - \|u_0\| - \|u_k - v_k\| \overset{2.3(ii)}{\geq} \tfrac{1}{1+\gamma L_{\varphi_1}}\|s_k - s_0\| - \|u_0\| - \|u_k - v_k\| \\ &\geq \tfrac{k-\|s_0\|}{1+\gamma L_{\varphi_1}} - \|u_0\| - \sqrt{\tfrac{2\gamma(\alpha-\varphi_\star)}{1-\gamma L_{\varphi_1}}} \to +\infty \quad \text{as } k \to \infty.\end{aligned}$$

This shows that $\mathbf{lev}_{\leq\alpha}\,\varphi$ is also unbounded. ∎

4. Convergence of Douglas-Rachford splitting. Closely related to the DRE, the augmented Lagrangian (3.6) (in fact, rather a *"reduced"* Lagrangian with negative penalty β) was used in [26] under the name of *Douglas-Rachford merit function* to analyze DRS for the special case $\lambda = 1$. It was shown that for sufficiently small γ there exists $c > 0$ such that the iterates generated by DRS satisfy

$$\mathscr{L}_{-1/\gamma}(u^{k+1}, v^{k+1}, \eta^{k+1}) \leq \mathscr{L}_{-1/\gamma}(u^k, v^k, \eta^k) - c\|u^k - u^{k+1}\|^2 \quad \text{with } \eta^k = \gamma^{-1}(u^k - s^k), \tag{4.1}$$

to infer that $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ have same accumulation points, all of which are stationary for φ. In [24], where also the case $\lambda = 2$ is addressed with $\mathscr{L}_{-3/\gamma}$ as penalty function, it was then shown that the sequence remains bounded and thus accumulation points exist in case φ is level bounded. We now generalize the decrease property (4.1) shown in [26, 24] by considering arbitrary relaxation parameters $\lambda \in (0, 4)$ (as opposed to $\lambda \in \{1, 2\}$) and providing tight ranges for the stepsize γ whenever $\lambda \in (0, 2]$. Thanks to the lower boundedness of $\varphi_\gamma^{\mathrm{DR}}$, it will be possible to show that the DRS residual vanishes without any coercivity assumption.

Theorem 4.1 (Sufficient decrease on the DRE). *Suppose that Assumption I is satisfied, and consider one DRS update $s \mapsto (u, v, s^+)$ for some stepsize $\gamma < \mathbf{min}\left\{\frac{2-\lambda}{2[\sigma_{\varphi_1}]_-}, \frac{1}{L_{\varphi_1}}\right\}$ and relaxation $\lambda \in (0, 2)$. Then,*

$$\varphi_\gamma^{\mathrm{DR}}(s) - \varphi_\gamma^{\mathrm{DR}}(s^+) \geq \tfrac{c}{(1+\gamma L_{\varphi_1})^2}\|s - s^+\|^2, \tag{4.2}$$

where, denoting $p_{\varphi_1} := \sigma_{\varphi_1}/L_{\varphi_1} \in [-1, 1]$, c is a strictly positive constant defined as[1]

$$c = \frac{2-\lambda}{2\lambda\gamma} - \begin{cases} L_{\varphi_1}\,\mathbf{max}\left\{\frac{[p_{\varphi_1}]_-}{2(1-[p_{\varphi_1}]_-)}, \frac{\gamma L_{\varphi_1}}{\lambda} - \frac{1}{2}\right\} & \text{if } p_{\varphi_1} \geq \frac{\lambda}{2} - 1, \\ \frac{[\sigma_{\varphi_1}]_-}{\lambda} & \text{otherwise.}\end{cases} \tag{4.3}$$

If φ_1 is strongly convex, then (4.2) also holds for

$$2 \leq \lambda < \tfrac{4}{1+\sqrt{1-p_{\varphi_1}}} \quad \text{and} \quad \tfrac{p_{\varphi_1}\lambda-\delta}{4\sigma_{\varphi_1}} < \gamma < \tfrac{p_{\varphi_1}\lambda+\delta}{4\sigma_{\varphi_1}}, \tag{4.4}$$

where $\delta := \sqrt{(p_{\varphi_1}\lambda)^2 - 8p_{\varphi_1}(\lambda-2)}$, in which case

$$c = \tfrac{2-\lambda}{2\lambda\gamma} + \tfrac{\sigma_{\varphi_1}}{\lambda}\left(\tfrac{1}{2} - \tfrac{\gamma L_{\varphi_1}}{\lambda}\right). \tag{4.5}$$

[1] A one-line expression for the constant is $c = \frac{2-\lambda}{2\lambda\gamma} - \mathbf{min}\left\{\frac{[p_{\varphi_1}]_-}{\lambda}, L_{\varphi_1}\,\mathbf{max}\left\{\frac{[\sigma_{\varphi_1}]_-}{2(1-[p_{\varphi_1}]_-)}, \frac{\gamma L_{\varphi_1}}{\lambda} - \frac{1}{2}\right\}\right\}$.

Proof. Let (u^+, v^+) be generated by one DRS iteration starting at s^+. Then,

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

and the minimum is attained at $w = v^+$. Therefore, letting ρ be as in Theorem 2.2,

$$\begin{aligned}
\varphi_\gamma^{\mathrm{DR}}(s^+) &\le \varphi_1(u^+) + \langle \nabla\varphi_1(u^+), v - u^+\rangle + \varphi_2(v) + \tfrac{1}{2\gamma}\|u^+ - v\|^2 \\
&= \underbrace{\varphi_1(u^+) + \langle \nabla\varphi_1(u^+), u - u^+\rangle} + \langle \nabla\varphi_1(u^+), v - u\rangle + \varphi_2(v) + \tfrac{1}{2\gamma}\|u^+ - v\|^2 \\
&\overset{\text{Thm. 2.2}}{\le} \varphi_1(u) - \rho(u, u^+) + \langle \nabla\varphi_1(u^+), v - u\rangle + \varphi_2(v) + \tfrac{1}{2\gamma}\|u^+ - v\|^2 \\
&= \varphi_1(u) - \rho(u, u^+) + \langle \nabla\varphi_1(u), v - u\rangle + \varphi_2(v) + \tfrac{1}{2\gamma}\|u^+ - v\|^2 + \langle \nabla\varphi_1(u^+) - \nabla\varphi_1(u), v - u\rangle \\
&= \varphi_\gamma^{\mathrm{DR}}(s) - \rho(u, u^+) + \langle \nabla\varphi_1(u^+) - \nabla\varphi_1(u), v - u\rangle + \tfrac{1}{2\gamma}\|u - u^+\|^2 + \tfrac{1}{\gamma}\langle u^+ - u, u - v\rangle.
\end{aligned}$$

Since $u - v = \frac{1}{\lambda}(s - s^+) = \frac{1}{\lambda}(u - u^+) + \frac{\gamma}{\lambda}(\nabla\varphi_1(u) - \nabla\varphi_1(u^+))$, see Proposition 2.3*(i)*, it all simplifies to

$$\varphi_\gamma^{\mathrm{DR}}(s) - \varphi_\gamma^{\mathrm{DR}}(s^+) \ge \tfrac{2-\lambda}{2\gamma\lambda}\|u - u^+\|^2 - \tfrac{\gamma}{\lambda}\|\nabla\varphi_1(u^+) - \nabla\varphi_1(u)\|^2 + \rho(u, u^+). \tag{4.6}$$

It will suffice to show that

$$\varphi_\gamma^{\mathrm{DR}}(s) - \varphi_\gamma^{\mathrm{DR}}(s^+) \ge c\|u - u^+\|^2;$$

inequality (4.2) will then follow from $\frac{1}{1+\gamma L_{\varphi_1}}$-strong monotonicity of $\mathbf{prox}_{\gamma\varphi_1}$, see Proposition 2.3*(ii)*. We now proceed by cases.

♠ **Case 1:** $\lambda \in (0, 2)$**.**

Let $\sigma := -[\sigma_{\varphi_1}]_- = \min\left\{\sigma_{\varphi_1}, 0\right\}$ and $L \ge L_{\varphi_1}$ be such that $L + \sigma > 0$; the value of such an L will be fixed later. Then, $\sigma \le 0$ and φ_1 is L-smooth and σ-hypoconvex. We may thus choose $\rho(u, u^+)$ as in Theorem 2.2*(ii)* with these values of L and σ. Inequality (4.6) then becomes

$$\tfrac{\varphi_\gamma^{\mathrm{DR}}(s) - \varphi_\gamma^{\mathrm{DR}}(s^+)}{L} \ge \left(\tfrac{2-\lambda}{2\lambda\xi} + \tfrac{p}{2(1+p)}\right)\|u^+ - u\|^2 + \tfrac{1}{L^2}\left(\tfrac{1}{2(1+p)} - \tfrac{\xi}{\lambda}\right)\|\nabla\varphi_1(u^+) - \nabla\varphi_1(u)\|^2,$$

where $\xi := \gamma L$ and $p := \sigma/L \in (-1, 0]$. Since $\nabla\varphi_1$ is L_{φ_1}-Lipschitz continuous, the constant c can be taken such that

$$\frac{c}{L} = \begin{cases} \frac{2-\lambda}{2\lambda\xi} + \frac{p}{2(1+p)} & \text{if } 0 < \frac{1}{2(1+p)} - \frac{\xi}{\lambda}, \\ \frac{2-\lambda}{2\lambda\xi} + \frac{p}{2(1+p)} + \frac{L_{\varphi_1}^2}{L^2}\left(\frac{1}{2(1+p)} - \frac{\xi}{\lambda}\right) & \text{otherwise.} \end{cases} \tag{4.7}$$

We will now select a suitable L so as to ensure that c is indeed strictly positive and as given in the statement. To this end, we consider two subcases:

- **Case 1a:** $0 < \lambda \le 2(1 + \sigma/L_{\varphi_1})$.
  Then, $\sigma \ge -\frac{2-\lambda}{2}L_{\varphi_1} > -L_{\varphi_1}$ and we can take $L = L_{\varphi_1}$. Consequently, $p = \sigma/L_{\varphi_1}$, $\xi = \gamma L_{\varphi_1}$, and (4.7) becomes

$$\frac{c}{L_{\varphi_1}} = \frac{2-\lambda}{2\lambda\gamma L_{\varphi_1}} + \begin{cases} \frac{p}{2(1+p)} & \text{if } \gamma L_{\varphi_1} < \frac{\lambda}{2(1+p)}, \\ \frac{1}{2} - \frac{\gamma L_{\varphi_1}}{\lambda} & \text{otherwise.} \end{cases} \tag{4.8}$$

  Let us verify that in this case any γ such that $\gamma < 1/L_{\varphi_1}$ yields a strictly positive coefficient c. If $0 < \gamma L_{\varphi_1} < \frac{\lambda}{2(1+p)} \le 1$, then

$$\tfrac{c}{L_{\varphi_1}} = \tfrac{2-\lambda}{2\lambda\gamma L_{\varphi_1}} + \tfrac{p}{2(1+p)} > \tfrac{2-\lambda}{2\lambda} + \tfrac{p}{\lambda} = \tfrac{1+p}{\lambda} - \tfrac{1}{2} \ge 0,$$

where the first inequality uses the facts that $\lambda < 2$, $p \leq 0$, and $\gamma L_{\varphi_1} < 1$. If instead $\frac{\lambda}{2(1+p)} \leq \gamma L_{\varphi_1} < 1$, then

$$\frac{c}{L_{\varphi_1}} = \frac{2-\lambda}{2\lambda\gamma L_{\varphi_1}} + \frac{1}{2} - \frac{\gamma L_{\varphi_1}}{\lambda} > \frac{2-\lambda}{2\lambda} + \frac{1}{2} - \frac{1}{\lambda} = 0.$$

Either way, the sufficient decrease constant c is strictly positive. Since $\sigma = -[\sigma_{\varphi_1}]_-$ and

$$\frac{2-\lambda}{2\lambda\gamma} + \frac{\sigma}{2(1+p)} \leq \frac{2-\lambda}{2\lambda\gamma} + \frac{L_{\varphi_1}}{2} - \frac{\gamma L_{\varphi_1}^2}{\lambda} \quad \Leftrightarrow \quad \gamma \leq \frac{\lambda}{2(L_{\varphi_1}+\sigma)},$$

from (4.8) we conclude that c is as in (4.2).

- **Case 1b:** $2(1 + \sigma/L_{\varphi_1}) < \lambda < 2$.
  Necessarily $\sigma < 0$, for otherwise the range of λ would be empty. In particular, $\sigma = \sigma_{\varphi_1}$, and the lower bound on λ can be expressed as $\sigma_{\varphi_1} < -\frac{2-\lambda}{2} L_{\varphi_1}$. Consequently, $L := \frac{-2\sigma_{\varphi_1}}{2-\lambda}$ is strictly larger than L_{φ_1}, and in particular $\sigma + L = \sigma_{\varphi_1} + L > 0$. The ratio of σ and L is thus $p = \frac{\lambda}{2} - 1$, and (4.7) becomes

$$c = \frac{2-\lambda}{2\lambda\gamma} + \begin{cases} \frac{\sigma_{\varphi_1}}{\lambda} & \text{if } \gamma < \frac{2-\lambda}{-2\sigma_{\varphi_1}}, \\ \frac{\sigma_{\varphi_1}}{\lambda} - \frac{\gamma L_{\varphi_1}^2}{\lambda} + \frac{2-\lambda}{-2\sigma_{\varphi_1}\lambda} L_{\varphi_1}^2 & \text{otherwise.} \end{cases} \tag{4.9}$$

  Let us show that, when $\gamma < \frac{2-\lambda}{-2\sigma_{\varphi_1}} = \frac{1}{L}$, also in this case the sufficient decrease constant c is strictly positive. We have

$$\frac{c}{L} = \frac{2-\lambda}{2\lambda\gamma L} + \frac{\sigma_{\varphi_1}}{\lambda}\frac{1}{L} > \frac{2-\lambda}{2\lambda} + \frac{\sigma_{\varphi_1}}{\lambda}\frac{2-\lambda}{-2\sigma_{\varphi_1}} = 0,$$

  hence the claim. This concludes the proof for the case $\lambda \in (0, 2)$.

