Douglas-Rachford splitting for nonconvex feasibility problems

Guoyin Li * Ting Kei Pong †

Jan 23, 2015

Abstract

We adapt the Douglas-Rachford (DR) splitting method to solve nonconvex feasibility problems by studying this method for a class of nonconvex optimization problem. While the convergence properties of the method for convex problems have been well studied, far less is known in the nonconvex setting. In this paper, for the direct adaptation of the method to minimize the sum of a proper closed function g and a smooth function f with a Lipschitz continuous gradient, we show that if the step-size parameter is smaller than a computable threshold and the sequence generated has a cluster point, then it gives a stationary point of the optimization problem. Convergence of the whole sequence and a local convergence rate are also established under the additional assumption that f and g are semi-algebraic. We also give simple sufficient conditions guaranteeing the boundedness of the sequence generated. We then apply our nonconvex DR splitting method to finding a point in the intersection of a closed convex set C and a general closed set D by minimizing the square distance to C subject to D. We show that if either set is bounded and the step-size parameter is smaller than a computable threshold, then the sequence generated from the DR splitting method is actually bounded. Consequently, the sequence generated will have cluster points that are stationary for an optimization problem, and the whole sequence is convergent under an additional assumption that C and D are semi-algebraic. We achieve these results based on a new merit function constructed particularly for the DR splitting method. Our preliminary numerical results indicate that the DR splitting method usually outperforms the alternating projection method in finding a sparse solution of a linear system, in terms of both the solution quality and the number of iterations taken.

1 Introduction

Many problems in diverse areas of mathematics, engineering and physics aim at finding a point in the intersection of two closed sets. This problem is often called the feasibility problem. Many practical optimization problems and reconstruction problems can be cast in this framework. We refer the readers to the comprehensive survey [7] and the recent monograph [8] for more details.

The Douglas-Rachford (DR) splitting method is an important and powerful algorithm that can be applied to solving problems with competing structures, such as finding a point in the intersection of two closed convex sets (feasibility problem), or, more generally, minimizing the sum of two proper closed convex functions. The latter problem is more general because the feasibility problem can be viewed as a minimization problem that minimizes the sum of the indicator functions of the two intersecting sets. In typical applications, the projection onto each of the constituent sets is

*Department of Applied Mathematics, University of New South Wales, Sydney 2052, Australia. E-mail: g.li@unsw.edu.au. This author was partially supported by a research grant from Australian Research Council.

†Department of Applied Mathematics, the Hong Kong Polytechnic University, Hong Kong. This author was supported partly by a research grant from Hong Kong Polytechnic University. E-mail: tk.pong@polyu.edu.hk.

simple to compute in the feasibility problem, and the so-called proximal operator of each of the constituent functions is also easy to compute in the case of minimizing the sum of two functions. Since these simple operations are usually the main computational parts of the DR splitting method, the method can be implemented efficiently in practice.

The DR splitting method was originally introduced in [14] to solve nonlinear heat flow problems, which aims at finding a point in the intersection of two closed sets in a Hilbert space. Later, Lions and Mercier [21] showed that the DR splitting method converges for two closed convex sets with nonempty intersection. This scheme was examined further in [15] again in the convex setting, and its relationship with another popular method, the proximal point algorithm, was revealed and explained therein. Recently, the DR splitting method has also been applied to various optimization problems that arise from signal processing and other applications, where the objective is the sum of two proper closed convex functions; see, for example, [13,16,18]. We refer the readers to the recent exposition [8] and references therein for a discussion about convergence in the convex setting.

While the behavior of the DR splitting method has been moderately understood in the convex cases, the theoretical justification is far from complete when the method is used in the nonconvex setting. Nonetheless, the DR splitting method has been applied very successfully to various important problems where the underlying sets are not necessarily convex [2,3]. This naturally motivates the following research direction:

Understand the DR splitting method when applied to possibly nonconvex sets.

As commented in [22], the DR splitting method is notoriously difficult to analyze compared with other projection type methods such as the alternating projection method [6,7,12,19]. Despite its difficulty, there has been some important recent progress towards understanding the behavior of the DR splitting method in the nonconvex setting. For example, it was shown in [22] that the DR splitting method exhibits local linear convergence for an affine set and a super-regular set (an extension of convexity which emphasizes local features), under suitable regularity conditions. There are also recent advances in dealing with specific structures such as the case where the two sets are finite union of convex sets [9], and the sparse feasibility problem where one seeks a sparse solution of a linear system [23]. On the other hand, in spite of the various local convergence results, global convergence of the method was only established in [1] for finding the intersection of a line and a circle.

In this paper, we approach the above basic problem from a new perspective. Recall that the alternating projection method for finding a point in the intersection of a closed convex set C and a closed set D can be interpreted as an application of the proximal gradient algorithm to the optimization problem

$$\min_{x \in D} \ \frac{1}{2} d_C^2(x) \tag{1}$$

with step-length equals 1, where $d_C(x)$ is the distance from x to C and $x \mapsto d_C^2(x)$ is smooth since C is convex. Motivated by this, we adapt the DR splitting method to solve the above optimization problem instead, which is conceivably easier to analyze due to the smooth objective. This is different from the common approach in the literature (see, for example, [1–3, 9, 23]) where the DR splitting method is applied to minimizing the sum of indicator functions of the intersecting sets. Notice that the feasibility problem is solved when the globally optimal value of (1) is zero.

In our analysis, we start with a more general setting: minimizing the sum of a smooth function f with Lipschitz continuous gradient, and a proper closed function g. We show that, if the step-size parameter is smaller than a computable threshold and the sequence generated from the DR splitting method has a cluster point, then it gives a stationary point of the optimization problem $\min_{x\in\mathbb{R}^n}\{f(x)+g(x)\}$. Moreover, under the additional assumption that f and g are semi-algebraic, we show convergence of the whole sequence and give a local convergence rate. In addition, we also give simple sufficient conditions guaranteeing the boundedness of the sequence generated, and hence the existence of cluster points. Our analysis relies heavily on the so-called Douglas-Rachford merit function (see Definition 2) we introduce, which is non-increasing along the sequence generated by the DR splitting method when the step-size parameter is chosen small enough.

We then apply our nonconvex DR splitting method to minimizing (1), whose objective is smooth with a Lipschitz continuous gradient. When the step-size parameter is smaller than a computable threshold and either set is compact, we show that the sequence generated from the DR splitting method is bounded. Thus, cluster points exist and they are stationary for (1). Furthermore, if C and D are in addition semi-algebraic, we show that the whole sequence is convergent. Finally, we perform numerical experiments to compare our method against the alternating projection method on finding a sparse solution of a linear system. Our preliminary numerical results show that the DR splitting method usually outperforms the alternating projection method, in terms of both the number of iterations taken and the solution quality.

The rest of the paper is organized as follows. We present notation and preliminary materials in Section 2. The DR splitting method applied to minimizing the sum of a smooth function f with Lipschitz continuous gradient and a proper closed function g is analyzed in Section 3, while its application to a nonconvex feasibility problem is discussed in Section 4. Numerical simulations are presented in Section 5. In Section 6, we present some concluding remarks.

2 Notation and preliminaries

We use $\mathbb{R}^n$ to denote the n-dimensional Euclidean space, $\langle \cdot, \cdot \rangle$ to denote the inner product and $\| \cdot \|$ to denote the norm induced from the inner product. For an extended-real-valued function f, the domain of f is defined as $\mathrm{dom} f := \{x \in \mathbb{R}^n : f(x) < +\infty\}$. The function is called proper if $\mathrm{dom} f \neq \emptyset$ and it is never $-\infty$. The function is called closed if it is lower semicontinuous. For a proper function $f : \mathbb{R}^n \to \overline{\mathbb{R}} := (-\infty, \infty]$, let $z \xrightarrow{f} x$ denote $z \to x$ and $f(z) \to f(x)$. Our basic *subdifferential* of f at $x \in \mathrm{dom}\, f$ (known also as the limiting subdifferential) is defined by

$$\partial f(x) := \left\{ v \in \mathbb{R}^n : \ \exists x^t \xrightarrow{f} x, \ v^t \to v \ \text{ with } \liminf_{z \to x^t} \frac{f(z) - f(x^t) - \langle v^t, z - x^t \rangle}{\|z - x^t\|} \geq 0 \text{ for each } t \right\}. \tag{2}$$

The above definition gives immediately the following robustness property:

$$\left\{ v \in \mathbb{R}^n : \ \exists x^t \xrightarrow{f} x, \ v^t \to v \ , v^t \in \partial f(x^t) \right\} \subseteq \partial f(x). \tag{3}$$

We also use the notation $\mathrm{dom}\, \partial f := \{x \in \mathbb{R}^n : \ \partial f(x) \neq \emptyset\}$. The subdifferential (2) reduces to the derivative of f denoted by ∇f if f is continuously differentiable. On the other hand, if f is convex, the subdifferential (2) reduces to the classical subdifferential in convex analysis (see, for example, [25, Proposition 8.12]), i.e.,

$$\partial f(x) = \{ v \in \mathbb{R}^n : \ \langle v, z - x \rangle \leq f(z) - f(x) \ \ \forall \ z \in \mathbb{R}^n \}.$$

For a function f with several groups of variables, we write $\partial_x f$ (resp., $\nabla_x f$) for the subdifferential (resp., derivative) of f with respect to the variable x. Finally, we say a function f is a strongly convex function with modulus $\omega > 0$ if $f - \frac{\omega}{2} \| \cdot \|^2$ is a convex function.

For a closed set $S \subseteq \mathbb{R}^n$, its indicator function δ_S is defined by

$$\delta_S(x) = \begin{cases} 0 & \text{if } x \in S, \\ +\infty & \text{if } x \notin S. \end{cases}$$

Moreover, the normal cone of S at $x \in S$ is given by $N_S(x) = \partial\delta_S(x)$. Furthermore, we use $\mathrm{dist}(x, S)$ or $d_S(x)$ to denote the distance from x to S, i.e., $\inf_{y\in S} \|x - y\|$. If a set S is closed and convex, we use $P_S(x)$ to denote the projection of x onto S.

A semi-algebraic set $S \subseteq \mathbb{R}^n$ is a finite union of sets of the form

$$\{x \in \mathbb{R}^n : h_1(x) = \cdots = h_k(x) = 0, g_1(x) < 0, \ldots, g_l(x) < 0\},$$

where $g_1, \ldots, g_l$ and $h_1, \ldots, h_k$ are real polynomials. A function $F : \mathbb{R}^n \to \mathbb{R}$ is semi-algebraic if the set $\{(x, F(x)) \in \mathbb{R}^{n+1} : \ x \in \mathbb{R}^n\}$ is semi-algebraic. Semi-algebraic sets and semi-algebraic functions can be easily identified and cover lots of possibly nonsmooth and nonconvex functions that arise in real world applications [4, 5, 10].

