



The ridge method for tame min-max problems

Edouard Pauwels

► To cite this version:

| Edouard Pauwels. The ridge method for tame min-max problems. 2022. hal-03186676v2

HAL Id: hal-03186676

<https://hal.science/hal-03186676v2>

Preprint submitted on 4 Feb 2022 (v2), last revised 23 Jun 2023 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The ridge method for tame min-max problems

Edouard Pauwels*

February 4, 2022

Abstract

We study the ridge method for min-max problems, and investigate its convergence without any convexity, differentiability or qualification assumption. The central issue is to determine whether the “parametric optimality formula” provides a conservative gradient, a notion of generalized derivative well suited for optimization. The answer to this question is positive in a semi-algebraic, and more generally definable, context. As a consequence, the ridge method applied to definable objectives is proved to have a minimizing behavior and to converge to a set of equilibria which satisfy an optimality condition. Definability is key to our proof: we show that for a more general class of nonsmooth functions, conservativity of the parametric optimality formula may fail, resulting in an absurd behavior of the ridge method.

Keywords. min-max problems, ridge algorithm, parametric optimality, conservative gradients, definable sets, o-minimal structures, Clarke subdifferential, First order methods

1 Introduction

1.1 Main result

We consider unconstrained minimization of an objective function:

$$f: x \rightarrow \max_{y \in \mathbb{R}^r} F(x, y), \quad (1)$$

where $F: \mathbb{R}^p \times \mathbb{R}^r \rightarrow \mathbb{R}$ is locally Lipschitz and achieves its maximum¹ in y for all $x \in \mathbb{R}^p$. The function f will be called the value function, note that in this situation, f is also locally Lipschitz. These notations and assumptions will be standing throughout this paper. We consider the Ridge Method (RM)², initialized with $x_0 \in \mathbb{R}^p$, it is defined recursively as

*IRIT, Université de Toulouse, CNRS. France.

¹The maximum in (1) is arbitrary and could be reversed into a minimum.

²We are not aware of detailed description of this method in the litterature, although the idea existed in discussions. We chose the name “Ridge Method” from oral transmission.

follows, for all $k \in \mathbb{N}$,

$$\begin{aligned} y_k &\in \arg \max_{y \in \mathbb{R}^r} F(x_k, y) \\ (u_k, 0) &\in \partial^c F(x_k, y_k) \\ x_{k+1} &= x_k - \alpha_k u_k, \end{aligned} \tag{RM}$$

where $(\alpha_k)_{k \in \mathbb{N}}$ is a nonsummable sequence of positive step sizes tending to 0 and ∂^c denotes the Clarke subdifferential [20]. The existence of the update direction u_k in (RM) is ensured by the Parametric Optimality (PO) formula for subgradient of partial maxima. The algorithm relies on the knowledge of

- **A partial maximization oracle**, which associates to any $x \in \mathbb{R}^p$, an element of the set $P(x) := \arg \max_{y \in \mathbb{R}^r} F(x, y)$, assumed to be nonempty.
- **A first order oracle**, which associates to any $(x, y) \in \mathbb{R}^p \times \mathbb{R}^r$ first order information about F in the form of its Clarke subdifferential, $\partial^c F$.

For most locally Lipschitz functions F , the Clarke subdifferential, $\partial^c F$ [20] carries absolutely no information about the function itself and is therefore useless from a computational perspective [43, 17, 18]. For our purpose, we need to restrict F to be in a subclass which is well behaved with respect to subdifferentiation, we choose the class of path differentiable functions, which was identified by several authors to be well behaved in terms of subgradient differential inclusion [42, 16, 23, 14].

The purpose of this paper is to investigate asymptotic behavior of (RM). The main results are the following.

- When applied to a large and widespread subclass of functions F , for example semi-algebraic functions, and more generally definable functions, Algorithm (RM) has a minimizing behavior. For bounded sequences, the value function, $f(x_k)$, is converging and accumulation points, $(\bar{x}, \bar{y})$, of the sequence $(x_k, y_k)_{k \in \mathbb{N}}$ are equilibria which satisfy the optimality condition:

$$\begin{aligned} 0 &\in \text{conv}\{u, (u, 0) \in \partial^c F(\bar{x}, y), y \in \arg \max_{z \in \mathbb{R}^r} F(\bar{x}, z)\} \\ \bar{y} &\in \arg \max_{y \in \mathbb{R}^r} F(\bar{x}, y) \end{aligned}$$

Let us stress that most objectives found in applications are definable, see [7, 13] for discussion and examples, and [15, 14] for a recent account in deep learning.

- Without definability assumption, for general path-differentiable functions, the algorithm may fail to have any minimizing property. We construct a Lipschitz function F which is path-differentiable, such that all inputs x are steady states for (RM) but not critical in any reasonable sense for f . This underlines the importance of the definability assumption in the previous result and shows that Algorithm (RM) requires to work with proper subclasses, beyond Lipschicity and path-differentiability.

1.2 Parametric optimality and nonsmooth differential calculus

In order to analyse algorithm (RM), we need a variational model for parametric optimality and calculus rules providing access to first order information for f , from the knowledge of the partial maximization oracle and the subgradient of F . Our main candidate is the parametric optimality formula, which was described in [37, Corollary 3.1.1] for partial minimization, see also [38, Theorem 10.13]. Clarke subdifferential is non directional, and therefore, the formula is also valid for partial maximization. Under our setting, it ensures that for all $x \in \mathbb{R}^p$

$$\partial^c f(x) \subset \text{conv}\{u, (u, 0) \in \partial^c F(x, y), y \in P(x)\}. \quad (\text{PO})$$

The formula is in fact a result describing the subgradient of f , but we will be interested mostly in its right hand side which will be referred to as the parametric optimality formula (PO formula)³. Indeed, any element given by the PO formula can be computed from the knowledge of maximization and first order oracles mentioned above and precisely corresponds to the search direction chosen in Algorithm (RM).

As noted in [38, Theorem 10.13], the PO formula is sharp, it holds with equality in the case of a concave F (jointly in (x, y)). However in general nonconvex settings, without further qualification assumptions, the PO formula does not hold with equality⁴. Therefore, the PO formula does not necessarily provide a subgradient, this constitutes the main difficulty in analyzing Algorithm (RM). We choose to use conservativity, a notion of generalized derivative which was recently introduced as a nonsmooth analysis tool which is compatible with differential calculus [14]. Most importantly, conservative gradients may be used in place of subgradients for first order optimization [19, 14, 15], making them a natural candidate for algorithmic oracles in our context. With this in mind, the proposed analysis of Algorithm (RM) boils down to the central question:

For a path-differentiable F , does the PO formula define a conservative gradient for f ?

The answer to this question is negative in general, we provide a counterexample. However for definable functions the answer turns out to be positive. The proof of the latter result relies on a characterization of definable conservative gradients based only on definable paths, which is of independent interest (see also remark 2 and [24]). In the context of conservativity, the definable case plays a special role as it is widespread in applications [14, 15] and many further properties are available [15, 32, 24]. The reader unfamiliar

³**Envelope formula:** the PO formula is very similar to the envelope formula: $\partial^c f(x) \subset \text{conv}\{u \in \partial_x^c F(x, y), y \in P(x)\}$, [20, Theorem 2.8.2] where ∂_x^c denotes the subgradient for fixed y . This formula holds with equality if F is convex in x , see [11, Proposition A.22][25, 22], and more generally regular [20, Theorem 2.8.2]. This is simpler and computationally more advantageous than the PO formula. However, it very much depends on convexity and is too coarse in general for our purposes, consider for example $F(x, y) = -|x - y|$, the envelope formula gives $[-1, 1]$ for all x and $f(x) = \max_{y \in \mathbb{R}} F(x, y) = 0$ for all x .

⁴**Failure of partial optimality formula:** Consider the function $F: (x, y) \mapsto \{-y \min\{|x|, 1\} + \min\{0, y\}\}$, which is concave in y . We have for all $x \in \mathbb{R}$, $f(x) = \max_{y \in \mathbb{R}} F(x, y) = 0$. Yet fixing $x = 0$, we have $F(0, y) = 0$ for all $y > 0$, all such y being partial maximizers. However $([-y, y], 0) \in \partial^c F(0, y)$ for any $y > 0$. This shows that any $s \in \mathbb{R}$ is compatible with the PO formula at $x = 0$, yet the corresponding Clarke subdifferential is only the singleton $\{0\}$, which coincides with the classical derivative.

with definability may consider instead semialgebraicity, which is a special case, a function being semialgebraic when its graph can be represented as the finite union of solution sets of polynomial systems involving finitely many equalities and inequalities. Section 3 exposes basic definitions and more details regarding definability.

Let us describe a few applications and consequences of these results:

- If F is given in the form of a composition of elementary definable functions, Lipschitz and concave in its second argument, then the maximization oracle could be given by an optimization solver which addresses the maximization problem to global optimality, and the subgradient oracle could be given by algorithmic differentiation (see remark 3). We ensure that the ridge method (RM) is attracted by stationary points in this case and actually most sequences are attracted by Clarke critical points [15, 9].
- In the previous example, concavity is instrumental, to ensure that global maximization is reasonable. Actually, the only requirement is to have a global maximization oracle, for which we could consider global optimization examples beyond concavity.
- In both cases, the (PO) formula evaluated at x is a singleton equal to $\nabla f(x)$, the classical gradient, everywhere outside of a finite union of differentiable manifolds of dimension strictly less than p (a negligible set). This is a consequence of [15, Theorem 1] and illustrates the fact that, although the PO formula may produce artifacts, it occurs very rarely.
- Finally, conservative gradients satisfy Fermat rule, using Carathéodory theorem, this translates in an optimality condition as follows: if $x \in \mathbb{R}^p$ is a local minimum of f , then there exists $(y_i, u_i, \lambda_i)_{i=1}^{p+1}$ such that

$$\begin{aligned} y_i \in \arg \max_{y \in \mathbb{R}^r} F(x, y), \quad (u_i, 0) \in \partial^c F(x, y_i), \quad \lambda_i \in [0, 1], \quad i = 1, \dots, p+1 \\ \sum_{i=1}^{p+1} \lambda_i u_i = 0, \quad \sum_{i=1}^{p+1} \lambda_i = 1. \end{aligned} \tag{2}$$

This constitutes a notion of stationarity for nonsmooth min-max problems which could be a target for future algorithmic developments.

