WORKING PAPERS

N° 1044

October 2019

“Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning”

Jérôme Bolte and Edouard Pauwels

Toulouse School of Economics

Conservative set valued fields, automatic differentiation, stochastic gradient methods and deep learning

Jérôme Bolte[*] and Edouard Pauwels[†]

October 7, 2019

Abstract

Modern problems in AI or in numerical analysis require nonsmooth approaches with a flexible calculus. We introduce generalized derivatives called conservative fields for which we develop a calculus and provide representation formulas. Functions having a conservative field are called path differentiable: convex, concave, Clarke regular and any semialgebraic Lipschitz continuous functions are path differentiable. Using Whitney stratification techniques for semialgebraic and definable sets, our model provides variational formulas for nonsmooth automatic differentiation oracles, as for instance the famous backpropagation algorithm in deep learning. Our differential model is applied to establish the convergence in values of nonsmooth stochastic gradient methods as they are implemented in practice.

Keywords. Deep Learning, Automatic differentiation, Backpropagation algorithm, Nonsmooth stochastic optimization, Definable sets, o-minimal structures, Stochastic gradient, Clarke subdifferential, First order methods

[*] Toulouse School of Economics, Université Toulouse 1 Capitole, France.

[†] IRIT, Université de Toulouse, CNRS. DEEL, IRT Saint Exupery, Toulouse, France.

Contents

1 Introduction

Classical approaches to solution methods for nonsmooth equations or nonsmooth optimization come from the calculus of variations [42, 46, 5, 12, 19, 32, 43, 48]. They have been successfully used in several contexts, from partial differential equations to machine learning. But most of the advances made in these last decades apply to classes revolving around convex-like non differentability phenomena: convex functions, semiconvex functions or (Clarke) regular problems. On the other hand several major problems arising in machine learning, numerical analysis, or non regular dynamical systems are not covered by these regularity models due to various calculus restrictions and the necessity of decomposing algorithms, see e.g., [17] and references therein. We propose a notion of generalized derivatives and identify a class of locally Lipschitz functions, called path differentiable functions, for which we obtain a flexible calculus, should we accept to use weaker generalized derivatives than standard ones. Our starting point is extremely elementary, we see derivation as an inverse operation to integration:

$$f(y) - f(x) = \int_x^y f'(t)\mathrm{d}t.$$

We thus introduce and study graph closed set valued mappings $D_f : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, and locally Lipschitz functions $f \colon \mathbb{R}^p \mapsto \mathbb{R}$ related by

$$f(\gamma(1)) - f(\gamma(0)) = \int_0^1 \langle D_f(\gamma(t)), \dot{\gamma}(t)\rangle \mathrm{d}t, \qquad \forall \gamma \in AC([0,1], \mathbb{R}^p),$$

where we use Aumann's integration while $AC([0,1], \mathbb{R}^p)$ is the set of absolutely continuous functions from $[0,1]$ to $\mathbb{R}^p$. Rephrasing this property yields a generalized form of the zero circulation property

$$\int_0^1 \langle D(\gamma(t)), \dot{\gamma}(t)\rangle \,\mathrm{d}t = \{0\}\,, \qquad \forall \gamma \in AC([0,1], \mathbb{R}^p).$$

We naturally call these objects conservative set valued fields and a function having a conservative field is called path differentiable. Convex, concave, Clarke regular, but also *any* semialgebraic Lipschitz continuous functions or Whitney stratifiable functions are path differentiable.

We provide a calculus and several characterizations of conservativity. First we show that conservative fields are classical gradients almost everywhere, which makes the Clarke subdifferential a minimal convex conservative field. Second, in the framework of semialgebraic or o-minimal structures, we provide conservative fields with a variational stratification formula [9]. This connection between Whitney stratification and conservativity allows to generalize known qualitative properties from the smooth world to definable conservative fields: Morse-Sard theorem, Kurdyka-Lojasiewicz inequality [34] and convergence of differential inclusions.

On a more applied side, conservative fields allow to analyze fundamental modern numerical algorithms in machine learning or numerical analysis based on automatic differentiation [49, 28] and decomposition [17, 24] in a nonsmooth context. Automatic differentiation is indeed proved to yield conservative fields which allows in turn to study discrete stochastic algorithms that are massively used to train AI systems. We illustrate this with the problem of training nonsmooth deep neural networks which are designed to perform prediction tasks based on a large labeled database [37].

Our work connects very applied concerns with the recent theory of o-minimal structures, by revealing surprising links between the massively used numerical libraries (Tensorflow [1], Pytorch, [44]), and Whitney stratifications.

Structure of the paper As a conclusion, let us mention that this article contains material that can be considered as having distinct and independent interests. Researchers working in analysis may focus on Section 2 and 3, which present conservative fields. Those having affinity with geometry can also go through Section 4 which provides insights into the semialgebraic and the definable cases. The more applied sections provide theorems which we believe useful to several communities: Section 5 is on nonsmooth automatic differentiation and a theoretical model for the corresponding "oracle", while Section 6 on studies stochastic gradient descent (with mini-batches) and deep learning.

2 Conservative set valued fields

Notations. We restrict our analysis to locally Lipschitz continuous functions in Euclidean spaces[1]. Take $p \in \mathbb{N}$. A locally Lipschitz continuous function, $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ is differentiable almost everywhere by Rademacher's theorem, see for example [26]. Denote by $R \subset \mathbb{R}^p$, the full measure set where f is differentiable, then the Clarke subgradient of f is given for any $x \in \mathbb{R}^p$, by

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

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

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

D is said to have *closed graph* or to be *graph-closed* if $\mathrm{graph}\, D$ 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

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

An *absolutely continuous curve* is a continuous function $x\colon \mathbb{R} \mapsto \mathbb{R}^p$ which admits a derivative $\dot{x}$ for Lebesgue almost all $t \in \mathbb{R}$, such that $\dot{x}$ is Lebesgue measurable and $x(t) - x(0)$ is the Lebesgue integral of $\dot{x}$ between 0 and t for all $t \in \mathbb{R}$. Absolutely continuous curves are well suited to generalize differential equations to differential inclusions [4].

[1] Although all results we provide are generalizable to complete Riemannian manifolds

Given a set valued map $D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$, $x_0 \in \mathbb{R}^p$, $x \colon \mathbb{R} \mapsto \mathbb{R}^p$ is *a solution to the differential inclusion problem*

$$\begin{aligned} \dot{x} &\in D(x) \\ x(0) &= x_0, \end{aligned}$$

if x is an absolutely continuous curve satisfying $x(0) = x_0$ and $\dot{x}(t) = D(x(t))$ for almost all t on a non trivial interval containing 0.

2.1 Definition and vanishing circulations

Throughout this section, we denote by $D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ a set valued map with closed graph and nonempty compact values. The following lemma is derived from results from the overview textbook [2].

Lemma 1 *Let $D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a set valued map with nonempty compact values and closed graph. Let $\gamma \colon [0,1] \mapsto \mathbb{R}^p$ be an absolutely continuous path. Then the following function*

$$t \mapsto \max_{v \in D(\gamma(t))} \langle \dot{\gamma}(t), v \rangle ,$$

defined almost everywhere on $[0,1]$*, is Lebesgue measurable.*

Proof : Consider the function $\Gamma \colon [0,1] \mapsto \mathbb{R}^p \times \mathbb{R}^p$ defined for almost all $t \in [0,1]$ by

$$\Gamma(t) = \begin{pmatrix} \gamma(t) \\ \dot{\gamma}(t) \end{pmatrix}.$$

Γ is, by definition, Lebesgue integrable, in particular, pre-images of Borel sets are Lebesgue sets. In addition, we consider the set-valued map $\tilde{D} \colon \mathbb{R}^p \times \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ defined as

$$\tilde{D} \colon (x, y) \rightrightarrows (D(x), y).$$

$\tilde{D}$ has closed graph and by [2, Theorem 18.20] it is measurable in the sense of [2, Definition 18.1]. Let $f \colon \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ be such that

$$f(x, y, v_1, v_2) = \langle y, v_1 \rangle$$

Then the map

$$m \colon (x, y) \mapsto \max_{w \in \tilde{D}(x,y)} f(x, y, w) = \max_{v \in D(x)} \langle y, v \rangle$$

is Borel measurable according to [2, Theorem 18.19]. This means that preimage of Borel sets are Borel sets. The function

$$t \mapsto \max_{v \in D(\gamma(t))} \langle \dot{\gamma}(t), v \rangle$$

is just $m \circ \Gamma$ and is hence Lebesgue measurable. □

We can now proceed with the central definition of a conservative set valued field.

Definition 1 (Conservative set valued fields) Let $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a set valued map. D is a *conservative (set valued) field* whenever it has closed graph, non empty compact values and for any absolutely continuous loop $\gamma\colon [0,1] \mapsto \mathbb{R}^p$, that is $\gamma(0) = \gamma(1)$, we have

$$\int_0^1 \max_{v \in D(\gamma(t))} \langle \dot\gamma(t), v\rangle \,\mathrm{d}t = 0$$

where the integral is understood in the Lebesgue sense[2]. It is equivalent to require

$$\int_0^1 \min_{v \in D(\gamma(t))} \langle \dot\gamma(t), v\rangle \,\mathrm{d}t = 0$$

for all loops γ.

Remark 1 (min, max circulations and conservativity) The min formula is indeed obtained by using the the reverse path $\tilde\gamma(t) = \gamma(1-t)$:

$$\begin{aligned}
\int_0^1 \max_{v \in D(\tilde\gamma(t))} \left\langle \dot{\tilde\gamma}(t), v\right\rangle \mathrm{d}t &= \int_0^1 \max_{v \in D(\gamma(1-t))} \langle -\dot\gamma(1-t), v\rangle \,\mathrm{d}t \\
&= -\int_0^1 \min_{v \in D(\gamma(1-t))} \langle \dot\gamma(1-t), v\rangle \,\mathrm{d}t \\
&= \int_1^0 \min_{v \in D(\gamma(t))} \langle \dot\gamma(t), v\rangle \,\mathrm{d}t \\
&= -\int_0^1 \min_{v \in D(\gamma(t))} \langle \dot\gamma(t), v\rangle \,\mathrm{d}t = 0.
\end{aligned}$$

We deduce that for almost all $t \in [0,1]$, $\max_{v \in D(\tilde\gamma(t))} \langle \dot\gamma(t), v\rangle = \min_{v \in D(\tilde\gamma(t))} \langle \dot\gamma(t), v\rangle$.

Remark 2 (Vanishing circulation and conservativity) There is a measurable argmax selection in Lemma 1 (see [2, Theorem 18.19]) so that for any measurable selection $v\colon [0,1] \mapsto \mathbb{R}^p$, $v(t) \in D(\gamma(t))$ for all t, we have $\int_0^1 \langle \dot\gamma(t), v(t)\rangle \,\mathrm{d}t = 0$. Thus, in the setting of Definition 1, an equivalent characterization is that the Aumann integral of $t \rightrightarrows \langle \dot\gamma(t), D(\gamma(t))\rangle$ is $\{0\}$. In short

$$\int_0^1 \langle D(\gamma(t)), \dot\gamma(t)\rangle \,\mathrm{d}t = \{0\}, \tag{1}$$

exactly means that D is conservative. We recover the standard definition of conservativity as fields with vanishing circulation.

