Generic Fréchet stationarity in constrained optimization

Edouard Pauwels

Generic Fréchet stationarity in constrained optimization

Edouard Pauwels*

February 13, 2024

Abstract

Minimizing a smooth function f on a closed subset C leads to different notions of stationarity: Fréchet stationarity, which carries a strong variational meaning, and criticallity, which is defined through a closure process. The latter is an optimality condition which may loose the variational meaning of Fréchet stationarity in some settings. We show that, while criticality is the appropriate notion in full generality, Fréchet stationarity is typical in practical scenarios. This is illustrated with two main results, first we show that if C is semi-algebraic, then for a generic smooth semi-algebraic function f, all critical points of f on C are actually Fréchet stationary. Second we prove that for small step-sizes, all the accumulation points of the projected gradient algorithm are Fréchet stationary, with an explicit global quadratic estimate of the remainder, avoiding potential critical points which are not Fréchet stationary, and some bad local minima.

Keywords— Constrained optimization, nonconvex optimization, optimality conditions, stationarity, semi-algebraic optimization, genericity, projected gradient algorithm.

1 Introduction

We consider the problem

$$\min_{x\in C} f(x) \tag{1}$$

where $f\colon \mathbb{R}^p \to \mathbb{R}$ is C^1 and $C \subset \mathbb{R}^p$ is closed. A point $x \in C$ is called Fréchet stationary for f on C if $f(y) - f(x) \geq o(\|y - x\|)$ for $y \in C$ and a vector v is called a regular normal vector to C at x whenever x is Fréchet stationary for the linear form $x \mapsto -\langle v, x\rangle$ on C. The notion of regular normal vector lacks basic continuity properties, and in particular limits of regular normals may not correspond to regular normals. The broader notion of criticality aims at recovering a form of continuity, see Section 2.1. But the price is the variational meaning of Fréchet stationarity which may be lost due to lack of regularity (see Example 1). The purpose of this work is to formaly show that despite the widespread use of the notion of criticality in non convex optimization, the vast majority of cases encountered relate to the stronger notion Fréchet stationarity, aligning formal guaranties with practical observations.

We first show that if f and C are assumed to be semi-algebraic, then considering the functions $\{f_v\colon x \mapsto f(x)+\langle v, x\rangle\}$, generically in v, all critical points of f_v on C are Fréchet stationary. This result illustrates the fact that the existence of critical points which are not Fréchet stationary is the consequence of a bad alignment of the objective function f and the constraint set C, which

*Toulouse School of Economics, Toulouse France. Institut Universitaire de France (IUF).

is very unlikelly under rigidity assumptions, modulo potential small perturbations of the loss function. This result is stated in the semi-algebraic setting which encompasses many practical scenarios, including sparse vectors, bounded rank matrices. The same result holds for broader classes of functions and constraint sets, definable in o-minimal structures, but we do not expand on this and stick to the semi-algebraic setting for simplicity.

Our second main result relates to the well known projected gradient algorithm. A typical feature is that the resulting sequences tend to be attracted by critical points and we consider the question of Fréchet stationarity for these limit points. It turns out that the answer is positive, the projected gradient algorithm produce sequences which are attracted by Fréchet stationary points, with an explicit global quadratic estimate of the negative variation remainder. Although simple, this result provides a much stronger variational guaranty for the resulting limit points compared to mere criticality. In particular, we obtain a global quadratic lower bound, which is reminescent of the optimality conditions from the convex setting. This is a desirable feature of the algorithm as one preserves a strong variational meaning for the accumulation points. The result follows from a more general analysis of the proximal gradient algorithm, of independent interest, for which the projected gradient algorithm is a special case. These results are obtained under very general assumptions, and in particular, they do not rely on semi-algebraicity.

1.1 Motivation and related work

Sparsity and rank constrained optimization: The phenomenon described above does not affect convex constraint sets C or more generally Clarke regular constraint sets C see for example [50, Definition 6.4]. This includes constraint sets defined by smooth inequalities in nonlinear programming, under qualification conditions [50, Theorem 6.14]. The most well known applications where this property fails involve cardinality constraints, such as sparsity and rank. The sparse setting was largely studied in [4] with a carefull analysis of optimality conditions and algorithms which were extended in [5]. For low rank matrices, it was remarked in [39] that rank deficiency is the source of an absence of regularity, with a potential detrimental effect on the interpretation of optimality conditions [31].

The consequences of lack of regularity on optimality conditions, and optimality measures, was further studied in [37] for low rank matrices under the name "apocalypses" with a very similar flavor as [4] for sparsity constraints. This constitutes further motivations to develop algorithmic schemes for low rank matrix optimization which avoid this pathology and are attracted by stationary points in [37] followed by [32, 44, 43, 45]. As mentioned in [46], the absence of regularity only has rare consequences in practice and our main motivation is to provide formal guaranties for this observation in the form of genericity results on problem data and convergence guaranties for the projected gradient algorithm.

Genericity in tame optimization: Our first main result, Theorem 1, relates to semi-algebraicity or tameness of the considered objective function f and the constraint set C. Studying nonconvex optimization and first order methods under such rigidity assumptions have a long history in optimization. Indeed, semi-algebraicity has numerous structural consequences on the optimization losses [10, 12, 11, 33]. Furthermore, virtually all losses met in an optimization context are covered by tameness assumptions, see the numerous examples in [2, 3], and the connection with deep learning in [16, 17]. One can take advantage of these properties, for which semi-algebraicity is a mild sufficient condition, to analyse optimization algorithms and optimization landscapes. Examples include sequential convergence of deterministic optimization algorithms [1, 2, 3, 18, 15, 48], as well as the analysis of stochastic first order methods [25, 7, 17, 16, 9].

Genericity is a notion that is used to express the fact that a certain behavior is typical. It is

most often expressed in measure theoretic terms (Lebesgue almost everywhere), or topological terms (residual sets are countable intersections of sets with dense interior). In general, the two notions do not coincide (see for example [47, Theorem 1.6]), but in the semi-algebraic setting they coincide and sometimes correspond to a stronger notion: being the complement of the union of finitely many lower dimensional embedded smooth manifolds. This is essentially due to the stratification property [53, 4.8]. Genericity results in an optimization context relate to the typical structure of the data of optimization problems [22, 13, 23, 49, 27, 14, 35] and generic desirable properties of optimization methods [42, 8, 24, 26]. Our first main result falls in this category, we show that for a generic semi-algebraic f and a fixed semi-algebraic set C, there is no critical point which is not Fréchet stationary for the resulting constrained minimization problem. A consequence of this result is that for a generic smooth semi-algebraic function f and a fixed closed set C, the "apocalypses" described in [37] do not exist.