1.3 Applications of the main results and existing literature

Min-max problems arise in machine learning applications with Generative Adversarial Networks (GANs) [28, 6], adversarial training of deep networks [40, 29], and further applications [1] bolstering research on algorithms for min-max problems, see for example [44] for a ridge-type method. Most of these applications amount to solving nonsmooth min-max problems, with non convex or concave definable objectives, based on first order methods and algorithmic differentiation oracles. Our results describe a notion of equilibrium for such problem as well as an algorithm to reach such equilibria, under assumptions which are general enough to encompass most machine learning applications [15].

Our main results are actually stated in terms of conservative gradients, a notion which has been shown to be compatible with the rules of differential calculus, contrary to the notion of subdifferential [14, 15]. Beyond min-max problems, the fact that the PO formula defines a conservative gradient can be used in compositional modeling, in combination with algorithmic differentiation. This is in close connection with emerging extensions of deep neural networks which include optimization problems within their formulation, some of the network layers being defined as partial maxima or minima [5, 2, 10]. Conservativity of the PO formula provides theoretical ground to develop algorithmic differentiation tools for compositional problems involving such max structured functions.

Stationarity for min-max problem is an important research topic, a large litterature is devoted to convex-concave min-max problems, which we will not describe as we consider a much broader class. In the nonconvex case, it is particularly interesting to restrict F to be twice continuously differentiable and strongly convex in y for each fixed x , as the value function f and the argmax mapping are differentiable in this case (using for example the implicit function theorem). Closer to our interest is the situation where the maximization problem is merely convex in y for each fixed x . In this case, even if F is smooth, the value function f may not be differentiable. The work of [35] consider min-max structured problem with F concave in y and weakly convex in x . They observe that in this case the value function is weakly convex and propose a dedicated algorithm. It turns out that the envelope formula remains valid with equality for such problems (see for example [33, Lemma 4.7] which is a special case of [20, Theorem 2.8.2]). Since functions with Lipschitz gradient are regular (actually weakly convex) these observations allowed to develop a variety of algorithms and analyses for twice differentiable F , concave in its second argument. Most proposed approach rely on equality in the envelope formula and the Moreau envelope of the value function [34, 31, 41, 33, 30, 35]. These provide a precise picture regarding stationarity and algorithms for smooth min max problems with concave maximization component. The present work departs from this litterature because it relies neither on smoothness nor on a form of concavity or weak convexity and the obtained results are only qualitative.

2 Presentation of the main results

2.1 Technical preliminary

For any integer $p \in \mathbb{N}$, we use the following notations. We denote by $\langle \cdot, \cdot \rangle$ be the canonical Euclidean scalar product on $\mathbb{R}^p$ and $\|\cdot\|$ its associated norm. A locally Lipschitz continuous function, $f: \mathbb{R}^p \rightarrow \mathbb{R}$ is differentiable almost everywhere by Rademacher's theorem, see for example [27]. Denote by $R \subset \mathbb{R}^p$, the full measure set where f is differentiable, then the Clarke subdifferential [20] of f is given for any $x \in \mathbb{R}^p$, by

$$\partial^c f(x) = \text{conv} \left\{ v \in \mathbb{R}^p, \exists y_k \xrightarrow[k \rightarrow \infty]{} x \text{ with } y_k \in R, v_k = \nabla f(y_k) \xrightarrow[k \rightarrow \infty]{} v \right\}.$$

A set valued map $D: \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ is a function from $\mathbb{R}^p$ to the set of subsets of $\mathbb{R}^q$. The graph of D is given by

$$\text{graph } D = \{(x, z) : x \in \mathbb{R}^p, z \in D(x)\}.$$

D is said to have *closed graph* or to be *graph closed* if $\text{graph } D$ is closed as a subset of $\mathbb{R}^{p+q}$. An equivalent characterization is that for any converging sequences $(x_k)_{k \in \mathbb{N}}$, $(v_k)_{k \in \mathbb{N}}$ in $\mathbb{R}^p$, with $v_k \in D(x_k)$ for all $k \in \mathbb{N}$, we have

$$\lim_{k \rightarrow \infty} v_k \in D(\lim_{k \rightarrow \infty} x_k).$$

D is said to be locally bounded if for each compact $K \subset \mathbb{R}^p$, there is $M > 0$ such that $\|v\| \leq M$ for all $v \in D(x)$ for all $x \in K$. An *absolutely continuous curve* is a continuous function $x: \mathbb{R} \rightarrow \mathbb{R}^p$ which admits a derivative $\dot{x}$ for Lebesgue almost all $t \in \mathbb{R}$, (in which case $\dot{x}$ is Lebesgue measurable), and $x(t) - x(0)$ is the Lebesgue integral of $\dot{x}$ between 0 and t for all $t \in \mathbb{R}$.

These elements allow to define the notion conservativity of set valued mappings [14].

Definition 1 (Conservative gradients) *Let $D: \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a set valued map with closed graph, non empty and locally bounded values and $f: \mathbb{R}^p \rightarrow \mathbb{R}$ a locally Lipschitz function. Then f is a potential for D if for all $x \in \mathbb{R}^p$, all $\gamma: [0, 1] \rightarrow \mathbb{R}^p$, absolutely continuous with $\gamma(0) = 0$ and $\gamma(1) = x$, and all measurable functions, $v: [0, 1] \rightarrow \mathbb{R}^p$, such that $v(t) \in D(\gamma(t))$ for all $t \in [0, 1]$,*

$$f(x) = f(0) + \int_0^1 \langle \dot{\gamma}(t), v(t) \rangle dt. \quad (3)$$

We shall also say that D is a conservative gradient for f or simply a conservative gradient. Such functions f are called path differentiable. The function v in (3) will be called a measurable selection of $D \circ \gamma$.

The result of [14, Corollary 1] ensures that for a path differentiable f , $\partial^c f$ is a conservative gradient.

2.2 Characterization of definable conservative gradients

Our main convergence result holds under definability assumptions, we start by showing that conservativity admits a simpler characterization in this context. From now on we fix an o-minimal structure (for example semialgebraic sets, see Section 3 for more details on definability), all definable objects we shall consider are implicitly definable in this structure. Recall that an o-minimal structure is a sequence of families of subsets $(\mathcal{O}_i)_{i \in \mathbb{N}}$ such that for each $i \in \mathbb{N}$, $\mathcal{O}_i$ contains subsets of $\mathbb{R}^i$ and satisfy axiomatic compatibility conditions.

Definition 2 (Definably conservative gradient) *Let $D: \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a set valued field with closed graph, non empty and locally bounded values. Assuming in addition that D is definable, D is called definably conservative if equation (3) holds only for definable C^1 loops γ and definable selections v .*

Following Definition 2, it is obvious that a definable conservative gradient is definably conservative since definable C^1 loops are absolutely continuous and definable selections are measurable. The following result provides a converse, a slightly more general proof is found in Section 3.

Theorem 1 *Let $D: \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be definably conservative, then D is conservative.*

Remark 1 *It is of primary importance in Definition 2, that the definable loops and definable set valued mapping D , are definable in the same o-minimal structure. For example consider the set E , the graph of the exponential function in $\mathbb{R}^2$. By a theorem of Wilkie [46], there exists an o-minimal structure which contains all semi-algebraic sets and such that E is definable in this structure, call it Wilkie's structure. Consider $D: \mathbb{R}^2 \rightrightarrows \mathbb{R}^2$ to be $\{0\}$ outside of E and the unit Euclidean ball on E , it has a closed graph, it is bounded with nonempty values and definable in Wilkie's structure. Consider any differentiable semialgebraic loop $\gamma: [0, 1] \rightarrow \mathbb{R}^2$. Since E is the graph of an analytic function, but is not a semialgebraic set, the intersection of γ and E must contain only finitely many points. Hence γ and D satisfy formula (3), but D is obviously not conservative as it does not satisfy the integral formula along any nontrivial absolutely continuous path which image is in E . Hence the importance of having a unique fixed o-minimal structure throughout the manuscript.*

Remark 2 *In the final stages of completing the first version of this manuscript, we became aware of the concurrent and independent work [24]. Theorems 3.2 and 3.5 in [24] proves equivalence between conservativity and definable conservativity provided that both D and its associated potential f are definable. The proposed result is very similar, it constitutes a minor improvement as it removes the requirement for definability of f and only requires definability of D . The equivalence is a byproduct of the analysis of [24] which contribution is orthogonal and in particular does not relate to parametric optimality and min-max problems.*

This result shows that definable conservativity is equivalent to conservativity in the definable world. Its proof is based on an equivalent characterization of conservativity in this context, variational stratification [12, 14]. The restriction to definable loops and selection in Definition 2 opens the possibility to use all results of o-minimal geometry [26, 21] in order to prove that a given field is conservative. This can be in particular useful to prove conservativity of the PO formula.

2.3 Definable parametric optimality formula

Taking advantage of the strong rigidity of definable objects, we obtain the following result, which proof is stated in Section 3.5.

Theorem 2 *Let $F: \mathbb{R}^p \times \mathbb{R}^r \rightarrow \mathbb{R}$ be locally Lipschitz and definable. Set*

$$\begin{aligned} f: \mathbb{R}^p &\rightarrow \mathbb{R} \\ x &\mapsto \max_{y \in \mathbb{R}^r} F(x, y), \end{aligned}$$

where the argmax is assumed to be locally bounded, call it $P(x) \subset \mathbb{R}^r$. Set

$$D_f: \mathbb{R}^p \rightrightarrows \mathbb{R}^p$$

$$x \mapsto \text{conv} \{u, \exists y \in P(x), (u, 0) \in \partial^c F(x, y)\},$$

then D_f is conservative for f .

This result is the consequence of the more general Theorem 8 stated in Example 1. The result is in fact stated for more general conservative gradients than the Clarke subdifferential, which opens the possibility to use objects defined through other calculus rules in place of subgradients, for example, outputs of algorithmic differentiation [14, 15]. This result implies that the PO formula can be used as a first order optimization oracle in the definable world, as illustrated in the next section.

