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

ANDREAS THEMELIS AND PANOS 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 by the authors 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-stepizes and over-relaxation parameters than previously known. In fact, our bounds are tight whenever the over-relaxation parameter ranges in $(0,2]$. Moreover, a novel primal equivalence of ADMM and DRS extends to any problem the known duality of the algorithms holding in the convex case.

1. Introduction

First introduced in [8] for finding numerical solutions of heat differential equations, the Douglas-Rachford splitting (DRS) is now considered 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:

$$
\minimize_{s \in \mathbb{R}^p} \varphi(s) \equiv \varphi_1(s) + \varphi_2(s). \quad (1.1)
$$

Here, $\varphi_1, \varphi_2 : \mathbb{R}^p \to \mathbb{R}$ 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 iteration of DRS applied to (1.1) with stepsize $\gamma > 0$ and relaxation parameter $\lambda > 0$ amounts to

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

Department of Electrical Engineering (ESAT-STADIUS) & Optimization in Engineering Center (OPTEC) – KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
IMT School for Advanced Studies Lucca – Piazza S. Francesco 19, 55100 Lucca, Italy
E-mail addresses: andreas.themelis@esat.kuleuven.be panos.patrinos@esat.kuleuven.be

Date: May 24, 2018.
1991 Mathematics Subject Classification. 90C06, 90C25, 90C26, 49J52, 49J53.
Key words and phrases. Nonsmooth optimization, nonconvex optimization, Douglas-Rachford splitting, Peaceman-Rachford splitting, ADMM.
Contributions.

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 4.8). Figures 1a and 1b highlight the extent of the improvement with respect to the state of the art.

1.1. Equivalence gap for nonconvex problems. 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. In order to compensate the lack of convexity some additional assumptions are in order, and all results in the literature seem to agree that a sufficient condition for ensuring convergence of DRS is Lipschitz-continuous differentiability of function \( \varphi_1 \). Due to the lack of a strong duality theory, this requirement could not be directly translated into equivalent conditions for ADMM applied to nonconvex problems, and indeed the literature only offers standalone results, possibly involving implicit constants and burdened with non-trivial assumptions.

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

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 DRS 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 \( \varphi_1 \) and \( \varphi_2 \) are convex functions, the condition is also sufficient and DRS iterations are known to converge for any \( \gamma > 0 \) and \( \lambda \in (0, 2) \), in the sense that the residual vanishes.

Closely related to DRS and possibly even more popular is the alternating direction method of multipliers (ADMM), first appeared in [12, 10], see also [11] for a recent historical overview. ADMM addresses linearly constrained optimization problems

\[
\minimize_{(x, z) \in \mathbb{R}^m \times \mathbb{R}^n} f(x) + g(z) \quad \text{subject to} \quad Ax + Bz = b \tag{1.2}
\]

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

\[
\begin{align*}
x^+ & \in \arg\min_{x, z, y} \mathcal{L}_\beta(x, z, y) \\
y^{1/2} & = y - \beta(1 - \lambda)(Ax^+ + Bz - b) \\
z^+ & \in \arg\min_{x^+, y^{1/2}} \mathcal{L}_\beta(x^+, z^+, y^{1/2}) \\
y^+ & = y^{1/2} + \beta(Ax^+ + Bz^+ - b).
\end{align*}
\tag{ADMM}
\]

Here, \( \beta > 0 \) is a penalty parameter, \( \lambda > 0 \) is a possible relaxation parameter, and

\[
\mathcal{L}_\beta(x, z, y) := f(x) + g(z) + (y, Ax + Bz - b) + \frac{\beta}{2} \|Ax + Bz - b\|^2 \tag{1.3}
\]

is the \( \beta \)-augmented Lagrangian of (1.2) with \( y \in \mathbb{R}^p \) as Lagrange equality multiplier.\(^1\) It is well known that for convex problems ADMM is simply DRS applied to a dual formulation [9], and its convergence properties for \( \lambda = 1 \) and arbitrary penalty parameters \( \beta > 0 \) are well documented in the literature, see e.g., [7].

\(^1\)The half-update \( y^{1/2} \) is introduced only for simplification purposes; in fact, replacing \( y^{1/2} \) with \( y \) and all occurrences of \( Ax^+ \) in the \( z \)- and \( y \)-update with \( \lambda Ax^+ - (1 - \lambda)(Bz - b) \) results in the same update. However, it is often the case that \( \lambda \) is set to 1, in which case \( y^{1/2} = y \).
2) **Novel primal equivalence of DRS and ADMM.** We prove the equivalence of DRS and ADMM for arbitrary problems, so extending the well known duality of the algorithms holding in the convex case.

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 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 [25] 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 the stepsize $\gamma$ and the penalty parameter $\beta$ are adaptively tuned, so that no prior knowledge of quantities such as Lipschitz moduli is needed.

1.3. **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.

1.3.1. **ADMM.** The primal equivalence of ADMM and DRS exploits a reformulation of problem (1.2) into the DRS-form (1.1) which, up to a sign switch, was used in [30] to show self-dual symmetry of ADMM in the convex case.

In [29] convergence of ADMM is studied for problems of the form

$$\min_{x=(x_0 \ldots x_p), z} g(x) + \sum_{i=0}^{p} f_i(x_i) + h(z) \quad \text{subject to} \quad Ax + Bz = 0.$$  

Despite 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) and uniform boundedness of the subgradient of the nonsmooth term, whereas we allow for multiple minimizers and make almost no requirement on the nonsmooth function. For instance, the requirements rule out interesting cases involving discrete variables or rank constraints. Moreover, the analysis is limited to showing convergence for ‘large enough’ penalty parameters, and the given bounds involve implicit constants that are not readily available.

In [18] 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 in ADMM, which agree with ours when the smooth function is convex but are more restrictive by a factor of $\sqrt{2}$ otherwise (cf. Fig. 1c for a more detailed comparison). The main limitation of the proposed approach, however, is that the theory only allows for functions either convex or smooth, differently from ours where the nonsmooth term can basically be anything. Once again, many interesting applications are not covered.

The work [20] 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
applied to projection onto set ϕ. This is done by applying DRS to the minimization of the sum in the intersection of nonempty closed sets at best. These mainly focus on feasibility problems, where the goal is to find points Douglas-Rachford splitting.

1.3.2. Douglas-Rachford splitting. Few exceptions apart [21, 19], 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 ϕ1 = δA and ϕ2 = δB, where δC is the indicator function of a set C (see §2.1). The minimization subproblems in DRS then reduce to (set-valued) projections onto either set, regardless of the stepsize parameter γ > 0. This is the case of [2], for instance, 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 [16, 15].

Although this particular application of DRS does not comply with our requirements, as ϕ1 fails to be Lipschitz-differentiable, however replacing δA with ϕ1 = 1 2 dist2A 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 ΠA, the projection onto set A, with a “relaxed” version ΠA,t := (1 − t)Id + tΠA for some t ∈ (0, 1). Then, it can be easily verified that for any α, β ∈ (0, +∞] one DRS-step applied to
\[
\min_{s \in \mathbb{R}^n} \frac{α}{2} \text{dist}_A^2(s) + \frac{β}{2} \text{dist}_B^2(s)
\] results in
\[
s^+ ∈ (1 − γ/2)s + γ/2 Π_{B,q} Π_{A,p} s
\]
for p = 2αγ 1+αγ and q = 2βγ 1+βγ. Notice that (1.5) is the γ/2-relaxation of the “method of alternating (p,q)-relaxed projections” ((p,q)-MARP) [4]. The (non-relaxed) (p,q)-MARP is recovered by setting λ = 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 1 2 dist2A is convex and α-Lipschitz differentiable; our theory then ensures convergence of the fixed-point residual and subsequent convergence of the iterations (1.5) for any λ ∈ (0, 2), p ∈ (0, 1) and q ∈ (0, 1], without any requirements on the (nonempty closed) set B. Here, q = 1 is obtained by replacing 1 2 dist2B with δB, which can be interpreted as the hard penalization obtained by letting β = ∞. Although the non-relaxed MARP is not covered due to the non-strong convexity of dist2A, however λ can be set arbitrarily close to 2.

The work [21] presents the first general analysis of global convergence of (non-relaxed) DRS for fully nonconvex problems where one function is Lipschitz differentiable. In [19] PRS is also considered under the additional requirement that the smooth function is strongly convex with strong-convexity/Lipschitz moduli ratio of at least 1 2α. Both papers show that for sufficiently small stepsizes one iteration of DRS or PRS yields a sufficient decrease on an augmented Lagrangian. However, due to the lower unboundedness of the augmented Lagrangian they cannot infer the very convergence of the algorithms, namely that the fixed-point residual vanishes and that therefore with finitely many iterations a stopping criterion is satisfied.
Figure 1. Maximum stepsize $\gamma$ ensuring convergence of DRS (Fig. 1a) and PRS (Fig. 1b), and maximum inverse of the penalty parameter $1/\beta$ in ADMM (Fig. 1c); comparison between our bounds (blue plot) and [21] for DRS, [19] for PRS and [13, 14, 18, 20] for ADMM; [29] is not considered due to the unknown range of parameters. On the $x$-axis the ratio between (possibly negative) strong convexity parameter $\sigma$ and the Lipschitz modulus $L$ of the gradient of the smooth function. On the $y$-axis, the supremum of stepsize $\gamma$ such that the algorithms converge. For ADMM, the analysis is made for a common framework: 2-block ADMM with no Bregman or proximal terms, $A$ full rank and $B$ identity; $L$ and $\sigma$ are relative to the transformed problem. Notice that, due to the proved analogy of DRS and ADMM, our theoretical bounds coincide in Fig. 1a and 1c.

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.

1.4. Organization of the paper. The paper is organized as follows. Section 2 introduces some notation and offers a brief recap of the needed theory. 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. For the sake of readability, some proofs and auxiliary results are deferred to the Appendix.

2. Background

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

The identity matrix of suitable size is denoted as $I$, and for a nonzero matrix $M \in \mathbb{R}^{p \times n}$ we let $\sigma_+(M)$ denote its smallest nonzero singular value. The linear subspaces parallel and orthogonal to an affine subspace $V$ are denoted as $V^\perp$ and $V^{\perp}$. For a set $E$ and a sequence $(x^k)_{k \in \mathbb{N}}$ we write $(x^k)_{k \in \mathbb{N}} \subset E$ with the obvious meaning of $x^k \in E$ 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.
The indicator function of a set $S \subseteq \mathbb{R}^n$ is denoted as $\delta_S : \mathbb{R}^n \to \mathbb{R}$, namely $\delta_S(x) = 0$ if $x \in S$ and $\delta_S(x) = \infty$ otherwise, and $\Pi_S : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ is the (set-valued) projection $x \mapsto \text{argmin}_{z \in S} \|z - x\|$. A function $h : \mathbb{R}^n \to \mathbb{R}$ is proper if $h > -\infty$ and $h \neq \infty$, in which case its domain is defined as the set $\text{dom} h := \{ x \in \mathbb{R}^n \mid h(x) < \infty \}$. For $a \in \mathbb{R}$, $\text{lev}_{\leq a} h$ is the $a$-level set of $h$, i.e., $\text{lev}_{\leq a} h := \{ x \in \mathbb{R}^n \mid h(x) \leq a \}$. We say that $h$ is level bounded if $\text{lev}_{\leq a} h$ is bounded for all $a \in \mathbb{R}$. A vector $v \in \partial h(x)$ is a subgradient of $h$ at $x$, where the subdifferential $\partial h(x)$ is considered in the sense of [26, Def. 8.3]

$$\partial h(x) = \left\{ v \in \mathbb{R}^n \mid \exists (x^k, v^k)_{k \in \mathbb{N}} \text{ s.t. } x^k \to x, \ h(x^k) \to h(x), \ \partial h(x^k) \ni v^k \to v \right\},$$

and $\hat{\partial} h(x)$ is the set of regular subgradients of $h$ at $x$, namely

$$\hat{\partial} h(x) = \left\{ v \in \mathbb{R}^n \mid h(z) \geq h(x) + \langle v, z - x \rangle + o(\|z - x\|), \ \forall z \in \mathbb{R}^n \right\}.$$  

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. For a function $h : \mathbb{R}^n \to \mathbb{R}$ we say that $h \in C^1(\text{dom} h)$ if the restriction $h|_{\text{dom} h} : \text{dom} h \to \mathbb{R}$ is continuously differentiable, and we define

$$\nabla h := \nabla (h|_{\text{dom} h}).$$

In case $\text{dom} h$ is an affine subspace of $\mathbb{R}^n$, then for all $u \in \text{dom} h$

$$\partial h(u) = \nabla h(u) + \text{dom} h^\perp \quad (2.1)$$

and in particular

$$\nabla h(u) = \Pi_{\text{dom} h^\perp} \partial h(u) = \partial h(u) \cap \text{dom} h^\perp \quad (2.2)$$

see e.g., [1, Prop. 3.2.3]. Alternatively, $\nabla h(x)$ can be defined as the unique vector in $\text{dom} h^\perp$ such that $h(y) = h(x) + \langle \nabla h(x), y - x \rangle + o(\|y - x\|)$ for all $y \in \text{dom} h$.

Definition 2.1 (Smoothness). We say that $h : \mathbb{R}^n \to \mathbb{R}$ is smooth on its domain, and we write $h \in C^{1,1}(\text{dom} h)$, if $h \in C^1(\text{dom} h)$ and $\nabla h := \nabla (h|_{\text{dom} h})$ is Lipschitz continuous, in which case the Lipschitz modulus is denoted $L_h$.

Remark 2.2. If $h$ is smooth on its domain, then there exists $\sigma_h \in [-L_h, L_h]$ satisfying any of the following equivalent properties:

(a) $h - \frac{\sigma_h}{2} \cdot \| \cdot \|^2$ is convex;

(b) $\sigma_h \| u - v \|^2 \leq \langle \nabla h(u) - \nabla h(v), u - v \rangle \leq L_h \| u - v \|^2$ for all $u, v \in \text{dom} h$;

(c) $\frac{\sigma_h}{2} \| u - v \|^2 \leq h(v) - h(u) - \langle \nabla h(u), v - u \rangle \leq \frac{L_h}{2} \| u - v \|^2$ for all $u, v \in \text{dom} h$.