We will also make use of the following Kurdyka-Łojasiewicz (KL) property that holds in particular for semi-algebraic functions.

Definition 1. (KL property & KL function) *We say that a proper function h has the Kurdyka-Łojasiewicz (KL) property at $\widehat{x} \in \mathrm{dom}\,\partial h$ if there exist a neighborhood $\mathcal{V}$ of $\widehat{x}$, $\nu \in (0, \infty]$ and a continuous concave function $\psi : [0, \nu) \to \mathbb{R}_+$ such that:*

(i) *$\psi(0) = 0$ and ψ is continuously differentiable on $(0, \nu)$ with $\psi' > 0$;*

(ii) *for all $x \in \mathcal{V}$ with $h(\widehat{x}) < h(x) < h(\widehat{x}) + \nu$, it holds that*

$$\psi'(h(x) - h(\widehat{x}))\,\mathrm{dist}(0, \partial h(x)) \geq 1.$$

A proper closed function h satisfying the KL property at all points in $\mathrm{dom}\,\partial h$ is called a KL function.

It is known from [4, Section 4.3] that a proper closed semi-algebraic function always satisfies the KL property. Moreover, in this case, the KL property is satisfied with a specific form; see also [11, Corollary 16] and [10, Section 2] for further discussions.

Proposition 1. (KL inequality in the semi-algebraic cases) *Let h be a proper closed semi-algebraic function on $\mathbb{R}^n$. Then, h satisfies the KL property at all points in $\mathrm{dom}\,\partial h$ with $\varphi(s) = cs^{1-\theta}$ for some $\theta \in [0, 1)$ and $c > 0$.*

3 Douglas-Rachford splitting for structured optimization

In this section, as a preparation for solving nonconvex feasibility problems, we consider the following structured optimization problem:

$$\min_u \ f(u) + g(u), \tag{4}$$

where f has a Lipschitz continuous gradient whose Lipschitz continuity modulus is bounded by L, and g is a proper closed function. In addition, we let $l \in \mathbb{R}$ be such that $f + \frac{l}{2}\|\cdot\|^2$ is convex. Notice that such an l always exists: in particular, one can always take $l = L$. Finally, for any given parameter $\gamma > 0$, which will be referred to as a step-size parameter throughout this paper, we assume that the proximal mapping of γg, is easy to compute, in the sense that it is simple to find a minimizer of the following problem for each given z:

$$\min_u \quad \gamma g(u) + \frac{1}{2}\|u - z\|^2. \tag{5}$$

Problems in the form of (4) arise naturally in many engineering and machine learning applications. Specifically, many sparse learning problems take the form of (4) where f is a loss function and g is a regularizer with (5) easy to compute; see, for example, [17] for the use of a difference-of-convex function as a regularizer, and [27] for the case where $g(x) = \sum_{i=1}^n |x_i|^{\frac{1}{2}}$. Below, we consider a direct adaptation of the DR splitting method to solve (4).

Douglas-Rachford splitting method

Step 0. Input an initial point x^0 and a step-size parameter $\gamma > 0$.

Step 1. Set

$$\begin{cases} y^{t+1} \in \operatorname*{Arg\,min}_y \left\{ f(y) + \frac{1}{2\gamma}\|y - x^t\|^2 \right\}, \\ z^{t+1} \in \operatorname*{Arg\,min}_z \left\{ g(z) + \frac{1}{2\gamma}\|2y^{t+1} - x^t - z\|^2 \right\}, \\ x^{t+1} = x^t + (z^{t+1} - y^{t+1}). \end{cases} \tag{6}$$

Step 2. If a termination criterion is not met, go to Step 1.

Using the optimality conditions and the subdifferential calculus rule [25, Exercise 8.8], we see from the y and z-updates in (6) that

$$\begin{aligned} 0 &= \nabla f(y^{t+1}) + \frac{1}{\gamma}(y^{t+1} - x^t), \\ 0 &\in \partial g(z^{t+1}) + \frac{1}{\gamma}(z^{t+1} - y^{t+1}) - \frac{1}{\gamma}(y^{t+1} - x^t). \end{aligned} \tag{7}$$

Hence, we have for all $t \geq 1$ that

$$0 \in \nabla f(y^t) + \partial g(z^t) + \frac{1}{\gamma}(z^t - y^t). \tag{8}$$

Thus, if

$$\lim_{t\to\infty} \|x^{t+1} - x^t\| = \lim_{t\to\infty} \|z^{t+1} - y^{t+1}\| = 0, \tag{9}$$

and for a cluster point (y^*, z^*, x^*) of $\{(y^t, z^t, x^t)\}$ with a convergent subsequence $\lim_{j\to\infty}(y^{t_j}, z^{t_j}, x^{t_j}) = (y^*, z^*, x^*)$, we have

$$\lim_{j\to\infty} g(z^{t_j}) = g(z^*), \tag{10}$$

then passing to the limit in (8) along the subsequence and using (3), it is not hard to see that (y^*, z^*) gives a stationary point of (4), in the sense that $y^* = z^*$ and

$$0 \in \nabla f(z^*) + \partial g(z^*).$$

In the next theorem, we establish convergence of the DR splitting method on (4) by showing that (9) and (10) hold. The proof of this convergence result heavily relies on the following definition of the Douglas-Rachford merit function.

Definition 2. **(DR merit function)** *Let $\gamma > 0$. The Douglas-Rachford merit function is defined by*

$$\mathfrak{D}_\gamma(y, z, x) := f(y) + g(z) - \frac{1}{2\gamma}\|y - z\|^2 + \frac{1}{\gamma}\langle x - y, z - y\rangle. \tag{11}$$

This definition was motivated by the so-called Douglas-Rachford envelop considered in [24, Eq. 31] in the convex case (that is, when f and g are both convex). Moreover, we see that $\mathfrak{D}_\gamma$ can be alternatively written as

$$\begin{aligned}\mathfrak{D}_\gamma(y,z,x) &= f(y)+g(z)+\frac{1}{2\gamma}\|2y-z-x\|^2-\frac{1}{2\gamma}\|x-y\|^2-\frac{1}{\gamma}\|y-z\|^2\\ &= f(y)+g(z)+\frac{1}{2\gamma}(\|x-y\|^2-\|x-z\|^2)\end{aligned} \tag{12}$$

where the first relation follows by applying the elementary relation $\langle u,v\rangle=\frac{1}{2}(\|u+v\|^2-\|u\|^2-\|v\|^2)$ in (11) with $u=x-y$ and $v=z-y$, while the second relation follows by completing the squares in (11).

Theorem 1. (Global subsequential convergence) *Suppose that the parameter $\gamma>0$ is chosen so that*

$$(1+\gamma L)^2+\frac{5\gamma l}{2}-\frac{3}{2}<0. \tag{13}$$

Then if a cluster point of the sequence $\{(y^t,z^t,x^t)\}$ exists, then (9) holds. Moreover, for any cluster point (y^*,z^*,x^*), we have $z^*=y^*$, and

$$0\in\nabla f(z^*)+\partial g(z^*).$$

Remark 1. *Notice that $\lim_{\gamma\downarrow 0}[(1+\gamma L)^2+\frac{5\gamma l}{2}-\frac{3}{2}]=-\frac{1}{2}<0$. Thus, given $l\in\mathbb{R}$ and $L>0$, the condition (13) will be satisfied for any sufficiently small $\gamma>0$. We also comment on the y and z-updates of the DR splitting method. Note that the z-update involves a computation of the proximal mapping of γg, which is simple by assumption. On the other hand, from the choice of γ in Theorem 1, we have $l<\frac{3}{5\gamma}<\frac{1}{\gamma}$. This together with the assumption $f+\frac{l}{2}\|\cdot\|^2$ is convex shows that the objective function in the unconstrained smooth minimization problem for the y-update is a strongly convex function with modulus $\frac{1}{\gamma}-l>0$.*

Proof. We first study the behavior of $\mathfrak{D}_\gamma$ along the sequence generated from the DR splitting method. First of all, notice from (11) that

$$\mathfrak{D}_\gamma(y^{t+1},z^{t+1},x^{t+1})-\mathfrak{D}_\gamma(y^{t+1},z^{t+1},x^t)=\frac{1}{\gamma}\langle x^{t+1}-x^t,z^{t+1}-y^{t+1}\rangle=\frac{1}{\gamma}\|x^{t+1}-x^t\|^2, \tag{14}$$

where the last equality follows from the definition of x-update. Next, using the first relation in (12), we obtain that

$$\begin{aligned}&\mathfrak{D}_\gamma(y^{t+1},z^{t+1},x^t)-\mathfrak{D}_\gamma(y^{t+1},z^t,x^t)\\ &=g(z^{t+1})+\frac{1}{2\gamma}\|2y^{t+1}-z^{t+1}-x^t\|^2-\frac{1}{\gamma}\|y^{t+1}-z^{t+1}\|^2\\ &\quad -g(z^t)-\frac{1}{2\gamma}\|2y^{t+1}-z^t-x^t\|^2+\frac{1}{\gamma}\|y^{t+1}-z^t\|^2\\ &\le\frac{1}{\gamma}(\|y^{t+1}-z^t\|^2-\|y^{t+1}-z^{t+1}\|^2)=\frac{1}{\gamma}(\|y^{t+1}-z^t\|^2-\|x^{t+1}-x^t\|^2),\end{aligned} \tag{15}$$

where the inequality follows from the definition of z^{t+1} as a minimizer, and the last equality follows from the definition of x^{t+1}. Next, notice from the first relation in (7) that

$$\frac{1}{\gamma}(x^t-y^{t+1})+ly^{t+1}=\nabla\left(f+\frac{l}{2}\|\cdot\|^2\right)(y^{t+1}).$$

Since $f + \frac{l}{2}\|\cdot\|^2$ is convex by assumption, using the monotonicity of the gradient of a convex function, we see that for all $t \geq 1$, we have

$$\begin{aligned}
&\left\langle \left(\frac{1}{\gamma}(x^t - y^{t+1}) + ly^{t+1}\right) - \left(\frac{1}{\gamma}(x^{t-1} - y^t) + ly^t\right), y^{t+1} - y^t \right\rangle \geq 0 \\
&\Longrightarrow \langle x^t - x^{t-1}, y^{t+1} - y^t\rangle \geq (1 - \gamma l)\|y^{t+1} - y^t\|^2.
\end{aligned}$$