Projected gradient algorithm: Our second main result, Theorem 2, concerns the projected gradient algorithm proposed independently by Glodstein [30] and Levitin Polyak [38] for convex optimization with subsequent contributions in the convex setting [6, 19, 28, 29]. Our analysis is a consequence of a detailed analysis of the proximal gradient algorithm, the proximal mapping generalizing the projection. Convergence of the proximal point algorithms in a non-convex setting was considered in [51, 34, 1], convergence of the proximal gradient algorithms under semi-algebraic assumptions was given in [3]. In the nonconvex setting, existing convergence guaranties are related to a notion of criticallity, weaker than Fréchet stationarity and we show that the analysis can be extended to obtain a quantitative version of Fréchet stationarity. This can be seen as an extension of the L-stationarity result obtained in [4] for the projected gradient algorithm under sparsity constraints to general sets, and constitutes an element of answer regarding the convergence guaranties of the projected gradient algorithm related to the concerns raised in [37].

1.2 Notations

Throughout the paper, the ambiant space is $\mathbb{R}^p$. We denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$, the Euclidean scalar product and Euclidean norm. We denote a set valued map F, from $\mathbb{R}^p$ to subsets of $\mathbb{R}^p$ with the notation $F \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$. For a subset $C \subset \mathbb{R}^p$, we denote by $T_C, \hat{N}_C, N_C$ the tangent, regular normal and normal cones respectively which are seen as set valued maps $\mathbb{R}^p \rightrightarrows \mathbb{R}^p$ with empty values outside C. Relevant definitions are introduced along the paper.

2 Main results

We introduce the required elements of variational geometry in Section 2.1 and state our two main results in Section 2.2 and Section 2.3.

2.1 Notions of stationarity

We use the notations and denominations of [50]. First recall the definitions of the objects of interest.

Definition 1 (Tangent and Normal Cones)**.** *For $x \in C$, $w \in \mathbb{R}^p$ is an element of the tangent cone of C at x, written $w \in T_C(x)$ if*

$$\frac{x_k - x}{\tau_k} \to w$$

for some sequence $(x_k)_{k\in\mathbb{N}}$*, in* C *and* $(\tau_k)_{k\in\mathbb{N}}$ *in* $\mathbb{R}_+$ *decreasing to* 0*. Furthermore,* $v \in \mathbb{R}^p$ *is an element of the regular normal cone of* C *at* x*, written* $w \in \hat{N}_C(x)$ *if*

$$\langle v, y - x\rangle \leq o(\|y - x\|), \qquad y \in C,$$

where the inequality is understood as $\limsup_{y\to x} \frac{\langle v, y-x\rangle}{\|y-x\|} \leq 0$*. Finally,* $v \in \mathbb{R}^p$ *is an element of the normal cone of* C *at* x*, written* $w \in N_C(x)$ *if*

$$\exists (x_k)_{k\in\mathbb{N}},\ (v_k)_{k\in\mathbb{N}},\ x_k \in C,\ v_k \in \hat{N}_C(x_k),\ k \in \mathbb{N},\ x_k \to x,\ v_k \to w,\ k \to \infty.$$

$T_C, \hat{N}_C$ *and* N_C *can be seen as set valued maps* $\mathbb{R}^p \rightrightarrows \mathbb{R}^p$ *by assigning empty values for* $x \notin C$*.*

We gather known facts about these cones, the following is taken from Theorem 6.21 and 6.28 [50].

Proposition 1. *Let* $C \subset \mathbb{R}^p$ *be closed, then for all* $x \in C$*,* $T_C(x)$ *is a closed cone and* $\hat{N}_C(x)$ *is the polar of* $T_C(x)$*:* $\hat{N}_C(x) = \{v \in \mathbb{R}^p,\ \langle v, w\rangle \leq 0,\ \forall w \in T_C(x)\}$*.*

Let $f\colon \mathbb{R}^p \to \mathbb{R}$ *be* C^1*. Suppose that* $x \in C$ *is a local minimum of* f *restricted to* C*, then the two equivalent conditions hold:*

$$\begin{aligned} -\nabla f(x) &\in \hat{N}_C(x) \\ \operatorname{proj}_{T_C(x)}(-\nabla f(x)) &= 0. \end{aligned} \tag{2}$$

A point $x \in C$ *which satisfy* (2) *is called Fréchet stationary for* f *on* C*, in which case,*

$$f(y) - f(x) \geq o(\|y - x\|). \tag{3}$$

This implies the stronger condition $-\nabla f(x) \in N_C(x)$*, an* $x \in C$ *satisfying this condition is called critical for* f *on* C*.*

Proposition 1 suggests to use (2) as an optimality condition for constrained optimization, however the proposed quantity lacks basic continuity in general, which is troublesome for many applications. This motivates the introduction of the normal cone to C, N_C which is the graph closure of the regular normal cone to C, recovering some form of continuity and the possibility to pass to limits. Typical optimization results fall in this scope and provide guaranties in terms of criticality, $-\nabla f(x) \in N_C(x)$, in the context of (1) which is necessary but not sufficient for (2). While this constitutes a bona fide optimality condition, in the sense that if it is not satisfied, x is not a local extremum, it may result in meaningless notion of criticality contrary to the interpretation of Fréchet stationarity in (3).

Example 1. *Let* $C \subset \mathbb{R}^2$ *be the set of* 1*-sparse vector, then* $N_C(0,0) = T_C(0,0) = C$ *while* $\hat{N}_C(0,0) = \{0\}$*. Set* $f\colon (x, y) \mapsto (x-1)^2 + y^2$*, then* $(0,0)$ *is critical for* f *on* C *but this does not have much variational meaning since* f *has directional derivative* 2 *in the* y *direction which is in* $T_C(0,0)$ *and is actually admissible with respect to the constraint induced by* C*.*

To ellaborate on this remark and illustrate the relevance of the normal cone in comparison to the regular normal cone, we quote Rockafellar and Wets [50] regarding the phenomenon presented in Example 1:

> *This possibility causes some linguistic discomfort over 'normality', but the cone of such limiting normal vectors comes to dominate technically in formulas and proofs,* [...] *Many key results would fail if we tried to make do with regular normal vectors alone.*

This absence of regularity is related to the main motivations in [37] to propose algorithms which do not suffer from it.

2.2 Genericity of Fréchet stationarity

We start by introducing the necessary tools from semi-algebraic geometry. An introduction to semi-algerbaic and tame geometry is found in [20, 21] and a comprehensive overview is given in [53], see also [52]. We recall all the required concepts with necessary bibliographic pointers.

2.2.1 Semi-algebraic geometry

Let us first introduce the required definitions.