This equivalence is readily proven by applying [23, Thm. 2.1.5] to the function $\psi|_{\text{dom} h}$, where $\psi := h - \frac{\sigma_h}{2} \cdot \| \cdot \|^2$. In particular, 2.2(a) emphasizes that the largest $\sigma_h$ is, the closer $h$ to a convex function: if $\sigma_h \geq 0$ then $h$ is convex, and if $\sigma_h > 0$ then $h$ is strongly convex, in which case $\sigma_h$ is its modulus of strong convexity.

For notational convenience, we denote the convexity-smoothness ratio of a smooth function $h$ as $p_h := \sigma_h/L_h \in [-1, 1]$.

2.3. Proximal mapping. The proximal mapping of $h : \mathbb{R}^n \to \mathbb{R}$ with parameter $\gamma > 0$ is the (possibly set-valued) map $\text{prox}_{\gamma h} : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ defined as

$$\text{prox}_{\gamma h}(x) := \text{argmin}_{w \in \mathbb{R}^n} \left\{ h(w) + \frac{1}{2\gamma} \| w - x \|^2 \right\}. \quad (2.3)$$

We say that a function $h$ is prox-bounded if $h + \frac{1}{2\gamma} \cdot \| \cdot \|^2$ is lower bounded for some $\gamma > 0$. The supremum of all such $\gamma$ is the threshold of prox-boundedness of $h$, denoted as $\gamma_h$. If $h$ is lsc, then $\text{prox}_{\gamma h}$ is nonempty- and compact-valued over $\mathbb{R}^n$ for all $\gamma \in (0, \gamma_h)$ [26, Thm. 1.25]. Consequently, the value function of the
minimization problem defining the proximal mapping, namely the Moreau envelope with parameter $\gamma$, denoted $h^\gamma : \mathbb{R}^n \to \mathbb{R}$ and defined as
\[
    h^\gamma(x) := \inf_{w \in \mathbb{R}^n} \left\{ h(w) + \frac{1}{2\gamma} \|w - x\|^2 \right\},
\]
is everywhere finite and, in fact, strictly continuous [26, Ex. 10.32].

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)). $\varphi_1, \varphi_2 : \mathbb{R}^p \to \mathbb{R}$ are proper and lsc functions such that

(i) (Smoothness) $\text{dom} \varphi_1$ is an affine set and $\varphi_1 \in \mathcal{C}^{1,1}(\text{dom} \varphi_1)$;

(ii) (DRS-feasibility) $\varphi_2$ is prox-bounded and $\gamma < \min \{\gamma_{\varphi_2}, 1/|\sigma_{\varphi_1}|\}$;

(iii) (Domain inclusion) $\text{dom} \varphi_2 \subseteq \text{dom} \varphi_1$.

We will say that a stepsize $\gamma$ is feasible if it complies with Assumption I(ii).

The bound $\gamma < \gamma_{\varphi_2}$ is equivalent to imposing that $\gamma$ is small enough such that the subproblem defining $v$ in (DRS) admits a (not necessarily unique) solution. Further constraining $\gamma < 1/|\sigma_{\varphi_1}|$ ensures that also the subproblem defining $u$ admits a solution; in fact, $u$ is unique and depends Lipschitz continuously on $s$, and for all $s \in \mathbb{R}^p$
\[
    \nabla \varphi_1(u) = \frac{1}{\gamma}(\Pi_{\text{dom} \varphi_1}(s) - u)
\]
where $u = \text{prox}_{\gamma \varphi_1}(s)$ (cf. Prop. A.3). Although the threshold $\gamma_{\varphi_1}$ might be larger than $1/|\sigma_{\varphi_1}|$, because of these favorable properties we constrain it as such. From (3.1) we can easily verify that the $v$-update in (DRS) is equivalent to selecting
\[
    v \in \arg\min_{w \in \mathbb{R}^p} \left\{ \varphi_2(w) + \varphi_1(u) + \langle \nabla \varphi_1(u), w - u \rangle + \frac{1}{2\gamma} \|w - u\|^2 \right\}.
\]
This shows that $v$ is the result of a minimization of a majorization model for the original function $\varphi = \varphi_1 + \varphi_2$, where the smooth function $\varphi_2$ is replaced by the quadratic upper bound emphasized by the under-bracket in (3.2).

Closely related to (3.2) and first introduced in [25] for convex problems, the Douglas-Rachford envelope (DRE) is the function $\varphi_{\gamma}^{DR} : \text{dom} \varphi_1 \to \mathbb{R}$ defined as
\[
    \varphi_{\gamma}^{DR}(s) := \min_{w \in \mathbb{R}^p} \left\{ \varphi_2(w) + \varphi_1(u) + \langle \nabla \varphi_1(u), w - u \rangle + \frac{1}{2\gamma} \|w - u\|^2 \right\}.
\]
where $u = \text{prox}_{\gamma \varphi_1}(s)$. Namely, rather than the minimizer $v$, $\varphi_{\gamma}^{DR}(s)$ is the value of the minimization problem (3.2) defining the $v$-update in (DRS). By plugging the minimizer $w = v$ in (3.3) we obtain the following useful interpretation of the DRE:
\[
    \varphi_{\gamma}^{DR}(s) = \mathcal{L}_{\gamma,\beta}(u, v, \gamma^{-1}(u - s))
\]
where $u$ and $v$ come from the DRS iteration and
\[
    \mathcal{L}_\beta(x, z, y) := \varphi_1(x) + \varphi_2(z) + \langle y, x - z \rangle + \frac{\beta}{2} \|x - z\|^2
\]
is the $\beta$-augmented Lagrangian relative to the equivalent problem formulation
\[
    \text{minimize}_{x,z} \varphi_1(x) + \varphi_2(z) \quad \text{subject to} \quad x - z = 0.
\]
Therefore, computing $\varphi_{\gamma}^{DR}(s)$ requires the same operations as performing one DRS update $s \mapsto (s, u, v)$. Moreover, by expressing $s$ in terms of the Lagrange multiplier
\[
    y := \frac{1}{\gamma}(u - s), \quad \text{so that} \quad \nabla \varphi_1(u) = -\Pi_{\text{dom} \varphi_1} y
\]
splitting (FBS) step at affine subspace of backward envelope definition (3.3) emphasizes the close connection that the DRE has with the
see e.g. formulation (3.2) shows that minimization Connections with the forward-backward envelope.

3.1. Connections with the forward-backward envelope. The majorization-minimization formulation (3.2) shows that $v$ is the result of a forward-backward splitting (FBS) step at $u = \text{prox}_{\gamma \varphi_1}(s)$, written more compactly as
\[
v \in \text{prox}_{\gamma \varphi_2}(u - \gamma \nabla \varphi_1(u)), \tag{3.9}
\]
see e.g. [6, 28] for an extensive discussion on nonconvex FBS. In particular, the definition (3.3) emphasizes the close connection that the DRE has with the forward-backward envelope (FBE) as in [28], namely
\[
\varphi^\text{DR}_\gamma(s) = \varphi^\text{FB}_\gamma(u) \quad \text{for } u = \text{prox}_{\gamma \varphi_1}(s). \tag{3.10}
\]
A first immediate consequence, due to Proposition A.4 in the Appendix, is that the DRE is “flat” along lines orthogonal to $\text{dom} \varphi_1$:
\[
\varphi^\text{DR}_\gamma(s + \nu) = \varphi^\text{DR}_\gamma(s) \quad \text{for all } s \in \mathbb{R}^p \text{ and } \nu \in \text{dom} \varphi_1^+. \tag{3.11}
\]

The FBE, first introduced in [24] and further studied and extended in [27, 28, 22], is an exact penalty function for FBS, in the same way as the DRE is for DRS, as we will see later on in this section. Strictly speaking, the FBE is defined when the smooth term $\varphi_1$ is full domain. However, when the domain of $\varphi_1$ is a proper affine subspace of $\mathbb{R}^p$ there is no ambiguity in regarding the FBE rather as a function $\varphi^\text{FB}_1 : \text{dom} \varphi_1 \to \mathbb{R}$. This causes no troubles, since the composition with $\text{prox}_{\gamma \varphi_1}$ in (3.10) ensures that the argument of $\varphi^\text{FB}_1$ is always in its domain. For the sake of notational simplicity we will then write $\varphi^\text{FB}_\gamma$ rather than $\varphi^\text{FB}_\gamma|_{\text{dom} \varphi_1}$, and accordingly we shall write $\nabla \varphi^\text{FB}_\gamma$ in place of $\nabla \varphi^\text{DR}_\gamma$ whenever the latter exists.

3.2. Properties. The equivalence (3.10) shows that the DRE $\varphi^\text{DR}_\gamma$ and the FBE $\varphi^\text{FB}_1$ are basically the same function up to a change of variable. As a consequence, regularity properties of the DRE can be shown with minimal effort by invoking analogous properties of the FBE. For instance, since the mapping $s \mapsto u = \text{prox}_{\gamma \varphi_1}(s) \in \text{dom} \varphi_1$ is Lipschitz continuous for $\gamma < 1/|\sigma_{\varphi_1}|$ (see Prop. A.3(ii)) it is immediate to infer that the DRE inherits the strict continuity property of the FBE [28, Prop. 4.2].

Proposition 3.1 (Strict continuity). For all feasible $\gamma$ the DRE $\varphi^\text{DR}_\gamma$ is a real-valued and strictly continuous function over $\mathbb{R}^p$.

We now investigate the fundamental connections between the DRE $\varphi^\text{DR}_\gamma$ and the original function $\varphi$. We show, for $\gamma$ 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 of $\varphi$ and $\varphi^\text{DR}_\gamma|_{\text{dom} \varphi_1}$. Due to the fact that $\varphi^\text{DR}_\gamma$ is constant on lines orthogonal to $\text{dom} \varphi_1$ (cf. (3.11)) we clearly cannot expect the DRE to have bounded level sets, unless $\varphi_1$ has full domain. Therefore, it should not be surprising that some properties of $\varphi$ are enjoyed by the DRE (and viceversa) up to a suitable restriction of its domain. All results are based on a key property of the DRE which we state below. The proof is similar to that of [28, Prop. 4.3], but we briefly outline it here for the sake of self-inclusiveness.

Proposition 3.2 (Sandwiching property). Let $s \in \mathbb{R}^p$ and a feasible $\gamma$ be fixed, and consider $u, v$ generated by one DRS iteration. Then,
(i) \( \varphi_{\gamma}^{DR}(s) \leq \varphi(u) \);
(ii) \( \varphi(v) + \frac{1-\gamma \sigma_{s}}{2}\|u-v\|^{2} \leq \varphi_{\gamma}^{DR}(s) + \frac{1-\gamma \sigma_{s}}{2}\|u-v\|^{2} \).

Equivalently, this holds for all \( u \in \text{dom} \varphi_{1} \), \( s \in u + \gamma \nabla \varphi_{1}(u) + \text{dom} \varphi_{1}^{+} \) and \( v \in \text{prox}_{\gamma \varphi_{2}}(u - \gamma \nabla \varphi_{1}(u)) \).

Proof. Plugging \( w = u \) in (3.3) proves 3.2(i). The minimizer is instead \( w = v \in \text{dom} \varphi_{2} \subseteq \text{dom} \varphi_{1} \), resulting in
\[
\varphi_{\gamma}^{DR}(s) = \varphi_{2}(v) + \varphi_{1}(u) + \langle \nabla \varphi_{1}(u), v - u \rangle + \frac{1}{2\gamma\|u-v\|^{2}} \]
\[
\text{Rem. 2.2(c)} \geq \varphi_{2}(v) + \varphi_{1}(v) - \frac{\gamma}{2\gamma\|u-v\|^{2}} \|v-u\|^{2} + \frac{1}{2\gamma\|u-v\|^{2}} \]
and one inequality of 3.2(ii) follows. The other inequality is readily proven by using the quadratic lower bound in Rem. 2.2(c) instead. Finally, the last claim follows from the expression (3.9) and Prop. A.3(i).

**Theorem 3.3** (Minimization and level-boundedness equivalence). For any feasible \( \gamma < 1/L_{\varphi_{1}} \) the following hold:
(i) \( \inf \varphi = \inf \varphi_{\gamma}^{DR} \);
(ii) \( \text{argmin} \varphi = \text{prox}_{\gamma \varphi_{1}}(\text{argmin} \varphi_{\gamma}^{DR}) \).
(iii) \( \varphi \) is level bounded iff so is \( \varphi_{\gamma}^{DR} \) on \( \text{dom} \varphi_{1} \).

Proof. That \( \varphi_{\gamma}^{DR} \) and \( \varphi \circ \text{prox}_{\gamma \varphi_{1}} \) have same infima and minimizers easily follows from Prop. 3.2, and clearly \( \inf \varphi \circ \text{prox}_{\gamma \varphi_{1}} = \inf \varphi \) since \( \text{prox}_{\gamma \varphi_{1}} \) is invertible on \( \text{dom} \varphi_{1} \cap \text{dom} \varphi_{1} \) for any feasible \( \gamma \) (cf. Prop. A.3(ii)). To show 3.3(iii), observe first that, due to the equivalence of infima and (3.11), if either \( \varphi \) or \( \varphi_{\gamma}^{DR} \) is level bounded then \( \varphi \text{=} \inf \varphi = \inf \varphi_{\gamma}^{DR} \) is finite and attained by both functions. Let \( \alpha > \varphi \), be arbitrary.

\( \blacklozenge \) Suppose first that \( \varphi_{\gamma}^{DR} \) is level bounded and let \( u \in \text{lev}_{\leq \alpha} \varphi \). Then, clearly \( u \in \text{dom} \varphi_{1} \) and \( s = u + \gamma \nabla \varphi_{1}(u) \) is such that \( \text{prox}_{\gamma \varphi_{1}}(s) = u \) (cf. Prop. A.3(i)). Then, from Prop. 3.2 it follows that \( s \in \text{lev}_{\leq \alpha} \varphi_{\gamma}^{DR} \). In particular,
\[
\text{lev}_{\leq \alpha} \varphi \subseteq \text{dom} \varphi_{1} \cap [I + \gamma \nabla \varphi_{1}](\text{lev}_{\leq \alpha} \varphi_{\gamma}^{DR}) \subseteq [I + \gamma \nabla \varphi_{1}](\text{lev}_{\leq \alpha} \varphi_{\gamma}^{DR} \text{dom} \varphi_{1}).
\]
Since \( I + \gamma \nabla \varphi_{1} \) is Lipschitz continuous, necessarily \( \text{lev}_{\leq \alpha} \varphi \) is bounded.

