Subgradient sampling for nonsmooth nonconvex minimization

Jérôme Bolte, Tam Le, Edouard Pauwels

Subgradient sampling for nonsmooth nonconvex minimization*

Jérôme Bolte† Tam Le† Edouard Pauwels ‡

February 18, 2022

Abstract

Risk minimization for nonsmooth nonconvex problems naturally leads to first-order sampling or, by an abuse of terminology, to stochastic subgradient descent. We establish the convergence of this method in the path-differentiable case, and describe more precise results under additional geometric assumptions. We recover and improve results from Ermoliev-Norkin [27] by using a different approach: conservative calculus and the ODE method. In the definable case, we show that first-order subgradient sampling avoids artificial critical point with probability one and applies moreover to a large range of risk minimization problems in deep learning, based on the backpropagation oracle. As byproducts of our approach, we obtain several results on integration of independent interest, such as an interchange result for conservative derivatives and integrals, or the definability of set-valued parameterized integrals.

Keywords. Subgradient sampling, stochastic gradient, online deep learning, conservative gradient, path-differentiability

AMS subject classifications. 68Q25, 68R10, 68U05

1 Introduction

We consider possibly nonconvex and nonsmooth risk minimization problems of the form

$$\min_{w\in\mathbb{R}^p} \mathcal{J}(w) := \mathbb{E}_{\xi\sim P}\left[f(w,\xi)\right], \tag{1}$$

where P is a probability distribution on some measurable space $(S,\mathcal{A})$. This type of problems has many applications, we refer for instance to [44, 45] and references therein, for various examples in several fields. Our specific interest goes in particular to online deep

*Submitted to the editors February 18, 2022.

†Toulouse School of Economics, Université de Toulouse, France.

‡IRIT, CNRS, Université de Toulouse, ANITI, Toulouse France.

learning [46] and machine learning more broadly [16]. We consider a minimization approach through first-order sampling: our model is that of the stochastic gradient method[1], which in its classical form generates iterates $(w_k)_{k\in\mathbb{N}}$ through

$$\begin{cases} w_0 \in \mathbb{R}^p \\ w_{k+1} = w_k - \alpha_k v(w_k, \xi_k) \qquad \text{for } k \in \mathbb{N}, \end{cases} \tag{2}$$

where $v(w_k, \xi_k)$ is a descent direction at w_k for the function $f(\cdot, \xi_k)$.

When f is smooth, v may be taken to be the gradient of f, so that, in expectation, the search direction boils down to the gradient:

$$\mathbb{E}_{\xi\sim P}\left[\nabla_w f(\cdot, \xi)\right] = \nabla \mathbb{E}_{\xi\sim P}\left[f(\cdot, \xi)\right] = \nabla \mathcal{J} \tag{3}$$

where the first equality follows under mild integrability conditions. This is a well known case for which many convergence results are available, see e.g., [31].

In the nonsmooth nonconvex world it has become classical to consider Clarke subgradient oracles. In that case the average update direction in (2) falls into $\mathbb{E}_{\xi\sim P}[\partial^c_w f(\cdot, \xi)]$; but contrary to (3), we merely have

$$\mathbb{E}_{\xi\sim P}\left[\partial^c_w f(\cdot, \xi)\right] \supset \partial^c \mathbb{E}_{\xi\sim P}\left[f(\cdot, \xi)\right] = \partial^c \mathcal{J}, \tag{4}$$

see [20, Theorem 2.7.3]. Without additional convexity or regularity assumptions, as in [32, 24], the inclusion is strict in general and unavoidable. In plain words, general subgradient sampling is not a stochastic subgradient method: expected increments are not necessarily subgradients of the loss. As a consequence, the very nature of a first-order sampling method may generate undesired directions that could result in absurd behaviors, such as sequences with occasionally erratic dynamics or artificial critical points which are irrelevant steady states [14]. As we shall see also, the situation is even worse in deep learning since in that case the oracle is based on backpropagation [43] which in a nonsmooth setting may add more artifacts to the dynamics, see e.g., [14] and references therein. Despite these issues, many real-world algorithms, are designed according to this model and it is necessary to provide theoretical support and guarantees to the practical success of these methods.

The goal of this paper is to study the first-order subgradient sampling method with vanishing step size, and precisely address these problems. Our answer is built upon conservative gradients and conservative calculus introduced in [13]. This approach makes rigorous the use of formal subdifferentiation in a wide mathematical framework encompassing most Lipschitz continuous nonconvex nonsmooth problems. One of the consequences is that sum, composition of conservative gradients are conservative, and under mild assumptions, we shall extend this property to parameterized integrals through the fact that expectation of conservative gradient is a conservative gradient. Another advantage in using this approach is its full compatibility with one of the core modern algorithms of large scale optimization: backpropagation [1, 39, 17].

An essential series of works on subgradient sampling are those developed in [27], followed also by [44, 45], using the generalized derivatives and the calculus introduced in Norkin

[1] Also known under the generic acronym *SGD*.

[35], see Section 6 for more details on this notion. These works address in particular the question of the interplay between expectations and subgradients under various assumptions, and provide as well convergence results to "generalized critical points". The approach relies on the notion of generalized derivatives in the sense of Norkin [35]. The ideas of [27], and some follow-up research [37, 38], have been surprisingly overlooked by the stochastic optimization and machine learning communities, at least until recently[2]. There is a strong common point between the present article and Norkin's research since, key to our work, is a new first-order calculus. But there are also important differences and additions that we highlight below:

— We follow the conservative approach of [13] instead of the generalized derivative approach of [35], and we thus work with different techniques. Our focus on the conservative case is motivated by a growing "theory" close to machine learning applications: implicit differentiation [12] with application to implicit neural networks [4], nondifferentiable programming [10], bi-level programming [41, 7]), differential equations [33] which are naturally connected to Neural ODEs [19, 26], or even partial minimization [40]. Let us also mention that conservativity is more general than differentiability in the sense of Norkin, see Section 6 — even though, both notions coincide in the semialgebraic case as recently established in [23].

— We provide a general and simple result on the avoidance of artificial critical points. First-order sampling and backpropagation create independently artificial critical points that can swamp the method. Under adequate subanalyticity assumptions, matching many practical problems, we establish that this does not occur, see Section 4.

— Our theory is developed in close connection to semialgebraicity and subanalyticity techniques. This allows for providing a "ready-to-use" flavor to our convergence results for online deep learning, and also to obtain new results in the definable world. We provide in particular conditions for set-valued integrands to result in definable parametrized integrals or expectation as well.

— Another specificity of our work is to rely for its asymptotic analysis on the "ODE approach" through the use of Benaim-Hofbauer-Sorin results [5] on stochastic approximations. The versatility and simplicity of this method allows for quite direct extension to other type of first-order methods as, see e.g., [9, 30, 18]. Definability allows to provide simple and explicit sufficient conditions for commonly used abstract hypotheses in this framework, such as path-differentiability of the risk and Sard's condition.

The paper is organized as follows: we first present representative samples of our general results in Section 2 with an emphasis on applications to online deep learning and the role of semialgebraic, or definable optimization. In Section 3, we present our main theoretical results with in particular a theorem of conservative differentiation for integral functions. Section 4 gathers results from o-minimal geometry and provides set-valued improvement

[2]As for us, it came to our knowledge in the finalization stage of this paper.

of Cluckers-Miller's integration result [21] as well as avoidance result for artificial critical points. Section 5 contains the proofs of the results presented in Section 2.

Notations For $q \in \mathbb{N}^*$, $\mathcal{B}(\mathbb{R}^q)$ denotes the Borel sigma algebra on $\mathbb{R}^q$. $\|\cdot\|$ denotes the Euclidean norm. For a subset A of a normed vector space, $\operatorname{conv} A$ denotes the convex hull of A and $\bar{A}$ its closure; $\dim A$ denotes its Hausdorff dimension. If in addition, A is bounded then $\|A\| := \sup\{\|y\| \mid y \in A\}$. For $x \in \mathbb{R}^p$ $r > 0$, $B(x, r)$ is the open ball of center x and radius r with respect to the Euclidean norm.

Let $f : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$ locally Lipschitz. f is differentiable almost everywhere, and we denote diff_f its differentiability domain. Its Clarke subgradient is defined for $x \in \mathbb{R}^p \times \mathbb{R}^m$ as

$$\partial^c f(x) = \operatorname{conv}\left\{ \lim_{k\to+\infty} \nabla f(x_k) \;\middle|\; x_k \in \operatorname{diff}_f, x_k \underset{k\to+\infty}{\to} x \right\}.$$

We denote $\partial^c_w f$ the projection of $\partial^c f(x)$ on w, for $(w, s) \in \mathbb{R}^p \times \mathbb{R}^m$,

$$\partial^c_w f(w, s) = \left\{A \in \mathbb{R}^p \mid (A, B) \in \partial^c f(w, s)\right\}.$$

For $F : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^q$ locally Lipschitz we can define as well the Clarke Jacobian and its projection on w:

$$\operatorname{Jac}^c F(x) = \operatorname{conv}\left\{ \lim_{k\to+\infty} \operatorname{Jac} F(x_k) \;\middle|\; x_k \in \operatorname{diff}_F, x_k \underset{k\to+\infty}{\to} x \right\},$$

$$\operatorname{Jac}^c_w F(w, s) = \left\{A \in \mathbb{R}^{q\times p} \mid [A\ B] \in \operatorname{Jac}^c F(w, s)\right\}.$$

2 Main results and online deep learning

This section provides simplified formulations of our results and also gives applications to online deep learning.

2.1 Subgradient sampling method

Framework Let us consider the stochastic minimization problem (1)

where P is a fixed probability distribution, f is possibly nonsmooth and nonconvex such that the risk function $\mathcal{J}$ is well defined. This problem is tackled through the *first-order sampling algorithm* (2). $(\xi_k)_{k\in\mathbb{N}}$ is a sequence of i.i.d. random variables with common distribution P and the mapping $v : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^p$ is a selection of $\partial^c_w f : \mathbb{R}^p \times \mathbb{R}^m \rightrightarrows \mathbb{R}^p$, i.e., for all $(w, s) \in \mathbb{R}^p \times \mathbb{R}^m, v(w, s) \in \partial^c_w f(w, s)$.

Assumption 1

1. *For all compact sets $K \subset \mathbb{R}^p$, $\exists \kappa : \mathbb{R}^m \to \mathbb{R}$ such that κ is square integrable with respect to P and*

$$\forall x, y \in K,\ |f(x, s) - f(y, s)| \le \kappa(s)\|x - y\|$$

for all $s \in \mathbb{R}^m$,

2. *For all* $k \in \mathbb{N}$, $\alpha_k > 0$, $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$ *and* $\lim_{k\to+\infty} \alpha_k = 0$,

3. f *and the selection* v *are semialgebraic,*

4. *There exists* $M > 0$ *such that* $\sup_{k\in\mathbb{N}} \|w_k\| \leq M$ *almost surely.*

Observe that the first condition implies that $\mathcal{J}$ is locally Lipschitz so that $\partial^c \mathcal{J}$ is well defined. We consider semialgebraicity for simplicity, many results of this work extends to definable functions f and selections v, which encompasses the vast majority of machine learning examples, using for example globally subanalytic sets (see Section 4.1 and [49]).

A glance at convergence results With this level of generality, we are not able to guarantee the almost sure convergence of $(w_k)_{k\in\mathbb{N}}$ to the set of critical points of $\mathcal{J}$, but we can describe its limit behaviour in a weaker sense introduced in [6] and extended to set-valued flows in [29]. This can be interpreted via the notion of *essential accumulation points*.

Definition 1 (Essential accumulation point) An accumulation point $\bar{w} \in \mathbb{R}^p$ is called *essential* if for every neighborhood $\mathcal{U}$ of $\bar{w}$ one has

$$\limsup_{k\to+\infty} \frac{\sum_{i=0}^{k} \alpha_i \mathbb{1}_{\{w_i \in U\}}}{\sum_{i=0}^{k} \alpha_i} > 0 \text{ almost surely.}$$

Intuitively, cluster points that are not essential accumulation points are hardly never seen. This is made more precise in [15, Lemma 21], through the use of occupation measures.

A first result on the criticality of the essential accumulation points of $(w_k)_{k\in\mathbb{N}}$ is as follows:

Theorem 1 (Criticality of essential accumulation points) *Let* $(w_k)_{k\in\mathbb{N}}$ *defined by (2). Then under Assumption 1, the following hold:*

(i) All essential accumulation points $\bar{w}$ *of* $(w_k)_{k\in\mathbb{N}}$ *satisfy the weak notion of criticality*

$$0 \in \mathbb{E}_{\xi\sim P}\left[\partial_w^c f(\bar{w}, \xi)\right] = \int_{\mathbb{R}^m} \partial_w^c f(\bar{w}, s)\, \mathrm{d}P(s) \tag{5}$$

where the integral is taken in the sense of Aumann (Definition 3).

(ii) If P *has a density with respect to Lebesgue, there exists a subset* $\Gamma \subset \mathbb{R}$ *whose complement is finite such that if* $\alpha_k \in \Gamma$ *for all* $k \in \mathbb{N}$*, then, for all initialization* w_0 *chosen in a residual full-measure, and with probability* 1*,* $(w_k)_{k\in\mathbb{N}}$ *verifies for all* $k \in \mathbb{N}$*,*

$$w_{k+1} = w_k - \alpha_k \left(\nabla \mathcal{J}(w_k) + \eta_k\right)$$

where $\eta_k := \nabla_w f(w_k, \xi_k) - \nabla \mathcal{J}(w_k)$ *is a martingale increment. Moreover, the essential accumulation points* $\bar{w}$ *of* $(w_k)_{k\in\mathbb{N}}$ *are Clarke critical, i.e.,*

$$0 \in \partial^c \mathcal{J}(\bar{w}). \tag{6}$$

Remark 1 (a)(Generalized criticality) Let us emphasize the fact that this criticality notion is unavoidable and inherent to the very nature of subgradient sampling methods. It is related to the notion of an artificial critical point as described in [14].
(b)(On proofs) In order to prove item (ii), we use stratification properties of f. The intermediary results leading to item (ii) are gathered in Section 4.