[2] which is possible thanks to Lemma 1.

2.2 Locally Lipschitz continuous potentials of conservative fields

Definition 2 (Potential functions of conservative fields) Let $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a conservative field. A function f defined through any of the equivalent forms

$$\begin{aligned} f(x) &= f(0) + \int_0^1 \max_{v \in D(\gamma(t))} \langle \dot\gamma(t), v\rangle \,\mathrm{d}t & (2)\\ &= f(0) + \int_0^1 \min_{v \in D(\gamma(t))} \langle \dot\gamma(t), v\rangle \,\mathrm{d}t & (3)\\ &= f(0) + \int_0^1 \langle \dot\gamma(t), D(\gamma(t))\rangle \,\mathrm{d}t & (4)\end{aligned}$$

where γ is an arbitrary absolutely continuous path joining 0 to x, is well and uniquely defined up to a constant. It is called a *potential function for* D. We shall also say that D *admits* f *as a potential,* or that D *is a conservative field for* f.

Remark 3 (a) To see that the definitions (2), (3) and (4) are indeed equivalent and independent of the chosen path, one adapts classical ideas as follows. Consider any $x \in \mathbb{R}^p$, and any absolutely continuous paths γ_1, γ_2 such that $\gamma_1(0) = \gamma_2(0) = 0$ and $\gamma_1(1) = \gamma_2(1) = x$. We have

$$\begin{aligned} &\int_0^1 \max_{v \in D(\gamma_1(t))} \langle \dot\gamma_1(t), v\rangle \,\mathrm{d}t - \int_0^1 \min_{v \in D(\gamma_2(t))} \langle \dot\gamma_2(t), v\rangle \,\mathrm{d}t\\ = &\int_0^1 \max_{v \in D(\gamma_1(t))} \langle \dot\gamma_1(t), v\rangle \,\mathrm{d}t + \int_0^1 \max_{v \in D(\gamma_2(t))} - \langle \dot\gamma_2(t), v\rangle \,\mathrm{d}t\\ = &\int_0^{\frac{1}{2}} \max_{v \in D(\gamma_1(2t))} \langle 2\dot\gamma_1(2t), v\rangle \,\mathrm{d}t + \int_{\frac{1}{2}}^1 \max_{v \in D(\gamma_2(2-2t))} \langle -2\dot\gamma_2(2-2t), v\rangle \,\mathrm{d}t\\ = &\, 0\end{aligned}$$

since the concatenation of $t \mapsto \gamma_1(2t)$ for $0 \le t \le 1/2$ and $t \mapsto \gamma_2(2-2t)$ for $1/2 \le t \le 1$ is an absolutely continuous loop. This shows that the value of the integral does not depend on the path. The "minimum and maximum integrals" are thus equal and we may set for any $x \in \mathbb{R}^p$:

$$f(x) = \int_0^1 \max_{v \in D(\gamma(t))} \langle \dot\gamma(t), v\rangle \,\mathrm{d}t = \int_0^1 \min_{v \in D(\gamma(t))} \langle \dot\gamma(t), v\rangle \,\mathrm{d}t$$

for any γ absolutely continuous with $\gamma(0) = 0$ and $\gamma(1) = x$. The right hand-side in (4) is thus a single number, and the identity is therefore well defined.
(b) If f is differentiable, ∇f is of course a conservative field (it is not unique). More examples and a discussion are provided in Subsection 3.2.
(c) The definition can be directly extended to star-shaped domains.
(d) The potential function f is locally Lipschitz continuous. Indeed take a bounded set S. Take $x, y \in S$ and use (c) above with the path $[0,1] \ni t \to \gamma(t) = tx + (1-t)y$,

$$|f(y) - f(x)| \le |y - x| \int_0^1 \max_{v \in D(\gamma(t))} |v| \mathrm{d}t \le M|x - y|$$

where M is such that

$$M \geq \max\{|v| : x \in \overline{\mathrm{conv}\, S}, v \in D(x)\}$$

with $\mathrm{conv}\, S$ being the convex envelope of S. From [15, Lemma 3], D is locally bounded and such a finite constant must exist.
(e) If D_1, D_2 are two graph-closed set valued mappings with compact nonempty values, then $D_1 \subset D_2$ and D_2 conservative implies that D_1 is conservative as well.
Observe also that if D is conservative $x \rightrightarrows \mathrm{conv}(D(x))$ is conservative as well.

Chain rule characterizes conservativity in the following sense:

Lemma 2 (Chain rule and conservativity) *Let $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a locally bounded, graph-closed set valued map and $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ a locally Lipschitz continuous function. Then D is a conservative field for f, if and only if for any absolutely continuous curve $x\colon [0,1] \mapsto \mathbb{R}^p$, the function $t \mapsto f(x(t))$ satisfies*

$$\frac{d}{dt} f(x(t)) = \langle v, \dot{x}(t) \rangle \qquad \forall v \in D_f(x(t)), \tag{5}$$

for almost all $t \in [0,1]$.

Proof : The reverse implication is obvious, using Lemma 1, integrating the characterization in equation (5) we obtain any of the equivalent equations of Definition 2. To prove the converse, assume now that D is a conservative field for f. For any $0 < s < 1$, we have

$$\begin{aligned} f(x(s)) - f(x(0)) &= \int_0^1 \max_{v \in D(x(st))} \langle s\dot{x}(st), v \rangle \,\mathrm{d}t \\ &= \int_0^s \max_{v \in D(x(t))} \langle \dot{x}(t), v \rangle \,\mathrm{d}t \\ &= \int_0^s \min_{v \in D(x(t))} \langle \dot{x}(t), v \rangle \,\mathrm{d}t \end{aligned}$$

The fundamental theorem of calculus states that $s \mapsto f(x(s))$ is differentiable almost everywhere and for almost all $s \in [0,1]$,

$$\frac{d}{ds} f(x(s)) = \max_{v \in D(x(s))} \langle \dot{x}(s), v \rangle = \min_{v \in D(x(s))} \langle \dot{x}(s), v \rangle = \langle \dot{x}(s), v \rangle \,,$$

for all $v \in D(x(s))$. This shows that f is a potential for D. □

3 A generalized differential calculus

3.1 Conservativity, Clarke subdifferential and gradient a.e.

We start with the following fundamental result.

Theorem 1 (A conservative field is a gradient almost everywhere) *Let $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a conservative field and $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ a (locally Lipschitz continuous) potential for D. Then $D = \{\nabla f\}$ Lebesgue almost everywhere.*

Proof : Fix a measurable selection $a\colon \mathbb{R}^p \mapsto \mathbb{R}^p$ of D and f a potential for D. Measurable selections exist because D has closed graph, with compact and nonempty values, and hence is measurable in the sense of [2, Definition 18.1] so that [2, Corollary 18.15] applies (see also Lemma 1). Fix a direction $v \in \mathbb{R}^p$, $x \in \mathbb{R}^p$ a base point, let $s < t$ be real numbers and γ the path $\gamma(\tau) = (1-\tau)(x+sv) + \tau(x+tv)$, then by using conservativity and an elementary change of variable, we obtain for all $x \in \mathbb{R}^p$

$$f(x+tv) - f(x+sv) = \int_s^t \langle v, a(x+\tau v)\rangle d\tau.$$

Using the fundamental theorem of calculus (in its Lebesgue form), one obtains

$$f'(y;v) = \langle v, a(y)\rangle \tag{6}$$

almost everywhere on the line $x + \mathbb{R}v$, where

$$f'(y;v) := \lim_{r\to 0,\, r\neq 0} \frac{f(y+rv) - f(y)}{r}$$

when the limit exists. Since f is continuous, the two functions defined for all $y \in \mathbb{R}^p$.

$$\begin{aligned} f_u'(y;v) &:= \lim \sup_{s\to 0,\, s\neq 0} \frac{f(y+sv)-f(y)}{s} = \lim_{k\to\infty} \sup_{0<|s|\leq 1/k} \frac{f(y+sv)-f(y)}{s} \\ f_l'(y;v) &:= \lim \inf_{s\to 0,\, s\neq 0} \frac{f(y+sv)-f(y)}{s} = \lim_{k\to\infty} \inf_{0<|s|\leq 1/k} \frac{f(y+sv)-f(y)}{s}, \end{aligned}$$

are Borel, hence Lebesgue measurable. Consider the following set

$$A = \left\{y \in \mathbb{R}^p,\ f_u'(y;v) \neq \langle v, a(y)\rangle \text{ or } f_l'(y;v) \neq \langle v, a(y)\rangle\right\}.$$

This set is Lebesgue measurable and for any $y \in \mathbb{R}^p \setminus A$, we have

$$f'(y;v) = f_u'(y;v) = f_l'(y;v) = \langle v, a(y)\rangle.$$

Furthermore, using (6) we have $\mathcal{H}^1(A \cap (x + \mathbb{R}v)) = 0$, where $\mathcal{H}^1$ is the Hausdorff measure of dimension 1. Since $x \in \mathbb{R}^p$ was arbitrary, we actually have $\mathcal{H}^1(A \cap L) = 0$ for any line L, parallel to v and since A is measurable, Fubini's theorem entails that A has zero Lebesgue measure, and hence we have $f'(y;v) = \langle v, a(y)\rangle$ for almost all $y \in \mathbb{R}^p$. Now the Rademacher Theorem [26, Theorem 3.2], ensures that f is differentiable almost everywhere, this implies that $f'(y;v) = \langle \nabla f(y), v\rangle$ for almost all $y \in \mathbb{R}^p$ and hence, $\langle \nabla f(y), v\rangle = f'(y;v) = \langle v, a(y)\rangle$ for almost all $y \in \mathbb{R}^p$.

The direction v was chosen arbitrarily, we repeat the same construction for every $v \in \mathbb{Q}^p$ (which is countable) and obtain that $\langle \nabla f(y), v\rangle = \langle v, a(y)\rangle$ for almost all $y \in \mathbb{R}^p$ and every $v \in \mathbb{Q}^p$, that is $a(y) = \nabla f(y)$ for almost all $y \in \mathbb{R}^p$.

Since a was chosen as an arbitrary measurable selection for D, we may use [2, Corollary 18.15] which states that there is a sequence of measurable selections for D, $(a_k)_{k\in\mathbb{N}}$ such that for any $x \in \mathbb{R}^p$, $D(x) = \mathrm{cl}\,\{a_i(x)\}_{i\in\mathbb{N}}$. Using the previous Rademacher's argument for each i in $\mathbb{N}$, there exists a sequence of measurable sets $(S_i)_{i\in\mathbb{N}}$ which have all full measure and such that $a_i = \nabla f$ on S_i. Setting $S = \cap_{i\in\mathbb{N}} S_i$, we have that $\mathbb{R}^p \setminus S$ has zero measure and $a_i = \nabla f$ on S for all i in $\mathbb{N}$ and hence, using [2, Corollary 18.15], $D = \{\nabla f\}$ on S. This proves the desired result. □

An important consequence of the above result is that Clarke subdifferential appears as a minimal conservative field among convex valued conservative fields.

