WEAKLY CONVEX DOUGLAS-RACHFORD SPLITTING AVOIDS STRICT SADDLE POINTS

FELIPE ATENAS*

Abstract. We prove that the Douglas-Rachford splitting method converges, almost surely, to local minimizers of semialgebraic weakly convex optimization problems, under the assumption of the strict saddle property. The approach consists of two steps: first, we prove a manifold identification result, and local smoothness of the involved iteration operator. Then, we proceed to show that strict saddle points are unstable fixed points of such operator, and thus the dynamics escape critical points of negative curvature. In this manner, Douglas-Ranchford splitting joins the family of *simple algorithms* that avoid saddle points, such as some first-order and proximal-type methods, including a close relative, the forward-backward splitting method.

Keywords. Douglas-Rachford splitting, weak convexity, semialgebraic, strict saddle, Moreau envelope

Mathematics Subject Classification. 65K05, 65K10, 90C26, 90C30

Funding. The author was supported in part by Australian Research Council grant DP230101749.

1. Introduction and Motivation

This work considers the following composite optimization problem

$$\min_{x\in\mathbb{R}^d} \varphi(x) = f(x) + g(x), \tag{1}$$

where $f, g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ are two functions, not necessarily convex. Problems with this type of structure appear in several applications in statistics, signal processing, and machine learning [Pen+16; PB+14; Boy+11].

One of the most common algorithms to solve problem (1) is the proximal point algorithm [Mar70; Roc76], see Section 3.2. In this context, applying this method amounts to compute a backward step to the sum $f + g$, in general an untractable operation. Splitting methods decompose the problem into simpler subproblems, circumventing the aforementioned difficulty. For problem (1), it means two separate subproblems are solved, one for each component function. Depending on the properties of the functions, the subproblems can be of proximal/backward type or of gradient/forward type.

Date: 18/1/2024

*School of Mathematics & Statistics, The University of Melbourne, Parkville VIC 3010, Australia (felipe.atenas@unimelb.edu.au).

We focus on the splitting method known as Douglas-Rachford (DR). Starting from $z^0 \in \mathbb{R}^d$ and given $\lambda \in (0,2)$, one iteration of the relaxed DR splitting method reads

$$\left\{ \begin{array}{rcl} x^k & = & \mathrm{prox}_{\gamma f}(z^k) \\ y^k & = & \mathrm{prox}_{\gamma g}(2x^k - z^k) \\ z^{k+1} & = & z^k + \lambda(x^k - y^k), \end{array} \right. \tag{2}$$

where prox denotes the proximal operator of a function, defined in (5) below.

In the convex case, the classical DR splitting method in (2) is known to converge to minimizers of (1). More specifically, for $\lambda = 1$, and f and g proper lsc convex functions, such that (1) has a nonempty solution set, and under a regularity assumption, the sequence $\{z^k\}$ converges to some $\bar{z}$, and $\{x^k\}$ converges to some $\bar{x}$ that solves (1), such that $\bar{x} = \mathrm{prox}_{\gamma f}(\bar{z})$ [Eck89, Theorem 3.15, Proposition 3.40]. In this setting, the convergence analysis fits into the framework of splitting methods for the sum of two maximal monotone operators [LM79], since the subdifferentials ∂f and ∂g are both maximal monotone. These results can be extended to finding a zero of the sum of two maximal monotone operators, see, for instance, [BC11, Theorem 25.6].

Customary arguments rely on the fact that the sequence $\{z^k\}$ in (2) is Féjer monotone with respect to the set of fixed points of the operator $(2\mathrm{prox}_{\gamma g} - I)(2\mathrm{prox}_{\gamma f} - I)$, since $\{z^k\}$ conforms a Krasnosel'skiĭ-Mann iteration scheme [BC11, Theorem 5.14]. Hence, for any fixed point $\bar{z}$ of $(2\mathrm{prox}_{\gamma g} - I)(2\mathrm{prox}_{\gamma f} - I)$, the sequence $\{\|z^k - \bar{z}\|\}$ is a nonincreasing sequence. As a result, the DR splitting method may not define a descent method in the usual sense, that is, it does not necessarily provide descent for the objective function φ.

For nonconvex problems, the subdifferentials of the involved functions are not necessarily maximal monotone, and thus the same line of reasoning cannot be employed. Instead, in [PSB14; TP20], a merit function is utilized to guide the convergence analysis. The advantage of this perspective is that such merit function is nonincreasing throughout the path defined by $\{z^k\}$, and thus the DR splitting method can be viewed as a method of descent for the merit function, even in nonconvex settings.

Depending on the assumptions for (1), the cluster points of the generated sequences may not solve the problem in the usual sense. In general, the expected behavior is that the sequences cluster at critical points. For convex problems, the limit of the sequence $\{x^k\}$ converges to a minimizer, since any critical point of a convex function is a global minimizer. It goes without saying that the condition $\nabla\varphi(x) = 0$ or $0 \in \partial\varphi(x)$ for nonconvex problems is no longer sufficient for global minimizers. Recent works have shown that more can be said under appropriate and rather generic assumptions. In the nonconvex smooth case, first-order methods for minimization of C^2 functions avoid critical points with negative curvature when randomly initialized [Lee+16; PP16]. For classical proximal splitting methods, this is also analyzed in the smooth case in [LY19].

In several modern applications, nonsmoothness rarely appears in an unstructured manner. This is further confirmed with the fact that numerical experiments show different types of methods in nonconvex optimization find critical points with function values not too far from a global minimum. This behavior suggests that classical methods are able to exploit the structure of the problem to produce critical points of *good quality*.

For nonsmooth nonconvex functions, as long as the problem satisfies the *strict saddle property* (cf. Section 5.4), some proximal-type methods provably converge to local minimizers almost surely [DD22], namely, the proximal point, the forward-backward (FB) splitting (see (10)) and the prox-linear methods. The key part of the analysis is to interpret these algorithms as a fixed-point iteration of a well-behaved operator, in such a way that fixed points of such operator are critical points of the associated problem. Moreover, strict saddle points corresponds to *unstable* fixed points of the operator, and thus a corollary of the Center-Stable manifold theorem yields the desired conclusion. In this work, we use the same type of arguments to address the same issue for the DR splitting method.

This paper is organized as follows. Section 2 starts defining the variational analysis concepts we use in relation to weak convexity, and the setting for smoothness of nonconvex value functions. We continue in Section 3 with proximal-type splitting methods, and define the respective envelopes inspired by their proximal point counterpart, the Moreau envelope. We study convergence properties of the FB and DR methods in Section 4 for nonconvex optimization problems. In particular, we establish that the corresponding generated sequences cluster in critical points of the original problem. The question of how to guarantee avoidance of saddle points, in particular for the DR splitting method, is addressed in Section 5, by resorting to the line of reasoning of [DD22].

2. Background material

A function $h : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is called proper when its domain is nonempty, that is, $\operatorname{dom}(h) := \{x \in \mathbb{R}^d : h(x) < +\infty\} \neq \emptyset$, and h is said to be lower semicontinuous (lsc) at $\bar{x}$ whenever $h(\bar{x}) \leq \liminf_{x\to\bar{x}} h(x)$, and lsc (on $\mathbb{R}^d$) if it is lsc at every $\bar{x}$. We say that the function h is level-bounded if it has bounded level sets, that is, if for any $\alpha \in \mathbb{R}$, the set $\{x \in \mathbb{R}^d : h(x) \leq \alpha\}$ is bounded.

A function h is locally Lipschitz continuous, if for all $\bar{x} \in \operatorname{dom}(h)$, there exists a neighborhood U of $\bar{x}$, and a constant $L = L(U) > 0$, such that for all $x, y \in U$, $|h(x) - h(y)| \leq L\|x - y\|$. By extension, we say h is (globally) Lipschitz continuous if the last estimate holds for any $x, y \in \mathbb{R}^d$, with a uniform constant $L > 0$ over the whole space.

Given a map $T : \mathbb{R}^d \to \mathbb{R}^d$, we say T is a lipeomorphism if T itself is globally Lipschitz continuous, and its inverse T^{-1} exists and is globally Lipschitz continuous as well. A point $\bar{z}$ is a fixed point of the operator T if $\bar{z} = T(\bar{z})$. If T is a C^1-smooth map around some fixed point $\bar{z}$, we say $\bar{z}$ is an unstable fixed point of T is the Jacobian $\nabla T(\bar{z})$ of T at $\bar{z}$ has at least one eigenvalue with magnitude strictly greater than one.

For nonconvex problems, the notion of global minimizer may be too strong. We say $\bar{x}$ is a local minimizer of h if there exists a neighborhood U of $\bar{x}$, such that for all $x \in U$, $h(\bar{x}) \leq h(x)$. Frequently used algorithms in nonconvex optimization can guarantee (sub)sequential convergence of the generated iterates to critical points, an even weaker notion. In order to define critical points, we need to first introduce the notion of subdifferential.

2.1. Subdifferentials and weak convexity.

For a proper lsc function $h : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$, the Fréchet subdifferential and the limiting subdifferential can be defined as in [RW09, Definition 8.3]. For a locally Lipschitz function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$

at $\bar{x}$, the Clarke subdifferential of h at $\bar{x}$ is defined as

$$\partial h(\bar{x}) = \overline{\text{co}} \left\{ \lim_{k \to +\infty} \nabla h(x^k) : x^k \to \bar{x}, \text{and } \nabla h(x^k) \text{ exists} \right\},$$

where $\overline{\text{co}}$ denotes the closed convex hull. Due to Rademacher's theorem, ∂h is well-defined.

We say $\bar{x} \in \mathbb{R}^d$ is a critical point of h if $0 \in \partial h(\bar{x})$, and $\bar{h} = h(\bar{x})$ is the corresponding critical value. For problems in composite form (1), $\bar{x}$ is a critical point if $0 \in \partial f(\bar{x}) + \partial g(\bar{x})$. Observe that, in particular, any local minimizer is a critical point, but the converse does not necessarily hold.