♠ **Case 2:** $\lambda \geq 2$**.**

In this case we need to assume that φ_1 is strongly convex, that is, that $\sigma_{\varphi_1} > 0$. Instead of considering a single expression of ρ, we will rather take a convex combination of those in Theorems 2.2*(i)* and 2.2*(ii)*, namely

$$\rho(u, u^+) = (1-\alpha)\frac{\sigma_{\varphi_1}}{2}\|u - u^+\|^2 + \alpha\frac{1}{2L_{\varphi_1}}\|\nabla\varphi_1(u) - \nabla\varphi_1(u^+)\|^2$$

for some $\alpha \in [0, 1]$ to be determined. (4.6) then becomes

$$\frac{\varphi_\gamma^{\mathrm{DR}}(s) - \varphi_\gamma^{\mathrm{DR}}(s^+)}{L_{\varphi_1}} \geq \left(\frac{2-\lambda}{2\lambda\xi} + \frac{(1-\alpha)p}{2}\right)\|u - u^+\|^2 + \frac{1}{L_{\varphi_1}^2}\left(\frac{\alpha}{2} - \frac{\xi}{\lambda}\right)\|\nabla\varphi_1(u) - \nabla\varphi_1(u^+)\|^2,$$

where $\xi := \gamma L_{\varphi_1}$ and $p := \sigma_{\varphi_1}/L_{\varphi_1} \in (0, 1]$. By restricting $\xi \in (0, 1)$, since $\lambda \geq 2$ one can take $\alpha := 2\xi/\lambda \in (0, 1)$ to make the coefficient multiplying the gradient norm vanish. We then obtain

$$\frac{c}{L_{\varphi_1}} = \frac{2-\lambda}{2\lambda\xi} + \frac{(\lambda - 2\xi)p}{2\lambda}. \tag{4.10}$$

Imposing $c > 0$ results in the following second-order equation in variable ξ,

$$2p\xi^2 - p\lambda\xi + (\lambda - 2) < 0. \tag{4.11}$$

The discriminant is $\Delta := (p\lambda)^2 - 8p(\lambda - 2)$, which, for $\lambda \geq 2$, is strictly positive iff

$$2 \leq \lambda < \frac{4}{1+\sqrt{1-p}} \quad \vee \quad \lambda > \frac{4}{1-\sqrt{1-p}}.$$

Denoting $\delta := \sqrt{\Delta} = \sqrt{(p\lambda)^2 - 8p(\lambda-2)}$, the solution to (4.11) is $\frac{p\lambda-\delta}{4p} < \xi < \frac{p\lambda+\delta}{4p}$. However, the case $\lambda \geq 4$ has to be discarded, as $\frac{p\lambda-\delta}{4p} > 1$ in this case, contradicting the fact that $p \leq 1$. To see this, suppose $\lambda \geq 4$. Then,

$$\begin{aligned} \tfrac{p\lambda-\delta}{4p} < 1 \quad &\Leftrightarrow \quad p(\lambda-4) < \delta \\ &\Leftrightarrow \quad p^2(\lambda-4)^2 < \Delta = (p\lambda)^2 - 8p(\lambda-2) \\ &\Leftrightarrow \quad p(2-\lambda) < 2-\lambda, \end{aligned}$$

hence $p > 1$, which contradicts the fact that $\sigma_{\varphi_1} \leq L_{\varphi_1}$. Thus, the only feasible ranges are the ones given in (4.4), hence the claimed sufficient decrease constant c, cf. (4.10). ∎

Remark 4.2 (Simpler bounds for DRS). By using the (more conservative) estimate $\sigma_{\varphi_1} = 0$ when the smooth function φ_1 is convex, and $\sigma_{\varphi_1} = -L_{\varphi_1}$ otherwise, the range of γ can be simplified as follows in case $\lambda \in (0,2]$:

$$\lambda \in (0,2) \begin{cases} \gamma < \frac{1}{L_{\varphi_1}} \quad \text{and} \quad c = \frac{2-\lambda}{2\lambda\gamma} - L_{\varphi_1}\left[\frac{\gamma L_{\varphi_1}}{\lambda} - \frac{1}{2}\right]_+ & \text{if } \varphi_1 \text{ is convex,} \\ \gamma < \frac{2-\lambda}{2L_{\varphi_1}} \quad \text{and} \quad c = \frac{2-\lambda}{2\lambda\gamma} - \frac{L_{\varphi_1}}{\lambda} & \text{otherwise.} \end{cases}$$

$$\lambda = 2 \begin{cases} \gamma < \frac{1}{L_{\varphi_1}} \quad \text{and} \quad c = \frac{\sigma_{\varphi_1}}{4}(1 - \gamma L_{\varphi_1}) & \text{if } \varphi_1 \text{ is strongly convex,} \\ \emptyset & \text{otherwise.} \end{cases}$$

∎

THEOREM 4.3 (Subsequential convergence of DRS). *Suppose that Assumption I is satisfied, and consider a sequence $(s^k, u^k, v^k)_{k\in\mathbb{N}}$ generated by DRS with stepsize γ and relaxation λ as in Theorem 4.1, starting from $s^0 \in \mathbb{R}^p$. The following hold:*

(i) The residual $(u^k - v^k)_{k\in\mathbb{N}}$ vanishes with rate $\mathbf{min}_{i\leq k} \|u^i - v^i\| = o(1/\sqrt{k})$.

(ii) $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ have same cluster points, all of which are stationary for φ and on which φ has same (finite) value, this being the limit of $(\varphi_\gamma^{\mathrm{DR}}(s^k))_{k\in\mathbb{N}}$. In fact, for each k one has $\mathbf{dist}(0, \hat{\partial}\varphi(v^k)) \leq \frac{1-\gamma\sigma_{\varphi_1}}{\gamma}\|u^k - v^k\|$.

(iii) If φ has bounded level sets, then the sequence $(s^k, u^k, v^k)_{k\in\mathbb{N}}$ is bounded.

Proof. To avoid trivialities, we assume that a fixed point is not found in a finite number of iterations, hence that $v^k \neq u^k$ for all $k \in \mathbb{N}$.

♠ 4.3*(i)* Let $c = c(\gamma, \lambda)$ be as in Theorem 4.1. Telescoping the inequality (4.2) yields

$$\tfrac{c\lambda^2}{(1+\gamma L_{\varphi_1})^2} \textstyle\sum_{k\in\mathbb{N}} \|u^k - v^k\|^2 \leq \sum_{k\in\mathbb{N}} [\varphi_\gamma^{\mathrm{DR}}(s^k) - \varphi_\gamma^{\mathrm{DR}}(s^{k+1})] \leq \varphi_\gamma^{\mathrm{DR}}(s^0) - \mathbf{inf}\, \varphi_\gamma^{\mathrm{DR}}.$$

Since $\mathbf{inf}\, \varphi_\gamma^{\mathrm{DR}} = \mathbf{inf}\, \varphi > -\infty$ and $\varphi_\gamma^{\mathrm{DR}}$ is real valued (Proposition 3.2 and Theorem 3.4), it follows that $(u^k - v^k)_{k\in\mathbb{N}}$ is square summable, hence the claimed rate of convergence. Moreover, since $\varphi_\gamma^{\mathrm{DR}}(s^k)$ is decreasing it admits a (finite) limit, be it $\varphi_\star$.

♠ 4.3*(ii)* Since $(u^k - v^k)_{k\in\mathbb{N}} \to 0$, necessarily $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ have same cluster points. Suppose that $(u^k)_{k\in K} \to u'$ for some $K \subseteq \mathbb{N}$ and $u' \in \mathbb{R}^p$. Then, $(v^k)_{k\in K} \to u'$, and since $s^k = u^k + \nabla\varphi_1(u^k)$ (Proposition 2.3*(i)*), continuity of $\nabla\varphi_1$ implies that $(s^k)_{k\in K} \to s' = u' + \gamma\nabla\varphi_1(u')$. From Proposition 2.3*(i)* we infer that $u' = \mathbf{prox}_{\gamma\varphi_1}(s')$.

Similarly, $(u^k - \gamma\nabla\varphi_1(u^k))_{k\in K} \to u' - \gamma\nabla\varphi_1(u')$, and the outer semicontinuity of $\mathbf{prox}_{\gamma\varphi_2}$ [32, Ex. 5.23(b)] combined with (3.1) implies that

$$u' = \lim_{K\ni k\to\infty} v^k \in \limsup_{K\ni k\to\infty} \mathbf{prox}_{\gamma\varphi_2}\left(u^k - \gamma\nabla\varphi_1(u^k)\right) \subseteq \mathbf{prox}_{\gamma\varphi_2}(u' - \gamma\nabla\varphi_1(u')).$$

From (2.7) we then have that $-\nabla\varphi_1(u') \in \hat{\partial}\varphi_2(u')$, hence $0 \in \hat{\partial}\varphi(u')$, as it follows from [32, Ex. 8.8]. Finally, since $v^k \to u'$,

$$\varphi(u') \le \liminf_{K\ni k\to\infty} \varphi(v^k) \le \limsup_{K\ni k\to\infty} \varphi(v^k) \le \limsup_{K\ni k\to\infty} \varphi^{\mathrm{DR}}_\gamma(s^k) = \varphi^{\mathrm{DR}}_\gamma(s') \le \varphi(u'),$$

where the first inequality is due to lower semicontinuity of φ, the third and the last to the sandwiching property (Proposition 3.3), and the equality to the continuity of $\varphi^{\mathrm{DR}}_\gamma$ (Proposition 3.2). This shows that $(\varphi(v^k))_{k\in K} \to \varphi(u') = \varphi^{\mathrm{DR}}_\gamma(s')$, and since $(\varphi^{\mathrm{DR}}_\gamma(s^k))_{k\in\mathbb{N}} \to \varphi_\star$, then necessarily $\varphi(u') = \varphi^{\mathrm{DR}}_\gamma(s') = \varphi_\star$ independently of the cluster point u'. As to the last assert, due to (2.7) the optimality condition of v^k as in (3.1) read $\frac{1}{\gamma}(u^k - v^k) - \nabla\varphi_1(u^k) \in \hat{\partial}\varphi_2(v^k)$. Let $F := \mathrm{id} - \gamma\nabla\varphi_1$ and observe that it is a $(1-\gamma\sigma_f)$-Lipschitz continuous mapping. From the above inclusion one has $\frac{1}{\gamma}(F(u^k) - F(v^k)) \in \hat{\partial}\varphi(v^k)$, hence $\mathbf{dist}(0, \hat{\partial}\varphi(v^k)) \le \frac{1}{\gamma}\|F(u^k) - F(v^k)\| \le \frac{1-\gamma\sigma_{\varphi_1}}{\gamma}\|u^k - v^k\|$.

♠ 4.3*(iii)* Suppose that φ has bounded level sets. Then, it follows from Theorem 3.4*(iii)* that so does $\varphi^{\mathrm{DR}}_\gamma$, and since $s^k \in \mathbf{lev}_{\le\varphi^{\mathrm{DR}}_\gamma(s^0)}\varphi^{\mathrm{DR}}_\gamma$ for all $k \in \mathbb{N}$, then the sequence $(s^k)_{k\in\mathbb{N}}$ is bounded. Due to Lipschitz continuity of $\mathbf{prox}_{\gamma\varphi_1}$ (Proposition 2.3*(ii)*), also $(u^k)_{k\in\mathbb{N}}$ is bounded. In turn, since $v^k - u^k \to 0$ we conclude that also $(v^k)_{k\in\mathbb{N}}$ is bounded. □

The Kurdyka-Łojasiewicz (KL) property is a powerful tool to establish global convergence (as opposed to subsequential convergence) of descent methods, see [1], and semialgebraic functions comprise a wide class of functions that enjoy this property. It was first observed in [26] that an augmented Lagrangian decreases along iterates generated by non-relaxed DRS, cf. (4.1), and global convergence was thus established when φ_1 and φ_2 are semialgebraic functions and the sequence remains bounded. The latter requirement was later shown to hold in [24] when φ has bounded level sets, as Theorem 3.4*(iii)* confirms. The key observation to extend the result of [26] to the tight ranges here provided is that for $\mathscr{L}_\beta$ as in (3.6) one has $\partial\mathscr{L}_{1/\gamma}(u^k, v^k, \gamma^{-1}(u^k - s^k)) \ni (\gamma^{-1}(u^k - v^k),\ 0,\ u^k - v^k)$, owing to the facts that $s^k = u^k + \gamma\nabla\varphi_1(u^k)$ and $\frac{2u^k - s^k - v^k}{\gamma} \in \partial\varphi_2(v^k)$. This ensures the bound $\mathbf{dist}(0, \partial\mathscr{L}_{1/\gamma}(u^k, v^k, \gamma^{-1}(u^k - s^k))) \le \sqrt{1 + 1/\gamma^2}\|u^k - v^k\|$ for all k, which together with the sufficient decrease of Theorem 4.1 and the identity $\varphi^{\mathrm{DR}}_\gamma(s^k) = \mathscr{L}_{1/\gamma}(u^k, v^k, \gamma^{-1}(u^k - s^k))$, allows to replicate the arguments of [26, Thm. 2] to infer global convergence when $\mathscr{L}_{1/\gamma}$ has the KL property.