Corollary 1 (Clarke subgradient as a minimal convex conservative field) *Let $f \colon \mathbb{R}^p \mapsto \mathbb{R}$ be locally Lipschitz continuous and $D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a conservative field for f. Then $\partial^c f$ is a conservative field for f, and for all $x \in \mathbb{R}^p$,*

$$\partial^c f(x) \subset \mathrm{conv}(D(x)).$$

Proof : Let $S \subset \mathbb{R}^p$ be a full measure set such that $D = \nabla f$ on S (such a set exists by Theorem 1). Using [51, Proposition 2.2], we have, for any $x \in \mathbb{R}^p$

$$\partial^c f(x) = \mathrm{cl}\ \mathrm{conv}\left(\left\{\lim_{k\to\infty} \nabla f(x_k),\, x_k \in S,\, x_k \underset{k\to\infty}{\to} x\right\}\right).$$

Since D has closed graph and $D = \nabla f$ on S, we have

$$\mathrm{cl}\ \mathrm{conv}\left(\left\{\lim_{k\to\infty} \nabla f(x_k),\, x_k \in S,\, x_k \underset{k\to\infty}{\to} x\right\}\right) \subset \mathrm{cl}\ \mathrm{conv}\left(D(x)\right) = \mathrm{conv}\left(D(x)\right),$$

which allows to conclude. The fact that $\partial^c f$ is conservative, follows right from the definition and the previous inclusion. □

We deduce from Corollary 1 a Fermat's rule for conservative fields.

Proposition 1 (Fermat's rule) *Let $f \colon \mathbb{R}^p \mapsto \mathbb{R}$ be a potential for $D_f \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ with nonempty compact values and closed graph. Let $x \in \mathbb{R}^p$ be a local minimum or local maximum of f. Then $0 \in \mathrm{conv}(D_f(x))$.*

Proof : This is a consequence of Corollary 1 since Fermat's rule holds for the Clarke subdifferential [48, Theorems 9.61 and 10.1]. □

Given a fixed conservative field D with f as a potential, we say that x is *D-critical* for f if $D(x) \ni 0$. The value $f(x)$ is then called *a D-critical value.* This idea originates in [18].

Remark 4 The convex envelope in Fermat's rule is necessary. For example, let $D \colon x \mapsto \mathrm{sign}(x)$ with $D(0) = \{-1, 1\}$, then D has closed graph and is conservative for the absolute value. The origin is a global minimum of the potential, but $0 \notin D(0)$.

3.2 Path differentiability

Conservative fields convey a natural notion of "generalized differentiability", a function being differentiable if it admits a conservative field for which Definition 2 holds true. We call such functions path differentiable and provide a characterization in this section.

Definition 3 (Path differentiability) We say that a locally Lipschitz continuous function $f : \mathbb{R}^p \mapsto \mathbb{R}$ is *path differentiable* if f is the potential of a conservative field on $\mathbb{R}^p$.

We deduce from Corollary 1 the following characterization of path differentiable functions.

Corollary 2 (Characterization of path differentiable functions) *Let $f \colon \mathbb{R}^p \mapsto \mathbb{R}$ be locally Lipschitz continuous, then f is path differentiable if and only if its Clarke subgradient is a conservative field (in which case it admits f as a potential).*

The following property is sometimes called integrability, it has been studied for convex functions in [47] and for broader classes in, e.g., [22, 52, 13, 53].

Corollary 3 (Integrability and Clarke subdifferential) *Let $f \colon \mathbb{R}^p \mapsto \mathbb{R}$ and $g \colon \mathbb{R}^p \mapsto \mathbb{R}$ be two locally Lipschitz path differentiable functions such that $\partial^c g(x) \subset \partial^c f(x)$ for all $x \in \mathbb{R}^p$, then $f - g$ is constant.*

Proof : Path differentiability entails that subdifferentials are conservative fields by Corollary 1. The result follows by definition of potential functions through integration in Definition 2. □

We briefly compare some standard subgradients notions and conservative fields. We use the vocabulary and notations of [48].

Proposition 2 (Some path differentiable functions) *Let $f : \mathbb{R}^p \to \mathbb{R}$ be Lipschitz continuous, the following are sufficient conditions for f to be path differentiable*

(i) f is convex or concave.

(ii) f or $-f$ is Clarke regular.

(iii) f or $-f$ is prox regular.

(iv) f is real semialgebraic (or more generally tame, i.e., definable in some o-minimal structure).

Proof : Using the chain rule characterization, all proofs boil down to providing a chain rule with the Clarke subdifferential for each of the above mentioned situation. We refer to [48] for convex, Clarke and prox regular functions, [24] for tame functions. □

In general, conservative fields may be distinct from all other classical subdifferentials, even in the tame case. Define for instance $D : \mathbb{R} \to \mathbb{R}$ by $D(0) = \{-1, 0, 1\}$, $D(1) = [0, 2]$ and $D(x) = 0$ otherwise. It is a conservative field on $\mathbb{R}$ with any constant function as a potential function.

Remark 5 (Historical aspects) Our effort to define a subclass of locally Lipschitz continuous functions which has favorable differentiability properties is one attempt among many others. The closest idea we found is due to Valadier who introduced in 1989 the notion of *"fonctions saines"* [54]. Although Definition 2 looks much more general than the notion given in [54], the equivalent characterization of Corollary 2 shows that path-differentiable and "saines" functions are actually the same! Later on, at the end of the nineties, Borwein and Moors introduced the notion of essentially smooth functions (strictly differentiable almost eveywhere) as a well-behaved subclass of locally Lipschitz continuous functions [13]. Interestingly, the notion of saines functions was as well reconsidered and slightly modified in [14] to describe the larger class of arcwise essentially smooth functions. Following [54] and Chapter 1 of [55], we see that, in the univariate case, saine and essentially smooth functions coincide. This is no longer true for $p \geq 2$, the set of "fonctions saines" is a subset of essentially smooth functions.

Remark 6 (Genericity: theory and practice) The work of Wang et al. [55, 15] allows to claim that generic 1-Lispchitz functions are not path differentiable. Paradoxically, we shall see in further sections that most functions arising in applications are path differentiable (e.g. any semialgebraic or tame function is path differentiable)[3].

3.3 Conservative mappings and calculus

In this part we often identify linear mappings to their matrices in the canonical basis. For general conservative mappings, we adopt here a definition through the chain rule rather than circulations in order to simplify the exposition. However it would be relevant to provide a direct extension of Definition 1 involving vanishing circulations through set valued integration, this is matter for future work.

Definition 4 (Conservative mappings) Let $F \colon \mathbb{R}^p \mapsto \mathbb{R}^m$ be a locally Lipschitz function. $J_F \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{m \times p}$ is called a *conservative mapping* for F, if for any absolutely continuous curve $\gamma\colon [0,1] \mapsto \mathbb{R}^p$, the function $t \mapsto F(\gamma(t))$ satisfies

$$\frac{d}{dt}F(\gamma(t)) = V\dot{\gamma}(t) \qquad \forall V \in J_F(\gamma(t))$$

for almost all $t \in [0,1]$.

Remark 7 (Conservative fields are conservative mappings) Note that if $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is a conservative field for $f\colon \mathbb{R}^p \mapsto \mathbb{R}$, it is of course also a conservative mapping for f.

The following lemma provides an elementary but essential way to construct conservative matrices.

[3] Valadier's terminology finds here a surprising justification, since "saine", healthy in English, is chosen as the opposite of pathological

Lemma 3 (Componentwise aggregation) *Let $F\colon \mathbb{R}^p \mapsto \mathbb{R}^m$ be a locally Lipschitz continuous function. Let $J_F\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ be given by:*

$$J_F(x) = \begin{pmatrix} v_1^T \\ \vdots \\ v_m^T \end{pmatrix}, \qquad v_i \in D_i(x),\ i = 1\ldots, m, \quad \forall x \in \mathbb{R}^p,$$

where D_i is a conservative field for the i-th coordinate of F, $i = 1, \ldots, m$. Then J_F is a conservative mapping for F.

Proof : This follows Lemma 2, the product structure of J_F and the fact that a finite union of Lebesgue null sets is a Lebesgue null set. □

A partial converse holds true (thanks to Lemma 2): projection on rows of conservative mappings have to be conservative mappings for the corresponding coordinate function.

Lemma 4 (Coordinates of conservative mappings) *Let $F\colon \mathbb{R}^p \mapsto \mathbb{R}^m$ be locally Lipschitz continuous. Let $J_F\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ be a conservative mapping for F, then the projection of J on the first row of J, is a conservative field for the first coordinate of F.*

Observe however, that these "generalized Jacobians" may have a more complex structure than the product structure outlined in Lemma 3.

The following chain rule of generalized differentiation follows readily from the definition.

Lemma 5 (The product of conservative mappings is conservative) *Let $F_1\colon \mathbb{R}^p \mapsto \mathbb{R}^m$ and $F_2\colon \mathbb{R}^m \mapsto \mathbb{R}^l$ be locally Lipschitz continuous mappings. Let $J_1\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ be a conservative mapping for F_1 and $J_2\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{l\times m}$ be a conservative mapping for F_2. Then the product mapping $J_2 \cdot J_1$ is a conservative mapping for $F_2 \circ F_1$.*

Proof : Consider any absolutely continuous curve $\gamma\colon [0,1] \mapsto \mathbb{R}^p$. By local Lipschitz continuity, $t \mapsto F_1(\gamma(t))$ is also absolutely continuous and by definition of J_1 we have,

$$\frac{d}{dt}F_1(\gamma(t)) = J_1(\gamma(t))\dot{\gamma}(t) \qquad \text{a.e. on } (0,1).$$

Furthermore, $F_2 \circ F_1 \circ \gamma$ is also absolutely continuous by the local Lipschitz continuity of F_2. From the definition of J_2, we have

$$\frac{d}{dt}\left(F_2 \circ F_1(\gamma(t))\right) = J_2(F_1(\gamma(t))) \times \frac{d}{dt}\left(F_1(\gamma(t))\right) \qquad \text{a.e. on } (0,1)$$

The last two identities lead to the conclusion. □

We deduce the following chain rule by enlargement.

