PROJECTED GRADIENT DESCENT ACCUMULATES AT BOULIGAND STATIONARY POINTS*

GUILLAUME OLIKIER† AND IRÈNE WALDSPURGER‡

Abstract. This paper considers 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 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, criticality, tangent and normal cones, Clarke regularity

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

1. Introduction.

Let $\mathcal{E}$ be a Euclidean vector space, C a nonempty closed subset of $\mathcal{E}$, and $f : \mathcal{E} \to \mathbb{R}$ a function satisfying at least the first of the following two properties:

(H1) f is differentiable on C, i.e., for every $x \in C$, there exists a (unique) vector in $\mathcal{E}$, denoted by $\nabla f(x)$, such that

$$\lim_{\substack{y \to x \\ y \in \mathcal{E} \setminus \{x\}}} \frac{f(y) - f(x) - \langle \nabla f(x), y - x \rangle}{\|y - x\|} = 0,$$

and $\nabla f : C \to \mathcal{E}$ is continuous;

(H2) f is differentiable on $\mathcal{E}$ and $\nabla f : \mathcal{E} \to \mathcal{E}$ is locally Lipschitz continuous.

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, finding an exact or approximate global minimizer of problem (1.1) is intractable. Even finding an approximate local minimizer is not always feasible in polynomial time (unless P = NP) [2]. Therefore, algorithms are only expected to return a point satisfying a condition called *stationarity*, which is a tractable surrogate for local optimality.

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 defines a different notion of stationarity, which is a surrogate for local optimality in the sense 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 [53, Definition 6.3 and Example 6.16], namely normality in the general sense, in the regular sense, and in the

*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).

This manuscript is for review purposes only.

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* $-\nabla f(x) \in \widehat{N}_C(x)$*;*
- 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 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 [13, §5], for problem (1.1) under the only assumption that f is differentiable on C, B-stationarity is the strongest necessary condition for local optimality. The same is true if f satisfies (H1). Indeed, by [53, Theorem 6.11], for all $x \in C$,

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

$$= \left\{ -\nabla h(x) \;\middle|\; \begin{array}{l} h : \mathcal{E} \to \mathbb{R} \text{ satisfies (H1)}, \\ x \text{ is a local minimizer of } h|_C \end{array} \right\}. \tag{1.4}$$

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 satisfies (H1).

P-stationarity is the strongest necessary condition for local optimality if f satisfies (H2). Indeed, by Theorem 2.5, for all $x \in C$,

$$\widehat{\widehat{N}}_C(x) = \left\{ -\nabla h(x) \;\middle|\; \begin{array}{l} h : \mathcal{E} \to \mathbb{R} \text{ satisfies (H2)}, \\ x \text{ is a local minimizer of } h|_C \end{array} \right\}. \tag{1.5}$$

The inclusion $\supseteq$ in (1.5) shows that, under (H2), 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 unsatisfactory in [27, §4], [32, §1], and [50, §2.1]. Furthermore, as explained in [32], 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 [32]) in the sense that it cannot be done based on standard measures of

B-stationarity because of their possible lack of lower semicontinuity at points where the feasible set is not Clarke regular, as also explained in [46, §§1.1 and 2.4].

Projected gradient descent, or PGD for short, is a basic algorithm aiming at solving problem (1.1). To the best of our knowledge, the first article to have considered PGD on a possibly nonconvex closed set was [6]. The nonmonotone backtracking version considered in this paper is defined as Algorithm 4.2 and is based on [30, Algorithm 3.1] and [11, Algorithm 3.1]. Given $x \in C$ as input, the iteration map of PGD, called the PGD map and defined as Algorithm 4.1, performs a backtracking 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 the step size $\alpha \in (0, \infty)$ 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 settings proposed in [30, 31] and [11], the value at the current iterate can be replaced with the maximum value of f over a prefixed number of the previous iterates ("max" rule) or with a weighted average of the values of f at the previous iterates ("average" rule). This version of PGD is called *nonmonotone*. By [31, 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 [11, Theorem 4.6], the same result holds for nonmonotone PGD with the "average" rule and, by [31, Theorem 4.1], also for nonmonotone PGD with the "max" rule 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 [32, §1], it is an open question whether the accumulation points of PGD are always B-stationary for (1.1).

This paper answers positively the question by proving Theorem 1.2.

THEOREM 1.2. *Consider a sequence generated by* PGD *(Algorithm 4.2) when applied to problem* (1.1)*.*

- *If this sequence is finite, then its last element is B-stationary for* (1.1) *under (H1), and even P-stationary for* (1.1) *under (H2).*
- *If this sequence is infinite, then all of its accumulation points, if any, are B-stationary for* (1.1) *under (H1), and even P-stationary for* (1.1) *under (H2).*

If ∇f is globally Lipschitz continuous, then it is known in the literature that every local minimizer of $f|_C$ is P-stationary for (1.1) [61, Proposition 3.5(ii)] (the result is given for a global minimizer but the proof shows that it also holds for a local minimizer) and that PGD with a constant step size smaller than the inverse of the Lipschitz constant accumulates at P-stationary points of (1.1) [61, Theorem 5.6(i)]. Indeed, the ZeroFPR algorithm proposed in [61] extends the proximal gradient (PG) algorithm with a constant step size [61, Remark 5.5] which itself extends PGD with a constant step size; problem (1.1) corresponds to [61, problem (1.1)] with g the indicator function of our set C. These results were rediscovered in [50] where, in addition, the distance from the negative gradient of the continuously differentiable function to the regular subdifferential of the other function is proven to converge to zero along the generated sequence, and a quadratic lower bound on $f - f(\overline{x})$ at every accumulation point $\overline{x}$ of PG is obtained. The two results cited from [61] were already stated in [5, Theorems 2.2 and 3.1] for C the set $\mathbb{R}^n_{\leq s}$ of vectors of $\mathbb{R}^n$ having at most s nonzero components for some positive integer $s < n$, and in [3, Proposition 1 and

Theorem 1] for C satisfying a regularity condition called *proximal smoothness* which none of the four examples studied in Section 7 satisfies.

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 hypothesis (H1) in Section 5 and under hypothesis (H2) in Section 6. Four practical 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. Theorem 1.2 is illustrated by a comparison between PGD and a first-order algorithm that is not guaranteed to accumulate at B-stationary points of (1.1) in Section 8. Concluding remarks are gathered in Section 9.

2. Elements of variational analysis. This section, mostly based on [53], reviews background material in variational analysis that is used in the rest of the paper. Section 2.1 concerns the projection map 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.

Recall that, throughout the paper, $\mathcal{E}$ is a Euclidean vector space and $C \subseteq \mathcal{E}$ is nonempty and closed. Moreover, 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 [53, §3B], a nonempty subset K of $\mathcal{E}$ is called a *cone* if $x \in K$ implies $\alpha x \in K$ for all $\alpha \in [0, \infty)$.

2.1. Projection map. 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 [53, 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 [53, 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 [53, 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 [53, Proposition 6.2] called the *tangent cone* to C at x and denoted by $T_C(x)$. Following [53, Definition 6.3 and Proposition 6.5], the *regular normal cone* to C at $x \in C$ is

$$\widehat{N}_C(x) \coloneqq \{v \in \mathcal{E} \mid \langle v, w \rangle \leq 0 \; \forall w \in T_C(x)\}$$

which is a closed convex cone. Following [53, Definition 6.3], a vector $v \in \mathcal{E}$ is said to be *normal (in the general sense)* 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 [53, Proposition 6.5] called the *normal cone* to C at x and denoted by $N_C(x)$. Following [53, 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 [53, 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 Section 7. For those examples, however, regular normals are proximal normals (Proposition 7.1). An example of a set C and a point $x \in C$ such that both inclusions in (1.2) are strict is given in Example 2.2.