When h is convex, all these three notions of subdifferential coincide with the subdifferential of convex analysis:

$$\partial h(\bar{x}) = \{s \in \mathbb{R}^d : h(x) \geq h(\bar{x}) + \langle s, x - \bar{x} \rangle \text{ for all } x\}.$$

However, in general, different types of subdifferential at any $\bar{x}$ may not agree. For a bening form of nonconvexity (using the parlance adopted in [Wri20]), it is possible to prove that the above nonconvex subdifferential notions coincide. This is the case of the family of weakly convex functions. We say a function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is $\rho-$weakly convex, for $\rho > 0$, if $h + \frac{\rho}{2}\|\cdot\|^2$ is a convex function. In this case, all the three nonconvex subdifferentials above are equal [Cla90, Proposition 2.1.5(d)], [Kru03, Proposition 1.40], and also coincide with the notion of proximal subdifferential [RW09, Definition 8.45]. This relationship provides an alternative characterization of weak convexity: h is $\rho-$weakly convex, if and only if, for any $x, z \in \text{dom}(h)$, whenever $s \in \partial h(z)$,

$$h(x) + \frac{\rho}{2}\|x - z\|^2 \geq h(z) + \langle s, x - z \rangle.$$

2.2. **Smoothness of value functions.** Under some regularity assumptions on problem (1), the problems in (11) and (17) and the respective iterations maps are sufficiently smooth locally, thus enabling the use of arguments and techniques of smooth optimization. First, we formulate a problem defining a value function in an abstract manner, and study differentiability properties in that framework. Based on [DD22], we construct the setting of smooth minimization on manifolds, suited for parametric merit functions such as the ones in (11) and (17).

We say that a set $\mathcal{M} \subseteq \mathbb{R}^d$ is a C^2-smooth manifold around $\bar{x} \in \mathcal{M}$ [DD22, Definition 2.2], if there exist an open neighborhood U of $\bar{x}$, and a C^2 function G defined on $\mathbb{R}^d$, such that $\nabla G(\bar{x})$ has full rank, and $\mathcal{M} \cap U = \{x \in \mathbb{R}^d : G(x) = 0\}$. Intuitively, such $\mathcal{M}$ can be locally described around $\bar{x}$ as the solution of smooth equations with linearly independent gradients at $\bar{x}$. The system of equations $G = 0$ is called the local defining equations for $\mathcal{M}$ around $\bar{x}$, and $T_{\mathcal{M}}(\bar{x}) = \ker\big(\nabla G(\bar{x})\big)$ denotes the tangent space to $\mathcal{M}$ at $\bar{x}$.

Consider a C^2 map $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and a C^2-smooth manifold $\mathcal{M}$ around $\bar{y}$ with defining equations $G = 0$. For any $x \in \mathbb{R}^d$, define the value function

$$\Phi(x) = \inf\{\phi(x, y) : y \in \mathcal{M}\}, \tag{3}$$

and the Lagrangian associated with the constraint $y \in \mathcal{M}$:

$$\mathcal{L}(x, y, \lambda) = \phi(x, y) + \langle G(y), \lambda \rangle, \tag{4}$$

for a multiplier vector λ. Given $\bar{x}$, if $\bar{y}$ denotes a minimizer of $\phi(\bar{x}, \cdot)$ over $\mathcal{M}$, then there exists $\bar{\lambda}$ such that $\nabla_y \mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}) = 0$. When ϕ is sufficiently smooth, the value

function Φ and the solution map inherit some differentiability properties, as the following result shows. This statement summarizes [DD22, Theorem 2.3 & Lemma 2.4], that we later use in Section 5.2.

Proposition 1 (Local smoothness of solution maps and envelopes)**.** *Consider a proper lsc function $\phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, and a point $\bar{x} \in \mathbb{R}^d$. Suppose that for all x in a neighborhood of $\bar{x}$, $\phi(x, \cdot)$ is an $\alpha-$strongly convex function, for some $\alpha > 0$, such that its (unique) minimizer $y(x)$ defines a continuous function $y : \mathbb{R}^d \to \mathbb{R}^d$. Then, for all x near $\bar{x}$,*

1. *$y(x)$ is a strong global minimizer of $\phi(x, \cdot)$.*
2. *There exists a neighborhood V of $\bar{x}$, such that for all $m > \phi(\bar{x}, y(\bar{x}))$, the following set is bounded:*

$$\bigcup_{x \in V} \{\tilde{y} \in \mathbb{R}^d : \ \phi(x, \tilde{y}) \leq m\}.$$

If, in addition, ϕ is a C^2 function, and $\mathcal{M}$ is a C^2-smooth manifold around $\bar{y} = y(\bar{x})$ with local defining equations $G = 0$, the following hold for the value function and Lagrangian defined in (3)-(4), respectively, and the (partial) Hessian matrices:

$$H_{xx} = \nabla^2_{xx}\mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}), \quad H_{xy} = \nabla^2_{xy}\mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}), \quad H_{yy} = \nabla^2_{yy}\mathcal{L}(\bar{x}, \bar{y}, \bar{\lambda}).$$

3. *The function y is locally C^1 around $\bar{x}$, such that*

$$\nabla y(\bar{x})h = \underset{u \in T_{\mathcal{M}}(\bar{y})}{\arg\min}\{2\langle H_{xy}u, h\rangle + \langle H_{yy}u, u\rangle\}$$

4. *The function Φ is C^2 around $\bar{x}$, such that $\nabla\Phi(\bar{x}) = \nabla_x\phi(\bar{x}, \bar{y})$, and*

$$\langle \nabla^2\Phi(\bar{x})h, h\rangle = \langle H_{xx}h, h\rangle + \min_{u \in T_{\mathcal{M}}(\bar{y})}\{2\langle H_{xy}u, h\rangle + \langle H_{yy}u, u\rangle\}.$$

We end this section by introducing the concept of active manifold, related to the identification results in Section 5.2, and the saddle point avoidance property analyzed in Section 5.4.

Let $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a proper lsc $\rho-$weakly convex function, $\bar{x} \in \mathbb{R}^d$ a critical point of h, and a set $\mathcal{M} \ni \bar{x}$. The set $\mathcal{M}$ is called an active C^2-smooth manifold around $\bar{x}$ [DD22, Definition 2.6], if there exists a neighborhood U of $\bar{x}$, such that the following two properties hold.

1. Smoothness: $\mathcal{M} \cap U$ is a C^2-smooth manifold around $\bar{x}$, such that h is a C^2 function on $\mathcal{M} \cap U$.
2. Sharpness: $\inf\{\|s\| : s \in \partial h(x),\ x \in U \setminus \mathcal{M}\} > 0$.

The sharpness condition essentially states that normal to the manifold, the function cannot be flat, that is, the norm of subgradients at points outside $\mathcal{M}$ are bounded away from 0.

3. Proximal maps and envelopes for prox-type methods

We next define the cornerstone of the DR method, the proximal operator. We also define a merit function tailored for the DR method, resembling the so-called Moreau envelope.

3.1. **Proximal operator and Moreau envelope.** Given a point $z \in \mathbb{R}^d$ and a proper function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, the proximal opetator of h at z with prox-parameter $\gamma > 0$ is defined as

$$\operatorname{prox}_{\gamma h}(z) = \operatorname*{arg\,min}_{x \in \mathbb{R}^d} \left\{ h(x) + \frac{1}{2\gamma}\|x - z\|^2 \right\}. \tag{5}$$

The optimal value of this minimization problem defines the Moreau envelope $e_\gamma h(z)$ of h at z with prox-parameter $\gamma > 0$:

$$e_\gamma h(z) = \inf_{x \in \mathbb{R}^d} \left\{ h(x) + \frac{1}{2\gamma}\|x - z\|^2 \right\}. \tag{6}$$

Both expressions in (5) and (6) are well-defined whenever h is $\mu-$prox-bounded, that is, when $h + (2\mu)^{-1}\|\cdot\|^2$ is bounded from below. Examples of functions in the prox-bounded family are convex and weakly convex functions.

When h is $\rho-$weakly convex, the Moreau envelope and the proximal map have remarkable continuity properties. More specifically, for any $\gamma \in (0, \rho^{-1})$, $\operatorname{prox}_{\gamma h}$ is Lipschitz continuous with constant $\rho/(\rho - \gamma)$ [RW09, Proposition 12.19], and $e_{\gamma h}$ is differentiable with gradient given by

$$\nabla e_{\gamma h}(z) = \frac{1}{\gamma}\big(z - \operatorname{prox}_{\gamma h}(z)\big), \tag{7}$$

in such a way that $\nabla e_{\gamma h}$ is Lipschitz continuous with constant $\max\{\rho(1-\rho\gamma)^{-1}, \gamma^{-1}\}$ [HLO, Corollary 3.4]. In this fashion, $e_{\gamma h}$ can be seen as a continuously differentiable smoothing of h. In the convex case, the Moreau envelope is also known as the Moreau-Yosida regularization [HL13, Ch. XI, Example 3.4.4].

3.2. **Proximal point algorithm.** The proximal operator defines the fixed-point iteration known as proximal point algorithm (PPA). For a $\rho-$weakly convex function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, consider the problem

$$\min_{x \in \mathbb{R}^d} h(x). \tag{8}$$

For $\gamma \in (0, \rho^{-1})$, a sequence $\{z^k\}$ is generated by the PPA if

$$z^{k+1} = \operatorname{prox}_{\gamma h}(z^k).$$

When h is convex with a nonempty set of minimizers, for any $\gamma > 0$, $\{z^k\}$ converges to a solution to (8) [Roc76, Theorem 1]. When h is $\rho-$weakly convex and $\gamma \in (0, \rho^{-1})$, $\{z^k\}$ subsequentially converges to a critical point of h, that is, all cluster points of $\{z^k\}$ are critical points of h [HLO, Proposition 5.1]. Due to this last result, the authors of this last article use smoothness of the Moreau envelope to reinterpret the PPA. In view of (7), the proximal point iteration scheme can be reformulated as

$$z^{k+1} = z^k - \gamma \nabla e_{\gamma h}(x^k). \tag{9}$$

Thus, the PPA applied to h corresponds to the gradient descent method for the Moreau envelope $e_{\gamma h}$ with fixed stepsize γ. In this way, classical convergence results can be retrieved through this point of view, see [HL96, Ch. XV, Sec. 4.2] and [Dru17] for more details. When h is convex, one notable property of the Moreau envelope $e_\gamma h$ is that it preserves the minimizers (if any) of h and the fixed-points of $\operatorname{prox}_{\gamma h}$ [HL96, Ch. XV, Theorem 4.1.7]:

$$\bar{x} \text{ minimizes } h \iff \bar{x} \text{ minimizes } e_{\gamma h} \iff \operatorname{prox}_{\gamma h}(\bar{x}) = \bar{x}.$$

Furthermore, in this case, $\inf h = \inf e_{\gamma h}$. This last property also holds when h is $\rho-$weakly convex and $\gamma \in (0, \rho^{-1})$. In order to guarantee that the cluster points of the generated sequence are local minimizers of problem (8), we need further assumptions, as discussed in Section 5.