Theorem 4.4 (Global convergence of DRS [26, Thm. 2]). *Suppose that Assumption I is satisfied, that φ is level bounded, and that φ_1 and φ_2 are semialgebraic. Then, the sequences $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ generated by DRS with γ and λ as in Theorem 4.3 converge to (the same) stationary point for φ.*

4.1. Adaptive variant. As described in Remark 4.2, when the hypoconvexity modulus σ_{φ_1} is not known one can always consider $\sigma_{\varphi_1} = -L_{\varphi_1}$; in case φ_1 is convex, the tighter estimate $\sigma_{\varphi_1} = 0$ is also feasible. In particular, for any $\lambda \in (0, 2)$ the knowledge of L_{φ_1} is enough for determining ranges of γ, although possibly conservative, that comply with Theorem 4.1 and thus make DRS iterations convergent.

When also the Lipschitz constant L_{φ_1} is not available, it is however possible to adjust the stepsize γ along the iterations without losing the convergence properties of Theorem 4.3. This can be done by selecting an initial estimate γ for the stepsize, and reduce it whenever a sufficient decrease condition is violated. Due to the fact that γ may be larger than the unknown threshold $1/[\sigma_{\varphi_1}]_-$ below which $\mathbf{prox}_{\gamma\varphi_1}$ is ensured to be single valued (Proposition 2.3), the DRE may fail to be a well-defined function of s. For this reason, we resort to the augmented Lagrangian interpretation given in (3.5).

The procedure is summarized in Algorithm 4.1. At each iteration, the stepsize γ is reduced whenever a sufficient decrease condition on the augmented Lagrangian is violated.

Algorithm 4.1 DRS with adaptive stepsize.
$\mathscr{L}_\beta$ is the augmented Lagrangian as defined in (3.6).
$\mathrm{DRS}_{\gamma,\lambda} : \mathbb{R}^p \rightrightarrows \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p$ maps $s \in \mathbb{R}^p$ to a triplet (u, v, s^+) as in (DRS).

REQUIRE $s^0 \in \mathbb{R}^p,\ L > 0,\ \lambda \in (0, 2), \gamma, c$ as in Remark 4.2 with L in place of L_{φ_1},
INITIALIZE $(u^0, v^0, s^1) \in \mathrm{DRS}_{\gamma,\lambda}(s^0),\ \mathcal{L}_0 = \mathscr{L}_{1/\gamma}(u^0, v^0, \gamma^{-1}(u^0 - s^0))$
For $k = 1, 2, \dots$ **do**
1: $(u^k, v^k, s^{k+1}) \in \mathrm{DRS}_{\gamma,\lambda}(s^k)$
2: $\mathcal{L}_k = \mathscr{L}_{1/\gamma}(u^k, v^k, \gamma^{-1}(u^k - s^k))$
3: **if** $\mathcal{L}_k > \mathcal{L}_{k-1} - \frac{c\lambda^2}{(1+\gamma L)^2}\|v^{k-1} - u^{k-1}\|^2$ **or** $\varphi(v^k) > \mathcal{L}_k$ **then**
4: $\gamma \leftarrow \gamma/2,\ c \leftarrow 2c,\ L \leftarrow 2L$
5: $(u^{k-1}, v^{k-1}, s^k) \in \mathrm{DRS}_{\gamma,\lambda}(s^{k-1})$
6: $\mathcal{L}_{k-1} \leftarrow \mathscr{L}_{1/\gamma}(u^{k-1}, v^{k-1}, \gamma^{-1}(u^{k-1} - s^{k-1}))$ and go back to step 1

This can happen only a finite number of times, since for γ small enough (3.5) holds and the sufficient decrease property as stated in Theorem 4.1 applies. The recomputation of the previous iterates at steps 5 and 6 is crucial: whenever γ is decreased the previous augmented Lagrangian value $\mathcal{L}_{k-1}$ has to be updated with the new value of γ, for no decrease can be guaranteed when comparing $\mathscr{L}_{1/\gamma}$ and $\mathscr{L}_{1/\gamma'}$ (or $\varphi^{\mathrm{DR}}_\gamma$ and $\varphi^{\mathrm{DR}}_{\gamma'}$) when $\gamma \neq \gamma'$. Finally, note that it may also be the case that γ remains high and lower boundedness cannot be inferred from Theorem 3.4*(i)*. The additional condition $\mathcal{L}_k \geq \varphi(v^k)$ at step 3 prevents the augmented Lagrangian from dropping arbitrarily low. This is a feasible requirement, since as soon as γ falls below $1/L_{\varphi_1}$ the bound of Proposition 3.3*(ii)* applies.

THEOREM 4.5 (Subsequential convergence of adaptive DRS). *Suppose that Assumption* I *is satisfied, and consider the iterates generated by Algorithm* 4.1*. The following hold:*
(i) The residual $(u^k - v^k)_{k\in\mathbb{N}}$ *vanishes with rate* $\min_{i\leq k} \|u^i - v^i\| = o(1/\sqrt{k})$.
(ii) $(u^k)_{k\in\mathbb{N}}$ *and* $(v^k)_{k\in\mathbb{N}}$ *have same cluster points, all of which are stationary for* φ *and on which* φ *has same (finite) value, this being the limit of* $(\mathcal{L}_k)_{k\in\mathbb{N}}$.
(iii) If φ *is level bounded, then the sequence* $(s^k, u^k, v^k)_{k\in\mathbb{N}}$ *is bounded.*

Proof. The sufficient decrease constant in Remark 4.2 satisfies $c(\gamma/2, 2L) = 2c(\gamma, L)$. Therefore, if $L \geq L_{\varphi_1}$ at iteration k, then it follows from (3.6) that $\mathcal{L}_k = \varphi^{\mathrm{DR}}_\gamma(s^k)$, and from Proposition 3.3 and Thm. 4.1 we infer that the condition at step 3 is never passed. Therefore, starting from iteration k the stepsize γ is never decreased, and the algorithm reduces to plain (nonadaptive) DRS. Either way, γ is decreased only a finite number of times; up to possibly discarding the first iterates we may assume that γ is constant (although possibly larger than or equal to $1/L_{\varphi_1}$). The iterates generated by Algorithm 4.1 then satisfy $\varphi(v^k) \leq \mathcal{L}_k$ and $\mathcal{L}_{k+1} \leq \mathcal{L}_k - c'\|u^k - v^k\|^2$ for some constant $c' > 0$. Due to lower boundedness of $\mathcal{L}_k$, by telescoping the second inequality we obtain that $(\|u^k - v^k\|^2)_{k\in\mathbb{N}}$ is summable, hence the claimed rate.

Since $u^k - v^k \to 0$, necessarily u^k and v^k have same cluster points. Suppose that a subsequence $(u^k)_{k\in K}$ converges to a point u'; then, so does $(v^k)_{k\in K}$. Moreover, it follows from [32, Ex. 10.2] that $s^k = u^k + \gamma\nabla\varphi_1(u^k)$ (due to the fact that γ may be larger than $1/[\sigma_{\varphi_1}]_-$, differently from the characterization given in Proposition 2.3*(i)* this condition is only necessary). Thus, for all k it holds that $v^k \in \mathbf{prox}_{\gamma\varphi_2}(2u^k - s^k) = \mathbf{prox}_{\gamma\varphi_2}\left(u^k - \gamma\nabla\varphi_1(u^k)\right)$. From the continuity of $\nabla\varphi_1$ and the outer semicontinuity of $\mathbf{prox}_{\gamma\varphi_2}$, cf. [32, Ex. 5.23(b)], it follows that the limit u' of $(v^k)_{k\in K}$ satisfies $u' \in \mathbf{prox}_{\gamma\varphi_2}(u' - \gamma\nabla\varphi_1(u'))$, and the same reasoning as in the proof of

Theorem 4.3*(ii)* shows that $0 \in \hat{\partial}\varphi(u')$.

Finally, since $\varphi(v^k) \le \mathcal{L}_k \le \mathcal{L}_0$, if φ is level bounded, then necessarily $(v^k)_{k\in\mathbb{N}}$ is bounded, hence so are $(u^k)_{k\in\mathbb{N}}$ and $(s^k)_{k\in\mathbb{N}}$ (since $u^k - v^k \to 0$ and $s^k = u^k + \gamma\nabla\varphi_1(u^k)$). ∎

Global convergence of adaptive DRS again falls as a consequence of [26, Thm. 2].

Theorem 4.6 (Global convergence of adaptive DRS). *Suppose that Assumption I holds, that φ is level bounded, and that φ_1 and φ_2 are semialgebraic. Then, the sequences $(u^k)_{k\in\mathbb{N}}$ and $(v^k)_{k\in\mathbb{N}}$ generated by Algorithm 4.1 converge to (the same) stationary point of φ.*

Apart from the re-evaluation of u- and v-steps whenever γ is decreased, the adaptive variant comes at the additional cost of computing $\varphi_1(u^k)$, $\varphi_1(v^k)$, and $\varphi_2(v^k)$ at each iteration, needed for the evaluation of $\mathcal{L}_k$ and for the test at step 3. The second condition at step 3 is what ensures the augmented Lagrangian to be lower bounded along the generated iterates even if the stepsize γ does not fall below the threshold $1/L_{\varphi_1}$. Whenever a lower bound for the optimal cost is known, the same condition can be guaranteed without the need to compute $\varphi_1(v^k)$. Letting φ_{LB} be a known quantity such that $-\infty < \varphi_{\text{LB}} \le \inf\varphi$, this can be achieved by modifying the condition at step 3 as follows:

3′: **if** $\mathcal{L}_k > \mathcal{L}_{k-1} - \frac{c\lambda^2}{(1+\gamma L)^2}\|v^{k-1} - u^{k-1}\|^2$ **or** $\mathcal{L}_k < \varphi_{\text{LB}}$ **then** ...

The modified condition $\mathcal{L}_k \ge \varphi_{\text{LB}}$ will eventually always be satisfied, owing to the fact that $\mathcal{L}_k = \varphi_\gamma^{\text{DR}}(s^k) \ge \inf\varphi$ for $\gamma < 1/L_{\varphi_1}$. Consequently, with the sole exception of 4.5*(iii)* all claims of Theorem 4.5 hold when step 3 is modified with the here proposed step 3′.

4.2. Tightness of the results. When both φ_1 and φ_2 are convex and $\varphi_1 + \varphi_2$ attains a minimum, well-known results of monotone operator theory guarantee that for any $\lambda \in (0, 2)$ and $\gamma > 0$ the residual $u^k - v^k$ generated by DRS iterations vanishes [4, Cor. 28.3]. In fact, the whole sequence $(u^k)_{k\in\mathbb{N}}$ converges and φ_1 need not even be differentiable. On the contrary, when φ_2 is nonconvex, the bound $\gamma < 1/L_{\varphi_1}$ plays a crucial role, as the next example shows.

Theorem 4.7 (Necessity of $\gamma < 1/L_{\varphi_1}$). *For any $L > 0$ and $\sigma \in [-L, L]$ there exist $\varphi_1, \varphi_2 : \mathbb{R} \to \overline{\mathbb{R}}$ satisfying the following properties*

P1 *φ_1 is L-smooth and σ-hypoconvex;*

P2 *φ_2 is proper and lsc;*

P3 $\arg\min(\varphi_1 + \varphi_2) \neq \emptyset$*;*

P4 *for all $s^0 \in \mathbb{R}$, $\gamma \ge 1/L$, and $\lambda > 0$, the sequence $(s^k)_{k\in\mathbb{N}}$ generated by DRS iterations with stepsize γ and relaxation λ starting from s^0 satisfies $\|s^k - s^{k+1}\| \not\to 0$ as $k \to \infty$.*

Proof. Fix $t > 1$, and let $\varphi = \varphi_1 + \varphi_2$, where $\varphi_2 = \delta_{\{\pm 1\}}$ and

$$\varphi_1(x) = \begin{cases} \frac{L}{2}x^2 & \text{if } x \le t, \\ \frac{L}{2}x^2 - \frac{L-\sigma}{2}(x-t)^2 & \text{otherwise.} \end{cases} \tag{4.12}$$

Notice that $\mathbf{dom}\,\varphi = \{\pm 1\}$, and therefore ± 1 are the unique stationary points of φ (in fact, they are also global minimizers). It can be easily verified that φ_1 and φ_2 satisfy properties 4.7P1, 4.7P2 and 4.7P3. Moreover, $\mathbf{prox}_{\gamma\varphi_1}$ is well defined iff $\gamma < 1/[\sigma]_-$, in which case

$$\mathbf{prox}_{\gamma\varphi_1}(s) = \begin{cases} \frac{s}{1+\gamma L} & \text{if } s \le t(1+\gamma L), \\ \frac{s-\gamma(L-\sigma)t}{1+\gamma\sigma} & \text{otherwise,} \end{cases} \qquad \text{and} \qquad \mathbf{prox}_{\gamma\varphi_2} = \mathbf{sgn}, \tag{4.13}$$

where $\mathbf{sgn}(0) = \{\pm 1\}$. Let now $s^0 \in \mathbb{R}$, $1/L \le \gamma < 1/[\sigma]_-$, and $\lambda > 0$ be fixed, and consider a sequence $(s^k)_{k\in\mathbb{N}}$ generated by DRS with stepsize γ and relaxation λ, starting at s^0. To arrive to a contradiction, suppose that $\|s^k - s^{k+1}\| = \lambda\|u^k - v^k\| \to 0$ as $k \to \infty$. For any $k \in \mathbb{N}$ we

have $v^k = -\mathbf{sgn}(s^k)$ if $s^k \le t(1+\gamma L)$, resulting in

$$u^k - v^k \in \begin{cases} \frac{s^k}{1+\gamma L} + \mathbf{sgn}(s^k) & \text{if } s^k \le t(1+\gamma L), \\ \frac{s^k}{1+\gamma\sigma} - \frac{\gamma(L-\sigma)t}{1+\gamma\sigma} - v^k & \text{otherwise,} \end{cases}$$

where v^k is either 1 or -1 in the second case. Since $u^k - v^k \to 0$, then

$$\mathbf{min}\left\{\left|\tfrac{s^k}{1+\gamma L} + \mathbf{sgn}(s^k)\right|, \left|\tfrac{s^k}{1+\gamma\sigma} - \tfrac{L-\sigma}{1+\gamma\sigma}\gamma t - 1\right|, \left|\tfrac{s^k}{1+\gamma\sigma} - \tfrac{L-\sigma}{1+\gamma\sigma}\gamma t + 1\right|\right\} \to 0.$$