Remark 2 (Extension to the conservative gradient formalism) *Under the same assumptions, Theorem 1 (i) holds, mutatis mutandis, by replacing the Clarke subgradient by a conservative gradient: $\partial^c_w f$ can be replaced by a semi-algebraic set-valued map D where for all $s \in \mathbb{R}^m$, $D(\cdot, s)$ is a conservative gradient for $f(\cdot, s)$. In this case, Theorem 1 (ii) also holds true* as is. *It states that most sequences generated by conservative gradient sampling are actually stochastic gradient sequences, so that any result holding true for stochastic subgradient method (subgradient plus zero mean noise [24, 32]) also holds true for conservative gradient sampling, Clarke criticality of essential accumulation point being an example. This allows to generalize to the backpropagation implementation of the algorithm (2) as done in Section 2.3.*

With an additional assumption on $(\alpha_k)_{k\in\mathbb{N}}$ and the distribution P all accumulation points are Clarke critical and the risk function converges.

Assumption 2 P has a semialgebraic density with respect to Lebesgue.

Assumption 3 $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty$.

Theorem 2 (Criticality of accumulation points and convergence) *Under Assumption 1-2-3, Theorem 1 holds, and in addition, $\mathcal{J}(w_k)$ converges almost surely while* (5) *holds for all accumulation points. Furthermore, under the assumptions of Theorem 1, Item (ii) on generic step sizes and initialization points, Clarke criticality* (6) *actually holds for all accumulation points.*

The proof of Theorem 2 uses results in [5]. In Section 3.3, we summarize the essential theoretical results we use from [5] to prove the convergence of (2).

Remark 3 (On Assumption 2) *The proofs leading to Theorem 1 and theorem 2 can easily be adapted to the case when P is the joint distribution between a discrete and a continuous random variable. This allows to take into account classification tasks in online deep learning (see theorem 3). We limit ourselves to assumption 2 for the sake of simplicity. Semialgebraic densities are extremely flexible: they approximate all continuous densities on compact sets. Although assumption 2 relates to the unknown distribution P, this is a reasonable proxy for a large class of distributions. The main reason for this assumption is that it provides enough rigidity to ensure a strong form of Sard's theorem for the risk function $\mathcal{J}$. Beyond semialgebraicity, P can be assumed to be in certain classes of definable sets, for example globally subanalytic sets, see Section 4.1, which includes analytic densities with semialgebraic compact support, like truncated Gaussian distributions.*

On calculus: chain rule for expectations A pivotal calculus result in our analysis is the following:

Lemma 1 (Chain rule for expected risk functions) *Under Assumption 1, $\mathcal{J}$ admits a chain rule, i.e., for any absolutely continuous curve $\gamma : [0,1] \to \mathbb{R}^p$,*

$$\frac{\mathrm{d}}{\mathrm{d}t}(\mathcal{J} \circ \gamma)(t) = \langle a, \dot{\gamma}(t) \rangle \text{ for all } a \in \partial^c \mathcal{J}(\gamma(t)) \text{ for almost all } t \in [0,1].$$

Let us recall that, functions admitting a chain rule with respect to their Clarke subgradient, or equivalently to a conservative gradient, are called *path-differentiable*. Lemma 1 is a consequence of the more general Theorem 4 which generalizes the outer sum rule [13, Corollary 4] and extends [36, Theorem 1] to conservative gradients. This result allows to interchange conservative gradient and integral operations: taking the expectation of $v(\cdot, \xi)$ with respect to $\xi \sim P$ gives $\mathbb{E}_{\xi \sim P}[v(\cdot, \xi)]$ which is a selection in a conservative gradient for $\mathcal{J}$.

2.2 Comparison with related works

Some recent works [24, 32] assume the Clarke subgradient and integral operations to be interchangeable, which in practice requires regularity assumptions. Without such assumptions, as here, first-order sampling leads by nature to spurious critical points. This was first observed in [27] and rediscovered in [13, 14]. To avoid converging to these undesirable points [27, Remark 4.2] suggests to perturb iterates by a random noise. Instead, our Theorem 2 ensures that "almost all" subgradient sampling sequences accumulates to Clarke critical points. We thus justify the correctness of the algorithm as implemented in practice. The avoidance of spurious critical point is also obtained in [8] through probabilistic initializations having a density with respect to the Lebesgue measure. The definable framework allows for a much sharper description of the set of stepsizes and initializations leading to spurious points. For instance, when we consider sequences with a finite horizon, i.e., $w_0, \ldots, w_K$ with $K \geq 0$, the set of bad initializations is a finite union of manifolds of dimension strictly lower than p, while the set of bad stepsizes is finite.

Due to the role of the ODE method in the analysis of stochastic optimization methods, considering risk functions that have a chain rule is not new in the literature, either as an assumption [8] or using restrictive sufficient conditions [24, 32]. With Lemma 1, we provide instead simple and explicit sufficient conditions for deriving this chain rule (see also Theorem 4 for a general form). Similarly, our Assumption 2 is sufficient to obtain a strong form of Sard's condition. To the best of our knowledge, in former works on risk minimization, Sard's theorem is part of the hypotheses in a weak abstract form [27, 5, 8].

2.3 First-order sampling, backpropagation and deep learning

Deep learning model We consider a probability distribution P on $\mathbb{R}^d \times \mathbb{R}^I$ called *population* distribution. The goal of machine learning is to build a predicting function $h \colon \mathbb{R}^d \mapsto \mathbb{R}^I$ such that $h(X) \simeq Y$, in an average sense, when (X, Y) is distributed according to P, which we will write $(X, Y) \sim P$. In deep learning, this predictor is taken in a

specific class: neural networks. The input X can be for instance an image, and Y can either be discrete for a classification task or continuous for a regression task.

A feed-forward neural network is defined via a compositional structure involving L layers and a parameters vector $w = (w^{(1)}, \dots, w^{(L)})$ seen as a vector of $\mathbb{R}^p$. For $l = 0, \dots, L$, the l-th layer is represented by a real vector $x^{(l)} \in \mathbb{R}^{p_l}$. We consider that layer 0 has the same dimension as $\mathbb{R}^d$, $p_0 = d$. Given input $x \in \mathbb{R}^d$ the predicting function encoded by the neural network with parameter $w \in \mathbb{R}^p$ is denoted by $h(w, \cdot)$ and is defined by the relations:

$$\begin{cases} x^{(0)} & = x, \\ x^{(l)} & = g_l(w^{(l)}, x^{(l-1)}), \qquad \text{for } l = 1, \dots, L, \\ h(w, x) & = x^{(L)}. \end{cases} \tag{7}$$

For $l = 1, \dots, L$, the function g_l can take several forms. Such a function can be an affine transformation composed with a nonlinear function $\sigma^{(l)} : \mathbb{R}^{p_l} \to \mathbb{R}^{p_l}$, i.e., $g_l(w^{(l)}, x^{(l-1)}) = \sigma^{(l)}(A^{(l)}x^{(l-1)} + b^{(l)})$ with weight-bias parameters $w^{(l)} = (A^{(l)}, b^{(l)})$ where $A^{(l)} \in \mathbb{R}^{p_l \times p_{l-1}}$, $b^{(l)} \in \mathbb{R}^{p_l}$. For example, σ_l could be an activation function applied componentwise, well known examples include the ReLU function, the sigmoid, or the softmax function. More general block structured nonlinear functions can be considered, such as max pooling, or less common nonlinearities such as "sorting". We can impose a special structure on the matrices $A^{(l)}$ and vectors $b^{(l)}$, for example constraining some of their entries to take specific values. This allows to capture with model (7), convolutional layers, which are sparse linear maps, parallel architectures or residual neural networks. With a slight abuse of notation, we will consider that w is a vector in $\mathbb{R}^p$, consisting of the concatenation of all vectors $(w^{(l)})_{l=1}^L$.

Online deep learning Given a loss function $\ell \colon \mathbb{R}^I \times \mathbb{R}^I \to \mathbb{R}$, training is formulated as a risk minimization problem:

$$\min_{w \in \mathbb{R}^p} \ \mathcal{J}(w) := \mathbb{E}_{(X,Y)\sim P}\left[\ell(h(w, X), Y)\right] \tag{8}$$

In general the expectation is unknown and only approximated through statistical sampling. We consider the situation in which we have a sequence $(x_k, y_k)_{k\in\mathbb{N}}$ of i.i.d. samples generated from the distribution P and we tackle problem (8) with the first-order sampling algorithm (2). In this setting the sequence $(\xi_k)_{k\in\mathbb{N}}$ corresponds to $(x_k, y_k)_{k\in\mathbb{N}}$ and f is the function $(w, x, y) \mapsto \ell(h(w, x), y)$.

Backpropagation and conservative gradients In deep learning first-order information is accessed using backpropagation [43]. It is an efficient application of the chain rule of differentiable calculus which provides a numerical evaluation of the derivative of f. We will consider differentiation with respect to the decision variable w in (8) and denote the output of backpropagation by $\mathrm{backprop}_w$. The function f writes as a composition $f = f_r \circ \ldots \circ f_1$ where the functions $f_1, \dots, f_r$ involve the functions ℓ and $g_1, \dots, g_L$ from (7) and (8). When applied to nondifferentiable functions $f_1, \dots, f_r$, $\mathrm{backprop}_w$ can be

considered as an oracle evaluating an element in the product of their Clarke Jacobians:

$$\begin{aligned}\text{backprop}_w(f(w,x,y)) \in \partial^c f_r(f_{r-1}\circ\ldots\circ f_1(w,x,y))^T \\ \times \text{Jac}^c\, f_{r-1}(f_{r-2}\circ\ldots\circ f_1(w,x,y))\times\ldots\times \text{Jac}^c_w\, f_1(w,x,y).\end{aligned}$$

With this definition in mind, we may define the backpropagation variant of (2).

Algorithm 1 First-order sampling with backpropagation

1: **Inputs**:
$w_0 \in \mathbb{R}^p$, $(x_k, y_k)_{k\in\mathbb{N}}$ i.i.d. with distribution P, $(\alpha_k)_{k\in\mathbb{N}}$ positive step sizes.
2: **for** $k = 0, 1, \ldots$ **do**
3: $\quad w_{k+1} = w_k - \alpha_k \,\text{backprop}_{w_k}(\ell(h(w_k, x_k), y_k))$
4: **end for**

As highlighted in introduction, backpropagation does not necessarily compute an element of the Clarke subgradient. When $f_1, \ldots, f_r$ are path-differentiable, backpropagation computes a selection of a conservative gradient of f. It is satisfied for instance if $\ell, g_1, \ldots, g_L$ are locally Lipschitz and semialgebraic, or more generally, globally subanalytic. We hence assume the following which is a condition satisfied by the vast majority of deep learning applications:

Assumption 4 (Locally Lipschitz continuity and semialgebraicity) The neural network training problem in (8) satisfies for $l = 1, \ldots, L$, the functions $g_l \colon \mathbb{R}^{p_{l-1}} \to \mathbb{R}^{p_l}$ and $\ell \colon \mathbb{R}^I \times \mathbb{R}^I \to \mathbb{R}$ are locally Lipschitz and semialgebraic functions.

In order to formulate our results, we further require an assumption on the distribution P. This assumption is quite mild as it encompasses a large class of probability distributions in classification or regression.

Assumption 5 (Semialgebraic distribution) The joint distribution P satisfies one of the following:

- Regression: P has a semialgebraic density ϕ with respect to Lebesgue on $\mathbb{R}^d \times \mathbb{R}^I$.
- Classification: $\mathcal{Y} = (e_i)_{i=1}^I$ is finite, where for $i = 1 \ldots I$, e_i denotes the i-th element of the canonical basis in $\mathbb{R}^I$. In this case the distribution P factorizes as $P(X, Y) = P(Y)P(X|Y)$. We then assume that P_Y is discrete over $(e_i)_{i=1}^I$ and $P(X|Y = e_i)$ has a semialgebraic density ϕ_i with respect to Lebesgue on $\mathbb{R}^d$.

With this setting, a direct consequence of Theorem 2 is the following result:

Theorem 3 (First-order sampling and training for online deep learning) *Under Assumptions 4 and 5, let $(w_k)_{k\in\mathbb{N}}$ be generated by Algorithm 1. We suppose that the sequence $(\alpha_k)_{k\in\mathbb{N}}$ is strictly positive and verifies $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty$. Assume that there exists $M > 0$ such that $\sup_{k\in\mathbb{N}} \|w_k\| \leq M$ almost surely.*

Then there exists a set $\Gamma \subset \mathbb{R}$ whose complement is finite, and $W_0 \subset \mathbb{R}^p$ of full measure and residual such that if $\forall k \in \mathbb{N}, \alpha_k \in \Gamma$ and $w_0 \in W_0$, $\mathcal{J}(w_k)$ converges almost surely as $k \to \infty$ and all accumulation points $\bar{w}$ of $(w_k)_{k\in\mathbb{N}}$ are Clarke critical, i.e., verify $0 \in \partial^c \mathcal{J}(\bar{w})$.