3.3. **Proximal splitting methods.** For composite problems in the form of (1), splitting methods of proximal type exploit the special structure of the objective function by performing separate gradient or proximal steps, one for each component. In this work, we will discuss two of these methods: the FB splitting method, also known as proximal-gradient method, and the DR splitting method. Naturally, these methods inherit some of the properties discussed for the proximal point algorithm. We first start discussing the FB method.

3.3.1. *Forward-backward splitting method.* . The FB method is a simple and direct extension of the PPA. Assuming that f is differentiable, first a gradient step is performed for f, and then a proximal step is made for g. More precisely, let $f : \mathbb{R}^d \to \mathbb{R}$ be a C^1 function with L_f-Lipschitz continuous gradient, and $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lsc $\rho-$weakly convex function. For $\gamma \in (0, \rho^{-1})$, a sequence $\{z^k\}$ is generated by the FB method if

$$z^{k+1} = \operatorname{prox}_{\gamma g}\big(z^k - \gamma \nabla f(z^k)\big). \tag{10}$$

Observe that the iterates defined in (10) can be equivalently computed by solving

$$\inf_{y \in \mathbb{R}^d} \left\{ f(z^k) + \langle \nabla f(z^k), y - z^k \rangle + g(y) + \frac{1}{2\gamma}\|y - z^k\|^2 \right\}.$$

The value function associated with this optimization problem defines the forward-backward envelope (FBE) [TSP18]:

$$\varphi_\gamma^{\mathrm{FB}}(z) = \inf_{y \in \mathbb{R}^d} \left\{ f(z) + \langle \nabla f(z), y - z \rangle + g(y) + \frac{1}{2\gamma}\|y - z\|^2 \right\}. \tag{11}$$

The FBE plays an analogous role for the FB method, as the Moreau envelope for the PPA: the solution map $z \mapsto \operatorname{prox}_{\gamma g}(z - \gamma \nabla f(z))$ is an adaptation of the proximal map (5), while the FBE is for the Moreau envelope (6), both tailored to the composite optimization problem (1) and the FB method.

As introduced in [PB13], when $f : \mathbb{R}^d \to \mathbb{R}$ is a twice continuously differentiable function with L_f-Lipschitz continuous gradient, and $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a proper lsc convex function, then $\varphi_\gamma^{\mathrm{FB}}$ is continuously differentiable with explicit gradient given by [STP17, Theorem 2.6]

$$\nabla \varphi_\gamma^{\mathrm{FB}}(z) = \gamma^{-1}\Big(I - \gamma \nabla^2 f(z)\Big)\Big(z - \operatorname{prox}_{\gamma g}\big(z - \gamma \nabla f(z)\big)\Big). \tag{12}$$

If $\gamma \in (0, L_f^{-1})$, then $\arg\min \varphi_\gamma^{\mathrm{FB}} = \arg\min \varphi$ and $\inf \varphi_\gamma^{\mathrm{FB}} = \inf \varphi$ [TSP18, Theorem 4.4]. Note that similar to the Moreau envelope, $\varphi_\gamma^{\mathrm{FB}}$ is a real-valued smoothing of φ, although it may fail to be convex.

Under weaker assumptions, namely when $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is only proper lsc $\mu-$prox-bounded, then $\varphi_\gamma^{\mathrm{FB}}$ is locally Lipschitz, and the relationship between minimizers and minimal values of $\varphi_\gamma^{\mathrm{FB}}$ and φ holds for $\gamma \in (0, \min\{L_f^{-1}, \mu\})$ [TSP18, Theorem 4.4]. In the latter case, although we may not be able to reformulate (10) as a gradient-like step using (12) as in (9) for the PPA, there still holds a sufficient

descent property that is the keystone to prove convergence results of the FB method in the nonconvex setting [TSP18, Proposition 4.3]:

$$\varphi_\gamma^{\mathrm{FB}}(z^{k+1}) + \frac{1-\gamma L_f}{2\gamma}\|z^{k+1}-z^k\|^2 \leq \varphi_\gamma^{\mathrm{FB}}(z^k). \tag{13}$$

Similar to the case of the PPA, without further regularity assumptions, convergence guarantees solely involve critical points of φ. The theory of avoidance of saddle points for the FB method is discussed in [DD22] and Section 5.

Much of the properties of the FB method and the FBE also hold true for the DR method, as we next discuss.

3.3.2. *Douglas-Rachford splitting method.* The DR method in (2) consists of two consecutive proximal steps, one for each function separately, and then a coordination-correction step. The sequences $\{x^k\}$ and $\{y^k\}$ represent copies of the same variable in problem (1), and the third step in (2) is a fixed-point iteration. In this way, we define the following iteration maps for the DR method. Given $z \in \mathbb{R}^d$, define

$$R_\gamma(z) = \mathrm{prox}_{\gamma g}(2\mathrm{prox}_{\gamma f}(z) - z), \tag{14}$$

and

$$S_\gamma(z) = z + \lambda\big(R_\gamma(z) - \mathrm{prox}_{\gamma f}(z)\big). \tag{15}$$

Hence, the DR scheme in (2) can be rewritten as the following fixed point iteration:

$$z^{k+1} = S_\gamma(z^k). \tag{16}$$

Observe that $y = R_\gamma(z)$ defined in (14) is the solution of the following problem, defining the Douglas-Rachford envelope (DRE) at z:

$$\varphi_\gamma^{\mathrm{DR}}(z) = \inf_{y\in\mathbb{R}^d}\left\{f\big(x(z)\big) + \langle \nabla f\big(x(z)\big), y - x(z)\rangle + g(y) + \frac{1}{2\gamma}\|y-x(z)\|^2\right\}, \tag{17}$$

where $x(z) = \mathrm{prox}_{\gamma f}(z)$. For $z = z^k$, problem (17) is equivalent to the problem solved when evaluating the second step in (17). Like the FBE, the DRE is an extension of the Moreau envelope adjusted to problem (21) and the DR method, and the maps R_γ and S_γ extend the proximal operator.

In principle, $\varphi_\gamma^{\mathrm{DR}}$ may not be well-defined in the general case. As shown in [TP20, Remark 3.1 & Proposition 3.2], if f is continuously differentiable with L_f-Lipschitz gradient, $g : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is proper and lsc, such that (1) has a nonempty set of minimizers, for any $\gamma \in (0, L_f^{-1})$, both f and g are prox-bounded, and $\varphi_\gamma^{\mathrm{DR}}$ is real-valued and locally Lipschitz continuous. Due to Rademacher's theorem, the DRE is almost everywhere differentiable, like the FBE. When, in addition, f is convex twice continuously differentiable and g proper lsc convex, the $\varphi_\gamma^{\mathrm{DR}}$ is differentiable, with a gradient given by a closed-formula (cf. (29)). Under weaker assumptions, just as for the FBE, we may not have the interpretation of DR as a gradient-like step, but we still have a sufficient decrease estimate that ultimately yields subsequential convergence of the iterates in the nonconvex setting. For any $\gamma \in (0, (2-\lambda)(2L_f)^{-1})$, then for some constant $c > 0$ depending on λ, γ and L_f, it holds [TP20, Themorem 4.1]:

$$\varphi_\gamma^{\mathrm{DR}}(z^{k+1}) + \frac{c}{(1+\gamma L_f)^2}\|z^{k+1}-z^k\|^2 \leq \varphi_\gamma^{\mathrm{DR}}(z^k). \tag{18}$$

Furthermore, regarding boundedness of level sets, in view of [TP20, Theorem 3.4(iii)], φ is level-bounded if and only if $\varphi_\gamma^{\mathrm{DR}}$ is level-bounded. It can be proven similarly that the same holds for $\varphi_\gamma^{\mathrm{FB}}$. This property is useful in Proposition 2 to prove that the sequences generated by the FB and DR splitting methods are bounded.

Note that both FBE and DRE can be cast, under appropriate assumptions, in the form of problem (3). Before going in depth into smoothness of these envelopes and its consequences, in the next section we first review how the FBE and the DRE are used to analyze convergence of the respective splitting methods in nonconvex optimization.

4. Convergence of splitting methods in nonconvex optimization

As mentioned above, most commonly used methods for nonconvex optimization can guarantee theoretical subsequential convergence to critical points, under appropriate assumptions. This corresponds to the first step in the convergence analysis of such method. In this work, we recall in Section 4.1, the conditions for which the FB and the DR methods generate sequences whose cluster points are critical points of problem (1). Then, in Section 4.2, we comment on standard assumptions in the literature that guarantee global convergence of such sequences to a unique limit. In particular, we discuss two conditions: a subdifferential-based error-bound and the Kurdyka-Łojasiewicz (KŁ) inequality.

4.1. Subsequential convergence.

We already stated that for weakly convex functions, the PPA generates a sequence that subsequentially converges to critical points of the problem at hand, for a sufficiently small prox-parameter. We next analyze the FB and DR splitting methods along the same lines. First, we state the assumptions for which we study subsequential convergence.

Assumption 1. *Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable function with L_f-Lipschitz continuous gradient, and $g : \mathbb{R}^d \to \mathbb{R}\cup\{+\infty\}$ is a proper lsc $\rho-$weakly convex function. Moreover, suppose $\varphi = f + g$ has a nonempty set of minimizers, and is level-bounded.*

Note that under Assumption 1, any critical point $\bar{x}$ of φ is characterized by $0 \in \nabla f(\bar{x}) + \partial g(\bar{x})$ [RW09, Exercise 8.8]. This is equivalent to the existence of $\bar{s} \in \partial g(\bar{x})$ such that $\bar{s} = -\nabla f(\bar{x})$.

The next result states that the FB and the DR methods subsequentially converge to critical points of problem (1). In [TP20, Theorem 4.3], the authors analyze the DR method, and the case of the FB is analogous. In the appendix, a summary of the proof can be found for completeness.