Hence, we see further that

$$\begin{aligned}
&\|y^{t+1} - z^t\|^2 = \|y^{t+1} - y^t + y^t - z^t\|^2 = \|y^{t+1} - y^t - (x^t - x^{t-1})\|^2 \\
&= \|y^{t+1} - y^t\|^2 - 2\langle y^{t+1} - y^t, x^t - x^{t-1}\rangle + \|x^t - x^{t-1}\|^2 \\
&\leq (-1 + 2\gamma l)\|y^{t+1} - y^t\|^2 + \|x^t - x^{t-1}\|^2,
\end{aligned} \tag{16}$$

where we made use of the definition of x^t for the second equality. Plugging (16) into (15), we obtain that whenever $t \geq 1$,

$$\mathfrak{D}_\gamma(y^{t+1}, z^{t+1}, x^t) - \mathfrak{D}_\gamma(y^{t+1}, z^t, x^t) \leq -\frac{1}{\gamma}\|x^{t+1} - x^t\|^2 + \frac{1}{\gamma}\left((-1 + 2\gamma l)\|y^{t+1} - y^t\|^2 + \|x^t - x^{t-1}\|^2\right). \tag{17}$$

Finally, using the second relation in (12), we obtain that

$$\begin{aligned}
\mathfrak{D}_\gamma(y^{t+1}, z^t, x^t) - \mathfrak{D}_\gamma(y^t, z^t, x^t) &= f(y^{t+1}) + \frac{1}{2\gamma}\|x^t - y^{t+1}\|^2 - f(y^t) - \frac{1}{2\gamma}\|x^t - y^t\|^2 \\
&\leq -\frac{1}{2}\left(\frac{1}{\gamma} - l\right)\|y^{t+1} - y^t\|^2,
\end{aligned} \tag{18}$$

where the inequality follows from the fact that $f + \frac{1}{2\gamma}\|x^t - \cdot\|^2$ is a strongly convex function with modulus $\frac{1}{\gamma} - l$ and the definition of y^{t+1} as a minimizer. Summing (14), (17) and (18), we see further that for any $t \geq 1$,

$$\mathfrak{D}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) - \mathfrak{D}_\gamma(y^t, z^t, x^t) \leq \frac{-3 + 5\gamma l}{2\gamma}\|y^{t+1} - y^t\|^2 + \frac{1}{\gamma}\|x^t - x^{t-1}\|^2. \tag{19}$$

Since we also have from the first relation in (7) and the Lipschitz continuity of ∇f that for $t \geq 1$

$$\|x^t - x^{t-1}\| \leq (1 + \gamma L)\|y^{t+1} - y^t\|, \tag{20}$$

we conclude further that for any $t \geq 1$

$$\mathfrak{D}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) - \mathfrak{D}_\gamma(y^t, z^t, x^t) \leq \frac{1}{\gamma}\left((1 + \gamma L)^2 + \frac{5\gamma l}{2} - \frac{3}{2}\right)\|y^{t+1} - y^t\|^2. \tag{21}$$

Summing the above relation from $t = 1$ to $N - 1 \geq 1$, we obtain that

$$\mathfrak{D}_\gamma(y^N, z^N, x^N) - \mathfrak{D}_\gamma(y^1, z^1, x^1) \leq \frac{1}{\gamma}\left((1 + \gamma L)^2 + \frac{5\gamma l}{2} - \frac{3}{2}\right)\sum_{t=1}^{N-1}\|y^{t+1} - y^t\|^2. \tag{22}$$

Recall that $(1+\gamma L)^2 + \frac{5\gamma l}{2} - \frac{3}{2} < 0$ by our choice of γ. Hence, using the existence of cluster points and taking limit along any convergent subsequence in (22), we conclude that $\lim_{t\to\infty}\|y^{t+1} - y^t\| = 0$. Combining this with (20), we conclude that (9) holds. Furthermore, combining these with the third relation in (6), we obtain further that $\lim_{t\to\infty}\|z^{t+1} - z^t\| = 0$. Thus, if (y^*, z^*, x^*) is a cluster point of $\{(y^t, z^t, x^t)\}$ with a convergent subsequence $\{(y^{t_j}, z^{t_j}, x^{t_j})\}$ so that $\lim_{j\to\infty}(y^{t_j}, z^{t_j}, x^{t_j}) = (y^*, z^*, x^*)$, then

$$\lim_{j\to\infty}(y^{t_j}, z^{t_j}, x^{t_j}) = \lim_{j\to\infty}(y^{t_j-1}, z^{t_j-1}, x^{t_j-1}) = (y^*, z^*, x^*). \tag{23}$$

From the definition of z^t as a minimizer, we have

$$g(z^t) + \frac{1}{2\gamma}\|2y^t - z^t - x^{t-1}\|^2 \leq g(z^*) + \frac{1}{2\gamma}\|2y^t - z^* - x^{t-1}\|^2. \tag{24}$$

Taking limit along the convergent subsequence and using (23) yields

$$\limsup_{j\to\infty} g(z^{t_j}) \leq g(z^*). \tag{25}$$

On the other hand, by the lower semicontinuity of g, we have $\liminf_{j\to\infty} g(z^{t_j}) \geq g(z^*)$. Consequently, (10) holds. Now passing to the limit in (8) along the convergent subsequence $\{(y^{t_j}, z^{t_j}, x^{t_j})\}$, and using (9), (10) and (3), we see that the conclusion of the theorem follows. □

Under the additional assumption that the functions f and g are semi-algebraic functions, we now show that if the whole sequence generated has a cluster point, then it is actually convergent.

Theorem 2. (Global convergence of the whole sequence) *Suppose that the step-size parameter $\gamma > 0$ is chosen as in* (13) *and the sequence $\{(y^t, z^t, x^t)\}$ generated has a cluster point (y^*, z^*, x^*). Suppose in addition that f and g are semi-algebraic functions. Then the whole sequence $\{(y^t, z^t, x^t)\}$ is convergent.*

Proof. The argument is very similar to the proof of [20, Theorem 2]; see also [5, Lemma 2.6]. We first consider the subdifferential of $\mathfrak{D}_\gamma$ at $(y^{t+1}, z^{t+1}, x^{t+1})$. Notice that for any $t \geq 0$, we have

$$\begin{aligned}
\nabla_x \mathfrak{D}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) &= \frac{1}{\gamma}(z^{t+1} - y^{t+1}) = \frac{1}{\gamma}(x^{t+1} - x^t),\\
\nabla_y \mathfrak{D}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) &= \nabla f(y^{t+1}) + \frac{1}{\gamma}(y^{t+1} - x^{t+1}) = \frac{1}{\gamma}(x^t - x^{t+1}),
\end{aligned}$$

where for the first gradient we made use of (11) and the definition of x^{t+1}, while for the second gradient we made use of the second relation in (12) and the first relation in (7). Moreover, for the subdifferential with respect to z, we have from the second relation in (12) that

$$\begin{aligned}
&\partial_z \mathfrak{D}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) = \partial g(z^{t+1}) - \frac{1}{\gamma}(z^{t+1} - x^{t+1})\\
&= \partial g(z^{t+1}) + \frac{1}{\gamma}(z^{t+1} - y^{t+1}) - \frac{1}{\gamma}(y^{t+1} - x^t) - \frac{1}{\gamma}(z^{t+1} - y^{t+1}) + \frac{1}{\gamma}(y^{t+1} - x^t) - \frac{1}{\gamma}(z^{t+1} - x^{t+1})\\
&\ni -\frac{2}{\gamma}(z^{t+1} - y^{t+1}) + \frac{1}{\gamma}(x^{t+1} - x^t) = -\frac{1}{\gamma}(x^{t+1} - x^t),
\end{aligned}$$

where the inclusion follows from the second relation in (7), and the last equality follows from the definition of x^{t+1}. The above relations together with (20) imply the existence of $\tau > 0$ so that whenever $t \geq 1$, we have

$$\mathrm{dist}(0, \partial \mathfrak{D}_\gamma(y^t, z^t, x^t)) \leq \tau \|y^{t+1} - y^t\|. \tag{26}$$

On the other hand, notice from (21) that there exists $K > 0$ so that

$$\mathfrak{D}_\gamma(y^t, z^t, x^t) - \mathfrak{D}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) \geq K\|y^{t+1} - y^t\|^2. \tag{27}$$

In particular, $\{\mathfrak{D}_\gamma(y^t, z^t, x^t)\}$ is non-increasing. Let $\{(y^{t_i}, z^{t_i}, x^{t_i})\}$ be a convergent subsequence that converges to (y^*, z^*, x^*). Then, from the lower semicontinuity of $\mathfrak{D}_\gamma$, we see that the sequence $\{\mathfrak{D}_\gamma(y^{t_i}, z^{t_i}, x^{t_i})\}$ is bounded below. This together with the non-increasing property of $\{\mathfrak{D}_\gamma(y^t, z^t, x^t)\}$ shows that $\{\mathfrak{D}_\gamma(y^t, z^t, x^t)\}$ is also bounded below, and so, $\lim_{t\to\infty} \mathfrak{D}_\gamma(y^t, z^t, x^t) = l^*$ exists.

We next claim that $l^* = \mathfrak{D}_\gamma(y^*, z^*, x^*)$. Let $\{(y^{t_j}, z^{t_j}, x^{t_j})\}$ be any subsequence that converges to (y^*, z^*, x^*). Then from lower semicontinuity, we readily have

$$\liminf_{j\to\infty} \mathfrak{D}_\gamma(y^{t_j}, z^{t_j}, x^{t_j}) \geq \mathfrak{D}_\gamma(y^*, z^*, x^*).$$

On the other hand, proceeding as in (23), (24) and (25), we can conclude further that

$$\limsup_{j\to\infty} \mathfrak{D}_\gamma(y^{t_j}, z^{t_j}, x^{t_j}) \leq \mathfrak{D}_\gamma(y^*, z^*, x^*).$$

These together with the existence of $\lim_{t\to\infty} \mathfrak{D}_\gamma(y^t, z^t, x^t)$ shows that $l^* = \mathfrak{D}_\gamma(y^*, z^*, x^*)$, as claimed. Note that if $\mathfrak{D}_\gamma(y^t, z^t, x^t) = l^*$ for some $t \geq 1$, then $\mathfrak{D}_\gamma(y^t, z^t, x^t) = \mathfrak{D}_\gamma(y^{t+k}, z^{t+k}, x^{t+k})$ for all $k \geq 0$ since the sequence is non-increasing. Then (27) gives $y^t = y^{t+k}$ for all $k \geq 0$. From (20), we see that $x^t = x^{t+k}$ for $k \geq 0$. These together with the third relation in (6) show that we also have $z^{t+1} = z^{t+k}$ for $k \geq 1$. Thus, the algorithm terminates finitely. Since this theorem holds trivially if the algorithm terminates finitely, from now on, we assume $\mathfrak{D}_\gamma(y^t, z^t, x^t) > l^*$ for all $t \geq 1$.