Definition 2. *Let $p, q \in \mathbb{N}$ be arbitrary.*

A basic semi-algebraic set $S \subset \mathbb{R}^p$ is the solution set of a polynomial system of inequalities.

$$S = \{x \in \mathbb{R}^p,\ P(x) = 0, Q_1(x) > 0, \dots Q_m(x) > 0\} \tag{4}$$

where $P, Q_1, \dots, Q_m$ are polynomials of p variables and $m \in \mathbb{N}$.

A semi-algebraic set is the finite union of basic semi-algebraic sets.

A function $f\colon \mathbb{R}^p \to \mathbb{R}^q$ is semi-algebraic if its graph $\{(x, z) \in \mathbb{R}^{p+q},\ z = f(x)\}$ is semi-algebraic.

A set-valued map $F\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ is semi-algebraic if its graph $\{(x, z) \in \mathbb{R}^{p+q},\ z \in F(x)\}$ is semi-algebraic.

Example 2 (semi-algebraic functions)**.** *Euclidean norm, square root, quotients, rational powers, matrix rank are semi-algebraic functions. Semi-algebraic functions are closed under composition.*

Semi-algebraic objects are closed under many relevant operations: intersection, unions, complementation and Cartesian product [20]. The Tarski-Seidenberg principle [21, Theorem 2.3] allows to characterize semi-algebraic sets as the smallest o-minimal structure [20, Exercise 1.17]. Therefore the general tools of o-minimal geometry apply to the semi-algebraic setting. Furthermore, many results for semi-algebraic sets naturaly extend to the o-minimal setting. In this spirit, we gather below properties which will be useful in order to prove our genericity result.

Proposition 2. *Let $p, m \in \mathbb{N}$ be arbitrary and $C \subset \mathbb{R}^p$ and $F\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^m$ be semi-algebraic.*

1. *The projection of C onto a subspace is semi-algebraic .*
2. *T_C, $\hat{N}_C$ and N_C are semi-algebraic set valued maps. .*
3. *The interior and closure of C are semi-algebraic .*
4. *If F is a differentiable function (single valued), then its Jacobian is a semi-algebraic function.*
5. *The image of C by F, $F(C) \subset \mathbb{R}^m$ is semi-algebraic.*
6. *C can be partitioned into a finite union of disjoint semi-algebraic smooth embedded submanifolds .*

Proof of Proposition 2: These are well known and can be found in [53, 21, 20], we provide proof arguments and detailed pointers for completeness.

1. This is Tarski-Seidenberg Theorem, up to a rotation, see [21, Theorem 2.3]. An equivalent formulation of this result is [21, Theorem 2.6] states that every first order formula (quantification on variables), involving semi-algebraic sets or functions, polynomials, inequalities, equalities and the logical negation, conjunction and disjunction, describes a semi-algebraic object. In the following, we describe each set of interest with such a first order formula, which implies that they are semi-algebraic (see [21, Section 2.1.2] and [20, Theorem 1.13]).

2.
$$\begin{aligned} z \in T_C(x) \quad &\Leftrightarrow \quad \forall \epsilon > 0,\, \exists y \in C,\, y \neq x,\, \|y - x\| \leq \epsilon,\, \left| \frac{y-x}{\|y-x\|} - \frac{z}{\|z\|} \right| \leq \epsilon \\ z \in \hat{N}_C(x) \quad &\Leftrightarrow \quad \forall w \in T_C(x),\, \langle z, w \rangle \leq 0 \\ z \in N_C(x) \quad &\Leftrightarrow \quad \forall \epsilon > 0, \exists y \in C,\, \exists v \in \hat{N}_C(y),\, \|x - y\| \leq \epsilon,\, \|v - z\| \leq \epsilon. \end{aligned}$$

3.
$$\begin{aligned} x \in \text{int } C \quad &\Leftrightarrow \quad \exists \epsilon > 0,\, \forall y \in \mathbb{R}^p,\, \|x - y\| > \epsilon \text{ or } y \in C \\ x \in \text{cl } C \quad &\Leftrightarrow \quad \forall \epsilon > 0, \exists y \in C, \|y - x\| \leq \epsilon. \end{aligned}$$

4.
$$\begin{aligned} M = J_F(x) \quad \Leftrightarrow \quad & M \in \mathbb{R}^{m \times p},\, \forall \epsilon > 0,\, \exists \delta > 0,\, \forall y \neq x, \\ & \|y - x\| > \delta \text{ or } \frac{\|F(y) - F(x) - M(y-x)\|}{\|y - x\|} \leq \epsilon \end{aligned}$$

5.
$$z \in F(C) \quad \Leftrightarrow \quad \exists x \in C,\, z \in F(x).$$

6. This is the geometric notion of stratification, see for example in [53, Claim 4.8].

□

Semi-algebraic sets come with a notion of integral dimension, denoted by dim C for a semi-algebraic set C, which agrees with the classical notion of dimension for affine sets or embedded manifolds. The following facts can be found in [20, Proposition 3.17, Theorem 3.22] and will be useful to prove our genericity result.

Proposition 3. *Let $p, m \in \mathbb{N}$ be arbitrary.*

1. *If $B \subset A \subset \mathbb{R}^p$ are semi-algebraic, then* $\dim B \leq \dim A$.
2. *If $A, B \subset \mathbb{R}^p$ are semi-algebraic, then* $\dim\, A \cup B = \max\{\dim\, A, \dim\, B\}$.
3. *For any $A \subset \mathbb{R}^p$ semi-algebraic,* $\dim \text{ cl } A = \dim\, A$, $\dim \text{ cl } A \setminus A < \dim\, A$.
4. *For $f \colon \mathbb{R}^p \to \mathbb{R}^m$ and $A \subset \mathbb{R}^p$, both semi-algebraic,* $\dim\, f(A) \leq \dim\, A$.

2.2.2 Main result

The following is our first main result. It is stated in the semi-algebraic setting, but it can be extended to functions and sets which are definable in the same o-minimal structure as the arguments rely on the elements described in Proposition 2 and Proposition 3 which hold for definable functions and sets [20, 53].

Theorem 1. *Let $f\colon \mathbb{R}^p \to \mathbb{R}$ be continuously differentiable and $C \subset \mathbb{R}^p$ be closed, both semi-algebraic. Then there is $V \subset \mathbb{R}^p$, which is a finite union of semi-algebraic embedded manifolds of dimension at most $p-1$, such that for all $v \notin V$, all critical point of $f_v\colon x \mapsto f(x) + \langle v, x\rangle$ on C are Fréchet stationary.*

Proof of Theorem 1: By Proposition 2 item 2, $\hat{N}_C$ and N_C are semi-algebraic.