Proposition 2 (Subsequential convegrence to critical points)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ comply to Assumption 1. Then,*

- *If $\{z^k\}$ is generated by the FB method (10) and $\gamma \in \big(0, \min\{L_f^{-1}, \rho^{-1}\}\big)$, then all cluster points of $\{z^k\}$ are critical points for problem (1).*
- *If $\{(x^k, y^k, z^k)\}$ is generated by the DR method (2) and $\gamma \in (0, \min\{(2-\lambda)(2L_f)^{-1}, \rho^{-1}\})$, then for all cluster points $\bar{z}$ of $\{z^k\}$, $\bar{x} = prox_{\gamma f}(\bar{z})$ is a cluster point of both $\{x^k\}$ and $\{y^k\}$, and is a critical point for problem (1).*

Remark 1. *The authors in [TP20, Theorem 4.3] prove the results of Proposition 2 for the DR splitting method when g is prox-bounded, a more general case. In the statement above, we assume that g is weakly convex, because this is the standing assumption in the following saddle point avoidance results (see Section 5, cf. [DD22]).*

4.2. **Global convergence in nonconvex case.** Once subsequential convergence to critical points is established, the next question is under what assumptions the method globally converges to a unique limit point. Here, we present two closely related approaches. The first one, examined in [Ate23], needs a subdifferential-based error bound for φ, an estimate for the distance from a point to the set of critical points, and a condition on critical values for close enough critical points.

Assumption 2. *Suppose that φ satisfies the following conditions.*

(1) *Error bound: for any $\bar{\varphi} > \inf \varphi$, there exist $\varepsilon, \ell > 0$, such that whenever $\varphi(x) \leq \bar{\varphi}$, and $s \in \partial\varphi(x) \cap B(0,\delta)$ for some $\delta > 0$, it holds*
$$dist\big(x, (\partial\varphi)^{-1}(0)\big) \leq \ell\|s\|.$$
(2) *Proper separation of isocost surfaces: there exist $\epsilon > 0$, such that whenever $\bar{x}, \bar{y} \in (\partial\varphi)^{-1}(0)$, $\|\bar{x} - \bar{y}\| < \epsilon$, then $\varphi(\bar{x}) = \varphi(\bar{y})$.*

Classical settings where Assumption 2 holds include strongly convex problems, the sum of a linear operator and the composition of a strongly convex function with Lipschitz continuous gradient with a linear operator, and the Fenchel conjugate of a strongly convex differentiable function with Lipschitz continuous gradient [LT93, Theorem 2.1].

The second approach corresponds to assuming the KŁ inequality holds [ABS13] around critical points:

Assumption 3. *Assume that φ satisfies the KŁ inequality with exponent $\theta \in [0,1)$ at $\bar{x}$, that is, there exist constants $c, \varepsilon, \eta > 0$, such that for all $x \in B(0,\varepsilon) \cap [f(\bar{x}) < f < f(\bar{x}) + \eta]$,*
$$dist\big(0, \partial f(x)\big) \geq c\big(f(x) - f(\bar{x})\big)^{\theta}.$$

Define $\Gamma(s) = \frac{c}{1-\theta} s^{1-\theta}$, so that the KŁ inequality with exponent θ can be reformulated as
$$\operatorname{dist}\big(0, \partial(\Gamma \circ \varphi)(x)\big) \geq 1.$$
From this estimate, we can interpret that the KŁ inequality states that φ is sharp up to a reparametrization via Γ.

One standard condition for Assumption 3 to hold is that f and g are semialgebraic [BDL07, Theorem 3.1], or more generally, when they are definable in o-minimal structures [Bol+07, Theorem 14]. Semialgebraic functions are characterized by their graphs being the solution of a finite system of polynomial inequalities. Furthermore, in view of [LP18, Theorem 4.1], Assumption 2 implies Assumption 3 with exponent $\theta = 1/2$. Another relationship between error bounds and the KŁ inequality can be found in [Bol+17, Theorem 5].

The next result is akin to [ABS13; Bol+17; Ate+23; AAS24] and references therein. The proof of convergence of DR splitting under Assumption 2 can be found in [Ate23]. The case for FB splitting is similar by resorting to the same type of arguments. A proof of global convergence of FB under Assumption 3 without resorting to envelopes can be found in [ABS13, Theorem 5.1], while for the DR

method the same type of result was obtained in [LP16] using a different type of merit function.

Theorem 1 (Global convergence of proximal splitting methods)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ comply to Assumption 1, and either satisfy Assumption 2 or Assumption 3. Then,*

- *If $\{z^k\}$ is generated by the FB method (10) and $\gamma \in \big(0, \min\{L_f^{-1}, \rho^{-1}\}\big)$, then $\{z^k\}$ converges to a critical point $\bar{z}$ of problem (1).*
- *If $\{(x^k, y^k, z^k)\}$ is generated by the DR method (2) and $\gamma \in (0, \min\{(2-\lambda)(2L_f)^{-1}, \rho^{-1}\})$, then both $\{x^k\}$ and $\{y^k\}$ converge to a critical point $\bar{x}$ for problem (1), $\{z^k\}$ converges to $\bar{z}$ such that $\bar{x} = prox_{\gamma f}(\bar{z})$.*

A consequence of Theorem 1 is that, for the DR method, $\bar{x} = R_\gamma(\bar{z})$. Indeed, whenever $\gamma \in (0, \min\{(2-\lambda)(2L_f)^{-1}, \rho^{-1}\})$, $\mathrm{prox}_{\gamma g}$ is a continuous operator. Therefore, $y^k = \mathrm{prox}_{\gamma g}\big(2x^k - z^k\big) \to \mathrm{prox}_{\gamma g}(2\bar{x} - \bar{z}) = \mathrm{prox}_{\gamma g}\big(2\mathrm{prox}_{\gamma g}(\bar{z}) - \bar{z}\big) = R_\gamma(\bar{z})$. Because $\{x^k\}$ and $\{y^k\}$ have the same limit under the assumptions of Theorem 1, then $\bar{x} = R_\gamma(\bar{z})$.

Since the setting for global convergence to critical points is established, in the next section we investigate some conditions yielding results of manifold identification, which later leads us to the theory of saddle point avoidance.

5. Saddle point avoidance of Douglas-Rachford splitting method

Throughout this section, we focus our attention on the DR method. The authors in [DD22] prove analogous results for the FB method. In turn, the latter work replicates the ideas of the smooth case of [Lee+16; Lee+19] for some first-order methods. The core of the analysis is to interpret a certain first-order method as a dynamical system that avoids eventually, almost surely, unstable fixed points. More precisely, as explained in [DD22], given a locally C^1-smooth map $T : \mathbb{R}^d \to \mathbb{R}^d$ around an unstable fixed point $\bar{z}$ in a neighborhood U, if the Jacobian $\nabla T(\bar{z})$ is invertible, then the set of initial points from which the dynamics stay close to $\bar{z}$

$$\{z \in U : T^k(z) \in U \text{ for } k \geq 1\} \text{ has zero Lebesgue measure.} \tag{19}$$

This is a consequence of the center stable-manifold theorem [Shu13, Theorem III.7].

The relationship between this approach and splitting methods is trough envelopes. As we discussed in Section 3.3, both the FB and DR methods enjoy properties of first-order methods, even in the nonconvex nonsmooth case, via their respective envelopes. Therefore, the underlying iteration map associated with such envelopes supplies the desired interpretation of these methods as dynamical systems, from which avoidance of saddle points of the original objective function can be derived.

We follow a two-step argument. First, we establish local smoothness of the iteration mapping near saddle points of our interest, despite φ begin nonsmooth. Then, we prove that such saddle points are unstable fixed points of the iteration mapping.

All the results of the upcoming sections assume the following standing condition for the stepsize γ:

$$\gamma \in \left(0, \min\left\{\frac{1}{L_f}, \frac{1}{\rho}\right\}\right). \tag{20}$$

5.1. **Characterization of critical points.** In order to follow the reasoning of [Lee+16; Lee+19] in a nonsmooth nonconvex setting, we first need to characterize critical points of φ in relation to the iteration maps and critical points of the envelopes, since ultimately both the iteration maps and envelopes are locally smooth whenever φ admits a smooth manifold around critical points. Furthermore, small neighborhoods of fixed points of the iteration maps are mapped onto such smooth manifold, a result known as manifold identification [HL07].

For the FB method, strict saddle points of φ correspond to unstable fixed points of the iteration map $z \mapsto \operatorname{prox}_{\gamma g}\big(z - \gamma \nabla f(z)\big)$, for sufficiently small $\gamma > 0$ [DD22, Theorem 4.1]. Our goal is to prove an analogous result for the DR method.

In the next result, we start by showing that critical points of φ correspond to fixed points of the DR method iteration map S_γ, through a proximal operation.

Proposition 3 (Characterization of critical points I)**.** *For any $\gamma > 0$ satisfying (20), $\bar{x}$ is a critical point of φ if and only if there exists $\bar{z}$ such that $\bar{x} = prox_{\gamma f}(\bar{z})$ and $\bar{z}$ is a fixed point of S_γ.*

Proof. For any critical point $\bar{x}$ of φ, there exists $\bar{s} \in \partial g(\bar{x})$ such that $\bar{s} = -\nabla f(\bar{x})$. Define $\bar{z} = \bar{x} - \gamma \bar{s}$, then $\nabla f(\bar{x}) = -\bar{s} = \gamma^{-1}(\bar{z} - \bar{x})$ implies $\bar{x} = \operatorname{prox}_{\gamma f}(\bar{z})$. Therefore,

$$\begin{aligned}
& 0 \in \nabla f(\bar{x}) + \partial g(\bar{x}) \\
\iff \quad & 0 \in \gamma^{-1}(\bar{z} - \bar{x}) + \partial g(\bar{x}) \\
\iff \quad & 0 \in \gamma^{-1}(\bar{x} - (2\bar{x} - \bar{z})) + \partial g(\bar{x}) \\
\iff \quad & \bar{x} = \operatorname{prox}_{\gamma g}(2\bar{x} - \bar{z}) \\
\iff \quad & \bar{x} = R_\gamma(\bar{z}) \\
\iff \quad & \bar{z} = S_\gamma(\bar{z}).
\end{aligned}$$

□

In order to relate critical points of the envelope $\varphi_\gamma^{\mathrm{DR}}$ with the results of Proposition 3, we would need to characterize the (Clarke) subdifferential of $\varphi_\gamma^{\mathrm{DR}}$, a task that can be considered cumbersome. Instead, here we exploit local smoothness of the problem under mild regularity assumptions to deduce smoothness of the envelopes, and thus reducing the subdifferential of $\varphi_\gamma^{\mathrm{DR}}$ to a singleton. One of the fundamental properties yielding local smoothness of these operators is manifold identification, a concept discussed in the following.