Next, from [4, Section 4.3] and our assumption on semi-algebraicity, the function $(y, z, x) \mapsto \mathfrak{D}_\gamma(y, z, x)$ is a KL function. From the property of KL functions, there exist $\nu > 0$, a neighborhood $\mathcal{V}$ of (y^*, z^*, x^*) and a continuous concave function $\psi : [0, \nu) \to \mathbb{R}_+$ as described in Definition 1 so that for all $(y, z, x) \in \mathcal{V}$ satisfying $l^* < \mathfrak{D}_\gamma(y, z, x) < l^* + \nu$, we have

$$\psi'(\mathfrak{D}_\gamma(y, z, x) - l^*)\,\mathrm{dist}(0, \partial\mathfrak{D}_\gamma(y, z, x)) \geq 1. \tag{28}$$

Pick $\rho > 0$ so that

$$\mathbf{B}_\rho := \{(y, z, x) : \ \|y - y^*\| < \rho, \ \|x - x^*\| < (2 + \gamma L)\rho, \ \|z - z^*\| < 2\rho\} \subseteq \mathcal{V}$$

and set $B_\rho := \{y : \ \|y - y^*\| < \rho\}$. Observe from the first relation in (7) that

$$\|x^t - x^*\| \leq \|x^t - x^{t-1}\| + \|x^{t-1} - x^*\| \leq \|x^t - x^{t-1}\| + (1 + \gamma L)\|y^t - y^*\|.$$

Since (9) holds by Theorem 1, there exists $N_0 \geq 1$ so that $\|x^t - x^{t-1}\| < \rho$ whenever $t \geq N_0$. Thus, it follows that $\|x^t - x^*\| < (2 + \gamma L)\rho$ whenever $y^t \in B_\rho$ and $t \geq N_0$. Next, using the third relation in (6), we see also that for all $t \geq N_0$,

$$\|z^t - z^*\| \leq \|y^t - y^*\| + \|x^t - x^{t-1}\| < 2\rho$$

whenever $y^t \in B_\rho$. Consequently, if $y^t \in B_\rho$ and $t \geq N_0$, then $(y^t, z^t, x^t) \in \mathbf{B}_\rho \subseteq \mathcal{V}$. Furthermore, using the facts that (y^*, z^*, x^*) is a cluster point, that $\lim_{t\to\infty} \mathfrak{D}_\gamma(y^t, z^t, x^t) = l^*$, and that $\mathfrak{D}_\gamma(y^t, z^t, x^t) > l^*$ for all $t \geq 1$, it is not hard to see that there exists (y^N, z^N, x^N) with $N \geq N_0$ such that

(i) $y^N \in B_\rho$ and $l^* < \mathfrak{D}_\gamma(y^N, z^N, x^N) < l^* + \nu$;

(ii) $\|y^N - y^*\| + \frac{\tau}{K}\psi(\mathfrak{D}_\gamma(y^N, z^N, x^N) - l^*) < \rho$.

Before proceeding further, we show that whenever $y^t \in B_\rho$ and $l^* < \mathfrak{D}_\gamma(y^t, z^t, x^t) < l^* + \nu$ for some fixed $t \geq N_0$, we have

$$\|y^{t+1} - y^t\| \leq \frac{\tau}{K}[\psi(\mathfrak{D}_\gamma(y^t, z^t, x^t) - l^*) - \psi(\mathfrak{D}_\gamma(y^{t+1}, z^{t+1}, x^{t+1}) - l^*)]. \tag{29}$$

Since $\{\mathfrak{D}_\gamma(y^t, z^t, x^t)\}$ is non-increasing and ψ is increasing, (29) clearly holds if $y^{t+1} = y^t$. Hence, suppose without loss of generality that $y^{t+1} \neq y^t$. Since $y^t \in B_\rho$ and $t \geq N_0$, we have $(y^t, z^t, x^t) \in$

$\mathbf{B}_\rho \subseteq \mathcal{V}$. Hence, (28) holds for (y^t, z^t, x^t). Making use of the concavity of ψ, (26), (27) and (28), we see that for all such t

$$
\begin{aligned}
&\tau\|y^{t+1}-y^t\|\cdot[\psi(\mathfrak{D}_\gamma(y^t,z^t,x^t)-l^*)-\psi(\mathfrak{D}_\gamma(y^{t+1},z^{t+1},x^{t+1})-l^*)]\\
&\geq \operatorname{dist}(0,\partial\mathfrak{D}_\gamma(y^t,z^t,x^t))\cdot[\psi(\mathfrak{D}_\gamma(y^t,z^t,x^t)-l^*)-\psi(\mathfrak{D}_\gamma(y^{t+1},z^{t+1},x^{t+1})-l^*)]\\
&\geq \operatorname{dist}(0,\partial\mathfrak{D}_\gamma(y^t,z^t,x^t))\cdot\psi'(\mathfrak{D}_\gamma(y^t,z^t,x^t)-l^*)\cdot[\mathfrak{D}_\gamma(y^t,z^t,x^t)-\mathfrak{D}_\gamma(y^{t+1},z^{t+1},x^{t+1})]\\
&\geq K\|y^{t+1}-y^t\|^2,
\end{aligned}
$$

from which (29) follows immediately.

We next show that $y^t \in B_\rho$ whenever $t \geq N$ by induction. The claim is true for $t = N$ by construction. Now, suppose the claim is true for $t = N, \dots, N+k-1$ for some $k \geq 1$; i.e., $y^N, \dots, y^{N+k-1} \in B_\rho$. Notice that as $\{\mathfrak{D}_\gamma(y^t, z^t, x^t)\}$ is a non-increasing sequence, our choice of N implies that $l^* < \mathfrak{D}_\gamma(y^t, z^t, x^t) < l^* + \nu$ for all $t \geq N$. In particular, (29) can be applied for $t = N, \dots, N+k-1$. Thus, for $t = N+k$, we have from this observation that

$$
\begin{aligned}
&\|y^{N+k}-y^*\| \leq \|y^N-y^*\| + \sum_{j=1}^{k}\|y^{N+j}-y^{N+j-1}\|\\
&\leq \|y^N-y^*\| + \frac{\tau}{K}\sum_{j=1}^{k}[\psi(\mathfrak{D}_\gamma(y^{N+j-1},z^{N+j-1},x^{N+j-1})-l^*)-\psi(\mathfrak{D}_\gamma(y^{N+j},z^{N+j},x^{N+j})-l^*)]\\
&\leq \|y^N-y^*\| + \frac{\tau}{K}\psi(\mathfrak{D}_\gamma(y^N,z^N,x^N)-l^*) < \rho,
\end{aligned}
$$

where the last inequality follows from the nonnegativity of ψ. Thus, we have shown that $y^t \in B_\rho$ for $t \geq N$ by induction.

Since $y^t \in B_\rho$ and $l^* < \mathfrak{D}_\gamma(y^t, z^t, x^t) < l^* + \nu$ for $t \geq N$, we can sum (29) from $t = N$ to $M \to \infty$, showing that $\{\|y^{t+1} - y^t\|\}$ is summable. Convergence of $\{y^t\}$ follows immediately from this. Convergence of $\{x^t\}$ follows from this and the first relation in (7). Finally, the convergence of $\{z^t\}$ follows from the third relation in (6). This completes the proof. □

Remark 2. (Comments on the proof) *Below, we make some comments about the proof of Theorem 2.*

(i) *Our proof indeed shows that, if the assumptions in Theorem 2 hold, then the sequence $\{(y^t, z^t, x^t)\}$ generated by the DR splitting method has a finite length, i.e.,*

$$\sum_{t=1}^{\infty}(\|y^{t+1}-y^t\|+\|z^{t+1}-z^t\|+\|x^{t+1}-x^t\|) < +\infty.$$

Precisely, the summability of $\|y^{t+1} - y^t\|$ and $\|x^{t+1} - x^t\|$ can be seen from (29) and (20). Moreover, notice from the third relation in (6) that

$$
\begin{aligned}
\|z^{t+1}-z^t\| &= \|(y^{t+1}+x^{t+1}-x^t)-(y^t+x^t-x^{t-1})\|\\
&\leq \|y^{t+1}-y^t\|+\|x^{t+1}-x^t\|+\|x^t-x^{t-1}\|.
\end{aligned}
$$

Therefore, the summability of $\|z^{t+1} - z^t\|$ follows from the summability of $\|y^{t+1} - y^t\|$ and $\|x^{t+1} - x^t\|$.

(ii) *The proof of Theorem 2 stays valid as long as the DR merit function $\mathfrak{D}_\gamma$ is a KL-function. We only state the case where f and g are semi-algebraic as this simple sufficient condition can be readily checked.*

Recall from Proposition 1 that a semi-algebraic function h satisfies the KL inequality with $\psi(s) = c\, s^{1-\theta}$ for some $\theta \in [0, 1)$ and $c > 0$. We now derive local convergence rates of the proposed nonconvex DR splitting method by examining the range of the exponent.

Theorem 3. (Local convergence rate) *Suppose that the step-size parameter $\gamma > 0$ is chosen as in (13) and the sequence $\{(y^t, z^t, x^t)\}$ generated has a cluster point (y^*, z^*, x^*). Suppose in addition that f and g are semi-algebraic functions so that the ψ in the KL inequality (28) takes the form $\psi(s) = c\, s^{1-\theta}$ for some $\theta \in [0, 1)$ and $c > 0$. Then, we have*

(i) *If $\theta = 0$, then there exists $t_0 \geq 1$ such that for all $t \geq t_0$, $0 \in \nabla f(z^t) + \partial g(z^t)$;*

(ii) *If $\theta \in (0, \frac{1}{2}]$, then there exist $\eta \in (0, 1)$ and $\kappa > 0$ so that $\mathrm{dist}\big(0, \nabla f(z^t) + \partial g(z^t)\big) \leq \kappa\, \eta^t$ for all large t;*

(iii) *If $\theta \in (\frac{1}{2}, 1)$, then there exists $\kappa > 0$ such that $\mathrm{dist}\big(0, \nabla f(z^t) + \partial g(z^t)\big) \leq \kappa\, t^{-\frac{1}{4\theta - 2}}$ for all large t.*

Proof. Let $l_t = \mathfrak{D}_\gamma(y^t, z^t, x^t) - \mathfrak{D}_\gamma(y^*, z^*, x^*)$. Then, we have from the proof of Theorem 2 that $l_t \geq 0$ for all $t \geq 1$ and $l_t \to 0$ as $t \to \infty$. Furthermore, from (27), we have

$$l_t - l_{t+1} \geq K \|y^{t+1} - y^t\|^2. \tag{30}$$

As $l_{t+1} \geq 0$, it follows that $K\|y^{t+1} - y^t\|^2 \leq l_t - l_{t+1} \leq l_t$ for all $t \geq 1$. This together with (20) implies that

$$\|y^t - z^t\| = \|x^t - x^{t-1}\| \leq (1 + \gamma L)\|y^{t+1} - y^t\| \leq \frac{(1 + \gamma L)}{\sqrt{K}} \sqrt{l_t},$$

where the first equality follows from the last relation in (6). Notice from (8) and the Lipschitz continuity of ∇f that

$$\mathrm{dist}(0, \nabla f(z^t) + \partial g(z^t)) \leq \left(L + \frac{1}{\gamma}\right) \|y^t - z^t\|.$$