We work with $\hat{N}_C\colon C \rightrightarrows \mathbb{R}^p$. It follows from Definition 1 that for any $S \subset C$, we have for all $x \in S$, $\hat{N}_C(x) \subset \hat{N}_S(x)$. We may consider a partition of C into $M_1, \dots, M_m$ embedded smooth manifolds by Proposition 2 item 6. For each $i = 1, \dots, m$, we have $M_i \subset C$ and therefore $\hat{N}_C(x) \subset \hat{N}_{M_i}(x)$. But $\hat{N}_{M_i}(x)$ is simply the normal space of M_i at x as described by differential geometry. Therefore the graph of $\hat{N}_C$ restricted to M_i is contained in the normal bundle of M_i which can be seen as an embedded submanifold of $\mathbb{R}^{2p}$ of dimension p, see for example [36, Theorem 6.23]. Therefore, the graph of the restriction of $\hat{N}_C$ to M_i is of dimension at most p by Proposition 3 item 1. The graph of $\hat{N}_C$ is the union of its restriction to M_i, $i = 1, \dots, m$ and it is therefore of dimension at most p by Proposition 3 item 2.

Set G the closure of graph $\hat{N}_C$ in $\mathbb{R}^{2p}$, it is semi-algebraic by Proposition 2 item 3. We have that $(x, z) \in G$ if and only if there is a sequence $x_k \to x$ and $z_k \to z$ such that $z_k \in \hat{N}_C(x_k)$ for all $k \in \mathbb{N}$. In other words, we have $G = \text{graph } N_C$. By Proposition 3 item 3, the semi-algebraic set $H = \text{cl}(\text{graph } \hat{N}_C) \setminus \text{graph } \hat{N}_C \subset \mathbb{R}^{2p}$ has dimension at most $p-1$. The set H can be understood as the graph of the possibly empty-valued map $S\colon x \rightrightarrows N_C(x) \setminus \hat{N}_C(x)$.

Now consider the set valued map $R\colon x \rightrightarrows S(x) + \nabla f(x)$. Using Proposition 3 item 4 the dimension of graph R is at most $p-1$ because it is the image of H, by the map $(x, z) \mapsto (x, z + \nabla f(x))$ which is semi-algebraic by Proposition 2 item 4. Now we have the following equivalence, for any $v \in \mathbb{R}^p$

$$\exists x \in C,\ -\nabla f(x) - v \in N_C(x) \setminus \hat{N}_C(x) \qquad \Leftrightarrow \qquad \exists x \in C, v \in R(x)$$

so that

$$\left\{v \in \mathbb{R}^p, \exists x \in C,\ -\nabla f(x) - v \in N_C(x) \setminus \hat{N}_C(x)\right\} \qquad = \qquad \text{proj}_v \text{ graph } R.$$

where $\text{proj}_v(x, z) = z$ for any $x, z \in \mathbb{R}^p$. Setting $V = \text{proj}_v$ graph R, we have that $\dim V \leq p-1$ by Proposition 3 item 4. By Proposition 2 item 6, V is a finite union of semi-algebraic embedded submanifolds of dimension $p-1$ at most by Proposition 3 item 2. □

Remark 1. *The result of Theorem 1 holds generically in v, as understood in both measure theoretic terms (almost everywhere), or topological terms (residual). We perturb f using a linear form, but one could consider a peturbation of the form $x \mapsto \epsilon\|x - c\|^2$ for small $\epsilon > 0$, and the same result would hold generically in $c \in \mathbb{R}^p$. It is easy to see that the critical point example in Example 1 would not persist under generic perturbation and Theorem 1 provides a general ground for this observation.*

2.3 Projected gradient is attracted by Fréchet stationary points

Given a non-empty closed set $C \subset \mathbb{R}^p$, the projection of $x \in \mathbb{R}^p$ on C, denoted by $\mathrm{proj}_C(x)$ is given by the non-empty set

$$\mathrm{proj}_C(x) = \arg\min_{z \in C} \|x - z\|.$$

Given an initial point $x_0 \in C$ and a step-size parameter, $\gamma > 0$, the projected gradient algorithm iterates.

$$x_{k+1} \in \mathrm{proj}_C(x_k - \gamma \nabla f(x_k)). \tag{5}$$

The following is our second main result. It is a consequence of the analysis of the proximal gradient algorithm proposed in Section 3.2

Theorem 2. *Let $f\colon \mathbb{R}^p \to \mathbb{R}$ be C^1 with L-Lipschitz gradient and $C \subset \mathbb{R}^p$ be non-empty and closed. Then for any step size $\gamma < 1/L$, any accumulation point of the projected gradient algorithm, $\bar{x}$, is Fréchet stationary such that*

$$f(y) \geq f(\bar{x}) - \frac{1}{\gamma}\|y - \bar{x}\|^2, \qquad \forall y \in C, \tag{6}$$

$$\mathrm{proj}_C(\bar{x} - s\nabla f(\bar{x})) = \{\bar{x}\}, \qquad \forall 0 < s < \gamma.$$

Furthermore $\mathrm{proj}_{T_C(x_k)}(-\nabla f(x_k)) \to 0$ *as* $k \to \infty$.

Proof : From [50, Exercise 8.14]: $\hat{\partial}\delta_C(\bar{x}) = \hat{N}_C(x)$, where δ_C is the indicator function of C with value 0 on C an $+\infty$ outside. Note that δ_C satisfies the hypotheses of Lemma 3 for any $\delta > 0$.

We have for any that $\tilde{f} = \gamma f$ has $L\gamma < 1$ Lipschitz gradient and the proximal gradient algorithm with unit step on $\tilde{f}$ and $g = \delta_C$ in (7) is equivalent to the projected gradient algorithm on f with step size γ so that Theorem 4 applies. We obtain that all accumulation points $\bar{x}$ are Fréchet stationary such that $\bar{x} \in \mathrm{proj}_C(\bar{x} - \gamma\nabla f(\bar{x}))$ from Theorem 4, which means that $\mathrm{dist}(\bar{x} - \gamma\nabla f(\bar{x})) = \gamma\|\nabla f(\bar{x})\|$. The quantitative statement on Fréchet stationarity follows from Theorem 4 applied to $\tilde{f} = \gamma f$.