5.2. **Manifold identification.** In modern applications, nonsmoothness frequently appear in a structured way, see [HL07; Lew02; LOS00; Wri93] and references therein. The key assumption for identification results is the existence of an active manifold around a critical point, where the function is smooth, while in normal directions it is sharp (cf. the concept of a partly smooth function in [Lew02]). Since f is itself smooth, we only assume such active manifold exists for g. Recall that $\bar{x}$ denotes a critical point of φ.

Assumption 4. *Suppose g admits a C^2-smooth active manifold $\mathcal{G}$ at $\bar{x}$, with defining equations $G = 0$.*

It turns out that Assumption 4 is satisfied by a large class of functions, as proven in [DD22, Theorem 2.9] and [DIL16, Theorem 4.16]. This family of functions include lsc weakly convex semialgebraic functions.

In [HL07] and [DD22, Theorem 4.1], the authors prove that under Assumptions 4, for all x near $\bar{x}$ and $\gamma \in (0, \rho^{-1})$, $\operatorname{prox}_{\gamma g}\big(x - \gamma \nabla f(x)\big) \in \mathcal{M}$. The following result

is an analogous statement for the DR method. In short, this method identifies the active manifold due to the sharpness condition.

Proposition 4 (Manifold identification)**.** *Consider the optimization problem* (1)*, and let* $\bar{x}$ *be a critical point of* φ*. Under Assumption* 4*, as long as* $\gamma > 0$ *satisfies* (20)*, there exists* $\bar{z}$ *such that* $\bar{x} = prox_{\gamma f}(\bar{z})$*, and for all* z *near* $\bar{z}$ *and* γ*,* $R_\gamma(z)$ *lies in* $\mathcal{G}$*.*

Proof. The existence of such $\bar{z}$ is guaranteed by Proposition 3. Let $\{z^i\}$ be a sequence such that $z^i \to \bar{z}$. The continuity of the proximal operators of f and g imply $x^i := \mathrm{prox}_{\gamma f}(z^i) \to \mathrm{prox}_{\gamma f}(\bar{z}) = \bar{x}$, and $y^i := R_\gamma(z^i) \to R_\gamma(\bar{z}) = \bar{x}$.

In view of the optimality condition of the problem in (17) at $z = z^i$, there exists $s^i \in \partial g(y^i)$ such that

$$s^i = -\nabla f(x^i) - \frac{1}{\gamma}(y^i - x^i).$$

Taking the limit, it follows that $s^i + \nabla f(x^i) \to 0$. Next, since ∇f is L_f-Lipschitz continuous, then

$$\begin{array}{rcl} \|\nabla f(y^i) + s^i\| & \leq & \|\nabla f(y^i) - \nabla f(x^i)\| + \|\nabla f(x^i) + s^i\| \\ & \leq & L_f\|y^i - x^i\| + \|\nabla f(x^i) + s^i\|. \end{array}$$

Taking the limit, it follows $\mathrm{dist}\big(0, \nabla f(y^i) + \partial g(y^i)\big) \to 0$. Therefore, in view of Assumption 4, the sharpness condition implies $y^i \in \mathcal{G}$ for all sufficiently large i. □

This identification result can be employed to relate the optimality conditions of problems (1) and (17). Under the assumptions of Proposition 4, problem (1) can be reformulated as the following constrained optimization problem:

$$\min_{x \in \mathcal{G}} \varphi(x) = f(x) + g(x). \tag{21}$$

Recall that $G = 0$ are the local defining equations for $\mathcal{G}$ around the critical point $\bar{x}$ of problem (1), and let $\hat{g} : \mathbb{R}^d \to \mathbb{R}$ be any C^2-smooth function that locally agrees with g on $\mathcal{G}$. Define the Lagrangian of problem (21) as

$$\mathcal{L}_0(x, \lambda) = f(x) + \hat{g}(x) + \langle G(x), \lambda \rangle. \tag{22}$$

In this manner, there exists a multiplier $\bar{\lambda} \in \mathbb{R}^m$ such that

$$\begin{array}{rcl} 0 & = & \nabla_x \mathcal{L}_0(\bar{x}, \bar{\lambda}) \\ & = & \nabla f(\bar{x}) + \nabla \hat{g}(\bar{x}) + \sum_{i \geq 1} \bar{\lambda}_i \nabla G_i(\bar{x}). \end{array}$$

Furthermore, since $R_\gamma(\bar{z}) = \bar{x}$, from Proposition 4, for all z close to $\bar{z}$, the problem in (17) is equal to

$$\min_{y \in \mathcal{G}} \left\{ f\big(\mathrm{prox}_{\gamma f}(z)\big) + \langle \nabla f\big(\mathrm{prox}_{\gamma f}(z)\big), y - \mathrm{prox}_{\gamma f}(z) \rangle + \hat{g}(y) + \frac{1}{2\gamma}\|y - \mathrm{prox}_{\gamma f}(z)\|^2 \right\}. \tag{23}$$

Define the parametric Lagrangian associated with (23) as:

$$\begin{array}{rcl} \mathcal{L}(z, y, \lambda) & = & f\big(\mathrm{prox}_{\gamma f}(z)\big) + \langle \nabla f\big(\mathrm{prox}_{\gamma f}(z)\big), y - \mathrm{prox}_{\gamma f}(z) \rangle + \hat{g}(y) \\ & & + \dfrac{1}{2\gamma}\|y - \mathrm{prox}_{\gamma f}(z)\|^2 + \langle G(y), \lambda \rangle. \end{array} \tag{24}$$

The optimality condition of problem (23) written in terms of the parametric Lagrangian are:

$$\begin{aligned} 0 &= \nabla_y \mathcal{L}(\bar{z}, \bar{x}, \bar{\lambda}) \\ &= \nabla f(\mathrm{prox}_{\gamma f}(\bar{z})) + \nabla \hat{g}(\bar{x}) + \frac{1}{\gamma}\big(\bar{x} - \mathrm{prox}_{\gamma f}(\bar{z})\big) + \textstyle\sum_{i\geq 1} \bar{\lambda}_i \nabla G_i(\bar{x}) \\ &= \nabla f(\bar{x}) + \nabla \hat{g}(\bar{x}) + \textstyle\sum_{i\geq 1} \bar{\lambda}_i \nabla G_i(\bar{x}), \end{aligned} \tag{25}$$

for some multiplier $\bar{\lambda}$. Observe that these optimality conditions coincide with the ones for problem (21).

In the next section, we exploit the manifold identification result to infer local smoothness of the iteration map and the envelope associated with the DR method.

5.3. **Smoothness of envelopes and iterations maps.** In accordance to Assumption 4, we need the following refinement of Assumption 1 in order to deduce local smoothness of the iteration map.

Assumption 5. *Additionally to the conditions of Assumption 1, suppose $f : \mathbb{R}^d \to \mathbb{R}$ is a C^2 function.*

Since the iteration map is defined based on the proximal operator, the next result is the first step towards establishing smoothness of S_γ and $\varphi_\gamma^{\mathrm{DR}}$.

Lemma 1 (Smoothness of proximal operator)**.** *Suppose Assumption 5 holds, and $\gamma > 0$ satisfies (20). Then, $prox_{\gamma f}$ is a continuously differentiable map and the Jacobian $\nabla prox_{\gamma f}(z)$ of $prox_{\gamma f}$ at z is a symmetric positive definite matrix.*

Proof. Choose any point $\bar{z} \in \mathbb{R}^d$. Since f is a C^2 function with $L-$Lipschitz continuous gradient, then for any $z \in \mathbb{R}^d$, $\phi(z, \cdot) = f(\cdot) + \frac{1}{2\gamma}\|\cdot - z\|^2$ is a strongly convex function, since $\gamma^{-1} > L_f$, and ϕ is C^2-smooth. Furthermore, $\mathrm{prox}_{\gamma f}$ is continuous, because f is, in particular, weakly convex. Taking $\mathcal{M} = \mathbb{R}^d$ in Proposition 1 yields that f^γ is C^2-smooth and $\mathrm{prox}_{\gamma f}$ is C^1-smooth around $\bar{z}$. Furthermore, in view of [RW09, 10.32 Example], for all z in a neighborhood of $\bar{z}$,

$$\nabla f^\gamma(z) = \gamma^{-1}\big(z - \mathrm{prox}_{\gamma f}(z)\big).$$

Therefore,

$$\nabla \mathrm{prox}_{\gamma f}(z) = I - \gamma \nabla^2 f^\gamma(z). \tag{26}$$

In particular, $\nabla \mathrm{prox}_{\gamma f}(z)$ is a symmetric matrix, and $\nabla \mathrm{prox}_{\gamma f}$ is continuous around $\bar{z}$. Furthermore, since ∇f is $L-$Lipschitz continuous, then ∇f^γ is $L-$Lipschitz continuous [TP20, Proposition 2.3], and thus the largest eigenvalue of $\nabla^2 f^\gamma(z)$ is bounded above by L_f. Consequently, for any $\gamma < \frac{1}{L_f}$, the smallest eigenvalue of $I - \gamma \nabla^2 f^\gamma(z)$ is positive, and thus $\nabla \mathrm{prox}_{\gamma f}(z)$ is positive definite. The result follows since $\bar{z}$ is arbitrary.