Consequently, for all $t \geq 1$,

$$\mathrm{dist}(0, \nabla f(z^t) + \partial g(z^t)) \leq \frac{(1 + \gamma L)^2}{\gamma \sqrt{K}} \sqrt{l_t}. \tag{31}$$

Moreover, from the convergence of $\{(y^t, z^t, x^t)\}$ to (y^*, z^*, x^*) guaranteed by Theorem 2, the relation (28), and the discussion that precedes it, we see that either

Case (i): the algorithm terminates finitely, i.e., there exists a $t_0 \geq 1$ such that $l_t = 0$ for some and hence all $t \geq t_0$; or

Case (ii): for all large t, we have $l_t > 0$ and

$$c\,(1 - \theta) l_t^{-\theta} \mathrm{dist}\left(0, \partial \mathfrak{D}_\gamma(y^t, z^t, x^t)\right) \geq 1. \tag{32}$$

For Case (i), (31) implies that the conclusion follows trivially. Therefore, we consider Case (ii). From (26), we obtain that $\mathrm{dist}(0, \partial \mathfrak{D}_\gamma(y^t, z^t, x^t)) \leq \tau \|y^{t+1} - y^t\|$. It then follows that

$$\|y^{t+1} - y^t\| \geq \frac{1}{c\,(1 - \theta)\tau}\, l_t^\theta. \tag{33}$$

Therefore, combining (30) and (33), we see that

$$l_t - l_{t+1} \geq M l_t^{2\theta} \text{ for all large } t,$$

where $M = K(\frac{1}{c\,(1-\theta)\tau})^2$. Without loss of generality, we assume $M \in (0,1)$. We now divide the discussion into three cases:

Case 1: $\theta = 0$. In this case, we have $l_t - l_{t+1} \geq M > 0$, which contradicts $l_t \to 0$. Thus, this case cannot happen.

Case 2: $\theta \in (0, \frac{1}{2}]$. In this case, as $l_t \to 0$, there exists $t_1 \geq 1$ such that $l_t^{2\theta} \geq l_t$ for all $t \geq t_1$. Then, for all large t

$$l_{t+1} \leq l_t - M l_t^{2\theta} \leq (1-M) l_t. \tag{34}$$

So, there exists $\mu > 0$ such that $l_t \leq \mu(1-M)^t$ for all large t. This together with (31) implies that the conclusion of Case 2 follows with $\kappa = \frac{\sqrt{\mu}(1+\gamma L)^2}{\gamma\sqrt{K}}$ and $\eta = \sqrt{1-M} \in (0,1)$.

Case 3: $\theta \in (\frac{1}{2}, 1)$. Define the non-increasing function $h : (0, +\infty) \to \mathbb{R}$ by $h(s) := s^{-2\theta}$. As there exists $i_0 \geq 1$ such that $M\, h(l_i)^{-1} = M l_i^{2\theta} \leq l_i - l_{i+1}$ for all $i \geq i_0$, then we get, for all $i \geq i_0$,

$$M \leq (l_i - l_{i+1}) h(l_i) \leq \int_{l_{i+1}}^{l_i} h(s)ds = \frac{l_i^{1-2\theta} - l_{i+1}^{1-2\theta}}{1-2\theta} = \frac{l_{i+1}^{1-2\theta} - l_i^{1-2\theta}}{2\theta - 1}.$$

Noting that $2\theta - 1 > 0$, this implies that for all $i \geq i_0$

$$l_{i+1}^{1-2\theta} - l_i^{1-2\theta} \geq M(2\theta - 1).$$

Summing for all $i = i_0$ to $i = t-1$ we have for all large t

$$l_t^{1-2\theta} - l_{i_0}^{1-2\theta} \geq M(2\theta-1)(t - i_0).$$

This gives us that for all large t

$$l_t \leq \frac{1}{\sqrt[2\theta-1]{l_{i_0}^{1-2\theta} + M(2\theta-1)(t-i_0)}}.$$

So, combining this with (31), we see that the conclusion of Case 3 follows. □

All our preceding convergence results rely on the existence of a cluster point. Before ending this section, we give some simple sufficient conditions that will guarantee the sequence generated from the DR splitting method is bounded.

Theorem 4. *Suppose that γ is chosen to satisfy* (13), *and so that there exists $\xi > \gamma$ with*

$$\inf_y \left\{ f(y) - \frac{\xi}{2}\|\nabla f(y)\|^2 \right\} =: \zeta_* > -\infty. \tag{35}$$

Suppose in addition that either

(i) *g is coercive, i.e.,* $\liminf_{\|z\|\to\infty} g(z) = \infty$; *or*

(ii) $\inf_z g(z) > -\infty$ *and that* $\liminf_{\|y\|\to\infty} \|\nabla f(y)\| = \infty$.

Then the sequence $\{(y^t, z^t, x^t)\}$ generated from (6) *is bounded.*

Proof. First, it is easy to see from the assumption on γ and (21) that for all $t \geq 1$,

$$\mathfrak{D}_\gamma(y^t, z^t, x^t) \leq \mathfrak{D}_\gamma(y^1, z^1, x^1). \tag{36}$$

In addition, using the second relation in (12), we have for $t \geq 1$ that

$$\begin{aligned} \mathfrak{D}_\gamma(y^t, z^t, x^t) &= f(y^t) + g(z^t) - \frac{1}{2\gamma}\|x^t - z^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2 \\ &= f(y^t) + g(z^t) - \frac{1}{2\gamma}\|x^{t-1} - y^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2, \end{aligned} \tag{37}$$

where the last equality follows from the definition of x^{t+1}, i.e., the third relation in (6). Next, from the first relation in (7), we have for $t \geq 1$ that

$$0 = \nabla f(y^t) + \frac{1}{\gamma}(y^t - x^{t-1}), \tag{38}$$

which implies that $\|x^{t-1} - y^t\|^2 = \gamma^2 \|\nabla f(y^t)\|^2$. Also, recall that $x^t - z^t = x^{t-1} - y^t$. Combining these with (37) and (36), we obtain further that

$$\begin{aligned} \mathfrak{D}_\gamma(y^1, z^1, x^1) \geq \mathfrak{D}_\gamma(y^t, z^t, x^t) &= f(y^t) + g(z^t) - \frac{1}{2\gamma}\|x^{t-1} - y^t\|^2 + \frac{1}{2\gamma}\|x^t - y^t\|^2 \\ &\geq f(y^t) - \frac{\gamma}{2}\|\nabla f(y^t)\|^2 + g(z^t) + \frac{1}{2\gamma}\|x^t - y^t\|^2 \\ &\geq \zeta_* + \frac{\xi - \gamma}{2}\|\nabla f(y^t)\|^2 + g(z^t) + \frac{1}{2\gamma}\|x^t - y^t\|^2, \end{aligned} \tag{39}$$

where the last inequality follows from (35).

Now, suppose first that (i) holds, i.e., g is coercive. Then it follows readily from (39) that $\{z^t\}$, $\{\nabla f(y^t)\}$ and $\{x^t - y^t\}$ are bounded. From (38) we see immediately that $\{y^t - x^{t-1}\}$ is also bounded. This together with the boundedness of $\{x^t - y^t\}$ shows that $\{x^t - x^{t-1}\}$ is also bounded. From the third relation in (6), this means that $\{z^t - y^t\}$ is bounded. Since we know already that $\{z^t\}$ is bounded, it follows that $\{y^t\}$ is also bounded. The boundedness of $\{x^t\}$ now follows from this and the boundedness of $\{x^t - y^t\}$.

Finally, suppose that (ii) holds. Then we see immediately from (39) that $\{y^t\}$ and $\{x^t - y^t\}$ are bounded. Consequently, the sequence $\{x^t\}$ is also bounded. The boundedness of $\{z^t\}$ then follows from the third relation in (6). This completes the proof. □

Remark 3. *The sufficient conditions given in Theorem 4 are similar to those discussed in [20, Theorem 2]. Following the arguments used there, it can be shown that these sufficient conditions are satisfied for a wide range of applications including the least squares problems with $\ell_{1/2}$ regularization. We refer the readers to [20, Remark 3] and the examples therein for the detailed discussion of the applicability of the aforementioned conditions. In the next section, we will also show that these sufficient conditions can be satisfied for nonconvex feasibility problems under a mild assumption.*

4 Douglas-Rachford splitting for nonconvex feasibility problems

In this section, we discuss how the nonconvex DR splitting method in Section 3 can be applied to solving a feasibility problem.

Let C and D be two nonempty closed sets, with C being convex. We also assume that a projection onto each of them is easy to compute. The feasibility problem is to find a point in $C \cap D$, if any. It is clear that $C \cap D \neq \emptyset$ if and only if the following optimization problem has a zero optimal value:

$$\begin{array}{cl} \min\limits_{u} & \frac{1}{2}d_C^2(u) \\ \text{s.t.} & u \in D. \end{array} \tag{40}$$

Since C is closed and convex, it is well known that the function $u \mapsto \frac{1}{2}d_C^2(u)$ is smooth with a Lipschitz continuous gradient whose Lipschitz continuity modulus is 1. Moreover, for each $\gamma > 0$, it is not hard to show that the infimum in

$$\inf_y \left\{ \frac{1}{2}d_C^2(y) + \frac{1}{2\gamma}\|y - x\|^2 \right\}$$

is attained at $y = \frac{1}{1+\gamma}(x + \gamma P_C(x))$. Hence, applying the DR splitting method in Section 3 to solving (40) gives the following algorithm:

Douglas-Rachford splitting method for feasibility problem

Step 0. Input an initial point x^0 and a step-size parameter $\gamma > 0$.

Step 1. Set

$$\begin{cases} y^{t+1} = \dfrac{1}{1+\gamma}(x^t + \gamma P_C(x^t)), \\ z^{t+1} \in \operatorname*{Arg\,min}_{z \in D} \left\{ \|2y^{t+1} - x^t - z\|^2 \right\}, \\ x^{t+1} = x^t + (z^{t+1} - y^{t+1}). \end{cases} \tag{41}$$

Step 2. If a termination criterion is not met, go to Step 1.