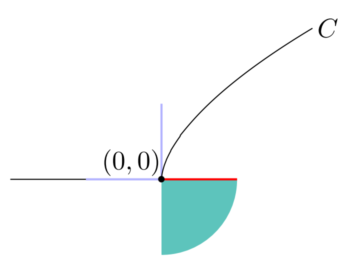
Example 2.2. *Let* $\mathcal{E} := \mathbb{R}^2$ *and* $C := \left\{\left(t, \max\left\{0, t^{3/5}\right\}\right) \mid t \in \mathbb{R}\right\}$ *(inspired by [53, Figure 6–12(a)]). Then,*

$$\begin{aligned}
T_C(0,0) &= (\{0\} \times [0,\infty)) \cup ((-\infty, 0] \times \{0\}), \\
\widehat{N}_C(0,0) &= [0,\infty) \times (-\infty, 0], \\
\widehat{\widehat{N}}_C(0,0) &= \widehat{N}_C(0,0) \setminus ((0,\infty) \times \{0\}), \\
N_C(0,0) &= \widehat{N}_C(0,0) \cup T_C(0,0).
\end{aligned}$$

Thus,

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

This is illustrated in Figure 1.



$$N_C(0,0) = \widehat{\widehat{N}}_C(0,0) \cup (\widehat{N}_C(0,0) \setminus \widehat{\widehat{N}}_C(0,0)) \cup T_C(0,0)$$

Fig. 1. *Tangent and normal cones from Example 2.2.*

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 [53, 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 satisfy (H1). In particular, under this assumption, P-stationarity is not necessary for local optimality, as illustrated by Example 2.3.

EXAMPLE 2.3. *Let* $\mathcal{E} := \mathbb{R}^2$, $C := \left\{(x_1, x_2) \in \mathbb{R}^2 \mid x_2 \geq \max\left\{0, x_1^{3/5}\right\}\right\}$ *[53, 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 on* $\mathcal{E}$*, hence on* C*, 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 assumed to satisfy (H2), that is, f is differentiable on $\mathcal{E}$ and ∇f is locally Lipschitz continuous. The latter means that, for every open or closed 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 [44, Lemma 1.2.3], that, for all $x, y \in \mathcal{B}$,

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

PROPOSITION 2.4. *Assume that* f *satisfies (H2). 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 strengthens [53, Proposition 8.46(d)] by stating that (1.5) is valid, which shows that P-stationarity is the strongest necessary condition for local optimality under hypothesis (H2).

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 $v \in \widehat{\widehat{N}}_C(x)$. By 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

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

The function 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

$$v \in \left\{ -\nabla h(x) \;\middle|\; \begin{array}{l} h : \mathcal{E} \to \mathbb{R} \text{ satisfies (H2)}, \\ x \text{ is a local minimizer of } h|_C \end{array} \right\},$$

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

Remark 2.6. From our proof, we see that (1.5) is also true if we replace "local minimizer" with "global minimizer". The same holds for equations (1.3)–(1.4) [53, Theorem 6.11]. However, in this section, we are interested in understanding the closeness between the notions of stationarity and *local* optimality.

3. Stationarity in the literature.

This section surveys the names given to the stationarity notions provided in Definition 1.1 and attempts to offer a brief historical perspective. The terms "B-stationarity" and "M-stationarity" first appeared in the literature about mathematical programs with equilibrium constraints (MPECs), as explained in Sections 3.1 and 3.2. In contrast, the term "P-stationarity" seems to be new in the literature. P-stationarity is called "criticality" in [61, Definition 3.1(ii)]; problem (1.1) corresponds to [61, problem (1.1)] with g the indicator function of our set C. We propose the name "P-stationarity" because this stationarity notion is based on the proximal normal cone. It is closely related to the so-called α-stationarity, as explained in Section 3.3.

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 [12, 13].

B-stationarity appears as a necessary condition for optimality in [62, Theorem 2.1] and [23, 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 [53, Corollary 6.21]. Neither "stationary" nor "critical" appears in [62] or [23].

The "Bouligand derivative", or "B-derivative" for short, was introduced in [52]. 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., [53, 41, 42] for recent references. Differentiability implies B-differentiability.

In [58, §4], a point where a real-valued function is B-differentiable is called a "Bouligand stationary (B-stationary) point" of the function if the B-derivative at that point is nonnegative. This is a stationarity concept for unconstrained optimization, which therefore does not apply to problem (1.1).

B-stationarity is called a "stationarity condition" and said to be "well known" in [38, §4.1] where [23] is cited.

In [54, §2.1], the term "B-stationarity" is used to name the stationarity concept for an MPEC that corresponds to the B-stationarity in the sense of [58, §4] for a nonsmooth reformulation of the MPEC [54, Proposition 6]. As pointed out in [65, §2.1] and [14, §3.3], the "B-stationarity" in the sense of [54, §2.1], which is specific to MPECs, is not B-stationarity in the sense of Definition 1.1 and is called "MPEC-linearized B-stationarity" in [14, §3.3] to avoid confusion. Nevertheless, this MPEC-linearized B-stationarity appears under the name "B-stationarity" in [28, §1.1], [24, Definition 2.2], and [64, Definition 3.2] which all cite [54].

The term "B-stationarity" was used to name the absence of descent directions in the tangent cone (as in Definition 1.1) first in [48, §1]. It was used in this sense in several subsequent works by various authors; see, e.g., [18, §2], [19, §2], [17, Definition 2.4], [65, Definition 2.2], [14, §§3.3 and 4], [15, §3], [47, §2], [59, Definition 2.4], [21, Definition 3.4], [49, (18)], [7, Definition 3(1)], [8, Definition 4(i)], [27, §4], and [9, Definition 6.1.1].

In [9, Definition 6.1.1], B-stationarity is defined for the problem of minimizing a real-valued function that is B-differentiable on a nonempty closed subset of a Euclidean vector space, thereby extending the concept introduced in [58, §4] to constrained optimization. This more general definition reduces to that from Definition 1.1 if the function is differentiable.

B-stationarity is also known under other names in the literature. First, in [16, Definition 1(b)] and [40, §3], B-stationarity is called "strong stationarity"; [16, problem (4)] and [40, (P2)] reduce to problem (1.1) for F the identity map on $\mathbb{R}^n$. Second, because the regular normal cone is also called the Fréchet normal cone, especially in infinite-dimensional spaces [53, 41, 42], B-stationarity is called "Fréchet stationarity", or "F-stationarity" for short, in [35, Definition 4.1(ii)], [36, Definition 5.1(i)], and [37, Definition 3.2(ii)]. Third, B-stationarity is simply called "stationarity" (or "criticality") in [55, §2.1], [25, §2.1.1], [32, Definition 2.3], [33, Definition 3.2(c)], and [20, Definition 1].

3.2. A brief history of Mordukhovich stationarity. According to [16, §2], the term "M-stationarity" was introduced in [56] for an MPEC. This name was chosen because the corresponding stationarity condition was derived from the generalized differential calculus of Mordukhovich. To the best of our knowledge, the term "M-stationarity" was used to indicate that the negative gradient is in the normal cone (as in Definition 1.1) first in [16, Definition 1(a)]; recall that [16, problem (4)] reduces to problem (1.1) for F the identity map on $\mathbb{R}^n$. There, the name is motivated by the presence of the normal cone which was introduced by Mordukhovich. M-stationarity appears, under this name, in several subsequent works by various authors; see, e.g., [7, Definition 3(3)], [8, Definition 4(iii)], [27, §4], [40, §3], [31, §2], [30, §3], and [29, §2.3].