The first element in the set above is always larger than 1, thus eventually s^k will be always close to either $(L-\sigma)\gamma t + (1+\gamma\sigma)$ or $(L-\sigma)\gamma t - (1+\gamma\sigma)$, both of which are strictly smaller than $t(1+\gamma L)$ (since $t > 1$). Therefore, eventually $s^k \le t(1+\gamma L)$ and the residual will be $u^k - v^k = \frac{s^k}{1+\gamma L} + \mathbf{sgn}(s^k)$ which is bounded away from zero, hence the contradiction. ∎

THEOREM 4.8 (Necessity of $0 < \lambda < 2(1+\gamma\sigma)$). *For any $L > 0$ and $\sigma \in [-L, L]$ there exist $\varphi_1, \varphi_2 : \mathbb{R} \to \overline{\mathbb{R}}$ satisfying the following properties*

P1 *φ_1 is L-smooth and σ-hypoconvex;*

P2 *φ_2 is proper, lsc, and strongly convex;*

P3 $\mathbf{arg\,min}(\varphi_1 + \varphi_2) \neq \emptyset$*;*

P4 *for all $0 < \gamma < 1/L$ and $\lambda \ge 2(1+\gamma\sigma)$, the sequence $(s^k)_{k\in\mathbb{N}}$ generated by DRS with stepsize γ and relaxation λ starting from a nonstationary point s^0 satisfies $\|s^k - s^{k+1}\| \not\to 0$.*

Proof. Let $\varphi = \varphi_1 + \varphi_2$, where φ_1 is as in (4.12) with $t = 1$, and $\varphi_2 = \delta_{\{p\}}$ for some $p > 1$. Clearly, properties 4.8P1, 4.8P2, and 4.8P3 are satisfied. Let $\gamma < 1/L$, $\lambda \ge 2(1+\gamma\sigma)$. Starting from $s^0 \neq (1+\gamma\sigma)p + \gamma(L-\sigma)$ (so that $u^0 \neq p$), consider DRS with stepsize γ and relaxation λ. To arrive to a contradiction, suppose that the residual vanishes. Since $v^k = \mathbf{prox}_{\gamma\varphi_2}(2u^k - s^k) = p$, necessarily $u^k \to p$; therefore, eventually $u^k > 1$ and in particular

$$u^{k+1} + \gamma\tfrac{L-\sigma}{1+\gamma\sigma} = \tfrac{1}{1+\gamma\sigma}s^{k+1} = \tfrac{1}{1+\gamma\sigma}(s^k + \lambda(p - u^k)) = u^k + \gamma\tfrac{L-\sigma}{1+\gamma\sigma} + \tfrac{\lambda}{1+\gamma\sigma}(p - u^k),$$

where the identity $s^k = (1+\gamma\sigma)u^k + \gamma(L-\sigma)$ was used, cf. (4.13). Therefore,

$$\left|u^{k+1} - p\right| = \left|1 - \tfrac{\lambda}{1+\gamma\sigma}\right|\left|u^k - p\right| \ge \left|u^k - p\right|,$$

where the inequality is due to the fact that $\lambda \ge 2(1+\gamma\sigma)$. Since $u^0 \neq p$ due to the choice of s^0, apparently $(u^k)_{k\in\mathbb{N}}$ is bounded away from p, hence the contradiction. ∎

Let us draw some conclusions:

- The nonsmooth function φ_2 is (strongly) convex in Theorem 4.8, therefore even for fully convex formulations the bound $0 < \lambda < 2(1+\gamma\sigma_{\varphi_1})$ need be satisfied.
- If $\lambda > 2$ (which is feasible only if φ_1 is strongly convex, i.e., if $\sigma_{\varphi_1} > 0$), then, regardless of whether also φ_2 is (strongly) convex or not, we obtain that *the stepsize must be lower bounded as* $\gamma > \frac{\lambda-2}{2\sigma_{\varphi_1}}$. In the more general setting of σ-strongly monotone operators in Hilbert spaces (with $\sigma \ge 0$) the similar bound $\lambda < \mathbf{min}\{2(1+\gamma\sigma), 2+\gamma\sigma+1/\gamma\sigma\}$ has been recently established in [28].
- Combined with the bound $\gamma < 1/L_{\varphi_1}$ shown in Theorem 4.7, we infer that (at least when φ_2 is nonconvex) necessarily $0 < \lambda < 2(1+\sigma_{\varphi_1}/L_{\varphi_1})$ and consequently $\lambda \in (0,4)$.

THEOREM 4.9 (Tightness). *Unless the generality of Assumption I is sacrificed, when $\lambda \in (0,2)$ or φ_1 is not strongly convex the bound $\gamma < \mathbf{min}\left\{\frac{1}{L_{\varphi_1}}, \frac{2-\lambda}{2[\sigma_{\varphi_1}]_-}\right\}$ is tight for ensuring convergence of DRS. Similarly, PRS (i.e., DRS with $\lambda = 2$) is ensured to converge iff φ_1 is strongly convex and $\gamma < 1/L_{\varphi_1}$.*

5. Alternating direction method of multipliers. While the classical interpretation of ADMM as DRS applied to the dual formulation is limited to convex problems, it has been recently observed that the two schemes are in fact related through a primal equivalence, when $\lambda = 1$. A proof of this fact can be found in [5, Rem. 3.14] when $A = -B = \mathrm{I}$; in turn, [36, Thm. 1] shows that there is no loss of generality in limiting the analysis to this case. Patterning the arguments of [36] in the next subsection we will show that the equivalence can be further extended to any relaxation parameter λ. To this end, we introduce the notion of *image function*, also known as *epi-composition* or *infimal post-composition* [2, 4, 32].

Definition 5.1 (Image function). *Given $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $C \in \mathbb{R}^{m\times n}$, the* image function $(Ch) : \mathbb{R}^m \to [-\infty, +\infty]$ *is defined as*

$$(Ch)(s) := \inf_{x\in\mathbb{R}^n} \{h(x) \mid Cx = s\}.$$

We now list some properties of image functions; the proofs are deferred to Appendix B.

Proposition 5.2. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $C \in \mathbb{R}^{p\times n}$. Suppose that for some $\beta > 0$ the set-valued mapping $X_\beta(s) := \operatorname{arg\,min}_{x\in\mathbb{R}^n} \left\{h(x) + \frac{\beta}{2}\|Cx - s\|^2\right\}$ is nonempty for all $s \in \mathbb{R}^p$. Then,*

(i) The image function (Ch) is proper.
(ii) $(Ch)(Cx_\beta) = h(x_\beta)$ for all $s \in \mathbb{R}^p$ and $x_\beta \in X_\beta(s)$.
(iii) $\operatorname{prox}_{(Ch)/\beta} = CX_\beta$.

Proposition 5.3. *For a function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $C \in \mathbb{R}^{p\times n}$, let $X : \mathbb{R}^p \rightrightarrows \mathbb{R}^n$ be defined as $X(s) := \operatorname{arg\,min}_{x\in\mathbb{R}^n} \{h(x) \mid Cx = s\}$. Then, for all $\bar{s} \in C\,\operatorname{dom} h$ and $\bar{x} \in X(\bar{s})$ it holds that*

$$C^\top \hat{\partial}(Ch)(\bar{s}) \subseteq \hat{\partial} h(\bar{x}).$$

Proposition 5.4 (Strong convexity of the image function). *Suppose that $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper, lsc, and σ_h-strongly convex. Then, for every $C \in \mathbb{R}^{p\times n}$ the image function (Ch) is $\sigma_{(Ch)}$-strongly convex with $\sigma_{(Ch)} = \sigma_h/\|C\|^2$.*

5.1. A universal equivalence of DRS and ADMM. Let us eliminate the linear coupling between x and z in the ADMM problem formulation (1.2), so as to bring it into DRS form (1.1). To this end, let us introduce a slack variable $s \in \mathbb{R}^p$ and rewrite (1.2) as

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

Invoking [32, Prop. 1.35], we may minimize first with respect to (x, z) to arrive to

$$\underset{s\in\mathbb{R}^p}{\text{minimize}} \; \inf_{x\in\mathbb{R}^m} \{f(x) \mid Ax = s\} + \inf_{z\in\mathbb{R}^n} \{g(z) \mid Bz = b - s\}.$$

The two parametric infima define two image functions, cf. Definition 5.1: indeed, ADMM problem formulation (1.2) can be expressed as

$$\underset{s\in\mathbb{R}^p}{\text{minimize}} \; (Af)(s) + (Bg)(b - s), \tag{5.1}$$

which is exactly (1.1) with $\varphi_1 = (Af)$ and $\varphi_2 = (Bg)(b - \cdot)$. Apparently, unless A and B are injective the correspondence between variable s in (5.1) and variables x, z in (1.2) may fail to be one to one, as s is associated to sets of variables $x \in X(s)$ and $z \in Z(s)$ defined as

$$X(s) := \operatorname*{arg\,min}_{x\in\mathbb{R}^m} \{f(x) \mid Ax = s\} \quad \text{and} \quad Z(s) := \operatorname*{arg\,min}_{z\in\mathbb{R}^n} \{g(z) \mid Bz = b - s\}.$$

THEOREM 5.5 (Primal equivalence of DRS and ADMM). *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 large enough penalty* $\beta > 0$ *so that any ADMM minimization subproblem has solutions. Let*

$$\begin{cases} s := Ax - y/\beta \\ u := Ax \\ v := b - Bz \end{cases} \quad \text{and, similarly,} \quad \begin{cases} s^+ := Ax^+ - y^+/\beta \\ u^+ := Ax^+ \\ v^+ := b - Bz^+. \end{cases} \tag{5.2}$$

Then, the variables are related as follows:

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

Moreover,

(i) $\varphi_1(u^+) = (Af)(Ax^+) = f(x^+)$,

(ii) $\varphi_2(v^+) = (Bg)(Bz^+) = g(z^+)$,

(iii) $-y^+ \in \hat{\partial}\varphi_1(u^+) = \hat{\partial}(Af)(Ax^+)$,

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

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

If, additionally, A has full row rank, $\varphi_1 \in C^{1,1}(\mathbb{R}^p)$ *is* L_{φ_1}*-smooth, and* $\beta > L_{\varphi_1}$*, then it also holds that*

(vi) $\varphi_\gamma^{\mathrm{DR}}(s^+) = \mathscr{L}_\beta(x^+, z^+, y^+)$.

Proof. Observe first that, as shown in Proposition 5.2*(iii)*, it holds that

$$\mathbf{prox}_{\gamma\varphi_1} = A\,\mathbf{arg\,min}\left\{f + \tfrac{1}{2\gamma}\|A\cdot - s\|^2\right\}. \tag{5.3a}$$

Similarly, with a simple change of variable one obtains that

$$\mathbf{prox}_{\gamma\varphi_2} = b - B\,\mathbf{arg\,min}\left\{g + \tfrac{1}{2\gamma}\|B\cdot + s - b\|^2\right\}. \tag{5.3b}$$

Let (s, u, v) and (s^+, u^+, v^+) be as in (5.2). We have

$$s + \lambda(v - u) = Ax - \tfrac{1}{\beta}y - \lambda(Ax + Bz - b) = Ax - \tfrac{1}{\beta}y^{+/2} - (Ax + Bz - b) = -\tfrac{1}{\beta}y^+ + Ax^+ = s^+,$$

where in the second and third equality the ADMM update rules for $y^{+/2}$ and y^+, respectively, were used. Moreover,

$$u^+ = Ax^+ \in A\,\mathbf{arg\,min}\,\mathscr{L}_\beta(\,\cdot\,, z, y^{+/2}) \overset{(5.3a)}{=} \mathbf{prox}_{\varphi_1/\beta}(b - Bz - y^{+/2}/\beta) = \mathbf{prox}_{\varphi_1/\beta}(s^+),$$

where the last equality uses the identity $b - Bz - y^{+/2}/\beta = v - \gamma y + (1-\lambda)(u - v) = s + \lambda(v - u) = s^+$. Next, observe that $2u^+ - s^+ = 2Ax^+ - (Ax^+ - y^+/\beta) = Ax^+ + y^+/\beta$, hence

$$v^+ = b - Bz^+ \in b - B\,\mathbf{arg\,min}\,\mathscr{L}_\beta(x^+, \,\cdot\,, y^+) \overset{(5.3b)}{=} \mathbf{prox}_{\varphi_2/\beta}(Ax^+ + y^+/\beta) = \mathbf{prox}_{\varphi_2/\beta}(2u^+ - s^+).$$

Let us now show the numbered claims.

♠ 5.5*(i)* & 5.5*(ii)* Follow from Proposition 5.2*(ii)*.

♠ 5.5*(iii)* Since $u^+ \in \mathbf{prox}_{\gamma\varphi_1}(s^+)$ and $-y^+ = \frac{1}{\gamma}(s^+ - u^+)$, the claim follows from (2.7).

♠ 5.5*(iv)* This follows from the optimality conditions of x^+ in the ADMM-subproblem defining the x-update (the claim can also be deduced from 5.5*(iii)* and Proposition 5.3).

♠ 5.5*(v)* The optimality conditions in the ADMM-subproblem defining the z-update read

$$0 \in \hat{\partial}_z \mathscr{L}_\beta(x^{k+1}, z^{k+1}, y^{k+1}) = \hat{\partial} g(z^{k+1}) + B^\top(Ax^{k+1} + Bz^{k+1} - b + y^{k+1}/\beta),$$

and the claim readily follows.

