Projected gradient descent accumulates at Bouligand stationary points

Guillaume Olikier, Irène Waldspurger

PROJECTED GRADIENT DESCENT ACCUMULATES AT BOULIGAND STATIONARY POINTS*

GUILLAUME OLIKIER† AND IRÈNE WALDSPURGER‡

Abstract. This paper concerns the projected gradient descent (PGD) algorithm for the problem of minimizing a continuously differentiable function on a nonempty closed subset of a Euclidean vector space. Without further assumptions, this problem is intractable and devoted algorithms are only expected to find a stationary point. PGD is known to generate a sequence whose accumulation points are Mordukhovich stationary. In this paper, these accumulation points are proven to be Bouligand stationary, and even proximally stationary if the gradient is locally Lipschitz continuous. These are the strongest stationarity properties that can be expected for the considered problem.

Key words. projected gradient descent, stationarity, tangent and normal cones, Clarke regularity

MSC codes. 65K10, 49J53, 90C26, 90C30, 90C46

1. Introduction.

Given a Euclidean vector space $\mathcal{E}$, a nonempty closed subset C of $\mathcal{E}$, and a function $f : \mathcal{E} \to \mathbb{R}$ that is differentiable on C, this paper considers the problem

$$\min_{x \in C} f(x) \tag{1.1}$$

of minimizing f on C. In general, without further assumptions on C or f, problem (1.1) is intractable and devoted algorithms are only expected to find a stationary point of this problem. A point $x \in C$ is said to be stationary for (1.1) if $-\nabla f(x)$ is normal to C at x. Several definitions of normality exist. Each one yields a definition of stationarity provided that, possibly under mild regularity assumptions on f, every local minimizer of $f|_C$ is stationary for (1.1). In particular, each of the three notions of normality in [42, Definition 6.3 and Example 6.16], namely normality in the general sense, in the regular sense, and in the proximal sense, yields an important definition of stationarity. The sets of general, regular, and proximal normals to C at $x \in C$ are respectively denoted by $N_C(x)$, $\widehat{N}_C(x)$, and $\widehat{\widehat{N}}_C(x)$. These sets are reviewed in Section 2.2. Importantly, they are nested as follows: for every $x \in C$,

$$\widehat{\widehat{N}}_C(x) \subseteq \widehat{N}_C(x) \subseteq N_C(x), \tag{1.2}$$

and C is said to be *Clarke regular* at x if the second inclusion is an equality. The definitions of stationarity based on these sets are given in Definition 1.1, and the terminology is discussed in Section 3.

DEFINITION 1.1. *For problem* (1.1)*, a point $x \in C$ is said to be:*

- Mordukhovich stationary (M-stationary) *if* $-\nabla f(x) \in N_C(x)$*;*
- Bouligand stationary (B-stationary) *if* $-\nabla f(x) \in \widehat{N}_C(x)$*;*

*This work was supported by the ERC grant #786854 G-Statistics from the European Research Council under the European Union's Horizon 2020 research and innovation program and by the French government through the 3IA Côte d'Azur Investments ANR-19-P3IA-0002 and the PRAIRIE 3IA Institute ANR-19-P3IA-0001, managed by the National Research Agency.

†Université Côte d'Azur and Inria, Epione Project Team, 2004 route des Lucioles - BP 93, 06902 Sophia Antipolis Cedex, France (guillaume.olikier@inria.fr).

‡CNRS, Université Paris Dauphine, équipe-projet Mokaplan (Inria), place du Maréchal de Lattre de Tassigny, 75016 Paris, France (waldspurger@ceremade.dauphine.fr).

- proximally stationary (P-stationary) *if* $-\nabla f(x) \in \widehat{\widehat{N}}_C(x)$.

There are many practical examples of a set C for which at least one of the inclusions in (1.2) is strict, especially the second one. This is notably shown by the four examples studied in [35] and Section 7, where the second inclusion is strict at infinitely many points. The three notions of stationarity are therefore not equivalent. Actually, as explained next, B-stationarity and P-stationarity are the strongest necessary conditions for local optimality under different sets of assumptions on f, while M-stationarity is a weaker condition.

As pointed out in [10, §5], for problem (1.1) without additional assumptions on C or f, B-stationarity is the strongest necessary condition for local optimality. The same is true if f is assumed to be continuously differentiable on $\mathcal{E}$. Indeed, by [42, Theorem 6.11], for all $x \in C$,

$$\widehat{N}_C(x) = \left\{ -\nabla h(x) \;\middle|\; \begin{array}{l} h : \mathcal{E} \to \mathbb{R} \text{ is differentiable at } x, \\ x \text{ is a local minimizer of } h|_C \end{array} \right\}, \tag{1.3}$$

$$= \left\{ -\nabla h(x) \;\middle|\; \begin{array}{l} h : \mathcal{E} \to \mathbb{R} \text{ is continuously differentiable}, \\ x \text{ is a local minimizer of } h|_C \end{array} \right\}. \tag{1.4}$$

The inclusion $\supseteq$ in (1.3) shows that every local minimizer of $f|_C$ is B-stationary for (1.1). Thus, $\widehat{N}_C(x)$ is sufficiently large to yield a necessary condition for local optimality. The inclusion $\subseteq$ in (1.3) shows that replacing $\widehat{N}_C(x)$ with one of its proper subsets would yield a condition that is not necessary for local optimality. The equality (1.4) shows that these observations also hold if f is continuously differentiable on $\mathcal{E}$.

P-stationarity is the strongest necessary condition for local optimality if f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous. Indeed, by Theorem 2.5, for all $x \in C$,

$$\widehat{\widehat{N}}_C(x) = \left\{ -\nabla h(x) \;\middle|\; \begin{array}{l} h : \mathcal{E} \to \mathbb{R} \text{ is differentiable}, \\ \nabla h \text{ is locally Lipschitz continuous}, \\ x \text{ is a local minimizer of } h|_C \end{array} \right\}. \tag{1.5}$$

The inclusion $\supseteq$ in (1.5) shows that, if f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous, then every local minimizer of $f|_C$ is P-stationary for (1.1). The inclusion $\subseteq$ in (1.5) shows that replacing $\widehat{\widehat{N}}_C(x)$ with one of its proper subsets would yield a condition that is not necessary for local optimality.

In comparison, M-stationarity is a weaker notion of stationarity which is considered as unsatisfactory in [19, §4] and [24, §1]. This is especially true when problem (1.1) is the main problem to be solved, and not merely a subproblem to be solved at every iteration of an algorithm like the augmented Lagrangian method proposed in [21, Algorithm 4.1]. Furthermore, as explained in [24], distinguishing convergence to a B-stationary point from convergence to an M-stationary point is difficult (a phenomenon formalized by the notion of *apocalypse* in [24]).

Projected gradient descent, or PGD for short, is a basic algorithm aiming at solving problem (1.1); see [21, Algorithm 3.1], [22, Remark 2.1 and Algorithm 4.1], or Algorithm 4.2 for a definition. To the best of our knowledge, the first article considering PGD on a possibly nonconvex closed set was [5]. Given $x \in C$ as input, the iteration map of PGD (Algorithm 4.1), called the PGD map, performs a projected line search along the direction of $-\nabla f(x)$, i.e., computes a projection y of $x - \alpha \nabla f(x)$

onto C for decreasing values of $\alpha \in (0, \infty)$ until y satisfies an Armijo condition. In the simplest version of PGD, called *monotone*, the Armijo condition ensures that the value of f at the next iterate is smaller by a specified amount than the value at the current iterate. Following the general setting proposed in [14], the value at the current iterate can be replaced with the maximum value of f over a prefixed number of the previous iterates. This version of PGD is called *nonmonotone*. By [22, Theorem 3.1], monotone PGD accumulates at M-stationary points of (1.1) if f is continuously differentiable on $\mathcal{E}$ and bounded from below on C. By [22, Theorem 4.1], the same holds for nonmonotone PGD if f is further uniformly continuous on the sublevel set

$$\{x \in C \mid f(x) \leq f(x_0)\}, \tag{1.6}$$

where $x_0 \in C$ is the initial iterate given to the algorithm. However, as pointed out in [24, §1], it is an open question whether the accumulation points of PGD can fail to be B-stationary for (1.1).

This paper answers negatively the question by proving Theorem 1.2.

THEOREM 1.2. *Let $(x_i)_{i\in\mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2) when applied to problem* (1.1).

- *If ∇f is continuous on C, then all accumulation points of $(x_i)_{i\in\mathbb{N}}$ are B-stationary for* (1.1).
- *If f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous, then all accumulation points of $(x_i)_{i\in\mathbb{N}}$ are P-stationary for* (1.1).

Under the assumption that ∇f is globally Lipschitz continuous, for some sets C, it has already been proven that every local minimizer of $f|_C$ is P-stationary for (1.1) and that PGD with a constant step size smaller than the inverse of the Lipschitz constant accumulates at P-stationary points of (1.1). The case where C satisfies a regularity condition called *proximal smoothness*—which none of the four examples studied in Section 7 satisfies—is considered in [2, Proposition 1 and Theorem 1]. The case where C is the set $\mathbb{R}^n_{\leq s}$ of vectors of $\mathbb{R}^n$ having at most s nonzero components for some positive integer $s < n$ is considered in [4, Theorems 2.2 and 3.1].

This paper is organized as follows. The necessary background in variational analysis is introduced in Section 2. The literature on stationarity is partially surveyed in Section 3. The PGD algorithm is reviewed in Section 4. It is analyzed under the assumption that ∇f is continuous on C in Section 5 and under the assumption that f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous in Section 6. Four examples of a set C for which the first inclusion in (1.2) is an equality for all $x \in C$ and the second is strict for infinitely many $x \in C$ are given in Section 7. Concluding remarks are gathered in Section 8.

When we were finishing writing this paper, we became aware of the independent work [39], which proves a theorem very close to the second item in Theorem 1.2. Our result applies to a more general version of PGD than [39], since we allow for (monotone or nonmonotone) backtracking line search, while [39] considers PGD with a constant step size only (which has the consequence that ∇f must be globally Lipschitz continuous, while it is only locally Lipschitz continuous in our work). On the other hand, [39] deduces the result from a more general theorem about the proximal gradient algorithm, which we do not consider in this work. It also states an explicit quadratic lower bound on $f - f(\bar{x})$ at any accumulation point $\bar{x}$.

2. Elements of variational analysis. This section, mostly based on [42], reviews background material in variational analysis that is used in the rest of the paper. Section 2.1 concerns the projection mapping onto C and its main properties.

Section 2.2 reviews the three notions of normality on which the three notions of stationarity provided in Definition 1.1 are based.

Throughout the paper, for every $x \in \mathcal{E}$ and $\rho \in (0,\infty)$, $B(x,\rho) \coloneqq \{y \in \mathcal{E} \mid \|x-y\| < \rho\}$ and $B[x,\rho] \coloneqq \{y \in \mathcal{E} \mid \|x-y\| \leq \rho\}$ are respectively the open and closed balls of center x and radius ρ in $\mathcal{E}$. Following [42, §3B], a nonempty subset K of $\mathcal{E}$ is called a *cone* if $x \in K$ implies $\alpha x \in K$ for all $\alpha \in [0,\infty)$. By [42, 6(14)], for every cone $K \subseteq \mathcal{E}$, the set

$$K^* \coloneqq \{w \in \mathcal{E} \mid \langle v, w\rangle \leq 0 \ \forall v \in K\}$$

is a closed convex cone called the *polar* of K.