\( \diamondsuit \) Suppose now that \( \varphi \) is level bounded, and contrary to the claim suppose that for all \( k \in \mathbb{N} \) there exists \( s_{k} \in \text{lev}_{\leq \alpha} \varphi_{\gamma}^{DR} \text{dom} \varphi_{1} \setminus \mathcal{B}(0; k) \). Let \( u_{k} = \text{prox}_{\gamma \varphi_{1}}(s_{k}) \) so that \( s_{k} = u_{k} + \gamma \nabla \varphi_{1}(u_{k}) \) (since \( s_{k} \in \text{dom} \varphi_{1} \)), and \( v_{k} \in \text{prox}_{\gamma \varphi_{2}}(u_{k} - \gamma \nabla \varphi_{1}(u_{k})) \); from Prop. 3.2 it follows that \( v_{k} \in \text{lev}_{\leq \alpha} \varphi \), and that
\[
\alpha - \varphi_{\gamma}^{DR}(s_{k}) - \varphi_{\gamma}^{DR}(v_{k}) = \varphi_{\gamma}^{DR}(s_{k}) - \varphi_{\gamma}^{DR}(s_{k}) - \varphi(v_{k}) \geq \frac{1-\gamma \sigma_{s}}{2\gamma \|u-v\|^{2}} \|u_{k} - v_{k}\|^{2}.
\]
Therefore,
\[
\|v_{k}\| \geq \|u_{k} - u_{0}\| - \|u_{0}\| + \|u_{k} - v_{k}\| \geq \frac{1}{\gamma \|u_{k}\|} \|s_{k} - s_{0}\| - \|u_{0}\| + \|u_{k} - v_{k}\| \geq \frac{k-\|u_{0}\|}{\gamma \|u_{k}\|} - \|u_{0}\| - \sqrt{\frac{2(\alpha - \varphi_{\gamma})}{\gamma \|u_{k}\|}} \to +\infty \text{ as } k \to \infty,
\]
and therefore \( \text{lev}_{\leq \alpha} \varphi \) cannot be bounded. \( \square \)

4. Convergence of Douglas-Rachford splitting

Closely related to the DRE, the augmented Lagrangian (3.5) was used in [21] under the name of Douglas-Rachford merit function to analyze DRS for the special
case $\lambda = 1$. It was shown that for sufficiently small $\gamma$ there exists $c > 0$ such that the iterates generated by DRS satisfy
\[
L_{\gamma}(u^{k+1}, v^{k+1}, y^{k+1}) \leq L_{\gamma}(u^{k}, v^{k}, y^{k}) - \gamma\|u^{k} - u^{k+1}\|^2, \quad (4.1)
\]
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 $\varphi$. However, it wasn’t actually proved that DRS does converge, namely the fact that the fixed-point residual vanishes and that therefore for any $\varepsilon > 0$ the stopping criterion $\|u^{k} - v^{k}\| \leq \varepsilon$ eventually is satisfied. This limitation comes from using $L_{\gamma} : \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ as a function of $u, v$ and $y$ (or $s$) separately in (4.1), which in fact is not guaranteed to be bounded below. Because of this, the standard practice of telescoping the inequality (4.1) does not ensure that $\|u^{k+1} - u^{k}\|$ vanishes, and consequently nor that (the proportional quantity) $\|u^{k} - v^{k}\|$ does. Similar remarks apply to [19], where the Peaceman-Rachford merit function was introduced to analyze PRS, that is, DRS with $\lambda = 2$.

On the contrary, the DRE is defined as a section of $L_{\gamma}$ which is tightly connected both to $\varphi$ and to DRS iterations: namely, $\varphi^{\text{DR}}(s) = L_{\gamma}(u_{\gamma}(s), v_{\gamma}(s), \gamma^{-1}(u_{\gamma}(s) - s))$, where $u_{\gamma}(s)$ and $v_{\gamma}(s)$ are respectively the $u$- and $v$-updates of point $s$ through DRS with stepsize $\gamma$ (cf. (3.4)). As shown in the previous section, our interpretation of the DRE as a function of the sole variable $s$ overcomes this limitation and preserves important properties the original function $\varphi$ may have such as level boundedness.

We now generalize the decrease property (4.1) shown in [21, 19] by considering arbitrary relaxation parameters $\lambda \in (0, 4)$ (as opposed to $\lambda \in \{1, 2\}$) and providing tight ranges for the stepsize $\gamma$ whenever $\lambda \in (0, 2]$. We are only interested in the case $\gamma < 1/L_{\varphi_{\gamma}}$, for otherwise the DRE may fail to be lower bounded. Morever, it will be shown in Section 4 that the bound $\gamma < 1/L_{\varphi_{\gamma}}$ is necessary for ensuring the convergence of DRS, unless the generality of Assumption I is sacrificed.

**Theorem 4.1 (Sufficient decrease).** Consider one DRS update $s \mapsto (u, v, s^+)$. Then,
\[
\varphi^{\text{DR}}(s) - \varphi^{\text{DR}}(s^+) \geq c \|u - v\|^2 \quad (4.2)
\]
where $c$ is a strictly positive constant defined as
\[
c := \frac{\lambda}{(1 + \gamma L_{\varphi_{\gamma}})^2 \left(\frac{2 - \lambda}{2 \gamma} - \max\left\{[\sigma_{\varphi_{\gamma}}], -1, L_{\varphi_{\gamma}}(\gamma L_{\varphi_{\gamma}} - \frac{1}{2})\right\}\right)}.
\]
If $\varphi_1$ is strongly convex, then (4.2) also holds for $2 \leq \lambda < \frac{4}{1 + \sqrt{1 - p_{\varphi_{\gamma}}}}$ and feasible $\gamma$ such that $\frac{p_{\varphi_{\gamma}} \lambda - 4}{4 \sigma_{\varphi_{\gamma}}} < \gamma < \frac{p_{\varphi_{\gamma}} \lambda + 4}{4 \sigma_{\varphi_{\gamma}}}$, where $\delta := \sqrt{(p_{\varphi_{\gamma}})^2 - 8 p_{\varphi_{\gamma}}(\lambda - 2)}$, in which case
\[
c := \frac{\lambda}{(1 + \gamma L_{\varphi_{\gamma}})^2 \left(\frac{2 - \lambda}{2 \gamma} + \sigma_{\varphi_{\gamma}}(\frac{\lambda}{2} - \gamma L_{\varphi_{\gamma}})\right)}.
\]

For the sake of readability, the proof of Theorem 4.1 is referred to Appendix B.

**Remark 4.2 (Simpler bounds for DRS).** By using the (more conservative) estimate $\sigma_{\varphi_{\gamma}} = 0$ when the smooth function $\varphi_1$ is convex, and $\sigma_{\varphi_{\gamma}} = -L_{\varphi_{\gamma}}$ otherwise, the range of $\gamma$ can be simplified as follows in case $\lambda \in (0, 2]$:
\[
\lambda \in (0, 2): \begin{cases} 
\gamma < \frac{1}{L_{\varphi_{\gamma}}} & \text{if } \varphi_1 \text{ is convex} \\
\gamma < \frac{2 \lambda}{4 \sigma_{\varphi_{\gamma}}} & \text{if } \varphi_1 \text{ is str. convex}
\end{cases} \quad \lambda = 2: \begin{cases} 
\gamma < \frac{1}{L_{\varphi_{\gamma}}} & \text{if } \varphi_1 \text{ is str. convex} \\
0 & \text{otherwise}.
\end{cases}
\]

**Theorem 4.3 (Subsequential convergence).** Suppose that the cost function $\varphi$ is lower bounded. Then, the following hold for the iterates generated by DRS with stepsize $\gamma$ and relaxation $\lambda$ as in Theorem 4.1:
\[
2\text{Although } \nu_{\gamma}(s) \text{ is not necessarily unique, this definition does not depend on the specific choice.}
\]
(i) the residuals $(u^k - u^k)_{k \in \mathbb{N}}$ vanish 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 $\varphi$ and on which $\varphi$ has same value, this being the limit of $(\varphi_{\gamma}^{DR}(s^k))_{k \in \mathbb{N}}$;
Moreover, if $\varphi$ has bounded level sets, then the sequences are bounded.

Proof. To avoid trivialities, we assume that the stopping criterion $v^k = u^k$ is never met, so that the algorithm runs infinite many iterations.

\[ \sum_{k \in \mathbb{N}} \| u^k - v^k \|^2 \leq \sum_{k \in \mathbb{N}} \left( \varphi_{\gamma}^{DR}(s^k) - \varphi_{\gamma}^{DR}(s^{k+1}) \right) \leq \varphi_{\gamma}^{DR}(0) - \inf_{\gamma} \varphi_{\gamma}^{DR}. \]

Since $\inf_{\gamma} \varphi_{\gamma}^{DR} = \inf \varphi > -\infty$ and $\varphi_{\gamma}^{DR}$ is real valued (cf. Prop. 3.1 and Thm. 3.3), it follows that $(u^k - v^k)_{k \in \mathbb{N}}$ is square summable, hence the claimed rate of convergence. Moreover, since $\varphi_{\gamma}^{DR}(s^k)$ is decreasing it admits a limit, be it $\varphi_* > -\infty$.

\[ |u^k - v^k| = \sum_{k \in \mathbb{N}} \frac{1}{2} \| k \| u^k - v^k \|^2 \leq \sum_{k \in \mathbb{N}} \left( \varphi_{\gamma}^{DR}(s^k) - \varphi_{\gamma}^{DR}(s^{k+1}) \right) \leq \varphi_{\gamma}^{DR}(0) - \inf_{\gamma} \varphi_{\gamma}^{DR}. \]

Let $\tilde{\varphi}_{\gamma}^{DR}$ denote the outer semicontinuous of $\varphi_{\gamma}^{DR}$ [26, Ex. 5.23(b)] implies that $\varphi_{\gamma}^{DR}(s^k)$ is bounded. Due to Lipschitz continuity of $\prox_{\gamma \varphi_{\gamma}^{DR}}$ (cf. Prop. A.3(i)), also $(u^k)_{k \in \mathbb{N}}$ is bounded and since $v^k - u^k \to 0$ we conclude that so is $(v^k)_{k \in \mathbb{N}}$.
In [21] it was shown that the augmented Lagrangian decreases along the iterates generated by the non-relaxed DRS. This fact was then used to prove global convergence in case the sequence remains bounded, which was later shown to be the case in [19] when \( \varphi \) has bounded level sets. Due to the equivalence of the DRE and the augmented Lagrangian evaluated at points generated by DRS, cf. (3.4), by invoking Theorem 4.1 we can extend their result to the tight ranges we provided. Notice that boundedness of the sequence for any \( \lambda \) is ensured by the level boundedness of the DRE, which holds for all \( \gamma < \frac{1}{L_{\varphi_1}} \) and independently of \( \lambda \) when \( \varphi \) is level bounded.

**Theorem 4.4** (Global convergence of DRS [21, Thm. 2]). Suppose that \( \varphi \) is level bounded and that \( \varphi_1 \) and \( \varphi_2 \) are semi-algebraic. Then, the sequences \((u^k)_{k \in \mathbb{N}}\) and \((v^k)_{k \in \mathbb{N}}\) generated by DRS with \( \gamma \) and \( \lambda \) as in Theorem 4.3 converge to (the same) stationary point for \( \varphi \).

**Remark 4.5** (Adaptive variant when \( L_{\varphi_1} \) is unknown). When \( \sigma_{\varphi_1} \) is not known, then a conservative estimate \( \sigma_{\varphi_1} = -L_{\varphi_1} \) is always feasible or, in case \( \varphi_1 \) is convex, \( \sigma_{\varphi_1} = 0 \) can be fixed (cf. Remark 4.2). In most applications, however, it is the very Lipschitz constant \( L_{\varphi_1} \) which is not known, in which case it can be adaptively retrieved. This is done by replacing \( L_{\varphi_1} \) with an initial estimate \( L > 0 \), and by checking at each iteration if the quadratic upper bound for \( \varphi_1 \) as in Remark 2.2(c) holds with \( L \) in place of the unknown \( L_{\varphi_1} \). Whenever the bound is violated, it suffices to, say, double the estimate \( L \) and decrease the stepsize \( \gamma \) accordingly. Notice that there is no need to compute \( \nabla \varphi_1 \), as it holds that

\[
\langle \nabla \varphi_1(u), u^+ - u \rangle = \frac{1}{L} \langle \Pi_{\operatorname{dom} \varphi_1} s - u, u^+ - u \rangle = \frac{1}{L} \langle s - u, u^+ - u \rangle
\]

where the last equality is due to the fact that \( u, u^+ \in \operatorname{dom} \varphi_1 \). The procedure is summarized in Algorithm 1. It is important to observe that, since replacing \( L_{\varphi_1} \),

**Algorithm 1** DRS with adaptive stepsize

```
REQUIRE \( s^0 \in \mathbb{R}^p \), \( L > 0 \), \( \gamma, \lambda \) as in Rem. 4.2 with \( L \) in place of \( L_{\varphi_1} \)
INITIALIZE \( u^0, v^0 \) and \( s^1 \) as in DRS
For \( k = 1, 2, \ldots \)
1: \( u^k = \operatorname{prox}_{\gamma\varphi_1}(s^k) \)
2: if \( \varphi_1(u^k) > \varphi_1(u^{k-1}) + \frac{1}{L} \langle s^{k-1} - u^{k-1}, u^k - u^{k-1} \rangle + \frac{L}{2} \| u^k - u^{k-1} \|^2 \) then
   \( L \leftarrow 2L \), \( \gamma \leftarrow \gamma/2 \), and go back to step 1
3: \( s^k \in \operatorname{prox}_{\gamma\varphi_2}(2u^k - s^k) \)
4: \( s^{k+1} = s^k + \lambda(u^k - u^k) \)
```

with any \( L \geq L_{\varphi_1} \), still satisfies the (upper) bound in Remark 2.2(c), it follows that \( L \) is incremented only a finite number of times. Therefore, there exists an iteration \( k_0 \) starting from which \( \gamma \) and \( L \) are constant; in particular, the convergence result stated in Theorem 4.3 covers this adaptive variant as well.