♠ 5.5*(vi)* Suppose now that φ_1 is L_{φ_1}-smooth (hence A is surjective, for otherwise φ_1 has not full domain), and that $\beta > L_{\varphi_1}$. Due to smoothness, the inclusion in 5.5*(iii)* can be strengthened to $\nabla\varphi_1(u^+) = -y^+$. We may then invoke the expression (3.3) of the DRE (recall that the minimum is attained at v^+) to obtain

$$\begin{aligned}\varphi_\gamma^{\mathrm{DR}}(s^+) &= \varphi_1(u^+) + \varphi_2(v^+) + \langle \nabla\varphi_1(u^+), v^+ - u^+\rangle + \tfrac{1}{2\gamma}\|v^+ - u^+\|^2 \\ &= f(x^+) + g(z^+) + \langle y^+, Ax^+ + Bz^+ - b\rangle + \tfrac{\beta}{2}\|Ax^+ + Bz^+ - b\|^2 = \mathscr{L}_\beta(x^+, z^+, y^+). \qquad \square\end{aligned}$$

5.2. Convergence of the ADMM. In order to extend the theory developed for DRS to ADMM we shall impose that φ_1 and φ_2 as in (5.1) comply with Assumption I. This motivates the following blanket requirement.

Assumption II (Requirements for the ADMM formulation (1.2)). *The following hold:*

A1 $f : \mathbb{R}^m \to \mathbb{R}$ *and* $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ *are proper and lsc.*

A2 A *is surjective.*

A3 $\varphi_1 := (Af) \in \mathrm{C}^{1,1}(\mathbb{R}^p)$ *is* $L_{(Af)}$*-smooth, hence* $\sigma_{(Af)}$*-hypoconvex with* $|\sigma_{(Af)}| \le L_{(Af)}$.

A4 $\varphi_2 := (Bg)$ *is lsc.*

A5 *Problem* (1.2) *has a solution:* $\operatorname{\mathbf{arg\,min}} \Phi \neq \emptyset$, *where* $\Phi(x, z) := f(x) + g(z) + \delta_S(x, z)$ *and* $S := \{(x, z) \in \mathbb{R}^m \times \mathbb{R}^n \mid Ax + Bz = b\}$ *is the feasible set.*

These requirements generalize Assumption I by allowing linear constraints more generic than $x - z = 0$, cf. (3.7). Surjectivity of A is as general as the inclusion $\mathbf{range}\, B \subseteq b + \mathbf{range}\, A$. In fact, (up to an orthogonal transformation) without loss of generality we may assume that $A = \binom{A'}{}$ for some surjective matrix $A' \in \mathbb{R}^{r\times m}$, where $r = \mathbf{rank}\, A$, stacked over a $(p - r) \times n$ zero matrix. Then, in light of the prescribed range inclusion necessarily $B = \binom{B'}{}$ and $b = \binom{b'}{}$, for some $B' \in \mathbb{R}^{r\times n}$ and $b \in \mathbb{R}^r$. Then, problem (1.2) can be simplified to the minimization of $f(x) + g(z)$ subject to $A'x + B'z = b'$, which satisfies the needed surjectivity property.

Notice further that Lipschitz differentiability of (Af) guarantees that all the ADMM subproblems admit minimizers when $\beta > L_{(Af)}$. This is a consequence of the $1/L_{(Af)}$-prox-boundedness of (Af) (which follows from Remark 3.1 and the fact that $\mathbf{inf}(Af)+(Bg)(b-\cdot) = \mathbf{inf}\,\Phi$), and the relation between ADMM subproblems and proximal mapping in Proposition 5.2*(iii)*.

Theorem 5.6 (Convergence of ADMM). *Suppose that Assumption II is satisfied, and let* φ_1, φ_2, *and* Φ *be as defined therein. Starting from* $(x^{-1}, y^{-1}, z^{-1}) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n$, *consider a sequence* $(x^k, y^k, z^k)_{k\in\mathbb{N}}$ *generated by ADMM with penalty* $\beta = 1/\gamma$ *and relaxation* λ, *where* γ *and* λ *are as in Theorem 4.1. The following hold:*

(i) $\mathscr{L}_\beta(x^{k+1}, z^{k+1}, y^{k+1}) \le \mathscr{L}_\beta(x^k, z^k, y^k) - \frac{c\lambda^2}{(1+\gamma L_{(Af)})^2}\|Ax^k + Bz^k - b\|^2$, *where* c *is as in Theorem 4.1, and the residual* $(Ax^k+Bz^k-b)_{k\in\mathbb{N}}$ *vanishes with* $\mathbf{min}_{i\le k}\|Ax^i+Bz^i-b\| = o(1/\sqrt{k})$.

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

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

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

(iii) 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,* $f \in \mathrm{C}^{1,1}(\mathbb{R}^m)$, *then the sequence* $(x^k, y^k, z^k)_{k\in\mathbb{N}}$ *is bounded.*

Proof. Let $s^0 := Ax^0 - y^0/\beta$, and consider the sequence $(s^k, u^k, v^k)_{k\in\mathbb{N}}$ generated by DRS applied to (5.1), with stepsize γ, relaxation λ, and starting from s^0. Then, for all $k \in \mathbb{N}$ it follows from Theorem 5.5 that the variables are related as

$$\begin{cases} s^k = Ax^k - y^k/\beta \\ u^k = Ax^k \\ v^k = b - Bz^k, \end{cases} \quad \text{and satisfy} \quad \begin{cases} \varphi_1(u^k) = f(x^k) \\ \varphi_2(v^k) = g(z^k) \\ \varphi_\gamma^{\mathrm{DR}}(s^k) = \mathscr{L}_\beta(x^k, z^k, y^k) \end{cases} \quad \text{and} \quad \begin{cases} y^k = -\nabla\varphi_1(u^k) \\ -A^\top y^k \in \hat{\partial} f(x^k) \\ \mathbf{dist}(-B^\top y^k, \hat{\partial} g(z^k)) \to 0. \end{cases}$$

♠ 5.6*(i)* Readily follows from Theorems 4.1 and 4.3.

♠ 5.6*(ii)* Suppose that for some $K \subseteq \mathbb{N}$ the subsequence $(x^k, y^k, z^k)_{k\in K}$ converges to (x, y, z); then, necessarily $Ax + Bz = b$. Moreover,

$$(Af)(Ax) \le f(x) \le \liminf_{K\ni k\to\infty} f(x^k) = \liminf_{K\ni k\to\infty} (Af)(Ax^k) \le \limsup_{K\ni k\to\infty} (Af)(Ax^k) = (Af)(Ax),$$

where the second inequality is due to the fact that f is lsc, and the last equality to the fact that (Af) is continuous. Therefore, $f(x^k) \to f(x)$, and the inclusion $-A^\top y^k \in \hat{\partial} f(x^k)$ in light of the definition of subdifferential results in $-A^\top y \in \partial f(x)$. In turn, since $\varphi_1(u^k) + \varphi_1(v^k)$ converges to $\varphi_1(Ax) + \varphi_2(b - Bz) = (Af)(Ax) + (Bg)(Bz)$ as it follows from Theorem 4.3*(ii)*, one has

$$\liminf_{K\ni k\to\infty} f(x^k) + g(z^k) = \liminf_{K\ni k\to\infty} (Af)(Ax^k) + (Bg)(Bz^k) = (Af)(Ax) + (Bg)(Bz).$$

The first term is lower bounded by $f(x) + g(z)$ due to lsc, and the last one is upper bounded by $f(x) + g(z)$ due to the definition of image function. Therefore $f(x^k) + g(z^k) \to f(x) + g(z)$ as $K \ni k \to \infty$, and since $f(x^k)$ converges to $f(x)$ we conclude that $g(z^k)$ converges to $g(z)$. In turn, since $\mathbf{dist}(-B^\top y^k, \hat{\partial} g(z^k)) \to 0$, g-attentive outer semicontinuity of ∂g, see [32, Prop. 8.7], implies that $-B^\top y \in \partial g(z)$. Finally, that $f(x) + g(z)$ equals the limit of the whole sequence $(\mathscr{L}_\beta(x^k, z^k, y^k))_{k\in\mathbb{N}}$ then follows from Theorem 4.3*(ii)* through the identity $\varphi_\gamma^{\mathrm{DR}}(s^k) = \mathscr{L}_\beta(x^k, z^k, y^k)$.

♠ 5.6*(iii)* Once we show that $\varphi = \varphi_1 + \varphi_2$ is level bounded, boundedness of $(Ax^k, Bz^k, y^k)_{k\in\mathbb{N}}$ will follow from Theorem 4.3*(iii)*. For $\alpha \in \mathbb{R}$ we have

$$\begin{aligned} \mathbf{lev}_{\le\alpha}\, \varphi &= \left\{ s \mid \inf_x \{f(x) \mid Ax = s\} + \inf_z \{g(z) \mid Bz = b - s\} \le \alpha \right\} \\ &= \left\{ s \mid \inf_{x,z} \{f(x) + g(z) \mid Ax = s,\ Bz = b - s\} \le \alpha \right\} \\ &= \{Ax \mid f(x) + g(z) \le \alpha,\ \exists z : Ax + Bz = b\} = \{Ax \mid (x, z) \in \mathbf{lev}_{\le\alpha}\, \Phi,\ \exists z\}. \end{aligned}$$

Since $\|Ax\| \le \|A\|\|x\| \le \|A\|\|(x, z)\|$ for any x, z, it follows that if $\mathbf{lev}_{\le\alpha}\, \Phi$ is bounded, then so is $\mathbf{lev}_{\le\alpha}\, \varphi$. Suppose now that $f \in \mathrm{C}^{1,1}(\mathbb{R}^n)$ is L_f-smooth, and for all $k \in \mathbb{N}$ let $\xi^k := x^k - A^\top(AA^\top)^{-1}(Ax^k + Bz^k - b)$. Then, $A\xi^k = b - Bz^k$, hence $f(\xi^k) + g(z^k) = \Phi(\xi^k, z^k)$, and $\xi^k - x^k \to 0$ as $k \to \infty$. We have

$$\begin{aligned} |\Phi(\xi^k, z^k) - (f(x^k) + g(z^k))| = \left|f(\xi^k) - f(x^k)\right| &\le |\langle \nabla f(x^k), \xi^k - x^k\rangle| + \tfrac{L_f}{2}\|\xi^k - x^k\|^2 \\ &\le \left|\langle y^k, Ax^k - A\xi^k\rangle\right| + \tfrac{L_f}{2}\|A^\top(AA^\top)^{-1}\|^2\|Ax^k + Bz^k - b\|^2, \end{aligned}$$

where the second inequality uses the identity $\nabla f(x^k) = -A^\top y^k$, cf. Theorem 5.5*(iv)*. In particular, $f(\xi^k) - f(x^k) \to 0$ as $k \to \infty$, and therefore $\Phi(\xi^k, z^k)$ converges to a finite quantity (the limit of $\mathscr{L}_\beta(x^k, z^k, y^k)$). Since Φ is level bounded, $(\xi^k, z^k)_{k\in\mathbb{N}}$ is bounded and thus so is $(x^k)_{k\in\mathbb{N}}$. □

Since no restriction is made on the initial triplet, the virtual iteration $k = -1$ is considered in the statement of Theorem 5.6 so as to ensure that for all $k \geq 0$ the triplets (x^k, y^k, z^k) are the output of an ADMM step whence the DRS equivalence of Theorem 5.5 can be invoked. Smoothness of f as required in Theorem 5.6*(iii)* is a standing assumption in the (proximal) ADMM analysis of [25], which, together with the restriction $A = \mathrm{I}$, ensures that $(Af) = f$ complies with Requirement IIA3. Our requirement of level boundedness of Φ to ensure boundedness of the sequences generated by ADMM is milder than that of [25, Thm. 3], which instead requires coercivity of either f or g.

Remark 5.7 (Simpler bounds for ADMM). In parallel with the simplifications outlined in Remark 4.2 for DRS, denoting $L := L_{(Af)}$ simpler (more conservative) bounds for the penalty parameter β in ADMM are, in case $\lambda \in (0, 2]$:

$$\lambda \in (0,2) \begin{cases} \beta > L & \text{and } c = \beta\frac{2-\lambda}{2\lambda} - L[\frac{L}{\beta\lambda} - \frac{1}{2}]_+ & \text{if } f \text{ is convex,} \\ \beta > \frac{2L}{2-\lambda} & \text{and } c = \beta\frac{2-\lambda}{2\lambda} - \frac{L}{\lambda} & \text{otherwise,} \end{cases}$$

$$\lambda = 2 \begin{cases} \beta > L & \text{and } c = \frac{\sigma_f}{4\|A\|^2}(1 - L/\beta) & \text{if } f \text{ is strongly convex,} \\ & \emptyset & \text{otherwise,} \end{cases}$$

The case $\lambda = 2$ uses Proposition 5.4 to infer strong convexity of (Af) from that of f. □

The Tarski-Seidenberg theorem ensures that $\varphi_1 := (Af)$ and $\varphi_2 := (Bg)(b - \cdot)$ are semi-algebraic functions provided f and g are, see e.g., [8]. Therefore, sufficient conditions for global convergence of ADMM follow from the similar result for DRS stated in Theorem 4.4, through the primal equivalence of the algorithms illustrated in Theorem 5.5. We should emphasize, however, that the equivalence identifies $u^k = Bz^k$ and $v^k = b - Ax^k$; therefore, only convergence of $(Ax^k, y^k, Bz^k)_{k\in\mathbb{N}}$ can be deduced, as opposed to that of $(x^k, y^k, z^k)_{k\in\mathbb{N}}$.