Lemma 6 (Outer chain rule) *Let $F\colon \mathbb{R}^p \mapsto \mathbb{R}^m$ and $g\colon \mathbb{R}^m \mapsto \mathbb{R}$ be locally Lipschitz continuous. Let $D_F\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^{m\times p}$ and $D_g\colon \mathbb{R}^m \rightrightarrows \mathbb{R}^m$ be some set valued mappings such that F_i, the i-th coordinate of F, is a potential for $[D_F(\gamma(t))]_i$, the i-th row of D_F for $i = 1\ldots m$ and g is a potential for D_g. Then $g\circ F$ is a potential of $D\colon x \rightrightarrows D_F(x)^T D_g(F(x))$.*

Proof : This is obtained combining Lemmas 3, 5 and Remark 7. □

A simple consequence is a "sum rule" by subdifferential enlargement which is fundamental in the study of the mini-batch stochastic gradient:

Corollary 4 (Outer sum rule) *Let $f_1, \ldots, f_n$ be locally Lipschitz continuous functions. Then $f = \sum_{i=1}^n f_i$ is a potential for $D_f = \sum_{i=1}^n D_i$ provided that f_i is a potential for each D_i, $i = 1, \ldots, n$.*

4 Tameness and conservativity

Let us beforehand provide two useful reading keys:

— The reader unfamiliar with definable objects can simply replace definability assumptions by semialgebraicity assumptions. It is indeed enough to treat major applications considered here, as for example deep learning with ReLU activation functions and square loss.

— Semialgebraicity and definability being easy to recognize in practice, the results in this section can be readily used as "black boxes" for applicative purposes.

4.1 Introduction and definition

We recall here the results of geometry that we use in the present work. Some references on this topic are [23, 25].

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 $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathcal{O}_{p+1}$;

(iii) if $\pi : \mathbb{R}^{p+1} \to \mathbb{R}^p$ 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 $g : \mathbb{R}^p \to \mathbb{R}$ is a polynomial function;

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

A subset of $\mathbb{R}^p$ which belongs to an o-minimal structure $\mathcal{O}$ is said to be *definable in* $\mathcal{O}$. Very often the o-minimal structure is fixed, so one simply says *definable* or *tame*. A set valued mapping is said to be definable in $\mathcal{O}$ whenever its graph is 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 $g_i, h_i : \mathbb{R}^p \to \mathbb{R}$ 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.

O-minimality is an extremely rich topological concept: major structures, such as globally subanalytic sets or sets belonging to the log-exp structure provides vast applicative opportunities (as deep learning with hyperbolic activation functions or entropic losses, see [24, 18] for some illustrations). We will not give proper definitions of these structures in this paper, but the interested reader may consult [25].

The tangent space at a point x of a manifold M is denoted by T_xM. Given a submanifold[4] M of a finite dimensional Riemannian manifold, it is endowed by the Riemanninan structure inherited from the ambient space. Given $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ and M a differentiable submanifold on which f is differentiable, we denote by $\operatorname{grad}_M f$ its Riemannian gradient or even, when no confusion is possible, $\operatorname{grad} f$.

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

Stratifications can have many properties, we refer to [25] and references therein for an account on this question and in particular for more on the idea of a Whitney stratification that we will use repeatedly. We pertain here to one basic definition: a C^r-stratification $\mathcal{S} = (M_i)_{i \in I}$ of a manifold M has the *Whitney-(a) property,* if for each $x \in \operatorname{cl} M_i \cap M_j$ (with $i \neq j$) and for each sequence $(x_k)_{k \in \mathbb{N}} \subset M_i$ we have:

$$\left.\begin{array}{ll} \lim\limits_{k\to\infty} x_k & = x \\ \text{and} & \\ \lim\limits_{k\to\infty} T_{x_k}M_i & = \mathcal{T} \end{array}\right\} \implies T_xM_j \subset \mathcal{T}$$

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.

4.2 Variational stratification and projection formulas

Let us fix an o-minimal structure $\mathcal{O}$, so that a set or a function will be called definable if it is definable in $\mathcal{O}$.

[4] We only consider embedded submanifolds

Definition 5 (Variational stratification) Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$, be locally Lipschitz continuous, let $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a set valued map and let $r \geq 1$. We say that the couple (f, D) has a C^r *variational stratification* if there exists a C^r Whitney stratification $\mathcal{S} = (M_i)_{i \in I}$ of $\mathbb{R}^p$, such that f is C^r on each stratum and for all $x \in \mathbb{R}^p$,

$$\mathrm{Proj}_{T_{M_x}(x)} D(x) = \{\mathrm{grad}\, f(x)\}\,, \tag{7}$$

where $\mathrm{grad}\, f(x)$ is the gradient of f restricted to the active strata M_x containing x.

The equations (7) are called *projection formulas* and are motivated by Corollary 9 in [9] which states that Clarke subgradients of tame functions have projection formulas.

Theorem 2 (Projection formula [9]) *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be definable, locally Lipschitz continuous*[5]*, and let $r \in \mathbb{N}$. Then there exists a finite C^r Whitney stratification $\mathcal{S} = (M_i)_{i \in I}$ of $\mathbb{R}^p$ such that for all $x \in \mathbb{R}^p$,*

$$\mathrm{Proj}_{T_x(M_x)} \partial^c f(x) = \{\mathrm{grad}\, f(x)\}\,.$$

In other words, the couple $(f, \partial^c f)$ has a a C^r variational stratification.

4.3 Characterization of tame conservative fields

The following is a direct extension of the chain rule result given in [24, Theorem 5.8]. It relies also on Theorem 2 and implies that a tame function f is a potential of its Clarke subgradient.

Theorem 3 (Integrability (from [24])**)** *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$, be locally Lispchitz continuous and let $D_f\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be compact valued and graph closed, with a C^1 projection formula for f. Then f is a potential of D_f.*

Proof : The proof given in [24] was proposed for the Clarke subdifferential but holds for larger classes of stratifiable set valued map. We reproduce the arguments here for clarity.

Let $\mathcal{S}$ be a stratification provided by the C^1 projection formula. Fix an absolutely continuous path $\gamma\colon [0,1] \mapsto \mathbb{R}^p$. Fix an arbitrary $t \in (0,1)$ and $M \in \mathcal{S}$ such that

$$\gamma(t) \in M, \quad \dot\gamma(t) \in T_M(\gamma(t)) \tag{8}$$

In this case, the projection formula ensures that for any $v \in D_f(\gamma(t))$

$$\langle \dot\gamma(t), v\rangle = \langle \dot\gamma(t), \mathrm{grad}\, f(\gamma(t))\rangle = \frac{d}{dt} f(\gamma(t)).$$

Set $\Omega_X = \{t \in [0,1], \gamma(t) \in M, \quad \dot\gamma(t) \not\in T_M(\gamma(t))\}$. Fix any $t_0 \in \Omega_M$, there exists a small closed interval I centered at t_0 such that $I \cap \Omega_M = \{t_0\}$, otherwise one would have $\dot\gamma(t_0) \in T_M(\gamma(t_0))$. The interval I may be chosen with rational endpoints and this gives

[5] In [9] the authors assume f to be arbitrary and obtain similar result, for simplicity we pertain to the Lipschitz case.

an injection from Ω_M to $\mathbb{Q}^2$. Hence Ω_X is countable. Since $\mathcal{S}$ contains only finitely many strata, for almost all $t \in [0, 1]$, relation (8) holds for some other $M \in \mathcal{S}$ and the result follows using the chain rule characterization of conservativity in Lemma 2. □

We aim at proving the following converse in the context of tame analysis.

Theorem 4 (Variational stratification for tame conservative fields) *Let $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ be a tame, nonempty compact valued, graph closed conservative field having a tame potential denoted by $f\colon \mathbb{R}^p \mapsto \mathbb{R}$. Then (f, D) has a C^r variational stratification, i.e. there exists a stratification $\{M_i\}_{i\in I}$ of $\mathbb{R}^p$ such that*

$$P_{T_xM_x} D(x) = \{\mathrm{grad}\, f(x)\}$$

whenever M_x is the active strata.

Proof : We actually establish a slightly stronger result and prove that the result holds by replacing the underlying space $\mathbb{R}^p$ by a C^r tame finite dimensional manifold M with $D(x) \subset T_xM$, i.e., D is a set valued section of the tangent bundle. We follow a classical pattern of stratification theory, see e.g., [25], and we establish that the failure of the projection formula may only occur on a convenient small set. More precisely, we shall provide a stratification $\mathcal{S} = \{M_1, \ldots, M_q\}$ of M for which each of the sets:

$$R_i = \{x \in M_i : \mathrm{grad}\, f(x) \neq P_{T_xM_i} D(x)\}, \text{ (with } i = 1, \ldots, q), \tag{9}$$

has a dimension strictly lower than M_i.

If we had such a stratification, let us see indeed how we would refine the stratification in order to downsize further the set of "bad points". For each R_i we would consider the stratification $R_i^1, \ldots, R_i^r$ of $f_{|R_i}$ into smooth functions. For each $j = 1, \ldots, r$ the couple $x \to f|_{R_i^j}(x), x \mapsto P_{T_xR_i^j} D(x)$ would satisfy the assumption of the theorem but on a manifold with strictly lower dimension. So we could then pursue the process and conclude by exhaustion.

To obtain $\mathcal{S}$, we consider a "constant-rank" Whitney stratification of f, $\mathcal{S} = \{M_1, \ldots, M_q\}$, i.e. such that f is smooth on each M_i and has a constant rank, 0 or 1 in our case (see [25]). Take M_i an arbitrary strata and R_i as in (9). We only need to prove that R_i has a lower dimension than M_i.

For simplicity set $R_i = R$, $M_i = M$. We consider first the rank 0 case and deduce the other case afterward.

Assume that $\mathrm{rank} f = 0$ on M, i.e., f is constant. We want to prove that for almost all x in M, $\max\{\|v\| : v \in P_{T_xM} Df(x)\} = 0$. We argue by contradiction and assume that for some $\bar{x}$ we have a ball of radius $\rho > 0$ on which the max is strictly greater than a positive real m. Consider the mapping $G\colon x \rightrightarrows \arg\max\{\|v\| : v \in P_{T_xM} Df(x)\}$ which is tame with nonempty compact values. Use the definable choice's theorem to obtain a tame single-valued selection H of G. H is a (nonsmooth) vector field on M that we may stratify in a way compatible with $B_M(\bar{x}, \rho)$. The ball $B_M(\bar{x}, \rho)$ must contain a stratum of maximal dimension and hence there exists $\hat{x} \in B_M(\bar{x}, r)$, $0 < \epsilon \leq \rho$ such that, $H(x) \in T_xM$ is smooth over $B(\hat{x}, \epsilon) \subset B(\bar{x}, \rho)$. We may thus consider a curve such that

$$\dot{\gamma} = H(\gamma),\ \gamma(0) = \hat{x}.$$

For this curve, which is non stationary, one has almost everywhere

$$\frac{d}{dt}f(\gamma(t)) = \max_{v \in Df(\gamma(t))} \langle \dot{\gamma}, v\rangle = \max_{v \in P_{T_xM}Df(\gamma(t))} \|v\|^2 \geq m^2 > 0,$$

which is in contradiction with the fact that f is constant. This concludes the null rank case.