**4.1. Tightness of the results.** When both \( \varphi_1 \) and \( \varphi_2 \) 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 (see e.g., [3]). In fact, the whole sequence \((u^k)_{k \in \mathbb{N}}\) converges and \( \varphi_1 \) needs not even be differentiable in this case. On the contrary, when \( \varphi_2 \) is nonconvex then the bound \( \gamma < \frac{1}{L_{\varphi_1}} \) plays a crucial role, as the next example shows.
Example 4.6 (Necessity of $\gamma < 1/L_{\varphi_1}$). Fix $L > 0$, $\sigma \in [-L, L]$ and $t > 1$, and let $\varphi = \varphi_1 + \varphi_2$, where $\varphi_2 = \delta_{\{0,1\}}$ and

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

Notice that $\text{dom } \varphi = \{\pm1\}$, and therefore $\pm1$ are the unique stationary points of $\varphi$ (in fact, they are also global minimizers). Moreover, $\varphi_1 \in C^{1,1}(\mathbb{R})$ with $L_{\varphi_1} = L$ and $\varphi_1 = \sigma$, and $\text{prox}_{\gamma \varphi_1}$ is well defined iff $\gamma < 1/|\sigma|$. More precisely,

$$\text{prox}_{\gamma \varphi_1}(s) = \begin{cases} \frac{s}{1 + \gamma L} & \text{if } s \leq t(1 + \gamma L) \\ \frac{s}{1 + \gamma \sigma} - \frac{\gamma(L-\sigma)}{1 + \gamma \sigma} s - v^k & \text{otherwise,} \end{cases}$$

where $\text{sgn}(0) = \{\pm1\}$. Suppose that $\gamma \geq 1/L$ and fix $\lambda > 0$. Then, one DRS iteration produces $v^k = -\text{sgn}(s^k)$ if $s^k \leq t(1 + \gamma L)$ and in particular the DRS-residual is

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

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

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

Notice that the first element in the set above is always larger than $1$, and therefore 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 $s^k \leq t(1 + \gamma L)$ and the residual will then be $u^k - v^k = \frac{s^k}{1 + \gamma L} + \text{sgn}(s^k)$ which is bounded away from zero, contradicting the fact that $u^k - v^k \to 0$. \hfill $\Box$

Notice that in Example 4.6 we actually showed that for $\gamma \geq 1/L_{\varphi_1}$ DRS fails to converge for any $\lambda > 0$ regardless of the starting point $s^0$. This was possible by allowing a too large stepsize $\gamma$; the next example shows the necessity of bounding $\lambda$.

Example 4.7 (Necessity of $0 < \lambda < 2(1 + \gamma \sigma)$). Fix $L > 0$ and $\sigma \in [-L, L]$, and consider $\varphi = \varphi_1 + \varphi_2$ where $\varphi_2 = \delta_{\{0\}}$ and

$$\varphi_1(x) = \begin{cases} \frac{\sigma}{2} x^2 & \text{if } x \leq 1 \\ \frac{\sigma}{2} x^2 + \frac{L-\sigma}{2} (x-1)^2 & \text{otherwise.} \end{cases}$$

For any $s^k$ one DRS iteration produces $v^k = 0$, and in particular the DR-residual is

$$u^k - v^k = u^k = \begin{cases} \frac{s^k}{1 + \gamma \sigma} & \text{if } s^k \leq 1 + \gamma \sigma \\ \frac{s^k + \gamma(L-\sigma)}{1 + \gamma \sigma} & \text{otherwise.} \end{cases}$$

Suppose that $\lambda \geq 2(1 + \gamma \sigma)$; then it is easy to check that starting from $s^0 \neq 0$ we have $s^k \neq 0$ for all $k$. Moreover, if the DRS-residual converges to $0$, then $\min \{|s^k|, |s^k + \gamma(L-\sigma)|\} \to 0$ and in particular, eventually $s^k \leq 1 + \gamma \sigma$. The iterations will then reduce to

$$s^{k+1} = s^k + \lambda(u^k - v^k) = \left(1 - \frac{\lambda}{1 + \gamma \sigma}\right)s^k.$$

Since $\lambda \geq 2(1 + \gamma \sigma)$ and $s^k \neq 0$ for all $k$, we have $|s^{k+1}/s^k| \geq 1$, contradicting the fact that $s^k \to 0$. \hfill $\Box$

Let us draw some conclusions:
the nonsmooth function $\varphi_2$ is (strongly) convex in Example 4.7, therefore even for fully convex formulations the bound $0 < \lambda < 2(1 + \sigma_{\varphi_1})$ needs be satisfied;

- if $\lambda > 2$ (which is feasible only if $\varphi_1$ is strongly convex, i.e., if $\sigma_{\varphi_1} > 0$), then, regardless of whether also $\varphi_2$ is (strongly) convex or not, we obtain that the stepsise must be lower bounded as $\gamma > \frac{\lambda - 2}{\sigma_{\varphi_1}}$;

- if $\varphi_1$ is not strongly convex, i.e., if $\sigma_{\varphi_1} \leq 0$, we infer the bound $\lambda \in (0, 2)$: this means, for instance, that even in the fully convex case, plain (nonstrong) convexity of $\varphi_1$ is not enough to guarantee convergence of the Peaceman-Rachford splitting;

- combined with the bound $\gamma < 1/L_{\varphi_1}$ shown in Example 4.6, we infer that (at least when $\varphi_2$ is nonconvex) necessarily $0 < \lambda < 2(1 + p_{\varphi_1})$ and consequently $\lambda \in (0, 4)$.

In particular we can infer the following:

**Theorem 4.8** (Tightness). Unless the generality of Assumption I is sacrificed, when $\lambda \in (0, 2)$ or $\varphi_1$ is not strongly convex the bound $\gamma < \min \left\{ \frac{1}{\lambda_{\varphi_1}}, \frac{2 - \lambda}{2\sigma_{\varphi_1}} \right\}$ is tight for ensuring convergence of DRS. Similarly, PRS (i.e., DRS with $\lambda = 2$) is ensured to converge iff $\varphi_1$ is strongly convex and $\gamma < 1/L_{\varphi_1}$.

5. Alternating Direction Method of Multipliers

In convex optimization, it is well known that ADMM and DRS are essentially the same algorithm applied to the respective dual formulations. Unfortunately, this equivalence does not extend to the nonconvex case due to the limitation of duality theory. In this section we provide a *primal* equivalence of the two algorithms that cope with nonconvexity and serves as universal framework. This will allow to extend the theory developed for DRS to ADMM. To this end, we invoke the notion of *image function*, also known as *infiomal post-composition* or *epi-composition* [1, 3, 26].

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

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

In formulation (1.2) the linear constraint between $x$ and $z$ can be decoupled by adding a slack variable $s \in \mathbb{R}^p$ and by rewriting the problem in the equivalent form

$$\begin{align*}
\min_{(x,z,s) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^p} & \quad f(x) + g(z) \\
\text{subject to} & \quad Ax = b - s, \quad Bz = s.
\end{align*}$$

Since the problem is independent of the order of minimization [26, Prop. 1.35] we may minimize first with respect to $(x,z)$ to arrive to

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

Therefore, ADMM problem formulation (1.2) can be expressed in the equivalent DRS formulation (1.1) as

$$\begin{align*}
\min_{s \in \mathbb{R}^p} \quad & (Bg)(s) + (Af)(b - s).
\end{align*}$$

(5.1)

Restricting the analysis to the convex case, from [30] it can be inferred that DRS on problem (5.1) is equivalent to ADMM on problem (1.2). However, the arguments still rely on the known dual equivalence of ADMM and DRS, which cannot be exploited in the general nonconvex framework here discussed. Our result is rather based on properties of image functions which also hold for nonconvex functions. It is convenient to eliminate the DRS variable $s$ by expressing the iteration in terms
of the dual multiplier $y$. To this end, we consider the equivalent DRS iteration described in (3.8).

For $\varphi_1 = (Bg)$ and $\varphi_2 = (Af)(b - \cdot)$, it follows from Proposition C.1(iii) in the Appendix that $\text{prox}_{\gamma\varphi_1}(s) = B \text{argmin}_z \{g(z) + \frac{1}{2\gamma}||Bz - s||^2\}$ and, by means of a simple change of variable, $\text{prox}_{\gamma\varphi_2}(s) = b - A \text{argmin}_x \{f(x) + \frac{1}{2\gamma}||Ax + s||^2\}$. Then, (3.8) becomes

$$
\begin{align*}
y^{k+1/2} &= y - \frac{1}{\mu}((1 - \lambda)(u^k - v^k)) \\
u^{k+1} &= B \text{argmin}_z \{g(z) + \frac{1}{2\gamma}||Bz - v^k + \gamma y^{k+1/2}||^2\} \\
y^{k+1} &= y^{k+1/2} + \frac{1}{\gamma}(u^{k+1} - u^k) \\
v^{k+1} &= v^k - \lambda(Au^{k+1} - \tilde{v}^{k+1/2})
\end{align*}
$$

By introducing suitable variables $x^k \in \mathbb{R}^m$ and $z^k \in \mathbb{R}^n$, so as to express $u^k = Bz^k$ and $v^k = b - Ax^k$, this reduces to

$$
\begin{align*}
y^{k+1/2} &= y - \frac{1}{\mu}((1 - \lambda)(Ax^k + Bz^k - b)) \\
z^{k+1} &= \text{argmin}_x \{g(x) + \frac{1}{2\gamma}||Ax + Bz - b + \gamma y^{k+1/2}||^2\} = \text{argmin}_{x,y} \{L_{1/\gamma}(x^k, z, y^{k+1/2})\} \\
y^{k+1} &= y^{k+1/2} + \frac{1}{\gamma}(Ax^k + Bz^{k+1} - b) \\
x^{k+1} &= \text{argmin}_x \{f(x) + \frac{1}{2\gamma}||Ax + Bz^{k+1} - b + \gamma y^{k+1}||^2\} = \text{argmin}_{x,y} \{L_{1/\gamma}(x, z^{k+1}, y^{k+1})\}
\end{align*}
$$

which is exactly ADMM with $\beta = 1/\gamma$, up to one update order switch (that is, starting from the $y^{k+1/2}$-update rather than from the $x^{k+1}$-update). To sum up, we have shown the following.

**Theorem 5.2** (Primal equivalence of ADMM and DRS). For proper and lsc functions $f$ and $g$ and matrices $A$, $B$ and $b$ of suitable size, up to an update order shift, one step of ADMM applied to

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

is equivalent to one step of DRS applied to

$$\text{minimize}_{s \in \mathbb{R}^n} \varphi(s) := (Bg)(s) + (Af)(b - s)$$

where $\gamma \leftrightarrow 1/\beta$ and $(s, u, v, y) \leftrightarrow (Bz - gy, Bz, b - Ax, y)$.

Moreover, for all $k \geq 1$ the following hold

(i) $\varphi_{i/\beta}(s^k) = \mathcal{L}_\beta(x^k, z^k, y^k)$;

(ii) $(Af)(Ax^k) = f(x^k)$;

(iii) $(Bg)(Bz^k) = g(z^k)$;

(iv) if $(Bg) \in C^{1,1}(\text{rng} B)$, then $\nabla (Bg)(Bz^k) = - \Pi_{\text{rng} B} y^k$.

**Proof.** By definition of $x^k$, we have that $\mathcal{L}_\beta(x^k, z^{k-1}, y^{k-1}) \leq \mathcal{L}_\beta(s, z^{k-1}, y^{k-1})$ for all $s \in \mathbb{R}^m$. If $s \in \mathbb{R}^m$ is such that $As = Ax^k$, then the inequality reduces to $f(x^k) \leq f(s)$, and 5.2(ii) follows from the definition of image function. A similar reasoning shows 5.2(iii); in turn, 5.2(i) follows from (3.4). Finally, 5.2(iv) follows from (3.7).

5.1. Convergence of ADMM. In order to extend the theory developed for DRS to ADMM we shall impose that $\varphi_1$ and $\varphi_2$ as in (5.1) comply with Assumption I. This motivates the following blank set requirement.
Assumption II (Requirements for the ADMM formulation (1.2)). Functions $f : \mathbb{R}^m \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$, and matrices $A \in \mathbb{R}^{p \times m}$, $B \in \mathbb{R}^{p \times n}$ and $b \in \mathbb{R}^p$ are such that $(Af)$ is a lsc function, and

(i) (Smoothness) $(Bg)$ is $C^{1,1}$ on its (affine) domain $\text{rng} \ B$;

(ii) (ADMM-feasibility) for $\beta$ large enough, the subproblems in (ADMM) admit a (not necessarily unique) solution;

(iii) (Domain inclusion) $A \text{ dom} \ f \subseteq b + \text{rng} \ B$.

These requirements generalize those in Assumption I by allowing linear constraints more generic than $x - z = 0$, cf. (3.6). Notice that prox-boundedness of $(Af)$ is ensured by ADMM-feasibility, as it follows from Proposition C.1(iii) in the Appendix. For $\varphi_1$ as in (5.1) it holds that $\text{dom} \ \varphi_1 = B \text{ dom} \ g$, and therefore unless $B$ is surjective $\varphi_1$ cannot have full domain. Thankfully, the theory for DRS developed in the previous sections accounted for this event by considering functions with affine domain. Although we could still work under this more generic assumption, at this stage this generalization is no longer needed, which is why we consider full domain functions $g$.

**Theorem 5.3** (Convergence of ADMM). Suppose that $f + g$ is lower bounded on the feasible set $\{(x, z) \in \mathbb{R}^m \times \mathbb{R}^n \mid Ax + Bz = b\}$. Let $\lambda$ and $\gamma$ be as in Theorem 4.1 for $\varphi_1 = (Bg)$ and $\varphi_2 = (Af)(b - \cdot)$. Then, the following hold for the iterates generated by ADMM with penalty $\beta = 1/\gamma$ and relaxation $\lambda$:

(i) $\mathcal{L}_\beta(x^{k+1}, z^{k+1}, y^{k+1}) \leq \mathcal{L}_\beta(x^k, z^k, y^k) - \frac{\beta - 1}{2\lambda} \|Ax^k + Bz^k - b\|^2$, and the residual $(Ax^k + Bz^k - b)_{k \in \mathbb{N}}$ vanishes with $\min_{k \in \mathbb{N}} \|Ax^k + Bz^k - b\| = o(1/\sqrt{k})$;

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

- $-A^T y = \partial f(x)$
- $-B^T y = \partial g(z)$
- $Ax + Bz = b$.