This theorem parallels [13, Theorem 9] for a general population distribution in the training of deep learning models: assumptions are simple and widespread, it is fully compatible with backpropagation and it shows that artificial critical points are avoided. Similar models have been considered in the literature. Let us mention [45] which require many stronger assumptions (in particular ruling out nonsmooth activation functions). Another important contribution is given in [38], a discussion is given in the introduction.

3 Nonsmooth analysis for stochastic approximation algorithms

3.1 Set-valued analysis and conservative gradients

Let $(X, \mathcal{F})$ be a measurable space. Let us recall some results from set-valued analysis.

Definition 2 (Measurable set-valued maps) Denote $\mathcal{K}_p$ the set of compact subsets of $\mathbb{R}^p$. It is a measurable space considering the Borel σ-algebra $\mathcal{B}_H(\mathcal{K}_p)$ induced by the topology of the Hausdorff distance. A nonempty compact valued map $F : X \rightrightarrows \mathbb{R}^p$ is called *measurable* if it is measurable from $(X, \mathcal{F})$ to $(\mathcal{K}_p, \mathcal{B}_H(\mathcal{K}_p))$. In this case, for all closed subset $A \subset \mathbb{R}^p$ the upper inverse $F^u(A) := \{x \in X \mid F(x) \subset A\}$ is measurable in $(X, \mathcal{F})$.

Proposition 1 (Measurable selection theorem [2]) *Let $F : X \rightrightarrows \mathbb{R}^p$ be a measurable nonempty and compact valued map. Then there exists a measurable selection of F, that is a measurable function $v : X \to \mathbb{R}^p$ satisfying for all $x \in X$, $v(x) \in F(x)$.*

Corollary 1 (Castaing's Theorem) *Let $F : X \rightrightarrows \mathbb{R}^p$ nonempty compact valued. Then F is measurable if and only if there exists a sequence of measurable selections $(F_n)_{n\in\mathbb{N}}$ such that $\forall x \in X, F(x) = \overline{\{F_1(x), F_2(x), ...\}}$.*

Remark 4 (Measurability of set-valued composition) Corollary 1 can be used to justify measurability of composed set-valued functions. For instance, given $g : \mathbb{R}^p \to \mathbb{R}$ continuous and $F : X \rightrightarrows \mathbb{R}^p$ compact valued and measurable, then $g \circ F$ is measurable. Indeed, let $(F_k)_{k\in\mathbb{N}}$ be a sequence of measurable selections given by Castaing's Theorem, such that $\forall x \in X, F(s) = \overline{\{F_1(x), F_2(x), ...\}}$. Then by continuity of g, we have for all $x \in X$, $g(F(x)) = g(\overline{\{F_1(x), F_2(x), ...\}}) = \overline{\{g(F_1(x)), g(F_2(x)), ...\}}$. The functions $g \circ F_i$ are all measurable and $g \circ F$ is compact valued by continuity of g hence by Castaing's corollary, $g \circ F$ is measurable.

Definition 3 (Aumann integral) Let $(X, \mathcal{F}, \mu)$ be a measure space and $F : X \rightrightarrows \mathbb{R}^p$ a set-valued map. Then the *integral* of F with respect to the measure μ is

$$\int_X F(x)\,\mathrm{d}\mu(x) = \left\{ \int_X v(x)\,\mathrm{d}\mu(x) \;\middle|\; v \text{ is measurable and for all } x \in X, v(x) \in F(x) \right\}.$$

We also use the expectation notation $\mathbb{E}_{\xi\sim P}\left[F(\xi)\right] = \int_X F(x)\,\mathrm{d}P(x)$ whenever $(X, \mathcal{F}, P)$ is a probability space and ξ is a random variable valued in X with distribution P.

Definition 4 Let $F : \mathbb{R}^m \rightrightarrows \mathbb{R}^p$ be a set-valued map.

(i) (Graph closedness) F is *graph closed* or to have closed graph if its graph

$$\operatorname{Graph} F := \{(x, y) \in \mathbb{R}^m \times \mathbb{R}^p \mid y \in F(x)\}$$

is a closed subset of $\mathbb{R}^m \times \mathbb{R}^p$.

(ii) (Local boundedness) F is *locally bounded* if for all $x \in \mathbb{R}^m$, there exist a neighborhood $\mathcal{U}$ of x and $M > 0$, such that for all $z \in U$ and $y \in F(z)$, $\|y\| < M$.

(iii) (Upper semicontinuity) F is *upper semicontinuous* at $x \in \mathbb{R}^m$, if for each open subset $\mathcal{V}$ containing $F(x)$, there exists a neighborhood $\mathcal{U}$ of x such that for all $z \in \mathcal{U}$, $F(z) \subset \mathcal{V}$.

Definition 5 (Conservative gradient) Let $f : \mathbb{R}^p \to \mathbb{R}$ be locally Lipschitz continuous. A locally bounded and graph closed set-valued map $D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is called *a conservative gradient* for f if for all absolutely continuous curve $\gamma : [0, 1] \to \mathbb{R}^p$, f admits a chain rule with respect to D along γ, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(t) = \langle v, \dot{\gamma}(t) \rangle, \text{ for all } v \in D(\gamma(t)) \text{ and almost all } t \in [0, 1].$$

Lipschitz continuous functions $f : \mathbb{R}^p \to \mathbb{R}$ admitting a conservative gradient are called *path-differentiable*. They are central to our analysis.

3.2 Conservative gradient of integral functions

In this part $(S, \mathcal{A}, \mu)$ is a complete measure space. We consider a function $f : \mathbb{R}^p \times S \to \mathbb{R}$ such that for almost all $s \in S$, $f(\cdot, s)$ is path-differentiable with conservative gradient $D(\cdot, s)$. Our goal is to show a result of "conservative differentiation" under the integral sign in order to get a conservative calculus for the parametrized integral $\int_S f(\cdot, s)\, \mathrm{d}\mu(s)$.

First, we provide a result of derivation under the integral sign when the integrand is absolutely continuous in its first variable. We shall use the following lemma.

Lemma 2 (Measurability of partial derivatives) *Let U an open subset of $\mathbb{R}$ and $f : U \times S \to \mathbb{R}$ a $(\mathcal{B}(\mathbb{R}) \times \mathcal{A})$-measurable function. We suppose that there exists $M \subset S$ of full measure such that for all $s \in M$, $f(\cdot, s)$ is absolutely continuous. Then $\frac{\partial f}{\partial x}$ is jointly measurable and its domain of existence $E \subset U \times S$ is measurable of full measure. Also, for almost all $x \in U$ we have for almost all $s \in S$, that $\frac{\partial f}{\partial x}(x, s)$ exists.*

Proof. Define the following quantities for all $x \in U$ and $s \in M$:

$$f'_u(x, s) = \limsup_{h \to 0} \frac{f(x + h, s) - f(x, s)}{h} \text{ and } f'_l(x, s) = \liminf_{h \to 0} \frac{f(x + h, s) - f(x, s)}{h}.$$

By continuity of f both limit operators may operate only in $\mathbb{Q}$ without changing the value of f'_u and f'_l. Whence f'_l and f'_u are measurable and so is $\frac{\partial f}{\partial x}$. Furthermore, the domain E of $\frac{\partial f}{\partial x}$ is $\{(x,s) \in U \times S \mid f'_l(x,s) = f'_u(x,s),\ -\infty < f_u(x,s) < +\infty\}$ which is measurable. Applying Fubini's Theorem yields

$$\int_{U\times S} \mathbb{1}_{E^c}(x,s)\,\mathrm{d}(\lambda\times\mu)\,(x,s) = \int_S \int_U \mathbb{1}_{E^c}(x,s)\,\mathrm{d}x\,\mathrm{d}\mu(s) = \int_U \int_S \mathbb{1}_{E^c}(x,s)\,\mathrm{d}\mu(s)\,\mathrm{d}x.$$

Since $f(\cdot,s)$ is absolutely continuous for $s \in M$, it is differentiable a.e., thus $\forall s \in M, \int_U \mathbb{1}_{E^c}(x,s)\,\mathrm{d}x = 0$ and the second integral is zero. The third integral vanishes, so for almost all $x \in U$, $\int_S \mathbb{1}_{E^c}(x,s)\,\mathrm{d}\mu(s) = 0$, i.e. $,\frac{\partial f}{\partial x}(x,s)$ exists for almost all s, which concludes the proof. □

Proposition 2 (Differentiation of absolutely continuous integral functions) *Let U be an open subset of $\mathbb{R}$ and $f : U \times S \to \mathbb{R}$ such that:*

1. *For all $x \in U$, $f(x,\cdot)$ is integrable.*
2. *For almost all $s \in S$, $f(\cdot,s)$ is absolutely continuous.*
3. *$\frac{\partial f}{\partial x}$ is locally integrable jointly in x and s. For any compact interval $[a,b] \subset U$*

$$\int_S \int_a^b \left|\frac{\partial f}{\partial x}(x,s)\right| \mathrm{d}x\,\mathrm{d}\mu(s) < +\infty.$$

Then, the function $g : x \mapsto \int_S f(x,s)\,\mathrm{d}\mu(s)$, is absolutely continuous, differentiable on almost all $x \in U$ with $g'(x) = \int_S \frac{\partial f}{\partial x}(x,s)\,\mathrm{d}\mu(s)$.

Proof. Let $f : U \times S \to \mathbb{R}$ verifying all the assumptions. We consider the function $g : x \in U \mapsto \int_S f(x,s)\,\mathrm{d}\mu(s)$ and $a < b$ in U. From Lemma 2, $\frac{\partial f}{\partial x}(x,s)$ and exists a.e. in $(x,s) \in U \times S$ and admits a measurable extension. The a.e. defined function $\frac{\partial f}{\partial x}$ is identified with some measurable extension. Since for almost all $s \in S$ $f(\cdot,s)$ is absolutely continuous, the Fundamental theorem of calculus for Lebesgue integration (see Theorem 14 in Section 4, Chapter 5 of [42]) implies that

$$g(b) - g(a) = \int_S [f(b,s) - f(a,s)]\,\mathrm{d}\mu(s) = \int_S \int_a^b \frac{\partial f}{\partial s}(t,s)\,\mathrm{d}t\,\mathrm{d}\mu(s).$$

Under Assumption 3, by measure completeness, Fubini-Lebesgue's Theorem applies:

$$g(b) - g(a) = \int_a^b \int_S \frac{\partial f}{\partial s}(t,s)\,\mathrm{d}\mu(s)\,\mathrm{d}t. \tag{9}$$

The function $x \mapsto \int_S \frac{\partial f}{\partial x}(x,s)\,\mathrm{d}\mu(s)$ is integrable on $[a,b]$ so g is absolutely continuous. The Fundamental Theorem of Calculus states that for almost all $x \in U$, g' exists with $g'(x) = \int_S \frac{\partial f}{\partial x}(x,s)\,\mathrm{d}\mu(s)$. □

The following result is one of the cornerstones of this paper and allows to interchange conservative gradient and integral operations.

Theorem 4 (Path-differentiability of parametrized integrals) *Let $D : \mathbb{R}^p \times S \rightrightarrows \mathbb{R}^p$ and $f : \mathbb{R}^p \times S \to \mathbb{R}$ such that:*

1. *For all $x \in \mathbb{R}^p$, $f(x, \cdot)$ is integrable.*
2. *For almost all $s \in S$, $f(\cdot, s)$ is locally Lipschitz continuous and $D(\cdot, s)$ is conservative for $f(\cdot, s)$.*
3. *$D : \mathbb{R}^p \times S \rightrightarrows \mathbb{R}^p$ is jointly measurable in $\mathcal{B}(\mathbb{R}^p) \times \mathcal{A}$.*
4. *For all compact subset $C \subset \mathbb{R}^p$ There exists an integrable function $\kappa : S \to \mathbb{R}^+$ such that*
$$\forall (x, s) \in C \times S,\ \|D(x, s)\| \leq \kappa(s)$$
where for $(x, s) \in \mathbb{R}^p \times S$, $\|D(x, s)\| := \sup_{y \in D(x,s)} \|y\|$.

Then $\int_S f(\cdot, s)\, \mathrm{d}\mu(s)$ is path-differentiable and $\int_S D(\cdot, s)\, \mathrm{d}\mu(s)$ is a conservative gradient for $\int_S f(\cdot, s)\, \mathrm{d}\mu(s)$.

Proof. We shall use Proposition 2. Proceeding this way, we reduce the problem to a simpler one and then, we can use Proposition 2 to conclude. Let $f : \mathbb{R}^p \times S \to \mathbb{R}$ and $D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ verifying assumptions 1 to 4.

First, we verify $\int_S D(\cdot, s)\, \mathrm{d}\mu(s)$ is graph closed, nonempty valued and locally bounded. From the Kuratowski–Ryll-Nardzewski measurable selection theorem [2, Theorem 18.13], $\int_S D(\cdot, s)\, \mathrm{d}\mu(s)$ is nonempty valued. It is locally bounded by Assumption 4. For almost all $s \in S$, $D(\cdot, s)$ is graph closed and locally bounded, hence it is upper semicontinuous by [3, Corollary 1 in Chapter 1, Section 1]. By Aumann's integral properties, see [48, Theorem 2], and since $D(\cdot, s)$ is upper semicontinuous, compact valued for all s, we have that $\int_S D(\cdot, s)\, \mathrm{d}\mu(s)$ is graph closed.