□

The next result states the local smoothness property of the iteration map S_γ. For that, we need to recall the Hessian matrices of the Lagrangian (24) defined in Proposition 1:

$$H_{zy} = \nabla^2_{zy} \mathcal{L}(\bar{z}, \bar{x}, \bar{\lambda}), \quad H_{yy} = \nabla^2_{yy} \mathcal{L}(\bar{z}, \bar{x}, \bar{\lambda}).$$

Theorem 2 (Smoothness of iteration maps and envelopes)**.** *Consider the optimization problem (1). Let $\bar{x}$ be a critical point of φ, and $\bar{z}$ such that $\bar{x} = prox_{\gamma f}(\bar{z})$. Under Assumptions 4 and 5, whenever $\gamma > 0$ satisfies (20), the maps R_γ and S_γ, defined in (14) and (15), respectively, are C^1-smooth locally around $\bar{z}$, and φ_γ^{DR} is a C^2-smooth map locally around $\bar{z}$.*

Furthermore, the gradient of the iteration map S_γ at $z = \bar{z}$ is given by

$$\nabla S_\gamma(\bar{z}) = I + \lambda\big(\nabla R_\gamma(\bar{z}) - \nabla prox_{\gamma f}(\bar{z})\big), \tag{27}$$

and the directional derivative of R_γ at $z = \bar{z}$ is given by:

$$\nabla R_\gamma(\bar{z})h = \operatorname*{arg\,min}_{u \in T_{\mathcal{G}}(\bar{x})} 2\langle H_{zy}u, h\rangle + \langle H_{yy}u, u\rangle. \tag{28}$$

Moreover, for all z near $\bar{z}$, the gradient of the DRE is

$$\nabla\varphi_\gamma^{DR}(z) = \gamma^{-1}\nabla prox_{\gamma f}(z)\Big(I - \gamma\nabla^2 f\big(prox_{\gamma f}(z)\big)\Big)\big(prox_{\gamma f}(z) - R_\gamma(z)\big). \tag{29}$$

Proof. Given $z \in \mathbb{R}^d$, define $\phi(z,\cdot)$ as the objective function of the minimization problem in (23). Our goal is to apply Proposition 1. Since $\gamma < \rho^{-1}$ and g is $\rho-$weakly convex, then $\phi(z,\cdot)$ is $(\gamma^{-1}-\rho)-$strongly convex. In view of the continuity of the proximal operators of f and g for $\gamma > 0$ satisfying (20), and Proposition 1, R_γ is C^1-smooth and $\varphi_\gamma^{\mathrm{DR}}$ is C^2-smooth on a neighborhood of $\bar{z}$. Furthermore, Lemma 1 implies the existence and continuity of $\nabla\mathrm{prox}_{\gamma f}$ on a neighborhood of $\bar{z}$. Thus (27) directly follows from (15), and thus ∇S_γ is also continuous around $\bar{z}$.

For the gradient of $\varphi_\gamma^{\mathrm{DR}}$, note that for points z close to $\bar{z}$, due to Proposition 4, it holds that

$$\begin{aligned}\varphi_\gamma^{\mathrm{DR}}(z) \;=\;& f\big(\mathrm{prox}_{\gamma f}(z)\big) + +\langle\nabla f\big(\mathrm{prox}_{\gamma f}(z)\big), R_\gamma(z) - \mathrm{prox}_{\gamma f}(z)\rangle\\ &+\hat{g}\big(R_\gamma(z)\big) + \frac{1}{2\gamma}\|R_\gamma(z) - \mathrm{prox}_{\gamma f}(z)\|^2.\end{aligned}$$

Next, using the chain rule and rearranging terms, we obtain

$$\begin{aligned}\nabla\varphi_\gamma^{\mathrm{DR}}(z) \;=\;& \nabla\mathrm{prox}_{\gamma f}(z)\nabla f(\mathrm{prox}_{\gamma f}(z))\\ &+\nabla\mathrm{prox}_{\gamma f}(z)\nabla^2 f\big(\mathrm{prox}_{\gamma f}(z)\big)\big(R_\gamma(z) - \mathrm{prox}_{\gamma f}(z)\big)\\ &+\big(\nabla R_\gamma(z) - \nabla\mathrm{prox}_{\gamma f}(z)\big)\nabla f\big(\mathrm{prox}_{\gamma f}(z)\big) + \nabla R_\gamma(z)\nabla\hat{g}\big(R_\gamma(z)\big)\\ &+\gamma^{-1}\big(\nabla R_\gamma(z) - \nabla\mathrm{prox}_{\gamma f}(z)\big)\big(R_\gamma(z) - \mathrm{prox}_{\gamma f}(z)\big)\\ =\;& \nabla R_\gamma(z)\Big(\nabla f(\mathrm{prox}_{\gamma f}(z)) + \nabla\hat{g}(R_\gamma(z)) + \gamma^{-1}\big(R_\gamma(z) - \mathrm{prox}_{\gamma f}(z)\big)\Big)\\ &\nabla\mathrm{prox}_{\gamma f}(z)\big(\nabla^2 f\big(\mathrm{prox}_{\gamma f}(z)\big) - \gamma^{-1}I\big)\big(R_\gamma(z) - \mathrm{prox}_{\gamma f}(z)\big).\end{aligned}$$

From the definition of R_γ and Proposition 4, for all z close to $\bar{z}$, the first-order optimality conditions of (17) take the following form

$$0 = \nabla f(\mathrm{prox}_{\gamma f}(z)) + \nabla\hat{g}(R_\gamma(z)) + \gamma^{-1}\big(R_\gamma(z) - \mathrm{prox}_{\gamma f}(z)\big).$$

Thus, substituting this identity in the above expression for $\nabla\varphi_\gamma^{\mathrm{DR}}(z)$, yields (29). □

The explicit form of the gradient of $\varphi_\gamma^{\mathrm{DR}}$ around $\bar{z}$ allows us to deduce the relationship between critical points of φ and $\varphi_\gamma^{\mathrm{DR}}$, thus extending Proposition 3.

Corollary 1 (Characterization of critical points II)**.** *Under Assumptions 4 and 5, if $\gamma > 0$ satisfies (20), $\bar{x}$ is a critical point of φ if and only if there exists $\bar{z}$ such that $\bar{x} = prox_{\gamma f}(\bar{z})$, and $\bar{z}$ is a critical point of φ_γ^{DR}.*

Proof. It follows from Theorem 2 and Proposition 3 that $0 \in \partial\varphi(\bar{x})$ if and only if $\bar{x} = R_\gamma(\bar{z})$. In turn, this is equivalent to $\operatorname{prox}_{\gamma f}(\bar{z}) = R_\gamma(\bar{z})$. In view of (29), it is also equivalent to $\nabla\varphi_\gamma^{\mathrm{DR}}(\bar{z}) = 0$, since $\nabla\operatorname{prox}_{\gamma f}(z)\big(I - \gamma\nabla^2 f(\operatorname{prox}_{\gamma f}(z))\big)$ is a non-singular matrix, in view of Lemma 1 and the fact that $\gamma < L_f^{-1}$. □

In this section, we have constructed the building blocks to analyze the avoidance of saddle points property of the DR method. In the next section, we specify in what sense these saddle points are avoided, and what type of saddle points are avoidable.

5.4. **Active saddle point avoidance of Douglas-Rachford.** Recall that in view of Theorem 1, the limit of the DR sequence $\{z^k\}$ is a critical point of φ. This limit can be guaranteed to be a local minimizer (almost surely) if φ possesses the strict saddle property, the main topic of this sections.

When a function h is C^2, avoidable critical points are the ones with negative curvature, that is, points such that $\nabla h(x) = 0$ and the Hessian $\nabla^2 h(x)$ has at least one negative eigenvalue. To generalize this idea for non-C^2 functions, we require a sharpness condition to guarantee that the local V-shape geometry leads the dynamics to a region of negative curvature.

We say that a critical point $\bar{x}$ of a weakly convex function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a strict saddle of h, if there exist an active C^2-smooth manifold $\mathcal{M}$ of h at $\bar{x}$, such that for some vector $u \in T_{\mathcal{M}}(\bar{x})$, the parabolic subderivative [RW09, Definition 13.59] satisfies $d^2 h_{\mathcal{M}}(\bar{x})(u) < 0$. A geometric interpretation of strict saddle points is that the function h restricted to $\mathcal{M}$ has negative curvature at such points. In the context of problem (21), recall that the parabolic subderivative of $\varphi_{\mathcal{G}} = \varphi + \delta_{\mathcal{G}}$ can be expressed in terms of the Hessian of the Lagrangian (22) as follows (see [DD22]):

$$d^2\varphi_{\mathcal{G}}(\bar{x})(u) = \langle \nabla^2_{xx}\mathcal{L}_0(\bar{x}, \bar{\lambda})u, u\rangle \tag{30}$$

for all $u \in T_{\mathcal{G}}(\bar{x})$.

Moreover, we say h satisfies the strict saddle property, if any critical point of φ is either a strict saddle or a local minimizer. Under the same properties that assure that Assumption 4 is fulfilled (cf. [DD22, Theorem 2.9] and [DIL16, Theorem 4.16]), the strict saddle property is also satisfied. For instance, in our setting, it accounts to $\varphi = f + g$ belonging to the family of lsc weakly convex semialgebraic functions. Observe that φ is already a $(L_f + \rho)-$weakly convex function.

As similarly proven in [DD22, Theorem 4.1] for the FB method, strict saddle points of φ are in correspondence with unstable fixed points of the DR method iteration map S_γ.

Theorem 3 (Unstable fixed points of the DR iteration operator)**.** *Consider the optimization problem (1), and $\gamma > 0$ satisfying (20). Let $\bar{x}$ be strict saddle point of φ. Under Assumptions 4 and 5, $\bar{z}$ is an unstable fixed point of S_γ, where $\bar{x} = \operatorname{prox}_{\gamma f}(\bar{z})$.*

Proof. First, $\bar{x}$ is in particular a critical point, then Proposition 1 and Theorem 2 imply (28) for

$$H_{zy} = \nabla\operatorname{prox}_{\gamma f}(\bar{z})\left(\nabla^2 f(\bar{x}) - \frac{1}{\gamma}I\right), \quad H_{yy} = \nabla^2\hat{g}(\bar{x}) + \frac{1}{\gamma}I + \sum_{i\geq 1}\lambda_i\nabla^2 G_i(\bar{x}).$$

Let W denote the orthogonal projection matrix onto $T_{\mathcal{G}}(\bar{x})$, and set $\overline{H}_{zy} = WH_{zy}W$ and $\overline{H}_{yy} = WH_{yy}W$. Note that since $R_\gamma(\bar{z}) = \bar{x}$ is a strong local minimizer of the problem in (23), then $\overline{H}_{yy}$ is (symmetric) positive definite. Due to Lemma 1, in particular, $\overline{H}_{yy}\nabla\mathrm{prox}_{\gamma f}(\bar{z}) = \nabla\mathrm{prox}_{\gamma f}(\bar{z})\overline{H}_{yy}$.