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

(iii) the sequence $(Ax^k, Bz^k, y^k)_{k \in \mathbb{N}}$ is bounded provided that $f(x) + g(z)$ is lower bounded on the feasible set $C := \{(x, z) \in \mathbb{R}^m \times \mathbb{R}^n \mid Ax + Bz = b\}$.

**Proof.**

\begin{itemize}
\item **5.3(i)** It follows from Thm. 5.2 that $s^k = Bz^k - \frac{1}{\beta} y^k$, $u^k = Bz^k$ and $v^k = b - Ax^k$ are the iterates generated by DRS with stepsize $\gamma = 1/\beta$ applied to $\varphi_1 + 2\varphi_2$, where $\varphi_1 = (Bg)$ and $\varphi_2 = (Af)(b - \cdot)$. Then the claim follows from Thms. 4.1 and 4.3, by observing that $u^k - v^k = Ax^k + Bz^k - b$.

\item **5.3(ii)** Suppose now that for some $K \subseteq \mathbb{N}$ the subsequence $((x^k, z^k, y^k))_{k \in K}$ converges to $(x, z, y)$, then, necessarily $Ax + Bz = b$. Moreover, the optimality conditions defining $z^k$ and $s^k$ in ADMM read

\[ 0 \in \partial f(x^k) + A^T y^{k-1} + \beta A^T (Ax^k + Bz^{k-1} - b) \]
\[ 0 \in \partial g(z^k) + B^T y^{k-1/2} + \beta B^T (Ax^k + Bz^k - b) = \partial g(z^k) + B^T y^{k}. \]

Continuity of $\nabla(Bg)$ on its domain then yields

\[ -B^T y = - B^T \Pi_{\text{rng}} B y \Longleftrightarrow - B^T \Pi_{\text{rng}} B y \]
\[ = B^T \nabla(Bg)(Bz^k) \Longleftrightarrow - B^T \nabla(Bg)(Bz) \]

and from Prop. C.2 we conclude that $-B^T y \in \partial g(z)$.

From Thms. 4.3 and 5.2(i) it follows that

\[ \mathcal{L}_\beta(x^k, z^k, y^k) = \varphi_2^D(s^k) \subseteq \varphi_1^D(s) = \mathcal{L}_\beta(x, z, y) = f(x) + g(z) \quad (5.2) \]
where \( s = Bz - \frac{1}{\lambda} y \) and the last equality follows from the fact that \( Ax + Bz = b \). Clearly, \( Bz^k \in \overline{\text{dom}} \, g = \overline{\text{dom}} (Bg) \) for all \( k \), and since \( g \) is lsc and \( (Bg) \) is continuous on its domain, we have

\[
(Bg)(Bz) \leq g(z) \leq \liminf_{k \to K} g(z_k) = \liminf_{k \to K} (Bg)(Bz_k) = (Bg)(Bz).
\]

Therefore, \( g(z^k) \to g(z) \) and (5.2) then implies that \( f(x^k) \to f(x) \). Notice that, since \( \varphi_1 = (Bg) \) is smooth, Prop. A.3(ii) implies that \( B(z^k - z^{k-1}) = u^k - u^{k-1} \) vanishes, and consequently so does \( Ax^k + Bz^{k-1} - b \). In particular,

\[
-A^T y = A^T (y^{k-1} + \beta(Ax^k + Bz^{k-1} - b)) \in \partial f(x^k),
\]

and \( f \)-attentive outer semicontinuity of \( \partial f \) [26, Prop. 8.7] ensures \( -A^T y \in \partial f(x) \).

\[\blacktriangleleft\] 5.3(iii) It suffices to show that \( \varphi = \varphi_1 + \varphi_2 \) is level bounded, as boundedness of the sequence \( (Ax^k, Bz^k, y^k) \) \( k \in \mathbb{N} \) will then follow from Thm. 4.3 in light of the DRS equivalence stated in Thm. 5.2. Let \( F(x, z) := f(x) + g(z) + \delta_C(x, z) \), where \( C = \{ (x, z) \in \mathbb{R} \times \mathbb{R}^m \mid Ax + Bz = b \} \) is the feasible set. For \( \alpha \in \mathbb{R} \) we have

\[
\text{lev}_{\leq \alpha} \varphi = \left\{ s \mid \inf_{x} \{ f(x) \mid Ax = b - s \} + \inf_{z} \{ g(z) \mid Bz = s \} \leq \alpha \right\} = \left\{ s \mid \inf_{x, z} \{ f(x) + g(z) \mid Ax = b - s, Bz = s \} \leq \alpha \right\} = \{ Bz \mid f(x) + g(z) \leq \alpha, \exists x : Ax + Bz = b \} = \{ Bz \mid (x, z) \in \text{lev}_{\leq \alpha} F, \exists x \}.
\]

Since \( \|Bz\| \leq \|B\| \|z\| \leq \|B\| \|(x, z)\| \) for any \( x, z \), it follows that if \( \text{lev}_{\leq \alpha} F \) is bounded then so is \( \text{lev}_{\leq \alpha} \varphi \).

\[\blacktriangleleft\]

As a consequence of the Tarski-Seidenberg theorem, functions \( \varphi_1 := (Bg) \) and \( \varphi_2 := (Af)(b - \cdot) \) are semialgebraic provided \( f \) and \( g \) are, see e.g., [5]. 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 5.2. We should emphasize, however, that the equivalence identifies \( u^k = Bz^k \) and \( \nu^k = b - Ax^k \), and thus only convergence of \( (Ax^k)_{k \in \mathbb{N}} \) and \( (Bz^k)_{k \in \mathbb{N}} \) can be deduced.

**Theorem 5.4** (Global convergence of ADMM). Suppose that \( f(x) + g(z) \) is level bounded on the feasible set \( \{ (x, z) \in \mathbb{R}^m \times \mathbb{R}^m \mid Ax + Bz = b \} \), and that \( f \) and \( g \) are semi-algebraic. Then, the sequence \( (Ax^k, Bz^k, y^k)_{k \in \mathbb{N}} \) generated by ADMM with \( \beta \) and \( \lambda \) as in Theorem 5.3 converges.

**Remark 5.5** (Simpler bounds for ADMM). In parallel with the simplifications outlined in Remark 4.2 for DRS, and by observing that \( (Bg) \) is strongly convex if so is \( g \), simpler (more conservative) bounds for the penalty parameter \( \beta \) in ADMM are

\[
\lambda \in (0, 2): \begin{cases} 
\beta > L & \text{if } g \text{ is convex} \\
\beta > \frac{2L}{\lambda} & \text{otherwise} \end{cases} \quad \lambda = 2: \begin{cases} \beta > L & \text{if } g \text{ is str. convex} \\
0 & \text{otherwise} \end{cases}
\]

where \( L = L_{(Bg)} \). In particular, if \( B \) is full-column rank or \( g \) is convex and \( B \) is full row-rank, then \( L = L_{g/\sigma_+(u^T B)} \) (cf. Thm. 5.8 and commentary thereafter).

As it was the case for DRS, the knowledge of \( L_{(Bg)} \) (or \( \sigma_1(Bg) \)) is actually not needed for ensuring convergence of ADMM iterations. In fact, the adaptive variant of DRS outlined in Remark 4.5 can be easily translated into an adaptive version of ADMM in which the penalty \( \beta \) is suitably increased whenever a quadratic upper bound fails to hold. Notice that, thanks to the equivalences outlined in Theorems 5.2(iii) and 5.2(iv), as it was the case for DRS this condition can be verified by
using quantities which are already available. Putting all the pieces together, the adaptive variant of ADMM is outlined in Algorithm 2. For the sake of simplicity, we only consider the case $\lambda = 1$, so that the half-update $g^Tz$ can be discarded.

**Algorithm 2** (Non-relaxed) ADMM with adaptive stepsize

**Require** $(x^0,z^0,y^0) \in \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^p$, $L > 0$, $\beta > 0$, $\beta > 2L$ otherwise

For $k = 0,1,2\ldots$

1. $x^{k+1} = \arg\min_x \mathcal{L}_\beta(x,z^k,y^k)$
2. $z^{k+1} = \arg\min_z \mathcal{L}_\beta(x^{k+1},z,y^k)$
3. if $g(z^{k+1}) > g(z^k) - \langle y^k, B(z^{k+1} - z^k) \rangle + \frac{L}{2}\|B(z^{k+1} - z^k)\|^2$ then
   $L \leftarrow 2L$, $\beta \leftarrow 2\beta$, and go back to step 1
4. $y^{k+1} = y^k + \beta(Ax^{k+1} + Bz^{k+1} - b)$

5.2. Sufficient conditions. We conclude the section by providing sufficient conditions on $f$ and $g$ ensuring that Assumption II is satisfied.

5.2.1. Lower semicontinuity of the image function.

**Proposition 5.6** (Lsc of $(Af)$). Suppose that $A \in \mathbb{R}^{p \times n}$ and the proper and lsc function $f : \mathbb{R}^m \to \mathbb{R}$ satisfy Assumption II(ii). Then, $(Af)$ is proper. Moreover, it is also lsc provided that for all $\bar{x} \in \text{dom} f$, either

(i) the set $X(s) := \arg\min_x \{f(x) \mid Ax = s\}$ is nonempty and $\text{dist}(0,X(s))$ is bounded for all $s \in \text{dom} f$ close to $A\bar{x}$,

(ii) or $\lim\inf_{d \to \bar{d}, t \to \infty} f(td) \geq \inf_{d \in \ker A} f(\bar{x} + d)$ for all $\bar{d} \in \ker A \setminus \{0\}$.

**Proof.** See Appendix C.

The requirement in Proposition 5.6(i) is much weaker than Lipschitz continuity of the map $s \mapsto X(s)$, which is the standing assumption in [29] for the analysis of ADMM. In fact, we do not even require uniqueness or boundedness of the sets of minimizers. Notice that Assumption II(ii) is only invoked to infer that $(Af)$ is proper. Moreover, whenever the inequality in Proposition 5.6(ii) holds, then it is actually an equality. Its role can be better visualized by considering $f : \mathbb{R}^2 \to \mathbb{R}$ defined as

$$f(x,y) = \begin{cases} 
1 & \text{if } y \leq 1, \\
-|x| & \text{if } 0 < |x| |y| < 1 \\
1 - q(|xy|)(1 + |x|) & \text{otherwise} 
\end{cases}$$

where $q(t) = \frac{1}{2}(1 - \cos \pi t)$. Notice that $f \in C^1(\mathbb{R}^2)$, and that $f$ and $A = [1 0]$ are ADMM-feasible, meaning that $\arg\min_{w \in \mathbb{R}^2} \{f(w) + \frac{\beta}{2}\|Aw - s\|^2\} \neq 0$ for any $\beta > 0$ and $s \in \mathbb{R}$. However, $(Af)(s) = -|s|$ if $s \neq 0$ while $(Af)(0) = 1$, resulting in the lack of lsc at $s = 0$. Along $\ker A = \{0\} \times \mathbb{R}$, by keeping $x$ constant $f$ attains minimum at $\{x\} \times \{x^{-1}\}$ for $x \neq 0$, which escapes to infinity as $x \to 0$, and $f(x,x^{-1}) = -|x| \to 0$. However, if instead $x = 0$ is fixed (as opposed to $x \to 0$), then the pathology comes from the fact that $f(0, \cdot) \equiv 1 > 0$, which contradicts the condition imposed in Proposition 5.6(ii). This example also shows that $f \in C^1(\mathbb{R}^m)$ is not enough a requirement for $(Af)$ to be lsc.

The limit inferior is somehow related to the *asymptotic function*, defined as $f_\infty(d) := \lim\inf_{d \to \bar{d}, t \to \infty} \frac{f(td)}{t}$, see e.g., [1]. The referenced book provides other sufficient conditions based on the behavior of $f_\infty$ on $\ker A \setminus \{0\}$. Unlike ours, such
conditions ensure also nonemptiness of the set $X(s)$ for all $s \in \text{dom } f$, and are in this sense less general. To see this, it suffices to modify (5.3) as follows

$$f(x, y) = \begin{cases} 1 & \text{if } y \leq 1, \\ -|x| & \text{if } 0 < |x| y < 1, \\ e^{-x^2} - q(|xy|)(e^{-x^2} + |x|) & \text{otherwise.} \end{cases}$$

The sufficient condition dictated by Proposition 5.6(ii) is satisfied. In fact, the function $(Af)(s) = -|s|$ is lsc, however argmin$_w \{f(w) \mid Aw = 0\}$ is empty.

5.2.2. Smoothness of the image function. We now turn to the smoothness requirement of $(Bg)$. To this end, we introduce the notion of smoothness with respect to a matrix, as follows

**Definition 5.7** (Smoothness relative to a matrix). We say that a function $h : \mathbb{R}^n \to \mathbb{R}$ is smooth relative to a matrix $C \in \mathbb{R}^{p \times n}$, and we write $h \in C^2_{r}(\text{dom } h)$, if $h$ is differentiable on its domain and $\nabla h$ satisfies the following Lipschitz condition: there exist $L_{h,C}$ and $\sigma_{h,C}$ with $|\sigma_{h,C}| \leq L_{h,C}$ such that

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

whenever $\nabla h(x), \nabla h(y) \in \text{rng } C^T$.

This condition is similar to that considered in [13], where $\Pi_{\text{rng } B^T} \nabla g$ is required to be Lipschitz. The paper analyzes convergence of a proximal ADMM; standard ADMM can be recovered when matrix $B$ is invertible, in which case both conditions reduce to Lipschitz differentiability of $g$. In general, our condition applies to a smaller set of points only, as it can be verified with $g(x, y) = \frac{1}{2} x^2 y^2$ and $B = [1 0]$. In fact, $\Pi_{\text{rng } B^T} \nabla g(x, y) = \begin{pmatrix} 2y^2 \\ 0 \end{pmatrix}$ is not Lipschitz continuous; however, $\nabla g(x, y) \in \text{rng } A^T$ iff $xy = 0$, in which case $\nabla g \equiv 0$. Then, $g$ is smooth relative to $B$ with $L_{g,B} = 0$.