2.4 Convergence of the ridge method

Getting back to the initial problem

$$\min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^r} F(x, y)$$

where $F: \mathbb{R}^p \times \mathbb{R}^r \rightarrow \mathbb{R}$ is locally Lipschitz and, in addition, definable. Assume that the mapping $x \rightrightarrows P(x) = \arg \max_{y \in \mathbb{R}^r} F(x, y)$ is nonempty and locally bounded. Consider the ridge algorithm, set $x_0 \in \mathbb{R}^p$ and iterate for $k \in \mathbb{N}$

$$y_k \in \arg \max_{y \in \mathbb{R}^r} F(x_k, y)$$

$$(u_k, 0) \in \partial^c F(x_k, y_k)$$

$$x_{k+1} = x_k - \alpha_k u_k$$

The following results ensures under mild assumptions that accumulation points $(\bar{x}, \bar{y})$ of $(x_k, y_k)_{k \in \mathbb{N}}$ are PO critical points for f such that

$$0 \in \text{conv} \{u, (u, 0) \in \partial^c F(\bar{x}, y), y \in \arg \max_{z \in \mathbb{R}^r} F(\bar{x}, z)\} \quad (4)$$

$$\bar{y} \in \arg \max_{y \in \mathbb{R}^r} F(\bar{x}, y)$$

Theorem 3 *Let $(\alpha_k)_{k \in \mathbb{N}}$ be a non summable sequence of positive step sizes tending to zero and assume that $(x_k)_{k \in \mathbb{N}}$ is bounded. Then $F(x_k, y_k)$ converges, and all accumulation points $(\bar{x}, \bar{y})$ of $(x_k, y_k)_{k \in \mathbb{N}}$ satisfy (4)*

Proof : This is a consequence of the results of [8], we sketch the proof arguments.

The sequence defined by the (RM) algorithm satisfies for all k , $u_k \in D_f(x_k)$ where D_f is a convex valued conservative gradient for f as described in Theorem 2. Furthermore, condition (4) can be equivalently read as $0 \in D_f(x)$, see also (2).

Convex valued conservative gradients can be used in place of subgradients in first order method, while allowing to deploy the general method of [8] for convergence analysis, we describe the main elements using the concepts and vocabulary in [8].

- [8, Proposition 1.3] states that the recursion (RM) produces a perturbed solution to the differential inclusion $\dot{x} \in -D_f(x)$. Indeed, we can apply definition III with a deterministic null perturbation term so that we meet step size requirements.
- [8, Theorem 4.2] ensures that perturbed solutions satisfy [8, Theorem 4.1 (ii)] and [8, Theorem 4.3] implies that limit sets of such solutions have to be internally chain transitive.
- Since D_f conservative for f , f is Lyapunov for the differential inclusion, and since both are definable, we have the Morse-Sard condition: $\{f(x), 0 \in D_f(x)\}$ is finite. Using [Proposition 3.27][8] we obtain that limit sets of the recurrence is contained in PO critical solutions (4) and that f is constant on this limit set, which is our result.

□

Remark 3 *The same result holds mutatis mutandis with a definable conservative gradient D in place of $\partial^c F$, for example one obtained by algorithmic differentiation [14]. Similarly minimization and maximization could be interchanged arbitrarily, modulo changes in the step sign.*

Remark 4 *The convex hull in (4) is necessary, for example setting $F(x, y) = xy - 2||y| - 1|$, we have $f(x) = \max_y F(x, y) = |x|$ for all $x \in [-1, 1]$. In this case, the algorithm reduces to subgradient descent and the convex hull is necessary to obtain a valid optimality condition at 0.*

2.5 Failure of parametric optimality formula in general

It was already shown that the PO formula does not necessarily provide elements of the subdifferential (see footnote⁴ on page 3). Yet the failure only occurred at the origin which does not prevent the PO formula to provide a conservative gradient for the value function f . The following result shows that this is not the case in general, its proof is given in Section 4.

Theorem 4 *There exists a Lipschitz path differentiable function $g: \mathbb{R}^2 \rightarrow \mathbb{R}$ such that*

- *for all $x \in [0, 1]$, $\max_{y \in \mathbb{R}} g(x, y) = x$ and the maximum is attained on $[0, 1]$.*
- *for all $x \in [0, 1]$, $0 \in \{v, (v, 0) \in \partial^c g(x, y), y \in \arg \max_{y \in \mathbb{R}} g(x, y)\}$*
- *for countably many $x \in [0, 1]$, $\arg \max_{y \in \mathbb{R}} g(x, y)$ is a pair, for the rest it is a singleton.*

The preceding result shows that the PO formula does not provide a conservative gradient as the value function is the identity on $\mathbb{R}$, but the formula may result in the constant 0, which is not compatible with the integration constraint in Definition 1. It is also obvious

that the ridge algorithm (RM) applied to minimization of $f: x \mapsto \max_{y \in [0,1]} g(x, y)$ based on the PO formula may get stuck at any initialization point $x \in [0, 1]$ since they are all steady states of the algorithm (RM). This illustrates the fact that failure of conservativity entails in this case failure for first order algorithms based on the PO formula. Obviously the function g given in Theorem 4 is not definable in any structure as otherwise Theorem 2 would apply.

3 Definably conservative gradients and parametric optimality formula

The result described in Theorem 1 is actually proved for conservative jacobians, which generalize conservative gradients in a similar way as jacobians generalize gradients. We start with an extension of Definition 1 to this setting and make the necessary technical connections with the work of [14], in particular the chain rule along absolutely continuous curves and the variational stratification. These preliminaries, although not explicitly stated this way in [14] are direct consequences of [14] and given here for completeness. We then proceed to the proof of the main results of this section: definably conservative jacobians are conservative and definable PO formula describes a conservative gradient.

3.1 Conservative jacobians

The following defines conservativity for matrix set valued functions using vanishing circulation. This was suggested in [14], it is equivalent to [14, Definition 4].

Definition 3 (Conservative jacobians) *Let $J: \mathbb{R}^p \rightrightarrows \mathbb{R}^{m \times p}$ be a set valued map with closed graph, non empty and locally bounded values. Then J is a conservative jacobian, if for all $x \in \mathbb{R}^p$, all $\gamma: [0, 1] \rightarrow \mathbb{R}^p$, absolutely continuous with $\gamma(0) = \gamma(1)$, and all measurable functions, $V: [0, 1] \rightarrow \mathbb{R}^{m \times p}$, such that $V(t) \in J(\gamma(t))$ for all $t \in [0, 1]$,*

$$\int_0^1 V(t) \dot{\gamma}(t) dt = 0. \quad (5)$$

If in addition, J is definable, then J is called definably conservative if (5) holds only for definable C^1 loops γ and definable selections V .

The following Lemma is a useful alternative characterization of conservativity through an operational chain rule corresponding to [14, Definition 4].

Lemma 1 *Let $J: \mathbb{R}^p \rightrightarrows \mathbb{R}^{m \times p}$, be a graph closed locally bounded non empty valued map. Then the following are equivalent*

- (i) *There exists $G: \mathbb{R}^p \rightarrow \mathbb{R}^m$, locally Lipschitz such that for any $\gamma: [0, 1] \rightarrow \mathbb{R}^p$ absolutely continuous, for almost all $t \in [0, 1]$*

$$\frac{d}{dt} G(\gamma(t)) = M \dot{\gamma}(t), \quad \forall M \in J(\gamma(t)).$$

(ii) J is a conservative jacobian as stated in Definition 3.

In the situation of Lemma 1 (i), we will say that J is a conservative jacobian for G , in which case G is defined up to constants using line integrals.

Remark 5 *Lemma 1 holds mutatis mutandis by taking γ to be definable C^1 curve and definable conservativity in Definition 2.*

3.2 O-minimal structures

Important references on this topic are [21, 26]. An *o-minimal structure* on $(\mathbb{R}, +, \cdot)$ is a collection of sets $\mathcal{O} = (\mathcal{O}_p)_{p \in \mathbb{N}}$ where each $\mathcal{O}_p$ is itself a family of subsets of $\mathbb{R}^p$, such that for each $p \in \mathbb{N}$:

- (i) $\mathcal{O}_p$ is stable by complementation, finite union, finite intersection and contains $\mathbb{R}^p$.
- (ii) if A belongs to $\mathcal{O}_p$, then both $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathcal{O}_{p+1}$;
- (iii) if $\pi : \mathbb{R}^{p+1} \rightarrow \mathbb{R}^p$ is the canonical projection onto $\mathbb{R}^p$ then, for any $A \in \mathcal{O}_{p+1}$, the set $\pi(A)$ belongs to $\mathcal{O}_p$;
- (iv) $\mathcal{O}_p$ contains the family of real algebraic subsets of $\mathbb{R}^p$, that is, every set of the form

$$\{x \in \mathbb{R}^p \mid g(x) = 0\}$$

where $g : \mathbb{R}^p \rightarrow \mathbb{R}$ is a polynomial function;

- (v) the elements of $\mathcal{O}_1$ are exactly the finite unions of intervals.

A subset of $\mathbb{R}^p$ which belongs to an o-minimal structure $\mathcal{O}$ is said to be *definable in $\mathcal{O}$* . A function is *definable in $\mathcal{O}$* whenever its graph is definable in $\mathcal{O}$. A set valued mapping (or a function) is said to be definable in $\mathcal{O}$ whenever its graph is definable in $\mathcal{O}$. The terminology *tame* refers to definability in an o-minimal structure without specifying which structure. From now on we fix an o-minimal structure $\mathcal{O}$, definable sets being implicitly definable in $\mathcal{O}$.

The simplest o-minimal structure is given by the class of real semialgebraic objects. Recall that a set $A \subset \mathbb{R}^p$ is called *semialgebraic* if it is a finite union of sets of the form

$$\bigcap_{i=1}^k \{x \in \mathbb{R}^p \mid g_i(x) < 0, h_i(x) = 0\}$$

where the functions $g_i, h_i : \mathbb{R}^p \rightarrow \mathbb{R}$ are real polynomial functions and $k \geq 1$. The key tool to show that these sets form an o-minimal structure is Tarski-Seidenberg principle which ensures that (iii) holds true. As detailed in [21] this result can be expressed in the following way.

Proposition 1 (Quantifier elimination) *Any first order formula (quantification on variables only) involving polynomials, equalities and inequalities, with definable functions and definable sets describes a definable set.*

3.3 Variational Stratification

The notion of Variational Stratification was first exposed in [12] and found an interesting application to conservativity [14]. The tangent space at a point x of a differentiable manifold M is denoted by $T_x M$ ⁵. Given a submanifold M of a finite dimensional Riemannian manifold, it is endowed by the Riemannian structure inherited from the ambient space. Given $G: \mathbb{R}^p \rightarrow \mathbb{R}^m$ and $M \subset \mathbb{R}^p$ a differentiable submanifold on which G is differentiable, we denote by $\text{Jac}_M G$ its Riemannian jacobian or even, when no confusion is possible, $\text{Jac } G$.