Notice that the above algorithm involves computation of $P_C(x^t)$ and a projection of $2y^{t+1} - x^t$ onto D, which are both easy to compute by assumption. Moreover, observe that as $\gamma \to \infty$, (41) reduces to the classical DR splitting method considered in the literature for finding a point in $C \cap D$, i.e., the DR splitting method in (6) applied to minimizing the sum of the indicator functions of C and D. Comparing with this, the version in (41) can be viewed as a *damped* DR splitting method for finding feasible points. The convergence of both versions are known in the convex scenario, i.e, when D is also convex. However, in our case, C is closed and convex while D is possibly nonconvex. Thus, the known results do not apply directly. Nonetheless, we have the following convergence result using Theorem 1. In addition, we can show in this particular case that the sequence $\{(y^t, z^t, x^t)\}$ generated from (41) is bounded, assuming C or D is compact.

Theorem 5. (Convergence of DR method for nonconvex feasibility problem involving two sets) *Suppose that $0 < \gamma < \sqrt{\frac{3}{2}} - 1$ and that either C or D is compact. Then the sequence $\{(y^t, z^t, x^t)\}$ generated from (41) is bounded, and any cluster point (y^*, z^*, x^*) of the sequence satisfies $z^* = y^*$, and z^* is a stationary point of (40). Moreover, (9) holds.*

Proof. From the above discussion, the algorithm (41) is just (6) as applied to (40). Thus, in particular, one can pick $L = 1$ and $l = 0$ using properties of $\frac{1}{2}d_C^2$. In view of Theorem 1, we only need to show that the sequence $\{(y^t, z^t, x^t)\}$ is bounded. We shall check that the conditions in Theorem 4 are satisfied.

First, note that $\nabla f(y^t) = y^t - P_C(y^t)$, where $f = \frac{1}{2}d_C^2$. Hence, we have

$$\inf_y \left\{ f(y) - \frac{1}{2}\|\nabla f(y)\|^2 \right\} \geq 0,$$

meaning that (35) holds with $\xi = 1 > \sqrt{\frac{3}{2}} - 1$ and $\zeta_* = 0$.

Now, if D is compact, then $g = \delta_D$ is coercive and hence condition (i) in Theorem 4 is satisfied. On the other hand, if C is compact, then $\|\nabla f(y)\| = d_C(y)$ is coercive and $g = \delta_D$ is clearly bounded below. Thus, condition (ii) in Theorem 4 is satisfied. Consequently, Theorem 4 is applicable, from which we conclude that the sequence $\{(y^t, z^t, x^t)\}$ generated from (41) is bounded. $\square$

We have the following immediate corollary if C and D are closed semi-algebraic sets.

Corollary 1. *Let C and D be closed semi-algebraic sets. Suppose that $0 < \gamma < \sqrt{\frac{3}{2}} - 1$ and that either C or D is compact. Then the sequence $\{(y^t, z^t, x^t)\}$ converges to a point (y^*, z^*, x^*) which satisfies $z^* = y^*$, and z^* is a stationary point of* (40).

Proof. As C is a semi-algebraic set, $y \mapsto \frac{1}{2}d_C^2(y)$ is a semi-algebraic function (see [5, Lemma 2.3]). Note that D is also a semi-algebraic set, and so, $z \mapsto \delta_D(z)$ is also a semi-algebraic function. Thus, the conclusion follows from Theorem 5 and Theorem 2 with $f(y) = \frac{1}{2}d_C^2(y)$ and $g(z) = \delta_D(z)$. ☐

Remark 4. (Practical computation consideration on the step-size parameter) *Though the upper bound on γ given in* (13) *might be too small in practice, it can be used in designing an update rule of γ so that the resulting algorithm is guaranteed to converge (in the sense described by Theorem 5). Indeed, similar to the discussion in [26, Remark 2.1], one could initialize the algorithm with a large γ, and decrease the γ by a constant ratio if γ exceeds $\sqrt{\frac{3}{2}} - 1$ and the iterate satisfies either $\|y^t - y^{t-1}\| > c_0/t$ for some prefixed $c_0 > 0$ or $\|y^t\| > c_1$ for some huge number $c_1 > 0$. In the worst case, one can obtain $0 < \gamma < \sqrt{\frac{3}{2}} - 1$ after finitely many decrease. Moreover, it is not hard to see from the proof of Theorem 5 that this heuristic ensures the boundedness of the sequence generated and its clustering at stationary points when either C or D is compact.*

In general, it is possible that the algorithm (41) gets stuck at a stationary point that is not a global minimizer. Thus, there is no guarantee that this algorithm will solve the feasibility problem. However, a zero objective value of $d_C(y^*)$ certifies that y^* is a solution of the feasibility problem, i.e., $y^* \in C \cap D$.

We next consider a specific case where $C = \{x \in \mathbb{R}^n : \; Ax = b\}$ for some matrix $A \in \mathbb{R}^{m\times n}$, $m \le n$, and D is a closed semi-algebraic set. We show below that, if $\{(y^t, z^t, x^t)\}$ generated by our DR splitting method converges to some (y^*, z^*, x^*) with z^* satisfying a certain constraint qualification, then the scheme indeed exhibits a local linear convergence rate.

Proposition 2. (Local linear convergence rate under constraint qualification) *Let $C = \{x \in \mathbb{R}^n : \; Ax = b\}$ and D be a closed semi-algebraic set where $A \in \mathbb{R}^{m\times n}$, $m \le n$, and $b \in \mathbb{R}^m$. Let $0 < \gamma < \sqrt{\frac{3}{2}} - 1$ and suppose that the sequence $\{(y^t, z^t, x^t)\}$ generated from* (41) *converges to (y^*, z^*, x^*). Suppose, in addition, that $C \cap D \neq \emptyset$ and the following constraint qualification holds:*

$$N_C\big(P_C(z^*)\big) \cap -N_D(z^*) = \{0\}. \tag{42}$$

Then, $z^* \in C \cap D$ and there exist $\eta \in (0, 1)$ and $\kappa > 0$ such that for all large t,

$$\operatorname{dist}\big(0, z^t - P_C(z^t) + N_D(z^t)\big) \le \kappa\, \eta^t.$$

Proof. Without loss of generality, we assume that $\mathfrak{D}_\gamma(y^t, z^t, x^t) > \mathfrak{D}_\gamma(y^*, z^*, x^*)$ and hence $(y^t, z^t, x^t) \neq (y^*, z^*, x^*)$ for all $t \ge 1$; since otherwise, the algorithm terminates finitely.

Recall that any limit (y^*, z^*, x^*) satisfies $y^* = z^* \in D$, and from the optimality condition, we see also that

$$0 = z^* - P_C(z^*) + N_D(z^*).$$

On the other hand, note also that $z^* - P_C(z^*) \in N_C\big(P_C(z^*)\big)$. Hence, our assumption implies that $z^* - P_C(z^*) = 0$ and so, $z^* \in C$. Thus, we have $y^* = z^* \in C \cap D$. Note that $y^* = \frac{1}{1+\gamma}(x^* + \gamma P_C(x^*))$, from which one can easily see that $x^* = y^*$.

Let $\mathfrak{D}_\gamma(y, z, x) = \delta_D(z) + \widehat{\mathfrak{D}}_\gamma(y, z, x)$ where

$$\begin{aligned} \widehat{\mathfrak{D}}_\gamma(y, z, x) \;&:=\; \frac{1}{2}d_C^2(y) - \frac{1}{2\gamma}\|y - z\|^2 + \frac{1}{\gamma}\langle x - y, z - y\rangle \\ &=\; \frac{1}{2}\|A^\dagger(Ay - b)\|^2 - \frac{1}{2\gamma}\|y - z\|^2 + \frac{1}{\gamma}\langle x - y, z - y\rangle. \end{aligned}$$

where $A^\dagger$ is the pseudo inverse of the matrix A. Let us consider the function h defined by

$$h(y,z,x) = \widehat{\mathfrak{D}}_\gamma(y+y^*, z+z^*, x+x^*) - \widehat{\mathfrak{D}}_\gamma(y^*, z^*, x^*).$$

Recall that $x^* = y^* = z^* \in C \cap D$. Hence, h is a quadratic function with $h(0,0,0) = 0$ and $\nabla h(0,0,0) = 0$. Thus, we have $h(u) = \frac{1}{2}u^T Bu$, where $u = (y,z,x)$ and

$$B = \nabla^2 h(0,0,0) = \begin{pmatrix} (A^\dagger A)^T(A^\dagger A) + \frac{1}{\gamma}I_n & 0 & -\frac{1}{\gamma}I_n \\ 0 & -\frac{1}{\gamma}I_n & \frac{1}{\gamma}I_n \\ -\frac{1}{\gamma}I_n & \frac{1}{\gamma}I_n & 0 \end{pmatrix}.$$

Here, we use I_n to denote the $n \times n$ identity matrix. Clearly, B is an indefinite $3n \times 3n$ matrix and so, $\{u : u^T Bu = 1\} \neq \emptyset$. Consider the following homogeneous quadratic optimization problem:

$$\begin{aligned} \alpha = \inf_{u \in \mathbb{R}^{3n}} \quad & \|Bu\|^2 \\ \text{s.t.} \quad & u^T Bu = 1. \end{aligned} \tag{43}$$

Clearly $\alpha \geq 0$. We now claim that $\alpha > 0$. To see this, we proceed by the method of contradiction and suppose that there exists a sequence $\{u^t\}$ such that $(u^t)^T Bu^t = 1$ and $\|Bu^t\|^2 \to 0$. Let $B = V^T \Sigma V$ be an eigenvalue decomposition of B, where V is an orthogonal matrix and Σ is a diagonal matrix. Letting $w^t = Vu^t$, we have

$$(w^t)^T \Sigma w^t = 1 \text{ and } (w^t)^T \Sigma^2 w^t \to 0.$$

Let $w^t = (w_1^t, \ldots, w_{3n}^t)$ and $\Sigma = \mathrm{Diag}(\lambda_1, \ldots, \lambda_{3n})$. Then we see further that

$$\sum_{i=1}^{3n} \lambda_i (w_i^t)^2 = 1 \text{ and } \sum_{i=1}^{3n} \lambda_i^2 (w_i^t)^2 \to 0.$$

The second relation shows that either $\lambda_i = 0$ or $w_i^t \to 0$ for each $i = 1, \ldots, 3n$. This contradicts the first relation. So, we must have $\alpha > 0$. Consequently, for any u satisfying $u^T Bu > 0$, we have

$$\|Bu\| \geq \sqrt{\alpha}\sqrt{u^T Bu}.$$

Recall that $h(u) = \frac{1}{2}u^T Bu$. It then follows from the definition of h that for all $t \geq 1$, we have

$$\begin{aligned} \|\nabla \widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t)\| &\geq \sqrt{2\alpha}\sqrt{\widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t) - \widehat{\mathfrak{D}}_\gamma(y^*, z^*, x^*)} \\ &= \sqrt{2\alpha}\sqrt{\mathfrak{D}_\gamma(y^t, z^t, x^t) - \mathfrak{D}_\gamma(y^*, z^*, x^*)}, \end{aligned} \tag{44}$$

where the equality follows from $\widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t) = \mathfrak{D}_\gamma(y^t, z^t, x^t) > \mathfrak{D}_\gamma(y^*, z^*, x^*) = \widehat{\mathfrak{D}}_\gamma(y^*, z^*, x^*)$ (thanks to $z^t \in D$). Finally, to finish the proof, we only need to justify the existence of $\beta > 0$ such that for all large t,

$$\mathrm{dist}(0, \partial \mathfrak{D}_\gamma(y^t, z^t, x^t)) \geq \beta \|\nabla \widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t)\|. \tag{45}$$

Then the conclusion of the proposition would follow from Theorem 3 with $\theta = \frac{1}{2}$.