To better understand how this notion of regularity comes into the picture, notice that if $g$ is differentiable, then $\nabla g(x) \in \text{rng } B^T$ on some domain $U$ if there exists a differentiable function $q : \mathcal{H} \to \mathbb{R}$ such that $g(x) = q(Bx)$. Then, it is easy to verify that $g$ is smooth relative to $B$ if the local “reparametrization” $q$ is smooth (on its domain). From an a posteriori perspective, if $(Bg)$ is smooth (in the classical sense), then due to the relation $B^T \nabla (Bg)(Bz) = \nabla g(z)$ holding for $z \in \text{argmin}_{z : Bz = s} g(z)$ (cf. Prop. C.2), it is apparent that $q$ serves as $(Bg)$. Therefore, smoothness relative to $B$ is somewhat a minimal requirement for ensuring smoothness of $(Bg)$.

**Theorem 5.8** (Smoothness of $(Bg)$). Suppose that $g : \mathbb{R}^n \to \mathbb{R}$ and $B \in \mathbb{R}^{p \times n}$ satisfy Assumption II(ii). Suppose that there exists $\beta > 0$ such that $g + \frac{\beta}{2} \|B \cdot -s\|^2$ is level bounded for all $s \in \mathbb{R}^p$. Then, the image function $(Bg)$ is smooth on its (affine) domain, provided that either

(i) $g \in C^1_{r}(\mathbb{R}^n)$, in which case $L_{(Bg)} = L_{g,B}$ and $\sigma_{(Bg)} = \sigma_{g,B}$,

(ii) $g \in C^1_{r}(\mathbb{R}^n)$, and $Z(s) := \text{argmin} \{g(z) \mid Bz = s\}$ is single-valued and Lipschitz continuous with modulus $M$, in which case $L_{(Bg)} = L_{g,M^2}$ and $\sigma_{(Bg)} = \sigma_{g,M^2}$,

(iii) $g \in C^1_{r}(\mathbb{R}^n)$ is convex and $B$ full-row rank, in which case $L_{(Bg)} = \frac{\sigma_{g}}{\|B\|^2}$ and $\sigma_{(Bg)} = \sigma_{g}/\|B\|^2$.

**Proof.** See Appendix C. \qed
Notice that the condition in Theorem 5.8(ii) covers the case when $g \in C^1(\mathbb{R}^n)$ and $B$ has full column rank, in which case $M = \frac{1}{\sigma_+(B)}$. This is somehow trivial, since necessarily $(Bg)(s) = g \circ (B^T B)^{-1} B^T$ if $B$ has full column rank.

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-step sizes and relaxation parameters than was previously known. The results are tight when $\lambda \in (0, 2]$ or when the differentiable function is nonconvex, covering in particular classical (non-relaxed) DRS and PRS. 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 new 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.

References

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

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

[3] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, 2011.

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

[5] Jacek Bochnak, Michel Coste, and Marie-Françoise Roy. *Real Algebraic Geometry*. A Series of Modern Surveys in Mathematics. Springer Berlin Heidelberg, 2013.

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

[7] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122, January 2011.

[8] Jim Douglas and Henry H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society*, 82(2):421–439, 1956.

[9] Daniel Gabay. Chapter ix applications of the method of multipliers to variational inequalities. In Michel Fortin and Roland Glowinski, editors, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, volume 15 of *Studies in Mathematics and Its Applications*, pages 299–331. Elsevier, 1983.

[10] Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976.

[11] Roland Glowinski. *On Alternating Direction Methods of Multipliers: A Historical Perspective*, pages 59–82. Springer Netherlands, Dordrecht, 2014.

[12] Roland Glowinski and A. Marroco. 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(R2):41–76, 1975.

[13] Max L. N. Goncalves, Jefferson G. Melo, and Renato D. C. Monteiro. Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems. *ArXiv e-prints*, February 2017.
Lemma A.1. For a function \( h \in C^{1,1}(\text{dom } h) \) with affine domain, let \( L \geq L_h \) and \( \sigma \leq \min \{0, \sigma_h\} \) be such that \( L + \sigma > 0 \). Then, for all \( x, z \in \text{dom } h \)

\[
h(z) \geq h(x) + \langle \nabla h(x), z - x \rangle + \frac{L \sigma}{2(L + \sigma)} \|x - z\|^2 + \frac{1}{2(L + \sigma)} \|\nabla h(x) - \nabla h(z)\|^2.
\]

Proof. The function \( \psi := h - \frac{\sigma}{2} \|\cdot\|^2 \in C^{1,1}(\text{dom } h) \) is convex with \( L_\psi \leq L - \sigma \). In particular, using the same reasoning as in [23, Thm. 2.1.12] it is easy to verify that

\[
\langle \nabla h(z) - \nabla h(x), z - x \rangle \geq \frac{\sigma L}{\sigma + L} \|z - x\|^2 + \frac{1}{\sigma + L} \|\nabla h(z) - \nabla h(x)\|^2.
\]
Therefore, \( \psi(z) \geq \psi(x) + \langle \nabla \psi(x), z - x \rangle + \frac{1}{2\nu} \| \nabla \psi(z) - \nabla \psi(x) \|^2 \) for any \( x, z \in \text{dom } h = \text{dom } \psi \), that is,

\[
\begin{align*}
\| h(z) & \geq \frac{\sigma}{2} \| z \|^2 + h(x) - \frac{\sigma}{2} \| x \|^2 + \langle \nabla h(x) - \sigma x, z - x \rangle + \frac{1}{2\nu} \| \nabla h(z) - \nabla h(x) \|^2 \\
& + \frac{1}{2\nu} \| \nabla h(x) - \nabla h(x) - \sigma (z - x) \|^2 \\
= h(x) & + \langle \nabla h(x), z - x \rangle + \frac{\sigma}{2} \| z - x \|^2 + \frac{1}{2\nu} \| \nabla h(z) - \nabla h(x) \|^2 \\
& + \frac{1}{2\nu} \| \nabla h(z) - \nabla h(x) - \sigma (z - x) \|^2 \\
& \quad - \frac{\sigma}{2\nu} \| \nabla h(z) - \nabla h(x) - \sigma (z - x) \|^2 \\
\end{align*}
\]

and since \( \sigma \leq 0 \), by using (A.1),

\[
\begin{align*}
\geq h(x) & + \langle \nabla h(x), z - x \rangle + \frac{\sigma L}{2\nu} \| z - x \|^2 + \frac{1}{2\nu} \| \nabla h(z) - \nabla h(x) \|^2 \\
& - \frac{\sigma}{2\nu} \left( \frac{\sigma L}{2\nu} \| z - x \|^2 + \frac{1}{2\nu} \| \nabla h(z) - \nabla h(x) \|^2 \right) \\
= h(x) & + \langle \nabla h(x), z - x \rangle + \frac{\sigma L}{2\nu} \| z - x \|^2 + \frac{1}{2\nu} \| \nabla h(z) - \nabla h(x) \|^2. \\
\end{align*}
\]

Proposition A.2 (Proximal inequalities). Given a proper function \( h : \mathbb{R}^n \to \mathbb{R} \) and \( x \in \mathbb{R}^n \) let \( \bar{x} \in \text{prox}_{\alpha h}(x) \) for some \( \gamma > 0 \). Then, for all \( w \in \mathbb{R}^n \)

\[
\| h(w) + \frac{1}{2\tau} \| w - x \|^2 \geq h(\bar{x}) + \frac{1}{2\tau} \| \bar{x} - x \|^2 + \frac{1}{2\tau} \| w - x \|^2 + r(w, \bar{x})
\]

where either

(i) \( r(w, \bar{x}) = - \frac{1}{2\tau} \| \bar{x} - w \|^2 \), or

(ii) \( r(w, \bar{x}) = \frac{\gamma}{2} \| x - w \|^2 \) in case \( h = \frac{\gamma}{2} \cdot \) \( ||x||^2 \) is convex for some \( \gamma \in \mathbb{R} \).

Moreover, if \( h \in C^{1,1}(\text{dom } h) \) with \( \text{dom } h \) affine, and \( x \in \text{dom } h \), then for any \( L \geq L_0 \) and \( \sigma \leq \min \{ 0, \sigma_h \} \) such that \( L + \sigma > 0 \) we can also take

(iii) \( r(w, \bar{x}) = \frac{\sigma L}{2(\sigma + L)} \| \bar{x} - w \|^2 + \frac{1}{2(\sigma + L)} \| \nabla h(\bar{x}) - \nabla h(w) \|^2 \).

If \( h \) is additionally strongly convex, so that \( \sigma_h > 0 \), then for any \( \alpha \in [0, 1] \) we can take

(iv) \( r(w, \bar{x}) = (1 - \alpha) \frac{\sigma_h}{2} \| \bar{x} - w \|^2 + \alpha \frac{1}{2\sigma_h} \| \nabla h(\bar{x}) - \nabla h(w) \|^2. \)

Proof. A.2(i) is a trivial consequence of the characterization of \( \text{prox}_{\alpha h}(x) \). In the other cases, from the optimality conditions of \( x \) it follows that \( \frac{1}{2}(x - \bar{x}) \in \partial h(\bar{x}) \).

In particular, by combining Lem. A.1 with the standard inequality for \( \sigma \)-strongly convex functions (allowing here \( \sigma \leq 0 \)) and for convex smooth functions (see e.g., [23, Thm. 2.1.5]), we have \( h(w) \geq h(\bar{x}) + \frac{1}{2}(x - \bar{x}, w - \bar{x}) + r(w, \bar{x}) \) for all \( w \in \mathbb{R}^n \).

The claim now follows from the identity \( 2(x - \bar{x}, w - \bar{x}) = \| x - \bar{x} \|^2 + \| w - \bar{x} \|^2 - \| x - w \|^2 \).

Proposition A.3 (Proximal properties of smooth functions). Let \( h : \mathbb{R}^n \to \mathbb{R} \) be a function with affine domain and suppose that \( h \in C^{1,1}(\text{dom } h) \). Then, \( h \) is prox-bounded with \( \gamma_h \geq 1/|\sigma_h| \) and for all \( \gamma < 1/|\sigma_h| \) the following hold

(i) for all \( s \in \mathbb{R}^n \), having \( u = \text{prox}_{\gamma s_h}(s) \) is equivalent to having \( \nu \in \text{dom } h^\perp \) such that \( s = u + \gamma \nabla h(u) + \nu \);

(ii) \( \text{prox}_{\gamma h}(s) \) is \( \frac{1}{1 + \gamma \sigma_h} \)-strongly monotone on \( \text{dom } h \) and \( \frac{1}{1 + \gamma \sigma_h} \)-Lipschitz continuous everywhere. In particular, for all \( s, s' \in \text{dom } h \)

\[
\frac{1}{1 + \gamma \sigma_h} \| s - s' \| \leq \| u - u' \| \leq \frac{1}{1 + \gamma \sigma_h} \| s - s' \|
\]

where \( u = \text{prox}_{\gamma h}(s) \) and \( u' = \text{prox}_{\gamma h}(s') \).
(iii) $h^\gamma \in C^{1,1}(\mathbb{R}^n)$ with $\sigma_{h^\gamma} = \frac{\sigma_h}{1+\gamma \sigma_h}$ and $L_{h^\gamma} = \max\left\{ \frac{L_h}{1+\gamma L_h}, \frac{[\gamma]_+}{1+\gamma [\gamma]_+} \right\}$, with

$$\nabla h(s) = \frac{1}{\gamma}(s - \text{prox}_{\gamma h}(s)) \text{ and } \nabla h(\text{prox}_{\gamma h}(s)) = \frac{1}{\gamma}(\Pi_{\text{dom} h} s - \text{prox}_{\gamma h}(s)).$$

Proof. For all $\gamma \in (0, \frac{1}{[\sigma_h]_-})$ the function $h + \frac{1}{2}\|\cdot\|^2$ is strongly convex and as such lower bounded; it follows that $\gamma h \geq \frac{1}{[\sigma_h]_-}.\n
\text{\textbullet} A.3(i) For all $s \in \mathbb{R}^n$ we have that $u = \text{prox}_{\gamma h}(s)$ is the (unique) minimizer of the strongly convex function $h(w) + \frac{1}{2\gamma}\|w - s\|^2$, and is therefore characterized by

$$0 \in \partial h(u) + \frac{1}{\gamma}(u - s) + \partial \delta_{\text{dom} h}(u) = \nabla h(u) + \frac{1}{\gamma}(u - s) + \text{ dom } h^\gamma.$$

\text{\textbullet} A.3(ii) For all $\gamma < \frac{1}{[\sigma_h]_-}$ the function $h(u) = \frac{1}{2}\|u\|^2 + \gamma h(u)$ is $(1 + \gamma \sigma_h)$-strongly convex and $(1 + \gamma L_h)$-smooth over $\text{ dom } h$. Moreover, for $u = \text{prox}_{\gamma h}(s)$ and $u' = \text{prox}_{\gamma h}(s')$, since $s - s' \in \text{ dom } h^\gamma A.3(i)$ implies that $s - s' = \nabla h(u) - \nabla h(u').$ Then,

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

This shows Lipschitz continuity, and from [23, Thm. 2.1.5 and 2.1.10] we also have

$$\frac{1}{1+\gamma \lambda} \|s - s'\|^2 \leq \langle s - s', u - u' \rangle \leq \frac{1}{1+\gamma \sigma_n} \|s - s'\|^2,$$

(A.2)

the first inequality of which shows strong monotonicity.