Let us prove unicity of the projection fix $0 < s < \gamma$. The case $\nabla f(\bar{x}) = 0$ is obvious so let us eliminate it. Denote by B_1 the ball of center $\bar{x} - \gamma\nabla f(\bar{x})$ and radius $\gamma\|\nabla f(\bar{x})\|$ and B_2 the ball of center $\bar{x} - s\nabla f(\bar{x})$ and radius $s\|\nabla f(\bar{x})\|$. Let us show that for any $x \in B_2$, $x \neq \bar{x}$, we have

$$\begin{aligned}\|x - \bar{x} + \gamma\nabla f(\bar{x})\| &= \|x - \bar{x} + s\nabla f(\bar{x}) + (\gamma - s)\nabla f(\bar{x})\| \\ &< s\|\nabla f(\bar{x})\| + (\gamma - s)\|\nabla f(\bar{x})\| = \gamma\|\nabla f(\bar{x})\|\end{aligned}$$

where the strict inequality is from the triangle inequality. Indeed, either the triangle inequality is strict, or $x - \bar{x} + s\nabla f(\bar{x}) = \alpha(\gamma - s)\nabla f(\bar{x})$ for some $\alpha \geq 0$. In this second case, since $x \in B_2$, by taking the norm, we obtain $s \geq \alpha(\gamma - s)$, so that $x = \bar{x} - t\nabla f(\bar{x})$ where $t = s - \alpha(\gamma - s) \geq 0$ and $t \leq s$. The case $t = 0$ is excluded because we assumed that $x \neq \bar{x}$ and we have $0 < t \leq s < \gamma$, so that

$$\|x - \bar{x} + \gamma\nabla f(\bar{x})\| = \|(\gamma - t)\nabla f(\bar{x})\| = |\gamma - t|\ \|\nabla f(\bar{x})\| < \gamma\|\nabla f(\bar{x})\|$$

We have shown that any $x \in B_2$ different from $\bar{x}$ is actually in $\mathrm{int}B_1$ and therefore at positive distance from C, otherwise this would contradict $\mathrm{dist}(\bar{x} - \gamma\nabla f(\bar{x})) = \|\nabla f(\bar{x})\|$. Since $\bar{x} \in C$, it is the unique element in $B_2 \cap C$ which proves unicity of the projection.

We conclude regarding the last statement

$$\mathrm{proj}_{T_C(x_k)}(-\nabla f(x_k)) \to 0$$

by combining the fact that $\mathrm{dist}(\hat{N}_C(x_k), -\nabla f(x_k)) \to 0$ from Theorem 4 and Lemma 1. □

Lemma 1. *Let $T \subset \mathbb{R}^p$ be a closed cone, not necessarily convex and N be its polar, $N = \{v \in \mathbb{R}^p, \langle w, v\rangle \leq 0, \forall w \in T\}$, then for any $x \in \mathbb{R}^p$, $\|\mathrm{proj}_T(x)\| \leq \mathrm{dist}(x, N)$.*

Proof : Set $z = \mathrm{proj}_T(x)$, if $\|z\| = 0$, then there is nothing to prove. Assume that $\|z\| > 0$. Set $\tilde{T} = \{\lambda z, \lambda \geq 0\}$ and $\tilde{N}$ its polar, we have

$$\tilde{T} \subset T, \qquad \tilde{N} \supset N, \qquad \mathrm{proj}_{\tilde{T}}(x) = z.$$

Both $\tilde{N}$ and $\tilde{T}$ are convex cones and by Moreau's identity [40, Section 4.b], we have $x = \mathrm{proj}_{\tilde{T}}(x) + \mathrm{proj}_{\tilde{N}}(x)$ so that

$$\|\mathrm{proj}_{\tilde{T}}(x)\| = \|x - \mathrm{proj}_{\tilde{N}}(x)\| = \mathrm{dist}(x, \tilde{N}) \leq \mathrm{dist}(x, N).$$

□

Remark 2 (Comments on Theorem 2)**.** *It was identified in [4] that for sparsity constraints, local minimizers need to be fixed point of the projected gradient algorithm (a condition termed L stationarity) and the projection has to be univalued. Theorem 2 shows that for general sets, the projected gradient algorithm will be attracted by such points, generalizing the result of [4] for the Iterative Hard Thresholding algorithm, as illustrated in Figure 1. This result is related to the notion of proximal normals [50, Exemple 6.16], the sequences are actually attracted by the set of points $\bar{x} \in C$ such that $-\nabla f(\bar{x})$ is a proximal normal of C at $\bar{x}$. The last assertion in Theorem 2 ensures that the so called "serendipity" phenomenon described in [37, Definition 2.8] does not affect the projected gradient algorithm. If we assume in addition that f and C are semi-algebraic, then the sequence actually converges, as shown in [3]. Finally if f is convex one can add a factor $\frac{1}{2}$ in front of the quadratic term in* (6)*.*

2.4 Numerical illustration

We illustrate the relevance of the result of Theorem 2, first with the avoidance of a critical point which is not Fréchet stationary as in Example 1, and second with the avoidance of bad local minima on a grid. These are illustrative toy examples, and in both cases the observed behavior could be justified with elementary dedicated arguments. Exploring consequences of Theorem 2 in practical application will be a matter of future research.

Sparsity constraints We consider as in Example 1 the set C of 1-sparse vectors in $\mathbb{R}^2$ and a loss function is $f\colon (x, y) \mapsto (x-1)^2 + y^2$ whose global minimum on the constraint set is $x = 1, y = 0$. We depict in Figure 1 the sequence generated by the projected gradient algorithm in (5) for various step sizes and initializations, representing both the gradient and the projection steps explicitly. The point $(0, 0)$ is critical but not Fréchet stationary, none of the three sequences converges to this point. Instead, they all converge to the global minimum, illustrating the result of Theorem 2.

Nonlinear optimization on a grid We consider the problem of minimizing a convex quadratic function, where the constraint set is a regular grid in $\mathbb{R}^2$. In this setting, all feasible points are local minimizers, hence Fréchet critical. Yet Theorem 2 predicts that not all of them are attractors of the projected gradient algorithm. We illustrate this with several projected gradient sequences in Figure 1 displaying explicitely the points which do not satisfy the quantitative estimate (5) (with factor $\frac{1}{2}$ for convex functions, see Remark 2). The sequences stop when they reach these stationary points as predicted by Theorem 2.