3.3. Proximal stationarity and α-stationarity. P-stationarity is related to α-stationarity which was introduced in [5, Definition 2.3] for $C = \mathbb{R}^n_{\leq s}$ and in [35, Definition 4.1(i)], [25, §2.1.1], [36, Definition 5.1(ii)], [34, (4.2)], and [37, 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 [32, Definition 2.6].

4. The PGD algorithm.

This section reviews the PGD algorithm, as defined in [30, Algorithm 3.1] except that the "average" rule is allowed as an alternative to the "max" rule. Its iteration map, called the PGD map, is defined as Algorithm 4.1. PGD is defined as Algorithm 4.2 which uses Algorithm 4.1 as a subroutine. The nonmonotonic behavior of PGD is described in Propositions 4.6 and 4.7.

Algorithm 4.1 PGD map

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: $y \in \mathrm{PGD}(x, \mu; \mathcal{E}, C, f, \underline{\alpha}, \overline{\alpha}, \beta, c)$.

Choose $\alpha \in [\underline{\alpha}, \overline{\alpha}]$ and $y \in P_C(x - \alpha\nabla f(x))$;
while $f(y) > \mu + c\,\langle \nabla f(x), y - x\rangle$ **do**
  $\alpha \leftarrow \alpha\beta$;
  Choose $y \in P_C(x - \alpha\nabla f(x))$;
end while
Return y.

Remark 4.1. 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 along 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 [31, Algorithms 3.1 and 4.1] and [11, Algorithm 3.1]. Importantly, all results from [31] hold for both conditions, as is clear from the proofs.

Remark 4.2. By Proposition 5.3, if f satisfies (H1) and x is not B-stationary for (1.1), then the while loop in Algorithm 4.1 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 satisfies (H2), then the while loop is guaranteed to terminate for every $x \in C$, by Corollary 6.2.

The PGD algorithm is defined as Algorithm 4.2. It is said to be monotone or nonmonotone depending on whether $\mu_i = f(x_i)$ for all i (that is, $l = 0$ for the "max" rule, or $p = 1$ for the "average" rule) or not.

Algorithm 4.2 PGD

Require: $(\mathcal{E}, C, f, \underline{\alpha}, \overline{\alpha}, \beta, c,$ "nonmonotonicity"$)$ 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 "nonmonotonicity" $\in \{(\text{"max"}, l), (\text{"average"}, p)\}$ with $l \in \mathbb{N}$ and $p \in (0, 1]$.

Input: $x_0 \in C$.

Output: a sequence in C.

1: $i \leftarrow 0$;
2: $\mu_{-1} \leftarrow f(x_0)$;
3: **while** $-\nabla f(x_i) \notin \widehat{N}_C(x_i)$ **do**
4:     **if** "nonmonotonicity" $= (\text{"max"}, l)$ **then**
5:         $\mu_i \leftarrow \max_{j \in \{\max\{0, i-l\}, \dots, i\}} f(x_j)$;
6:     **else if** "nonmonotonicity" $= (\text{"average"}, p)$ **then**
7:         $\mu_i \leftarrow (1 - p)\mu_{i-1} + p f(x_i)$;
8:     **end if**
9:     Choose $x_{i+1} \in \mathrm{PGD}(x_i, \mu_i; \mathcal{E}, C, f, \underline{\alpha}, \overline{\alpha}, \beta, c)$;
10:     $i \leftarrow i + 1$;
11: **end while**

Remark 4.3. For simplicity, we use a constant weight p in the "average" rule. However, we could allow the weight to change from one iteration to the other. It would then be denoted by p_i. The main results of the article would hold true in this more general setting, under the additional assumption that $\inf_{i \in \mathbb{N}} p_i > 0$.

Remark 4.4. If f satisfies (H2), then $\widehat{N}_C(x_i)$ should be replaced with $\widehat{\widehat{N}}_C(x_i)$ in line 3.

Examples of a set C for which the projection map and the regular and proximal normal cones can be described explicitly abound in the literature; see Section 7. For such examples, Algorithm 4.2 can be practically implemented.

Remark 4.5. From Remark 4.2, under (H1), the call to Algorithm 4.1 in line 9 of PGD never results in an infinite loop. Consequently, by running PGD, one always encounters one of the following two situations:

- PGD generates a finite sequence, and the last element of this sequence is B-stationary for (1.1) if f satisfies (H1), and even P-stationary for (1.1) if f satisfies (H2);
- PGD generates an infinite sequence.

The rest of this section and the next two concern the nontrivial case where PGD 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 [51, 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 imposed. As a reminder, a sequence $(x_i)_{i \in \mathbb{N}}$ in $\mathcal{E}$ has at least one accumulation point if and only if $\liminf_{i \to \infty} \|x_i\| < \infty$.

A property of monotone PGD that is helpful for its analysis is the fact that f is strictly decreasing along the generated sequence. For nonmonotone PGD, this is not true. However, weaker properties, stated in the following two propositions, will be enough for our purposes.

PROPOSITION 4.6. *Let $(x_i)_{i\in\mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2) using the "max" rule. 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 monotonically 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 [31, Lemma 4.1 and Corollary 4.1]. For the third one, 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 monotonically nonincreasing, 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 $p_j := g(i_j + l) - i_j \in \{0,\ldots,l\}$. Then, there exist $p \in \{0,\ldots,l\}$ and a subsequence $(p_{j_k})_{k\in\mathbb{N}}$ such that, for all $k \in \mathbb{N}$, $p_{j_k} = p$. By [31, (27)] or [30, (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 $k \in \mathbb{N}$, $f(x_{g(i_{j_k}+l)-p}) = f(x_{i_{j_k}})$. ∎

PROPOSITION 4.7. *Let $(x_i)_{i\in\mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2) using the "average" rule. Then:*

1. *$(x_i)_{i\in\mathbb{N}}$ is contained in the sublevel set* (1.6)*;*
2. *if $(x_i)_{i\in\mathbb{N}}$ has an accumulation point, then $(f(x_i))_{i\in\mathbb{N}}$ and $(\mu_i)_{i\in\mathbb{N}}$ converge, toward the same (finite) value.*

Proof. The sequence $(\mu_i)_{i\in\mathbb{N}}$ is monotonically nonincreasing since, for all $i \in \mathbb{N}$, $f(x_i) \leq \mu_{i-1}$, hence $\mu_i = (1-p)\mu_{i-1} + pf(x_i) \leq \mu_{i-1}$. Therefore, for all $i \in \mathbb{N}$,

$$f(x_i) \leq \mu_{i-1} \leq \mu_{-1} = f(x_0),$$

meaning that $(x_i)_{i\in\mathbb{N}}$ is contained in the sublevel set (1.6).

Now, we prove the second item of the proposition. Let us assume that $(x_i)_{i\in\mathbb{N}}$ has an accumulation point x. Let $(x_{i_k})_{k\in\mathbb{N}}$ be a subsequence converging to x. Observe that

$$\lim_{k\to\infty} f(x_{i_k}) = f(x),$$

since f is differentiable, and in particular continuous, at x. As $(\mu_i)_{i\in\mathbb{N}}$ is monotonically nonincreasing, it has a limit $\varphi \in \mathbb{R} \cup \{-\infty\}$. For all $k \in \mathbb{N}$,

$$f(x_{i_k}) \leq \mu_{i_k-1}.$$

Letting k tend to infinity yields

$$f(x) = \lim_{k\to\infty} f(x_{i_k}) \leq \lim_{k\to\infty} \mu_{i_k-1} = \varphi.$$

This manuscript is for review purposes only.