Assume now that $\mathrm{rank}\, f = 1$, so that $\mathrm{grad}\, f$ is nonzero all throughout M. Consider $\tilde{D} = D - \mathrm{grad}\, f$ which is tame, convex valued and has a closed graph. By linearity of the integral $\tilde{D}$ is conservative and has zero as a potential function over M. Indeed if $\gamma : [0,1] \to M$ is an arbitrary absolutely continuous curve, we have the set valued identity:

$$\begin{aligned} \int_0^1 \langle \tilde{D}\gamma(t), \dot{\gamma}(t)\rangle \mathrm{d}t &= \int \langle D(\gamma(t)), \dot{\gamma}(t)\rangle \mathrm{d}t - \int_0^1 \langle \mathrm{grad}\, f(\gamma(t)), \dot{\gamma}(t)\rangle \mathrm{d}t \\ &= f(\gamma(1)) - f(\gamma(0)) - (f(\gamma(1)) - f(\gamma(0))) \\ &= 0. \end{aligned}$$

Since the null function has rank 0 on M, we deduce as above that $P_{T_xM}\tilde{D}(x) = \{0\}$ for almost all $x \in M$. Since $\tilde{D} = Df - \mathrm{grad}\, f$ and $\mathrm{grad}\, f(x) \in T_xM$ for all $x \in M$, we deduce that $P_{T_xM}Df(x) = \{\mathrm{grad}\, f(x)\}$ for almost all $x \in M$ which is what we needed to prove.□

Remark 8 (Alternative proof) Another method for proving Theorem 2 relies on the repeated use of Theorem 1. We chose to avoid the use of strong analysis results, as Rademacher theorem, and pertain to standard self-contained definable arguments.

4.4 Geometric and dynamical properties of definable conservative fields

This section describes some properties of definable conservative fields (with definable potential function). The ideas and proofs are direct generalizations of [9].

Theorem 5 (Nonsmooth Morse-Sard for D critical values) *Let $f \colon \mathbb{R}^p \mapsto \mathbb{R}$ be a definable locally Lipschitz continuous function and $D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ a definable conservative field for f. Then the set of D critical values $\{f(x),\, x \in \mathbb{R}^p,\, 0 \in D(x)\}$ is finite.*

Proof : The proof is as in [9, Corollary 5] and follows from the variational stratification property, applying the definable Sard theorem to each strata. This ensures that the set of critical values has zero Lebesgue measure in $\mathbb{R}$ and since it is definable, it is a finite set. □

The following is a generalization of the result of Kurdyka [34]

Theorem 6 (A nonsmooth KL inequality for conservative fields) *Let $f \colon \mathbb{R}^p \mapsto \mathbb{R}$ be a definable locally Lipschitz continuous function and $D \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ a definable conservative field for f. Then there exists $\rho > 0$, $\varphi \colon [0, \rho) \mapsto \mathbb{R}_+$, definable strictly increasing,*

C^1 on $(0,\rho)$ with $\varphi(0)=0$ and a continuous definable function $\chi\colon \mathbb{R}_+ \mapsto (0,+\infty)$, such that for all $x \in \mathbb{R}^p$ with $0 < |f(x)| \leq \chi(\|x\|)$ and $v \in D(x)$,

$$\|v\|\varphi'(|f(x)|) \geq 1.$$

Proof : This is deduced from the variational stratification property as in Theorem 14 of [9]. □

The following convergence result is a consequence and calls for many questions regarding nonsmooth generalized gradient systems [35].

Theorem 7 (Conservative fields curves have finite length) *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be a tame locally Lipschitz function and $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ a tame conservative field for f. Let $x\colon \mathbb{R}_+ \mapsto \mathbb{R}^p$ be a solution of the differential inclusion*

$$\dot{x} \in -\mathrm{conv}(D(x)),$$

then if x is bounded, x has finite length: $\displaystyle\int_0^{+\infty} \|\dot{x}\| < +\infty$ and in particular x is a convergent trajectory.

Proof :

Assume without loss of generality that D has convex values, set $\|D(x)\| := \min_{v\in D(x)} \|v\|$ for any $x \in \mathbb{R}^p$. We have

$$\frac{d}{dt} f(x(t)) = -\|D(x(t))\|^2$$

for almost all $t \geq 0$. We deduce that $t \mapsto f(x(t))$ has a limit, say 0. The limit points ω of x are entirely contained in a compact zone of $[f=0]$. Uniformize the nonsmooth KL inequality on a tubular neighborhood, say Z, of this zone (see [10, Lemma 6]), and finally assume that $x(t) \in Z$ for some $t \geq t_1$. On Z, set $\tilde{D}(x) = D(x) \times \varphi'(f(x))$ and observe that $\tilde{D}$ is a conservative field for $\varphi \circ f$, hence for almost all $t \geq t_1$

$$\begin{aligned}\frac{d}{dt}\varphi \circ f(x(t)) &= \langle \dot{x}(t), \varphi'(f(x))D(x)\rangle \\ &= -\|\dot{x}(t)\|^2\varphi'(f(x)) \leq -\|\dot{x}(t)\|\varphi'(f(x))\|D(x)\| \leq -\|\dot{x}(t)\|.\end{aligned}$$

Since $\varphi(f(x(t))$ tends to 0, we obtain that $\displaystyle\int_0^{+\infty} \|\dot{x}\| \leq \varphi(f(x(0)))$. □

5 Automatic differentiation

Automatic differentiation emerged in the 70's as a computational framework which allows to compute efficiently gradients of multivariate functions expressed through smooth elementary functions. When the function formula involves nonsmooth elementary functions the automatic differentiation approach fails to provide gradients. This issue is largely studied in [28, chapter 14] which discusses connections with Clarke generalized derivatives using notions such as "piecewise analyticity" or "stable domain". Let us mention [29] which developed piecewise linear approximation for functions which can be expressed using absolute value, min or max operators. This approach led to successful algorithmic developments [30] but may suffer from a high computational complexity and a lack of versatility (the Euclidean norm cannot be dealt with within this framework). Another attempt using the same model of branching programs was described in [33] where a qualification assumption is used to compute Clarke generalized derivatives automatically[6].

We provide now a transparent interpretation of automatic differentiation which provides conservative fields which do not correspond to any known sort of subgradients.

5.1 A functional framework: "closed formula functions"

Automatic differentiation deals essentially with composed functions, that is functions coming as "closed formulas". It presumes the existence of a chain rule and aggregates the basic derivation operations according to this principle. We refer to [28] for a detailed account. The purpose of this section is to demonstrate that our nonsmooth differentiation model is perfectly fit to deal with this approach.

The function f we consider now is accessible through a recursive algorithm which materializes an evaluation process built on a directed graph. This graph[7] is modelled by a discrete map called `parents` and a collection of known "elementary functions" g_k:

a) $q \in \mathbb{N}$, $q > p$

b) `parents` maps the set $\{p+1, \ldots, q\}$ into the set of tuples of the form $(i_1, \ldots, i_m)$ where $m \in \mathbb{N}$ and $i_1, \ldots, i_m$ range over $\{1, \ldots, q-1\}$ without repetition. It has the property that for any $k \in \{p+1, \ldots, q\}$, $\texttt{parents}(k)$ is a tuple without repetition over the indices $\{1, \ldots, k-1\}$.

c) $(g_i)_{i=p+1}^{q}$ such that for any $i = p+1, \ldots, q$, $g_i \colon \mathbb{R}^{|\texttt{parents}(i)|} \mapsto \mathbb{R}$.

```
Algorithm 1: Definition program of $f \colon \mathbb{R}^p \mapsto \mathbb{R}$
Input: $x = (x_1, \ldots x_p)$
for $k = p+1, p+2, \ldots q$ do
    Set:
        $x_k = g_k(x_{\texttt{parents}(k)})$
    where $x_{\texttt{parents}(k)} = (x_i)_{i \in \texttt{parents}(k)}$.
end for
Return: $x_q =: f(x)$.
```

This defines the function f through an operational evaluation program.

[6] From a practical point of view, qualification is hard to enforce or even check.

[7] Which we shall not define formally since it is not essential to our purpose.

Algorithm 2: Forward mode of automatic differentiation for f

```
Input: variables $(x_1, \dots x_q)$; $d_i = (d_{ij})_{j=1}^{|\texttt{parents}(k)|} \in D_i(x_{\texttt{parents}}(i))$, $i = p+1 \dots q$
1: Initialize: $\frac{\partial x_k}{\partial x_k} = 1$, $k = 1, \dots, p$.
2: for $k = p+1, \dots P$ do
3:     Compute:
           $$\frac{\partial x_k}{\partial x} = \sum_{j \in \texttt{parents}(k)} \frac{\partial x_j}{\partial x} d_{kj}$$
       where $x = (x_1, \dots, x_q)$.
4: end for
Return: $\frac{\partial x_P}{\partial x_{1,\dots,p}}$.
```

Example 1 The idea behind automatic differentiation is that the original function is given through a closed formula, which is then interpreted as a composed function in order to make its differentiation (or "subdifferentiation") amenable to simple chain rule computations. For instance for $f(x) = (x_1 x_2 + \tan x_2)(|x_1| + x_1 x_2 x_3)$, we may choose

$$\begin{aligned} & x_4 = g_4(x_1, x_2) = x_1 x_2,\ x_5 = g_5(x_2) = \tan x_2,\ x_6 = g_6(x_1) = |x_1|, \\ & x_7 = g_7(x_3, x_4) = x_3 x_4, \\ & x_8 = g_8(x_4, x_5, x_6, x_7) = (x_4 + x_5)(x_7 + x_6) \end{aligned}$$

where the $\texttt{parents}$ function is in evidence $\texttt{parents}(4) = \{1, 2\}$, $\texttt{parents}(5) = \{2\}$, $\texttt{parents}(6) = \{1\}$, $\texttt{parents}(7) = \{3, 4\}$, $\texttt{parents}(8) = \{4, 5, 6, 7\}$. Observe that the derivatives or subdifferentials of $g_4, \dots, g_8$ are known in closed form. Concerning $g_6 = |\cdot|$ one has $\partial^c g_6(0) = [-1, 1]$. Thus in practice we need to choose a specific element in that set, as 0, and perform the computation with this choice (see below the forward or backward differentiation modes).

5.2 Forward and backward nonsmooth automatic differentiation

In order to compute a conservative field for f, we need in addition the following:

d) For any $i = p+1, \dots, q$, $D_i \colon \mathbb{R}^{|\texttt{parents}(i)|} \rightrightarrows \mathbb{R}^{|\texttt{parents}(i)|}$ is a conservative field for g_i.

For example, D_i could be the Clarke subgradient of g_i if g_i is definable (a mere definable selection in the Clarke would also work). For instance in Example 1, one may set $D_6(0) = \{0\}$ or $D_6(0) = [0, 1]$. Given $(x_i)_{i=1}^q$ as computed in Algorithm 1, an algorithm to compute a conservative field of f is described in Algorithm 2. This is a direct implementation of the chain rule as described in Lemma 6. This ensures that the output of Algorithm 2 is a conservative field for the function f described in Algorithm 1. Furthermore, the reverse mode of automatic differentiation described in Algorithm 3 computes essentially the same quantity but with a lower memory and time footprint.