In order to prove that $\nabla S_\gamma(\bar{z})$ has a real eigenvalue strictly greater than one, it suffices to show there exists $\mu > 1$ such that

$$\big(\lambda\nabla\mathrm{prox}_{\gamma f}(z) + (\mu-1)I\big)\overline{H}_{yy} + \lambda\overline{H}_{zy} \text{ is singular.}$$

Indeed, the optimality condition of (28) implies

$$\nabla R_\gamma(\bar{z})h = -\overline{H}_{yy}^{-1}\overline{H}_{zy}h,$$

and thus (27) yields

$$\nabla S_\gamma(\bar{z})h = h - \lambda\big(\overline{H}_{yy}^{-1}\overline{H}_{zy} + \nabla\mathrm{prox}_{\gamma f}(\bar{z})\big)h.$$

In this manner, $\mu \in \mathbb{R}$ is a a real eigenvalue of $\nabla S(\bar{z})$ if and only if there exists $h \in \mathbb{R}^d \setminus \{0\}$ such that

$$\begin{aligned}
\nabla S_\gamma(\bar{z})h &= \mu h\\
h - \lambda\big(\overline{H}_{yy}^{-1}\overline{H}_{zy} + \nabla\mathrm{prox}_{\gamma f}(\bar{z})\big)h &= \mu h\\
\overline{H}_{yy}h - \lambda\big(\overline{H}_{zy}h + \overline{H}_{yy}\nabla\mathrm{prox}_{\gamma f}(\bar{z})h\big) &= \mu\overline{H}_{yy}h\\
\Big[\overline{H}_{yy}\big(\lambda\nabla\mathrm{prox}_{\gamma f}(\bar{z}) + (\mu-1)I\big) + \lambda\overline{H}_{zy}\Big]h &= 0\\
\Big[\big(\lambda\nabla\mathrm{prox}_{\gamma f}(\bar{z}) + (\mu-1)I\big)\overline{H}_{yy} + \lambda\overline{H}_{zy}\Big]h &= 0
\end{aligned}$$

We focus on proving that, for some $\mu > 1$, $\big(\lambda\nabla\mathrm{prox}_{\gamma f}(\bar{z}) + (\mu-1)I\big)\overline{H}_{yy} + \lambda\overline{H}_{zy}$ is singular. First, since $\overline{H}_{yy}$ and $\nabla\mathrm{prox}_{\gamma f}(\bar{z})$ are positive definite, then for sufficiently large $\mu > 1$, $\big(\lambda\nabla\mathrm{prox}_{\gamma f}(\bar{z}) + (\mu-1)I\big)\overline{H}_{yy} + \lambda\overline{H}_{zy}$ has positive eigenvalues. Secondly, for $\mu = 1$,

$$\begin{aligned}
\nabla\mathrm{prox}_{\gamma f}(\bar{z})\overline{H}_{yy} + \overline{H}_{zy} &= \nabla\mathrm{prox}_{\gamma f}(\bar{z})\left(\nabla^2\hat{g}(\bar{x}) + \tfrac{1}{\gamma}I + \textstyle\sum_{i\geq 1}\lambda_i\nabla^2 G_i(\bar{x})\right)\\
&\quad + \nabla\mathrm{prox}_{\gamma f}(\bar{z})\left(\nabla^2 f(\bar{x}) - \tfrac{1}{\gamma}I\right)\\
&= \nabla\mathrm{prox}_{\gamma f}(\bar{z})\left(\nabla^2 f(\bar{x}) + \nabla^2\hat{g}(\bar{x}) + \textstyle\sum_{i\geq 1}\lambda_i\nabla^2 G_i(\bar{x})\right)\\
&= \nabla\mathrm{prox}_{\gamma f}(\bar{z})\nabla^2_{xx}\mathcal{L}_0(\bar{x},\bar{\lambda}),
\end{aligned}$$

where in the last line we used (22). In view of Lemma 1, $\nabla\mathrm{prox}_{\gamma f}(\bar{z})\overline{H}_{yy} + \overline{H}_{zy}$ is similar to

$$\nabla\mathrm{prox}_{\gamma f}(\bar{z})^{1/2}\nabla^2_{xx}\mathcal{L}_0(\bar{x},\bar{\lambda})\nabla\mathrm{prox}_{\gamma f}(\bar{z})^{1/2}. \tag{31}$$

The assumption on $\bar{x}$ and (30) implies there exists $u \in T_{\mathcal{G}}(\bar{x})$, such that

$$\langle\nabla^2_{xx}\mathcal{L}_0(\bar{x},\bar{\lambda})u, u\rangle < 0.$$

Hence, the matrix in (31) has a negative eigenvalue, and consequently, so does $\nabla\mathrm{prox}_{\gamma f}(\bar{z})\overline{H}_{yy} + \overline{H}_{zy}$.

It follows from the continuity of the minimal eigenvalue function, the existence of a real $\mu > 1$ such that $\big(\lambda\nabla\mathrm{prox}_{\gamma f}(\bar{z}) + (\mu-1)I\big)\overline{H}_{yy} + \lambda\overline{H}_{zy}$ has a null eigenvalue, and thus, is singular. Hence, $\nabla S_\gamma(\bar{z})$ has real eigenvalue greater than one.

□

From the geometric nature of active strict saddle points, it is possible to prove that the set of initial points of the DR dynamic system $z^{k+1} = S_\gamma(z^k)$ that converge to an unstable fixed point, has measure zero. In order to globalize the result in (19), we need to add a small momentum term to the dynamics (cf. [DD22, Theorem 4.2]), so that the resulting iteration map is invertible, and thus be able to apply the Center Stable Manifold theorem. More specifically, the following result is a consequence of [DD22, Corollary 2.12]: given a lipeomorphism $T : \mathbb{R}^d \to \mathbb{R}^d$, if $\mathcal{U}_T$ denotes the set of unstable fixed points of T at which the Jacobian of T is invertible, then

$$\left\{ x \in \mathbb{R}^d : \lim_{k\to+\infty} T^k(x) \in \mathcal{U}_T \right\} \text{ has zero Lebesgue measure.} \tag{32}$$

Theorem 4 (Douglas-Rachford method: global escape of strict saddles)**.** *Consider the optimization problem (1). Under Assumptions 4 and 5, suppose φ has the strict saddle property. Choose $\gamma > 0$ satisfying (20), and a damping parameter $\alpha \in (0, 1)$ such that $\alpha\mu < 1$, where*

$$\mu = |1-\lambda| + \max\left\{\frac{\rho\gamma}{1-\rho\gamma}, 1\right\}\left(\frac{L_f}{L_f-\gamma} + \max\left\{\frac{L_f\gamma}{1-L_f\gamma}, 1\right\}\right) + \max\left\{\frac{L_f\gamma}{1-L_f\gamma}, 1\right\}.$$

Consider the damped Douglas-Rachford splitting method

$$z^{k+1} = T_\gamma(z^k) := (1-\alpha)z^k + \alpha S_\gamma(z^k).$$

For almost all initial points z^0, if the limit of $\{z^k\}$ exists and is denoted $\bar{z}$, then $\bar{x} = prox_{\gamma f}(\bar{z})$ is a local minimizer of φ.

Proof. First, note that (7) implies $\mathrm{prox}_{\gamma f} - R_\gamma = \gamma\nabla e_{\gamma g}\big(\mathrm{prox}_{\gamma f} - \gamma\nabla e_{\gamma f}\big) + \gamma\nabla e_{\gamma f}$. Therefore, $I - S_\gamma = (1-\lambda)I + \lambda(\mathrm{prox}_{\gamma f} - R_\gamma)$ is a Lipschitz map with constant μ. Given $z^0 \in \mathbb{R}^d$, suppose $\{z^k\}$ converges to some $\bar{z}$. Continuity of S_γ implies that $\bar{z}$ is a fixed point of S_γ, and $\bar{x} = \mathrm{prox}_{\gamma f}(\bar{z})$ is a critical point of φ, due to Proposition 3. Furthermore, Theorem 2 implies that $\nabla S_\gamma(\bar{z})$ and $\nabla T_\gamma(\bar{z})$ exist. From [DD22, Lemma 2.14], it follows

(i) $\bar{z}$ is a fixed point of T_γ,
(ii) if $\bar{z}$ is an unstable fixed point of S_γ, so it is of T_γ, and
(iii) T_γ is a lipeomorphism.

By way of contradiction, suppose $\bar{x}$ is a strict saddle point of φ. In view of Theorem 3 and (ii), then $\bar{z}$ is an unstable fixed point of T_γ. Due to (iii) and Rademacher's theorem, $\nabla T(\bar{z})$ is invertible almost surely, yielding a contradiction with (32). Therefore, $\bar{x}$ has to be a local minimizer of φ, since φ has the strict saddle property.

□

6. Concluding remarks

In this work, we show that the Douglas-Rachford splitting method enjoys the acclaimed property of saddle point avoidance for nonconvex nonsmooth problems, just as the gradient descent method and the proximal point algorithm. In principle, such property seems straightforward to prove, since the algorithm is based on simpler methods satisfying said property. However, the structure of the steps of one Douglas-Rachford iteration makes the analysis more intricate and less direct. The same question remains open for other methods that use first-order information

or of proximal-type. Moreover, although our arguments provide a satisfactory explanation to a behavior that has been observed numerically, there are still some drawbacks. For instance, in practice we never compute exact solutions to the subproblems, and thus convergence to local minimizers with inexact subproblem solutions still remains to be theoretically justified.

References

[AAS24] W. van Ackooij, F. Atenas, and C. Sagastizábal. "Weak convexity and approximate subdifferentials". In: *Optimization Online* (2024).

[ABS13] 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". In: *Mathematical Programming* 137.1-2 (2013), pp. 91–129.

[Ate+23] F. Atenas, C. Sagastizábal, P. J. Silva, and M. Solodov. "A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods". In: *SIAM Journal on Optimization* 33.1 (2023), pp. 89–115.

[Ate23] F. Atenas. "Convergence Rate of Nonconvex Douglas-Rachford splitting via merit functions, with applications to weakly convex constrained optimization". In: *arXiv preprint arXiv:2303.16394* (2023).

[BC11] H. Bauschke and P. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer New York, 2011.