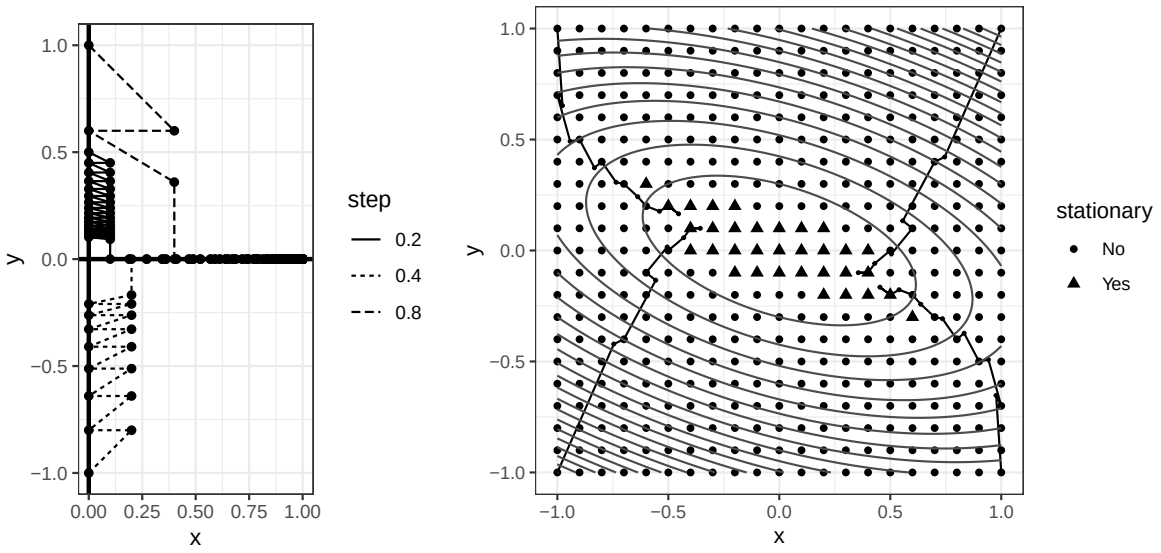

Figure 1: Left: Illustration of the avoidance of the non Clarke regular point of Example 1, the constraint set is depicted by the thick black lines and the thiner lines display several projected gradient sequences with different step-sizes. Right: Avoidance of bad local minima. The feasible set is a grid and the contour of the convex quadratic objective is displayed. Every feasible point is a local minimum and the shape of the point indicate those which satisfies the quantitative estimate (6) (with an additional factor $\frac{1}{2}$ from Remark 2). We display several projected gradient sequences which all stop at the points satisfying these estimates.

3 The proximal gradient algorithm

In this section we provide a general result for the proximal gradient algorithm, from which Theorem 2 follows. We first recall the necessary notations and concepts. This section can be seen of independent interest.

3.1 Technical results from nonmooth analysis

The following extends the notion of gradient in a natural way.

Definition 3 (Regular subdifferential)**.** *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R} \cup \{+\infty\}$ and consider $x \in \mathbb{R}^p$ such that $f(x) < +\infty$. Then $v \in \hat{\partial} f(x)$ if*

$$f(y) \geq f(x) + \langle v, y - x\rangle + o(\|y - x\|).$$

This notation means that $\liminf_{y \to x} \frac{f(y) - f(x) - \langle v, y - x\rangle}{\|y - x\|} \geq 0$.

We obtain an optimality condition as a consequence of the definition in [50, Theorem 10.1].

Theorem 3 (Fermat Rule)**.** *If $x \in \mathbb{R}^p$ is a local minimizer of a lower semicontinuous function $f\colon \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, then $0 \in \hat{\partial} f(x)$.*

Conversely, a point $x \in \mathbb{R}^p$ with $f(x)$ finite satisfying $0 \in \hat{\partial} f(x)$ has non-negative first order variations around x, in the sense that $f(y) - f(x) \geq o(\|y - x\|)$. Such a point is called Fréchet critical. While calculus is in general out of scope for this type of object, it is possible to obtain sum rules when combined with a C^1 function.

Lemma 2 (Smooth sum rule)**.** *Let $g\colon \mathbb{R}^p \mapsto \mathbb{R} \cup \{+\infty\}$ be lower semicontinuous and consider $x \in \mathbb{R}^p$ such that $g(x) < +\infty$. Let $f\colon \mathbb{R}^p \to \mathbb{R}$ be C^1, then $\hat{\partial}(f+g)(x) = \hat{\partial}g(x) + \nabla f(x)$.*

Proof : From [50, Corollary 10.9] we have $\hat{\partial}(f+g)(x) \supset \hat{\partial}g(x) + \nabla f(x)$. Let us prove the reverse inclusion. Choose $v \in \hat{\partial}(f+g)(x)$, we have by Definition 3 and continuous differentiability.

$$\liminf_{y \to x} \frac{f(y) + g(y) - f(x) - g(x) - \langle v, y - x\rangle}{\|y - x\|} \geq 0$$
$$\lim_{y \to x} \frac{f(y) - f(x) - \langle \nabla f(x), y - x\rangle}{\|y - x\|} = 0.$$

We deduce by a substraction that

$$\liminf_{y \to x} \frac{g(y) - g(x) - \langle v - \nabla f(x), y - x\rangle}{\|y - x\|} \geq 0,$$

which shows that $v - \nabla f(x) \in \hat{\partial}g(x)$ which is the desired result. □

3.2 The proximal gradient algorithm and Fréchet stationarity

Given a lower-semi continuous function $g\colon \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$, the proximity operator of g is defined as the possibly empty valued mapping

$$\mathrm{prox}_g(x) = \arg\min_{y \in \mathbb{R}^p} g(y) + \frac{1}{2}\|y - x\|^2.$$

The following Lemma provides a sufficient condition for prox_g to be well behaved. This is [50, Theorem 1.25], we provide a short proof for completeness.

Lemma 3. *Let $g\colon \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be lower semicontinuous, finite at least at one point, such that $g + \frac{1+\delta}{2}\|\cdot\|^2$ is bounded below for some $\delta > 0$. Then $\mathrm{prox}_g\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ has non-empty values, is locally bounded and upper semi-continuous, in the sense that for any converging sequences $y_k \in \mathrm{prox}_g(x_k)$, $k \in \mathbb{N}$, $x_k \to x$, $y_k \to y$, we have $y \in \mathrm{prox}_g(x)$.*

Proof : By assumption, there is no escape at infinity, the prox operation is compact valued and locally bounded. Let $(x_k)_{k\in\mathbb{N}}$ and $(y_k)_{k\in\mathbb{N}}$ be sequences such that $y_k \in \mathrm{prox}_g(x_k)$ for all $k \in \mathbb{N}$, and $x_k \to x$, $y_k \to y$ as $k \to \infty$. For any $z \in \mathbb{R}^p$, and any $k \in \mathbb{N}$, we have

$$g(z) + \frac{1}{2}\|z - x_k\|^2 \geq g(y_k) + \frac{1}{2}\|y_k - x_k\|^2.$$

Hence for any $z \in \mathbb{R}^p$, we have by lower semi-continuity

$$g(z) + \frac{1}{2}\|z - x_k\|^2 \geq \liminf_{k\to\infty} g(y_k) + \frac{1}{2}\|y_k - x_k\|^2 \geq g(y) + \frac{1}{2}\|y - x\|^2$$

which is what we wanted to prove. □

We now state the main result of this section.