Note first that

$$\mathrm{dist}(0, \partial \mathfrak{D}_\gamma(y^t, z^t, x^t)) = \mathrm{dist}\left(0, \{\nabla_y \widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t)\} \times \left(\nabla_z \widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t) + N_D(z^t)\right) \times \{\nabla_x \widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t)\}\right),$$

To establish (45), we only need to consider the partial subgradients with respect to z. To this end, define $w^t := \nabla_z \widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t) = -\frac{1}{\gamma}(z^t - x^t)$ and let $v^t \in N_D(z^t)$ be such that

$$\mathrm{dist}\left(0, \nabla_z \widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t) + N_D(z^t)\right) = \left\|w^t + v^t\right\|.$$

We now claim that, there exists $\theta \in [0,1)$ such that for all large t

$$\langle w^t, v^t \rangle \geq -\theta \|w^t\| \cdot \|v^t\|. \tag{46}$$

Otherwise, there exist $t_k \to \infty$ and $\theta_k \uparrow 1$ such that

$$\langle w^{t_k}, v^{t_k} \rangle < -\theta_k \|w^{t_k}\| \cdot \|v^{t_k}\|. \tag{47}$$

In particular, $w^{t_k} \neq 0$ and $v^{t_k} \neq 0$. Furthermore, note that $v^t \in N_D(z^t)$ and

$$w^t = -\frac{1}{\gamma}(z^t - x^t) = -\frac{1}{\gamma}(y^t - x^{t-1}) = \frac{1}{1+\gamma}\left(x^{t-1} - P_C(x^{t-1})\right),$$

where the second equality follows from the third relation in (41) and the last relation follows from the first relation in (41). By passing to a subsequence if necessary, we may assume that

$$\frac{w^{t_k}}{\|w^{t_k}\|} \to w^* \in N_C(P_C(x^*)) \cap S = N_C(P_C(z^*)) \cap S \text{ and } \frac{v^{t_k}}{\|v^{t_k}\|} \to v^* \in N_D(z^*) \cap S,$$

where S is the unit sphere. Dividing $\|w^{t_k}\|\|v^{t_k}\|$ on both sides of (47) and passing to the limit, we see that

$$\langle w^*, v^* \rangle \leq -1.$$

This shows that $\|w^* + v^*\|^2 = 2 + 2\langle w^*, v^* \rangle \leq 0$ and hence, $w^* = -v^*$. This contradicts (42) and thus (46) holds for some $\theta \in [0,1)$ and for all large t.

Now, using (46), we see that for all large t

$$\begin{aligned}
\left\| -\frac{1}{\gamma}(z^t - x^t) + v^t \right\|^2 &= \|w^t + v^t\|^2 = \|w^t\|^2 + \|v^t\|^2 + 2\langle w^t, v^t \rangle \\
&\geq \|w^t\|^2 + \|v^t\|^2 - 2\theta \|w^t\|\|v^t\| \geq (1-\theta)(\|w^t\|^2 + \|v^t\|^2) \\
&\geq (1-\theta)\left\| -\frac{1}{\gamma}(z^t - x^t) \right\|^2.
\end{aligned}$$

Therefore, for all large t

$$\begin{aligned}
&\operatorname{dist}^2\left(0, \nabla_z \widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t) + N_D(z^t)\right) = \left\| -\frac{1}{\gamma}(z^t - x^t) + v^t \right\|^2 \\
&\geq (1-\theta)\left\| -\frac{1}{\gamma}(z^t - x^t) \right\|^2 = (1-\theta)\|\nabla_z \widehat{\mathfrak{D}}_\gamma(y^t, z^t, x^t)\|^2.
\end{aligned}$$

Therefore, (45) holds with $\beta = \sqrt{1-\theta}$. Thus, the conclusion follows. □

For a general nonconvex feasibility problem, i.e., to find a point in $\bigcap_{i=1}^M D_i$, with each D_i being a closed set whose projection is easy to compute, it is classical to reformulate the problem as finding a point in the intersection of $H \cap (D_1 \times D_2 \times \cdots \times D_M)$, where

$$H = \{(x_1, \ldots, x_M) : \ x_1 = \cdots = x_M\}. \tag{48}$$

The algorithm (41) can thus be applied. In addition, if it is known that $\bigcap_{i=1}^M D_i$ is bounded, one can further reformulate the problem as finding a point in the intersection of $H_R \cap (D_1 \times D_2 \times \cdots \times D_N)$, where

$$H_R = \{(x_1, \ldots, x_M) : \ x_1 = \cdots = x_M, \quad \|x_1\| \leq R\}, \tag{49}$$

and R is an upper bound on the norms of the elements in $\bigcap_{i=1}^M D_i$. We note that both the projections onto H and H_R can be easily computed.

We next state a corollary concerning the convergence of our DR splitting method as applied to finding a point in the intersection of $H \cap (D_1 \times D_2 \times \cdots \times D_M)$, assuming compactness of D_i, $i = 1, \ldots, M$.

Corollary 2. (DR method for general nonconvex feasibility problem) *Let $D_1, \ldots, D_M$ be compact semi-algebraic sets in $\mathbb{R}^n$. Let $C = H$, where H is defined as in* (48)*, and let $D = D_1 \times \cdots \times D_M$. Let $0 < \gamma < \sqrt{\frac{3}{2}} - 1$ and let the sequence $\{(y^t, z^t, x^t)\}$ be generated from* (41)*. Then,*

(i) *the sequence $\{(y^t, z^t, x^t)\}$ converges to a point (y^*, z^*, x^*), with $y^* = z^*$ and $z^* = (z_1^*, \ldots, z_M^*) \in D_1 \times \cdots D_M$ satisfying*

$$0 \in z_i^* - \frac{1}{M}\sum_{i=1}^{M} z_i^* + N_{D_i}(z_i^*),\ \ i = 1, \ldots, M.$$

(ii) *Suppose, in addition, that $\bigcap_{i=1}^M D_i \neq \emptyset$ and the following constraint qualification holds:*

$$a_i \in N_{D_i}(z_i^*),\, i = 1, \ldots, M,\ \sum_{i=1}^{M} a_i = 0 \ \Rightarrow\ a_i = 0, i = 1, \ldots, M. \tag{50}$$

Then, $z_1^* = \cdots = z_M^* \in \bigcap_{i=1}^M D_i$ and there exist $\eta \in (0,1)$ and $\kappa > 0$ such that for all large t,

$$\operatorname{dist}\left(0, z_i^t - \frac{1}{M}\sum_{i=1}^{M} z_i^t + N_{D_i}(z_i^t)\right) \le \kappa\, \eta^t,\ \ i = 1, \ldots, M. \tag{51}$$

Proof. Clearly, C and D are both semi-algebraic sets. As D is compact, from Corollary 1, we see that the sequence $\{(y^t, z^t, x^t)\}$ converges to a point (y^*, z^*, x^*) satisfying $y^* = z^*$ and

$$0 \in z^* - P_C(z^*) + N_D(z^*).$$

Moreover, recall from [25, Proposition 10.5] and the definition of normal cone as the subdifferential of the indicator function that

$$N_D(z^*) = N_{D_1}(z_1^*) \times \cdots \times N_{D_1}(z_M^*). \tag{52}$$

The conclusion in (i) now follows from these and the observation that

$$P_C(z^*) = \left(\frac{1}{M}\sum_{i=1}^{M} z_i^*, \ldots, \frac{1}{M}\sum_{i=1}^{M} z_i^*\right) \in \mathbb{R}^{Mn}.$$

We now prove (ii). Note that for any $x \in C$,

$$N_C(x) = \left\{(a_1, \ldots, a_M) : \sum_{i=1}^{M} a_i = 0\right\}.$$

Thus, the constraint qualification (50), together with (52), imply that $N_C\big(P_C(z^*)\big) \cap -N_D(z^*) = \{0\}$. Consequently, Proposition 2 gives $z^* = (z_1^*, \ldots, z_M^*) \in C \cap D$ and (51) holds. Finally, from the definitions of C and D, we have $z_1^* = \cdots = z_M^* \in \bigcap_{i=1}^M D_i$. Hence, the conclusion in (ii) also follows. $\square$

Remark 5. *The constraint qualification* (50) *is known as the linear regularity condition. It plays an important role in quantifying the local linear convergence rate of the alternating projection method [19].*

Before closing this section, we use the example given in [9, Remark 6] to illustrate the difference in the behavior of our DR splitting method (41) and the usual DR splitting method considered in the literature for the feasibility problem, i.e., (6) applied to minimizing the sum of the indicator functions of the two sets.