2.1. Projection mapping. Given $x \in \mathcal{E}$, the distance from x to C is $d(x,C) \coloneqq \min_{y\in C} \|x-y\|$ and the projection of x onto C is $P_C(x) \coloneqq \operatorname{argmin}_{y\in C} \|x-y\|$. By [42, Example 1.20], the function $\mathcal{E} \to \mathbb{R} : x \mapsto d(x,C)$ is continuous and, for every $x \in \mathcal{E}$, the set $P_C(x)$ is nonempty and compact. Proposition 2.1 is invoked frequently in the rest of the paper.

PROPOSITION 2.1. *For all $x \in C$, $v \in \mathcal{E}$, and $y \in P_C(x-v)$,*

$$\|y-x\| \leq 2\|v\|, \tag{2.1}$$

$$2\langle v, y-x\rangle \leq -\|y-x\|^2, \tag{2.2}$$

and the inequalities are strict if $x \notin P_C(x-v)$.

Proof. By definition of the projection, $\|y-(x-v)\| \leq \|x-(x-v)\| = \|v\|$ and the inequality is strict if $x \notin P_C(x-v)$. Thus, on the one hand,

$$\|y-x\| = \|y-(x-v)-v\| \leq \|y-(x-v)\| + \|-v\| \leq \|v\| + \|v\| = 2\|v\|,$$

and, on the other hand, $\|y-(x-v)\|^2 \leq \|v\|^2$, which is equivalent to (2.2). ∎

2.2. Normality and stationarity. Based on [42, Chapter 6], this section reviews the three notions of normality on which the three notions of stationarity given in Definition 1.1 are based.

Following [42, Definition 6.1], a vector $v \in \mathcal{E}$ is said to be *tangent* to C at $x \in C$ if there exist sequences $(x_i)_{i\in\mathbb{N}}$ in C converging to x and $(t_i)_{i\in\mathbb{N}}$ in $(0,\infty)$ such that the sequence $(\frac{x_i-x}{t_i})_{i\in\mathbb{N}}$ converges to v. The set of all tangent vectors to C at $x \in C$ is a closed cone [42, Proposition 6.2] called the *tangent cone* to C at x and denoted by $T_C(x)$. Following [42, Definition 6.3 and Proposition 6.5], for every $x \in C$, the polar

$$\widehat{N}_C(x) \coloneqq T_C(x)^*$$

is called the *regular normal cone* to C at x. Following [42, Definition 6.3], a vector $v \in \mathcal{E}$ is said to be *normal* to C at $x \in C$ if there exist sequences $(x_i)_{i\in\mathbb{N}}$ in C converging to x and $(v_i)_{i\in\mathbb{N}}$ converging to v such that, for all $i \in \mathbb{N}$, $v_i \in \widehat{N}_C(x_i)$. The set of all normal vectors to C at $x \in C$ is a closed cone [42, Proposition 6.5] called the *normal cone* to C at x and denoted by $N_C(x)$. Following [42, Example 6.16], a vector $v \in \mathcal{E}$ is called a *proximal normal* to C at $x \in C$ if there exists $\overline{\alpha} \in (0,\infty)$ such that $x \in P_C(x+\overline{\alpha}v)$, i.e., $\overline{\alpha}\|v\| = d(x+\overline{\alpha}v, C)$, which implies that, for all $\alpha \in [0,\overline{\alpha})$, $P_C(x+\alpha v) = \{x\}$. The set of all proximal normals to C at $x \in C$ is a convex cone called the *proximal normal cone* to C at x and denoted by $\widehat{\widehat{N}}_C(x)$. As stated in (1.2),

for all $x \in C$,

$$\widehat{\widehat{N}}_C(x) \subseteq \widehat{N}_C(x) \subseteq N_C(x).$$

Following [42, Definition 6.4], C is said to be *Clarke regular* at $x \in C$ if $\widehat{N}_C(x) = N_C(x)$. Thus, M-stationarity is equivalent to B-stationarity at a point $x \in C$ if and only if C is Clarke regular at x, which is not the case in many practical situations, as shown by the four examples given in [35]. For those examples, however, regular normals are proximal normals, as established in Section 7. Note that there exist a set C and a point $x \in C$ such that both inclusions in (1.2) are strict, as illustrated by Example 2.2.

EXAMPLE 2.2. *Let* $\mathcal{E} := \mathbb{R}^2$ *and* $C := \left\{\left(x, \max\left\{0, x^{3/5}\right\}\right) \mid x \in \mathbb{R}\right\}$ *(inspired by [42, Figure 6–12(a)]). Then,*

$$\widehat{\widehat{N}}_C(0,0) \subsetneq \widehat{N}_C(0,0) \subsetneq N_C(0,0).$$

As pointed out in Section 1, the regular and proximal normal cones enjoy gradient characterizations which imply that B- and P-stationarity are the strongest necessary conditions for local optimality under different sets of assumptions on f. Those given in (1.3)–(1.4) come from [42, Theorem 6.11]. That given in (1.5) comes from Theorem 2.5, established at the end of this section.

As shown by (1.4), for problem (1.1), B-stationarity is the strongest necessary condition for local optimality if f is only assumed to be continuously differentiable on $\mathcal{E}$. In particular, under this assumption, P-stationarity is not necessary for local optimality, as illustrated by Example 2.3.

EXAMPLE 2.3. *Let* $\mathcal{E} := \mathbb{R}^2$, $C := \left\{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 \geq \max\left\{0, x_1^{3/5}\right\}\right\}$ *[42, Figure 6–12(a)], and* $f : \mathbb{R}^2 \to \mathbb{R} : (x_1, x_2) \mapsto \frac{1}{2}(x_1-1)^2+|x_2|^{3/2}$. *Then,* f *is continuously differentiable and, for all* $(x_1, x_2) \in \mathbb{R}^2$, $\nabla f(x_1, x_2) = (x_1-1, \frac{3}{2}\mathrm{sgn}(x_2)|x_2|^{1/2})$. *Thus,* $-\nabla f(0,0) = (1,0) \in \widehat{N}_C(0,0) \setminus \widehat{\widehat{N}}_C(0,0)$, *yet* $\operatorname{argmin}_C f = \{(0,0)\}$.

Proposition 2.4 states that P-stationarity is necessary for local optimality if f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous. The latter means that, for every ball $\mathcal{B} \subsetneq \mathcal{E}$,

$$\operatorname*{Lip}_{\mathcal{B}}(\nabla f) := \sup_{\substack{x,y \in \mathcal{B} \\ x \neq y}} \frac{\|\nabla f(x) - \nabla f(y)\|}{\|x - y\|} < \infty,$$

which implies, by [34, Lemma 1.2.3], that, for all $x, y \in \mathcal{B}$,

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{\operatorname{Lip}_{\mathcal{B}}(\nabla f)}{2}\|y - x\|^2. \tag{2.3}$$

PROPOSITION 2.4. *Assume that* f *is differentiable on* $\mathcal{E}$ *and* ∇f *is locally Lipschitz continuous. If* $x \in C$ *is a local minimizer of* $f|_C$, *then* $-\nabla f(x) \in \widehat{\widehat{N}}_C(x)$.

Proof. By contrapositive. Assume that $-\nabla f(x) \notin \widehat{\widehat{N}}_C(x)$ for some $x \in C$. Let $\rho \in (0, \infty)$. Then, for all $\alpha \in (0, \frac{\rho}{2\|\nabla f(x)\|}]$,

$$x \notin P_C(x - \alpha\nabla f(x)) \subseteq B(x, 2\alpha\|\nabla f(x)\|) \subseteq B(x, \rho),$$

where the first inclusion holds by (2.1). Thus, by (2.3) and (2.2), for all $\alpha \in (0, \min\{\frac{\rho}{2\|\nabla f(x)\|}, \frac{1}{\mathrm{Lip}_{B(x,\rho)}(\nabla f)}\}]$ and $y \in P_C(x - \alpha\nabla f(x))$,

$$\begin{aligned} f(y) - f(x) &\leq \langle \nabla f(x), y - x\rangle + \frac{\mathrm{Lip}_{B(x,\rho)}(\nabla f)}{2}\|y - x\|^2 \\ &< \left(-\frac{1}{2\alpha} + \frac{\mathrm{Lip}_{B(x,\rho)}(\nabla f)}{2}\right)\|y - x\|^2 \\ &\leq 0. \end{aligned}$$

Hence, x is not a local minimizer of $f|_C$. ∎

Theorem 2.5 states that (1.5) is valid, which shows that P-stationarity is the strongest necessary condition for local optimality if f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous.

THEOREM 2.5 (gradient characterization of proximal normals). *For every $x \in C$, (1.5) holds.*

Proof. Let $x \in C$. The inclusion $\supseteq$ holds by Proposition 2.4. For the inclusion $\subseteq$, let us fix $v \in \widehat{\widehat{N}}_C(x)$. From the definition of the proximal normal cone, there exists $\overline{\alpha} \in (0, \infty)$ such that $x \in P_C(x + \overline{\alpha}v)$. This is equivalent to the fact that x is a minimizer of $h|_C$, where $h : \mathcal{E} \to \mathbb{R}$ is defined by

$$h(y) \coloneqq \frac{1}{2\overline{\alpha}}\|y - (x + \overline{\alpha}v)\|^2 \quad \forall y \in \mathcal{E}.$$

The map h is differentiable, its gradient is locally Lipschitz continuous (actually, globally Lipschitz continuous, since it is an affine map), and

$$-\nabla h(x) = v.$$

Since x is a global minimizer of $h|_C$, it is also a local minimizer of $h|_C$. This shows that

$$v \in \left\{ -\nabla h(x) \;\middle|\; \begin{array}{l} h : \mathcal{E} \to \mathbb{R} \text{ is differentiable,} \\ \nabla h \text{ is locally Lipschitz continuous,} \\ x \text{ is a local minimizer of } h|_C \end{array} \right\},$$

which implies the inclusion $\subseteq$ in (1.5). ∎

3. Stationarity in the literature. This section surveys the names given to the stationarity notions provided in Definition 1.1. The name "M-stationarity" is unanimous in the literature. It is legitimate because this stationarity notion is based on the normal cone introduced by Mordukhovich and often called the Mordukhovich normal cone.

In contrast, B-stationarity is known under other names in the literature. On the one hand, because the regular normal cone is also called the Fréchet normal cone, especially in infinite-dimensional spaces [42, 31, 32], B-stationarity is called Fréchet stationarity, or F-stationarity for short, in [26, Definition 4.1(ii)], [27, Definition 5.1(i)], and [28, Definition 3.2(ii)]. On the other hand, B-stationarity is simply called stationarity in [44], [17], [24, Definition 2.3], [23, Definition 3.2(c)], and [13, Definition 1]. This is legitimate since this notion of stationarity is the natural one, as explained in Section 1. Furthermore, in the literature about mathematical programs

with equilibrium constraints, the name "B-stationarity" sometimes indicates a notion of stationarity that is not B-stationarity. This is detailed in Section 3.1 which provides a brief history of B-stationarity.