Theorem 8 (Forward and backward "autodiff" are conservative fields) *Let f be given as in Algorithm 1.*

Algorithm 3: Reverse Mode of automatic differentiation for f

```
Input: variables $(x_1, \dots x_q)$; a the map $\{\texttt{parents}(t)\}_{t\in\{1,\dots q\}}$; associated derivatives
    $d_i = (d_{ij})_{j=1}^{|\texttt{parents}(k)|} \in D_i(x_{\texttt{parents}}(i))$, $i = p+1 \dots q$
Initialize: $v = (0, 0, \dots, 0, 1) \in \mathbb{R}^q$
for $t = q, \dots p+1$ do
    for $j \in \texttt{parents}(t)$ do
        Update coordinate $j$ of $v$:
            $v[j] := v[t] d_{tj}$
    end for
end for
Return: $(v[1], v[2], \dots, v[p])$.
```

(i) Set for any $x \in \mathbb{R}^p$, $D(x) = \{v \in \mathbb{R}^p;\ \text{ output of Algorithm 2}\}$, for any choice of $d_k \in D_k$, $k = p+1, \dots, P$. Then $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is a conservative field for f.

(ii) The same results holds for Algorithm 3.

Proof : We substitute the functions $(g_k)_{k=p+1}^q$ by functions $(G_k)_{k=p+1}^q$, such that for each $k = p+1, \dots, q$

$$\begin{aligned} G_k \colon \mathbb{R}^q &\mapsto \mathbb{R}^q \\ x &\mapsto x + e_k(g_k(x_{\texttt{parents}(k)}) - x_k), \end{aligned}$$

where e_k is the k-th element of the canonical basis. Similarly, for all $k \in p+1, \dots, q$, we still denote by $D_k \colon \mathbb{R}^q \rightrightarrows \mathbb{R}^q$ the conservative field of g_k seen as a function of $x_1, \dots, x_q$ (simply add zeros to coordinates which do not correspond to parents of k). Then f as computed in Algorithm 1, is equivalently given by

$$f(x) = [(G_q \circ G_{q-1} \circ \dots \circ G_{p+1}(x_1, \dots, x_p, 0, \dots, 0))]_q = [G_q \circ G_{q-1} \circ \dots \circ G_p(x)]_q$$

where G_p maps the first p coordinates $(x_k)_{k=1}^p$ to the vector $((x_i)_{i=1}^p, (0)_{i=p+1}^q) \in \mathbb{R}^p$ and indexation $[\cdot]_q$ denotes the q-th coordinate of a q vector. For each $k = p+1 \dots, q$, $x \in \mathbb{R}^q$, the following "componentwise derivative" of G_k in a matrix form

$$L_k \colon x \mapsto \left\{ I - e_k e_k^T + e_k d^T, \quad d \in D_k(x) \right\}, \tag{10}$$

is a conservative mapping for G_k by Lemma 3. For each $k = p+1 \dots, q$, we choose one such matrix according to a fixed input of Algorithm 2, $d_k \in D_k(x_1, \dots, x_q)$,

$$J_k = I - e_k e_k^T + e_k d_k^T \in L_k(x_1, \dots, x_q) \tag{11}$$

For each $k = p+1, \dots q$, denote by M_k the matrix defined by blocks as follows

$$M_k = \begin{pmatrix} I_p \\ \frac{\partial x_1}{\partial x_{1,\dots,p}} \\ \vdots \\ \frac{\partial x_k}{\partial x_{1,\dots,p}} \\ 0 \end{pmatrix} \in \mathbb{R}^{q \times p}$$

where $\frac{\partial x_k}{\partial x_{1,\ldots,p}}$ is computed by Algorithm 2. Denote also by $J_p \in \mathbb{R}^{q \times p}$ the diagonal matrix which diagonal elements are 1 and the remainders are 0, the Jacobian of G_p. One can see that

$$M_k = J_k \times J_{k-1} \times \ldots \times J_{p+1} \times J_p$$

for all $k = p+1, \ldots q$. This is easily seen for M_{p+1} as Algorithm 2 computes $\frac{\partial x_1}{\partial x_{1,\ldots,p}} = d_1$. The rest is a simple recursion. In the end Algorithm 2 computes

$$\begin{aligned} e_q^T M_q &= e_q^T \times J_q \times J_{P-1} \times \ldots \times J_{p+1} \times J_p \\ &\in e_q^T \times L_q \times L_{q-1} \times \ldots \times L_{p+1} \times J_p \end{aligned}$$

Combining Lemma 5 and Lemma 4, the right hand side is a conservative field for f. Actually it can be seen from equations (10) and (11) that the right hand side consists precisely of all possible outputs of Algorithm 2 for all possible choices of d_k, $k = p+1, \ldots, q$. This proves the claim for Algorithm 2.

Regarding Algorithm 3, we will show that it computes the same quantity reversing the order of the products. Set for all $t = q, \ldots, p+1$, set $v_t \in \mathbb{R}^q$ to be the vector v obtained after step t of the "for loop" of Algorithm 3. We have $v_q = e_q + d_q = (I + d_q e_q^T)e^q$. An induction shows that for all $t = q, \ldots, p+1$

$$v_t = (I + d_t e_t^T) \ldots (I + d_q e_q^T) e^q.$$

Using the same notations as in equation (11), set for $t = q, \ldots p+1$

$$w_t = J_t^T \times \ldots \times J_q^T \times e_q$$

It is easy to see that w_q and v_q agree on the first $q-1$ coordinates. By recursion, for $t = q, \ldots p+1$, w_t and v_t agree on the first $t-1$ coordinates (recall that d_t is supported on the first $t-1$ coordinates). We deduce that w_{p+1} and v_{p+1} agree on the first p coordinates so that the output of Algorithm 3 is

$$J_{G_p}^T v_{p+1} = J_{G_p}^T w_{p+1} = J_{G_p}^T \times J_{G_{p-1}}^T \times \ldots \times J_{G_q}^T \times e^q = M_q^T e_q$$

which is the same quantity as computed by Algorithm 3. The claim follows. □

Remark 9 Automatic differentiation is not necessarily convex valued. Consider the function

$$f \colon (x, y) \mapsto |\max(x, y)|$$

Both the max and absolute value functions are convex so that their respective convex subgradients are conservative fields. Applying the chain rule in Lemma 5 at $x = y = 0$ we obtain a conservative field for f evaluated at zero of the form

$$D = \{tv;\ t \in [-1, 1],\ v \in \Delta\}$$

where Δ is the one dimensional simplex in $\mathbb{R}^2$. The set D is not convex.

The following corollary is a direct consequence of Theorems 2 and 8. Note that a result close to equation (12) was already guessed in [28, Proposition 14.2].

Corollary 5 (automatic differentiation for definable functions) *Assume that all the g_k defining f and their conservative fields D_k, are definable. Then f is differentiable almost everywhere, more precisely*

$$D_f = \{\nabla f\} \tag{12}$$

on the complement of finitely many smooth manifolds with dimension at most $p-1$. Furthermore, for any v, w in $\mathbb{R}^p$,

$$f(w) - f(v) = \int_0^1 \langle D_f((1-t)v + tw), w - v\rangle \,. \tag{13}$$

Proof : From Theorem 8, D_f is a conservative field for f. Basic closedness properties of definable objects ensure that both f and D_f are definable so that Theorem 4 ensures the existence of a variational stratification 5. The fact that f is differentiable almost everywhere is a basic result of tame geometry. To obtain (12), use the stratification provided in Theorem 4, and consider the dense open set given by the union of the finite number of strata of maximal dimensions. The integration formula is the application of Definition 2 along any segment. □

Remark 10 (The limitations of the smooth chain rule) It is surprising to use Theorem 4 which is non trivial to obtain (12). It is instead tempting to simply use the expression of f provided in Theorem 8:

$$f(x) = e_q^T G_p \circ \ldots \circ G_q(x)$$

and to differentiate it "almost everywhere" to obtain

$$f'(x) = e_q^T G_p' \left(G_{p-1}(\ldots G_q(x))\ldots)\right) \circ \ldots \circ G_q'(x),$$

which would give the desired result. Unfortunately this expression has no obvious meaning, since for instance, the image of G_q may be entirely contained in the points of non-differentiability of G_{q-1}, so that $G_{q-1}'(G_q(x))$ has no meaning. This result is illustrated further in the deep learning section through an experimental example.

6 Algorithmic consequences and deep learning

What follows is in the line of many works on decomposition methods [16], in particular those involving nonconvex problems, see e.g., [41, 17, 24, 39, 20]. Our study uses connections with dynamical systems, see e.g., [38, 36, 8, 11] in order to take advantage of continuous-like properties as null circulation in (2). Using our formalism, we gather ideas from [24, 18], and use the Benaïm-Hofbauer-Sorin approach [8], to obtain almost sure subsequential convergence to steady states that are carefully defined. To our knowledge, this provides the first proof for the subsequential convergence of SGD with mini-batches in deep learning when the actual backpropagation model is used instead of the subgradient one's, which is the case in almost all applications involving nonsmooth objects. As outlined in a conclusion, many more algorithms could be considered along this perspective.

All sets and functions we consider in this section are definable in the same o-minimal structure.

6.1 Mini-batch stochastic approximation for finite nonsmooth nonconvex sums aka "nonsmooth nonconvex SGD"

We consider the following loss function on $\mathbb{R}^p$

$$\mathcal{J} \colon w \mapsto \frac{1}{n}\sum_{i=1}^{n} f_i(w) \tag{14}$$

where each $f_i \colon \mathbb{R}^p \mapsto \mathbb{R}$ is definable and locally Lipschitz continuous. We assume that for each $i = 1 \ldots n$, $D_i \colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is a definable conservative field for f_i, for example the ones provided by automatic differentiation. We consider the following recursive process, given a sequence of nonempty mini-batches subsets of $\{1, \ldots, n\}$, $(B_k)_{k\in\mathbb{N}}$, taken independently, uniformly at random, $(\alpha_k)_{k\in\mathbb{N}}$ a deterministic sequence of positive step sizes, and $w_0 \in \mathbb{R}^p$, iterate

$$\begin{aligned} w_{k+1} &= w_k - \alpha_k d_k \\ d_k &\in \frac{1}{|B_k|}\sum_{i\in B_k} D_i(w_k) \end{aligned} \tag{15}$$

We set

$$D_{\mathcal{J}} \colon w \rightrightarrows \frac{1}{n}\mathrm{conv}\left(\sum_{i=1}^{n} D_i(w)\right)$$

and $\mathrm{crit}_J = \{w \in \mathbb{R}^p,\ 0 \in D_{\mathcal{J}}(w)\}$, the set of $D_{\mathcal{J}}$-critical points. Combining our results with the approach of [8], we obtain the following asymptotic characterization.