Now, we have to verify the chain rule property. Let $\gamma : [0, 1] \to \mathbb{R}^p$ be any absolutely continuous curve. By hypothesis, there exists a set of full measure $M \subset S$ such that for all $s \in M$, $f(\cdot, s)$ has conservative gradient $D(\cdot, s)$. We have $\forall s \in M$, $f(\gamma(\cdot), s)$ is absolutely continuous because f is locally Lipschitz in $x \in C$ and γ is absolutely continuous. Thus, $\forall s \in M$ $f(\gamma(\cdot), s)$ is differentiable a.e. and the chain rule property (10) holds for almost all $t \in [0, 1]$, i.e.,
$$\forall v \in D(\gamma(t), s),\ \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t), s) = \langle v, \dot{\gamma}(t) \rangle. \tag{10}$$
Let $E \subset [0, 1] \times S$ be the domain of existence of $\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t), s)$. E is measurable and of full measure according to Lemma 2. We want to verify the measurability of the domain of validity of eq. (10) which is
$$E \cap \{(t, s) \in [0, 1] \times S \mid \varphi(t, s) = 0\}$$
where $\varphi(t, s) = \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t), s) - \langle D(\gamma(t), s), \dot{\gamma}(t) \rangle$ for all $(t, s) \in E$ and $\varphi(t, s) = 1$ elsewhere. By Castaing's Theorem (see Remark 4) φ is measurable. The set $\{(t, s) \in [0, 1] \times S \mid \varphi(t, s) = 0\}$ is exactly the upper inverse of $\{0\}$ by φ, $\varphi^u(\{0\})$ hence it is jointly measurable in

$(\mathbb{R} \times S, \mathcal{B}(\mathbb{R}) \otimes \mathcal{A})$. By the same arguments using Fubini's Theorem in the proof of Lemma 2, $\varphi^u(\{0\})$ is of full measure and there exists $I_1 \subset [0,1]$ of full measure such that for all $t \in I$ eq. (10) holds for almost all $s \in S$.

Let $t \in I_1$. From eq. (10) we can say that for any measurable selection $v : s \to \mathbb{R}^p$ of $D(\gamma(t), \cdot)$ we have for almost all $s \in S$

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t), s) = \langle v(s), \dot{\gamma}(t) \rangle. \tag{11}$$

Integrating (11) over $s \in S$ we have for any a in the Aumann integral $\int_S D(\gamma(t), s)\,\mathrm{d}s$ and measurable selection v such that $a = \int_S v(s)\,\mathrm{d}s$,

$$\int_S \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t), s)\,\mathrm{d}\mu(s) = \int_S \langle v(s), \dot{\gamma}(t) \rangle\,\mathrm{d}\mu(s) = \langle a, \dot{\gamma}(t) \rangle. \tag{12}$$

In the other hand, by continuity of γ, $\gamma([0,1])$ is compact so Cauchy-Schwarz inequality gives us for all $(t, s) \in [0,1] \times S$

$$|\langle v(s), \dot{\gamma}(t) \rangle| \leq \|D(\gamma(t), s)\| \|\dot{\gamma}(t)\| \leq \kappa(s) \|\dot{\gamma}(t)\|.$$

Since γ is absolutely continuous, $\dot{\gamma}$ is integrable on $[0,1]$. Thus, the function $(t, s) \mapsto \kappa(s)\|\dot{\gamma}(t)\|$ is locally integrable and so is $(t, s) \mapsto \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t), s)$. Using Proposition 2, there exists I_2 of full measure such that

$$\forall t \in I_2,\ \int_S \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t), s)\,\mathrm{d}\mu(s) = \frac{\mathrm{d}}{\mathrm{d}t} \int_S f(\gamma(t), s)\,\mathrm{d}\mu(s). \tag{13}$$

Combining eq. (12) which holds on I_1 and eq. (13) which holds on I_2 we have

$$\forall t \in I_1 \cap I_2,\ \forall a \in \int_S D(\gamma(t), s)\,\mathrm{d}\mu(s),\ \frac{\mathrm{d}}{\mathrm{d}t} \int_S f(\gamma(t), s)\,\mathrm{d}\mu(s) = \langle a, \dot{\gamma}(t) \rangle$$

and $I_1 \cap I_2$ is of full measure. Finally we have shown that $\int_S D(\cdot, s)\,\mathrm{d}\mu(s)$ is nonempty compact valued graph closed, and verifies the chain rule property, hence it is a conservative gradient for $\int_S f(\cdot, s)\,\mathrm{d}\mu(s)$. □

3.3 Application to stochastic approximation

Let P be a probability measure on $(S, \mathcal{A})$, and consider a set-valued map $D : \mathbb{R}^p \times S \to \mathbb{R}^p$ such that for almost all $s \in S$, $f(\cdot, s)$ is locally Lipschitz continuous and $D(\cdot, s)$ is a conservative gradient for $f(\cdot, s)$. We consider the sequence $(w_k)_{k \in \mathbb{N}}$ defined by (2) where v is now a selection of D, i.e., for all $(w, s) \in \mathbb{R}^p \times S$, $v(w, s) \in D(w, s)$.

In order to study the sequence $(w_k)_{k \in \mathbb{N}}$, we will use the nonsmooth ODE method developed in [5]. To this end, one considers the set-valued map

$$D_{\mathcal{J}} : w \mapsto \operatorname{conv}\left(\int_S D(w, s)\,\mathrm{d}P(s) \right)$$

and the differential inclusion

$$\dot{w} \in -D_{\mathcal{J}}(w). \tag{14}$$

A *solution* to the differential inclusion (14) with initial point $w_0 \in \mathbb{R}^p$ is an absolutely continuous curve $w : \mathbb{R}_+ \to \mathbb{R}^p$ such that $w(0) = w_0$ and for almost all $t \in \mathbb{R}_+$, $\dot{w}(t) \in -D_{\mathcal{J}}(w(t))$. It is known that for any initial condition, (14) has at least a solution whenever $D_{\mathcal{J}}$ satisfies the following conditions (see Chapter 2, Theorem 3 in [3]):

- $D_{\mathcal{J}}$ is graph closed,
- $D_{\mathcal{J}}$ is nonempty and convex valued,
- $D_{\mathcal{J}}$ is locally bounded

Define Φ_t be the *set-valued flow* at $t \in \mathbb{R}_+$ defined for $w_0 \in \mathbb{R}^p$ as

$$\Phi_t(w_0) := \{w(t) \mid w : \mathbb{R}_+ \to \mathbb{R}^p \text{ is a solution of (14) with } w(0) = w_0\}.$$

Definition 6 (Lyapunov function for a set) A function $\mathcal{F}$ is a *Lyapunov function for a set* $S \subset \mathbb{R}^p$ and for the dynamical system (14) if

$$\begin{aligned} &\forall x \in \mathbb{R}^p \setminus S\ , \forall t > 0, \forall y \in \Phi_t(x),\ \mathcal{F}(y) < \mathcal{F}(x) \\ &\forall x \in S,\ \forall t \geq 0, \forall y \in \Phi_t(x),\ \mathcal{F}(y) \leq \mathcal{F}(x). \end{aligned}$$

Lemma 3 (Conservative gradient and Lyapunov function) *Let $D_{\mathcal{J}}$ a conservative gradient for $\mathcal{J}$. $\mathcal{J}$ is a Lyapunov function for* $\operatorname{crit} D_{\mathcal{J}}$ *and the differential inclusion (14)*

Proof. Let $x \in \mathbb{R}^p$, $t \geq 0$ and $y \in \Phi_t(x)$. By definition of Φ_t, there exists $w : \mathbb{R}_+ \to \mathbb{R}^p$ a solution to the differential inclusion $\dot{w} \in -D_{\mathcal{J}}(w)$ with initial value $w(0) = x \in \mathbb{R}^p$ such that $y = w(t)$. By definition of a conservative gradient and since w is absolutely continuous we have:

$$\mathcal{J}(w(t)) - \mathcal{J}(w(0)) = \int_0^t \langle D_{\mathcal{J}}(w(u)), \dot{w}(u)\rangle \,\mathrm{d}u. \tag{15}$$

Since w is a solution of (1), $\dot{w}(u) \in -D_{\mathcal{J}}(z(u))$ for almost all $u \in [0, t]$ and we have $\mathcal{J}(w(t)) - \mathcal{J}(w(0)) = -\int_0^t \|\dot{w}(u)\|^2 \,\mathrm{d}u$, hence $\mathcal{J}(w(t)) = \mathcal{J}(y) \leq \mathcal{J}(x)$.

Now we suppose that $x \in \mathbb{R}^p \setminus \operatorname{crit} D_{\mathcal{J}}$ and $t > 0$. By upper semicontinuity, $\exists \epsilon > 0, \exists \delta > 0, \forall y \in \mathbb{R}^p$ such that $\|y - x\| \leq \delta$, we have $\forall v \in D_{\mathcal{J}}(y), \|v\| \geq \epsilon$. By continuity of w, $\exists t_0 > 0, \forall u \in [0, t_0]\,, \|w(u) - x\| \leq \delta$ hence $\|\dot{w}(u)\| \geq \epsilon$ for almost all $u \in [0, t_0]$. Thus, by integration, $\int_0^t \|\dot{w}(u)\|^2 \,\mathrm{d}u$ is strictly positive and $\mathcal{J}(y) < \mathcal{J}(x)$. □

Assumption 6 *The following hold for D and f:*

1. *For all $w \in \mathbb{R}^p$, $f(w, \cdot)$ is integrable.*
2. *$D : \mathbb{R}^p \times S \rightrightarrows \mathbb{R}^p$ is jointly measurable in $\mathbb{R}^p \times S$.*

3. *For all compact subset $K \subset S$, there exists a measurable function $\kappa : S \to \mathbb{R}^+$ such that κ is square integrable with respect to P and:*

$$\forall (w, s) \in K \times S, \|D(w, s)\| \leq \kappa(s)$$

where for $(w, s) \in \mathbb{R}^p \times S$, $\|D(w, s)\| := \sup_{y \in D(w,s)} \|y\|$.

Remark 5 Assumption 6.3 is sufficient to apply Theorem 4 because if κ is square integrable with respect to the probability measure P, it is also P-integrable.

We have shown in Section 3 that $D_{\mathcal{J}}$ is a conservative gradient for $\mathcal{J}$. Following the work of Benaim-Hofbauer-Sorin [5], let us make a Morse-Sard assumption:

Assumption 7 (Morse-Sard) $\mathcal{J}(\operatorname{crit} D_{\mathcal{J}})$ has empty interior.

We now arrive to one of the central theorems of this work

Theorem 5 (Convergence of the subgradient sampling method) *Let $(w_k)_{k\in\mathbb{N}}$ given by (2). Suppose Assumption 6, Assumption 7. Assume furthermore that $(\alpha_k)_{k\in\mathbb{N}}$ is strictly positive, $\sum_{k\in\mathbb{N}} \alpha_k = +\infty$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty$ and $\sup_{k\in\mathbb{N}} \|w_k\| = M < +\infty$ almost surely. Then almost surely, $\mathcal{J}(w_k)$ converges as $k \to +\infty$ and all accumulation point $\bar{w}$ of $(w_k)_{k\in\mathbb{N}}$ satisfies $0 \in D_{\mathcal{J}}(\bar{w})$.*

Proof.

We set for all $k \in \mathbb{N}$ the quantities

$$a_k = \int_S v(w_k, s)\, \mathrm{d}P(s) \qquad \text{and} \qquad \epsilon_k = v(w_k, \xi_k) - a_k.$$

With these notations, (2) writes

$$w_{k+1} = w_k - \alpha_k(a_k + \epsilon_k) \text{ for all } k \in \mathbb{N},$$

where we have $a_k \in \int_S D(w_k, s)\, \mathrm{d}P(s) \subset D_{\mathcal{J}}(w_k)$. By independence of the ξ_k, $k \in \mathbb{N}$, we have for all $k \in \mathbb{N}^*$, $\mathbb{E}\left[\epsilon_k \mid \epsilon_0, \ldots, \epsilon_{k-1}\right] = \int_S v(w_k, s)\, \mathrm{d}P(s) - a_k = 0$. We have to verify the sequence $(\epsilon_k)_{k\in\mathbb{N}}$ is bounded in second order moment. By assumption $\sup_{k\in\mathbb{N}} \|w_k\| = M < +\infty$ almost surely, whence by Assumption 6, we have almost surely $\|D(w_k, \xi)\| \leq \kappa(\xi)$ where $\xi \sim P$. It follows that

$$\begin{aligned}
\mathbb{E}\left[\|\epsilon_k\|^2 \mid \epsilon_0, \ldots, \epsilon_{k-1}\right] = \mathbb{E}_{\xi\sim P}\left[\|v(w_k, \xi) - a_k\|^2\right] &\leq \mathbb{E}_{\xi\sim P}\left[(\|v(w_k, \xi)\| + \|a_k\|)^2\right] \\
&\leq 4\mathbb{E}_{\xi\sim P}\left[\|D(w_k, \xi)\|^2\right] \\
&\leq 4\mathbb{E}_{\xi\sim P}\left[\kappa(\xi)^2\right] \\
&< +\infty \text{ almost surely,}
\end{aligned}$$

hence $\sup_{k\in\mathbb{N}} \mathbb{E}\left[\|\epsilon_k\|^2\right] < +\infty$ and by [5, Remark 1.5 (i)] $(w_k)_{k\in\mathbb{N}}$ satisfies [5, Definition (II)], i.e., is a perturbed solution to (14) with probability 1. We can now combine several

results from [5] to deduce our convergence result. By [5, Theorem 4.2] $(w_k)_{k\in\mathbb{N}}$ satisfies [5, Theorem 4.1 (ii)] for (14) which implies by [5, Theorem 4.3] since $\sup_{k\in\mathbb{N}} \|w_k\| = M < +\infty$, that the set of accumulation points of $(w_k)_{k_i n\mathbb{N}}$ is internally chain transitive with probability 1. Then by Lemma 3 $\mathcal{J}$ is a Lyapunov function for crit $D_{\mathcal{J}}$ and the differential inclusion (14). By Assumption 7, $\mathcal{J}(\text{crit}\, D_{\mathcal{J}})$ has an empty interior. All the conditions are satisfied to apply [5, Proposition 3.27] hence almost surely the set of accumulation points of $(w_k)_{k\in\mathbb{N}}$ is contained in crit $D_{\mathcal{J}}$ and $\mathcal{J}(w_k)$ converges. □

4 On the geometry of stochastic optimization problems

4.1 Definable sets

Definition 7 (o-minimal structure) Let $\mathcal{O} = (\mathcal{O}_p)_{p\in\mathbb{N}}$ be a collection of sets such that, for all $p \in \mathbb{N}$, $\mathcal{O}_p$ is a set of subsets of $\mathbb{R}^p$. $\mathcal{O}$ is an o-minimal structure on $(\mathbb{R}, +, \cdot)$ if it satisfies the following axioms:

1. For all $p \in \mathbb{N}$, $\mathcal{O}_p$ is stable by finite intersection and union, complementation, and contains $\mathbb{R}^p$.
2. If $A \in \mathcal{O}_p$ then $A \times \mathbb{R}$ and $\mathbb{R} \times A$ belong to $\mathcal{O}_{p+1}$.
3. Denoting by π the projection on the p first coordinates, if $A \in \mathcal{O}_{p+1}$ then $\pi(A) \in \mathcal{O}_p$.
4. For all $p \in \mathbb{N}$, $\mathcal{O}_p$ contains the algebraic subsets of $\mathbb{R}^p$, i.e., sets of the form $\{x \in \mathbb{R}^p : P(x) = 0\}$, where $P : \mathbb{R}^p \to \mathbb{R}$ is a polynomial function.
5. The elements of $\mathcal{O}_1$ are exactly the finite unions of intervals.