Theorem 5.8 (Global convergence of ADMM). *Suppose Assumption II is satisfied, and let Φ be as defined therein. If Φ is level bounded and f and g are semialgebraic, then the sequence $(Ax^k, y^k, Bz^k)_{k\in\mathbb{N}}$ generated by ADMM with β and λ as in Theorem 5.6 converges.*

5.3. Adaptive variant. Similar to what done for DRS, one can ensure a sufficient decrease on the augmented Lagrangian without knowing the exact value of $L_{(Af)}$, when $\lambda \in (0, 2)$. However, due to the implicitness of $\varphi_1 = (Af)$, enforcing the inequality $\varphi(v^k) \leq \mathcal{L}_k$ as in step 3 of Algorithm 4.1, needed to ensure the lower boundedness of $(\mathscr{L}_{1/\gamma}(x^k, z^k, y^k))_{k\in\mathbb{N}}$, may not be possible. Indeed, although we may exploit (5.2) and Theorem 5.5*(ii)* to arrive to

$$\varphi(v^k) = \varphi_1(v^k) + \varphi_2(v^k) = (Af)(b - Bz^k) + g(z^k),$$

the value of $(Af)(b-Bz^k)$ may not be readily available. For this reason we adopt the alternative proposed at the end of Subsection 4.1, which requires prior knowledge of a constant $\Phi_{\text{LB}} \leq \inf \Phi$. This detail apart, the adaptive variant of DRS can be easily translated into the adaptive ADMM version of Algorithm 5.1, in which the penalty β is suitably adjusted. For the sake of simplicity, we only consider the case $\lambda = 1$, so that the half-update $y^{+/2}$ can be discarded.

Theorem 5.9 (Subsequential convergence of adaptive ADMM). *Suppose that Assumption II is satisfied. Then, the following hold for the iterates generated by Algorithm 5.1:*

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

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

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

Algorithm 5.1 ADMM with adaptive stepsize ($\lambda = 1$ for simplicity).
$\mathrm{ADMM}_\beta : \mathbb{R}^p \times \mathbb{R}^n \rightrightarrows \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R}^n$ maps (y, z) to a triplet (x^+, y^+, z^+) as in (ADMM) with $\lambda = 1$ (since $\lambda = 1$, the update does not depend on x).

REQUIRE $(y^{-1}, z^{-1}) \in \mathbb{R}^p \times \mathbb{R}^n$, $L > 0$, β, c as in Remark 5.7 with $\lambda = 1$,
$\Phi_{\text{LB}} \in (-\infty, \inf \Phi]$ (a known lower bound for the optimal cost)
INITIALIZE $(x^0, y^0, z^0) \in \mathrm{ADMM}_\beta(y^{-1}, z^{-1})$, $\mathcal{L}_0 = \mathscr{L}_\beta(x^0, z^0, y^0)$
For $k = 0, 1, \dots$ **do**
1: $(x^{k+1}, y^{k+1}, z^{k+1}) \in \mathrm{ADMM}_\beta(y^k, z^k)$
$\mathcal{L}_{k+1} = \mathscr{L}_\beta(x^{k+1}, y^{k+1}, z^{k+1})$
2: **if** $\mathcal{L}_{k+1} > \mathcal{L}_k - \frac{c\lambda^2}{(1+L/\beta)^2}\|Ax^k + Bz^k - b\|^2$ **or** $\mathcal{L}_{k+1} < \Phi_{\text{LB}}$ **then**
3: $\beta \leftarrow 2\beta$, $c \leftarrow 2c$, $L \leftarrow 2L$
$(x^k, y^k, z^k) \in \mathrm{ADMM}_\beta(y^{k-1}, z^{k-1})$
$\mathcal{L}_k \leftarrow \mathscr{L}_\beta(x^k, y^k, z^k)$ and go back to step 1

(ii) The residual $(\|Ax^k + Bz^k - b\|)_{k\in\mathbb{N}}$ *vanishes with rate* $\min_{i\le k} \|Ax^i + Bz^i - b\| \le o(1/\sqrt{k})$. *In particular, the claims hold if at some iteration the inequality* $\beta > L_{(Af)}$ *is satisfied. In this case, and if the cost function* Φ *is level bounded, the following also hold:*
(iii) the sequence $(Ax^k, y^k, Bz^k)_{k\in\mathbb{N}}$ *is bounded.*
(iv) the sequence $(Ax^k, y^k, Bz^k)_{k\in\mathbb{N}}$ *is convergent if* f *and* g *are semialgebraic.*

5.4. Sufficient conditions. This subsection provides sufficient conditions on f and g ensuring that Assumption II is satisfied.

5.4.1. Lower semicontinuity of the image function.

PROPOSITION 5.10 (Lsc of (Bg)). *Suppose that Requirements* IIA1 *and* IIA2 *are satisfied. Then,* (Bg) *is proper. It is also lsc provided that the set* $Z(s) := \arg\min_z \{g(z) \mid Bz = s\}$ *is nonempty for all* $\bar{z} \in \mathbf{dom}\, g$, *and that* $\mathbf{dist}(0, Z(s))$ *is bounded for all* $s \in B\,\mathbf{dom}\, g$ *close to* $B\bar{z}$.

Proof. Properness is shown in Proposition 5.2*(i)*. Suppose that $(s_k)_{k\in\mathbb{N}} \subseteq \mathbf{lev}_{\le\alpha}(Bg)$ for some $\alpha \in \mathbb{R}$ and that $s_k \to \bar{s}$. Then, due to [32, Thm. 1.6] it suffices to show that $\bar{s} \in \mathbf{lev}_{\le\alpha}(Bg)$. The assumption ensures the existence of a bounded sequence $(z_k)_{k\in\mathbb{N}}$ such that eventually $Bz_k = s_k$ and $(Bg)(s_k) = g(z_k)$. By possibly extracting, $z_k \to \bar{z}$ and necessarily $B\bar{z} = \bar{s}$. Then,

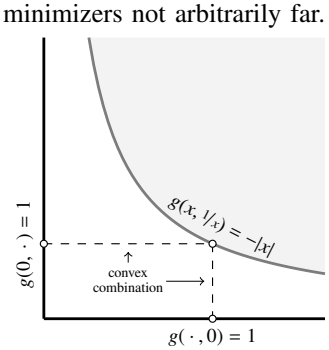
$$(Bg)(\bar{s}) \le g(\bar{z}) \le \liminf_{k\to\infty} g(z_k) = \liminf_{k\to\infty} (Bg)(s_k) \le \alpha,$$

hence $\bar{s} \in \mathbf{lev}_{\le\alpha}(Bg)$. □

The requirement in Proposition 5.10 is weaker than Lipschitz continuity of the map $s \mapsto Z(s)$, which is the standing assumption in [35]. In fact, no uniqueness or boundedness of the sets of minimizers is required, but only the existence of minimizers not arbitrarily far. The pathology occurring when this condition is not met can be well visualized by considering $g : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$g(x, y) = \begin{cases} -|x| & \text{if } |xy| \ge 1, \\ 1 - q(|xy|)(1 + |x|) & \text{otherwise,} \end{cases} \tag{5.4}$$

where $q(t)$ is any function such that $q(0) = 0 < q(t) < 1 = q(1)$ for all $t \in (0, 1)$. On the right, a graphical representation of the piecewise definition on the positive orthant of $\mathbb{R}^2$ (the



function is mirrored in all other orthants). On the axes, f achieves its maximum value, that is, 1. In the gray region $|xy| \geq 1$, $f(x, y) = -|x|$. In the white portion, f is extended by means of a convex combination of 1 and $-|x|$. Function g and $B := [1\ 0]$ are ADMM-feasible, meaning that $\mathbf{arg\,min}_{w\in\mathbb{R}^2} \left\{g(w) + \frac{\beta}{2}\|Bw - s\|^2\right\} \neq \emptyset$ for all $s \in \mathbb{R}$ and β large enough (in fact, for all $\beta > 0$, being $g(\,\cdot\,, y) + \frac{\beta}{2}\|\,\cdot\, - s\|^2$ coercive for any $y \in \mathbb{R}$). However, $(Bg)(s) = -|s|$ if $s \neq 0$ while $(Bg)(0) = 1$, resulting in the lack of lsc at $s = 0$. For every $s \neq 0$, the minimizers of g with smallest norm in the set $\{(x, y) \mid x = s\} = \{s\} \times \mathbb{R}$ are $(s, \pm s^{-1})$, and while escaping to infinity (in norm) they satisfy $g(s, \pm s^{-1}) = -|s| \to 0$ as $s \to 0$. However, if instead $s = 0$ is fixed (as opposed to $s \to 0$), then the pathology comes from the fact that $g(0, \cdot\,) \equiv 1 > 0$. The *interpolating* function q simply models the transition from a constant function on the axes and a linear function in the regions delimited by the hyperbolae. For any $k \in \mathbb{N}$ it can thus be chosen such that g is k times continuously differentiable; the choice $q(t) = \frac{1}{2}(1 - \mathbf{cos}\,\pi t)$, for instance, makes $g \in C^1(\mathbb{R}^2)$. In particular, (high-order) continuous differentiability is not enough for (Bg) to be lsc.

The next result provides necessary and sufficient conditions ensuring the image function (Bg) to inherit lower semicontinuity from that of g. It will be evident that pathological cases such as the one depicted in (5.4) may only occur due to the behavior of g at infinity.

THEOREM 5.11. *For any lsc function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $B \in \mathbb{R}^{p\times n}$, (Bg) is lsc iff*

$$\liminf_{\substack{\|d\|\to\infty \\ Bd\to 0}} g(\bar{z} + d) \geq \inf_{d\in\mathbf{ker}\,B} g(\bar{z} + d) \qquad \forall \bar{z} \in \mathbf{dom}\, g. \tag{5.5}$$

In particular, for any lsc and level bounded function $g : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $B \in \mathbb{R}^{p\times n}$, (Bg) is lsc.

Proof. Observe first that the right-hand side in (5.5) is $(Bg)(B\bar{z})$. Suppose now that (5.5) holds, and given $\bar{s} \in \mathbf{dom}(Bg)$ consider a sequence $(s_k)_{k\in\mathbb{N}} \subseteq \mathbf{lev}_{\leq\alpha}(Bg)$ for some $\alpha \in \mathbb{R}$ and such that $s_k \to \bar{s}$. Then, it suffices to show that $\bar{s} \in \mathbf{lev}_{\leq\alpha}(Bg)$. Let $(z_k)_{k\in\mathbb{N}}$ be such that $Bz_k = s_k$ and $g(z_k) \leq (Bg)(s_k) + 1/k$ for all $k \in \mathbb{N}$. If, up to possibly extracting, there exists z such that $z^k \to z$ as $k \to \infty$, then the claim follows with a similar reasoning as in the proof of Proposition 5.10. Suppose, instead, that $t_k := \|z_k\| \to \infty$ as $k \to \infty$, and let $d_k := z_k - \bar{z}$, where $\bar{z} \in \mathbf{dom}\, g$ is any such that $B\bar{z} = s$ (such a $\bar{z}$ exists, being $\bar{s} \in \mathbf{dom}(Bg) = B\,\mathbf{dom}\, g$). Since $Bd_k = B(z_k - \bar{z}) = s_k - \bar{s} \to 0$, we have

$$(Bg)(\bar{s}) = \inf_{d\in\mathbf{ker}\,B} g(\bar{z} + d) \leq \liminf_{k\to\infty} g(\bar{z} + d_k) = \liminf_{k\to\infty} g(z_k) \leq \liminf_{k\to\infty} (Bg)(s_k) + \tfrac{1}{k} \leq \alpha,$$

proving that $\bar{s} \in \mathbf{lev}_{\leq\alpha}(Bg)$. To show the converse implication, suppose that (5.5) does not hold. Thus, there exist $\bar{z} \in \mathbf{dom}\, g$ and $(d^k)_{k\in\mathbb{N}} \subset \mathbb{R}^n$ such that $Bd^k \to 0$ and $\|d^k\| \to \infty$ as $k \to \infty$, and such that, for some $\varepsilon > 0$,

$$g(\bar{z} + d^k) + \varepsilon \leq \inf_{d\in\mathbf{ker}\,B} g(\bar{z} + d) = (Bg)(B\bar{z}) \quad \forall k.$$

Then, $s_k := B(\bar{z} + d^k)$ satisfies $s_k \to B\bar{z}$ as $k \to \infty$, and

$$(Bg)(B\bar{z}) \geq \liminf_{k\to\infty} g(\bar{z} + d^k) + \varepsilon \geq \liminf_{k\to\infty} (Bg)(s^k) + \varepsilon,$$

hence (Bg) is not lsc at $B\bar{z}$. ∎

The *asymptotic function* $g_\infty(\bar{d}) := \mathbf{lim\,inf}_{d\to\bar{d},\, t\to\infty} \frac{g(td)}{t}$ is a tool used in [2] to analyze the behavior of g at infinity and derive sufficient properties ensuring lsc of (Bg). These all ensure that the set of minimizers $Z(s)$ as defined in Proposition 5.10 is nonempty, although

this property is not necessary as long as lower semicontinuity is concerned. To see this, it suffices to modify (5.4) as follows

$$g(x,y) = \begin{cases} -|x| & \text{if } |xy| \geq 1, \\ e^{-y^2} - q(|xy|)(e^{-y^2} + |x|) & \text{otherwise,} \end{cases}$$

that is, by replacing the constant value 1 on the y axis with e^{-y^2}. Then, $(Bg)(s) = -|s|$ is lsc, but the set of minimizers $\mathbf{arg\,min}_w \{g(w) \mid Bw = 0\} = \{0\} \times \mathbf{arg\,min}_y\, e^{-y^2}$ is empty at $s = 0$.

5.4.2. Smoothness of the image function. We now turn to the smoothness requirement of (Af). To this end, we introduce the following notion of *smoothness with respect to a matrix*.

Definition 5.12 (Smoothness relative to a matrix). *We say that $h : \mathbb{R}^n \to \mathbb{R}$ is* smooth relative to a matrix $C \in \mathbb{R}^{p\times n}$*, and we write $h \in C^{1,1}_C(\mathbb{R}^n)$, if h is differentiable and ∇h satisfies the following Lipschitz condition: there exist $L_{h,C}$ and $\sigma_{h,C}$ with $|\sigma_{h,C}| \leq L_{h,C}$ such that*

$$\sigma_{h,C}\|C(x-y)\|^2 \leq \langle \nabla h(x) - \nabla h(y), x - y\rangle \leq L_{h,C}\|C(x-y)\|^2 \tag{5.6}$$

whenever $\nabla h(x), \nabla h(y) \in \mathbf{range}\, C^\top$.