Theorem 9 (Convergence of mini-batch SGD) *Assume that $\alpha_k = o(1/\log(k))$. For any $M > 0$, conditioning on the event $\sup_{k\in\mathbb{N}} \|w_k\| \leq M$, setting, $\bar{w} \subset \mathbb{R}^p$, the set of accumulation points of $(w_k)_{k\in\mathbb{N}}$. We have, almost surely, $\emptyset \neq \bar{w} \subset \mathrm{crit}_{\mathcal{J}}$ and $\mathcal{J}$ is constant on $\bar{w}$.*

Proof : We condition on the event $\sup_{k\in\mathbb{N}} \|w_k\| \leq M$. $D_{\mathcal{J}}$ is a conservative field for $\mathcal{J}$. Hence $\mathcal{J}$ is a Lyapunov function for $\mathrm{crit}_{\mathcal{J}}$ and the differential inclusion

$$\dot{w} \in -D_{\mathcal{J}}(w),$$

which admits solutions according to [4, Chapter 2, Theorem 3]. We have by uniform randomness

$$\mathbb{E}_{B_k}\left[\frac{1}{|B_k|}\sum_{i\in B_k} D_i(w_k)\right] = \frac{1}{n}\sum_{i=1}^{n} D_i(w_k) \subset D_{\mathcal{J}}(w_k).$$

By conditioning, everything remains bounded so that there exists a constant $C(M)$ which only depends on M such that almost surely

$$\begin{aligned} &\sup_k \|d_k - v\| \leq C(M) \\ &\text{s.t.} \quad v \in \mathbb{E}_{B_k}\left[\frac{1}{|B_k|}\sum_{i\in B_k} D_i(w_k)\right] \end{aligned}$$

Theorem 2 implies that $\mathcal{J}(\mathrm{crit}_{\mathcal{J}})$ is finite, and hence has empty interior. The result follows by combining Theorem 3.6, Remark 1.5 and Proposition 3.27 of [8], see also [7, Proposition 4.4] for discussion on the step size. □

Remark 11 (Convergence) We conjecture that, beyond subsequential convergence, iterates should converge in the case of definable potentials.

6.2 Deep Neural Networks and nonsmooth backpropagation

We pertain to feed forward neural networks even though much more general cases are adapted to our auto-differentiation setting and to our definability assumptions.

Let us consider two finite dimensional real vector spaces spaces $\mathcal{X}, \mathcal{Y}$. The space $\mathcal{X}$ models input objects of interest (images, economical data, sentences, texts) while $\mathcal{Y}$ is an output space of properties of interest for the objects under consideration. The points y in $\mathcal{Y}$ are often called labels. The goal of deep learning is to label automatically objects in $\mathcal{X}$ by "learning the labelling principles" from a large dataset of known paired vectors $(x_i, y_i)_{i=1,\ldots,n}$. Given x in $\mathcal{X}$, we thus wish to discover its label y. This is done by designing a predictor function whose parameters are organized in L layers, each of which is represented by an affine function $A_j \colon \mathbb{R}^{p_j} \mapsto \mathbb{R}^{p_{j+1}}$ for values $p_j \in \mathbb{N}$, $j = 1, \ldots, L$. Our predictor function has then the form

$$\mathcal{X} \ni x \to \sigma_L(A_L(\sigma_{L-1}(A_{L-1}(\ldots\sigma_2(A_2(\sigma_1(A_1(x))))\ldots)))) \in \mathcal{Y} \tag{16}$$

where for each j, the functions $\sigma_j \colon \mathbb{R}^{p_j} \mapsto \mathbb{R}^{p_j}$, is locally Lipschitz continuous. These functions are called *activation function* and usually apply univariate functions coordinatewise. Very often one simply takes a single activation function $\sigma \colon \mathbb{R} \mapsto \mathbb{R}$ and apply it to coordinates of each layer. Classical choices for σ include:

1. identity: $t \mapsto t$,
2. sigmoid: $t \mapsto \frac{1}{1+e^{-t}}$,
3. hyperbolic tangent: $t \mapsto \tanh(t)$,
4. softplus: $t \mapsto \log(1 + \exp(t))$,
5. ReLU: $t \mapsto \max\{0, t\}$, aka positive part,
6. "Leaky-ReLU": $t \mapsto \max\{0, t\} + \alpha \min\{t, 0\}$, $\alpha > 0$, parameter.

7. piecewise polynomial activations.

Examples 1, 5, 6, 7 are semialgebraic, the others are definable in the same o-minimal structure ($\mathbb{R}$-exp definable sets). Among these examples, the ReLU activation function [27] played a crucial role in the development of deep learning architectures as it was found to be efficient in reducing "vanishing gradient" issues (those being related to the flatness of the commonly used sigmoid). This activation function is still widely used nowadays and constitutes one of the motivations for studying in more details automatic differentiation oracles applied to nonsmooth functions.

In order to lighten the notations, the weights of all the A_i in (16) are concatenated into a global weight vector w in $\mathbb{R}^p$, so we may simply write the parametrized predictor with parameter w,

$$g(w,x) := \sigma_L(A_L(\sigma_{L-1}(\dots\sigma_1(A_1(x)))))).$$

Learning a predictor function is finding an adequate collection of weights w. To do so one trains the neural networks by minimizing a loss of the form:

$$\mathcal{J}(w) = \frac{1}{n}\sum_{i=1}^{n} l(g(w,x_i),y_i) \tag{17}$$

where l is some elementary loss function, typical choices include the square loss $l(a,b) = \frac{1}{2}\|a-b\|^2$, $(a,b) \in \mathbb{R}^2$ for regression or binary cross entropy for classification: $l(a,b) = b\log(a) + (1-b)\log(1-a)$, where $a \in (0,1)$, $b \in \{0,1\}$. In view of matching the abstract model (14), set $f_i(x) = l(g(w,x_i),y_i)$ for all i. It is obvious to see that:

Lemma 7 (Deep Learning loss in algorithmic form) *Given $\sigma_1,\dots,\sigma_L$ and l, each term f_i of the deep learning loss $\mathcal{J}$ has a representation as in Algorithm 1.*

Let us now fix $\sigma_1,\dots,\sigma_L$ and l. Choose a conservative map[8] D_i for each σ_i, $i = 1\dots,L$, and D_l for l. An index i being fixed, the backpropagation algorithm applied to f_i is exactly backward auto-differentiation over f_i based on the data of $\{D_i\}_{i=1\dots L}$ and D_l. We denote by BP_{f_i} the output mapping. We have:

Corollary 6 (Backpropagation defines a conservative field) *With the above conventions, assume that l and $\sigma_1,\dots,\sigma_L$ as well as the corresponding conservative maps are definable in the same o-minimal structure, then the mapping BP_{f_i} is a conservative field. As a consequence*

$$BP_{f_i} = \nabla f_i$$

save on a finite union of manifolds of dimension at most $d-1$.

As a consequence, setting

$$BP_{\mathcal{J}} = \frac{1}{n}\sum_{i=1}^{n} BP_{f_i} \tag{18}$$

[8] If a unique $\sigma\colon \mathbb{R} \mapsto \mathbb{R}$ is applied to each coordinate of each layer, this amounts to consider a conservative field for σ, for example its Clarke subgradient.

we obtain a conservative field, and thus

$$BP_{\mathcal{J}} = \nabla \mathcal{J} \ \ a.\ e. \tag{19}$$

$$\mathcal{J}(w) - \mathcal{J}(v) = \int_0^1 \langle BP_{\mathcal{J}}((1-t)v + tw), w - v \rangle \,, \tag{20}$$

for any v, w in $\mathbb{R}^p$.

Remark 12 (Backpropagation and differentiability a.e.) (a) The backpropagation algorithm was popularized in the context of neural networks by [49] and is at the heart of virtually the totality of numerical algorithms for training deep learning architectures [37, 1, 44]. Most importantly, and this was the main motivation for our work, the backpropagation algorithm is used even for network built with non differentiable activation functions one of the most well known example being ReLU [27]. Using such nondifferentiable functions completely destroys the interpretation of backpropagation algorithm as computing a gradient or even a subgradient. Our results says that, although not computing any kind of known subdifferential, the nonsmooth backpropagation algorithm computes elements of a conservative field. As a consequence, it satisfies the operational chain rule given in Lemma 2. Note also that virtually all deep network architectures used in applications are actually definable, see e.g. [24].
(b) Despite our efforts we do not see any means to obtain Corollary 6 easily. In the "compositional course of loss differentiation" (recall Algorithm 1 and 3), one can indeed get trapped in "nondifferentiability zones" and thus speaking of the derivative of the active layers at this point has no meaning. Thus the smooth chain rule is of no use (see Remark 10) and the nonsmooth chain rules, for limiting or Clarke subdifferential are simply false in general, see for example [33].

To illustrate the fact that nonsmooth zones can be significantly activated during the training phase, we present now a numerical experiment. Let us consider a very simple feed forward architecture composed of L layers of fixed size p. Each layer is computed from the previous layer by application of a linear map, from $\mathbb{R}^p$ to $\mathbb{R}^p$, composed with coordinatewise application of ReLU. The input layer is the first element of the canonical basis and we sample the weights matrices with iid uniform entries in $[-1, 1]$. We repeat this sampling many times and estimate empirically the probability of computing ReLU(0) during forward propagation of the network (this would require to use the derivative of ReLU at 0 during backpropagation).

The results are depicted in Figure 1. It appears very clearly that for some architectures, with nonvanishing probability, we sample weight matrices resulting in the computation of ReLU(0). This means that, although the output of the network is piecewise polynomial as a function of weight matrices, and hence almost everywhere differentiable, *we still need to evaluate intermediate functions at points where they are not differentiable with non zero proability.* Hence, as we already mentioned, one cannot assert that the fact that the output is differentiable almost everywhere implies that the classical chain rule of differentiation applies almost everywhere. This assertion is just false.

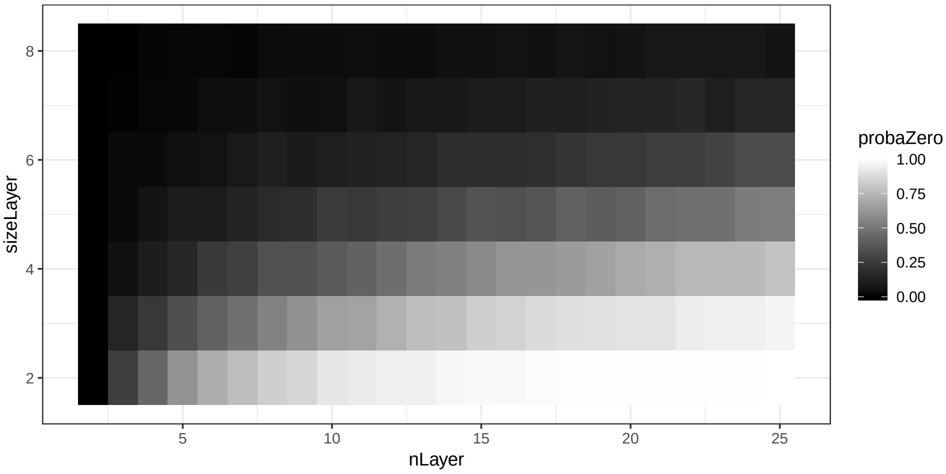

Figure 1: Estimation of the probability of applying ReLU to 0, as a function of the size and number of layers in a feedfoward network. The input is set to the first element of the canonical basis and we then propagate application ReLU layers with linear functions. The weights of the linear term are sampled uniformly at random between -1 and 1.