Example 1. (Different behavior: our DR method vs the usual DR method) *We consider* $C = \{x \in \mathbb{R}^2 : \ x_2 = 0\}$ *and* $D = \{(0,0),(7+\eta,\eta),(7,-\eta)\}$*, where* $\eta \in (0,1]$*. It was discussed in [9, Remark 6] that the DR splitting method, initialized at* $x^0 = (7,\eta)$ *and applied to minimizing the sum of indicator functions of the two sets, is not convergent; indeed, the sequence generated has a discrete limit cycle. On the other hand, convergence of* (41) *applied to this pair of sets is guaranteed by Corollary 1, as long as* $0 < \gamma < \sqrt{\frac{3}{2}} - 1$*. Below, we show explicitly that the generated sequence is convergent, and the limit is* $y^* = z^* = (7+\eta,\eta)$ *and* $x^* = (7+\eta,(1+\gamma)\eta)$*.*

To this end, we first consider a sequence $\{a_t\}$ *defined by* $a_1 = 2 - \frac{1}{1+\gamma} > 0$ *and for any* $t \geq 1$*,*

$$a_{t+1} = \frac{\gamma}{1+\gamma} a_t + 1. \tag{53}$$

Then $a_t > 0$ *for all* t *and we have*

$$a_{t+1} - a_t = \frac{\gamma}{1+\gamma}(a_t - a_{t-1}) = \cdots = \left(\frac{\gamma}{1+\gamma}\right)^{t-1}(a_2 - a_1).$$

Consequently, $\{a_t\}$ *is a Cauchy sequence and is thus convergent. Furthermore, it follows immediately from* (53) *that* $\lim_{t\to\infty} a_t = 1+\gamma$*.*

Now, we look at (41) *initialized at* $x^0 = (7,\eta)$*. Then* $y^1 = \left(7, \frac{\eta}{1+\gamma}\right)$ *and* $2y^1 - x^0 = \left(7, \left[\frac{2}{1+\gamma} - 1\right]\eta\right)$*. Since* $\gamma < \sqrt{\frac{3}{2}} - 1 < \frac{3}{5}$*, it is not hard to show that* $z^1 = (7+\eta,\eta)$ *and consequently* $x^1 = (7+\eta, a_1\eta)$*. Inductively, one can show that for all* $t \geq 1$*,*

$$y^{t+1} = \left(7+\eta, \frac{a_t}{1+\gamma}\eta\right), \quad z^{t+1} = (7+\eta,\eta) \quad \text{and} \quad x^{t+1} = (7+\eta, a_{t+1}\eta)\,.$$

Consequently, $y^* = z^* = (7+\eta,\eta)$ *and* $x^* = (7+\eta,(1+\gamma)\eta)$*.*

5 Numerical simulations

In this section, we perform numerical experiments to test the DR splitting method on solving a nonconvex feasibility problem. All codes are written in MATLAB, and the experiments were performed in MATLAB version R2014a on a cluster with 32 processors (2.9 GHz each) and 252G RAM.

We consider the problem of finding an r-sparse solution of a linear system $Ax = b$. To apply the DR splitting method, we let $C = \{x \in \mathbb{R}^n : \ Ax = b\}$ and $D = \{x \in \mathbb{R}^n : \ \|x\|_0 \leq r\}$, where $\|x\|_0$ denotes the cardinality of x. We benchmark our algorithm against the alternating projection method, which is an application of the proximal gradient algorithm with step-length 1 to solve (40). Specifically, in this latter algorithm, one initializes at an x^0 and updates

$$x^{t+1} \in \operatorname*{Arg\,min}_{\|x\|_0 \leq r} \left\{\|x - (x^t + A^\dagger(b - Ax^t))\|\right\}.$$

We initialize both algorithms at the origin and terminate them when

$$\frac{\max\{\|x^t - x^{t-1}\|, \|y^t - y^{t-1}\|, \|z^t - z^{t-1}\|\}}{\max\{\|x^{t-1}\|, \|y^{t-1}\|, \|z^{t-1}\|, 1\}} < 10^{-8} \quad \text{and} \quad \frac{\|x^t - x^{t-1}\|}{\max\{\|x^{t-1}\|, 1\}} < 10^{-8}$$

respectively, for the DR splitting method and the alternating projection method. Furthermore, for the DR splitting method, we adapt the heuristics described in Remark 4: we initialize $\gamma = 150 \cdot \gamma_0$ and update γ as $\max\{\frac{\gamma}{2}, 0.9999 \cdot \gamma_0\}$ whenever $\gamma > \gamma_0 := \sqrt{\frac{3}{2}} - 1$, and the sequence satisfies either $\|y^t - y^{t-1}\| > \frac{1000}{t}$ or $\|y^t\| > 10^{10}$.[1]

We generate random linear systems with sparse solutions. We first generate an $m \times n$ matrix A with i.i.d. standard Gaussian entries. We then randomly generate an $\widehat{x} \in \mathbb{R}^r$ with $r = \lceil \frac{m}{5} \rceil$, again with i.i.d. standard Gaussian entries. A random sparse vector $\widetilde{x} \in \mathbb{R}^n$ is then generated by first setting $\widetilde{x} = 0$ and then specifying r random entries in $\widetilde{x}$ to be $\widehat{x}$. Finally, we set $b = A\widetilde{x}$.

In our experiments, for each m = 100, 200, 300, 400 and 500, and n = 4000, 5000 and 6000, we generate 50 random instances as described above. The computational results are reported in Table 1, where we report the number of iterations averaged over the 50 instances, as well as the maximum and minimum function values at termination.[2] We also report the number of successes and failures, where we declare a success if the function value at termination is below 10^{-12}, and a failure if the value is above 10^{-6}. We observe that both methods fail more often for harder instances (smaller m), and the DR splitting method clearly outperforms the alternating projection method in terms of both the number of iterations and the solution quality.

Data		DR					Alt Proj				
m	n	iter	fval$_{\max}$	fval$_{\min}$	succ	fail	iter	fval$_{\max}$	fval$_{\min}$	succ	fail
100	4000	1967	3e-02	6e-17	30	20	1694	8e-02	4e-03	0	50
100	5000	2599	2e-02	2e-16	18	32	1978	7e-02	5e-03	0	50
100	6000	2046	1e-02	1e-16	12	38	2350	5e-02	4e-05	0	50
200	4000	836	2e-15	2e-16	50	0	1076	3e-01	3e-05	0	50
200	5000	1080	3e-15	2e-16	50	0	1223	2e-01	2e-03	0	50
200	6000	1279	7e-02	1e-16	43	7	1510	2e-01	2e-13	1	49
300	4000	600	3e-15	2e-16	50	0	872	4e-01	6e-14	3	46
300	5000	710	4e-15	4e-16	50	0	1068	3e-01	9e-14	3	45
300	6000	812	3e-15	2e-16	50	0	1252	3e-01	1e-13	1	49
400	4000	520	2e-15	3e-17	50	0	818	6e-01	8e-14	30	19
400	5000	579	3e-15	5e-16	50	0	946	4e-01	9e-14	12	36
400	6000	646	4e-15	6e-16	50	0	1108	3e-01	1e-13	4	44
500	4000	499	1e-16	1e-18	50	0	640	4e-01	6e-14	38	10
500	5000	519	1e-15	4e-17	50	0	846	4e-01	9e-14	37	13
500	6000	556	3e-15	3e-16	50	0	1071	5e-01	1e-13	22	28

Table 1: Comparing Douglas-Rachford splitting and alternating projection on random instances.

6 Concluding remarks

In this paper, we examine the convergence behavior of the Douglas-Rachford splitting method when applied to solving nonconvex optimization problems, particularly, the nonconvex feasibility problem. By introducing the Douglas-Rachford merit function, we prove the global convergence and establish local convergence rates for the DR splitting method when the step-size parameter γ is chosen sufficiently small (with an explicit threshold) and the sequence generated is bounded. Preliminary numerical experiments are performed, which indicate that the DR splitting method

[1]We also solved a couple instances using directly a small $\gamma < \gamma_0$ in the DR splitting method. The sequence generated tends to get stuck at stationary points that are not global minimizers.

[2]We report $\frac{1}{2}d_C^2(z^t)$ for DR splitting, and $\frac{1}{2}d_C^2(x^t)$ for alternating projection.

usually outperforms the alternating projection method in finding a sparse solution of a linear system, in terms of both solution quality and number of iterations taken.

References

[1] A. F. Aragón and J.M. Borwein. Global convergence of a non-convex Douglas-Rachford iteration. *J. Global Optim.* 57, pp. 1–17 (2012).

[2] A. F. Aragón, J. M. Borwein and M. K. Tam. Douglas-Rachford feasibility methods for matrix completion problems. *ANZIAM J.* 55, pp. 299–326 (2014).

[3] A. F. Aragón, J. M. Borwein and M. K. Tam, Recent results on Douglas-Rachford methods for combinatorial optimization problems. *J. Optim. Theory & Appl.* 163, pp. 1–30 (2014).

[4] H. Attouch, J. Bolte, P. Redont and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems. An approach based on the Kurdyka-Lojasiewicz inequality. *Math. Oper. Res.* 35, pp. 438–457 (2010).

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

[6] H. H. Bauschke and J. M. Borwein. On the convergence of von Neumann's alternating projection algorithm for two sets. *Set-Valued Anal.* 1, pp. 185–212 (1993).

[7] H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Rev.* 38, pp. 367–426 (1996).

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

[9] H. H. Bauschke and D. Noll. On the local convergence of the Douglas-Rachford algorithm. Preprint (2014). Available at `http://arxiv.org/abs/1401.6188`.

[10] J. Bolte, A. Daniilidis and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. Optim.* 17, pp. 1205–1223 (2007).

[11] J. Bolte, A. Daniilidis, A. Lewis and M. Shiota. Clarke subgradients of stratifiable functions. *SIAM J. Optim.* 18, pp. 556–572 (2007).

[12] J. M. Borwein, G. Li and L. J. Yao. Analysis of the convergence rate for the cyclic projection algorithm applied to basic semialgebraic convex sets. *SIAM J. Optim.* 24, pp. 498–527 (2014).

[13] P. L. Combettes and J.-C. Pesquet. A Douglas-Rachford splitting approach to nonsmooth convex variational signal recovery. *IEEE J. Sel. Top. Signal Proces.* 1, pp. 564–574 (2007).

[14] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two or three space variables, *T. Am. Math. Soc.* 82, pp. 421–439 (1956).

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

[16] S. Gandy, B. Recht and I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Probl.* 27, 025010 (2011).

[17] P. Gong, C. Zhang, Z. Lu, J. Huang and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. The 30th International Conference on Machine Learning (ICML 2013).

[18] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM J. Numer. Anal.* 50, pp. 700–709 (2012).

[19] A. S. Lewis, D. R. Luke and J. Malick. Local convergence for alternating and averaged nonconvex projections. *Found. Comput. Math.* 9, pp. 485–513 (2009).

[20] G. Li and T. K. Pong. Global convergence of splitting methods for nonconvex composite optimization. Preprint (2014). Available at `http://arxiv.org/abs/1407.0753`.

[21] P.-L. Lions, B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.* 16, pp. 964–979 (1979).

[22] R. Hesse and D. R. Luke. Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems. *SIAM J. Optim.* 23, pp. 2397–2419 (2013).

[23] R. Hesse, D. R. Luke and P. Neumann. Alternating projections and Douglas-Rachford for sparse affine feasibility. *IEEE T. Signal. Proces.* 62, pp. 4868–4881 (2014).

[24] P. Patrinos, L. Stella and A. Bemporad. Douglas-Rachford splitting: complexity estimates and accelerated variants. Preprint (2014). Available at `http://arxiv.org/abs/1407.6723`.

[25] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis.* Springer (1998).

[26] D. Sun, K.-C. Toh and L. Yang. A convergent proximal alternating direction method of multipliers for conic programming with 4-block constraints. Preprint (2014).

[27] J. Zeng, S. Lin, Y. Wang and Z. Xu. $L_{1/2}$ regularization: convergence of iterative half thresholding algorithm. *IEEE Trans. Signal Process.* 62, pp. 2317–2329 (2014).