\text{\textbullet} A.3(iii) From [26, Ex. 10.32] it follows that $h^\gamma$ is a strictly continuous function on $\mathbb{R}^n$ with $\partial h^\gamma(s) \subseteq \{(s - s')/\gamma\}$, where $u := \text{prox}_{\gamma h}(s).$ By invoking [26, Thm. 9.18] we conclude that $h^\gamma$ is everywhere differentiable with $\nabla h^\gamma(s) = \frac{1}{\gamma}(s - u).$

Moreover, $u$ minimizes $h + \frac{1}{2\gamma}\|s - u\|^2$ and in particular $\frac{1}{\gamma}(s - u) \in \partial h(u).$ Therefore, $\nabla h(u) = \Pi_{\text{dom} h} \partial h(u) = \frac{1}{\gamma}(\Pi_{\text{dom} h} s - u)$, where the last equality follows from the fact that $u \in \text{ dom } h.$ By using the first inequality in (A.2) we then obtain

$$\langle \nabla h^\gamma(s) - \nabla h^\gamma(s'), s - s' \rangle = \frac{1}{\gamma}(\|s - s'\|^2 - \langle s - s', u - u' \rangle) \leq \frac{1}{\gamma} (1 - \frac{\lambda}{1+\gamma \lambda}) \|s - s'\|^2 = \frac{L_\sigma}{1+\gamma \lambda} \|s - s'\|^2.$$

Similarly, by using also the second inequality in (A.2) we arrive to

$$\frac{\sigma_n}{1+\gamma \sigma_n} \|s - s'\|^2 \leq \langle \nabla h^\gamma(s) - \nabla h^\gamma(s'), s - s' \rangle \leq \frac{L_\sigma}{1+\gamma \lambda} \|s - s'\|^2$$

and the claimed values of $\sigma_{h^\gamma}$ and $L_{h^\gamma}$ follow. \hfill \Box

**Proposition A.4.** Let $h : \mathbb{R}^n \rightarrow \mathbb{R}$ and $\gamma > 0$. Then, $\text{prox}_{\gamma h}(x) = \text{prox}_{\gamma h}(x + \nu)$ for all $x \in \mathbb{R}^n$ and $\nu \in \mathbb{V}^\perp$, where $\mathbb{V}$ is the affine hull of $\text{ dom } h.$

Proof. Trivial. \hfill \Box

**Appendix B. Proof of Theorem 4.1**

We first prove a preliminary result.

**Lemma B.1.** Suppose that $\varphi_1$ and $\varphi_2$ comply with Assumption 1, and starting from any $s \in \mathbb{R}^p$ consider one DRS update $(s, u, v) \mapsto (s^+, u^+, v^+)$ for some feasible $\gamma$ and $\lambda > 0$. Then,

$$\varphi_{\gamma}(s) - \varphi_{\gamma}(s^+) \geq \frac{2\alpha}{L_{\varphi_1}}\|u - u^+\|^2 - \frac{2\alpha}{\lambda}\|y - y^+\|^2 + r(s, s^+)$$

(B.1)

where

$$r(s, s^+) = \begin{cases} (1 - \alpha)\frac{\sigma_2}{2\lambda}\|u - u^+\|^2 + \alpha\frac{1}{2L_{\varphi_1}}\|y - y^+\|^2 & \text{if } \sigma_{\varphi_1} > 0 \\ \frac{\sigma_1L}{2(\lambda + \sigma_{\varphi_1})}\|u - u^+\|^2 + \frac{1}{2L + \sigma_{\varphi_1}}\|y - y^+\|^2 & \text{otherwise} \end{cases}$$

with $\alpha \in [0, 1]$ and $L \geq L_{\varphi_1}$ such that $L + \sigma_{\varphi_1} > 0.$
Proof. Observe first that
\[ \varphi_{\gamma}^{\text{DR}}(s) = \varphi_1(u) + \varphi_2(v) + \frac{1}{2\gamma} \|u - v + \gamma y\|^2 - \frac{1}{2\gamma} \|v\|^2. \]  
\hbox{(B.2)}

By applying Prop. A.2 with \( w = u, \ x = v - \gamma y^{+}\) and \( \bar{x} = u^+ \) we obtain
\[ \varphi_1(u) + \frac{1}{2\gamma} \|u - v + \gamma y^{+}\|^2 \geq \varphi_1(u^+) + \frac{1}{2\gamma} \|u^+ - v + \gamma y^{+}\|^2 + \frac{1}{2\gamma} \|u^+ - u\|^2 + r \]
and again Prop. A.2(i) with \( w = u, \ x = u^+ + \gamma y^{+}\) and \( \bar{x} = v^+ \) we obtain
\[ \varphi_2(v) + \frac{1}{2\gamma} \|v - u^+ - \gamma y^{+}\|^2 \geq \varphi_2(v^+) + \frac{1}{2\gamma} \|v^+ - u^+ - \gamma y^{+}\|^2 \]
where \( r = r(s, s^+) \). Summing and invoking the expression (B.2) of the DRE yield
\[ \varphi_{\gamma}^{\text{DR}}(s) - \varphi_{\gamma}^{\text{DR}}(s^+) \geq \frac{3}{2} \|y^+\|^2 - \frac{3}{2\gamma} \|y\|^2 + \frac{2}{2\gamma} \|v - u - \gamma y\|^2 - \frac{1}{2\gamma} \|v - u^+ - \gamma y^{+}\|^2 + \frac{1}{2\gamma} \|u^+ - u\|^2 + r. \]

We now expand all the squares with three terms to arrive to
\[ \varphi_{\gamma}^{\text{DR}}(s) - \varphi_{\gamma}^{\text{DR}}(s^+) \geq -\langle u - v, y \rangle - \langle u - v, y^{+}\rangle + \langle v - u^+, y^+ \rangle 
+ \langle u^- - v, y^{+}\rangle + \frac{1}{2\gamma} \|u^+ - u\|^2 + r 
= \langle u - v, y - y^{+}\rangle + \langle v - u^+, y^+ - y^{+}\rangle + \frac{1}{2\gamma} \|u^+ - u\|^2 + r 
= \frac{1}{2\gamma} \|u - v\|^2 - \frac{1}{2\gamma} \|v - u^+\|^2 + \frac{1}{2\gamma} \|u^+ - u\|^2 + r 
= -\frac{1}{2\gamma} \|u - v\|^2 - \frac{1}{2\gamma} \|u^+ - u\|^2 - \frac{2}{\gamma} \langle v - u, u - u^+ \rangle + r. \]

By replacing \( u - v = \frac{1}{\gamma} (s - s^+) = \frac{1}{\gamma} (u - u^+) = -\frac{1}{\gamma} (y - y^+) \), after simple algebra the claimed inequality follows.

For the sake of simplicity, we split the proof of Theorem 4.1 into two cases, depending on whether the smooth function \( \varphi_1 \) is strongly convex or not. In both cases, we will rather show that \( \varphi_{\gamma}^{\text{DR}}(s) - \varphi_{\gamma}^{\text{DR}}(s^+) \geq \bar{c} \|u^+ - u\|^2 \), for some \( \bar{c} > 0 \) and Theorem 4.1 will follow from the inequality
\[ \|u^+ - u\|^2 \geq \left( \frac{\lambda}{1+\lambda L_{\varphi_1}} \right)^2 \|s^+ - s\|^2 = \left( \frac{\lambda}{1+\lambda L_{\varphi_1}} \right)^2 \|y - y^+\|^2 \]  
\hbox{(B.3)}

due to Proposition A.3(ii), so that one can consider \( c = \frac{1}{1+\gamma L_{\varphi_1}} \bar{c} \).

**Theorem B.2** (Sufficient decrease – Non-strongly convex case). Suppose that \( \sigma_{\varphi_1} \leq 0 \). Starting from \( s \in \mathbb{R}^p \) consider one DRS update \( (s, u, v) \rightarrow (s^+, u^+, v^+) \) for some \( \lambda \in (0, 2) \) and some feasible \( \gamma < \min \left\{ \frac{1}{L_{\varphi_1}}, \frac{2\lambda - \lambda}{2\lambda - 2\gamma L_{\varphi_1}} \right\} \). Then, \( \varphi_1(s) - \varphi_1(s^+) \geq \bar{c} \|u - u^+\|^2 \) where \( \bar{c} \) is a positive constant defined as
\[ \bar{c} = \frac{2\lambda - \lambda}{2\lambda - 2\gamma L_{\varphi_1}} \min \left\{ \frac{p \lambda}{1+\gamma L_{\varphi_1}}, \lambda - 2\gamma L_{\varphi_1} \right\} \]  
if \( 0 < \lambda < 2(1 + p_{\varphi_1}) \), otherwise.

**Proof.** From Lem. B.1 it follows that for any feasible \( \gamma \) and any \( L \geq L_{\varphi_1} \) such that \( L + \sigma_{\varphi_1} > 0 \) we have
\[ \varphi_1(s) - \varphi_1(s^+) \geq \left( \frac{2\lambda - \lambda}{2\lambda - 2\gamma L_{\varphi_1}} \right) \|u^+ - u\|^2 + \frac{1}{2\gamma} \left( \frac{1}{2(1+p)} - \frac{\xi}{\lambda} \right) \|y^+ - y^+\|^2 \]
where \( \xi := \gamma L \) and \( p := \sigma_{\varphi_1}/L \in (-1, 0) \). Since \( \|y^+ - y^+\|^2 = \|\bar{\varphi}_1(u^+) - \bar{\varphi}_1(u)\|^2 \leq L_{\varphi_1} \|u^+ - u\|^2 \), the claim holds provided that
\[ \frac{\bar{c}}{L} = \left\{ \begin{array}{ll} \frac{2\lambda - \lambda}{2\lambda - 2\gamma L_{\varphi_1}} + \frac{p}{2(1+p)} & \text{if } 0 < \frac{1}{2(1+p)} - \frac{\xi}{\lambda} \text{ otherwise.} \end{array} \right. \]  
\hbox{(B.4)}

Now, let \( \lambda \in (0, 2) \) be fixed and let us consider two cases:
we conclude that that is, and with simple algebra we arrive to

\[
\ell = \frac{2 - \lambda}{2\lambda\gamma} + \begin{cases} 
\frac{\sigma_{\varphi_1}}{2(1 + p_{\varphi_1})} & \text{if } \frac{\lambda}{\sigma_{\varphi_1}} \leq \frac{L_{\varphi_1}}{\lambda} \\
\frac{L_{\varphi_1}}{2} - \frac{\gamma L_{\varphi_1}^2}{\lambda} & \text{otherwise}.
\end{cases}
\]

By imposing \( \ell < 0 \) we obtain the following conditions on \( \gamma \) and \( \lambda \):

\[
\begin{cases} 
0 < \frac{2 - \lambda}{2\lambda\gamma} + \frac{\sigma_{\varphi_1}}{2(1 + p_{\varphi_1})} & \text{if } \frac{\lambda}{\sigma_{\varphi_1}} \leq \frac{L_{\varphi_1}}{\lambda} \\
\gamma < \frac{2 - \lambda}{\lambda + \sigma_{\varphi_1}} & \text{otherwise}
\end{cases}
\]

and since we are considering the case \( \sigma_{\varphi_1} \geq -\frac{2 - \lambda}{2\lambda} L_{\varphi_1} \), this reduces to

\[
\gamma < \frac{2(1 + p_{\varphi_1})}{L_{\varphi_1} + \sigma_{\varphi_1}} \quad \text{and} \quad \gamma < \frac{1}{L_{\varphi_1}}
\]

that is, \( \gamma < \frac{1}{L_{\varphi_1}} \). As to the coefficient \( \ell \), since

\[
\frac{2 - \lambda}{2\lambda\gamma} + \frac{\sigma_{\varphi_1}}{2(1 + p_{\varphi_1})} \leq \frac{2 - \lambda}{2\lambda\gamma} + \frac{L_{\varphi_1}}{2} - \frac{\gamma L_{\varphi_1}^2}{\lambda} \quad \Leftrightarrow \quad \gamma \leq \frac{\lambda}{2(1 + p_{\varphi_1})},
\]

we conclude that \( \ell = \frac{2 - \lambda}{2\lambda\gamma} + \min \left\{ \frac{\sigma_{\varphi_1}}{2(1 + p_{\varphi_1})}, \frac{L_{\varphi_1}}{2} - \frac{\gamma L_{\varphi_1}^2}{\lambda} \right\} \)

\[\text{Case 2: } \lambda > 2(1 + p_{\varphi_1}).\]

Then, \( \sigma_{\varphi_1} < -\frac{2 - \lambda}{2\lambda} L_{\varphi_1} \) and we can take \( L = -\frac{2\sigma_{\varphi_1}}{2 - \lambda} > L_{\varphi_1} \). Therefore, \( p = -\frac{2 - \lambda}{2} \)
and \( \xi = -\frac{2\sigma_{\varphi_1}}{2 - \lambda} \gamma \), and (B.4) becomes

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

By imposing \( \ell > 0 \) this reduces to

\[
\begin{cases} 
0 < \frac{2 - \lambda}{2\lambda\gamma} + \frac{\sigma_{\varphi_1}}{\lambda} & \text{if } \gamma < \frac{2 - \lambda}{2\sigma_{\varphi_1}} \\
\gamma < \frac{2 - \lambda}{\sigma_{\varphi_1}} & \text{otherwise}
\end{cases}
\]

and with simple algebra we arrive to

\[
\gamma < \frac{2 - \lambda}{2\sigma_{\varphi_1}} \quad \text{and} \quad \frac{2 - \lambda}{2\sigma_{\varphi_1}} \leq \gamma < \frac{2 - \lambda}{2\sigma_{\varphi_1}}
\]

that is, \( \gamma < \frac{2 - \lambda}{2\sigma_{\varphi_1}} \). From (B.5) it follows that \( \ell = \frac{2 - \lambda}{2\lambda\gamma} + \frac{\sigma_{\varphi_1}}{\lambda} \)

The proof of the first part of Theorem 4.1 now follows from (B.3) by observing that the constant \( c \) in the statement of Theorem 4.1 satisfies 0 < \( c \leq \left( \frac{\lambda}{1 + \gamma L_{\varphi_1}} \right)^2 \ell \).