The definition allows in particular to prove definability through the use of first-order formula (see [22] for more details). We also recall a few useful results below.

Proposition 3 (Definable choice [22]) *Let $A \subset \mathbb{R}^p \times \mathbb{R}^q$ a definable set. Denote $\mathcal{P}_p$ the projection on the p first coordinates. Then there exists a definable function $h : \mathcal{P}_p A \to \mathbb{R}^q$ such that for all $x \in \mathcal{P}_p A$, $(x, h(x)) \in A$.*

Definition 8 (Semianalytic sets and functions [49]) (i) (Semianalyticity) A subset A of $\mathbb{R}^n$ is semianalytic if for any point $x \in \mathbb{R}^d$, there exists a neighborhood U of x such that $U \cap A$ has the form

$$\bigcup_{i=1}^{l} \bigcap_{j=1}^{k} \{x \in \mathbb{R}^n \mid g_{ij}(x) < 0,\ h_{ij}(x) = 0\} \text{ where the } g_{ij} \text{ and } h_{ij} \text{ are real analytic.}$$

(ii) (Subanalyticity) A subset A of $\mathbb{R}^n$ is called a *subanalytic set* if there exists $m \in \mathbb{N}$ such that A is the projection of a semianalytic set $M \subset \mathbb{R}^{n+m}$ on the n first coordinates. A function $f : \mathbb{R}^l \to \mathbb{R}^k$ *is subanalytic* if its graph is a subanalytic set of $\mathbb{R}^l \times \mathbb{R}^k$.

A useful notion to get an o-minimal structure containing "compactly defined" subanalytic functions is global subanalyticity. Let

$$\tau_n : x \mapsto \left(\frac{x_1}{\sqrt{1+x_1^2}}, \dots, \frac{x_n}{\sqrt{1+x_n^2}} \right)$$

and define the *globally subanalytic* sets of $\mathbb{R}^n$ as the subsets of $\mathbb{R}^n$ whose images by τ_n are subanalytic.

The collection of globally subanalytic sets on the spaces $\mathbb{R}^n$, $n \geq 1$ forms an o-minimal structure denoted $\mathbb{R}_{\text{an}}$. Also, there exists an o-minimal structure denoted $\mathbb{R}_{\text{an,exp}}$ containing $\mathbb{R}_{\text{an}}$ and the graph of the exponential function (see [25]).

4.2 Definability and set-valued integration

The following result is a set-valued version of [21, Theorem 1.3].

Theorem 6 (Definable set-valued integrals) *Let ϕ a globally subanalytic density function on $\mathbb{R}^m$, $D : \mathbb{R}^p \times \mathbb{R}^m \rightrightarrows \mathbb{R}^p$ globally subanalytic, graph closed, locally bounded and convex valued. Then the set-valued map $D_{\mathcal{J}} : w \mapsto \int_{\mathbb{R}^m} D(w,s)\phi(s)\,\mathrm{d}s$ is definable in $\mathbb{R}_{\text{an,exp}}$.*

Proof. In order to prove the definability of $D_{\mathcal{J}}$ using Theorem 1.3 in [21], we use a property of the support function. For a compact convex subset $C \subset \mathbb{R}^p$ define the support function as $h_C : q \mapsto \max_{v \in C} \langle v, q \rangle$. Then by duality, we have

$$C = \left\{ z \in \mathbb{R}^p \;\middle|\; \sup_{v \in \mathbb{R}^p} \langle v, z \rangle - h_C(v) = 0 \right\}.$$

By assumption D is definable in $\mathbb{R}_{\text{an}}$, so the set-valued map $(w,q,s) \mapsto \langle D(w,s), q \rangle$ is definable. Denote G its graph. The function

$$\begin{aligned} H \colon \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^m &\to \mathbb{R} \\ (w,q,s) &\mapsto \max_{v \in D(w,s)} \langle v, q \rangle \end{aligned}$$

is definable in $\mathbb{R}_{\text{an}}$ because its graph can be written as a first-order formula:

$$\operatorname{Graph} H = \{(w,q,s,y) \in G \mid \forall (w',q',s',y') \in G, (w',q',s') = (w,q,s) \implies y \geq y'\}.$$

By definition of the Aumann integral, we have for all $w \in \mathbb{R}^p$

$$D_{\mathcal{J}}(w) = \left\{ \int_{\mathbb{R}^m} g(s)\phi(s)\,\mathrm{d}s \;\middle|\; g \text{ is a measurable selection of } D(w,\cdot) \right\}.$$

Then by linearity, for $(w,q) \in \mathbb{R}^p \times \mathbb{R}^p$ we have

$$\begin{aligned} h_{D_{\mathcal{J}}(w)}(q) &= \max_{v \in D_{\mathcal{J}}(w)} \langle v, q \rangle \\ &= \max \left\{ \int_{\mathbb{R}^m} \langle g(s), q \rangle \phi(s)\,\mathrm{d}s \;\middle|\; g \text{ is a measurable selection of } D(w,\cdot) \right\}. \end{aligned}$$

Using Theorem 18.19. in [2], there exists a measurable selection $\tilde{g} : \mathbb{R}^m \mapsto \mathbb{R}^p$ of $D(w,\cdot)$ such that $\forall s \in \mathbb{R}^m, \langle \tilde{g}(s), q\rangle = \max_{v \in D(w,s)} \langle v, q\rangle = H(w,q,s)$, and thus, $\tilde{g}$ achieves the maximum $h_{D_{\mathcal{J}}(w)}(q)$, i.e., $h_{D_{\mathcal{J}}(w)}(q) = \int_{\mathbb{R}^m} H(w,q,s)\phi(s)\,\mathrm{d}s$.

Using the property of the support function, we can finally write the graph of $D_{\mathcal{J}}$:

$$\begin{aligned}\operatorname{Graph} D_{\mathcal{J}} &= \{(w,z) \in \mathbb{R}^p \times \mathbb{R}^p \mid z \in D_{\mathcal{J}}(w)\} \\ &= \left\{(w,z) \in \mathbb{R}^p \times \mathbb{R}^p \;\middle|\; \sup_{q \in \mathbb{R}^p} \langle q, z\rangle - h_{D_{\mathcal{J}}(w)}(q) = 0\right\} \\ &= \left\{(w,z) \in \mathbb{R}^p \times \mathbb{R}^p \;\middle|\; \sup_{q \in \mathbb{R}^p} \langle q, z\rangle - \int_{\mathbb{R}^m} H(w,q,s)\phi(s)\,\mathrm{d}s = 0\right\}.\end{aligned}$$

Using [21, Theorem 1.3], the function $(w,q) \in \mathbb{R}^p \times \mathbb{R}^p \mapsto \int_S H(w,q,s)\phi(s)\,\mathrm{d}s$ is definable in $\mathbb{R}_{\text{an,exp}}$, hence $D_{\mathcal{J}}$ is definable in $\mathbb{R}_{\text{an,exp}}$ (see [25]). □

4.3 Consequences in stochastic optimization

Preliminary results Before providing our stochastic results, let us establish some technical lemmas:

Lemma 4 *Let $(r,q) \in \mathbb{N}^* \times \mathbb{N}^*$, and $L \subset \mathbb{R}^r \times \mathbb{R}^q$ a dense definable set. Then for all $w \in \mathbb{R}^r$ setting $L_w := \{s \in \mathbb{R}^q \mid (w,s) \in L\}$, there is a dense definable set $W \subset \mathbb{R}^r$ such that for all $w \in W$, L_w is dense in $\mathbb{R}^q$.*

Proof. Set $W = \{w \in \mathbb{R}^r \mid \forall z \in \mathbb{R}^q, \forall \epsilon > 0, \exists s \in \mathbb{R}^q, (w,s) \in L, \|s - z\| < \epsilon\}$. This set is definable and is precisely the set of w such that L_w is dense in $\mathbb{R}^q$. Assume that W^c has non empty interior. This means that there is an open set $U \subset \mathbb{R}^r$, such that for all $w \in U$

$$\exists z \in \mathbb{R}^q, \exists \epsilon > 0, \forall s \in \mathbb{R}^q, (w,s) \in L, \|s - z\| \geq \epsilon.$$

By definable choice, there are definable functions $z\colon U \to \mathbb{R}^q$ and $\epsilon\colon U \to \mathbb{R}_+^*$, such that for all $w \in U$, we have

$$\{(w,v) \in U \times \mathbb{R}^q \mid \|v - z(w)\| < \epsilon(w)\} \subset L^c.$$

By stratification, reducing U if necessary, z and ϵ can be chosen Lipschitz continuous and thus L^c has nonempty interior which contradicts the density of L. □

Claim 1 *Let $g\colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^p$ be a definable function. Then there exists a definable dense open sets $L \subset \mathbb{R}^p \times \mathbb{R}^m$ a subset $\Gamma \subset \mathbb{R}$ which complement is finite as well as a definable dense set $\Delta \subset \Gamma \times \mathbb{R}^m$, such that g is C^2 on L, for every $\alpha \in \Gamma$, the definable set $\{s \in \mathbb{R}^m \mid (\alpha,s) \in \Delta\}$ is dense open in $\mathbb{R}^m$ and for all $(\alpha,s) \in \Delta$, denoting by $\Phi_{\alpha,s} = \mathrm{Id} - \alpha\nabla_w g(\cdot,s)$ from $L_s := \{w \in \mathbb{R}^p \mid (w,s) \in L\}$, dense open, to $\mathbb{R}^p$, we have*

$$\forall Z \subset \mathbb{R}^p \text{ definable}, \ \dim Z \leq p-1 \implies \dim \Phi_{\alpha,s}^{-1}(Z) \leq p-1.$$

Proof. Denote by L a definable dense open set such that g is C^2 on L (such sets exist by stratification). Let $\lambda : L \subset \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^p$ be a definable representation of the eigenvalues of $\nabla^2_w g$, where $\nabla^2_w g$ denotes the partial Hessian of g with respect to the variable w. Refine L so that λ is jointly differentiable in (w, s). L is open and dense by definability of g. We further set $S_0 \subset \mathbb{R}^m$ the definable dense set obtained from Lemma 4 such that for all $s \in S_0$ the set $L_s := \{w \in \mathbb{R}^p \mid (w, s) \in L\}$ is open dense in $\mathbb{R}^p$.

Set F to be the complement of the critical values of the function λ_i, for $i = 1, \ldots, p$ on L. The critical values set F^c is finite by the definable Sard's theorem [11]. Set $\Gamma := \{\alpha \in \mathbb{R} \mid \alpha \neq 0, \alpha^{-1} \in F\}$. For $i = 1, \ldots, p$ set

$$E_i := \{(\alpha, s) \in \Gamma \times S_0 \mid \exists w \in L, \alpha\lambda_i(w, s) = 1, \nabla_w \lambda_i(w, s) = 0\}.$$

This set is definable because it is defined by a first-order formula involving definable functions and L, F, S_0 which are definable sets. Let us fix an arbitrary $\alpha \in \Gamma$, and show that the set $E_{\alpha,i} := \{s \in \mathbb{R}^m \mid (\alpha, s) \in E_i\}$ has empty interior. By [22, Theorem 3.1] there exists a definable function $\tilde{w} : E_{\alpha,i} \to \mathbb{R}^p$ such that $\forall s \in E_{\alpha,i}$, $\alpha\lambda_i(\tilde{w}(s), s) = 1$ and $\nabla_w \lambda_i(\tilde{w}(s), s) = 0$. Suppose by contradiction that there exists a nonempty open subset $U \subset E_{\alpha,i}$. By definability of $\tilde{w}$ and a stratification argument, U can be chosen so that $\tilde{w}$ is continuously differentiable on U. Then denoting $\tilde{\lambda}_i : s \mapsto \lambda_i(\tilde{w}(s), s)$ we have for all $s \in U$, $\nabla\tilde{\lambda}_i(s) = 0$. The chain rule applied on $\tilde{\lambda}_i$ yields

$$\forall s \in U,\ \nabla\tilde{\lambda}_i(s) = \operatorname{Jac}\tilde{w}(s)^{\mathsf{T}}\nabla_w\lambda_i(\tilde{w}(s), s) + \nabla_s\lambda_i(\tilde{w}(s), s) = \nabla_s\lambda_i(\tilde{w}(s), s) = 0.$$

Hence we have or all $s \in U$, $\nabla\lambda_i(\tilde{w}(s), s) = 0$. In other words since $\lambda_i(\tilde{w}(s), s) = \alpha^{-1}$, for all $s \in U$ then α^{-1} is a critical value of λ_i which contradicts the fact that $\alpha \in \Gamma$. This shows that $E_{\alpha,i}$ has empty interior for all α in Γ therefore E_i also has empty interior.