In particular, φ is finite.

Now, we show that $\varphi = \liminf_{i\to\infty} f(x_i)$. Let $(x_{j_k})_{k\in\mathbb{N}}$ be a subsequence such that

$$\lim_{k\to\infty} f(x_{j_k}) = \liminf_{i\to\infty} f(x_i).$$

For all $k \in \mathbb{N}$, it holds that

$$\mu_{j_k} = (1-p)\mu_{j_k-1} + pf(x_{j_k}).$$

The two sides of this equality must have the same limit:

$$\varphi = (1-p)\varphi + p\liminf_{i\to\infty} f(x_i).$$

As $p > 0$, this implies $\varphi = \liminf_{i\to\infty} f(x_i)$ (and, in particular, $\liminf_{i\to\infty} f(x_i) > -\infty$). To conclude, we observe that, for all $k \in \mathbb{N}$,

$$f(x_k) \le \mu_{k-1}.$$

Hence,

$$\limsup_{k\to\infty} f(x_k) \le \lim_{k\to\infty} \mu_{k-1} = \varphi = \liminf_{k\to\infty} f(x_k).$$

Therefore, $(f(x_k))_{k\in\mathbb{N}}$ converges to φ. ∎

5. Convergence analysis for a continuous gradient. In this section, PGD (Algorithm 4.2) is analyzed under hypothesis (H1). As mentioned after Remark 4.5, only the nontrivial case where an infinite sequence is generated is considered here. Specifically, the first part of the second item 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 satisfies (H1), 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 every 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: roughly, 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. The intuition for this proof is that, for every x close to $\underline{x}$ and for every $y \in P_C(x - \alpha\nabla f(x))$,

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \text{some remainder}.$$

The inner product $\langle \nabla f(x), y - x\rangle$ is negative, and larger in absolute value than some fraction of $\|\nabla f(x)\|\|y-x\|$ (Proposition 5.2). On the other hand, if α is small enough, the remainder (upper bounded in Proposition 5.3) is smaller than some arbitrarily small fraction of $\|\nabla f(x)\|\|y - x\|$. Therefore, for α small enough,

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

PROPOSITION 5.2. *Assume that f satisfies (H1). 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 $\kappa := \sqrt{1 - \frac{\beta\langle w, -\nabla f(\underline{x})\rangle^2}{8\|w\|^2\|\nabla f(\underline{x})\|^2}} \in (0, 1)$. For every $\varepsilon \in (0,\infty)$, there exist $\alpha_{\underline{x}} \in (0, \varepsilon]$ and $\bar{\rho}(\alpha_{\underline{x}}) \in (0,\infty)$ such that, for all $x \in B(\underline{x}, \bar{\rho}(\alpha_{\underline{x}})) \cap C$ and $\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$,

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

which implies, for all $y \in P_C(x - \alpha\nabla f(x))$,

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

Proof. Let $\varepsilon \in (0,\infty)$ be fixed. We show that there exist $\alpha_{\underline{x}} \in (0, \varepsilon]$ and $\bar{\rho}(\alpha_{\underline{x}}) \in (0,\infty)$ 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

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

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 $\frac{1}{t_i}\frac{\|w_i-\underline{x}\|^2}{\langle w_i-\underline{x},-\nabla f(\underline{x})\rangle} \xrightarrow{i\to\infty} \frac{\|w\|^2}{\langle w,-\nabla f(\underline{x})\rangle}$ and $t_i \xrightarrow{i\to\infty} 0$, 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 $\bar{\rho}(\alpha_{\underline{x}})$, it holds for all $x \in B(\underline{x}, \bar{\rho}(\alpha_{\underline{x}})) \cap C$ 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 $\bar{\rho}(\alpha_{\underline{x}}) \in (0, \rho_0]$. For all $x \in B(\underline{x}, \bar{\rho}(\alpha_{\underline{x}})) \cap C$, $\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$, and $y \in P_C(x - \alpha\nabla f(x))$,

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

where the last inequality follows from (5.5b) and the fact that $\alpha_{\underline{x}} \leq \alpha \leq \frac{\alpha_{\underline{x}}}{\beta}$. Choose $\bar{\rho}(\alpha_{\underline{x}}) \in (0, \rho_0]$ small enough to ensure

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

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

$$\begin{aligned}\|x-\alpha\nabla f(x)-y\|^2 &\leq \alpha^2\|\nabla f(x)\|^2 - \frac{\alpha}{2}\langle w_i-\underline{x}, -\nabla f(x)\rangle \\ &= \alpha^2\|\nabla f(x)\|^2\left(1-\frac{\langle w_i-\underline{x}, -\nabla f(x)\rangle}{2\alpha\|\nabla f(x)\|^2}\right) \\ &\leq \alpha^2\|\nabla f(x)\|^2\left(1-\frac{\beta\langle w_i-\underline{x}, -\nabla f(x)\rangle}{2\alpha_{\underline{x}}\|\nabla f(x)\|^2}\right) \text{ as } \alpha \leq \frac{\alpha_{\underline{x}}}{\beta} \\ &\leq \alpha^2\|\nabla f(x)\|^2\left(1-\frac{\beta\langle w_i-\underline{x}, -\nabla f(x)\rangle^2}{4\|w_i-\underline{x}\|^2\|\nabla f(x)\|^2}\right) \text{ from (5.5b)} \\ &\leq \alpha^2\|\nabla f(x)\|^2\left(1-\frac{\beta\langle w, -\nabla f(\underline{x})\rangle^2}{8\|w\|^2\|\nabla f(\underline{x})\|^2}\right) \text{ from (5.5c)} \\ &= \kappa^2\alpha^2\|\nabla f(x)\|^2.\end{aligned}$$

In other words, for all $x \in B(\underline{x}, \bar{\rho}(\alpha_{\underline{x}})) \cap C$, $\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$, 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. *Let $\underline{\alpha} \in (0,\infty)$ and $c \in (0,1)$. Assume that f satisfies (H1). Let $\underline{x} \in C$ be non-B-stationary for* (1.1)*. There exists $\alpha_{\underline{x}} \in (0, \underline{\alpha}]$ and $\rho \in (0,\infty)$ such that, for all $x \in B(\underline{x}, \rho) \cap C$, $\alpha \in [\alpha_{\underline{x}}, \alpha_{\underline{x}}/\beta]$, and $y \in P_C(x-\alpha\nabla f(x))$,*

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

Proof. Define κ as in Proposition 5.2. Let $\delta \in (0,\infty)$ be small enough to ensure

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

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

These inequalities are satisfied by all δ 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.