Finally, P-stationarity seems to be new in the literature, although it is closely related to the so-called α-stationarity, as explained in Section 3.2. We propose the name "P-stationarity" because this stationarity notion is based on the proximal normal cone. In the context of optimization problems with complementarity constraints, [8] defines a stationarity notion called strong stationarity, or S-stationarity for short, which involves a proximal normal cone. However, the definition is specific to problems with complementarity constraints, and does not seem to have a clear relation with our "P-stationarity".

3.1. A brief history of Bouligand stationarity. Peano already knew that B-stationarity is a necessary condition for optimality. The statement is implicit in his 1887 book *Applicazioni geometriche del calcolo infinitesimale* and explicit in his 1908 book *Formulario Mathematico* where the formulation is based on the tangent cone and the derivative defined in the same book; see the historical investigation in [9, 10].

B-stationarity appears as a necessary condition for optimality in [46, Theorem 2.1] and [15, Theorem 1], without any reference to Peano's work. The latter theorem uses the polar of the closure of the convex hull of the tangent cone which equals the polar of the tangent cone by [42, Corollary 6.21]. Neither "stationary" nor "critical" appears in [46] or [15].

The Bouligand derivative is introduced in [41]. It is a special case of the contingent derivative introduced by Aubin based on the tangent cone. The name "Bouligand derivative" was chosen because the tangent cone is generally attributed to Bouligand; see, e.g., [42, 31, 32] for recent references.

B-stationarity is called a "stationarity condition" and said to be "well known" in [29, §4.1] where [15] is cited. The linearized cone is also introduced which always contains the tangent cone; the stationarity condition [29, (28)] associated with the linearized cone is called L-stationarity in this section. In general, these cones are equal only if a constraint qualification (CQ) is satisfied.

The name "Bouligand stationarity", or "B-stationarity" for short, is introduced in [43], without any comment on its origin. In [43, §2.1], "B-stationarity" is defined for a mathematical program with complementarity constraints. However, the definition looks more like L-stationarity than B-stationarity and yields a necessary condition for local optimality only if a CQ is satisfied, while B-stationarity is always necessary for local optimality. This suggests that this "B-stationarity", which is specific to problems with complementarity constraints, is not B-stationarity. In [43, §2.3], "B-stationarity" is defined for the minimization of an exact penalty function. The definition relies on the Bouligand derivative but not on the tangent cone since the problem is unconstrained. Thus, in this case, the name "B-stationarity" is arguably due to the Bouligand derivative.

B-stationarity appears, under this name, in [37]. Moreover, it is indicated in [37, §1] that "the condition" introduced in [29] is called "B-stationarity" in [43]; this condition is arguably L-stationarity since B-stationarity is said to be well known in [29] and is thus not "introduced" in [29]. Hence, on the one hand, [37] confirms that what is called "B-stationarity" in [43] is actually L-stationarity but, on the other hand, "B-stationarity" means both B-stationarity and L-stationarity in [37].

B-stationarity appears, under this name, in [12, Definition 2.4], in the proof of [11, Theorem 3.9], in [49, Definition 2.2], in [36, §2], in [38, (18)], in [19, §4], and

in [6, Definition 6.1.1]. Moreover, [49] confirms that, in [43], "B-stationarity" means L-stationarity and not B-stationarity. Remarkably, among all stationarity notions appearing in [49], B-stationarity is the only one defined based on a tangent or normal cone; formulating the other notions, such as those of Clarke and Mordukhovich, based of the corresponding normal cones would be an interesting contribution to this investigation.

However, L-stationarity is called "B-stationarity" in [16, Definition 2.2] and [48, Definition 3.2] which both cite [43].

In conclusion, although the name "B-stationarity" was introduced in [43], it means B-stationarity everywhere in the literature except in [43] and some papers that cite it.

3.2. Proximal stationarity and α-stationarity. Proximal stationarity is related to α-stationarity which was introduced in [4, Definition 2.3] for $C = \mathbb{R}^n_{\leq s}$ and in [26, Definition 4.1(i)], [17], [27, Definition 5.1(ii)], [25, (4.2)], and [28, Definition 3.2(i)] for several low-rank sets. By definition of the proximal normal cone, a point $x \in C$ is P-stationary for (1.1) if and only if there exists $\alpha \in (0, \infty)$ such that $x \in P_C(x - \alpha\nabla f(x))$. In contrast, given $\alpha \in (0, \infty)$, a point $x \in C$ is said to be α-stationary for (1.1) if $x \in P_C(x - \alpha\nabla f(x))$. Thus, while α-stationarity prescribes the number $\alpha \in (0, \infty)$, P-stationarity merely requires the existence of such a number. Furthermore, α-stationarity should not be confused with the approximate stationarity from [24, Definition 2.6].

4. The PGD algorithm. This section reviews the PGD algorithm, as defined in [21, Algorithm 3.1], based on its iteration map, called the PGD map and defined as Algorithm 4.1. The nonmonotonic behavior of PGD is described in Proposition 4.1.

Algorithm 4.1 PGD map (iteration map of [21, Algorithm 3.1])

Require: $(\mathcal{E}, C, f, \underline{\alpha}, \overline{\alpha}, \beta, c)$ where $\mathcal{E}$ is a Euclidean vector space, C is a nonempty closed subset of $\mathcal{E}$, $f : \mathcal{E} \to \mathbb{R}$ is differentiable on C, $0 < \underline{\alpha} \leq \overline{\alpha} < \infty$, and $\beta, c \in (0, 1)$.
Input: (x, μ) with $x \in C$ and $\mu \in [f(x), \infty)$.
Output: $y \in \mathrm{PGD}(x, \mu; \mathcal{E}, C, f, \underline{\alpha}, \overline{\alpha}, \beta, c)$.
1: Choose $\alpha \in [\underline{\alpha}, \overline{\alpha}]$ and $y \in P_C(x - \alpha\nabla f(x))$;
2: **while** $f(y) > \mu + c\,\langle \nabla f(x), y - x\rangle$ **do**
3: $\quad \alpha \leftarrow \alpha\beta$;
4: $\quad$ Choose $y \in P_C(x - \alpha\nabla f(x))$;
5: **end while**
6: Return y.

Two remarks should be made about Algorithm 4.1. First, the Armijo condition

$$f(y) \leq \mu + c\,\langle \nabla f(x), y - x\rangle$$

ensures that the decrease $\mu - f(y)$ is at least a fraction c of the opposite of the directional derivative of f at x with respect to the update vector $y - x$. By (2.2), this condition implies that

$$f(y) \leq \mu - \frac{c}{2\alpha}\|y - x\|^2, \tag{4.1}$$

which is the condition used in [22, Algorithms 3.1 and 4.1]. Importantly, all results from [22] hold for both conditions, as is clear from the proofs.

Second, by Proposition 5.3, if ∇f is continuous on C and x is not B-stationary for (1.1), then the while loop is guaranteed to terminate, thereby producing a point y such that $f(y) < \mu$; $y \neq x$ holds because x is not B-stationary and hence not P-stationary. If f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous, then the while loop is guaranteed to terminate, by Corollary 6.2.

The PGD algorithm is defined as Algorithm 4.2. It is said to be monotone or nonmonotone depending on whether $l = 0$ or $l > 0$.

Algorithm 4.2 PGD [21, Algorithm 3.1]

Require: $(\mathcal{E}, C, f, \underline{\alpha}, \overline{\alpha}, \beta, c, l)$ where $\mathcal{E}$ is a Euclidean vector space, C is a nonempty closed subset of $\mathcal{E}$, $f : \mathcal{E} \to \mathbb{R}$ is differentiable on C, $0 < \underline{\alpha} \leq \overline{\alpha} < \infty$, $\beta, c \in (0, 1)$, and $l \in \mathbb{N}$.
Input: $x_0 \in C$.
Output: a sequence in C.
1: $i \leftarrow 0$;
2: **while** $-\nabla f(x_i) \notin \widehat{N}_C(x_i)$ **do**
3: $\quad \mu_i \leftarrow \max_{j \in \{\max\{0, i-l\}, \ldots, i\}} f(x_j)$;
4: $\quad$ Choose $x_{i+1} \in \mathrm{PGD}(x_i, \mu_i; \mathcal{E}, C, f, \underline{\alpha}, \overline{\alpha}, \beta, c)$;
5: $\quad i \leftarrow i + 1$;
6: **end while**

If f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous, then $\widehat{N}_C(x_i)$ should be replaced with $\widehat{\widehat{N}}_C(x_i)$ in line 2. If PGD generates a finite sequence, then the last element of this sequence is B-stationary for (1.1), and even P-stationary for (1.1) if f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous. The rest of the paper concerns the case where it generates an infinite sequence. In that case, the stationarity of the accumulation points of the generated sequence, if any, is studied in Sections 5 and 6. Following [40, Remark 14], which states that it is usually better to determine whether an algorithm generates a sequence having at least one accumulation point by examining the algorithm in the light of the specific problem to which one wishes to apply it, no condition ensuring the existence of a convergent subsequence is made. As a reminder, a sequence $(x_i)_{i \in \mathbb{N}}$ has at least one accumulation point if and only if $\liminf_{i \to \infty} \|x_i\| < \infty$.

Monotone PGD generates a sequence along which f is strictly decreasing. Nonmonotone PGD generates a sequence containing a subsequence along which f is nonincreasing, as stated in Proposition 4.1.