Set $\Delta = (\bigcup_{i=1}^p E_i)^c$, Δ is the complement of a finite union of definable sets with empty interiors so it is definable and dense. Lemma 4 implies that there are only finitely many values α such that $\{s \in \mathbb{R}^m \mid (\alpha, s) \in \Delta\}$ is not dense in $\mathbb{R}^m$. Therefore, we may refine further Γ by removing finitely many points, and refine Δ accordingly such that it satisfies the desired projection property: for every $\alpha \in \Gamma$, the set $\{s \in \mathbb{R}^m \mid (\alpha, s) \in \Delta\}$ is dense in $\mathbb{R}^m$.

Now, fix $\alpha \in \Gamma$ and s such that $(\alpha, s) \in \Delta$. Consider the set

$$K_{\alpha,s} = \left\{w \in L_s \mid \Phi'_{\alpha,s}(w) = I_p - \alpha\nabla^2_w g(w, s) \quad \text{is not invertible}\right\}$$

where I_p is the identity matrix of size p. Diagonalizing $\nabla^2_w g(w, s)$, the determinant of $\Phi'_{\alpha,s}(w)$ is $\prod_{i=1}^p (1 - \alpha\lambda_i(w, s))$. It is equal to zero if and only if there exists $i \in \{1, \ldots, p\}$ such that $\alpha\lambda_i(w, s) = 1$ hence

$$K_{\alpha,s} = \bigcup_{i=1}^{p} \{w \in L_s \mid \alpha\lambda_i(w, s) = 1\}.$$

Since $\alpha \in \Gamma$ and $(\alpha, s) \in \Delta$, by construction of Δ, α^{-1} is a regular value for the functions $w \mapsto \lambda_i(w, s)$, defined for $w \in L_s$, for all $i = 1, \ldots, p$. So the set $K_{\alpha,s}$ is a union of $p - 1$ dimensional submanifolds in L_s and $K^c_{\alpha,s}$ is open and dense set in L_s. Then, let $Z \subset \mathbb{R}^p$

definable and such that $\dim Z \leq p-1$. Suppose by contradiction that there exists a nonempty open set $V \subset \Phi_{\alpha,s}^{-1}(Z)$. The intersection $V \cap K_{\alpha,s}^c$ is open and nonempty because $K_{\alpha,s}^c$ is dense and both sets are open. Since $\Phi_{\alpha,s}$ is a local diffeomorphism on $K_{\alpha,s}^c$, the image $\Phi_{\alpha,s}(V \cap K_{\alpha,s}^c)$ has a nonempty interior but is included in Z of dimension $p-1$, which is a contradiction. The claim is proved. □

The following claim is a consequence of Lemma 4.

Claim 2 *Let $g\colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$ be a definable function and $v\colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^p \times \mathbb{R}^m$ be a definable map such that $\nabla g = v$ on a definable dense open set $C \subset \mathbb{R}^p \times \mathbb{R}^m$. Then there is a definable set $Z \subset \mathbb{R}^p$, dense, such that for all $w \in Z$, $\nabla g(w,s) = v(w,s)$ for all s in a definable dense open set in $\mathbb{R}^m$.*

Theorem 7 (Most subgradient sequences are gradient sequences) *Let g a definable function and $(s_k)_{k\in\mathbb{N}}$ a sequence of $\mathbb{R}^m$. Consider the sequence generated by*

$$w_{k+1} = w_k - \alpha_k v(w_k, s_k) \text{ for all } k \in \mathbb{N} \tag{16}$$

where $v : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^p$ is a definable map such that there exists a definable dense open set $C \subset \mathbb{R}^p \times \mathbb{R}^m$, where for all $(w,s) \in C$, $\nabla_w g(w,s) = v(w,s)$. Then there exist a set $W_0 \subset \mathbb{R}^p$ of full measure and residual, and a set $\Gamma \subset \mathbb{R}$ whose complement is finite such that if $w_0 \in W_0$ and $\alpha_k \in \Gamma$ for all $k \in \mathbb{N}$, then for all $k \in \mathbb{N}$

- *$w_{k+1} = w_k - \alpha_k \nabla_w g(w_k, s_k)$ for s_k in a definable dense open set.*
- *$g(\cdot, s_k)$ is C^2 in a neighborhood of w_k for s_k in a definable dense open set.*

Proof. Let $L \subset \mathbb{R}^p \times \mathbb{R}^m$, $\Gamma \subset \mathbb{R}$, $\Delta \subset \Gamma \times \mathbb{R}^m$, given by Claim 1. g is C^2 on L which is definable dense open, and for every $\alpha \in \Gamma$, the definable set $\{s \in \mathbb{R}^m \mid (\alpha, s) \in \Delta\}$ is dense open in $\mathbb{R}^m$.

L is definable and dense, therefore by Lemma 4 there exists a definable dense set $W \subset \mathbb{R}^p$ such that for all $w \in W$, $\{s \in \mathbb{R}^m \mid (w,s) \in L\}$ is dense in $\mathbb{R}^m$. By Claim 2 there exist definable dense and open subsets $V \subset \mathbb{R}^p$, such that for all $w \in V$, $\nabla_w g(w,s) = v(w,s)$ for all s in $\{s \in \mathbb{R}^m \mid (w,s) \in C\}$ definable and dense. For all $w \in V \cap W$ the following are satisfied:

- $\nabla_w g(w,s) = v(w,s)$ for s in $\{s \in \mathbb{R}^m \mid (w,s) \in C\}$, definable and dense.
- $g(\cdot, s)$ is C^2 in a neighborhood of w for s in $\{s \in \mathbb{R}^m \mid (w,s) \in L\}$ definable and dense.

Set for all $k \in \mathbb{N}$ $\Delta_k := \{s \in \mathbb{R}^m \mid (\alpha_k, s) \in \Delta\}$, $C_k := \{s \in \mathbb{R}^m \mid (w_k, s) \in C\}$ and $L_k := \{s \in \mathbb{R}^m \mid (w_k, s) \in L\}$ which are definable and dense. Suppose for all $k \in \mathbb{N}$ $\alpha_k \in \Gamma$ and $\xi_k \in \Delta_k \cap C_k \cap L_k$. We have to show for each $k \in \mathbb{N}$, w_k defined by (16) belongs to $V \cap W$.

Denote for $(\alpha, s) \in \Gamma \times \mathbb{R}^m$ $\Phi_{\alpha,s} := \mathrm{Id}_p - \alpha \nabla_w g(\cdot, s)$ and $\Psi_k := \prod_{i=0}^{k} \Phi_{\alpha_i, \xi_i}$ and remark that $\Psi_k(w_0) = w_{k+1}$ for all $k \in \mathbb{N}$.

For all $k \in \mathbb{N}$, by definition of Δ_k, if $\xi_k \in \Delta_k \cap C_k \cap L_k$ then $(\alpha_k, \xi_k) \in \Delta$. Following Claim 1 the function $\Phi_{\alpha_k,\xi_k} := \mathrm{Id}_p - \alpha_k \nabla_w g(\cdot, \xi_k)$ from $\{w \in \mathbb{R}^p \mid (w, \xi_k) \in L\}$, dense and open, to $\mathbb{R}^p$ verifies

$$\forall Z \subset \mathbb{R}^p \text{ definable, } \dim Z \leq p-1 \implies \dim \Phi^{-1}_{\alpha_k,\xi_k}(Z) \leq p-1. \tag{17}$$

Define the sequence $N_0 := (V \cap W)^c$ and $N_{k+1} := \Phi^{-1}_{\alpha_k,\xi_k}(N_k)$ for $k \in \mathbb{N}$ and remark that for $k \in \mathbb{N}$, $N_{k+1} = \Psi_k^{-1}((V \cap W)^c)$. Furthermore, the set of initialization points w_0 such that for all $k \in \mathbb{N}$, $w_k \in V \cap W$ is $W_0 := \bigcap_{k\in\mathbb{N}} N_k^c$. Since by definability of $V \cap W$ we have $\dim(V \cap W)^c \leq p-1$, we can apply the property (17) recursively on the $(N_k)_{k\in\mathbb{N}}$ and obtain for all $k \in \mathbb{N}$ $\dim N_{k+1} = \dim \Psi_k^{-1}((V \cap W)^c) \leq p-1$ provided that $\xi_k \in \Delta_k \cap C_k \cap L_k$. For all $k \in \mathbb{N}$, N_k is a definable closed set with empty interior so it has zero measure. $W_0 = \bigcap_{k\in\mathbb{N}} N_k^c$ is a countable intersection of dense open sets with empty interiors so by Baire's theorem it is also a dense open set. It is a residual set by countable intersection of residual sets, and it has full measure as a countable intersection of full measure sets.

If $w_0 \in W_0$ which is of full measure and residual, if for all $k \in \mathbb{N}$, $\alpha_k \in \Gamma$ whose complement is finite and $\xi_k \in \Delta_k \cap C_k \cap L_k$ which is definable dense, then the following hold for all $k \in \mathbb{N}$:

- $\forall k \in \mathbb{N}$, $w_{k+1} = w_k - \alpha_k \nabla_w g(w_k, \xi_k)$.
- $g(\cdot, \xi_k)$ is C^2 in a neighborhood of w_k.

□

5 Proofs of Section 2

Proof of Lemma 1. Since f is semialgebraic, for all $s \in S$ the function $f(\cdot, s)$ is path-differentiable [13, Proposition 2] with conservative gradient $\partial^c_w f(\cdot, s)$. By Assumption 1, for all compact set $K \subset \mathbb{R}^p$ there exists a square P-integrable (hence P-integrable) function $\kappa : \mathbb{R}^m \to \mathbb{R}$ such that $\forall (x, y) \in K, |f(x, s) - f(y, s)| \leq \kappa(s)\|x - y\|$ for all $s \in \mathbb{R}^m$ which implies $\|\partial^c_w f(x, s)\| \leq \kappa(s)$ for all $(x, s) \in K \times \mathbb{R}^m$. All the conditions are satisfied to apply Theorem 4 hence $\mathcal{J}$ admits a chain rule. □

Proof of Theorem 1. According to Lemma 1 $\mathcal{J}$ is path-differentiable with respect to the set-valued map $\mathbb{E}_{\xi\sim P}[\partial^c_w f(\cdot, \xi)]$. (i) is a consequence of [15, Lemma 21], [29, Theorem 3.2] and [8, Proposition 9] which apply to the setting of conservative gradients. We rewrite the sequence (2) as

$$w_{k+1} = w_k - \alpha_k(a(w_k) + \eta_k),$$

where $a = \mathbb{E}_{\xi\sim P}[v(\cdot, \xi)]$ and $\eta_k = v(w_k, \xi_k) - a(w_k)$. By joint measurability of v, a is a measurable selection of the Aumann integral $\mathbb{E}_{\xi\sim P}[\partial^c_w f(\cdot, \xi)]$, see Definition 3. $(\eta_k)_{k\in\mathbb{N}}$ verifies for all $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}[\eta_{k+1}|\xi_0, \ldots, \xi_k] &= \mathbb{E}[v(w_{k+1}, \xi_{k+1}) - a(w_{k+1})|\xi_0, \ldots, \xi_k] \\ &= \mathbb{E}_{\xi\sim P}[v(w_{k+1}, \xi)] - a(w_{k+1}) = 0.\end{aligned}$$

Following [29, Remark 4.6], see also [6], we have to verify

$$\lim_{R\to+\infty} \sup_{k\in\mathbb{N}} \mathbb{E}\left[\|\eta_{k+1}\|\mathbb{1}_{\{\|\eta_k\|\geq R\}}|\xi_0,\dots,\xi_k\right] = 0.$$

By Assumption 1 there exists an integrable function $\kappa : \mathbb{R}^m \to \mathbb{R}_+$ with respect to P such that $\forall x, y \in \overline{B(0,M)}$, $|f(x,s) - f(y,s)| \leq \kappa(s)\|x-y\|$. It implies that for all $(w,s) \in \overline{B(0,M)} \times \mathbb{R}^m$, $\|\partial_w^c f(w,s)\| \leq \kappa(s)$, where we recall that $\|\partial_w^c f(w,s)\| := \sup_{y\in\partial_w^c f(w,s)} \|y\|$. By assumption, there exists $M > 0$ such that $\sup_{k\in\mathbb{N}}\|w_k\| \leq M$ almost surely. Since $\forall (w,s) \in \mathbb{R}^p \times \mathbb{R}^m, v(w,s) \in \partial_w^c f(w,s)$, we have for all $w \in \overline{B(0,M)}$, $\|a(w)\| \leq \mathbb{E}_{\xi\sim P}\left[\|v(w,\xi)\|\right] \leq \mathbb{E}_{\xi\sim P}\left[\kappa(\xi)\right] < +\infty$. On the event $\sup_{k\in\mathbb{N}}\|w_k\| \leq M$ we have for all $k \in \mathbb{N}$,

$$\begin{aligned}\|\eta_k\| = \|v(w_k,\xi_k) - a(w_k)\| &\leq \|\partial_w^c f(w_k,\xi_k)\| + \|a(w_k)\| \\ &\leq \kappa(\xi_k) + \|a(w_k)\| \\ &\leq \kappa(\xi_k) + \mathbb{E}_{\xi\sim P}\left[\kappa(\xi)\right]. \end{aligned} \tag{18}$$