Theorem 4. *Let $f\colon \mathbb{R}^p \to \mathbb{R}$ be C^1 with $1-\delta$ Lipschitz gradient for some $\delta \in (0,1)$ and g be as in Lemma 3. Fix $x_0 \in \mathbb{R}^p$; and consider the recursion*

$$x_{k+1} = \mathrm{prox}_g(x_k - \nabla f(x_k)). \tag{7}$$

Any accumulation point $\bar{x}$ of $(x_k)_{k\in\mathbb{N}}$ are Fréchet critical for $f+g$ such that

$$\bar{x} \in \mathrm{prox}_g(\bar{x} - \nabla f(\bar{x})) \tag{8}$$
$$f(y) + g(y) \geq f(\bar{x}) + g(\bar{x}) - \|y - \bar{x}\|^2, \qquad \forall y \in \mathbb{R}^p.$$

Furthermore, $\mathrm{dist}(-\nabla f(x_{k+1}), \hat{\partial}g(x_{k+1})) \to 0$ *as* $k \to \infty$.

Proof : One can check that $x_{k+1} \in \arg\min_y f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{1}{2}\|y - x_k\|^2 + g(y)$ by completing the square.

Combining with the descent lemma for Lipschitz gradient functions[41, Lemma 1.2.3], we have

$$\begin{aligned} f(x_k) + g(x_k) &\geq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{1}{2}\|x_{k+1} - x_k\|^2 + g(x_{k+1}) \\ &= f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{1-\delta}{2}\|x_{k+1} - x_k\|^2 + g(x_{k+1}) + \frac{\delta}{2}\|x_{k+1} - x_k\|^2 \\ &\geq f(x_{k+1}) + g(x_{k+1}) + \frac{\delta}{2}\|x_{k+1} - x_k\|^2. \end{aligned}$$

Now suppose that the sequence $(x_k)_{k\in\mathbb{N}}$ has an accumulation point $\bar{x}$. In this case $f(x_k)+g(x_k)$ is decreasing, and it converges to a finite value. Therefore the increments $x_{k+1} - x_k$ tend to 0 and $\mathrm{prox}_g(x_k - \nabla f(x_k))$ also tends to $\bar{x}$. Using Lemma 3, we have that $\bar{x} \in \mathrm{prox}_g(\bar{x} - \nabla f(\bar{x}))$ so that, using Fermat rule in Theorem 3 and Lemma 2

$$\begin{aligned} &\bar{x} \in \arg\min_y g(y) + \frac{1}{2}\|y - \bar{x} + \nabla f(\bar{x})\|^2 \\ &0 \in \hat{\partial}\left(g(y) + \frac{1}{2}\|y - \bar{x} + \nabla f(\bar{x})\|^2\right)_{y=\bar{x}} = \hat{\partial} g(\bar{x}) + \nabla f(\bar{x}) \end{aligned} \tag{9}$$

which is the Fréchet stationarity. This actually ensures that $-\nabla f(\bar{x})$ is a proximal subgradient of g, and using [50, Proposition 8.46] and the descent Lemma, for all $y \in \mathbb{R}^p$,

$$\begin{aligned} g(y) &\geq g(\bar{x}) + \langle -\nabla f(\bar{x}), y - \bar{x}\rangle - \frac{1}{2}\|y - \bar{x}\|^2 \\ f(y) &\geq f(\bar{x}) + \langle \nabla f(\bar{x}), y - \bar{x}\rangle - \frac{1}{2}\|y - \bar{x}\|^2 \\ f(y) + g(y) &\geq f(\bar{x}) + g(\bar{x}) - \|y - \bar{x}\|^2. \end{aligned} \tag{10}$$

For the last point, using Fermat rule in Theorem 3 for the prox operator leads to

$$\begin{aligned} x_{k+1} - x_k + \nabla f(x_k) &= x_{k+1} - x_k + \nabla f(x_k) - \nabla f(x_{k+1}) + \nabla f(x_{k+1}) \\ &\in -\hat{\partial} g(x_{k+1}) \end{aligned}$$

so that

$$\mathrm{dist}(-\nabla f(x_{k+1}), \hat{\partial} g(x_{k+1})) \leq \|x_{k+1} - x_k\| + \|\nabla f(x_k) - \nabla f(x_{k+1})\| \underset{k\to\infty}{\to} 0,$$

which is the second result. □

Remark 3 (Comments on Theorem 4)**.** *If in addition, the function f and the set C are assumed to be semi-algebraic, then the sequence actually converges [3]. The quadratic lower bound provides a quantitative estimate of Fréchet stationarity. Furthermore, if f is convex, then one can add a factor $\frac{1}{2}$ in front of the quadratic term in* (8)*, since the inequality* (10) *can be tightened.*

4 Conclusion

Our main results in Theorem 1 and Theorem 2 ensure that the lack of Clarke regularity has minimal effect on the optimality conditions in nonconvex constrained optimization, and that it does not affect the projected gradient algorithm. A by product of the analysis is a global quantitative estimate for Fréchet stationarity of accumulation points of the projected gradient algorithm with a strong variational interpretation and a natural connection to the convex setting

where the negative quadratic term vanishes. This generalizes the analysis of Iterative Hard Thresholding in [4] and illustrates the fact that the projected gradient algorithms constitutes a strong baseline in light of the observations made in [37]. Future work will be dedicated to the exploration of the consequences of this observation. Finally it is a natural to ask if this type of favorable property would extend to different proximal decomposition algorithms in a nonconvex setting, such as alternating methods or momentum methods.

Acknowledgements

The authors acknowledge the support of the AI Interdisciplinary Institute ANITI funding, through the French "Investments for the Future – PIA3" program under the grant agreement ANR-19-PI3A0004, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA8655-22-1-7012, ANR Chess (ANR-17-EURE-0010), ANR Regulia and ANR Bonsai.

References

[1] Attouch, H., and Bolte, J. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming 116* (2009), 5–16.

[2] Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of operations research 35*, 2 (2010), 438–457.

[3] Attouch, H., Bolte, J., and Svaiter, B. F. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming 137*, 1-2 (2013), 91–129.

[4] Beck, A., and Eldar, Y. C. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization 23*, 3 (2013), 1480–1509.

[5] Beck, A., and Hallak, N. On the minimization over sparse symmetric sets: projections, optimality conditions, and algorithms. *Mathematics of Operations Research 41*, 1 (2016), 196–223.

[6] Bertsekas, D. P. On the goldstein-levitin-polyak gradient projection method. *IEEE Transactions on automatic control 21*, 2 (1976), 174–184.

[7] Bianchi, P., Hachem, W., and Schechtman, S. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis 30*, 3 (2022), 1117–1147.

[8] Bianchi, P., Hachem, W., and Schechtman, S. Stochastic subgradient descent escapes active strict saddles on weakly convex functions. *Mathematics of Operations Research* (2023).

[9] Bolte, J., Boustany, R., Pauwels, E., and Pesquet-Popescu, B. Nonsmooth automatic differentiation: a cheap gradient principle and other complexity results. In *International Conference on Learning Representations* (2023).