A C^k stratification of a (sub)manifold M (of $\mathbb{R}^p$) is a partition $\mathcal{S} = (M_1, \dots, M_m)$ of M into C^k manifolds having the property that $\text{cl } M_i \cap M_j \neq \emptyset$ implies that M_j is entirely contained in the boundary of M_i whenever $i \neq j$. Assume that a function $G: M \rightarrow \mathbb{R}^m$ is given and that M is stratified into manifolds on which G is differentiable. For x in M , we denote by M_x the strata containing x and we simply write $\text{Jac } G(x)$ for the jacobian of G with respect to M_x .

Stratifications can have many properties, we refer to [26] and references therein for an account on this question and in particular for more on the idea of a Whitney stratification. The definition is as follows: a C^r -stratification $\mathcal{S} = (M_i)_{i \in I}$ of a manifold M has the *Whitney-(a) property*, if for each $x \in \text{cl } M_i \cap M_j$ (with $i \neq j$) and for each sequence $(x_k)_{k \in \mathbb{N}} \subset M_i$ we have:

$$\left. \begin{array}{l} \lim_{k \rightarrow \infty} x_k = x \\ \lim_{k \rightarrow \infty} T_{x_k} M_i = \mathcal{T} \end{array} \right\} \implies T_x M_j \subset \mathcal{T}$$

where the second limit is to be understood in the Grassmanian, i.e., “directional”, sense. In the sequel we shall use the term *Whitney stratification* to refer to a C^1 -stratification with the Whitney-(a) property. The following can be found for example in [26, 4.8].

Theorem 5 (Whitney stratification) *Let $A_1, \dots, A_k$ be definable subsets of $\mathbb{R}^p$, then there exists a definable Whitney stratification $(M_i)_{i \in I}$ compatible with $A_1, \dots, A_k$, i.e. such that for each $i \in I$, and $t \in \{1, \dots, k\}$, either $M_i \subset A_t$ or $M_i \cap A_t = \emptyset$.*

For the rest of this section, k denotes an arbitrary positive integer.

Definition 4 (Variational stratification [12]) Let $G: \mathbb{R}^p \rightarrow \mathbb{R}^m$, be locally Lipschitz continuous, let $J: \mathbb{R}^p \rightrightarrows \mathbb{R}^{m \times p}$ be a set valued map and let $k \geq 1$. We say that the couple (G, J) has a C^k *variational stratification* if there exists a C^k Whitney stratification $\mathcal{S} = (M_i)_{i \in I}$ of $\mathbb{R}^p$, such that G is C^k on each stratum and for all $x \in \mathbb{R}^p$,

$$J(x) \text{Proj}_{T_{M_x}(x)} = \{\text{Jac } G(x)\}, \quad (6)$$

where $\text{Jac } G(x)$ is the jacobian of G restricted to the active strata M_x containing x .

⁵We only consider embedded manifolds and submanifolds

Theorem 6 (Characterization of conservativity) *Let $J: \mathbb{R}^p \rightrightarrows \mathbb{R}^{m \times p}$ be a definable, nonempty, locally bounded, graph closed set valued mapping and $G: \mathbb{R}^p \rightarrow \mathbb{R}$ be a definable locally Lipschitz function. Then the following are equivalent*

- *J is conservative for G .*
- *(G, J) admit a C^k variational stratification.*

For the reverse implication, J and G need not to be definable.

Proof : This result is essentially known and we point out the arguments for completeness. First, J is conservative for G if and only if, the projection of each row of J is conservative for the corresponding coordinate of G , this is Lemma 3 and 4 in [14] in combination with Lemma 1 above. Hence we may reason coordinatewise.

For the direct implication, it results from [14, Theorem 4] that each coordinate of G_i and the corresponding line of J , J_i admit a variational projection formula ([14, Definition 5]) for each $i = 1, \dots, m$, this corresponds to the variational formula introduced in [12], which is limited to univariate outputs. The variational projection formula is stable by considering submanifolds and hence is stable when refining a given stratification. Hence thanks to Theorem 5, we may find a common Whitney stratification such that the projection formula holds for each coordinate G_i of G and the corresponding row J_i of J . This results in the formula given in Definition 4.

For the reverse implication, similarly as above, the variational stratification in Definition 4 implies the projection formula of [14, Definition 5] for each coordinate of G with the corresponding line of J_i . By [14, Theorem 3] (see also [23] which states the result for the Clarke subdifferential), each row of J is conservative for the corresponding coordinate of G which implies that J is conservative for G by [14, Lemma 4]. This does not require definability. $\square$

3.4 Definably conservative jacobians

We start with a preliminary lemma which will then be applied recursively toward a proof of a variational stratification property from which Theorem 1 will follow.

Lemma 2 *Let $J: \mathbb{R}^p \rightrightarrows \mathbb{R}^{m \times p}$, be a definably conservative jacobian. Let $G: x \mapsto \int_0^1 V(\gamma(t))\dot{\gamma}(t)dt$ for any C^1 definable $\gamma: [0, 1] \rightarrow \mathbb{R}^p$ with $\gamma(0) = 0$ and $\gamma(1) = x$ and any definable selection V as in Definition 3. Then there exists a finite number of definable open sets $U_1, \dots, U_N$ in $\mathbb{R}^p$ such that $\cup_{i=1}^N \text{cl}(U_i) = \mathbb{R}^p$, G is continuously differentiable on each set and $J = \{J_G\}$, the jacobian of G , on each set.*

Proof : Note that it is not known *a priori* if G is definable and therefore the results of [14, 24] do not directly applies. Denote by $\mathcal{R}$ the set where J is single valued, this set is definable by Proposition 1 and we are going to show that its complement has empty interior. Toward a contradiction, suppose that the complement of $\mathcal{R}$ has nonempty interior. Then, definable choice [26, 4.5] ensures that there exists two definable selections

V_1 and V_2 such that $V_1 \neq V_2$ on a small open ball B . Let v be a unit norm definable selection in $\text{Im}(V_1 - V_2)^T$, which exists thanks to [26, 4.5]. Since v, V_1, V_2 are definable, we may stratify using [26, 4.8] so that v, V_1, V_2 can be assumed to be Lipschitz continuous on B (reducing and translating B if necessary). Call r the radius of B and assume without loss of generality that it is centered at 0. Consider the solution to

$$\dot{\gamma}(t) = v(\gamma(t)), \quad \gamma(0) = 0.$$

γ is C^1 and stays in a neighborhood of 0 for small values of t , let's say that $\|\gamma(t)\| < r/2$ for all $t \in [-\alpha, \alpha]$ for a certain $\alpha > 0$. Let $0 < \epsilon < r/2$ be arbitrary and fix $\tilde{\gamma}$, a C^1 definable path with $\tilde{\gamma}(0) = 0$ such that

$$\max_{t \in \alpha[0,1]} \max \{ \|\gamma(t) - \tilde{\gamma}(t)\|, \|\gamma'(t) - \tilde{\gamma}'(t)\| \} \leq \epsilon,$$

take for example a polynomial approximation of γ' and its integral, using Weierstrass approximation Theorem. Since $\tilde{\gamma}$ is a C^1 definable arc which remains in B by construction, for almost all t around 0, we have using Lemma 1 (see remark 5),

$$\frac{d}{dt} G(\tilde{\gamma}(t)) = V_1(\tilde{\gamma}(t))(v(\gamma(t)) + u(t)) = V_2(\tilde{\gamma}(t))(v(\gamma(t)) + u(t))$$

where $u(t) := \tilde{\gamma}'(t) - v(\gamma(t))$ for all t . By construction, we have $\|u(t)\| \leq \epsilon$ for all t . We can let $\epsilon \rightarrow 0$ along a sequence of such approximations $\tilde{\gamma}$, and using continuity of V_1 and V_2 , we obtain for almost all t ,

$$V_1(\gamma(t))v(\gamma(t)) = V_2(\gamma(t))v(\gamma(t))$$

and therefore $(V_1(\gamma(t)) - V_2(\gamma(t)))v(\gamma(t)) = 0$, for all t by continuity of v, V_1, V_2 and γ . Since $v(\gamma(t)) \in \text{Im}(V_1(\gamma(t)) - V_2(\gamma(t)))^T$, this shows that $v(\gamma(t)) = 0$ for all t around 0. This is contradictory with the fact that v has unit norm.

This shows that the complement of $\mathcal{R}$ has empty interior. By stratification, using Theorem 5, there exists $U_1, \dots, U_N$ strata of maximal dimension such that the complement of $\mathcal{R}$ does not intersect any U_i , $i = 1, \dots, N$, and $\cup_{i=1}^N \text{cl}(U_i) = \mathbb{R}^p$.

By graph closedness and local boundedness, J can be identified with a continuous function on each U_i . Let $x \in U_i$ for some i , there is a small ball around x such that J is continuous on the ball. By the definition of G , for any $v \in \mathbb{R}^p$ and $t > 0$ such that $x + tv$ remains in this ball,

$$\begin{aligned} \frac{G(x + tv) - G(x)}{t} &= \int_0^1 J(x + stv) v ds \\ &= \frac{1}{t} \int_0^t J(x + sv) v ds. \end{aligned}$$

Letting $t \rightarrow 0$, we have

$$\lim_{t \rightarrow 0, t > 0} \frac{G(x + tv) - G(x)}{t} = J(x)v$$

where the limit is by continuity of J at x . This formula allows to identify the partial derivatives at x of each of the m coordinate components of G with entries of $J(x)$. Since these are continuous at x , G is differentiable at x , and since $x \in U_i$ was arbitrary and J is continuous on each U_i , G is C^1 on each U_i with jacobian J . $\square$

Theorem 7 *Let $J: \mathbb{R}^p \rightrightarrows \mathbb{R}^{m \times p}$ be a definably conservative jacobian. Let $G: \mathbb{R}^p \rightarrow \mathbb{R}$ be defined as in Lemma 2. Then (G, J) admits a C^k variational stratification: $(M_i)_{i \in I}$ a definable Whitney stratification of $\mathbb{R}^p$ such that G is C^k on each stratum with*

$$J(x)Q(x) = \{J_G(x)\},$$

where $Q(x)$ is the matrix representing orthogonal projection to the tangent space of the active stratum $M(x)$ at x , seen as a subspace of $\mathbb{R}^p$.