Then, define $\varepsilon \coloneqq \min\{\underline{\alpha}, \delta\}$ and let $\alpha_{\underline{x}} \in (0, \varepsilon]$ and $\bar{\rho}(\alpha_{\underline{x}}) \in (0, \infty)$ be as in Proposition 5.2. Define

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

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

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

so that from (5.7b), $\|\nabla f(x) - \nabla f(\underline{x})\| < \frac{\|\nabla f(\underline{x})\|}{4}$, which implies

$$\begin{aligned}
\frac{3}{4}\|\nabla f(\underline{x})\| &< \|\nabla f(\underline{x})\| - \|\nabla f(x) - \nabla f(\underline{x})\| \\
&\leq \|\nabla f(x)\| \\
&\leq \|\nabla f(\underline{x})\| + \|\nabla f(x) - \nabla f(\underline{x})\| \\
&< \frac{5}{4}\|\nabla f(\underline{x})\|. \qquad (5.8)
\end{aligned}$$

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

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

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

$$\begin{aligned}
\|y - \underline{x}\| &\leq \|y - x\| + \|x - \underline{x}\| \\
&\leq 2\alpha\|\nabla f(x)\| + \rho \text{ from (2.1)} \\
&\leq \frac{2\alpha_{\underline{x}}}{\beta}\|\nabla f(x)\| + \alpha_{\underline{x}}\|\nabla f(\underline{x})\| \\
&< \frac{5\alpha_{\underline{x}}}{2\beta}\|\nabla f(\underline{x})\| + \alpha_{\underline{x}}\|\nabla f(\underline{x})\| \text{ from (5.8)} \\
&= \frac{7\alpha_{\underline{x}}}{2\beta}\|\nabla f(\underline{x})\|.
\end{aligned}$$

We continue from (5.9):

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

Inequality (a) is true because

$$
\begin{aligned}
\|y-x\| &\ge \alpha\|\nabla f(x)\| - \|x - \alpha\nabla f(x) - y\| \text{ by the triangle inequality}\\
&= \alpha\|\nabla f(x)\| - d(x-\alpha\nabla f(x), C)\\
&\ge (1-\kappa)\alpha\|\nabla f(x)\| \text{ from Proposition 5.2}\\
&\ge \frac{3}{4}(1-\kappa)\rho \text{ from (5.8), the definition of } \rho, \text{ and } \alpha_{\underline{x}} \le \alpha\\
&> \frac{3}{4}(1-\kappa)\|x-\underline{x}\|.
\end{aligned}
$$

□

5.2. Second part: convergence of successive iterates.

PROPOSITION 5.4. *Assume that f satisfies (H1). Let $(x_i)_{i\in\mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2), and x be an accumulation point. Then, for every 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.

If the nonmonotonicity rule is set to "average", this is a direct consequence of Proposition 4.7. Indeed, for all $i \in \mathbb{N}$, from (4.1),

$$f(x_{i+1}) \le \mu_i - \frac{c}{2\overline{\alpha}}\|x_{i+1}-x_i\|^2 \le \mu_i.$$

From Proposition 4.7, $(f(x_{i+1}))_{i\in\mathbb{N}}$ and $(\mu_i)_{i\in\mathbb{N}}$ converge to the same limit. Therefore,

$$\left(\mu_i - \frac{c}{2\overline{\alpha}}\|x_{i+1}-x_i\|^2\right)_{i\in\mathbb{N}}$$

also converges to this limit. This implies that $(\|x_{i+1}-x_i\|)_{i\in\mathbb{N}}$ converges to 0, hence $(\|x_{i_k+1}-x_{i_k}\|)_{k\in\mathbb{N}}$ converges to 0, and $(x_{i_k+1})_{k\in\mathbb{N}}$ converges to the same limit as $(x_{i_k})_{k\in\mathbb{N}}$, that is, x.

Now, let us consider the "max" rule case. 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 4.6. By the third statement of Proposition 4.6, 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 all $k \in \mathbb{N}$, 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, $x_{i_k} \in B(\underline{x}, \rho) \cap C$. Thus, 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 all $k \in \mathbb{N}$, 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 every 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 [43, Lemma 4.10].

6. Convergence analysis for a locally Lipschitz continuous gradient. In this section, PGD (Algorithm 4.2) is analyzed under hypothesis (H2). As mentioned after Remark 4.5, only the nontrivial case where an infinite sequence is generated is considered here. Specifically, the second part of the second item of Theorem 1.2, restated in 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 satisfies (H2). Let $\underline{x} \in C$, $\overline{\alpha} \in (0, \infty)$, $c \in (0, 1)$, and $\rho \in (0, \infty)$. Let $\bar{\rho} \in \left[\rho + 2\overline{\alpha} \max_{x \in B[\underline{x},\rho] \cap C} \|\nabla f(x)\|, \infty\right)$ and define $\alpha_* := (1-c)/\operatorname{Lip}_{B[\underline{x},\bar{\rho}]}(\nabla f)$. Then, 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. For all $x \in B[\underline{x}, \rho] \cap C$ and $\alpha \in [0, \overline{\alpha}]$, $P_C(x - \alpha \nabla f(x)) \subseteq B[\underline{x}, \bar{\rho}]$; indeed, for all $y \in P_C(x - \alpha \nabla f(x))$,

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

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

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

∎

COROLLARY 6.2. *Consider Algorithm 4.1 under hypothesis (H2). Given $\underline{x} \in C$ and $\rho \in (0, \infty)$, let $\bar{\rho}$ 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 $\alpha \in \left[\min\left\{\underline{\alpha}, \frac{\beta(1-c)}{\operatorname{Lip}_{B[\underline{x},\bar{\rho}]}(\nabla f)}\right\}, \overline{\alpha}\right]$ and hence after at most*

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

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

Proof. At the latest, the while loop ends after iteration $i \in \mathbb{N} \setminus \{0\}$ with $\alpha = \alpha_0 \beta^i$ such that $\frac{\alpha}{\beta} > \frac{1-c}{\mathrm{Lip}_{B[\underline{x},\bar{\rho}]}(\nabla f)}$. In that case, $i < 1 + \ln(\frac{1-c}{\alpha_0 \mathrm{Lip}_{B[\underline{x},\bar{\rho}]}(\nabla f)})/\ln(\beta)$ and thus $i \le \lceil \ln(\frac{1-c}{\alpha_0 \mathrm{Lip}_{B[\underline{x},\bar{\rho}]}(\nabla f)})/\ln(\beta) \rceil$. ∎

THEOREM 6.3. *Assume that f satisfies (H2). Let $(x_i)_{i\in\mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2). Then, all accumulation points of $(x_i)_{i\in\mathbb{N}}$ are P-stationary for* (1.1). *Moreover, for every convergent subsequence $(x_{i_j})_{j\in\mathbb{N}}$,*

$$\lim_{j\to\infty} d(-\nabla f(x_{i_j+1}), \hat{\hat{N}}_C(x_{i_j+1})). \tag{6.1}$$

Proof. Assume that a subsequence $(x_{i_j})_{j\in\mathbb{N}}$ converges to $\underline{x} \in C$. Given $\rho \in (0,\infty)$, let $\bar{\rho}$ be as in Proposition 6.1. Define

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

There exists $j_* \in \mathbb{N}$ such that, for all integers $j \ge j_*$, $x_{i_j} \in B[\underline{x}, \rho]$, 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 $j_{k_*} \ge j_*$. Furthermore, by Proposition 5.4, $(x_{i_j+1})_{j\in\mathbb{N}}$ converges to $\underline{x}$. Therefore, for all integers $k \ge k_*$,

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

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 $-\nabla f(\underline{x}) \in \hat{\hat{N}}_C(\underline{x})$.