PROPOSITION 4.1. *Let $(x_i)_{i \in \mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2). For every $i \in \mathbb{N}$, let $g(i) \in \operatorname{argmax}_{j \in \{\max\{0, i-l\}, \ldots, i\}} f(x_j)$. Then:*

1. *$(f(x_{g(i)}))_{i \in \mathbb{N}}$ is nonincreasing;*
2. *$(x_i)_{i \in \mathbb{N}}$ is contained in the sublevel set* (1.6)*;*
3. *if $x \in C$ is an accumulation point of $(x_i)_{i \in \mathbb{N}}$, then $(f(x_{g(i)}))_{i \in \mathbb{N}}$ converges to $\varphi \in [f(x), f(x_0)]$;*
4. *if f is bounded from below and uniformly continuous on a set that contains $(x_i)_{i \in \mathbb{N}}$, then $(f(x_i))_{i \in \mathbb{N}}$ converges to $\varphi \in \mathbb{R}$.*

Proof. The first two statements are [22, Lemma 4.1 and Corollary 4.1]. Let $(x_{i_k})_{k \in \mathbb{N}}$ be a subsequence converging to x. Since the sequence $(f(x_{g(i)}))_{i \in \mathbb{N}}$ is non-

increasing, it has a limit in $\mathbb{R} \cup \{-\infty\}$. Thus,

$$\lim_{i\to\infty} f(x_{g(i)}) = \lim_{k\to\infty} f(x_{g(i_k)}) \geq \liminf_{k\to\infty} f(x_{i_k}) = f(x) > -\infty.$$

It remains to prove the fourth statement. From the first statement, and because f is bounded from below, $(f(x_{g(i)}))_{i\in\mathbb{N}}$ converges to some limit $\varphi \in \mathbb{R}$. Assume, for the sake of contradiction, that $(f(x_i))_{i\in\mathbb{N}}$ does not converge to φ. Then, there exist $\rho \in (0,\infty)$ and a subsequence $(f(x_{i_j}))_{j\in\mathbb{N}}$ contained in $\mathbb{R} \setminus [\varphi-\rho, \varphi+\rho]$. For all $j \in \mathbb{N}$, define $p_j \coloneqq g(i_j + l) - i_j \in \{0,\dots,l\}$. Then, there exist $p \in \{0,\dots,l\}$ and a subsequence $(p_{j_k})_{k\in\mathbb{N}}$ such that, for all $k \in \mathbb{N}$, $p_{j_k} = p$. By [22, (27)] or [21, (A.9)], $(f(x_{g(i)-p}))_{i\in\mathbb{N}}$ converges to φ. Therefore, $(f(x_{g(i+l)-p}))_{i\in\mathbb{N}}$ converges to φ. Hence, $(f(x_{g(i_{j_k}+l)-p}))_{k\in\mathbb{N}}$ converges to φ. This is a contradiction since, for all $k \in \mathbb{N}$, $f(x_{g(i_{j_k}+l)-p}) = f(x_{i_{j_k}})$. ∎

5. Convergence analysis for a continuous gradient. In this section, PGD (Algorithm 4.2) is analyzed under the assumption that ∇f is continuous on C. Specifically, the first statement of Theorem 1.2, restated in Theorem 5.1 for convenience, is proven.

THEOREM 5.1. *Let $(x_i)_{i\in\mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2). If ∇f is continuous on C, then all accumulation points of $(x_i)_{i\in\mathbb{N}}$ are B-stationary for* (1.1)*. If, moreover, $(x_i)_{i\in\mathbb{N}}$ has an isolated accumulation point, then $(x_i)_{i\in\mathbb{N}}$ converges.*

The proof is divided into three parts. First, in Section 5.1, we show that, in a neighborhood of any point that is not B-stationary for (1.1), the PGD map (Algorithm 4.1) terminates after a bounded number of iterations. Then, in Section 5.2, we prove that, if a subsequence $(x_{i_k})_{k\in\mathbb{N}}$ converges, then $(x_{i_k+1})_{k\in\mathbb{N}}$ also does, to the same limit. Finally, we combine the first two parts in Section 5.3: if $(x_{i_k})_{k\in\mathbb{N}}$ converges to x, then, from the second part,

$$\|x_{i_k+1} - x_{i_k}\| \to 0 \quad \text{when } k \to \infty,$$

but, from the first part, if x is not B-stationary for (1.1), then the iterates of PGD move by at least a constant amount at each iteration. It is therefore impossible that $(x_{i_k})_{k\in\mathbb{N}}$ converges to a point that is not B-stationary for (1.1).

5.1. First part: analysis of the PGD map. In this section, we show that, if $\underline{x} \in C$ is not B-stationary for (1.1), then the while loop in Algorithm 4.1 terminates, in some neighborhood of $\underline{x}$, for nonvanishing values of α. To show it, we first prove that, in a neighborhood of $\underline{x}$, the first-order term of the Taylor expansion of f is "large" compared to the remainder. This is the goal of Proposition 5.2.

PROPOSITION 5.2. *Assume that ∇f is continuous on C. Let $\underline{x} \in C$ be non-B-stationary for* (1.1)*, and $w \in T_C(\underline{x})$ be such that*

$$\langle w, -\nabla f(\underline{x})\rangle > 0. \tag{5.1}$$

Define $\kappa \coloneqq \sqrt{1 - \frac{\beta\langle w, -\nabla f(\underline{x})\rangle^2}{8\|w\|^2\|\nabla f(\underline{x})\|^2}} \in (0,1)$. There exist arbitrarily small numbers $\alpha_{\underline{x}} \in (0, \underline{\alpha}]$ such that, for some $\rho(\alpha_{\underline{x}}) \in (0,\infty)$, it holds for any $x \in B(\underline{x}, \rho(\alpha_{\underline{x}})) \cap C$ and $\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$ that, for all $y \in P_C(x - \alpha\nabla f(x))$,

$$d(x - \alpha\nabla f(x), C) \leq \kappa\alpha\|\nabla f(x)\|,$$

which implies

$$\langle \nabla f(x), y - x\rangle \leq -\sqrt{1-\kappa^2}\|\nabla f(x)\|\|y-x\|.$$

Proof. Let $\varepsilon \in (0, \underline{\alpha}]$. We show that there exists $\alpha_{\underline{x}} \in (0, \varepsilon)$ satisfying the required property.

Let $(w_i)_{i\in\mathbb{N}}$ be a sequence in C converging to $\underline{x}$, and $(t_i)_{i\in\mathbb{N}}$ be a sequence in $(0,\infty)$ such that

$$\frac{w_i - \underline{x}}{t_i} \overset{i\to\infty}{\longrightarrow} w.$$

From the definition of w in (5.1), it holds for all $i \in \mathbb{N}$ large enough that

$$\langle w_i - \underline{x}, -\nabla f(\underline{x})\rangle > 0. \tag{5.2}$$

As $\frac{1}{t_i}\frac{\|w_i-\underline{x}\|^2}{\langle w_i-\underline{x}, -\nabla f(\underline{x})\rangle} \overset{i\to\infty}{\longrightarrow} \frac{\|w\|^2}{\langle w, -\nabla f(\underline{x})\rangle}$ and $t_i \overset{i\to\infty}{\longrightarrow} 0$, it also holds for all $i \in \mathbb{N}$ large enough that

$$\frac{\|w_i - \underline{x}\|^2}{\langle w_i - \underline{x}, -\nabla f(\underline{x})\rangle} < \varepsilon. \tag{5.3}$$

Similarly, it holds for all $i \in \mathbb{N}$ large enough that

$$\frac{\langle w_i - \underline{x}, -\nabla f(\underline{x})\rangle^2}{\|w_i - \underline{x}\|^2} > \frac{\langle w, -\nabla f(\underline{x})\rangle^2}{2\|w\|^2}. \tag{5.4}$$

Fix $i \in \mathbb{N}$ satisfying (5.2), (5.3), and (5.4). Pick $\alpha_{\underline{x}}$ such that

$$\frac{\alpha_{\underline{x}}}{2} < \frac{\|w_i - \underline{x}\|^2}{\langle w_i - \underline{x}, -\nabla f(\underline{x})\rangle} < \alpha_{\underline{x}} < \varepsilon.$$

Since ∇f is continuous at $\underline{x}$, there exists $\rho_0 \in (0,\infty)$ such that, for all $x \in B[\underline{x}, \rho_0]\cap C$,

$$\langle w_i - \underline{x}, -\nabla f(x)\rangle > 0, \tag{5.5a}$$

$$\frac{\alpha_{\underline{x}}}{2} < \frac{\|w_i - \underline{x}\|^2}{\langle w_i - \underline{x}, -\nabla f(x)\rangle} < \alpha_{\underline{x}}, \tag{5.5b}$$

$$\frac{\langle w_i - \underline{x}, -\nabla f(x)\rangle^2}{\|w_i - \underline{x}\|^2\|\nabla f(x)\|^2} > \frac{\langle w, -\nabla f(\underline{x})\rangle^2}{2\|w\|^2\|\nabla f(\underline{x})\|^2}. \tag{5.5c}$$

We now establish the first inequality we have to prove: for an adequate value of $\rho(\alpha_{\underline{x}})$, it holds for any $x \in B(\underline{x}, \rho(\alpha_{\underline{x}})) \cap C$ and $\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$ that

$$\|x - \alpha\nabla f(x) - y\| \leq \kappa\alpha\|\nabla f(x)\| \quad \forall y \in P_C(x - \alpha\nabla f(x)),$$

which is equivalent to $d(x - \alpha\nabla f(x), C) \leq \kappa\alpha\|\nabla f(x)\|$.

Let us for the moment consider any $\rho(\alpha_{\underline{x}}) \in (0, \rho_0]$. For any $x \in B(\underline{x}, \rho(\alpha_{\underline{x}})) \cap C$,

$\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$, and $y \in P_C(x - \alpha\nabla f(x))$,

$$\begin{aligned}
\|x - \alpha\nabla f(x) - y\|^2 &\leq \|x - \alpha\nabla f(x) - w_i\|^2 \\
&= \|\underline{x} - \alpha\nabla f(x) - w_i\|^2 + 2\left\langle \underline{x} - x, \alpha\nabla f(x) + w_i - \underline{x}\right\rangle + \|\underline{x} - x\|^2 \\
&\leq \|\underline{x} - \alpha\nabla f(x) - w_i\|^2 \\
&\qquad + 2\rho(\alpha_{\underline{x}})\left(\alpha\|\nabla f(x)\| + \|w_i - \underline{x}\|\right) + \rho(\alpha_{\underline{x}})^2 \\
&\leq \|\underline{x} - \alpha\nabla f(x) - w_i\|^2 \\
&\qquad + 2\rho(\alpha_{\underline{x}})\left(\alpha \max_{z \in B[\underline{x}, \rho_0] \cap C} \|\nabla f(z)\| + \|w_i - \underline{x}\|\right) + \rho(\alpha_{\underline{x}})^2 \\
&= \alpha^2\|\nabla f(x)\|^2 - 2\alpha\left\langle w_i - \underline{x}, -\nabla f(x)\right\rangle + \|w_i - \underline{x}\|^2 \\
&\qquad + 2\rho(\alpha_{\underline{x}})\left(\alpha \max_{z \in B[\underline{x}, \rho_0] \cap C} \|\nabla f(z)\| + \|w_i - \underline{x}\|\right) + \rho(\alpha_{\underline{x}})^2 \\
&\leq \alpha^2\|\nabla f(x)\|^2 - \alpha\left\langle w_i - \underline{x}, -\nabla f(x)\right\rangle \\
&\qquad + 2\rho(\alpha_{\underline{x}})\left(\frac{\alpha_{\underline{x}}}{\beta} \max_{z \in B[\underline{x}, \rho_0] \cap C} \|\nabla f(z)\| + \|w_i - \underline{x}\|\right) + \rho(\alpha_{\underline{x}})^2.
\end{aligned}$$

The last inequality is true from (5.5b) and the fact that $\alpha_{\underline{x}} \leq \alpha \leq \frac{\alpha_{\underline{x}}}{\beta}$. Choose $\rho(\alpha_{\underline{x}}) \in (0, \rho_0]$ small enough so that

$$\begin{aligned}
&2\rho(\alpha_{\underline{x}})\left(\frac{\alpha_{\underline{x}}}{\beta} \max_{z \in B[\underline{x}, \rho_0] \cap C} \|\nabla f(z)\| + \|w_i - \underline{x}\|\right) + \rho(\alpha_{\underline{x}})^2 \\
&\leq \frac{\alpha_{\underline{x}}}{2} \min_{z \in B[\underline{x}, \rho_0] \cap C} \left\langle w_i - \underline{x}, -\nabla f(z)\right\rangle.
\end{aligned}$$