**Theorem B.3** (Sufficient decrease – Strongly convex case). Suppose that \( \sigma_{\varphi_1} > 0 \). Starting from any \( s \in \text{dom} \varphi_1 \), consider one DRS update \( (s, u, v) \mapsto (s^+, u^+, v^+) \) for some feasible \( \gamma < \frac{1}{L_{\varphi_1}} \) and \( \lambda > 0 \). Then,

\[
\varphi_1(s) - \varphi_1(s^+) \geq \ell \|u - u^+\|^2
\]

for \( \ell = \frac{2 - \lambda}{2\lambda\gamma} + \sigma_{\varphi_1} \left( \frac{1}{2} - \frac{\gamma L_{\varphi_1}^2}{\lambda} \right) > 0 \), provided that

\[
p_{\varphi_1} \frac{\lambda - \delta}{4\sigma_{\varphi_1}} < \gamma < p_{\varphi_1} \frac{\lambda + \delta}{4\sigma_{\varphi_1}} \quad \text{and} \quad 2 \leq \lambda < \frac{4}{1 - \sqrt{p_{\varphi_1}}}
\]

where \( \delta := \sqrt{p_{\varphi_1} \lambda^2 - 8 p_{\varphi_1}(\lambda - 2)} \) is a strictly positive constant.
Proof. Suppose \( \lambda \geq 2 \) and let \( \xi := \gamma L_{\varphi_1} \). From Lem. B.1 it follows that for all \( \alpha \in [0,1] \)
\[
\varphi_{\lambda}(s) - \varphi_{\lambda}(s^*) \geq \left( \frac{2-\lambda}{2\alpha} + \frac{1}{2\alpha} \right) \|u - u^+\|^2 + \frac{1}{\sqrt{2}} \left( \frac{\alpha}{\sqrt{2}} - \lambda \right) \|y - y^+\|^2.
\]
Since \( \lambda \geq 2 \), by imposing \( \xi < 1 \) we can make the coefficient of \( \|y - y^+\|^2 \) vanish by taking \( \alpha = 2\xi/\lambda \in (0,1) \). Therefore, we obtain \( \frac{\alpha}{\sqrt{2}} \geq \frac{\lambda - 2\xi}{2\lambda} \frac{|p_{\varphi_1}|}{\|v_{\varphi_1}\|} \). By imposing \( \bar{c} > 0 \) we are left with the condition
\[
\frac{p_{\varphi_1}}{4\|v_{\varphi_1}\|} < \gamma < \frac{p_{\varphi_1}}{4\|v_{\varphi_1}\|} \quad \text{and} \quad \left( 2 \leq \lambda < \frac{4}{1+\sqrt{1-p_{\varphi_1}}} \lor \lambda > \frac{4}{1+\sqrt{1-p_{\varphi_1}}} \right)
\]
where \( \delta = \sqrt{(p_{\varphi_1})^2 - 8p_{\varphi_1}(\lambda - 2)} \). However, since \( \frac{p_{\varphi_1}}{4\|v_{\varphi_1}\|} \geq \frac{1}{L_{\varphi_1}} \) for \( \lambda \geq 4 \), it follows that the lower bound \( \lambda > \frac{4}{1+\sqrt{1-p_{\varphi_1}}} \) is to be discarded.

\[\square\]

Appendix C. Proofs of Section 5

We first prove two useful results.

Proposition C.1. Let \( h : \mathbb{R}^n \to \mathbb{R} \) be a proper lsc function and \( C \in \mathbb{R}^{p \times n} \). Suppose there exists \( \beta \geq 0 \) such that the set \( X_\beta(s) := \arg\min_x \left\{ h(x) + \frac{\beta}{2} \| Cx - s \|^2 \right\} \)

is nonempty for all \( \beta > \beta \) and \( s \in \mathbb{R}^p \). Then,

(i) the image function \( (Ch) \) is proper;

(ii) \( (Ch)(C\beta) = h(x_\beta) \) whenever \( x_\beta \in X_\beta(s) \) for some \( \beta > \beta \) and \( s \in \mathbb{R}^p \);

(iii) \( CX_\beta(s) = \text{prox}_{(C\beta)\delta}(s) \) for all \( \beta > \beta \) and \( s \in \mathbb{R}^p \).

Proof. \(
\star\ C.1(i).\) Let \( \tilde{s} \in \text{dom} \ h \) and \( \beta > \beta \) be fixed. Then,

\[
-\infty < \min_{x \in \partial h \tilde{s}} \left\{ h(x) + \frac{\beta}{2} \| Cx - \tilde{s} \|^2 \right\} \leq \inf_{x : Cx = \tilde{s}} \left\{ h(x) + \frac{\beta}{2} \| Cx - \tilde{s} \|^2 \right\} = (Ch)(\tilde{s})
\]

and since \( (Ch)(\tilde{s}) < \infty \) for all \( \tilde{s} \in \text{dom} \ h \neq \emptyset \) we conclude that \( (Ch) \) is proper.

\(\star\ C.1(ii).\) Since \( C(x_\beta + v) = Cx_\beta \) iff \( v \in \text{ker} \ C \), for all \( s \in \mathbb{R}^p \) and \( x_\beta \in X_\beta(s) \) necessarily \( h(x_\beta) \leq h(x_\beta + v) \). Consequently,

\[
(Ch)(C\beta) \leq h(x_\beta) \leq \inf_{x : Cx = Cx_\beta} h(x_\beta + v) = \inf_{x : Cx = Cx_\beta} h(x) = (Ch)(C\beta).
\]

\(\star\ C.1(iii).\) Fix \( \beta > \beta \) and \( \tilde{s} \in \mathbb{R}^p \), and let \( x_\beta \in X_\beta(\tilde{s}) \). Then, from \( C.1(ii) \) and the optimality of \( x_\beta \) we have

\[
(Ch)(C\beta) + \frac{\beta}{2} \| Cx_\beta - \tilde{s} \|^2 = h(x_\beta) + \frac{\beta}{2} \| Cx_\beta - \tilde{s} \|^2 \leq h(x) + \frac{\beta}{2} \| Cx - \tilde{s} \|^2
\]

for all \( x \in \mathbb{R}^n \). In particular, this holds for all \( s \in \mathbb{R}^p \) and \( x \) such that \( Cx = s \), hence

\[
(Ch)(C\beta) + \frac{\beta}{2} \| Cx_\beta - \tilde{s} \|^2 \leq \inf_{x : Cx = s} \left\{ h(x) + \frac{\beta}{2} \| Cx - \tilde{s} \|^2 \right\} = (Ch)(s) + \frac{\beta}{2} \| s - \tilde{s} \|^2
\]

proving \( CX_\beta(s) \subseteq \text{prox}_{(C\beta)\delta}(s) \). Similar reasonings yield the other inclusion too.

\[\square\]

Proposition C.2. Let \( h : \mathbb{R}^n \to \mathbb{R} \) and \( C \in \mathbb{R}^{p \times n} \) be such that \( (Ch)(C\bar{x}) = h(\bar{x}) \) for some \( \bar{x} \in \mathbb{R}^n \). Then, \( C^\top \hat{\partial}(Ch)(C\bar{x}) \subseteq \hat{\partial}h(\bar{x}) \).

Proof. Let \( \hat{v} \in \hat{\partial}(Ch)(C\bar{x}) \). Then, for all \( x \in \mathbb{R}^n \)

\[
h(x) - h(\bar{x}) - (C^\top \hat{v}, x - \bar{x}) \geq (Ch)(C\bar{x}) - (Ch)(C\bar{x}) - \langle \hat{v}, C(x - \bar{x}) \rangle \geq o(\|C(x - \bar{x})\|)
\]

Therefore, \( \liminf_{x \to \bar{x}, x \neq \bar{x}} \frac{h(x) - h(\bar{x}) - (C^\top \hat{v}, x - \bar{x})}{\|x - \bar{x}\|} \geq 0 \), proving that \( C^\top \hat{v} \in \hat{\partial}h(\bar{x}) \).

\[\square\]
Proof of Proposition 5.6. Properness is shown in Prop. C.1. Suppose now that $(s_k)_{k \in \mathbb{N}} \subseteq \text{lev}_{\leq \alpha} f$ for some $\alpha \in \mathbb{R}$ and that $s_k \to \tilde{s}$. Then, due to the characterization of [26, Thm. 1.6] it suffices to show that $\tilde{s} \in \text{lev}_{\leq \alpha} f$.

\* 5.6(ii). The assumption ensures the existence of a bounded sequence $(x_k)_{k \in \mathbb{N}}$ such that eventually $Ax_k = s_k$ and $(Af)(s_k) = f(x_k)$. By possibly extracting, $x_k \to \bar{x}$ and necessarily $Ax = \bar{s}$. Then,

$$\tag{5.6(i)} f(\bar{s}) \leq f(\bar{x}) \leq \liminf_{k \to \infty} f(x_k) = \liminf_{k \to \infty} (Af)(s_k) \leq \alpha,$$

hence $\bar{s} \in \text{lev}_{\leq \alpha} f$.

Proof of Theorem 5.8 (Smoothness of $(Bg)$). That $(Bg)$ is proper follows from Prop. C.1(i). The level boundedness condition ensures that for all $\alpha \in \mathbb{R}$ and $s \in B \text{ dom } g$ the set $\{z \mid g(x) \leq \alpha, \|Bz - s\| < \varepsilon\}$ is bounded for some $\varepsilon > 0$ (in fact, for all $\varepsilon > 0$). Then, we may invoke [26, Thm. 1.32] to infer that $(Bg)$ is lsc, that the set $Z(s) : = \text{argmin}_{z} \{g(z) \mid Bz = s\}$ is nonempty for all $s \in \text{dom}(Bg)$, and that the function $H(z, s) : = g(z) + \delta_{\{0\}}(Bz - s)$ is uniformly level bounded in $z$ locally uniformly in $s$, in the sense of [26, Def. 1.16].

\* 5.8(i) and 5.8(ii). Observe that

$$\partial^\infty H(z, Bz) = \left\{ (\nu - B^T y) \mid \nu \in \text{ dom } g^\perp, y \in \mathbb{R}^p \right\}$$

for all $z \in \text{ dom } g$. We may then apply [26, Cor. 10.14] to arrive to

$$\partial^\infty (Bg)(s) \subseteq \bigcup_{z \in Z(s)} \{y \mid (0, y) \in \partial^\infty H(z, s)\} = \{y \mid B^T y \in \text{ dom } g^\perp\} = (\text{ dom}(Bg))^\perp$$

for all $s \in \text{ dom}(Bg) = B \text{ dom } g$. It follows that $\partial^\infty (Bg)|_{\text{ dom}(Bg)}(s) = \{0\}$, and by virtue of [26, Thm. 9.13] from the arbitrariness of $s \in \text{ dom}(Bg)$ we conclude that $(Bg)$ is strictly continuous on its domain and has nonempty subdifferential.

Now, fix $s_i \in \text{ dom}(Bg) = \text{ rng } B$ and $y_i \in \partial (Bg)(s_i)$, $i = 1, 2$. It follows from Prop. C.2 and (2.1) that $B^T y_i = \nabla g(z_i) + \nu_i$ for some $\nu_i \in \text{ dom } g^\perp$ and $z_i \in Z(s_i)$, $i = 1, 2$. Moreover, from the optimality conditions $z_i \in Z(s_i) = \text{ argmin } g + \delta_{\{0\}}(B - s_i)$, we obtain $\nabla g(z_i) \in \text{ rng } B^T$ and $Bz_i = s_i$, $i = 1, 2$. Therefore,

$$\langle y_1 - y_2, s_1 - s_2 \rangle = (y_1 - y_2, Bz_1 - Bz_2) = (B^T y_1 - B^T y_2, z_1 - z_2)$$

$$= (\nabla g(z_1) + \nu_1 - \nabla g(z_2) - \nu_2, z_1 - z_2)$$

$$= (\nabla g(z_1) - \nabla g(z_2), z_1 - z_2).$$

If 5.8(i) holds, since $\nabla g(z_i) \in \text{ rng } B^T$, $i = 1, 2$, smoothness of $g$ relative to $B$ implies

$$\sigma_{g, B} \langle s_1 - s_2 \rangle = \sigma_{g, B} \|B(x_1 - x_2)\|^2$$

$$\leq \langle y_1 - y_2, s_1 - s_2 \rangle \leq L_{g, B}\|Bz_1 - Bz_2\|^2 = L_{g, B}\|s_1 - s_2\|^2$$

$$\leq \langle y_1 - y_2, s_1 - s_2 \rangle$$

Therefore,

$$\sigma_{g, B} \langle s_1 - s_2 \rangle \leq L_{g, B}\|s_1 - s_2\|^2.$$
for all $s_i \in \text{dom}(Bg)$ and $y_i \in \partial(Bg)(s_i)$, $i = 1, 2$. Otherwise, if 5.8(ii) holds, then
\[ \sigma_y \| z_1 - z_2 \|^2 \leq (y_1 - y_2, s_1 - s_2) \leq L_g \| z_1 - z_2 \|^2 \]
and from the bound
\[ \frac{1}{\|B\|} \| s_1 - s_2 \| \leq \| z_1 - z_2 \| \leq M \| s_1 - s_2 \| \]
we obtain
\[ \sigma(Bg) \| s_1 - s_2 \|^2 \leq (y_1 - y_2, s_1 - s_2) \leq L(Bg) \| s_1 - s_2 \|^2 \]
with the constants $\sigma(Bg)$ and $L(Bg)$ as in the statement.

Either way, for all $s_i$ there is at most one subgradient $y_i \in \partial(Bg) \big|_{\text{dom}(Bg)}$, and from [26, Thm. 9.18] we conclude that necessarily such is $y_i = \nabla(Bg)(s_i)$. In particular, the inequalities also show that $(Bg) \in C^{1,1}(\text{dom}(Bg))$ with moduli as in the statement.

\[ 5.8(iii). \] We know from [3, Prop.s 12.34(ii)] and [17, Thm. D.4.5.1] that $(Bg)$ is a convex function and that $\partial(Bg)(s) = \{ y \mid B^T y = \nabla g(z) \}$ where $z \in Z(s)$. Since $B^T$ is injective and $\partial(Bg)(s) \neq \emptyset$, it follows that $(Bg)$ is differentiable. Using Fenchel duality, the claim will follow once we show that $(Bg)^*$ is strongly convex with modulus $\frac{\sigma_r(B^T B)}{L_g}$ and smooth with modulus $\| B \|^2 / \sigma_y$ if $\sigma_y > 0$.

Let $s_i = \nabla(Bg)^*(y_i)$, $i = 1, 2$. In particular, $y_i = \nabla(Bg)(s_i)$ and $B^T y_i = \nabla g(z_i)$ where $Bz_i = s_i$. We have
\[ (s_1 - s_2, y_1 - y_2) = (Bz_1 - Bz_2, y_1 - y_2) = (z_1 - z_2, B^T y_1 - B^T y_2) \]
\[ \geq \frac{1}{L_g} \| B^T y_1 - B^T y_2 \|^2 \geq \frac{\sigma(B^T B)}{L_g} \| B_{\text{rng}}(y_1 - y_2) \|^2 \]
\[ = \frac{\sigma(B^T B)}{L_g} \| y_1 - y_2 \|^2 \]
where the first inequality follows from [23, Thm. 2.1.5], the second is a well known property (see e.g., [13, Lem. A.2]), and the last equality is due to the fact that $\text{rng} B = \mathbb{R}^n$. These inequalities prove the sought properties. ☐