We now establish (6.1). Recall that, for all integers $j \ge j_*$, since $x_{i_j+1} \in P_C(x_{i_j} - \alpha_{i_j} \nabla f(x_{i_j}))$ with $\alpha_{i_j} \in I$, it holds that $\frac{1}{\alpha_{i_j}}(x_{i_j} - x_{i_j+1}) - \nabla f(x_{i_j}) \in \hat{\hat{N}}_C(x_{i_j+1})$, and thus

$$\begin{aligned}
d(-\nabla f(x_{i_j+1}), \hat{\hat{N}}_C(x_{i_j+1})) &\le \| - \nabla f(x_{i_j+1}) - (\frac{1}{\alpha_{i_j}}(x_{i_j} - x_{i_j+1}) - \nabla f(x_{i_j}))\| \\
&\le \frac{1}{\alpha_{i_j}} \|x_{i_j+1} - x_{i_j}\| + \|\nabla f(x_{i_j+1}) - \nabla f(x_{i_j})\| \\
&\to 0 \text{ when } j \to \infty,
\end{aligned}$$

by Proposition 5.4 and the fact that $(\alpha_{i_j})_{j\in\mathbb{N}}$ is bounded away from zero. ∎

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

PROPOSITION 6.4. *Assume that f satisfies (H2). Let $(x_i)_{i\in\mathbb{N}}$ be a sequence generated by* PGD *(Algorithm 4.2). If $(x_i)_{i\in\mathbb{N}}$ is bounded, 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, and*

$$\lim_{i\to\infty} d(-\nabla f(x_i), \hat{\hat{N}}_C(x_i)) = 0. \tag{6.2}$$

Proof. Assume that $(x_i)_{i\in\mathbb{N}}$ is bounded. It suffices to establish (6.2) and to prove that all accumulation points of $(x_i)_{i\in\mathbb{N}}$ have the same image by f; the other statements follow from Theorem 6.3.

The proof that all accumulation points of $(x_i)_{i\in\mathbb{N}}$ have the same image by f is based on the argument given in the proof of [51, 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 Propositions 4.6 and 4.7, the sequence $(f(x_i))_{i\in\mathbb{N}}$ converges; Proposition 4.6 applies because a continuous real-valued function is bounded from below and uniformly continuous on every compact set [63, Propositions 1.3.3 and 1.3.5]. Therefore, $f(\underline{x}) = \lim_{k\to\infty} f(x_{i_k}) = \lim_{i\to\infty} f(x_i) = \lim_{k\to\infty} f(x_{j_k}) = f(\overline{x})$.

Let us establish (6.2). Assume, for the sake of contradiction, that (6.2) does not hold. Then, there exist $\varepsilon \in (0,\infty)$ and a subsequence $(x_{i_j})_{j\in\mathbb{N}}$ such that $i_0 \geq 1$ and $d(-\nabla f(x_{i_j}), \widehat{\widehat{N}}_C(x_{i_j})) > \varepsilon$ for all $j \in \mathbb{N}$. Since $(x_{i_j-1})_{j\in\mathbb{N}}$ is bounded, it contains a subsequence $(x_{i_{j_k}-1})_{k\in\mathbb{N}}$ that converges to a point $\underline{x} \in C$. Therefore, by (6.1),

$$\lim_{k\to\infty} d(-\nabla f(x_{i_{j_k}}), \widehat{\widehat{N}}_C(x_{i_{j_k}})) = 0,$$

a contradiction. ∎

7. Examples of feasible sets on which PGD can be practically implemented.

Examples of a set C on which PGD can be practically implemented include:

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 [26, Lecture 9]
$$\mathbb{R}^{m\times n}_{\leq r} := \{X \in \mathbb{R}^{m\times n} \mid \operatorname{rank} X \leq r\},$$
m, n, and r being positive integers such that $r < \min\{m, n\}$;
4. the closed cone
$$\mathrm{S}^+_{\leq r}(n) := \{X \in \mathbb{R}^{n\times n}_{\leq r} \mid X^\top = X,\, X \succeq 0\}$$
of order-n real symmetric positive-semidefinite matrices of rank at most r, n and r being positive integers such that $r < n$.

Indeed, for every set in this list, the projection map, the tangent cone, the regular normal cone, and the normal cone are explicitly known; see [45, §§6 and 7.4] and the references therein. In particular, it is known that these sets are not Clarke regular at infinitely many points. In this section, we prove that, for these sets, regular normals are proximal normals.

As detailed in [45], if C is a set in this list, then there exist a positive integer p and disjoint nonempty smooth submanifolds $S_0, \dots, S_p$ of $\mathcal{E}$ such that $\overline{S_p} = C$ and, for all $i \in \{0, \dots, p\}$, $\overline{S_i} = \bigcup_{j=0}^{i} S_j$. This implies that $\{S_0, \dots, S_p\}$ is a *stratification* of C satisfying the *condition of the frontier* [39, §5]. Thus, C is called a *stratified set* and $S_0, \dots, S_p$ are called the *strata* of $\{S_0, \dots, S_p\}$.

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

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

Since the proof of Proposition 7.1 relies on significantly different concepts than those previously used, we present it in Appendix A.

8. Comparison of PGD and P^2GD on a simple example. P^2GD, which is short for projected-projected gradient descent, was introduced in [55, Algorithm 3] for $C := \mathbb{R}_{\leq r}^{m \times n}$ and extended to an arbitrary set C in [45, Algorithm 5.1]. It works like PGD except that it involves an additional projection: given $x \in C$ as input, the P^2GD map [45, Algorithm 5.1] performs a backtracking projected line search along a projection g of $-\nabla f(x)$ onto $T_C(x)$, i.e., computes a projection y of $x + \alpha g$ onto C for decreasing values of the step size $\alpha \in (0, \infty)$ until y satisfies an Armijo condition.

As pointed out in [57, §3.2], the convergence of optimization algorithms that use descent directions in the tangent cone, such as P^2GD, often suffers from the discontinuity of the tangent cone. In [32, §2.2], on an instance of (1.1) where $\mathcal{E} := \mathbb{R}^{3\times 3}$ and $C := \mathbb{R}_{\leq 2}^{3\times 3}$, P^2GD is proven to generate a sequence converging to a point of rank one that is M-stationary but not B-stationary. Several methods are compared numerically on this instance in [46, §8.2].

In this section, monotone PGD and P^2GD are compared analytically on the instance of (1.1) where $\mathcal{E} := \mathbb{R}^2$, $C := \mathbb{R}_{\leq 1}^2$, $f(x) := \frac{1}{2}\|x - x_*\|^2$ for all $x \in \mathbb{R}^2$, $x_* := (a, 0)$, and $a \in \mathbb{R} \setminus \{0\}$. For all $x \in \mathbb{R}^2$, $\nabla f(x) = x - x_*$. Thus, the global Lipschitz constant of ∇f is 1; in particular, f satisfies (H2). Both algorithms are used with $\underline{\alpha} := \overline{\alpha} := \alpha \in (0, 2)$ and an arbitrary $\beta \in (0, 1)$. The initial iterate is $(0, b)$ for some $b \in \mathbb{R} \setminus \{0\}$.

We recall from [45, Proposition 7.13] that $T_{\mathbb{R}_{\leq 1}^2}(0, 0) = \mathbb{R}_{\leq 1}^2$ and, for all $t \in \mathbb{R}\setminus\{0\}$,

$$T_{\mathbb{R}_{\leq 1}^2}(0, t) = \{0\} \times \mathbb{R}, \qquad T_{\mathbb{R}_{\leq 1}^2}(t, 0) = \mathbb{R} \times \{0\},$$

from [45, Propositions 7.16 and 7.17] that

$$\widehat{N}_{\mathbb{R}_{\leq 1}^2}(0, 0) = \{(0, 0)\} \subsetneq \mathbb{R}_{\leq 1}^2 = N_{\mathbb{R}_{\leq 1}^2}(0, 0)$$

and, for all $t \in \mathbb{R} \setminus \{0\}$,

$$\widehat{N}_{\mathbb{R}_{\leq 1}^2}(0, t) = \mathbb{R} \times \{0\}, \qquad \widehat{N}_{\mathbb{R}_{\leq 1}^2}(t, 0) = \{0\} \times \mathbb{R},$$

and from Proposition 7.1 that $\widehat{\widehat{N}}_{\mathbb{R}_{\leq 1}^2}(x) = \widehat{N}_{\mathbb{R}_{\leq 1}^2}(x)$ for all $x \in \mathbb{R}_{\leq 1}^2$.