Now set for $\xi \in \mathbb{R}^m$ $u(\xi) := \kappa(\xi) + \mathbb{E}_{\xi\sim P}\left[\kappa(\xi)\right]$. Then we have for all $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[u(\xi_{k+1})\mathbb{1}_{\{u(\xi_{k+1})\geq R\}}|\xi_0,\dots,\xi_k\right] =& \mathbb{E}_{\xi\sim P}\left[u(\xi)\mathbb{1}_{\{u(\xi)\geq R\}}\right] \\ &+ \mathbb{E}_{\xi\sim P}\left[\kappa(\xi)\right]\mathbb{E}_{\xi\sim P}\left[\mathbb{1}_{\{u(\xi)\geq R\}}\right].\end{aligned}$$

Finally,

$$\begin{aligned}\sup_{k\in\mathbb{N}} \mathbb{E}\left[\|\eta_{k+1}\|\mathbb{1}_{\{\|\eta_{k+1}\|\geq R\}}|\xi_0,\dots,\xi_k\right] \leq& \mathbb{E}_{\xi\sim P}\left[u(\xi)\mathbb{1}_{\{u(\xi)\geq R\}}\right] \\ &+ \mathbb{E}_{\xi\sim P}\left[\kappa(\xi)\right]\mathbb{E}_{\xi\sim P}\left[\mathbb{1}_{\{u(\xi)\geq R\}}\right]\end{aligned} \tag{19}$$

where we used the inequality (18) which implies $\mathbb{1}_{\{\|\eta_{k+1}\|\geq R\}} \leq \mathbb{1}_{\{u(\xi_{k+1})\geq R\}}$. By the Dominated Convergence Theorem, the right side in (19) converges to 0, hence

$$\lim_{R\to+\infty} \sup_{k\in\mathbb{N}} \mathbb{E}\left[\|\eta_{k+1}\|\mathbb{1}_{\{\|\eta_k\|\geq R\}}|\xi_0,\dots,\xi_k\right] = 0.$$

and [29, Theorem 3.2] applies. Combining [15, Lemma 21] (see Remark 2 for its validity in this setting), [29, Theorem 3.2] and [8, Proposition 9], item (i) is proved.

In order to prove item (ii) we first apply Theorem 7 which gives a set $W_0 \subset \mathbb{R}^p$ of full measure and residual, and a set $\Gamma \subset \mathbb{R}$ whose complement is finite such that if $w_0 \in W_0$ and $\alpha_k \in \Gamma$ for all $k \in \mathbb{N}$, then with probability 1 we have for all $k \in \mathbb{N}$:

- $w_{k+1} = w_k - \alpha_k \nabla_w f(w_k, \xi_k)$.
- $f(\cdot, \xi_k)$ is C^2 in a neighborhood of w_k.

Let $k \in \mathbb{N}$, $h \in \mathbb{R}^p$. By assumption there exists a square integrable function $\kappa : \mathbb{R}^m \to \mathbb{R}_+$ such that for all $a \in [-1,1]$, $|f(w_k + ah, s) - f(w_k, s)| \leq \kappa(s)|a|\|h\|$ for P-almost all

$s \in \mathbb{R}^m$. We can apply the Dominated Convergence Theorem to obtain

$$\begin{aligned}\lim_{a\to 0} \frac{\mathcal{J}(w_k + ah) - \mathcal{J}(w_k)}{a} &= \lim_{a\to 0} \frac{1}{a} \int_S f(w_k + ah, s) - f(w_k, s) \, \mathrm{d}P(s) \\ &= \int_S \lim_{a\to 0} \frac{f(w_k + ah, s) - f(w_k, s)}{a} \, \mathrm{d}P(s) \\ &= \int_S \langle \nabla_w f(w_k, s), h \rangle \, \mathrm{d}P(s) \\ &= \langle \int_S \nabla_w f(w_k, s) \, \mathrm{d}P(s), h \rangle.\end{aligned}$$

Thus, for all $k \in \mathbb{N}$, with probability 1, $\mathcal{J}$ is differentiable at w_k with the equality $\nabla \mathcal{J}(w_k) = \mathbb{E}_{\xi\sim P}\left[\nabla_w f(w_k, \xi)\right]$. $(w_k)_{k\in\mathbb{N}}$ hence verifies with probability 1

$$w_{k+1} = w_k - \alpha_k \left(\nabla \mathcal{J}(w_k) + \eta_{k+1}\right) \text{ for all } k \in \mathbb{N},$$

where $\eta_k := \nabla_w f(w_k, \xi_k) - \nabla \mathcal{J}(w_k)$ is a martingale increment. The result then follows from [15, Lemma 21], [29, Theorem 3.2] and [8, Proposition 9]. □

Proof of Theorem 2. Under Assumption 2, the function $\mathcal{J}$ is definable using [21, Theorem 1.3]. From Theorem 6 the set-valued map $D_{\mathcal{J}} = \operatorname{conv} \mathbb{E}_{\xi\sim P}\left[\partial^c f(\cdot, \xi)\right]$ is definable and it is a conservative gradient for $\mathcal{J}$ by Lemma 1. By definable Sard's theorem the set of $D_{\mathcal{J}}$-critical values of $\mathcal{J}$ which is $\mathcal{J}(\operatorname{crit} D_{\mathcal{J}})$ is definable. Assuming for all $k \in \mathbb{N}$ $\alpha_k > 0$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty$, all conditions are satisfied to apply Theorem 5 hence all accumulations points $\bar{w}$ satisfy $0 \in D_{\mathcal{J}}(\bar{w})$ and $\mathcal{J}(w_k)$ converges.

Theorem 1 gives us $\Gamma \subset \mathbb{R}$ whose complement is finite, $W_0 \subset \mathbb{R}^p$ of full measure and residual, such that if $\alpha_k \in \Gamma$ for all $k \in \mathbb{N}$ then for all initialization w_0 in W_0 we have with probability 1 the relation

$$w_{k+1} = w_k - \alpha_k(\nabla \mathcal{J}(w_k) + \eta_k) \text{ for all } k \in \mathbb{N}$$

where $\nabla \mathcal{J}(w_k) = \mathbb{E}_{\xi\sim P}\left[\nabla_w f(w_k, \xi_k)\right]$ and $\eta_k = \nabla_w f(w_k, \xi_k) - \nabla \mathcal{J}(w_k)$. Under Assumption 1.1, as in (18) we have $\|\eta_k\| \leq \kappa(\xi_k) + \mathbb{E}_{\xi\sim P}\left[\kappa(\xi)\right]$ for all $k \in \mathbb{N}$, hence $\sup_{k\in\mathbb{N}} \mathbb{E}[\|\eta_k\|^2] < +\infty$ by square integrability of κ. Since the function $\mathcal{J}$ is definable, $\partial^c \mathcal{J}$ is also definable. Then by definable Sard's theorem the Clarke critical values of $\mathcal{J}$ are finite. By Lemma 1 and Lemma 3 $\mathcal{J}$ is a Lyapunov function for $\operatorname{crit} \partial^c \mathcal{J}$ and the differential inclusion $\dot{w} \in -\partial^c \mathcal{J}(w)$. Assume furthermore that for all $k \in \mathbb{N}$ $\alpha_k > 0$, $\sum_{k\in\mathbb{N}} \alpha_k^2 < +\infty$. Then all conditions are satisfied to apply the results in [5] as in the proof of Theorem 5 and deduce that with probability 1, the sequence $(\mathcal{J}(w_k))_{k\in\mathbb{N}}$ converges as $k \to +\infty$ and all accumulation points $\bar{w}$ of $(w_k)_{k\in\mathbb{N}}$ are Clarke critical, i.e.,verify $0 \in \partial^c \mathcal{J}(\bar{w})$. □

Proof of Theorem 3. The regression case, when P has a semialgebraic density with respect to the Lebesgue measure, is a direct application of Theorem 2 which holds with the backpropagation oracle, see Remark 2. The classification case uses similar arguments as mentioned in Remark 3. □

6 Generalized gradients of Norkin and conservativity

6.1 Definitions

Throughout this section, $f\colon \mathbb{R}^p \to \mathbb{R}$ is Lipschitz continuous and $D\colon \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ is locally bounded nonempty convex valued and upper semicontinuous. Convex values are indeed required by Norkin in [34, 35, 36].

Definition 9 (Semismooth generalized gradients) The set-valued mapping D is a *generalized gradient of* f if for all $x \in \mathbb{R}^p$, we have

$$\limsup_{y \to x,\, g \in D(y)} \frac{f(y) - f(x) - \langle g, y - x\rangle}{\|y - x\|} = 0.$$

The lim sup property in the definition is referred to as the *semismoothness property* of the generalized gradients. On the other hand, conservative gradients can be defined through the chain rule along absolutely continuous curves as in Definition 5. In both cases, the corresponding set-valued gradient maps have to be singletons almost everywhere while containing the Clarke subgradient of f everywhere [34, 13]. Functions with generalized gradient are called *differentiable in the generalized sense*. Recall that functions with conservative gradients are called path-differentiable.

6.2 Relations between the two notions

The following strenghtens [44, Theorem 1] which shows a chain rule with respect to a class of semismooth curves, strictly smaller than the class of absolutely continuous curves.

Proposition 4 *If D is a generalized gradient of f in the sense of Definition 9, then it is a conservative gradient of f.*

Proof. We shall use the chain rule along absolutely continuous path characterization of conservative gradients in Definition 5. Let D be a generalized gradient of f as in Definition 9, and $\gamma\colon [0,1] \to \mathbb{R}^p$ be an absolutely continuous path. Then both γ and $f \circ \gamma$ are absolutely continuous, hence differentiable almost everywhere. Therefore, there exists a full measure subset $R \subset [0,1]$ such that both are differentiable at every point on R.

Suppose, toward a contradiction, that the chain rule is not valid along γ, that is, there exists a non zero set $E_1 \subset R$ such that for all $t \in E_1$, there is $g \in D(\gamma(t))$ such that $\frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(t) \neq \langle g, \dot\gamma(t)\rangle$. Note that this implies that $\dot\gamma(t) \neq 0$ for all $t \in E_1$, since if $\dot\gamma(t) = 0$ then $0 = \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(t) = \langle g, \dot\gamma(t)\rangle$. Reducing E_1 and changing sign if necessary, we may assume without loss of generality that for all $t \in E_1$, there is $g \in D(\gamma(t))$ such that $\frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(t) < \langle g, \dot\gamma(t)\rangle$.

Consider the measurable function (measurability is justified in [13]), $g\colon [0,1] \to \mathbb{R}^p$, defined for all $t \in R$ by $g(t) = \underset{v \in D(\gamma(t))}{\arg\max} \langle \dot\gamma(t), v\rangle$ and $g(t) = 0$ otherwise.

We have for all $t \in E_1$, $0 < \langle \dot{\gamma}(t), g(t) \rangle - \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(t)$. This means that there is $\epsilon > 0$ and a nonzero set $E_2 \subset E_1$ such that $\epsilon \leq \langle \dot{\gamma}(t), g(t) \rangle - \frac{\mathrm{d}}{\mathrm{d}t} f \circ \gamma(t)$ for all $t \in E_2$ (otherwise, one would have $\langle \dot{\gamma}, g \rangle - \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma) = 0$ almost everywhere on E_1).

Let us apply Lusin's theorem (see, e.g., [42, Section 3.3]), fix an arbitrary $\alpha > 0$, such that $\lambda(E_2) > \alpha$, there is a closed subset $E_3 \subset E_2$ such that $\lambda(E_2 \setminus E_3) < \alpha$ and g restricted to E_3 is continuous. The set E_3 has positive measure since $\lambda(E_3) = \lambda(E_2) - \lambda(E_2 \setminus E_3) > \alpha - \alpha = 0$. Let us summarize, $E_3 \subset [0, 1]$ is closed with positive measure and we have the following on E_3:

- Both $f \circ \gamma$ and γ have derivatives and $\dot{\gamma} \neq 0$.
- $\frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma) + \epsilon \leq \langle \dot{\gamma}, g \rangle$.
- g restricted to E_3 is continuous.