This condition is similar to that considered in [18], where $\mathbf{\Pi}_{\mathbf{range}\, A^\top}\nabla f$ is required to be Lipschitz. The paper analyzes convergence of a proximal ADMM; standard ADMM can be recovered when matrix A is invertible, in which case both conditions reduce to Lipschitz differentiability of f. In general, our condition applies to a smaller set of points only, as it can be verified with $f(x,y) = \frac{1}{2}x^2y^2$ and $A = [1\ 0]$. In fact, $\mathbf{\Pi}_{\mathbf{range}\, A^\top}\nabla f(x,y) = \binom{xy^2}{0}$ is *not* Lipschitz continuous; however, $\nabla f(x,y) \in \mathbf{range}\, A^\top$ iff $xy = 0$, in which case $\nabla f \equiv 0$. Then, f is smooth relative to A with $L_{f,A} = 0$.

To better understand how this notion of regularity comes into the picture, notice that if f is differentiable, then $\nabla f(x) \in \mathbf{range}\, A^\top$ on some domain $\mathcal{U}$ if there exists a differentiable function $q : A\mathcal{U} \to \mathbb{R}$ such that $f(x) = q(Ax)$. Then, it is easy to verify that f is smooth relative to A if the local "reparametrization" q is smooth (on its domain). From an a posteriori perspective, if (Af) is smooth, then due to the relation $A^\top\nabla(Af)(Az_s) = \nabla f(z_s)$ holding for $z_s \in \mathbf{arg\,min}_{z:Az=s} f(z)$ (Proposition 5.3), it is apparent that q serves as (Af). Therefore, smoothness relative to A is somewhat a minimal requirement ensuring smoothness of (Af).

Theorem 5.13 (Smoothness of (Af)). *Let $A \in \mathbb{R}^{p\times n}$ be surjective and $f : \mathbb{R}^n \to \mathbb{R}$ be lsc. Suppose that there exists $\beta \geq 0$ such that the function $f + \frac{\beta}{2}\|A\cdot - s\|^2$ is level bounded for all $s \in \mathbb{R}^p$. Then, the image function (Af) is smooth on $\mathbb{R}^p$, provided that either*

(i) $f \in C^{1,1}_A(\mathbb{R}^n)$, in which case $L_{(Af)} = L_{f,A}$ and $\sigma_{(Af)} = \sigma_{f,A}$,

(ii) or $f \in C^{1,1}(\mathbb{R}^n)$, and $X(s) := \mathbf{arg\,min}\{f(x) \mid Ax = s\}$ is single valued and Lipschitz continuous with modulus M, in which case

$$L_{(Af)} = L_f M^2 \quad \text{and} \quad \sigma_{(Af)} = \begin{cases} \sigma_f/\|A\|^2 & \text{if } \sigma_f \geq 0, \\ \sigma_f M^2 & \sigma_f < 0; \end{cases}$$

(iii) or $f \in C^{1,1}(\mathbb{R}^n)$ is convex, in which case $L_{(Af)} = \frac{L_f}{\sigma_+(A^\top A)}$ and $\sigma_{(Af)} = \sigma_f/\|A\|^2$.

Proof. As shown in Proposition 5.2*(i)*, (Af) is proper. Surjectivity of A and level boundedness ensure that for all $\alpha \in \mathbb{R}$ and $s \in \mathbb{R}^p$ the set $\{x \mid f(x) \leq \alpha,\ \|Ax - s\| < \varepsilon\}$ is bounded for some $\varepsilon > 0$ (in fact, for all $\varepsilon > 0$). Then, we may invoke [32, Thm. 1.32] to infer that (Af) is lsc, that the set $X(s) := \mathbf{arg\,min}_x \{f(x) \mid Ax = s\}$ is nonempty for all $s \in \mathbb{R}^p$ (owing to surjectivity of A and the fact that $\mathbf{dom}\, f = \mathbb{R}^p$), and that the function $H(x,s) := f(x) + \delta_{\{0\}}(Ax - s)$ is uniformly level bounded in x locally uniformly in s, in the sense of [32, Def. 1.16]. Moreover, since f is differentiable, observe that $\partial^\infty H(x, Ax) = \mathbf{range}\binom{A^\top}{1}$ for all $x \in \mathbb{R}^m$. Hence,

for all $s \in \mathbb{R}^p$ it holds that

$$\partial^\infty(Af)(s) \subseteq \bigcup_{x \in X(s)} \{y \mid (0,y) \in \partial^\infty H(x,s)\} = \mathbf{ker}\, A^\top = \{0\},$$

where the inclusion follows from [32, Thm. 10.13]. By virtue of [32, Thm. 9.13], we conclude that (Af) is strictly continuous and has nonempty subdifferential on $\mathbb{R}^p$. Fix $s_i \in \mathbb{R}^p$ and $y_i \in \partial(Af)(s_i)$, $i = 1,2$, and let us proceed by cases.

♠ 5.13*(i)* and 5.13*(ii)* It follows from Proposition 5.3 and continuous differentiability of f that $A^\top y_i \in \partial f(x_i) = \{\nabla f(x_i)\}$, for some $x_i \in X(s_i)$, $i = 1,2$. We have

$$\begin{aligned} \langle y_1 - y_2, s_1 - s_2\rangle &= \langle y_1 - y_2, Ax_1 - Ax_2\rangle = \langle A^\top y_1 - A^\top y_2, x_1 - x_2\rangle \\ &= \langle \nabla f(x_1) - \nabla f(x_2), x_1 - x_2\rangle. \end{aligned} \tag{5.7}$$

If 5.13*(i)* holds, since $\nabla f(x_i) = A^\top y_i \in \mathbf{range}\, A^\top$, $i = 1,2$, smoothness of f relative to A implies

$$\begin{aligned} \sigma_{f,A}\|s_1 - s_2\|^2 &= \sigma_{f,A}\|Ax_1 - Ax_2\|^2 \\ &\le \langle y_1 - y_2, s_1 - s_2\rangle \le L_{f,A}\|Ax_1 - Ax_2\|^2 = L_{f,A}\|s_1 - s_2\|^2 \end{aligned}$$

for all $s_i \in \mathbb{R}^p$ and $y_i \in \partial(Af)(s_i)$, $i = 1,2$. Otherwise, if 5.13*(ii)* holds, then

$$\sigma_f\|x_1 - x_2\|^2 \le \langle y_1 - y_2, s_1 - s_2\rangle \le L_f\|x_1 - x_2\|^2$$

and from the bound $\frac{1}{\|A\|}\|s_1 - s_2\| \le \|x_1 - x_2\| \le M\|s_1 - s_2\|$ we obtain

$$\sigma_{(Af)}\|s_1 - s_2\|^2 \le \langle y_1 - y_2, s_1 - s_2\rangle \le L_{(Af)}\|s_1 - s_2\|^2$$

with the constants $\sigma_{(Af)}$ and $L_{(Af)}$ as in the statement. The claimed smoothness and hypoconvexity then follow by invoking Lemma 2.1.

♠ 5.13*(iii)* It follows from [22, Thm. D.4.5.1 and Cor. D.4.5.2] that (Af) is convex and differentiable, and satisfies $\nabla(Af)(s) = y$, where for any $x \in X(s)$, y is such that $A^\top y = \nabla f(x)$. For $y_i = \nabla(Af)(s_i)$ and $x_i \in X(s_i)$, $i = 1,2$, the equalities in (5.7) hold. In turn,

$$\langle s_1 - s_2, y_1 - y_2\rangle \ge \tfrac{1}{L_f}\|A^\top(y_1 - y_2)\|^2 \ge \tfrac{\sigma_+(A^\top A)}{L_f}\|\mathbf{\Pi}_{\mathbf{range}\, A}(y_1 - y_2)\|^2 = \tfrac{\sigma_+(A^\top A)}{L_f}\|y_1 - y_2\|^2,$$

where the first inequality is due to $1/L_f$-cocoercivity of ∇f, see [29, Thm. 2.1.5], the second inequality is a known fact (see e.g., [18, Lem. A.2]), and the equality is due to the fact that A is surjective. We may again invoke [29, Thm. 2.1.5] to infer the claimed $\frac{L_f}{\sigma_+(A^\top A)}$-smoothness of (Af). Since (Af) is convex (thus 0-hypoconvex), if $\sigma_f = 0$ there is nothing more to show. The case $\sigma_f > 0$ follows from Proposition 5.4. ∎

Notice that the condition in Theorem 5.13*(ii)* covers the case when $f \in C^{1,1}(\mathbb{R}^n)$ and A has full column rank (hence is invertible), in which case $M = 1/\sigma_+(A)$. This is somehow trivial, since necessarily $(Af)(s) = f \circ A^{-1}$ in this case.

6. Conclusive remarks. This paper provides new convergence results for nonconvex Douglas-Rachford splitting (DRS) and ADMM with an all-inclusive analysis of all possible relaxation parameters $\lambda \in (0,4)$. Under the only assumption of Lipschitz differentiability of one function, convergence is shown for larger prox-stepsizes and relaxation parameters than was previously known. The results are tight when $\lambda \in (0,2]$, covering in particular classical (non-relaxed) DRS and PRS, or when the differentiable function is nonconvex. The necessity of $\lambda < 4$ and of a lower bound for the stepsize when $\lambda > 2$ is also shown.

Our theory is based on the Douglas-Rachford envelope (DRE), a continuous, real-valued, exact penalty function for DRS, and on a primal equivalence of DRS and ADMM that extends the well-known connection of the algorithms to arbitrary (nonconvex) problems. The DRE is shown to be a better Lyapunov function for DRS than the augmented Lagrangian, due to its closer connections with the cost function and with DRS iterations.

Acknowledgements. The authors thankfully acknowledge the superlative work carried out by the anonymous reviewers that led to this final version of the paper.

A. Proofs of Section 2.

Proof of Lemma 2.1 *(Subdifferential characterization of smoothness).* The claimed hypoconvexity follows from [32, Ex. 12.28]. It suffices to show that h is continuously differentiable, so that $\partial h = \nabla h$ and the claim then follows from (2.3). To this end, without loss of generality we may assume that $\sigma \geq 0$, since h is continuously differentiable iff so is $h - \frac{\sigma}{2}\|\cdot\|^2$. Thus, for all $x_i \in \mathbb{R}^n$, $v_i \in \partial h(x_i)$, $i = 1, 2$, one has

$$\begin{aligned} h(x_1) &\geq h(x_2) + \langle v_2, x_1 - x_2\rangle = h(x_2) + \langle v_2 - v_1, x_1 - x_2\rangle + \langle v_1, x_1 - x_2\rangle \\ &\geq h(x_2) - L\|x_1 - x_2\|^2 + \langle v_1, x_1 - x_2\rangle, \end{aligned}$$

where the first inequality follows from convexity of h (being it 0-hypoconvex by assumption). Rearranging,

$$h(x_2) \leq h(x_1) + \langle v_1, x_2 - x_1\rangle + L\|x_1 - x_2\|^2 \quad \forall x_i \in \mathbb{R}^n,\ v_1 \in \partial h(x_1),\ i = 1, 2.$$

Let $\tilde{h} := h - \langle v_1, \cdot\rangle$, so that $0 \in \partial h(x_1)$. Due to convexity, $x_1 \in \mathbf{arg\,min}\, \tilde{h}$, hence for all $w \in \mathbb{R}^n$ and $v_1' \in \partial h(x_1)$ one has

$$\tilde{h}(x_1) \leq \tilde{h}(w) \leq h(x_1) + \langle v_1', w - x_1\rangle + L\|w - x_1\|^2 - \langle v_1, w\rangle = \tilde{h}(x_1) + \langle v_1' - v_1, w - x_1\rangle + L\|w - x_1\|^2.$$

By selecting $w = x_1 - \frac{1}{2L}(v_1' - v_1)$, one obtains $\|v_1 - v_1'\|^2 \leq 0$, hence necessarily $v_1 = v_1'$. From the arbitrarity of $x_1 \in \mathbb{R}^n$ and $v_1, v_1' \in \partial h(x_1)$ it follows that ∂h is everywhere single valued, and the sought continuous differentiability of h then follows from [32, Cor. 9.19]. □

Proof of Theorem 2.2 *(Lower bounds for smooth functions).*

♠ 2.2*(i)* This is the lower bound in (2.2).