Note that the right-hand side of this inequality is positive, from (5.5a). Combining this definition with the previous inequality, we arrive at

$$\begin{aligned}
\|x - \alpha\nabla f(x) - y\|^2 &\leq \alpha^2\|\nabla f(x)\|^2 - \frac{\alpha}{2}\left\langle w_i - \underline{x}, -\nabla f(x)\right\rangle \\
&= \alpha^2\|\nabla f(x)\|^2\left(1 - \frac{\left\langle w_i - \underline{x}, -\nabla f(x)\right\rangle}{2\alpha\|\nabla f(x)\|^2}\right) \\
&\leq \alpha^2\|\nabla f(x)\|^2\left(1 - \frac{\beta\left\langle w_i - \underline{x}, -\nabla f(x)\right\rangle}{2\alpha_{\underline{x}}\|\nabla f(x)\|^2}\right) \\
&\leq \alpha^2\|\nabla f(x)\|^2\left(1 - \frac{\beta\left\langle w_i - \underline{x}, -\nabla f(x)\right\rangle^2}{4\|w_i - \underline{x}\|^2\|\nabla f(x)\|^2}\right) \\
&\leq \alpha^2\|\nabla f(x)\|^2\left(1 - \frac{\beta\left\langle w, -\nabla f(\underline{x})\right\rangle^2}{8\|w\|^2\|\nabla f(\underline{x})\|^2}\right) \\
&= \kappa^2\alpha^2\|\nabla f(x)\|^2.
\end{aligned}$$

In other words, for any $x \in B(\underline{x}, \rho(\alpha_{\underline{x}})) \cap C$, $\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$, and $y \in P_C(x - \alpha\nabla f(x))$, it holds that

$$\|x - \alpha\nabla f(x) - y\| \leq \kappa\alpha\|\nabla f(x)\|.$$

To conclude, we show that this inequality implies

$$\left\langle \frac{y - x}{\|y - x\|}, \frac{\nabla f(x)}{\|\nabla f(x)\|} \right\rangle \leq -\sqrt{1 - \kappa^2}. \tag{5.6}$$

Indeed, if we define $\theta \in \mathbb{R}$ such that $\left\langle \frac{y-x}{\|y-x\|}, \frac{\nabla f(x)}{\|\nabla f(x)\|} \right\rangle = \cos(\theta)$, we have

$$\|y-x\|^2 + 2\alpha\|\nabla f(x)\|\|y-x\|\cos(\theta) + \alpha^2\|\nabla f(x)\|^2 \leq \alpha^2\kappa^2\|\nabla f(x)\|^2.$$

This already shows that $\cos(\theta) < 0$. In addition, if we minimize the left-hand side over all possible values of $\|y-x\|$, we get

$$-\alpha^2\|\nabla f(x)\|^2\cos^2(\theta) + \alpha^2\|\nabla f(x)\|^2 \leq \alpha^2\kappa^2\|\nabla f(x)\|^2,$$

hence $\cos^2(\theta) \geq 1-\kappa^2$, which establishes (5.6). ∎

PROPOSITION 5.3. *Assume that ∇f is continuous on C. Let $\underline{x} \in C$ be non-B-stationary for (1.1). There exists $\alpha_{\underline{x}} \in (0, \underline{\alpha}]$ such that, for some $\rho \in (0,\infty)$, it holds for any $x \in B(\underline{x},\rho)\cap C$, $\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$, and $y \in P_C(x-\alpha\nabla f(x))$ that*

$$f(y) < f(x) + c\,\langle \nabla f(x), y-x\rangle\,.$$

Proof. Fix $\alpha_{\underline{x}}$ as in Proposition 5.2, small enough so that

$$\sup_{y \in B\left[\underline{x}, \frac{7\alpha_{\underline{x}}}{2\beta}\|\nabla f(\underline{x})\|\right]\cap C} \frac{|f(y) - f(\underline{x}) - \langle \nabla f(\underline{x}), y-\underline{x}\rangle\,|}{\|y-\underline{x}\|} < \frac{(1-c)\sqrt{1-\kappa^2}\|\nabla f(\underline{x})\|}{4\left(1+\frac{8}{3(1-\kappa)}\right)}, \tag{5.7a}$$

$$\sup_{y \in B\left[\underline{x}, \frac{7\alpha_{\underline{x}}}{2\beta}\|\nabla f(\underline{x})\|\right]\cap C} \|\nabla f(y) - \nabla f(\underline{x})\| < \frac{(1-c)\sqrt{1-\kappa^2}}{4}\|\nabla f(\underline{x})\|. \tag{5.7b}$$

These inequalities are satisfied by all $\alpha_{\underline{x}}$ small enough, from the definition of the gradient for the first one, and because the gradient is continuous at $\underline{x}$ for the second one.

Let $\rho(\alpha_{\underline{x}})$ be as in Proposition 5.2. Define

$$\rho \coloneqq \min\left\{\rho(\alpha_{\underline{x}}), \alpha_{\underline{x}}\|\nabla f(\underline{x})\|\right\}.$$

Note that, for all $x \in B(\underline{x},\rho)\cap C$,

$$\|x-\underline{x}\| < \rho \leq \alpha_{\underline{x}}\|\nabla f(\underline{x})\| < \frac{7\alpha_{\underline{x}}}{2\beta}\|\nabla f(\underline{x})\|,$$

so that from (5.7b),

$$\frac{3}{4}\|\nabla f(\underline{x})\| < \|\nabla f(x)\| < \frac{5}{4}\|\nabla f(\underline{x})\|. \tag{5.8}$$

For any $x \in B(\underline{x},\rho)\cap C$, $\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$, and $y \in P_C(x-\alpha\nabla f(x))$,

$$\begin{aligned} f(y) &= f(x) + \langle \nabla f(\underline{x}), y-x\rangle \\ &\quad + (f(\underline{x}) - f(x) - \langle \nabla f(\underline{x}), \underline{x}-x\rangle) \\ &\quad + (f(y) - f(\underline{x}) - \langle \nabla f(\underline{x}), y-\underline{x}\rangle) \\ &\leq f(x) + \langle \nabla f(\underline{x}), y-x\rangle + \frac{(1-c)\sqrt{1-\kappa^2}\|\nabla f(\underline{x})\|}{4\left(1+\frac{8}{3(1-\kappa)}\right)}\left(\|\underline{x}-x\| + \|y-\underline{x}\|\right). \end{aligned} \tag{5.9}$$

The last inequality follows from (5.7a); observe that

$$\begin{aligned}\|y-\underline{x}\| &\leq \|y-x\| + \|x-\underline{x}\| \\ &\leq 2\alpha\|\nabla f(x)\| + \rho \text{ from (2.1)} \\ &\leq \frac{2\alpha_{\underline{x}}}{\beta}\|\nabla f(x)\| + \alpha_{\underline{x}}\|\nabla f(\underline{x})\| \\ &< \frac{5\alpha_{\underline{x}}}{2\beta}\|\nabla f(\underline{x})\| + \alpha_{\underline{x}}\|\nabla f(\underline{x})\| \\ &\leq \frac{7\alpha_{\underline{x}}}{2\beta}\|\nabla f(\underline{x})\|.\end{aligned}$$

We continue from (5.9):

$$\begin{aligned}f(y) &\leq f(x) + \langle \nabla f(\underline{x}), y-x\rangle + \frac{(1-c)\sqrt{1-\kappa^2}\|\nabla f(\underline{x})\|}{4\left(1+\frac{8}{3(1-\kappa)}\right)}\left(2\|\underline{x}-x\| + \|y-x\|\right) \\ &\overset{(a)}{<} f(x) + \langle \nabla f(\underline{x}), y-x\rangle + \frac{(1-c)\sqrt{1-\kappa^2}\|\nabla f(\underline{x})\|}{4}\|y-x\| \\ &\leq f(x) + \langle \nabla f(x), y-x\rangle + \|\nabla f(\underline{x}) - \nabla f(x)\|\|y-x\| \\ &\quad + \frac{(1-c)\sqrt{1-\kappa^2}\|\nabla f(\underline{x})\|}{4}\|y-x\| \\ &\leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{(1-c)\sqrt{1-\kappa^2}\|\nabla f(\underline{x})\|}{2}\|y-x\| \text{ from (5.7b)} \\ &< f(x) + \langle \nabla f(x), y-x\rangle + (1-c)\sqrt{1-\kappa^2}\|\nabla f(x)\|\|y-x\| \text{ from (5.8)} \\ &\leq f(x) + \langle \nabla f(x), y-x\rangle - (1-c)\langle \nabla f(x), y-x\rangle \text{ from Proposition 5.2} \\ &= f(x) + c\langle \nabla f(x), y-x\rangle.\end{aligned}$$

Inequality (a) is true because

$$\begin{aligned}\|y-x\| &\geq \alpha\|\nabla f(x)\| - \|x-\alpha\nabla f(x) - y\| \\ &= \alpha\|\nabla f(x)\| - d(x-\alpha\nabla f(x), C) \\ &\geq (1-\kappa)\alpha\|\nabla f(x)\| \\ &\geq \frac{3}{4}(1-\kappa)\rho \\ &> \frac{3}{4}(1-\kappa)\|x-\underline{x}\|.\end{aligned}$$

∎

5.2. Second part: convergence of successive iterates.

PROPOSITION 5.4. *Assume that ∇f is continuous on C. Let $(x_i)_{i\in\mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2), and x be an accumulation point. Then, for any subsequence $(x_{i_k})_{k\in\mathbb{N}}$ converging to x, the sequence $(x_{i_k+1})_{k\in\mathbb{N}}$ also converges to x.*

Proof. Let $(x_{i_k})_{k\in\mathbb{N}}$ be a subsequence converging to x. We show that $(x_{i_k+1})_{k\in\mathbb{N}}$ also converges to x.

It suffices to show that x is an accumulation point of every subsequence of $(x_{i_k+1})_{k\in\mathbb{N}}$. In other words, we show the following: for every subsequence $(i_{j_k})_{k\in\mathbb{N}}$ of $(i_k)_{k\in\mathbb{N}}$, there exists a subsequence of $(x_{i_{j_k}+1})_{k\in\mathbb{N}}$ that converges to x. Let $(i_{j_k})_{k\in\mathbb{N}}$ be a subsequence of $(i_k)_{k\in\mathbb{N}}$. For all $i\in\mathbb{N}$, define $g(i) \in \operatorname{argmax}_{j\in\{\max\{0,i-l\},\dots,i\}} f(x_j)$,

as in Proposition 4.1. By the third statement of Proposition 4.1, the sequence $(f(x_{g(i)}))_{i\in\mathbb{N}}$ converges to $\varphi \in [f(x), f(x_0)]$. For every $k \in \mathbb{N}$, letting $\alpha_{i_{j_k}} \in (0, \overline{\alpha}]$ be the number such that $x_{i_{j_k}+1} \in P_C(x_{i_{j_k}} - \alpha_{i_{j_k}} \nabla f(x_{i_{j_k}}))$, by (2.1),

$$\|x_{i_{j_k}+1} - x_{i_{j_k}}\| \leq 2\alpha_{i_{j_k}} \|\nabla f(x_{i_{j_k}})\| \leq 2\overline{\alpha}\|\nabla f(x_{i_{j_k}})\|.$$

Thus, since $(x_{i_{j_k}})_{k\in\mathbb{N}}$ is bounded and ∇f is locally bounded (as it is continuous), the sequence $(x_{i_{j_k}+1})_{k\in\mathbb{N}}$ is bounded. If we replace $(i_{j_k})_{k\in\mathbb{N}}$ by a subsequence, we can assume that $(x_{i_{j_k}+1})_{k\in\mathbb{N}}$ converges.