Proof : We shall prove that (G, J) has a C^1 variational stratification. The C^k variational stratification follows by definability and the existence of C^k stratification of G using [26, 4.8], which allows to refine a potential C^1 stratification to obtain differentiability up to order k .

Let M be a connected definable C^1 submanifold embedded in $\mathbb{R}^p$ which is also a connected C^1 cell (see [26, 4.2]). Since M is a connected C^1 cell, there is a definable C^1 diffeomorphism $\theta: M \rightarrow \mathbb{R}^{\dim M}$, see for example [21, Section 6.2]. Set $\phi: \mathbb{R}^{\dim M} \rightarrow M$ such that $\phi = \theta^{-1}$. Consider now

$$\begin{aligned} \tilde{G}: \mathbb{R}^{\dim M} &\rightarrow \mathbb{R}^m \\ \tilde{x} &\mapsto G(\phi(\tilde{x})) \\ \tilde{J}: \mathbb{R}^{\dim M} &\rightrightarrows \mathbb{R}^{m \times \dim M} \\ \tilde{x} &\rightrightarrows J(\phi(x))J_\phi(\tilde{x}) \end{aligned}$$

For any definable C^1 path $\gamma: [0, 1] \rightarrow \mathbb{R}^{\dim M}$, $\phi \circ \gamma$ is definable and so $(\tilde{G}, \tilde{J})$ satisfy the hypotheses of Lemma 2 since by Lemma 1 (see remark 5), for almost all t

$$\frac{d}{dt} \tilde{G}(\gamma(t)) = \frac{d}{dt} G(\phi \circ \gamma(t)) = J(\phi \circ \gamma(t))(\phi \circ \gamma)'(t) = J(\phi \circ \gamma(t))J_\phi(\gamma(t))\dot{\gamma}(t) = \tilde{J}(\gamma(t))\dot{\gamma}(t),$$

and γ was an arbitrary C^1 definable path.

Hence there exists $\tilde{U}_1 \dots \tilde{U}_N$ open in $\mathbb{R}^{\dim M}$ such that the union of their closure is equal the whole $\mathbb{R}^{\dim M}$, and on each set U_i , $\tilde{J}$ is single valued and $\tilde{G}$ is C^1 with $\tilde{J} = \{J_{\tilde{G}}\}$. For $i = 1 \dots N$, set $U_i = \phi(\tilde{U}_i)$ we have that the union of their relative closure is M : $\cup_{i=1}^N \text{cl}_M(U_i) = M$. Furthermore, $G = \tilde{G} \circ \theta$ is differentiable on each U_i , relative to M , and

$$\{J_G(x)\} = \{J_{\tilde{G}}(\theta(x))J_{\theta(x)}\} = J(x)J_\phi(\theta(x))J_\theta(x).$$

We remark that $J_\phi(\theta(x))J_\theta(x)$ is the projection on the tangent space of M at x so that each line of J has a single valued projection on the tangent space of M on each U_i .

Since M was arbitrary, we may start with $M = \mathbb{R}^p$ and proceed by induction on the dimension by applying Lemma 2 with the above reasoning. We have a C^1 variational projection on a dense set of strata (cells) of dimension p . We may obtain a Whitney stratification, compatible with this set, so that the projection formula does not hold only on a finite union of strata (cells) of dimension at most $p - 1$ [26, 4.8]. Each stratum M , being a C^1 embedded submanifold as well as a C^1 cell and we may repeat the process recursively until the dimension of the set where the projection formula does not hold is zero, i.e. a finite set of points, to obtain the desired Whitney stratification. $\square$

The following corollary combines Theorems 6 and 7. Theorem 1 is a special case for $m = 1$.

Corollary 1 *Let $J: \mathbb{R}^p \rightrightarrows \mathbb{R}^{m \times p}$ be a definably conservative jacobian, then J is a conservative jacobian.*

3.5 Application to PO formula

This section describe how Corollary 1 can be used to prove that the PO formula provides a conservative gradient in the definable world. We start with a slightly more general result which generalizes the finite selection process described in [15] from the discrete to the continuous setting.

Theorem 8 *Let $F: \mathbb{R}^p \times \mathbb{R}^r \rightarrow \mathbb{R}$ be locally Lipschitz, D a definable conservative jacobian for F and $P: \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a definable set valued field, with closed graph nonempty locally bounded values, such that for all $x \in \mathbb{R}^p$ and $y \in P(x)$, there exists $u \in \mathbb{R}^p$ such that $(u, 0) \in D(x, y)$ and $y \mapsto F(x, y)$ is constant on $P(x)$. We set*

$$f: x \mapsto F(x, P(x)).$$

We have that f is continuous, set

$$D_f: x \mapsto \{u, \exists y \in P(x), (u, 0) \in D(x, y)\}.$$

Then D_f is conservative for f .

Proof : One can check that D_f is definable thanks to Proposition 1. Furthermore, it has closed graph and is locally bounded with nonempty values. Hence, by Theorem 1 and Lemma 1, we only have to prove a chain rule along definable C^1 curves.

Let $t \mapsto x(t) \in \mathbb{R}^p$ be a C^1 definable path we will obtain definable selections of interest thanks to [26, 4.5], and draw conclusions thanks to Lemma 8 which asserts that definable set valued fields have countable dense definable selections.

Let $t \mapsto y(t) \in \mathbb{R}^p$ be a definable selection in $t \mapsto P(x(t))$ and let $t \mapsto T(t) = (u(t), v(t))$ be a definable selection in $t \rightrightarrows D(x(t), y(t))$. Definable curves are piecewise continuously differentiable. Hence the functions $t \mapsto f(x(t)) = F(x(t), y(t))$, $t \mapsto x(t)$ and $t \mapsto y(t)$ are differentiable everywhere except at finitely many points, call them $t_1, \dots, t_M$. Bounded definable curves have left and right limits everywhere, we have for $i = 2, \dots, M$ that $t \mapsto y(t)$ is continuous on (t_{i-1}, t_i) and can be extended to an absolutely continuous path on $[t_{i-1}, t_i]$. Hence we can use the fact that D is conservative for F , which yields using graph closedness of P and continuity of F , for $i = 2, \dots, M$

$$\begin{aligned} f(x(t_i)) - f(x(t_{i-1})) &= F(x(t_i), \lim_{t \uparrow t_i} y(t)) - F(x(t_{i-1}), \lim_{t \downarrow t_{i-1}} y(t)) \\ &= \int_{t=t_{i-1}}^{t=t_i} \frac{d}{dt} F(x(t), y(t)) dt = \int_{t=t_{i-1}}^{t=t_i} \langle \dot{x}(t), u(t) \rangle + \langle \dot{y}(t), v(t) \rangle dt. \end{aligned}$$

Finally, by removing the set of discontinuity points of y , we obtain for almost all t .

$$\frac{d}{dt}f(x(t)) = \frac{d}{dt}F(x(t), y(t)) = \langle \dot{x}(t), u(t) \rangle + \langle \dot{y}(t), v(t) \rangle$$

Since T was an arbitrary definable selection of $t \rightrightarrows D(x(t), y(t))$, which admits a countable dense sequence of definable selections thanks to Lemma 8, we have for almost all t

$$\frac{d}{dt}f(x(t)) = \langle \dot{x}(t), u \rangle + \langle \dot{y}(t), v \rangle, \quad \forall (u, v) \in D(x(t), y(t)).$$

By the hypotheses, for each $u \in D_f(x)$, one can choose $v = 0$ and hence for almost all t

$$\frac{d}{dt}f(x(t)) = \langle \dot{x}(t), u \rangle \quad \forall u \text{ such that } (u, 0) \in D(x(t), y(t)).$$

Note that y is an arbitrary definable selection in $t \rightarrow P(x(t))$. By Lemma 9, there is a countable family of such selections $(y_i)_{i \in \mathbb{N}}$, such that for all t ,

$$D_f(x(t)) = \text{cl} \{u, \exists i \in \mathbb{N}, (u, 0) \in D(x(t), y_i(t))\}.$$

This implies that for almost all t

$$\frac{d}{dt}f(x(t)) = \langle \dot{x}(t), u \rangle \quad \forall u \in D_f(x(t)).$$

Hence, we have a chain rule along definable arcs which is the desired result. Indeed, repeating the proof of direct implication in Lemma 1 for definable arcs, we have that D_f is definably conservative and hence conservative thanks to Theorem 1 and admits f as a potential. $\square$

This result can be applied to partial maximization as the following example shows. It could also be applied to partial minimization or differentiation of more general critical values such as local minima or local maxima (under suitable assumptions). The following example is a repetition of Theorem 2 which is based on Theorem 8.

Example 1 (Partial maximization) *Let $F: \mathbb{R}^p \times \mathbb{R}^r \rightarrow \mathbb{R}$ be definable locally Lipschitz and $D: \mathbb{R}^{p+r} \rightrightarrows \mathbb{R}^{p+r}$ be definable and conservative for F with convex values. Set*

$$\begin{aligned} f: \mathbb{R}^p &\rightarrow \mathbb{R} \\ x &\mapsto \max_{y \in \mathbb{R}^r} F(x, y), \end{aligned}$$

where the argmax is assumed to be nonempty and locally bounded, call it $P(x) \subset \mathbb{R}^r$. P has a closed graph, set

$$\begin{aligned} D_f: \mathbb{R}^p &\rightrightarrows \mathbb{R}^p \\ x &\mapsto \text{conv} \{u, \exists y \in P(x), (u, 0) \in D(x, y)\}, \end{aligned}$$

then D_f is conservative for f . By [38, Example 10.12], for all $x \in \mathbb{R}^p$ and $y \in P(x)$ there is $u \in \mathbb{R}^p$ such that $(u, 0) \in \partial^c F^c(x, y) \subset D(x, y)$ (the inclusion follows from [14, Corollary 1] and convexity of $D(x, y)$). This shows that D and P comply with Theorem 8.