Proposition 8.1 explicitly describes the sequences generated by PGD and P^2GD for small values of c. We omit its proof, which consists in elementary computations.

PROPOSITION 8.1. *If $\alpha = 1$ and $c \in (0, \frac{1}{2}]$, then* PGD *and* P^2GD *generate the finite sequences $((0, b), (a, 0))$ and $((0, b), (0, 0), (a, 0))$, respectively. If $\alpha \neq 1$, then both algorithms generate infinite sequences.*

- *For every $c \in (0, \frac{2-\alpha}{2}]$,* P^2GD *generates the sequence $((0, (1-\alpha)^i b))_{i \in \mathbb{N}}$ which converges to $(0, 0)$.*
- *For every $c \in (0, \frac{2-\alpha}{4})$:*
  - *if $\alpha|a|/|b| > |1 - \alpha|$, then* PGD *generates the sequence*

$$((0, b), (a(1 - (1 - \alpha)^{i+1}, 0))_{i \in \mathbb{N}});$$

  - *if $\alpha|a|/|b| \leq |1-\alpha|$, then $i_* := \left\lfloor \frac{\ln(\alpha|a|/|b|)}{\ln(|1-\alpha|)} \right\rfloor \in \mathbb{N}\setminus\{0\}$ and* PGD *generates the sequence*

$$(((0, (1 - \alpha)^i b))_{i=0}^{i=i_*}, (a(1 - (1 - \alpha)^{i+1}, 0))_{i \in \mathbb{N}})$$

$$\text{if } \tfrac{\ln(\alpha|a|/|b|)}{\ln(|1-\alpha|)} \notin \mathbb{N} \text{ and}$$

$$(((0,(1-\alpha)^i b))_{i=0}^{i=i_*}, (a(1-(1-\alpha)^{i+1},0))_{i\in\mathbb{N}})$$
$$\text{or } (((0,(1-\alpha)^i b))_{i=0}^{i=i_*+1}, (a(1-(1-\alpha)^{i+1},0))_{i\in\mathbb{N}})$$

$$\text{if } \tfrac{\ln(\alpha|a|/|b|)}{\ln(|1-\alpha|)} \in \mathbb{N}.$$

Thus, every sequence generated by PGD *converges to* $(a,0)$.

In conclusion, if $\alpha \neq 1$, then P^2GD converges to $(0,0)$, which is M-stationary but not B-stationary, while PGD converges to $(a,0)$, which is P-stationary and even a global minimizer of $f|_{\mathbb{R}^2_{\leq 1}}$ (and f). This is illustrated in Figure 2 for some choice of a, b, and α.

By Proposition 8.1, for every sequence $(x_i)_{i\in\mathbb{N}}$ generated by P^2GD, it holds that

$$\lim_{i\to\infty} d(-\nabla f(x_i), \widehat{N}_{\mathbb{R}^2_{\leq 1}}(x_i)) = 0.$$

Thus, the measure of B-stationarity $\mathbb{R}^2_{\leq 1} \to \mathbb{R} : x \mapsto d(-\nabla f(x), \widehat{N}_{\mathbb{R}^2_{\leq 1}}(x))$ is not lower semicontinuous at $(0,0)$, and the convergence to an M-stationary point that is not B-stationary cannot be suspected based on the mere observation of this limit. In the terminology of [32], $((0,0),(x_i)_{i\in\mathbb{N}}, f)$ is an apocalypse.

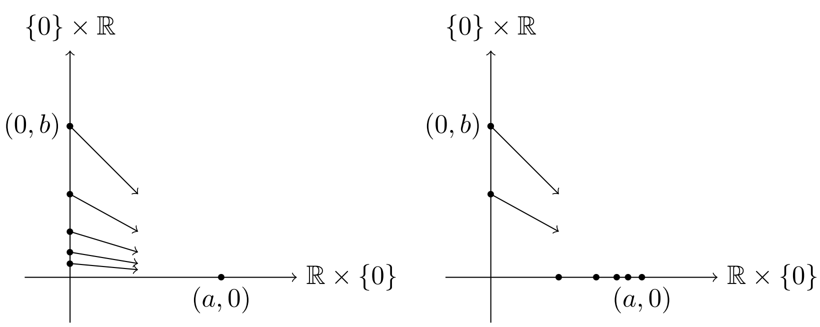


FIG. 2. *First few iterates generated by* PGD *(right) and* P^2GD *(left) on the instance of* (1.1) *studied in Section 8 with* $a := b := 1$ *and* $\alpha := 0.45$. *The arrows represent* $x_i - \alpha\nabla f(x_i)$. *The point* $(a,0)$, *which is the unique global minimizer, is also represented. It is already visible from the first few iterates that* P^2GD *converges to the M-stationary point* $(0,0)$ *while* PGD *converges to the global minimizer.*

9. Conclusion.

The main contribution of this paper is the proof of Theorem 1.2. This theorem ensures that PGD (Algorithm 4.2) enjoys the strongest stationarity properties that can be expected for problem (1.1) under the considered assumptions.

A sufficient condition for the convergence of a 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 [29] for monotone PGD under the assumption that f satisfies (H2) and a Kurdyka–Łojasiewicz property.

Two possible extensions of this work are left for future research. First, can Theorem 1.2 be extended to an algorithm that uses more general search directions than PGD? For example, a search direction at a point $x \in C$ that is not B-stationary for (1.1) could be a vector $v \notin \widehat{N}_C(x)$ that satisfies [22, 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 $c_1, c_2 \in (0,\infty)$.

Second, can Theorem 1.2 be extended to the proximal gradient algorithm as defined in [31, Algorithm 4.1] or [11, Algorithm 3.1]? The first step toward such an extension would be defining suitable stationarity notions for the corresponding problem whose objective function is not differentiable. Furthermore, significant adaptations would be needed, e.g., because inequality (2.1), which plays an instrumental role in our analysis, does not seem to admit a straightforward extension.

Appendix A. Proof of Proposition 7.1. The strict inclusion follows from [45, Proposition 7.16] and [4, Theorem 3.9] if $C = \mathbb{R}^n_{\leq s}$, from [45, Proposition 6.7] and [60, Theorem 3.4] if $C = \mathbb{R}^n_{\leq s} \cap \mathbb{R}^n_+$, from [27, Corollary 2.3 and Theorem 3.1] if $C = \mathbb{R}^{m\times n}_{\leq r}$, and from [45, Proposition 6.28] and [60, 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 [45, Proposition 7.16] and [27, 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 [45, Proposition 6.7] and [60, Proposition 3.2] or [45, Proposition 6.28] and [10, Corollary 17]; the detail is given below for completeness.