6.3 Training nonsmooth neural networks with nonsmooth SGD

To our knowledge the following result is the first genuine analysis of *nonsmooth* SGD for deep learning taking into account the real nature of backpropagation and the use of mini-batches. Note that the steady states below, $BP_{\mathcal{J}}$ critical points (see (18)), are the *actual steady states* of the corresponding dynamics. For simplicity of reading, we consider the special case of Relu networks with squared loss.

Corollary 7 (Convergence of SGD for Deep Learning) *Consider a feed forward neural network with mean squared error and* ReLU *activation function. Then the bounded sequences generated by the mini-batch SGD algorithm using the backpropagation oracle approach the* $BP_{\mathcal{J}}$ *critical set of the loss function with probability one.*

This is a direct consequence of Theorem 9 since the squared norm and ReLU are semialgebraic. The same result holds with any functions $\sigma_1, \ldots, \sigma_L$ and l definable in the same structure. As mentioned previously more complex architectures are accessible since our results rely only on abstract automatic differentiation and definability.

7 Conclusion

We introduced new tools for *nonsmooth* nonconvex problems, based on the idea that the choice of a fixed notion of subdifferential right from the start can be extremely limiting in terms of analysis and even of representation (e.g., automatic differentiation).

Our approach eventually consists in the following protocol. Consider an optimization problem involving an automatic differentiation oracle. We focused on the example of deep learning, but other application fields are possible (numerical simulations, optimal control solvers or partial differential equations [40, 21]).

- **"Choose your optimization method and then choose your subdifferential".** Evaluate precisely your decomposition requirements, in terms of sum or product, e.g., mini-batches for SGD. Infer from the decomposition method and the use of nonsmooth automatic differentiation a conservative field matched to the considered algorithm, e.g., coming back to SGD, set $D_f = \sum D_{f_i}$.
- **"Verify definability or tameness assumption".** Check that the various objects are definable in some common adequate structure. The problems we met are covered by one of the following, by order of frequency: semialgebraicity, global subanalyticity or log-exp structures.
- **"Identify Lyapunov/dissipative properties".** Use a Lyapunov approach, e.g. à la Benaïm-Hofbauer-Sorin, to conclude that the algorithm under consideration has dissipative properties and thus fine asymptotic properties.

To feel the generality of this protocol one can for instance consider mini-batch stochastic approximation strategies based on discretization of standard continuous time dynamical systems with known Lyapunov stability. Prominent examples include the heavy ball momentum method [3] commonly proposed in deep learning libraries, as well as INDIAN introduced and studied in [18].

Acknowledgements. The authors 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. J. Bolte acknowledges the support of ANR Chess, grant ANR-17-EURE-0010 and ANR OMS.

References

[1] Abadi M., Barham P., Chen J., Chen Z., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M., Kudlur M., Levenberg J., Monga R., Moore S., Murray D., Steiner B., Tucker P., Vasudevan V., Warden P., Wicke M., Yu Y. and Zheng X. (2016). Tensorflow: A system for large-scale machine learning. In Symposium on Operating Systems Design and Implementation.

[2] Aliprantis C.D., Border K.C. (2005) Infinite Dimensional Analysis (3rd edition) Springer

[3] Attouch H., Goudou X. and Redont P. (2000). The heavy ball with friction method, I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. Communications in Contemporary Mathematics, 2(01), 1-34.

[4] Aubin, J. P., Cellina, A. (1984). Differential inclusions: set-valued maps and viability theory (Vol. 264). Springer.

[5] Aubin, J.-P., and Frankowska, H. (2009). Set-valued analysis. Springer Science & Business Media.

[6] Baydin A., Pearlmutter B., Radul A. and Siskind J. (2018). Automatic differentiation in machine learning: a survey. Journal of machine learning research, 18(153).

[7] Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In Séminaire de probabilités XXXIII (pp. 1-68). Springer, Berlin, Heidelberg.

[8] Benaïm, M., Hofbauer, J., Sorin, S. (2005). Stochastic approximations and differential inclusions. SIAM Journal on Control and Optimization, 44(1), 328-348.

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

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

[11] Borkar, V. (2009). Stochastic approximation: a dynamical systems viewpoint (Vol. 48). Springer.

[12] Borwein J. and Lewis A. S. (2010). Convex analysis and nonlinear optimization: theory and examples. Springer Science & Business Media.

[13] Borwein J. M. and Moors W. B. (1997). Essentially smooth Lipschitz functions. Journal of functional analysis, 149(2), 305-351.

[14] Borwein J. M. and Moors, W. B. (1998). A chain rule for essentially smooth Lipschitz functions. SIAM Journal on Optimization, 8(2), 300-308.

[15] Borwein, J., Moors, W. and Wang, X. (2001). Generalized subdifferentials: a Baire categorical approach. Transactions of the American Mathematical Society, 353(10), 3875-3893.

[16] Bottou L. and Bousquet O. (2008). The tradeoffs of large scale learning. In Advances in neural information processing systems (pp. 161-168).

[17] Bottou L., Curtis F. E. and Nocedal J. (2018). Optimization methods for large-scale machine learning. Siam Review, 60(2), 223-311.

[18] Castera C., Bolte J., Févotte C., Pauwels E. (2019). An Inertial Newton Algorithm for Deep Learning. arXiv preprint arXiv:1905.12278.

[19] Clarke F. H. (1983). Optimization and nonsmooth analysis. Siam.

[20] Chizat, L., and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. In Advances in neural information processing systems, 3036-3046.

[21] Corliss G., Faure C., Griewank A., Hascoet L. and Naumann U. (Editors) (2002). Automatic differentiation of algorithms: from simulation to optimization. Springer Science & Business Media.

[22] Correa R. and Jofre, A. (1989). Tangentially continuous directional derivatives in nonsmooth analysis. Journal of optimization theory and applications, 61(1), 1-21.

[23] Coste M., *An introduction to o-minimal geometry.* RAAG notes, Institut de Recherche Mathématique de Rennes, 81 pages, November 1999.

[24] Davis, D., Drusvyatskiy, D., Kakade, S., Lee, J. D. (2018). Stochastic subgradient method converges on tame functions. Foundations of Computational Mathematics.

[25] van den Dries L. and Miller C. (1996). Geometric categories and o-minimal structures. Duke Math. J, 84(2), 497-540.

[26] Evans, L. C. and Gariepy, R. F. (2015). Measure theory and fine properties of functions. Revised Edition. Chapman and Hall/CRC.

[27] Glorot X., Bordes A. and Bengio Y. (2011). Deep sparse rectifier neural networks. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 315-323).

[28] Griewank, A., Walther, A. (2008). Evaluating derivatives: principles and techniques of algorithmic differentiation (Vol. 105). SIAM.

[29] Griewank A. (2013). On stable piecewise linearization and generalized algorithmic differentiation. Optimization Methods and Software, 28(6), 1139-1178.

[30] Griewank A., Walther A., Fiege S. and Bosse T. (2016). On Lipschitz optimization based on gray-box piecewise linearization. Mathematical Programming, 158(1-2), 383-415.

[31] Ioffe A. D. (1981). Nonsmooth analysis: differential calculus of nondifferentiable mappings. Transactions of the American Mathematical Society, 266(1), 1-56.

[32] Ioffe, A. D. (2017). Variational analysis of regular mappings. Springer Monographs in Mathematics. Springer, Cham.

[33] Kakade, S. M. and Lee, J. D. (2018). Provably correct automatic subdifferentiation for qualified programs. In Advances in Neural Information Processing Systems (pp. 7125-7135).

[34] Kurdyka, K. (1998). On gradients of functions definable in o-minimal structures. In Annales de l'institut Fourier 48(3), 769-783.

[35] Kurdyka, K., Mostowski, T. and Parusinski, A. (2000). Proof of the gradient conjecture of R. Thom. Annals of Mathematics, 152(3), 763-792.

[36] Kushner H. and Yin, G. G. (2003). Stochastic approximation and recursive algorithms and applications (Vol. 35). Springer Science & Business Media.

[37] LeCun Y., Bengio Y., Hinton, G. (2015). Deep learning. Nature, 521(7553).

[38] Ljung L. (1977). Analysis of recursive stochastic algorithms. IEEE transactions on automatic control, 22(4), 551-575.

[39] Majewski, S., Miasojedow, B. and Moulines, E. (2018). Analysis of nonsmooth stochastic approximation: the differential inclusion approach. arXiv preprint arXiv:1805.01916.

[40] Mohammadi, B. and Pironneau, O. (2010). Applied shape optimization for fluids. Oxford university press.

[41] Moulines E. and Bach, F. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In Advances in Neural Information Processing Systems (pp. 451-459).

[42] Moreau J.-J. (1963). Fonctionnelles sous-différentiables.

[43] Mordukhovich B. S. (2006). Variational analysis and generalized differentiation I: Basic theory. Springer Science & Business Media.

[44] Paszke A., Gross S., Chintala S., Chanan G., Yang E., DeVito Z., Lin Z., Desmaison A., Antiga L. and Lerer A. (2017). Automatic differentiation in pytorch. In NIPS workshops.

[45] Robbins H. and Monro, S. (1951). A stochastic approximation method. The annals of mathematical statistics, 400-407.

[46] Rockafellar R. T. (1963). Convex functions and dual extremum problems. Doctoral dissertation, Harvard University.

[47] Rockafellar R. (1970). On the maximal monotonicity of subdifferential mappings. Pacific Journal of Mathematics, 33(1), 209-216.

[48] Rockafellar, R. T., Wets, R. J. B. (1998). Variational analysis. Springer.

[49] Rumelhart E., Hinton E., Williams J. (1986). Learning representations by back-propagating errors. Nature 323:533-536.

[50] Speelpenning, B. (1980). Compiling fast partial derivatives of functions given by algorithms (No. COO-2383-0063; UILU-ENG-80-1702; UIUCDCS-R-80-1002). Illinois Univ., Urbana (USA). Dept. of Computer Science.

[51] Thibault, L. (1982). On generalized differentials and subdifferentials of Lipschitz vector-valued functions. Nonlinear Analysis: Theory, Methods & Applications, 6(10), 1037-1053.

[52] Thibault, L. and Zagrodny, D. (1995). Integration of subdifferentials of lower semicontinuous functions on Banach spaces. Journal of Mathematical Analysis and Applications, 189(1), 33-58.

[53] Thibault, L. and Zlateva, N. (2005). Integrability of subdifferentials of directionally Lipschitz functions. Proceedings of the American Mathematical Society, 2939-2948.

[54] Valadier, M. (1989). Entraînement unilatéral, lignes de descente, fonctions lipschitziennes non pathologiques. Comptes rendus de l'Académie des Sciences, 308, 241-244.

[55] Wang X. (1995). Pathological Lipschitz functions in $\mathbb{R}^n$. Master Thesis, Simon Fraser University.