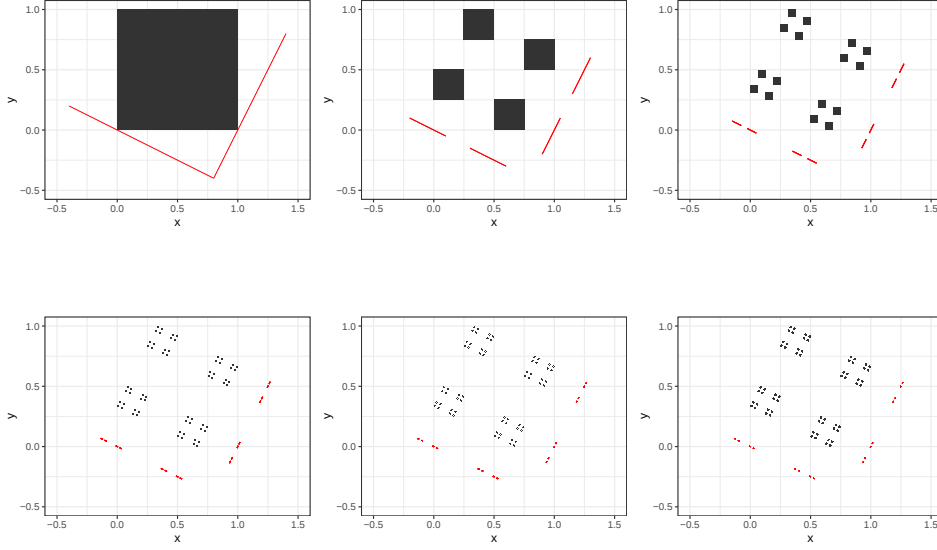


Figure 1: The fractal construction in Section 4. We start with the closed unit square in black. It is split into four copies of size one fourth the original square. This process is repeated recursively on each square ad infinitum. The additional red lines represent projection of these sets on rotated axes. Considering C_i , $i \in \mathbb{N}$, the set obtained after i steps (C_0 is the original square), we have $C_{i+1} \subset C_i$ for all i , we let the limiting set to be $C = \bigcap_{i \in \mathbb{N}} C_i$, which is closed. The projection of C on each axes are full segments. Furthermore, in the limit, both projections on rotated axes are Cantor sets of zero measure.

4 Failure of parametric optimality formula

This section is dedicated to the construction of the function g in Theorem 4. We start with the construction of a fractal set C and then describe the counterexample which will be based on the distance functions to C .

4.1 A fractal set

Let C be the fractal set which construction is described in Figure 1. The construction of C is similar to the one described in [45] to provide a counterexample to Morse-Sard theorem. This construction was also used in [36] to provide a subgradient sequence on a path differentiable function which fails to have dissipative and minimizing properties.

The set is defined as $C = \bigcap_{i \in \mathbb{N}} C_i$, where for each $i \in \mathbb{N}$, C_i is the union of 4^i squares of size $1/4^i$. Furthermore, these sets form a nested decreasing sequence for the inclusion partial order. The set C has the following properties.

- C is closed as an intersection of closed sets, hence it is compact.
- The projection of C on each axis are full segments $[0, 1] \times \{0\}$ and $\{0\} \times [0, 1]$, we denote them by $\text{proj}_x(C)$ and $\text{proj}_y(C)$. Indeed, since we have a nested sequence $\text{proj}_x(C) = \bigcap_{i \in \mathbb{N}} \text{proj}_x(C_i) = \bigcap_{i \in \mathbb{N}} [0, 1] = [0, 1]$.

- In particular, C is nonempty.
- The projection on each axis rotated clockwise by an angle of $\arctan(2)$ is a Cantor set of zero measure (see Figure 1). Indeed, at each step there is a constant proportion of each segment which is removed from the projection, this is the simplest construction of Cantor sets.
- For each $i \geq 1$, there is a finite number of $x \in [0, 1]$ for which the vertical line starting at x intersects C_i at two distinct squares, call this set X_i . We have $X_i \subset X_{i+1}$ and for each $x \in [0, 1]$, $x \notin X_i$, the intersection of the vertical line at x and C_i is at a single square. Set $X = \bigcup_{i \in \mathbb{N}} X_i$, X is denumerable and for each $x \in X$, the vertical line starting at x intersects C at two distinct points, for all other $x \in [0, 1]$ this intersection is a singleton.
- Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be a function such that for some $i \in \mathbb{N}$, $(x, f(x)) \in C_i$ for all $x \in [0, 1]$, then the total variation of f is at least i . Hence if f satisfies this property for all i , that is $(x, f(x)) \in C$ for all $x \in [0, 1]$, then f has infinite total variation.

4.2 Construction and proof of the counterexample

Consider the following Lipschitz function

$$\begin{aligned} f: \mathbb{R}^2 &\rightarrow \mathbb{R} \\ z &\mapsto -\text{dist}(z, C). \end{aligned}$$

Set $O = \mathbb{R}^2 \setminus C$, O is open and we have for all $z \in O$

$$f(z) = -\min_{c \in C} \|z - c\| = \max_{c \in C} -\|z - c\|.$$

Each function $f_c: z \mapsto -\|z - c\|$ is C^1 on O and both f_c and ∇f_c are jointly continuous with respect to z and c on $O \times C$. This shows that f is lower C^1 on O [38, Definition 10.29]. Hence we have that f is subdifferentially regular on O [38, Theorem 10.31]. Combining with Lemma 5 we have

Lemma 3 *f is path differentiable.*

Proof : We will prove that $\partial^c f$ satisfies the chain rule along absolutely continuous curves [23, 14], since f is Lipschitz, this is sufficient to conclude, see also Lemma 1. Let $\gamma: [0, 1] \rightarrow \mathbb{R}^2$ be an absolutely continuous curve and $R \subset [0, 1]$ be the full measure set where γ and $f \circ \gamma$ are differentiable, we will show that γ satisfies the chain rule for almost all $t \in R$ which is sufficient to conclude. We set

$$E = \{t \in R, \gamma(t) \notin C\}$$

We also consider $\tilde{E} \subset E$ with

$$\tilde{E} = \left\{ t \in E, \max_{v \in \partial^c f(\gamma(t))} \left| \langle v, \dot{\gamma}(t) \rangle - \frac{d}{dt} f(\gamma(t)) \right| > 0 \right\}.$$

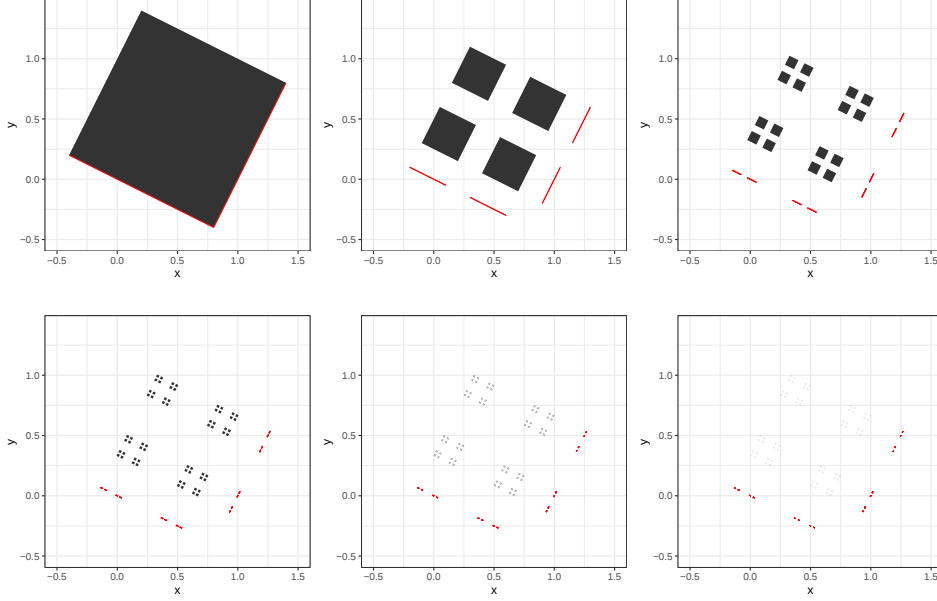


Figure 2: Same construction as in Figure 1, except that we keep the whole product set, the red line representing the projections on rotated axes and the black squares representing their products. In the limit, we obtain the same set C as in Figure 1 since the distance between the set represented here and those of Figure 1 goes to zero.

Fix $t \in \tilde{E}$ arbitrary, since $t \in E$, choosing a small enough, we have by continuity of γ , $\gamma([t-a, t+a]) \cap C = \emptyset$ and $\gamma([t-a, t+a]) \subset O$. Since f is lower C^1 on O , by [38, Theorem 10.31], f is subdifferentially regular in a neighborhood of $\gamma([t-a, t+a]) \subset O$. We may apply [23, Lemma 5.4] which shows that f satisfies the chain rule along the curve γ restricted to the closed segment $I = [t-a, t+a]$, in other words $[t-a, t+a] \cap \tilde{E}$ has measure zero. The segment can be taken of arbitrarily small length, therefore, such intervals form a Vitali covering set (see for example [39, Section 6.2]). By the Vitali covering theorem [39, Section 6.2] for any $\epsilon > 0$, there is a finite collection of such segments $I_1, \dots, I_K$ such that $\tilde{E} \setminus \cup_{k=1}^K I_k$ has (outer) measure at most ϵ which shows that $\tilde{E}$ has (outer) measure arbitrary small and therefore has measure zero.

Now set

$$\begin{aligned} E_1 &= \{t \in R, \gamma(t) \in C, \gamma'(t) \neq 0\} \\ E_2 &= \{t \in R, \gamma(t) \in C, \gamma'(t) = 0\}. \end{aligned}$$

We have $R = E \cup E_1 \cup E_2$, we have shown that the chain rule holds for almost all t in E , Lemma 6 ensures that E_1 has zero measure and the chain rule holds trivially for all $t \in E_2$ because f is Lipschitz. This shows that the chain rule holds for almost all $t \in \mathbb{R}$ and the conclusion follows. $\square$

Now, we should characterize the subdifferential of f , which directly relates to the normal cone to C . We will use the notion of normal cone as described in [38, Definition 6.3]. As depicted in Figure 2, it turns out that C is actually a product of Cantor set which are closed and have empty interior. Using Lemma 7 we obtain by [38, Proposition 6.41] that the normal cone to C is $\mathbb{R}^2$ everywhere on C . We deduce the following.