♠ 2.2*(ii)* Let $L \geq L_h$ and $\sigma \in (-L, \mathbf{min}\{0, \sigma_h\}]$ be fixed. Then, h is L-smooth and σ-hypoconvex, and from [29, Thm. 2.1.12] we obtain that

$$\langle \nabla h(y) - \nabla h(x), y - x\rangle \geq \tfrac{\sigma L}{L+\sigma}\|x - y\|^2 + \tfrac{1}{L+\sigma}\|\nabla h(x) - \nabla h(y)\|^2 \tag{A.1}$$

for all $x, y \in \mathbb{R}^n$. (Although [29, Thm. 2.1.12] assumes $\sigma > 0$, the given proof does not necessitate this restriction). Moreover, $\psi := h - \frac{\sigma}{2}\|\cdot\|^2$ is convex and L_ψ-smooth, with $L_\psi = L - \sigma$. Consequently, for all $x, y \in \mathbb{R}^n$ one has $\psi(y) \geq \psi(x) + \langle \nabla\psi(x), y - x\rangle + \frac{1}{2L_\psi}\|\nabla\psi(y) - \nabla\psi(x)\|^2$, see [29, Thm. 2.1.5], resulting in

$$\begin{aligned} h(y) \geq h(x) &+ \langle \nabla h(x), y - x\rangle + \tfrac{\sigma L}{2(L-\sigma)}\|y - x\|^2 + \tfrac{1}{2(L-\sigma)}\|\nabla h(y) - \nabla h(x)\|^2 \\ &- \tfrac{\sigma}{L-\sigma}\langle \nabla h(y) - \nabla h(x), y - x\rangle. \end{aligned}$$

Since $\sigma \leq 0$, the coefficient of the scalar product in the second line is positive. We may thus invoke the inequality (A.1) to arrive to

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

$$- \tfrac{\sigma}{L-\sigma}\Big[\tfrac{\sigma L}{L+\sigma}\|x-y\|^2 + \tfrac{1}{L+\sigma}\|\nabla h(x) - \nabla h(y)\|^2\Big]$$
$$= h(x) + \langle \nabla h(x), y - x\rangle + \tfrac{\sigma L}{2(L+\sigma)}\|y-x\|^2 + \tfrac{1}{2(L+\sigma)}\|\nabla h(y) - \nabla h(x)\|^2,$$

hence the claimed inequality. □

Proof of Proposition 2.3 *(Proximal properties of smooth functions).* Fix $\gamma \in (0, 1/[\sigma_h]_-)$, and let $\psi := \gamma h + \frac{1}{2}\|\cdot\|^2$. Observe that $\psi \in \mathrm{C}^{1,1}(\mathbb{R}^n)$ is L_ψ-smooth and σ_ψ-strongly convex, with $L_\psi = 1 + \gamma L_h$ and $\sigma_\psi = 1 + \gamma\sigma_h$. In particular, due to strong convexity $\mathbf{inf}\,\psi > -\infty$, and by definition of prox-boundedness it then follows that $\gamma_h \geq 1/[\sigma_h]_-$.

♠ 2.3*(i)* Follows from (2.7), by observing that $h + \frac{1}{2\gamma}\|\cdot - s\|^2$ is strongly convex, hence that a minimizer is characterized by stationarity.

♠ 2.3*(ii)* For $s, s' \in \mathbb{R}^n$, let $u = \mathbf{prox}_{\gamma h}(s)$ and $u' = \mathbf{prox}_{\gamma h}(s')$. Then,

$$\langle s - s', u - u'\rangle = \langle \nabla\psi(u) - \nabla\psi(u'), u - u'\rangle \geq \sigma_\psi\|u - u'\|^2 = (1 + \gamma\sigma_h)\|u - u'\|^2,$$

where the first equality was shown in 2.3*(i)* and the inequality follows from (2.3). By using the $\frac{1}{L_\psi}$-cocoercivity of $\nabla\psi$ [29, Thm. 2.1.10], also the claimed strong monotonicity follows. In turn, the Cauchy-Schwartz inequality on the inner product yields (2.8).

♠ 2.3*(iii)* From [32, Ex. 10.32] it follows that h^γ is strictly continuous and that $\partial h^\gamma(s) \subseteq \frac{1}{\gamma}(s - \mathbf{prox}_{\gamma h}(s))$. Because of single valuedness of $\mathbf{prox}_{\gamma h}$, by invoking [32, Thm. 9.18] we conclude that h^γ is everywhere differentiable with $\nabla h^\gamma(s) = \frac{1}{\gamma}(s - \mathbf{prox}_{\gamma h}(s))$. Thus,

$$\langle \nabla h^\gamma(s) - \nabla h^\gamma(s'), s - s'\rangle = \tfrac{1}{\gamma}\Big(\|s - s'\|^2 - \langle s - s', u - u'\rangle\Big),$$

and from the bounds in 2.3*(ii)* we conclude that

$$\tfrac{\sigma_h}{1+\gamma\sigma_h}\|s - s'\|^2 \leq \langle \nabla h^\gamma(s) - \nabla h^\gamma(s'), s - s'\rangle \leq \tfrac{L_h}{1+\gamma L_h}\|s - s'\|^2.$$

The claimed smoothness and hypoconvexity follow from the characterization of (2.3). □

B. Proofs of Section 5.

Proof of Proposition 5.2.

♠ 5.2*(i)* If $\bar{s} \notin C\,\mathbf{dom}\,h$, then $(Ch)(\bar{s}) = \infty$. If instead $\bar{s} = C\bar{x}$ for some $\bar{x} \in \mathbf{dom}\,h$, then

$$-\infty < \min_x \Big\{h(x) + \tfrac{\beta}{2}\|Cx - \bar{s}\|^2\Big\} \leq \inf_{x:\,Cx=\bar{s}} \Big\{h(x) + \tfrac{\beta}{2}\|Cx - \bar{s}\|^2\Big\} \overset{(def)}{=} (Ch)(\bar{s}),$$

which is upper bounded by the finite quantity $h(\bar{x})$.

♠ 5.2*(ii)* Since $C(x_\beta + v) = Cx_\beta$ iff $v \in \mathbf{ker}\,C$, for all $s \in \mathbb{R}^p$ and $x_\beta \in X_\beta(s)$ necessarily $h(x_\beta) \leq h(x_\beta + v)$. Consequently,

$$(Ch)(Cx_\beta) \leq h(x_\beta) \leq \inf_{v\in\mathbf{ker}\,C} h(x_\beta + v) = \inf_{x:\,Cx=Cx_\beta} h(x) = (Ch)(Cx_\beta).$$

♠ 5.2*(iii)* For any $\bar{s} \in \mathbb{R}^p$ one has

$$\mathbf{prox}_{\gamma(Ch)}(\bar{s}) = \mathbf{arg\,min}_{w\in\mathbb{R}^p}\Big\{(Ch)(w) + \tfrac{1}{2\gamma}\|w - \bar{s}\|^2\Big\} = \mathbf{arg\,min}_{w\in\mathbb{R}^p}\Big\{\inf_{x\in\mathbb{R}^n} f(x) + \tfrac{1}{2\gamma}\|w - \bar{s}\|^2 \mid Cx = w\Big\}$$
$$= C\,\mathbf{arg\,min}_{x\in\mathbb{R}^n}\Big\{f(x) + \tfrac{1}{2\gamma}\|Cx - \bar{s}\|^2\Big\} = CX(\bar{s}),$$

where the third equality uses the change of variable $w = Cx$. □

Proof of Proposition 5.3. Let $\bar{v} \in \hat{\partial}(Ch)(C\bar{x})$. Then,

$$\begin{aligned}\liminf_{\substack{x\to\bar{x}\\ x\neq\bar{x}}} \frac{h(x)-h(\bar{x})-\langle C^\top\bar{v}, x-\bar{x}\rangle}{\|x-\bar{x}\|} &= \liminf_{\substack{x\to\bar{x}\\ x\neq\bar{x}}} \frac{h(x)-(Ch)(C\bar{x})-\langle \bar{v}, C(x-\bar{x})\rangle}{\|x-\bar{x}\|}\\ &\geq \liminf_{\substack{x\to\bar{x}\\ x\neq\bar{x}}} \frac{(Ch)(Cx)-(Ch)(C\bar{x})-\langle \bar{v}, C(x-\bar{x})\rangle}{\|x-\bar{x}\|}\\ &\geq \liminf_{\substack{x\to\bar{x}\\ x\neq\bar{x}}} \frac{o(\|C(x-\bar{x})\|)}{\|x-\bar{x}\|} = 0,\end{aligned}$$

where the last inequality follows from the fact that $\bar{v} \in \hat{\partial}(Ch)(C\bar{x})$. □

Proof of Proposition 5.4 *(Strong convexity of the image function).* That (Ch) is convex follows from [4, Prop. 12.36(ii)]. Due to strong convexity of h, for every $s \in C\,\mathbf{dom}\,h = \mathbf{dom}(Ch)$ there exists a unique $x_s \in \mathbb{R}^n$ such that $Cx_s = s$ and $(Ch)(s) = h(x_s)$. Let $v_s \in \partial(Ch)(s)$. Then, Proposition 5.3 ensures that $C^\top v_s \in \partial h(x_s)$, hence, for all $s' \in \mathbf{dom}(Ch)$

$$h(x_{s'}) \geq h(x_s) + \langle C^\top v_s, x_{s'} - x_s\rangle + \tfrac{\sigma_h}{2}\|x_{s'} - x_s\|^2 \geq h(x_s) + \langle v_s, s' - s\rangle + \tfrac{\sigma_h}{2\|C\|^2}\|s' - s\|^2.$$

Strong convexity then follows by observing that $h(x_s) = (Ch)(s)$ and $h(x_{s'}) = (Ch)(s')$. □

REFERENCES

[1] H. Attouch, J. Bolte, and B. F. Svaiter, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods*, Mathematical Programming, 137 (2013), pp. 91–129.

[2] A. Auslender and M. Teboulle, *Asymptotic Cones and Functions in Optimization and Variational Inequalities*, Springer Monographs in Mathematics, Springer New York, 2002.

[3] H. Bauschke and D. Noll, *On the local convergence of the Douglas-Rachford algorithm*, Archiv der Mathematik, 102 (2014), pp. 589–600.

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

[5] H. H. Bauschke and V. R. Koch, *Projection methods: Swiss army knives for solving feasibility and best approximation problems with halfspaces*, in Infinite Products of Operators and Their Applications, S. Reich and A. J. Zaslavski, eds., vol. 636, American Mathematical Society, 2015, pp. 1–40.

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

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

[8] J. Bochnak, M. Coste, and M.-F. Roy, *Real Algebraic Geometry*, A Series of Modern Surveys in Mathematics, Springer Berlin Heidelberg, 2013.

[9] J. Bolte, S. Sabach, and M. Teboulle, *Proximal Alternating Linearized Minimization for nonconvex and nonsmooth problems*, Mathematical Programming, 146 (2014), pp. 459–494.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning, 3 (2011), pp. 1–122.

[11] J. Douglas and H. H. Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Transactions of the American Mathematical Society, 82 (1956), pp. 421–439.

[12] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Mathematical Programming, 55 (1992), pp. 293–318.

[13] D. Gabay, *Chapter IX applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems, M. Fortin and R. Glowinski, eds., vol. 15 of Studies in Mathematics and Its Applications, Elsevier, 1983, pp. 299–331.

[14] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite element approximation*, Computers & Mathematics with Applications, 2 (1976), pp. 17–40.

[15] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Scientific Computation, Springer, Berlin Heidelberg, 2013.

[16] R. Glowinski, *On alternating direction methods of multipliers: A historical perspective*, in Modeling, Simulation and Optimization for Science and Technology, W. Fitzgibbon, Y. A. Kuznetsov, P. Neittaanmäki, and O. Pironneau, eds., Springer Netherlands, Dordrecht, 2014, pp. 59–82.

[17] R. Glowinski and A. Marrocco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires*, ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique, 9 (1975), pp. 41–76.

[18] M. L. N. Goncalves, J. G. Melo, and R. D. C. Monteiro, *Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems*, ArXiv e-prints, (2017), https://arxiv.org/abs/1702.01850.

[19] K. Guo, D. Han, and T.-T. Wu, *Convergence of alternating direction method for minimizing sum of two nonconvex functions with linear constraints*, International Journal of Computer Mathematics, 94 (2017), pp. 1653–1669.

[20] R. Hesse and R. Luke, *Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems*, SIAM Journal on Optimization, 23 (2013), pp. 2397–2419.

[21] R. Hesse, R. Luke, and P. Neumann, *Alternating projections and Douglas-Rachford for sparse affine feasibility*, IEEE Transactions on Signal Processing, 62 (2014), pp. 4868–4881.

[22] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Grundlehren Text Editions, Springer Berlin Heidelberg, 2012.

[23] M. Hong, Z.-Q. Luo, and M. Razaviyayn, *Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems*, SIAM Journal on Optimization, 26 (2016), pp. 337–364.

[24] G. Li, T. Liu, and T. K. Pong, *Peaceman–Rachford splitting for a class of nonconvex optimization problems*, Computational Optimization and Applications, 68 (2017), pp. 407–436.

[25] G. Li and T. K. Pong, *Global convergence of splitting methods for nonconvex composite optimization*, SIAM Journal on Optimization, 25 (2015), pp. 2434–2460.

[26] G. Li and T. K. Pong, *Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems*, Mathematical Programming, 159 (2016), pp. 371–401.

[27] T. Liu and T. K. Pong, *Further properties of the forward-backward envelope with applications to difference-of-convex programming*, Computational Optimization and Applications, 67 (2017), pp. 489–520.

[28] R. D. C. Monteiro and C.-K. Sim, *Complexity of the relaxed Peaceman-Rachford splitting method for the sum of two maximal strongly monotone operators*, Computational Optimization and Applications, 70 (2018), pp. 763–790.

[29] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer, 2003.

[30] P. Patrinos and A. Bemporad, *Proximal Newton methods for convex composite optimization*, in 52nd IEEE Conference on Decision and Control, 2013, pp. 2358–2363.

[31] P. Patrinos, L. Stella, and A. Bemporad, *Douglas-Rachford splitting: Complexity estimates and accelerated variants*, in 53rd IEEE Conference on Decision and Control, Dec 2014, pp. 4234–4239.

[32] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*, vol. 317, Springer, 2011.

[33] L. Stella, A. Themelis, and P. Patrinos, *Forward-backward quasi-Newton methods for nonsmooth optimization problems*, Computational Optimization and Applications, 67 (2017), pp. 443–487.

[34] A. Themelis, L. Stella, and P. Patrinos, *Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms*, SIAM Journal on Optimization, 28 (2018), pp. 2274–2303.

[35] Y. Wang, W. Yin, and J. Zeng, *Global convergence of ADMM in nonconvex nonsmooth optimization*, Journal of Scientific Computing, 78 (2019), pp. 29–63.

[36] M. Yan and W. Yin, *Self equivalence of the alternating direction method of multipliers*, in Splitting Methods in Communication, Imaging, Science, and Engineering, R. Glowinski, S. J. Osher, and W. Yin, eds., Springer International Publishing, Cham, 2016, pp. 165–194.