Lebesgue density theorem (see, e.g., [28, Theorem 1.35]) ensures that almost all $t \in E_3$ have density 1, that is,

$$\frac{\lambda([t - \delta, t + \delta] \cap E_3)}{\lambda([t - \delta, t + \delta])} \underset{\delta \to 0}{\to} 1.$$

Since E_3 has positive measure, there exists $\bar{t} \in E_3$, a point of density 1 in E_3. We have for all $t \neq \bar{t}$, such that $\gamma(t) \neq \gamma(\bar{t})$,

$$\begin{aligned}
&\frac{f(\gamma(t)) - f(\gamma(\bar{t}))}{(t - \bar{t})} \\
&= \frac{\|\gamma(t) - \gamma(\bar{t})\|}{t - \bar{t}} \frac{f(\gamma(t)) - f(\gamma(\bar{t}))}{\|\gamma(t) - \gamma(\bar{t})\|} \\
&= \frac{\|\gamma(t) - \gamma(\bar{t})\|}{t - \bar{t}} \left(\frac{f(\gamma(t)) - f(\gamma(\bar{t})) - \langle g(t), \gamma(t) - \gamma(\bar{t}) \rangle}{\|\gamma(t) - \gamma(\bar{t})\|} + \frac{\langle g(t), \gamma(t) - \gamma(\bar{t}) \rangle}{\|\gamma(t) - \gamma(\bar{t})\|} \right) \\
&= \frac{\|\gamma(t) - \gamma(\bar{t})\|}{(t - \bar{t})} \left(\frac{f(\gamma(t)) - f(\gamma(\bar{t})) - \langle g(t), \gamma(t) - \gamma(\bar{t}) \rangle}{\|\gamma(t) - \gamma(\bar{t})\|} \right) + \frac{\langle g(t), \gamma(t) - \gamma(\bar{t}) \rangle}{t - \bar{t}}.
\end{aligned}$$

Letting $t \to \bar{t}$ with $t \in E_3$, $t \neq \bar{t}$ and $\gamma(t) \neq \gamma(\bar{t})$, which is possible because $\bar{t}$ has density 1 in E_3 and $\dot{\gamma}(\bar{t}) \neq 0$, we have

$$\begin{aligned}
&\frac{f(\gamma(t)) - f(\gamma(\bar{t}))}{(t - \bar{t})} \to \frac{\mathrm{d}}{\mathrm{d}t}(f \circ \gamma)(\bar{t}) \\
&\frac{\|\gamma(t) - \gamma(\bar{t})\|}{(t - \bar{t})} \to \|\dot{\gamma}(\bar{t})\| \\
&\frac{f(\gamma(t)) - f(\gamma(\bar{t})) - \langle g(t), \gamma(t) - \gamma(\bar{t}) \rangle}{\|\gamma(t) - \gamma(\bar{t})\|} \to 0 \\
&\frac{\langle g(t), \gamma(t) - \gamma(\bar{t}) \rangle}{t - \bar{t}} \to \langle g(\bar{t}), \dot{\gamma}(\bar{t}) \rangle,
\end{aligned}$$

where the first two identities follow from the differentiability of $f \circ \gamma$ and γ at $\bar{t} \in E_3$, the third equality stems from the semismooth property of generalized gradients (Definition 9) while the last one is by differentiability of γ and continuity of g restricted to E_3 at $\bar{t}$. We

obtain that $\frac{\mathrm{d}}{\mathrm{d}t}(f\circ\gamma)(\bar{t}) = \langle g(\bar{t}), \dot{\gamma}(\bar{t})\rangle \geq \frac{\mathrm{d}}{\mathrm{d}t}(f\circ\gamma)(\bar{t}) + \epsilon$, where the equality follows by the previous limit and the inequality is because $\bar{t} \in E_3$. This is contradictory since $\epsilon > 0$, which concludes the proof. □

Whence the class of functions differentiable in the generalized sense is contained in the class of path-differentiable functions. In the semialgebraic case, both notions coincide [23], but in general the inclusion is strict as the following example shows.

Proposition 5 *Consider the closed set $C \subset [-1,1]$ defined through $C = \{1/k \mid k \in \mathbb{Z}, k \neq 0\} \cup \{0\}$. Then the distance function to C is path-differentiable but not differentiable in the generalized sense.*

Proof. First let us recall a substitution formula for absolutely continuous function [47, Corollary 7]. If $g\colon \mathbb{R} \to \mathbb{R}$ is absolutely continuous and $f\colon \mathbb{R} \to \mathbb{R}$ is measurable and bounded, then for all α, β

$$\int_{g(\alpha)}^{g(\beta)} f(x)dx = \int_{\alpha}^{\beta} f(g(s))\dot{g}(s)ds.$$

It is clear that $\partial^c F$ is locally constant ($+1$ or -1) out of a closed countable set (the set C and its cut locus). Therefore, choosing f to be any measurable selection in $\partial^c F$, the previous formula allows to conclude that F is path-differentiable. Indeed F is path-differentiable if and only it satisfies the change of variable formula for any absolutely continuous g, which is the case, because F is 1-Lipschitz so that $|f| \leq 1$.

On the other hand, $F(0) = 0$ and for all $k \in \mathbb{N}^*$, $F(1/k) = 0$ and $\partial^c F(1/k) = [-1,1]$ so that $-1 \in \partial^c F(1/k)$. The equality

$$\frac{F(1/k) - F(0) - \langle -1, 1/k - 0\rangle}{\|1/k - 0\|} = 1,$$

contradicts the semismoothness property for the Clarke subgradient of F. Since F is differentiable in the generalized sense if and only if its Clarke subgradient is a generalized gradient, we conclude that F is not differentiable in the generalized sense at 0. □

Acknowledgments

The authors acknowledge the financial support of the AI Interdisciplinary Institute ANITI funding under the grant agreement ANR-19-PI3A-0004, Air Force Office of Scientific Research, Air Force Material Command, USAF, under grant numbers FA9550-19-1-7026, and ANR MaSDOL 19-CE23-0017-01. J. Bolte also acknowledges the support of ANR Chess, grant ANR-17-EURE-0010 and TSE-P.

References

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga,

S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, *Tensorflow: A system for large-scale machine learning*, in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), 2016, pp. 265–283.

[2] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis (3rd edition)*, vol. 264, 2005.

[3] J. P. Aubin and A. Cellina, *Differential inclusions: set-valued maps and viability theory*, vol. 264, 1984.

[4] S. Bai, J. Z. Kolter, and V. Koltun, *Deep equilibrium models*, in Advances in Neural Information Processing Systems, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., vol. 32, Curran Associates, Inc., 2019.

[5] M. Benaim, J. Hofbauer, and S. Sorin, *Stochastic Approximations and Differential Inclusions*, vol. 44, 2005.

[6] M. Benaïm and S. J. Schreiber, *Ergodic properties of weak asymptotic pseudotrajectories for semiflows*, Journal of Dynamics and Differential Equations, 12 (2000), pp. 579–598, https://doi.org/10.1023/A:1026463628355.

[7] Q. Bertrand, Q. Klopfenstein, M. Blondel, S. Vaiter, A. Gramfort, and J. Salmon, *Implicit differentiation of lasso-type models for hyperparameter optimization*, in International Conference on Machine Learning, PMLR, 2020, pp. 810–821.

[8] P. Bianchi, W. Hachem, and S. Schechtman, *Convergence of constant step stochastic gradient descent for non-smooth non-convex functions*, (2020).

[9] P. Bianchi and R. Rios-Zertuche, *A closed-measure approach to stochastic approximation*, arXiv preprint arXiv:2112.05482, (2021).

[10] M. Blondel, Q. Berthet, M. Cuturi, R. Frostig, S. Hoyer, F. Llinares-López, F. Pedregosa, and J.-P. Vert, *Efficient and modular implicit differentiation*, arXiv preprint arXiv:2105.15183, (2021).

[11] J. Bolte, A. Daniilidis, A. Lewis, and M. Shiota, *Clarke subgradients of stratifiable functions*, SIAM Journal on Optimization, 18 (2007), pp. 556–572, https://doi.org/10.1137/060670080.

[12] J. Bolte, T. Le, E. Pauwels, and T. Silveti-Falls, *Nonsmooth implicit differentiation for machine-learning and optimization*, Advances in Neural Information Processing Systems, 34 (2021).

[13] J. Bolte and E. Pauwels, *Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning*, 2019, https://arxiv.org/abs/1909.10300.

[14] J. Bolte and E. Pauwels, *A mathematical model for automatic differentiation in machine learning*, 2020, https://arxiv.org/abs/2006.02080.

[15] J. Bolte, E. Pauwels, and R. Rios-Zertuche, *Long term dynamics of the subgradient method for lipschitz path differentiable functions*, ArXiv, abs/2006.00098 (2020).

[16] L. Bottou, F. E. Curtis, and J. Nocedal, *Optimization methods for large-scale machine learning*, SIAM Review, 60 (2018), p. 223–311, https://doi.org/10.1137/16m1080173.

[17] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang, *JAX: composable transformations of Python+NumPy programs*, 2018, http://github.com/google/jax.

[18] C. Castera, J. Bolte, C. Févotte, and E. Pauwels, *An inertial newton algorithm for deep learning*, Journal of Machine Learning Research, 22 (2021), pp. 1–31.

[19] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, *Neural ordinary differential equations*, Advances in neural information processing systems, 31 (2018).

[20] F. Clarke, *Optimization and Nonsmooth Analysis*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1990.

[21] R. Cluckers and D. J. Miller, *Stability under integration of sums of products of real globally subanalytic functions and their logarithms*, Duke Mathematical Journal, 156 (2011), p. 311–348, https://doi.org/10.1215/00127094-2010-213.

[22] M. Coste, *An introduction to o-minimal geometry*, 1999.

[23] D. Davis and D. Drusvyatskiy, *Conservative and semismooth derivatives are equivalent for semialgebraic maps*, Set-Valued and Variational Analysis, (2021), pp. 1–11.

[24] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee, *Stochastic subgradient method converges on tame functions*, Foundations of Computational Mathematics, 20 (2020), pp. 119–154, https://doi.org/10.1007/s10208-018-09409-5.

[25] L. Dries and C. Miller, *On the real exponential field with restricted analytic functions*, Israel Journal of Mathematics, 92 (1995), p. 427.

[26] E. Dupont, A. Doucet, and Y. W. Teh, *Augmented neural odes*, Advances in Neural Information Processing Systems, 32 (2019).

[27] Y. M. Ermol'ev and V. Norkin, *Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization*, Cybernetics and Systems Analysis, 34 (1998), pp. 196–215.

[28] L. C. Evans and R. F. Gariepy, *Measure theory and fine properties of functions*, vol. 5, CRC press Boca Raton, 1992.

[29] M. Faure and G. Roth, *Ergodic properties of weak asymptotic pseudotrajectories for set-valued dynamical systems*, 2011, https://arxiv.org/abs/1101.2154.

[30] S. Gadat, F. Panloup, and S. Saadane, *Stochastic heavy ball*, Electronic Journal of Statistics, 12 (2018), pp. 461–529.

[31] H. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, vol. 35, Springer Science & Business Media, 2003.

[32] S. Majewski, B. Miasojedow, and É. Moulines, *Analysis of nonsmooth stochastic approximation: the differential inclusion approach*, arXiv: Optimization and Control, (2018).

[33] S. Marx and E. Pauwels, *Path differentiability of ode flows*, arXiv preprint arXiv:2201.03819, (2022).

[34] V. Norkin, *Nonlocal minimization algorithms of nondifferentiable functions*, Cybernetics, 14 (1978), pp. 704–707.

[35] V. Norkin, *Generalized-differentiable functions*, Cybernetics and Systems Analysis - CYBERN SYST ANAL-ENGL TR, 16 (1980), pp. 10–12, https://doi.org/10.1007/BF01099354.

[36] V. Norkin, *Stochastic generalized-differentiable functions in the problem of nonconvex nonsmooth stochastic optimization*, Cybernetics and Systems Analysis - CYBERN SYST ANAL-ENGL TR, 22 (1986), pp. 804–809, https://doi.org/10.1007/BF01068698.

[37] V. Norkin, *Substantiation of the backpropagation technique via the hamilton—pontryagin formalism for training nonconvex nonsmooth neural networks*, 12 (2019), pp. 19–26, https://doi.org/10.15407/dopovidi2019.12.019.

[38] V. I. Norkin, *Stochastic generalized gradient methods for training nonconvex nonsmooth neural networks*, Cybernetics and Systems Analysis, 57 (2021), pp. 714–729, https://doi.org/10.1007/s10559-021-00397-z.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, *Pytorch: An imperative style, high-performance deep learning library*, in Advances in Neural Information Processing Systems 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., Curran Associates, Inc., 2019, pp. 8024–8035.

[40] E. Pauwels, *The ridge method for tame min-max problems*, arXiv preprint arXiv:2104.00283, (2021).

[41] F. Pedregosa, *Hyperparameter optimization with approximate gradient*, Proceedings of the 33 rd International Conference on Machine Learning, (2016).

[42] H. Royden, *Real Analysis*, Collier Macmillan international editions, Macmillan, 1968.

[43] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Learning representations by back-propagating errors*, Nature, 323 (1986), pp. 533–536.

[44] A. Ruszczyński, *Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization*, Optimization Letters, 14 (2020), pp. 1615–1625, https://doi.org/10.1007/s11590-020-01537-8.

[45] A. Ruszczynski, *A stochastic subgradient method for nonsmooth nonconvex multilevel composition optimization*, SIAM J. Control. Optim., 59 (2021), pp. 2301–2320.

[46] D. Sahoo, Q. Pham, J. Lu, and S. C. Hoi, *Online deep learning: Learning deep neural networks on the fly*, arXiv preprint arXiv:1711.03705, (2017).

[47] J. Serrin and D. E. Varberg, *A general chain rule for derivatives and the change of variables formula for the lebesgue integral*, The American Mathematical Monthly, 76 (1969), pp. 514–520.

[48] A. Shapiro and H. Xu, *Uniform laws of large numbers for set-valued mappings and subdifferentials of random functions*, Journal of Mathematical Analysis and Applications, 325 (2007), pp. 1390–1399, https://doi.org/10.1016/j.jmaa.2006.02.078.

[49] L. van den Dries and C. Miller, *Geometric categories and o-minimal structures*, Duke Math. J., 84 (1996), pp. 497–540, https://doi.org/10.1215/S0012-7094-96-08416-1.