Lemma 4 *For all $z \in C$, $\partial^c f(z) = B$, where B is the unit ball.*

Proof : As shown in Figure 1, C is contained in a product of cantor sets, $\tilde{C}$. Actually C is equal to a product of Cantor sets. Indeed, as detailed in Figure 2, the distance between the sets constructed in Figure 1 and those of Figure 2 decreases to 0, so that limiting intersections are the same. By [38, Proposition 6.41] and using Lemma 7, $N_{\tilde{C}} = \mathbb{R}^2$ everywhere on $\tilde{C}$ hence on C . The result follows from [38, Example 8.53] using the fact that the Clarke subdifferential is the convex closure of the limiting subdifferential and it commutes with multiplication by scalars. $\square$

For the function g in Theorem 4, one could take for example $g: (x, y) \mapsto 2f(x, y) + x$ which satisfies all the required properties

- Since the projection of C on x is $[0, 1]$ (Section 4.1), we have $\max_{y \in [0, 1]} g(x, y) = x$.
- Since for all but countably many x , the intersection of C with the vertical line at x is a singleton (Section 4.1), for such x , the argmax is unique, for the others it is a pair.
- The sum rule holds for f , so for every $x \in [0, 1]$ and $y \in \arg \max_{y \in [0, 1]} g(x, y)$, we have $(x, y) \in C$ so $\partial^c f(x, y) = B$, by Lemma 4. By sum rule, $\partial^c g(x, y) = 2B + (1, 0)$ so that

$$\left\{ v, (v, 0) \in \partial^c g(x, y), y \in \arg \max_{y \in [0, 1]} g(x, y) \right\} = [-1, 3]$$

which contains 0.

4.3 Need for a better behaved subclass

In the definition of the function g , the fractal nature of the construction makes the argmin mapping, $P: \mathbb{R}^p \rightrightarrows \mathbb{R}^r$, although almost everywhere a singleton, highly non regular. In this example, it is not even of bounded variation in the sense that it is not possible to obtain bounded variation selections in P , a fortiori not absolutely continuous. This explains why conservativity is destroyed, the connection between g , F and its value function f , through P in the PO formula takes place outside of the absolutely continuous world and hence outside of the conservative world which is built on absolutely continuous paths.

Therefore additional restrictions on the function F have to be enforced if one wants a calculus rule for the PO formula which preserves conservativity. An intuitive direction, is to ensure that the argmax mapping admits selections which are absolutely continuous, or close to absolute continuity, in order to apply and use the definition of conservativity in Definition 1. There would be potentially many possible such classes, in Section 3.5 we focus on one of them, definable functions [26, 21], for which we have access to definable selections which are piecewise differentiable, which is close enough to absolute continuity for our purpose.

5 Lemmas and proofs

Proof of Lemma 1: The fact that (i) implies (ii) is direct by integration. Indeed, $G \circ \gamma$ is absolutely continuous. For any $\gamma: [0, 1] \rightarrow \mathbb{R}^p$ absolutely continuous, with $\gamma(0) = \gamma(1)$, and measurable selection V as in the statement of the Lemma,

$$\int_0^1 V(\gamma(t))\dot{\gamma}(t)dt = \int_0^1 \frac{d}{dt}G(\gamma(t))dt = G(\gamma(1)) - G(\gamma(0)) = 0,$$

where the first inequality uses (i) and the fact that $V(\gamma(t)) \in J(t)$ for all t and the second is absolute continuity of $G \circ \gamma$.

For the reverse implication, fix $x \in \mathbb{R}^p$, an absolutely continuous path $\gamma: [0, 1] \rightarrow \mathbb{R}^p$, $\gamma(0) = 0$, $\gamma(1) = x$ and a measurable selection $V: \mathbb{R}^p \rightarrow \mathbb{R}^p$ such that for all $z \in \mathbb{R}^p$, $V \in J(z)$. We define

$$G(x) = \int_0^1 V(t)\dot{\gamma}(t)dt.$$

By (ii), the value of G does not depend on the choice of γ and on the measurable selection V . Furthermore, we have by Lebesgue differentiation theorem, for almost all $t \in [0, 1]$,

$$\frac{d}{dt}G(\gamma(t)) = V(t)\dot{\gamma}(t) \quad (7)$$

Now since J is nonempty compact valued, from [3, corollary 18.15], there exists a sequence $(V_i)_{i \in \mathbb{N}}$ of measurable selections such that for all $t \in [0, 1]$

$$J(\gamma(t)) = \text{cl}\{V_i(\gamma(t))\}_{i \in \mathbb{N}}. \quad (8)$$

The result follows by combining (7) and (8). $\square$

The following Lemma is essentially a repetition of [36, Lemma 16] which we reproduce here for completeness.

Lemma 5 *Let $C \subset \mathbb{R}^2$ be a closed set which projections on the x and y axes have measure zero respectively. Then for any Lipschitz curve, $\gamma: \mathbb{R} \rightarrow \mathbb{R}^2$, the set*

$$E = \{t \in \mathbb{R}, \gamma(t) \in C\} \setminus \{t \in \mathbb{R}, \gamma'(t) = 0\}$$

has measure zero.

Proof : Write γ_1, γ_2 the coordinates of γ and P_1, P_2 the projection of C on the x and y axes respectively, by hypothesis they have measure zero. Set $B \subset \mathbb{R}$ the set where γ' is either not defined or well defined and different from zero. We have $E = \gamma^{-1}(C) \cap B$.

Set $A_1 \subset B$ and $A_2 \subset B$ the sets where $\gamma'_1 \neq 0$ and $\gamma'_2 \neq 0$ respectively. Set $A_3 \subset B$ the zero measure set where γ' is not defined. We have that, $A_1 \cup A_2 \cup A_3 = B$.

Consider an enumeration of intervals $(I_j)_{j \in \mathbb{N}}$ of the form $[p_j, q_j]$, with $p_j, q_j \in \mathbb{Q}$ and $p_j < q_j$ for $j \in \mathbb{N}$. Fix $j \in \mathbb{N}$ and $i = 1, 2$ and consider the following function on I_j

$$f_j^i: t \rightarrow \min_{s \in I_j, \gamma_i(s) = \gamma_i(t)} s.$$

The set valued function $t \mapsto \{s \in I_j, \gamma_i(s) = \gamma_i(t)\}$ has closed graph and nonempty compact values on I_j and hence f_j^i is measurable by [3, Theorems 18.19 and 18.20]. Set

$$Q_j^i = \{t \in I_j, t = f_j^i(t)\} \cap A_i \cap \gamma_i^{-1}(P_i).$$

Q_j^i is measurable and γ_i is injective on Q_j^i by construction and $\gamma_i(Q_j^i) \subset P_i$.

Using the injectivity of γ_i on Q_j^i and a change of variable formula [27, Theorem 3.8], we have for any $j \in \mathbb{N}$ and $i = 1, 2$,

$$0 \leq \int_{Q_j^i} |\gamma_i'(t)| dt = \int_{\gamma_i(Q_j^i)} 1 dt \leq \int_{P_i} 1 dt = 0$$

Since the first integrand is strictly positive, this means that Q_j^i has measure zero.

Now consider $t \in \gamma_i^{-1}(P_i) \cap A_i$, this means that there exists $p \in P_i$ such that $t \in \gamma_i^{-1}(p) \cap A_i$. Since $t \in A_i$ this means that $\gamma_i'(t) \neq 0$ and so there exists an interval I_j containing t , such that $\{t\} = \{s \in I_j, \gamma_i(s) = \gamma_i(t)\}$. This shows that $t \in Q_j^i$ and since t was arbitrary, we have that $\gamma_i^{-1}(P_i) \cap A_i = \cup_{j \in \mathbb{N}} Q_j^i$ and hence $\gamma_i^{-1}(P_i) \cap A_i$ has zero measure as the countable union of zero measure sets.

We have

$$\begin{aligned} E &= \gamma^{-1}(C) \cap B = \gamma^{-1}(C) \cap (A_1 \cup A_2 \cup A_3) \\ &\subset \gamma_1^{-1}(P_1) \cap \gamma_2^{-1}(P_2) \cap (A_1 \cup A_2 \cup A_3) \\ &\subset (\gamma_1^{-1}(P_1) \cap A_1) \cup (\gamma_2^{-1}(P_2) \cap A_2) \cup A_3. \end{aligned}$$

All three sets on the right hand side have zero measures so E has zero measure. □

Lemma 6 *The result of Lemma 5 holds for any absolutely continuous curve $\gamma: [0, 1] \rightarrow \mathbb{R}^2$.*

Proof : From [4, Lemma 1.1.4], γ admits a Lipschitz reparametrization. That is, there exists an increasing absolutely continuous function $s: \mathbb{R} \rightarrow \mathbb{R}$ with Lipschitz inverse t and a Lipschitz curve $\hat{\gamma}$, such that

$$\begin{aligned} \hat{\gamma} &= \gamma \circ t \\ \|\hat{\gamma}'\| \circ s &= \frac{\|\gamma'\|}{1 + \|\gamma'\|} \end{aligned} \tag{9}$$

where the second identity holds almost everywhere. Lemma 6 holds true for $\hat{\gamma}$. We have for almost all a using (9) and the fact that t is the inverse of s ,

$$\begin{aligned} &\hat{\gamma}(a) \in C, \|\hat{\gamma}'\|(a) \neq 0 \\ \Leftrightarrow &\gamma(t(a)) \in C, \|\hat{\gamma}'\| \circ s(t(a)) \neq 0 \\ \Leftrightarrow &\gamma(t(a)) \in C, \|\gamma'(t(a))\| \neq 0. \end{aligned}$$

Set A the full measure set where the previous equivalence holds, we have

$$\{t, \gamma(t) \in C, \|\gamma'(t)\| \neq 0\} \subset t(\{a, \hat{\gamma}(a) \in C, \|\hat{\gamma}'\|(a) \neq 0\}) \cup t(A^c).$$

The set $\{s, \hat{\gamma}(s) \in C, \|\hat{\gamma}'\|(s) \neq 0\}$ has measure zero by Lemma 6 because $\hat{\gamma}$ is Lipschitz. The Lebesgue measure of the image of a zero measure set by a Lipschitz map is zero, and, since t is Lipschitz, the right hand side has measure zero. This is the desired result. $\square$

Lemma 7 *Let $C \subset \mathbb{R}$ be a closed set with empty interior. Then for all $x \in C$, $N_C(x) = \mathbb{R}$.*

Proof : Denote by f the distance function to C . f is 1-Lipschitz and hence differentiable almost everywhere. Furthermore, $f(x) = 0$ if and only if $x \in C$ since C is closed. Fix $x \in C$ and construct a sequence $(x_k)_{k \in \mathbb{N}}$ as follows, choose $\epsilon_k = 1/(k+1)$ for $k \in \mathbb{N}$,

- If $(x, x + \epsilon_k) \cap C = \emptyset$, then $x_k = x$ and we have that $1 \in N_C(x_k)$.
- Otherwise choose $z_k \in (x, x + \epsilon_k) \cap C$ and $y_k = \arg \max_{y \in [x, z_k]} f(y)$ where the maximum is positive by closedness of C and the fact that C has empty interior. y_k has a projection on the left and on the right on C (otherwise it cannot be in the argmax). Choose x_k to be the projection on the left. We have that $1 \in N_C(x_k)$, $x \leq x_k \leq x + \epsilon_k$.