Iterating the reasoning, we can assume that $(x_{i_{j_k}+s})_{k\in\mathbb{N}}$ converges to some $x^s \in C$ for every $s \in \{0, \dots, l+1\}$. By definition of x, $x^0 = x$.

Observe that, from the continuity of f,

$$\begin{aligned} f(x_{g(i_{j_k}+l+1)}) &= \max\{f(x_{i_{j_k}+1}), \dots, f(x_{i_{j_k}+l+1})\} \\ &\to \max\{f(x^1), \dots, f(x^{l+1})\} \text{ when } k \to \infty. \end{aligned}$$

In particular, there exists $s_1 \in \{1, \dots, l+1\}$ such that

$$f(x^{s_1}) = \varphi. \tag{5.10}$$

Let s_1 be the smallest such integer. For any $k \in \mathbb{N}$, from the condition in line 2 of Algorithm 4.1 and (4.1),

$$f(x_{i_{j_k}+s_1}) \leq f(x_{g(i_{j_k}+s_1-1)}) - \frac{c}{2\overline{\alpha}}\|x_{i_{j_k}+s_1} - x_{i_{j_k}+s_1-1}\|^2.$$

Letting k tend to infinity yields

$$\varphi = f(x^{s_1}) \leq \varphi - \frac{c}{2\overline{\alpha}}\|x^{s_1} - x^{s_1-1}\|^2.$$

Consequently, $x^{s_1} = x^{s_1-1}$. In particular, $f(x^{s_1-1}) = f(x^{s_1}) = \varphi$. Therefore, $s_1 = 1$, otherwise it would not be the smallest integer satisfying (5.10). The equality $x^{s_1} = x^{s_1-1}$ then rewrites as $x^1 = x^0 = x$ and, when $k \to \infty$,

$$x_{i_{j_k}+1} \to x^1 = x.$$

∎

5.3. Third part: proof of Theorem 5.1. Let $\underline{x}$ be an accumulation point of $(x_i)_{i\in\mathbb{N}}$. Assume, for the sake of contradiction, that $\underline{x}$ is not B-stationary for (1.1). Let $(x_{i_k})_{k\in\mathbb{N}}$ be a subsequence converging to $\underline{x}$.

Let $\alpha_{\underline{x}}$ and ρ be as in Proposition 5.3. For all $k \in \mathbb{N}$ large enough, $x_{i_k} \in B(\underline{x}, \rho) \cap C$. Then, when Algorithm 4.1 is called at point x_{i_k}, the condition in line 2 stops being fulfilled for some $\alpha_{i_k} \geq \alpha_{\underline{x}}$, meaning that

$$x_{i_k+1} \in P_C(x_{i_k} - \alpha_{i_k} \nabla f(x_{i_k})) \text{ for some } \alpha_{i_k} \in [\alpha_{\underline{x}}, \overline{\alpha}].$$

If we replace $(i_k)_{k\in\mathbb{N}}$ with a subsequence, we can assume that $(\alpha_{i_k})_{k\in\mathbb{N}}$ converges to some $\alpha_{\lim} \in [\alpha_{\underline{x}}, \overline{\alpha}]$.

For any $k \in \mathbb{N}$, we have

$$\|x_{i_k} - \alpha_{i_k} \nabla f(x_{i_k}) - x_{i_k+1}\| = d(x_{i_k} - \alpha_{i_k} \nabla f(x_{i_k}), C)$$

and since the distance to a nonempty closed set is a continuous function, we can take this equality to the limit. We use the fact that $x_{i_k+1} \to \underline{x}$ when $k \to \infty$, from Proposition 5.4. This yields

$$\|\alpha_{\lim} \nabla f(\underline{x})\| = d(\underline{x} - \alpha_{\lim} \nabla f(\underline{x}), C),$$

which means that $\underline{x} \in P_C(\underline{x} - \alpha_{\lim} \nabla f(\underline{x}))$. In particular, $-\nabla f(\underline{x}) \in \widehat{\widehat{N}}_C(\underline{x}) \subseteq \widehat{N}_C(\underline{x})$, which contradicts our assumption that $\underline{x}$ is not B-stationary for (1.1). We have therefore proven that any accumulation point is B-stationary.

Finally, if $(x_i)_{i\in\mathbb{N}}$ has an isolated accumulation point, then the sequence $(x_i)_{i\in\mathbb{N}}$ converges, from Proposition 5.4 and [33, Lemma 4.10].

6. Convergence analysis for a locally Lipschitz continuous gradient. In this section, PGD (Algorithm 4.2) is analyzed under the assumption that f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous. Specifically, the second statement of Theorem 1.2, restated as Theorem 6.3 for convenience, is proven based on Proposition 6.1 and Corollary 6.2 which state that, for every $\underline{x} \in C$ and every input x sufficiently close to $\underline{x}$, the PGD map (Algorithm 4.1) terminates after at most a given number of iterations which depends only on $\underline{x}$.

PROPOSITION 6.1. *Assume that f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous. For every $\underline{x} \in C$, $\overline{\alpha} \in (0,\infty)$, and $c \in (0,1)$, there exists $\rho \in (0,\infty)$ such that, with $\overline{\rho} \coloneqq \rho + 3\overline{\alpha}\|\nabla f(\underline{x})\|$ and $\alpha_* \coloneqq (1-c)/\operatorname{Lip}_{B[\underline{x},\overline{\rho}]}(\nabla f)$, for all $x \in B[\underline{x},\rho] \cap C$, $\alpha \in [0, \min\{\alpha_*, \overline{\alpha}\}]$, and $y \in P_C(x - \alpha\nabla f(x))$,*

$$f(y) \leq f(x) + c\,\langle \nabla f(x), y - x\rangle\,.$$

Proof. Let $\underline{x} \in C$, $\overline{\alpha} \in (0,\infty)$, and $c \in (0,1)$. Since ∇f is continuous at $\underline{x}$, there exists $\rho \in (0,\infty)$ such that, for all $x \in B[\underline{x},\rho]$, $\|\nabla f(x) - \nabla f(\underline{x})\| \leq \frac{1}{2}\|\nabla f(\underline{x})\|$ and hence, as $|\|\nabla f(x)\| - \|\nabla f(\underline{x})\|| \leq \|\nabla f(x) - \nabla f(\underline{x})\|$,

$$\frac{1}{2}\|\nabla f(\underline{x})\| \leq \|\nabla f(x)\| \leq \frac{3}{2}\|\nabla f(\underline{x})\|.$$

For all $x \in B[\underline{x},\rho]$ and $\alpha \in [0,\overline{\alpha}]$, $P_C(x - \alpha\nabla f(x)) \subseteq B[\underline{x},\overline{\rho}]$; indeed, for all $y \in P_C(x - \alpha\nabla f(x))$,

$$\|y - \underline{x}\| \leq \|y - x\| + \|x - \underline{x}\| \leq 2\alpha\|\nabla f(x)\| + \rho \leq \overline{\rho},$$

where the second inequality follows from (2.1). Thus, by (2.3) and (2.2), for all $x \in B[\underline{x},\rho]$, $\alpha \in [0, \min\{\alpha_*, \overline{\alpha}\}]$, and $y \in P_C(x - \alpha\nabla f(x))$,

$$\begin{aligned} f(y) &\leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2} \operatorname*{Lip}_{B[\underline{x},\overline{\rho}]} (\nabla f)\|y - x\|^2 \\ &\leq f(x) + \left(1 - \alpha \operatorname*{Lip}_{B[\underline{x},\overline{\rho}]} (\nabla f)\right) \langle \nabla f(x), y - x\rangle \\ &\leq f(x) + c\,\langle \nabla f(x), y - x\rangle\,. \end{aligned}$$

∎

COROLLARY 6.2. *Consider Algorithm 4.1 under the assumption that f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous. Given $\underline{x} \in C$, let ρ and $\overline{\rho}$ be as*

in Proposition 6.1. Then, for every $x \in B[\underline{x}, \rho] \cap C$, the while loop terminates with a step size $\alpha \in \left[\min\left\{\underline{\alpha}, \frac{\beta(1-c)}{\operatorname{Lip}_{B[\underline{x},\overline{\rho}]}(\nabla f)}\right\}, \overline{\alpha}\right]$ and hence after at most

$$\max\left\{0, \left\lceil \ln\left(\frac{1-c}{\alpha_0 \operatorname{Lip}_{B[\underline{x},\overline{\rho}]}(\nabla f)}\right) / \ln(\beta) \right\rceil\right\}$$

iterations, where α_0 is the step size chosen in line 1.

Proof. Either the initial step size chosen in $[\underline{\alpha}, \overline{\alpha}]$ satisfies the Armijo condition or the while loop ends after iteration $i \in \mathbb{N} \setminus \{0\}$ with $\alpha = \alpha_0 \beta^i$ such that $\frac{\alpha}{\beta} > \frac{1-c}{\operatorname{Lip}_{B[\underline{x},\overline{\rho}]}(\nabla f)}$. In the second case, $i < 1 + \ln(\frac{1-c}{\alpha_0 \operatorname{Lip}_{B[\underline{x},\overline{\rho}]}(\nabla f)}) / \ln(\beta)$ and thus $i \leq \lceil \ln(\frac{1-c}{\alpha_0 \operatorname{Lip}_{B[\underline{x},\overline{\rho}]}(\nabla f)}) / \ln(\beta) \rceil$. ∎

THEOREM 6.3. *Let $(x_i)_{i\in\mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2). If f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous, then all accumulation points of $(x_i)_{i\in\mathbb{N}}$ are P-stationary for* (1.1).