Assume that C is $\mathbb{R}^n_{\leq s} \cap \mathbb{R}^n_+$. Let $\mathrm{supp}(x) := \{i \in \{1,\dots,n\} \mid x_i \neq 0\}$. By [45, Proposition 6.7],

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

Thus, by [60, 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 [45, 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 $\mathrm{S}(n) := \{X \in \mathbb{R}^{n\times n} \mid X^\top = X\}$, $\mathrm{S}(n)^\perp = \{X \in \mathbb{R}^{n\times n} \mid X^\top = -X\}$, and $\mathrm{S}^-(n) := \{X \in \mathrm{S}(n) \mid X \preceq 0\}$. Let $Z \in \widehat{N}_{\mathrm{S}^+_{\leq r}(n)}(X)$ and $Z_{\mathrm{sym}} := \frac{1}{2}(Z + Z^\top)$. Then, by [10, Corollary 17], $P_{\mathrm{S}^+_{\leq r}(n)}(X+Z) = P_{\mathrm{S}^+_{\leq r}(n)}(X+Z_{\mathrm{sym}})$. Let $\underline{r} := \mathrm{rank}\, X$ and $\tilde{r} := \mathrm{rank}\, Z_{\mathrm{sym}}$. Since $\mathrm{im}\, Z_{\mathrm{sym}} \subseteq \ker X$, $\tilde{r} \leq n - \underline{r}$ and there exists $U \in \mathrm{O}(n)$ such that

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

and

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

are eigendecompositions. Thus,

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

is an eigendecomposition. Hence, by [10, Corollary 17], $P_{\mathrm{S}^+_{\leq r}(n)}(X + Z_{\mathrm{sym}}) = \{X\}$.

Acknowledgments. The authors thank two anonymous referees for several helpful comments that improved the quality of the paper.

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] A. A. Ahmadi and J. Zhang, *On the complexity of finding a local minimizer of a quadratic function over a polytope*, Mathematical Programming, 195 (2022), pp. 783–792, https://doi.org/10.1007/s10107-021-01714-2.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] M. Benko and H. Gfrerer, *On estimating the regular normal cone to constraint systems and stationarity conditions*, Optimization, 66 (2017), pp. 61–92, https://doi.org/10.1080/02331934.2016.1252915.

[8] M. Benko and H. Gfrerer, *New verifiable stationarity concepts for a class of mathematical programs with disjunctive constraints*, Optimization, 67 (2018), pp. 1–23, https://doi.org/10.1080/02331934.2017.1387547.

[9] 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.

[10] 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.

[11] A. De Marchi, *Proximal gradient methods beyond monotony*, Journal of Nonsmooth Analysis and Optimization, 4 (2023), https://doi.org/10.46298/jnsao-2023-10290.

[12] 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.

[13] 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.

[14] M. L. Flegel and C. Kanzow, *Abadie-type constraint qualification for mathematical programs with equilibrium constraints*, Journal of Optimization Theory and Applications, 124 (2005), pp. 595–614, https://doi.org/10.1007/s10957-004-1176-x.

[15] M. L. Flegel and C. Kanzow, *On the Guignard constraint qualification for mathematical programs with equilibrium constraints*, Optimization, 54 (2005), pp. 517–534, https://doi.org/10.1080/02331930500342591.

[16] M. L. Flegel, C. Kanzow, and J. V. Outrata, *Optimality conditions for disjunctive programs with application to mathematical programs with equilibrium constraints*, Set-Valued and Variational Analysis, 15 (2007), pp. 139–162, https://doi.org/10.1007/s11228-006-0033-5.

[17] 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.

[18] M. Fukushima and J.-S. Pang, *Convergence of a smoothing continuation method for mathematical progams with complementarity constraints*, in Ill-posed Variational Problems and Regularization Techniques, M. Théra and R. Tichatschke, eds., vol. 477 of Lecture Notes in Economics and Mathematical Systems, Springer, Berlin, Heidelberg, 1999, pp. 99–110, https://doi.org/10.1007/978-3-642-45780-7_7.

[19] M. Fukushima and P. Tseng, *An implementable active-set algorithm for computing a B-stationary point of a mathematical program with linear complementarity constraints*, SIAM Journal on Optimization, 12 (2002), pp. 724–739, https://doi.org/10.1137/S1052623499363232.

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

[21] H. Gfrerer, *Optimality conditions for disjunctive programs based on generalized differentiation with application to mathematical programs with equilibrium constraints*, SIAM Journal

on Optimization, 24 (2014), pp. 898–931, https://doi.org/10.1137/130914449.
[22] 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.
[23] 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.
[24] 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.
[25] 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.
[26] 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.
[27] 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.
[28] X. M. Hu and D. Ralph, *Convergence of a penalty method for mathematical programming with complementarity constraints*, Journal of Optimization Theory and Applications, 123 (2004), pp. 365–390, https://doi.org/10.1007/s10957-004-5154-0.
[29] 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.
[30] 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.
[31] 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.
[32] 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.
[33] E. Levin, J. Kileel, and N. Boumal, *The effect of smooth parametrizations on nonconvex optimization landscapes*, Mathematical Programming, (2024), https://doi.org/10.1007/s10107-024-02058-3.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] 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.
[40] P. Mehlitz, *Stationarity conditions and constraint qualifications for mathematical programs with switching constraints*, Mathematical Programming, 181 (2020), pp. 149–186, https://doi.org/10.1007/s10107-019-01380-5.
[41] 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.
[42] B. S. Mordukhovich, *Variational Analysis and Applications*, Springer Monographs in Math-

ematics, Springer Cham, 2018, https://doi.org/10.1007/978-3-319-92775-6.
[43] 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.
[44] 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.
[45] G. Olikier, K. A. Gallivan, and P.-A. Absil, *First-order optimization on stratified sets*, (2023), https://arxiv.org/abs/2303.16040.
[46] G. Olikier, K. A. Gallivan, and P.-A. Absil, *Low-rank optimization methods based on projected-projected gradient descent that accumulate at bouligand stationary points*, (2024), https://arxiv.org/abs/2201.03962v2.
[47] 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.
[48] 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.
[49] 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.
[50] E. Pauwels, *Generic Fréchet stationarity in constrained optimization*, (2024), https://arxiv.org/abs/2402.09831v1.
[51] E. Polak, *Computational Methods in Optimization*, vol. 77 of Mathematics in Science and Engineering, Academic Press, 1971.
[52] 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.
[53] 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.
[54] 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.
[55] 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.
[56] S. Scholtes, *Convergence properties of a regularization scheme for mathematical programs with complementarity constraints*, SIAM Journal on Optimization, 11 (2001), pp. 918–936, https://doi.org/10.1137/S1052623499361233.
[57] S. Scholtes, *Nonconvex structures in nonlinear programming*, Operations Research, 52 (2004), pp. 368–383, https://doi.org/10.1287/opre.1030.0102.
[58] S. Scholtes and M. Stöhr, *Exact penalization of mathematical programs with equilibrium constraints*, SIAM Journal on Optimization, 37 (1999), pp. 617–652, https://doi.org/10.1137/S0363012996306121.
[59] S. Steffensen and M. Ulbrich, *A new relaxation scheme for mathematical programs with equilibrium constraints*, SIAM Journal on Optimization, 20 (2010), pp. 2504–2539, https://doi.org/10.1137/090748883.
[60] 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.
[61] A. Themelis, L. Stella, and P. Patrinos, *Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms*, SIAM Journal on Optimization, 28 (2018), pp. 2274–2303, https://doi.org/10.1137/16M1080240.
[62] P. P. Varaiya, *Nonlinear programming in Banach space*, SIAM Journal on Applied Mathematics, 15 (1967), pp. 284–293, https://doi.org/10.1137/0115028.
[63] M. Willem, *Functional Analysis: Fundamentals and Applications*, Cornerstones, Birkhäuser New York, 2013, https://doi.org/10.1007/978-1-4614-7004-5.
[64] 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.
[65] 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.