In all cases we have $x_k \rightarrow x$ and $1 \in N_C(x_k)$ which shows that $1 \in N_C(x)$. Similarly one could show that $-1 \in N_C(x)$ and hence $N_C(x) = \mathbb{R}$ since it is a cone. Since x was arbitrary in C , this proves the desired result. $\square$

Lemma 8 *Let $J: \mathbb{R}^p \rightarrow \mathbb{R}^m$, be a definable compact valued map with nonempty values. Then there exists a sequence of definable selections $(V_i)_{i \in \mathbb{N}}$ for J such that for all $x \in \mathbb{R}^p$*

$$J(x) = \text{cl}\{V_i(x)\}_{i \in \mathbb{N}}.$$

Proof : Let V_1 be any definable selection of J , such a V_1 exists by [26, 4.5]. Set by recursion, for $i = 2, \dots$, V_i a definable selection of

$$x \rightrightarrows \arg \max_{v \in J(x)} \text{dist} \left(v, \{V_j(x)\}_{j=1}^{i-1} \right).$$

which is definable. By a simple covering argument, using compactity, for all x

$$\lim_{i \rightarrow \infty} \max_{v \in J(x)} \text{dist} \left(v, \{V_j(x)\}_{j=1}^{i-1} \right) = 0,$$

which shows that the constructed sequence has the desired property. $\square$

Lemma 9 *Using the notations of Theorem 8, there is a sequence of definable selections, $(y_i)_{i \in \mathbb{N}}$ such that for all $i \in \mathbb{N}$, and all $x \in \mathbb{R}^p$, $y_i(x) \in P(x)$ and*

$$D_f(x) = \{u, \exists y \in P(x), (u, 0) \in D(x, y)\} = \text{cl} \{u, (u, 0) \in D(x, y_i(x))\}.$$

Proof : Note that D_f has compact values and is definable by Proposition 1. Using Lemma 8 we have a definable sequence $(u_i)_{i \in \mathbb{N}}$ of selections in D_f such that $D_f(x) = \text{cl}\{u_i(x)\}_{i \in \mathbb{N}}$. By definability, for $i \in \mathbb{N}$ using [26, 4.5], we can choose a definable sequence $(y_i)_{i \in \mathbb{N}}$ such that for all $x \in \mathbb{R}^p$,

$$\begin{aligned} y_i(x) &\in P(x) \\ (u_i(x), 0) &\in D(x, y_i(x)). \end{aligned}$$

We have for all $x \in \mathbb{R}^p$

$$\lim_{i \rightarrow \infty} \max_{u \in D_f(x)} \text{dist} \left(u, \{ \tilde{u}, (\tilde{u}, 0) \in D(x, y_j(x)) \}_{j=1}^{i-1} \right) \leq \lim_{i \rightarrow \infty} \max_{u \in D_f(x)} \text{dist} \left(u, \{u_j(x)\}_{j=1}^{i-1} \right) = 0,$$

which shows that the constructed sequence has the desired property. □

Acknowledgments. The author would like to thank Jérôme Bolte and Rodolfo Rios-Zeruche for interesting discussions which helped putting this work together. The author acknowledge the support of ANR-3IA Artificial and Natural Intelligence Toulouse Institute, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA9550-19-1-7026, FA9550-18-1-0226, and ANR MaSDOL - 19-CE23-0017-01.

References

- [1] Ablin, P. and Peyré, G. and Moreau, T. (2020). Super-efficiency of automatic differentiation for functions defined as a minimum. In International Conference on Machine Learning.
- [2] Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., Kolter, Z. (2019). Differentiable convex optimization layers. Advances in neural information processing systems.
- [3] Aliprantis C.D., Border K.C. (2005) Infinite Dimensional Analysis (3rd edition) Springer
- [4] Ambrosio L., Gigli N. and Savaré G. (2008). Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media.
- [5] Amos, B., Kolter, J. Z. (2017). Optnet: Differentiable optimization as a layer in neural networks. In International Conference on Machine Learning.
- [6] Arjovsky, Chintala, Bottou (2017). Wasserstein GAN. International Conference on Machine Learning.
- [7] Attouch, H., Bolte, J. and Svaiter, B. F. (2013). Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Mathematical Programming, 137(1), 91-129.

- [8] Benaïm M., Hofbauer J. and Sorin S. (2005). Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1), 328-348.
- [9] Bianchi, P., Hachem, W. and Schechtman, S. (2020). Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *arXiv preprint arXiv:2005.08513*.
- [10] Berthet, Q., Blondel, M., Teboul, O., Cuturi, M., Vert, J. P., Bach, F. (2020). Learning with differentiable perturbed optimizers. *Advances in neural information processing systems*.
- [11] Bertsekas D. P. (1971). Control of uncertain systems with a set-membership description of the uncertainty. Doctoral dissertation, Massachusetts Institute of Technology.
- [12] Bolte J., Daniilidis A., Lewis A. and Shiotani, M. (2007). Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2), 556-572.
- [13] Bolte, J., Sabach, S., & Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1), 459-494.
- [14] Bolte, J. and Pauwels, E. (2020). Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning. *Mathematical Programming*.
- [15] Bolte, J. and Pauwels, E. (2020). A mathematical model for automatic differentiation in machine learning. *Proceedings of the conference on neural information processing systems*.
- [16] Borwein J. M. and Moors, W. B. (1998). A chain rule for essentially smooth Lipschitz functions. *SIAM Journal on Optimization*, 8(2), 300-308.
- [17] Borwein J., Moors W. and Wang, X. (2001). Generalized subdifferentials: a Baire categorical approach. *Transactions of the American Mathematical Society*, 353(10), 3875-3893.
- [18] Borwein J. M. (2017). Generalisations, Examples, and Counter-examples in Analysis and Optimisation. *Set-Valued and Variational Analysis*, 25(3), 467-479.
- [19] C. Castera, J. Bolte, C. Févotte and E. Pauwels (2019). An inertial newton algorithm for deep learning. *arXiv preprint arXiv:1905.12278*.
- [20] Clarke F. H. (1983). *Optimization and nonsmooth analysis*. Siam.
- [21] Coste M. (1999) An introduction to ϕ -minimal geometry. RAAG notes, Institut de Recherche Mathématique de Rennes.

- [22] Danskin J. M. (1966). The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4), 641-664.
- [23] Davis D., Drusvyatskiy D., Kakade S., and Lee J. D. (2020). Stochastic sub-gradient method converges on tame functions, 20(1), 119-154. *Foundations of Computational Mathematics*.
- [24] Davis, D. and Drusvyatskiy, D. (2021). Conservative and semismooth derivatives are equivalent for semialgebraic maps. *arXiv preprint arXiv:2102.08484*.
- [25] Dem'Yanov V. F. (1966). On the solution of several minimax problems. I. *Cybernetics*, 2(6), 47-53.
- [26] van den Dries L. and Miller C. (1996). Geometric categories and o-minimal structures. *Duke Math. J*, 84(2), 497-540.
- [27] Evans L. C. and Gariepy R. F. (2015). *Measure theory and fine properties of functions*. Revised Edition. Chapman and Hall/CRC.
- [28] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*.
- [29] Goodfellow, I. J., Shlens, J., Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International conference on learning representations*.
- [30] Jin, C., Netrapalli, P. and Jordan, M. (2020). What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning* (pp. 4880-4889). PMLR.
- [31] Kong, W. and Monteiro, R. D. (2019). An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *arXiv preprint arXiv:1905.13433*.
- [32] Lewis, A. and Tian, T. (2021). The structure of conservative gradient fields. *arXiv preprint arXiv:2101.00699*.
- [33] Lin, T., Jin, C. and Jordan, M. (2020). On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning* (pp. 6083-6093). PMLR.
- [34] Ostrovskii, D. M., Lowy, A. and Razaviyayn, M. (2021). Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4), 2508-2538.
- [35] Rafique, H., Liu, M., Lin, Q. and Yang, T. (2021). Weakly-convex-concave min-max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, 1-35.

- [36] Rios-Zertuche R. (2020). Examples of pathological dynamics of the subgradient method for Lipschitz path-differentiable functions. arXiv preprint arXiv:2007.11699.
- [37] Rockafellar R. T. (1985). Extensions of subgradient calculus with applications to optimization. *Nonlinear Analysis: Theory, Methods & Applications*, 9(7), 665-698.
- [38] Rockafellar R. T. and Wets R. J. B. (1998). *Variational analysis* (Vol. 317). Springer Science & Business Media.
- [39] H. Royden, P. Fitzpatrick (2010) *Real Analysis* Prentice Hall
- [40] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. (2014). Intriguing properties of neural networks. *International Conference on Learning Representations*.
- [41] Thekumparampil, K. K., Jain, P., Netrapalli, P. and Oh, S. (2019). Efficient algorithms for smooth minimax optimization. arXiv preprint arXiv:1907.01543.
- [42] Valadier M. (1989). Entraînement unilatéral, lignes de descente, fonctions lipschitziennes non pathologiques. *Comptes rendus de l'Académie des Sciences*, 308, 241-244.
- [43] Wang X. (1995). Pathological Lipschitz functions in $\mathbb{R}^n$. Master Thesis, Simon Fraser University.
- [44] Wang Y. and Zhang G. and Ba J. (2020) On Solving Minimax Optimization Locally: A Follow-the-Ridge Approach *International Conference on Learning Representations*.
- [45] Whitney H. (1935). A function not constant on a connected set of critical points. *Duke Mathematical Journal*, 1(4), 514-517.
- [46] Wilkie A. J. (1999). A theorem of the complement and some new o-minimal structures. *Selecta Mathematica*, 5(4), 397-421.