Proof. Assume that a subsequence $(x_{i_j})_{j\in\mathbb{N}}$ converges to $\underline{x} \in C$. Let ρ and $\overline{\rho}$ be as in Proposition 6.1. Define

$$I \coloneqq \left[\min\left\{\underline{\alpha}, \frac{\beta(1-c)}{\operatorname{Lip}_{B[\underline{x},\overline{\rho}]}(\nabla f)}\right\}, \overline{\alpha}\right].$$

There exists $j_* \in \mathbb{N}$ such that, for all integers $j \geq j_*$, $x_{i_j} \in B[\underline{x}, \rho]$, thus, by Corollary 6.2, $x_{i_j+1} \in P_C(x_{i_j} - \alpha_{i_j} \nabla f(x_{i_j}))$ with $\alpha_{i_j} \in I$, and hence

$$\|x_{i_j+1} - (x_{i_j} - \alpha_{i_j} \nabla f(x_{i_j}))\| = d(x_{i_j} - \alpha_{i_j} \nabla f(x_{i_j}), C).$$

Since I is compact, a subsequence $(\alpha_{i_{j_k}})_{k\in\mathbb{N}}$ converges to $\alpha \in I$. Moreover, there exists $k_* \in \mathbb{N}$ such that $j_{k_*} \geq j_*$. Furthermore, by Proposition 5.4, $(x_{i_j+1})_{j\in\mathbb{N}}$ converges to $\underline{x}$. Therefore, for all integers $k \geq k_*$,

$$\|x_{i_{j_k}+1} - (x_{i_{j_k}} - \alpha_{i_{j_k}} \nabla f(x_{i_{j_k}}))\| = d(x_{i_{j_k}} - \alpha_{i_{j_k}} \nabla f(x_{i_{j_k}}), C),$$

and letting k tend to infinity yields

$$\|\underline{x} - (\underline{x} - \alpha \nabla f(\underline{x}))\| = d(\underline{x} - \alpha \nabla f(\underline{x}), C).$$

It follows that $\underline{x} \in P_C(\underline{x} - \alpha \nabla f(\underline{x}))$, which implies that $-\nabla f(\underline{x}) \in \widehat{\widehat{N}}_C(\underline{x})$. ∎

Proposition 6.4 considers the case where PGD generates a bounded sequence.

PROPOSITION 6.4. *If* PGD *(Algorithm 4.2) generates a bounded sequence $(x_i)_{i\in\mathbb{N}}$, which is the case if the sublevel set* (1.6) *is bounded, then all of its accumulation points, of which there exists at least one, are P-stationary for* (1.1) *and have the same image by f.*

Proof. Assume that PGD (Algorithm 4.2) generates a bounded sequence $(x_i)_{i\in\mathbb{N}}$. It suffices to prove that all of its accumulation points have the same image by f; the other statements follow from Theorem 6.3. The proof is based on the argument given in the proof of [40, Theorem 65]. Assume that $(x_{i_k})_{k\in\mathbb{N}}$ and $(x_{j_k})_{k\in\mathbb{N}}$ converge respectively to $\underline{x}$ and $\overline{x}$. Being bounded, the sequence $(x_i)_{i\in\mathbb{N}}$ is contained in a compact set. By Proposition 4.1, since a continuous, real-valued function is bounded from below

and uniformly continuous on every compact set [47, Propositions 1.3.3 and 1.3.5], the sequence $(f(x_i))_{i\in\mathbb{N}}$ converges. Therefore, $f(\underline{x}) = \lim_{k\to\infty} f(x_{i_k}) = \lim_{i\to\infty} f(x_i) = \lim_{k\to\infty} f(x_{j_k}) = f(\overline{x})$. ∎

7. Proximal normal cones to some stratified sets. The following examples of a set C that is not Clarke regular at infinitely many points are studied in [35]:

1. the closed cone $\mathbb{R}^n_{\leq s}$ of s-sparse vectors of $\mathbb{R}^n$, i.e., those having at most s nonzero components, n and s being positive integers such that $s < n$;
2. the closed cone $\mathbb{R}^n_{\leq s} \cap \mathbb{R}^n_+$ of nonnegative s-sparse vectors of $\mathbb{R}^n$;
3. the determinantal variety [18, Lecture 9]

$$\mathbb{R}^{m\times n}_{\leq r} \coloneqq \{X \in \mathbb{R}^{m\times n} \mid \operatorname{rank} X \leq r\},$$

   m, n, and r being positive integers such that $r < \min\{m, n\}$;
4. the closed cone

$$\mathrm{S}^+_{\leq r}(n) \coloneqq \{X \in \mathbb{R}^{n\times n}_{\leq r} \mid X^\top = X,\, X \succeq 0\}$$

   of order-n real symmetric positive-semidefinite matrices of rank at most r, n and r being positive integers such that $r < n$.

In this section, we prove that, for these sets, regular normals are proximal normals.

As detailed in [35], if C is a set in this list, then there exist a positive integer p and disjoint nonempty smooth submanifolds $S_0, \ldots, S_p$ of $\mathcal{E}$ such that $\overline{S_p} = C$ and, for all $i \in \{0, \ldots, p\}$, $\overline{S_i} = \bigcup_{j=0}^{i} S_j$. This implies that $\{S_0, \ldots, S_p\}$ is a *stratification* of C satisfying the *condition of the frontier* [30, §5]. Thus, C is called a *stratified set* and $S_0, \ldots, S_p$ are called the *strata* of $\{S_0, \ldots, S_p\}$.

Proposition 7.1. *Let C be a set in the list. For all $x \in C$,*

$$\widehat{\widehat{N}}_C(x) = \widehat{N}_C(x)$$

and, if $x \notin S_p$, then

$$\widehat{N}_C(x) \subsetneq N_C(x).$$

Proof. The strict inclusion follows from [35, Proposition 7.16] and [3, Theorem 3.9] if $C = \mathbb{R}^n_{\leq s}$, from [35, Proposition 6.7] and [45, Theorem 3.4] if $C = \mathbb{R}^n_{\leq s} \cap \mathbb{R}^n_+$, from [19, Corollary 2.3 and Theorem 3.1] if $C = \mathbb{R}^{m\times n}_{\leq r}$, and from [35, Proposition 6.28] and [45, Theorem 3.12] if $C = \mathrm{S}^+_{\leq r}(n)$. By (1.2), it remains to prove that, for all $x \in C$, $\widehat{\widehat{N}}_C(x) \supseteq \widehat{N}_C(x)$. This follows from [1, Lemma 4] if $x \in S_p$. Let $x \in C \setminus S_p$. If C is $\mathbb{R}^n_{\leq s}$ or $\mathbb{R}^{m\times n}_{\leq r}$, then, by [35, Proposition 7.16] and [19, Corollary 2.3], $\widehat{N}_C(x) = \{0\}$ and the result follows. If C is $\mathbb{R}^n_{\leq s} \cap \mathbb{R}^n_+$ or $\mathrm{S}^+_{\leq r}(n)$, then the result follows from [35, Proposition 6.7] and [45, Proposition 3.2] or [35, Proposition 6.28] and [7, Corollary 17]; the detail is given below for completeness.

Assume that C is $\mathbb{R}^n_{\leq s} \cap \mathbb{R}^n_+$. Let $\operatorname{supp}(x) \coloneqq \{i \in \{1, \ldots, n\} \mid x_i \neq 0\}$. By [35, Proposition 6.7],

$$\widehat{N}_{\mathbb{R}^n_{\leq s}\cap\mathbb{R}^n_+}(x) = \{v \in \mathbb{R}^n_- \mid \operatorname{supp}(v) \subseteq \{1, \ldots, n\} \setminus \operatorname{supp}(x)\}.$$

Thus, by [45, Proposition 3.2], for every $v \in \widehat{N}_{\mathbb{R}^n_{\leq s}\cap\mathbb{R}^n_+}(x)$, $P_{\mathbb{R}^n_{\leq s}\cap\mathbb{R}^n_+}(x + v) = \{x\}$.

Assume now that C is $\mathrm{S}^+_{\leq r}(n)$. By [35, Proposition 6.28],

$$\widehat{N}_{\mathrm{S}^+_{\leq r}(n)}(X) = \mathrm{S}(n)^\perp + \{Z \in \mathrm{S}^-(n) \mid XZ = 0_{n\times n}\},$$

with $\mathrm{S}(n) \coloneqq \{X \in \mathbb{R}^{n\times n} \mid X^\top = X\}$, $\mathrm{S}(n)^\perp = \{X \in \mathbb{R}^{n\times n} \mid X^\top = -X\}$, and $\mathrm{S}^-(n) \coloneqq \{X \in \mathrm{S}(n) \mid X \preceq 0\}$. Let $Z \in \widehat{N}_{\mathrm{S}^+_{\leq r}(n)}(X)$ and $Z_{\mathrm{sym}} \coloneqq \frac{1}{2}(Z + Z^\top)$. Then, by [7, Corollary 17], $P_{\mathrm{S}^+_{\leq r}(n)}(X + Z) = P_{\mathrm{S}^+_{\leq r}(n)}(X + Z_{\mathrm{sym}})$. Let $\underline{r} \coloneqq \operatorname{rank} X$ and $\tilde{r} \coloneqq \operatorname{rank} Z_{\mathrm{sym}}$. Since $\operatorname{im} Z_{\mathrm{sym}} \subseteq \ker X$, $\tilde{r} \leq n - \underline{r}$ and there exists $U \in \mathrm{O}(n)$ such that

$$X = U \operatorname{diag}(\lambda_1(X), \ldots, \lambda_{\underline{r}}(X), 0_{n-\underline{r}})U^\top$$

and

$$Z_{\mathrm{sym}} = U \operatorname{diag}(0_{n-\tilde{r}}, \lambda_{n-\tilde{r}+1}(Z_{\mathrm{sym}}), \ldots, \lambda_n(Z_{\mathrm{sym}}))U^\top$$

are eigendecompositions. Thus,

$$X + Z_{\mathrm{sym}} = U \operatorname{diag}(\lambda_1(X), \ldots, \lambda_{\underline{r}}(X), 0_{n-\underline{r}-\tilde{r}}, \lambda_{n-\tilde{r}+1}(Z_{\mathrm{sym}}), \ldots, \lambda_n(Z_{\mathrm{sym}}))U^\top$$

is an eigendecomposition. Hence, by [7, Corollary 17], $P_{\mathrm{S}^+_{\leq r}(n)}(X + Z_{\mathrm{sym}}) = \{X\}$. ∎

8. Conclusion. In this paper, PGD is proven to accumulate at B-stationary points of (1.1) if ∇f is continuous on C, and even at P-stationary points of (1.1) if f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous. These are the strongest stationarity properties that can be expected for problem (1.1) under the considered assumptions.

A sufficient condition for the convergence of the sequence generated by PGD is provided in Theorem 5.1. However, if satisfied, this condition does not offer a characterization of the rate of convergence. This important matter is addressed in [20] for monotone PGD under the assumption that f is differentiable on $\mathcal{E}$, ∇f is locally Lipschitz continuous, and f satisfies a Kurdyka–Łojasiewicz property.

This paper opens several avenues of research.

1. Is it possible to extend the results to more general search directions? For example, a search direction at a point $x \in C$ that is not B-stationary for (1.1) could be any $v \notin \widehat{N}_C(x)$ that satisfies [14, conditions (2) and (3)], i.e., $\langle \nabla f(x), v\rangle \leq -c_1\|\nabla f(x)\|^2$ and $\|v\| \leq c_2\|\nabla f(x)\|$ with $c_1, c_2 \in (0, \infty)$.
2. Are there necessary or sufficient conditions on C for the equality $\widehat{\widehat{N}}_C(x) = \widehat{N}_C(x)$ to hold at all $x \in C$?

REFERENCES

[1] P.-A. Absil and J. Malick, *Projection-like retractions on matrix manifolds*, SIAM Journal on Optimization, 22 (2012), pp. 135–158, https://doi.org/10.1137/100802529.