[10] Bolte, J., Daniilidis, A., and Lewis, A. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization 17*, 4 (2007), 1205–1223.

[11] Bolte, J., Daniilidis, A., and Lewis, A. Tame functions are semismooth. *Mathematical Programming 117*, 1-2 (2009), 5–19.

[12] Bolte, J., Daniilidis, A., Lewis, A., and Shiota, M. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization 18*, 2 (2007), 556–572.

[13] Bolte, J., Daniilidis, A., and Lewis, A. S. Generic optimality conditions for semialgebraic convex programs. *Mathematics of Operations Research 36*, 1 (2011), 55–70.

[14] Bolte, J., Hochart, A., and Pauwels, E. Qualification conditions in semi-algebraic programming. *SIAM journal on Optimization 28*, 2 (2018), 1867–1891.

[15] Bolte, J., and Pauwels, E. Majorization-minimization procedures and convergence of sqp methods for semi-algebraic and tame programs. *Mathematics of Operations Research 41*, 2 (2016), 442–465.

[16] Bolte, J., and Pauwels, E. A mathematical model for automatic differentiation in machine learning. In *Advances in Neural Information Processing Systems* (2020).

[17] Bolte, J., and Pauwels, E. Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming 188*, 1 (2021), 19–51.

[18] Bolte, J., Sabach, S., and Teboulle, M. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming 146*, 1-2 (2014), 459–494.

[19] Calamai, P. H., and Moré, J. J. Projected gradient methods for linearly constrained problems. *Mathematical programming 39*, 1 (1987), 93–116.

[20] Coste, M. *An introduction to o-minimal geometry*. Istituti editoriali e poligrafici internazionali Pisa, 2000.

[21] Coste, M. An introduction to semialgebraic geometry, 2000.

[22] Daniilidis, A., Bolte, J., and Lewis, A. Generic identifiability and second-order sufficiency in tame convex optimization. *Mathematics of Operations Research* (2009), 1–30.

[23] Daniilidis, A., and Pang, J. C. Continuity and differentiability of set-valued maps revisited in the light of tame geometry. *Journal of the London Mathematical Society 83*, 3 (2011), 637–658.

[24] Davis, D., and Drusvyatskiy, D. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics 22*, 2 (2022), 561–606.

[25] Davis, D., Drusvyatskiy, D., Kakade, S., and Lee, J. D. Stochastic subgradient method converges on tame functions. *Foundations of computational mathematics 20*, 1 (2020), 119–154.

[26] Davis, D., and Jiang, L. A nearly linearly convergent first-order method for nonsmooth functions with quadratic growth. *arXiv preprint arXiv:2205.00064* (2022).

[27] Drusvyatskiy, D., Ioffe, A. D., and Lewis, A. S. Generic minimizing behavior in semialgebraic optimization. *SIAM Journal on Optimization 26*, 1 (2016), 513–534.

[28] Dunn, J. C. Global and asymptotic convergence rate estimates for a class of projected gradient processes. *SIAM Journal on Control and Optimization 19*, 3 (1981), 368–400.

[29] Dunn, J. C. On the convergence of projected gradient processes to singular critical points. *Journal of Optimization Theory and Applications 55* (1987), 203–216.

[30] Goldstein, A. Convex programming in hilbert space. *Bulletin of the American Mathematical Society 70*, 5 (1964), 709–710.

[31] Hosseini, S., Luke, D. R., and Uschmajew, A. Tangent and normal cones for low-rank matrices. *Nonsmooth optimization and its applications* (2019), 45–53.

[32] Hou, T. Y., Li, Z., and Zhang, Z. Asymptotic escape of spurious critical points on the low-rank matrix manifold. *arXiv preprint arXiv:2107.09207* (2021).

[33] Ioffe, A. D. An invitation to tame optimization. *SIAM Journal on Optimization 19*, 4 (2009), 1894–1917.

[34] Kaplan, A., and Tichatschke, R. Proximal point methods and nonconvex optimization. *Journal of global Optimization 13* (1998), 389–406.

[35] Lee, G. M., and Pham, T. S. Generic properties for semialgebraic programs. *SIAM Journal on Optimization 27*, 3 (2017), 2061–2084.

[36] Lee, J. M. *Introduction to Smooth Manifolds*. Springer, 2012.

[37] Levin, E., Kileel, J., and Boumal, N. Finding stationary points on bounded-rank matrices: A geometric hurdle and a smooth remedy. *Mathematical Programming 199*, 1-2 (2023), 831–864.

[38] Levitin, E. S., and Polyak, B. T. Constrained minimization methods. *USSR Computational mathematics and mathematical physics 6*, 5 (1966), 1–50.

[39] Luke, D. R. Prox-regularity of rank constraint sets and implications for algorithms. *Journal of Mathematical Imaging and Vision 47* (2013), 231–238.

[40] Moreau, J.-J. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France 93* (1965), 273–299.

[41] Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media, 2003.

[42] Nie, J. Optimality conditions and finite convergence of lasserre's hierarchy. *Mathematical programming 146* (2014), 97–121.

[43] Olikier, G., and Absil, P.-A. An apocalypse-free first-order low-rank optimization algorithm with at most one rank reduction attempt per iteration. *SIAM Journal on Matrix Analysis and Applications 44*, 3 (2023), 1421–1435.

[44] Olikier, G., Gallivan, K. A., and Absil, P.-A. An apocalypse-free first-order low-rank optimization algorithm. *arXiv preprint arXiv:2201.03962* (2022).

[45] Olikier, G., Gallivan, K. A., and Absil, P.-A. First-order optimization on stratified sets. *arXiv preprint arXiv:2303.16040* (2023).

[46] Olikier, G., Uschmajew, A., and Vandereycken, B. Gauss-southwell type descent methods for low-rank matrix optimization. *arXiv preprint arXiv:2306.00897* (2023).

[47] Oxtoby, J. C. *Measure and category*, vol. 2. Springer Science & Business Media, 1971.

[48] Pauwels, E. The value function approach to convergence analysis in composite optimization. *Operations Research Letters 44*, 6 (2016), 790–795.

[49] Pham, T. S., and Vui, H. H. *Genericity in polynomial optimization*, vol. 3. World Scientific, 2016.

[50] Rockafellar, R. T., and Wets, R. J.-B. *Variational analysis*, vol. 317. Springer Science & Business Media, 1998.

[51] Spingarn, J. E. Submonotone mappings and the proximal point algorithm. *Numerical Functional Analysis and Optimization 4*, 2 (1982), 123–150.

[52] Van den Dries, L. *Tame topology and o-minimal structures*, vol. 248. Cambridge university press, 1998.

[53] Van den Dries, L., and Miller, C. Geometric categories and o-minimal structures.