[BDL07] J. Bolte, A. Daniilidis, and A. Lewis. "The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems". In: *SIAM Journal on Optimization* 17.4 (2007), pp. 1205–1223.

[Bol+07] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota. "Clarke subgradients of stratifiable functions". In: *SIAM Journal on Optimization* 18.2 (2007), pp. 556–572.

[Bol+17] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. "From error bounds to the complexity of first-order descent methods for convex functions". In: *Mathematical Programming* 165 (2017), pp. 471–507.

[Boy+11] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers". In: *Foundations and Trends® in Machine learning* 3.1 (2011), pp. 1–122.

[Cla90] F. H. Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

[DD22] D. Davis and D. Drusvyatskiy. "Proximal methods avoid active strict saddles of weakly convex functions". In: *Foundations of Computational Mathematics* 22.2 (2022), pp. 561–606.

[DIL16] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis. "Generic minimizing behavior in semialgebraic optimization". In: *SIAM Journal on Optimization* 26.1 (2016), pp. 513–534.

[Dru17] D. Drusvyatskiy. "The proximal point method revisited". In: *arXiv preprint arXiv:1712.06038* (2017).

[Eck89] J. Eckstein. "Splitting methods for monotone operators with applications to parallel optimization". PhD thesis. Massachusetts Institute of Technology, 1989.

[HL07] W. L. Hare and A. S. Lewis. "Identifying active manifolds". In: *Algorithmic Operations Research* 2.2 (2007), pp. 75–82.

[HL13] J. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2013.

[HL96] J. Hiriart-Urruty and C. Lemarechal. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1996.

[HLO] T. HOHEISEL, M. LABORDE, and A. OBERMAN. "On proximal point-type algorithms for weakly convex functions and their connection to the backward euler method". In: *Optimization Online* ().

[Kru03] A. Y. Kruger. "On Fréchet subdifferentials". In: *J. Math Sci.* 116.3 (2003), pp. 3325–3358.

[Lee+16] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. "Gradient descent only converges to minimizers". In: *Conference on learning theory*. PMLR. 2016, pp. 1246–1257.

[Lee+19] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. "First-order methods almost always avoid strict saddle points". In: *Mathematical programming* 176 (2019), pp. 311–337.

[Lew02] A. S. Lewis. "Active sets, nonsmoothness, and sensitivity". In: *SIAM Journal on Optimization* 13.3 (2002), pp. 702–725.

[LM79] P.-L. Lions and B. Mercier. "Splitting algorithms for the sum of two nonlinear operators". In: *SIAM Journal on Numerical Analysis* 16.6 (1979), pp. 964–979.

[LOS00] C. Lemaréchal, F. Oustry, and C. Sagastizábal. "The -Lagrangian of a convex function". In: *Transactions of the American mathematical Society* 352.2 (2000), pp. 711–729.

[LP16] G. Li and T. K. Pong. "Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems". In: *Mathematical programming* 159 (2016), pp. 371–401.

[LP18] G. Li and T. K. Pong. "Calculus of the exponent of Kurdyka–Łojasiewicz inequality and its applications to linear convergence of first-order methods". In: *Foundations of computational mathematics* 18.5 (2018), pp. 1199–1232.

[LT93] Z.-Q. Luo and P. Tseng. "Error bounds and convergence analysis of feasible descent methods: a general approach". In: *Annals of Operations Research* 46.1 (1993), pp. 157–178.

[LY19] Y. Liu and W. Yin. "An envelope for Davis–Yin splitting and strict saddle-point avoidance". In: *Journal of Optimization Theory and Applications* 181 (2019), pp. 567–587.

[Mar70] B. Martinet. "Regularisation d'inequations variationelles par approximations successives". In: *Revue Francaise d'informatique et de Recherche operationelle* 4 (1970), pp. 154–159.

[PB+14] N. Parikh, S. Boyd, et al. "Proximal algorithms". In: *Foundations and trends® in Optimization* 1.3 (2014), pp. 127–239.

[PB13] P. Patrinos and A. Bemporad. "Proximal Newton methods for convex composite optimization". In: *52nd IEEE Conference on Decision and Control*. IEEE. 2013, pp. 2358–2363.

[Pen+16] Z. Peng, T. Wu, Y. Xu, M. Yan, and W. Yin. "Coordinate friendly structures, algorithms and applications". In: *SIAM Journal on Control and Optimization* 1.1 (2016), pp. 57–119.

[PP16] I. Panageas and G. Piliouras. "Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions". In: *arXiv preprint arXiv:1605.00405* (2016).

[PSB14] P. Patrinos, L. Stella, and A. Bemporad. "Douglas-Rachford splitting: Complexity estimates and accelerated variants". In: *53rd IEEE Conference on Decision and Control*. IEEE. 2014, pp. 4234–4239.

[Roc76] R. T. Rockafellar. "Monotone operators and the proximal point algorithm". In: *SIAM journal on control and optimization* 14.5 (1976), pp. 877–898.

[RW09] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Vol. 317. Springer Science & Business Media, 2009.

[Shu13] M. Shub. *Global stability of dynamical systems*. Springer Science & Business Media, 2013.

[STP17] L. Stella, A. Themelis, and P. Patrinos. "Forward–backward quasi-Newton methods for nonsmooth optimization problems". In: *Computational Optimization and Applications* 67.3 (2017), pp. 443–487.

[TP20] A. Themelis and P. Patrinos. "Douglas–Rachford splitting and admm for nonconvex optimization: Tight convergence results". In: *SIAM Journal on Optimization* 30.1 (2020), pp. 149–181.

[TSP18] A. Themelis, L. Stella, and P. Patrinos. "Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms". In: *SIAM Journal on Optimization* 28.3 (2018), pp. 2274–2303.

[Wri20] S. Wright. *Some Perspectives on Nonconvex Optimization*. IPAM workshop on Intersections between Control, Learning and Optimization. 2020.

[Wri93] S. J. Wright. "Identifiable surfaces in constrained optimization". In: *SIAM Journal on Control and Optimization* 31.4 (1993), pp. 1063–1079.

Appendices

Proof of Proposition 2.

Suppose $\{z^k\}$ is generated by the FB method. Then, (13) holds, and $\inf \varphi_\gamma^{\mathrm{FB}} = \inf \varphi$, for all $\gamma \in \big(0, \min\{L_f^{-1}, \rho^{-1}\}\big)$. As a consequence, $\{\varphi_\gamma^{\mathrm{FB}}(z^k)\}$ is a nonincreasing sequence bounded from below, thus convergent to some $\bar{\varphi}$. The same estimate yields the following inequality:

$$\sum_{k=1}^{\infty} \|z^{k+1} - z^k\|^2 \leq \frac{2\gamma}{1-\gamma L_f}(\bar{\varphi} - \inf \varphi),$$

which implies $z^{k+1} - z^k \to 0$ as $k \to +\infty$. In view of $\{z^k\} \subseteq \{z \in \mathbb{R}^d : \varphi_\gamma^{\mathrm{FB}}(z) \leq \varphi_\gamma^{\mathrm{FB}}(z^0)\}$, then $\{z^k\}$ is bounded. Take any cluster point $\bar{z}$, and a subsequence $z^{k_j} \to \bar{z}$ as $j \to +\infty$. Then, by continuity of the gradient and of $\mathrm{prox}_{\gamma g}$, $z^{k_j+1} \to \mathrm{prox}_{\gamma g}\big(\bar{z} - \gamma \nabla f(\bar{z})\big)$. Since $z^{k+1} - z^k \to 0$, then $z^{k_j+1} \to \bar{z}$ as well, yielding $\bar{z} = \mathrm{prox}_{\gamma g}\big(\bar{z} - \gamma \nabla f(\bar{z})\big)$, that is, any cluster point of $\{z^k\}$ is a fixed point of the iteration operator $z \mapsto \mathrm{prox}_{\gamma g}\big(z - \gamma \nabla f(z)\big)$, and thus a critical point for problem (1). Indeed, the first-order optimality conditions of the problem defining $\mathrm{prox}_{\gamma g}$ imply

$$-\nabla f(\bar{z}) = \frac{1}{\gamma}(\bar{z} - \gamma \nabla f(\bar{z}) - \bar{z}) \in \partial g(\bar{z}),$$

and hence $0 \in \nabla f(\bar{z}) + \partial g(\bar{z})$.

Now suppose $\{z^k\}$ is generated using the DR method. We replicate the arguments above step-by-step. First, (18) and $\inf \varphi_\gamma^{\mathrm{DR}} = \inf \varphi$ hold for $\gamma \in (0, \min\{(2-\lambda)(2L_f)^{-1}, \rho^{-1}\})$. Then, $\{\varphi_\gamma^{\mathrm{DR}}(z^k)\}$ is a convergent sequence (to some $\bar{\varphi}$) as a nonincreasing sequence bounded from below. As for FB, (18) yields the following inequality:

$$\sum_{k=1}^{\infty} \|z^{k+1} - z^k\|^2 \leq \frac{(1+\gamma L_f)^2}{c}(\bar{\varphi} - \inf \varphi),$$

which implies $z^{k+1} - z^k \to 0$ as $k \to +\infty$, and also $x^k - y^k \to 0$. Furthermore, $\{z^k\}$ is bounded because $\varphi_\gamma^{\mathrm{DR}}$ is level-bounded [TP20, Theorem 3.4(iii)], and thus both $\{x^k\}$ and $\{y^k\}$ are bounded as well, due to Lipschitz continuity of the proximal operator. In particular, $\{x^k\}$ and $\{y^k\}$ have the same set of cluster points. Take any cluster point $\bar{z}$ of $\{z^k\}$, and a subsequence $z^{k_j} \to \bar{z}$ as $j \to +\infty$, and thus $x^{k_j} \to \mathrm{prox}_{\gamma f}(\bar{z}) = \bar{x}$, and $y^{k_j} \to \mathrm{prox}_{\gamma g}(2\bar{x} - \bar{z})$. Note that the first-order optimality condition of the problem defining $\mathrm{prox}_{\gamma f}$ yields $\frac{1}{\gamma}(\bar{z} - \bar{x}) = \nabla f(\bar{x})$, and thus $\bar{x} = \mathrm{prox}_{\gamma g}\big(\bar{x} - \gamma \nabla f(\bar{x})\big)$. In other words, $\bar{x}$ is a fixed point of the FB iteration operator, and thus is a critical point of (1).