[2] M. V. Balashov, B. T. Polyak, and A. A. Tremba, *Gradient projection and conditional gradient methods for constrained nonconvex minimization*, Numerical Functional Analysis and Optimization, 41 (2020), pp. 822–849, https://doi.org/10.1080/01630563.2019.1704780.

[3] H. H. Bauschke, D. R. Luke, H. M. Phan, and X. Wang, *Restricted normal cones and sparsity optimization with affine constraints*, Foundations of Computational Mathematics, 14 (2014), pp. 63–83, https://doi.org/10.1007/s10208-013-9161-0.

[4] A. Beck and Y. C. Eldar, *Sparsity constrained nonlinear optimization: Optimality conditions and algorithms*, SIAM Journal on Optimization, 23 (2013), pp. 1480–1509, https://doi.org/10.1137/120869778.

[5] A. Beck and M. Teboulle, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, vol. 49 of Springer Optimization and Its Applications, Springer New York, 2011, ch. A Linearly Convergent Algorithm for Solving a Class of Nonconvex/Affine Feasibility Problems, pp. 33–48, https://doi.org/10.1007/978-1-4419-9569-8_3.

[6] Y. Cui and J.-S. Pang, *Modern Nonconvex Nondifferentiable Optimization*, MOS-SIAM Series on Optimization, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2021, https://doi.org/10.1137/1.9781611976748.

[7] A. Dax, *Low-rank positive approximants of symmetric matrices*, Advances in Linear Algebra & Matrix Theory, 4 (2014), pp. 172–185, https://doi.org/10.4236/alamt.2014.43015.

[8] C. Ding, D. Sun, and J. J. Ye, *First order optimality conditions for mathematical programs with semidefinite cone complementarity constraints*, Mathematical Programming, 147 (2014), pp. 539–579, https://doi.org/10.1007/s10107-013-0735-z.

[9] S. Dolecki and G. H. Greco, *Towards historical roots of necessary conditions of optimality: Regula of Peano*, Control and Cybernetics, 36 (2007), pp. 491–518.

[10] S. Dolecki and G. H. Greco, *Tangency vis-à-vis differentiability by Peano, Severi and Guareschi*, Journal of Convex Analysis, 18 (2011), p. 301–339.

[11] M. L. Flegel and C. Kanzow, *On M-stationary points for mathematical programs with equilibrium constraints*, Journal of Mathematical Analysis and Applications, 310 (2005), pp. 286–302, https://doi.org/10.1016/j.jmaa.2005.02.011.

[12] M. Fukushima and G.-H. Lin, *Smoothing methods for mathematical programs with equilibrium constraints*, in International Conference on Informatics Research for Development of Knowledge Society Infrastructure, 2004 (ICKS 2004), Kyoto, Japan, 2004, IEEE, pp. 206–213, https://doi.org/10.1109/ICKS.2004.1313426.

[13] B. Gao, R. Peng, and Y.-x. Yuan, *Low-rank optimization on Tucker tensor varieties*, (2023), https://arxiv.org/abs/2311.18324.

[14] L. Grippo, F. Lampariello, and S. Lucidi, *A nonmonotone line search technique for Newton's method*, SIAM Journal on Numerical Analysis, 23 (1986), pp. 707–716, https://doi.org/10.1137/0723046.

[15] M. Guignard, *Generalized Kuhn–Tucker conditions for mathematical programming problems in a Banach space*, SIAM Journal on Control, 7 (1969), pp. 232–241, https://doi.org/10.1137/0307016.

[16] L. Guo and G.-H. Lin, *Notes on some constraint qualifications for mathematical programs with equilibrium constraints*, Journal of Optimization Theory and Applications, 156 (2013), pp. 600–616, https://doi.org/10.1007/s10957-012-0084-8.

[17] W. Ha, H. Liu, and R. F. Barber, *An equivalence between critical points for rank constraints versus low-rank factorizations*, SIAM Journal on Optimization, 30 (2020), pp. 2927–2955, https://doi.org/10.1137/18M1231675.

[18] J. Harris, *Algebraic Geometry*, vol. 133 of Graduate Texts in Mathematics, Springer-Verlag New York, 1992, https://doi.org/10.1007/978-1-4757-2189-8.

[19] S. Hosseini, D. R. Luke, and A. Uschmajew, *Nonsmooth Optimization and Its Applications*, vol. 170 of International Series of Numerical Mathematics, Birkhäuser Cham, 2019, ch. Tangent and Normal Cones for Low-Rank Matrices, pp. 45–53, https://doi.org/10.1007/978-3-030-11370-4_3.

[20] X. Jia, C. Kanzow, and P. Mehlitz, *Convergence analysis of the proximal gradient method in the presence of the Kurdyka–Łojasiewicz property without global Lipschitz assumptions*, SIAM Journal on Optimization, 33 (2023), pp. 3038–3056, https://doi.org/10.1137/23M1548293.

[21] X. Jia, C. Kanzow, P. Mehlitz, and G. Wachsmuth, *An augmented Lagrangian method for optimization problems with structured geometric constraints*, Mathematical Programming, 199 (2023), pp. 1365–1415, https://doi.org/10.1007/s10107-022-01870-z.

[22] C. Kanzow and P. Mehlitz, *Convergence properties of monotone and nonmonotone proximal gradient methods revisited*, Journal of Optimization Theory and Applications, 195 (2022), pp. 624–646, https://doi.org/10.1007/s10957-022-02101-3.

[23] E. Levin, J. Kileel, and N. Boumal, *The effect of smooth parametrizations on nonconvex optimization landscapes*, (2023), https://arxiv.org/abs/2207.03512v4.

[24] E. Levin, J. Kileel, and N. Boumal, *Finding stationary points on bounded-rank matrices: a geometric hurdle and a smooth remedy*, Mathematical Programming, 199 (2023), pp. 831–864, https://doi.org/10.1007/s10107-022-01851-2.

[25] X. Li and Z. Luo, *Normal cones intersection rule and optimality analysis for low-rank matrix*

optimization with affine manifolds, SIAM Journal on Optimization, 33 (2023), pp. 1333–1360, https://doi.org/10.1137/22M147863X.

[26] X. Li, W. Song, and N. Xiu, *Optimality conditions for rank-constrained matrix optimization*, Journal of the Operations Research Society of China, 7 (2019), pp. 285–301, https://doi.org/10.1007/s40305-019-00245-0.

[27] X. Li, N. Xiu, and S. Zhou, *Matrix optimization over low-rank spectral sets: Stationary points and local and global minimizers*, Journal of Optimization Theory and Applications, 184 (2020), pp. 895–930, https://doi.org/10.1007/s10957-019-01606-8.

[28] Z. Luo and L. Qi, *Optimality conditions for Tucker low-rank tensor optimization*, Computational Optimization and Applications, 86 (2023), pp. 1275–1298, https://doi.org/10.1007/s10589-023-00465-4.

[29] Z.-Q. Luo, J.-S. Pang, D. Ralph, and S.-Q. Wu, *Exact penalization and stationarity conditions of mathematical programs with equilibrium constraints*, Mathematical Programming, 75 (1996), pp. 19–76, https://doi.org/10.1007/BF02592205.

[30] J. Mather, *Notes on topological stability*, Bulletin of the American Mathematical Society, 49 (2012), pp. 475–506, https://doi.org/10.1090/S0273-0979-2012-01383-6.

[31] B. S. Mordukhovich, *Variational Analysis and Generalized Differentiation I*, vol. 330 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag Berlin Heidelberg, 2006, https://doi.org/10.1007/3-540-31247-1.

[32] B. S. Mordukhovich, *Variational Analysis and Applications*, Springer Monographs in Mathematics, Springer Cham, 2018, https://doi.org/10.1007/978-3-319-92775-6.

[33] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM Journal on Scientific and Statistical Computing, 4 (1983), pp. 553–572, https://doi.org/10.1137/0904038.

[34] Y. Nesterov, *Lectures on Convex Optimization*, vol. 137 of Springer Optimization and Its Applications, Springer, Cham, 2nd ed., 2018, https://doi.org/10.1007/978-3-319-91578-4.

[35] G. Olikier, K. A. Gallivan, and P.-A. Absil, *First-order optimization on stratified sets*, (2023), https://arxiv.org/abs/2303.16040.

[36] J.-S. Pang, *Partially B-regular optimization and equilibrium problems*, Mathematics of Operations Research, 32 (2007), pp. 687–699, https://doi.org/10.1287/moor.1070.0262.

[37] J.-S. Pang and M. Fukushima, *Complementarity constraint qualifications and simplified B-stationarity conditions for mathematical programs with equilibrium constraints*, Computational Optimization and Applications, 13 (1999), pp. 111–136, https://doi.org/10.1023/A:1008656806889.

[38] J.-S. Pang, M. Razaviyayn, and A. Alvarado, *Computing B-stationary points of nonsmooth DC programs*, Mathematics of Operations Research, 42 (2017), pp. 95–118, https://doi.org/10.1287/moor.2016.0795.

[39] E. Pauwels, *Generic Fréchet stationarity in constrained optimization*, (2024), https://arxiv.org/abs/2402.09831.

[40] E. Polak, *Computational Methods in Optimization*, vol. 77 of Mathematics in Science and Engineering, Academic Press, 1971.

[41] S. M. Robinson, *Nonlinear Analysis and Optimization*, vol. 30 of Mathematical Programming Studies, Springer Berlin Heidelberg, 1987, ch. Local structure of feasible sets in nonlinear programming, Part III: Stability and sensitivity, pp. 45–66, https://doi.org/10.1007/BFb0121154.

[42] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer-Verlag Berlin Heidelberg, 1998, https://doi.org/10.1007/978-3-642-02431-3. Corrected 3rd printing 2009.

[43] H. Scheel and S. Scholtes, *Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity*, Mathematics of Operations Research, 25 (2000), pp. 1–22, https://doi.org/10.1287/moor.25.1.1.15213.

[44] R. Schneider and A. Uschmajew, *Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality*, SIAM Journal on Optimization, 25 (2015), pp. 622–646, https://doi.org/10.1137/140957822.

[45] M. K. Tam, *Regularity properties of non-negative sparsity sets*, Journal of Mathematical Analysis and Applications, 447 (2017), pp. 758–777, https://doi.org/10.1016/j.jmaa.2016.10.040.

[46] P. P. Varaiya, *Nonlinear programming in Banach space*, SIAM Journal on Applied Mathematics, 15 (1967), pp. 284–293, https://doi.org/10.1137/0115028.

[47] M. Willem, *Functional Analysis: Fundamentals and Applications*, Cornerstones, Birkhäuser New York, 2013, https://doi.org/10.1007/978-1-4614-7004-5.

[48] J. Wu, L. Zhang, and Y. Zhang, *Mathematical programs with semidefinite cone complementarity constraints: Constraint qualifications and optimality conditions*, Set-Valued and Variational Analysis, 22 (2014), pp. 155–187, https://doi.org/10.1007/s11228-013-0242-7.

[49] J. J. Ye, *Necessary and sufficient optimality conditions for mathematical programs with equilibrium constraints*, Journal of Mathematical Analysis and Applications, 307 (2005), pp. 350–369, https://doi.org/10.1016/j.jmaa.2004.10.032.