Subgradient sampling for nonsmooth nonconvex minimization

Jérôme Bolte, Tam Le, Edouard Pauwels

Subgradient sampling for nonsmooth nonconvex minimization*

Jérôme Bolte† Tam Le† Edouard Pauwels ‡

October 28, 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 [31] 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 $(S,\mathcal{A})$. We assume throughout the paper that $\mathcal{J}$ is bounded below. This type of problems has many applications, we refer for instance to [51, 52] and references therein, for various examples

*Submitted to the editors February 18, 2022.

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

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

in several fields. Our specific interest goes in particular to online deep learning [53] and machine learning more broadly [19]. 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., [36].

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 [23, Theorem 2.7.3]. Without additional convexity or regularity assumptions, as in [38, 27], the inclusion is strict in general. In plain words, general subgradient sampling is not a stochastic subgradient method: expected increments may not be subgradients of the loss. As a consequence, the very nature of a first-order sampling method may generate undesired directions that could result in sporadically erratic movements or artificial and irrelevant steady states [16]. The situation is even worse in deep learning where the oracle is based on backpropagation [50], which may add more artifacts to the dynamics, see, e.g., [16] and references therein. Despite these issues, many real-world algorithms are designed according to this model, justifying the need for theoretical support and guarantees to the practical success of these methods.

This paper aims at addressing these problems in the vanishing step size regime. We build upon conservative calculus introduced in [15]. This approach makes rigorous the use of formal subdifferentiation in a wide mathematical framework encompassing most nonconvex nonsmooth problems. One of its main features is that sum or composition of conservative gradients are conservative. Under mild assumptions, we shall extend this property to integration, parameterized integrals of conservative gradients being conservative. An important advantage of this approach is its full compatibility with one of the core modern algorithms of large-scale optimization: backpropagation [1, 46, 20].

An essential series of works on subgradient sampling are those developed in [31], followed also by [51, 52], using the generalized derivatives and the calculus introduced in Norkin [41], see Section 6 for more details. They address, in particular, the interplay between expectations and subgradients under various assumptions, and ensure as well convergence

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

to "generalized critical points". The ideas of [31], and some follow-up research [43, 44], 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 [15] instead of the generalized derivative approach of [41], which are different techniques. Focus on the conservative case is motivated by a growing "theory" close to machine learning applications: implicit differentiation [14] with application to implicit neural networks [4], nondifferentiable programming [12], bilevel programming [48, 9]), differential equations [39] which are naturally connected to Neural ODEs [22, 29], or even partial minimization [47]. 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 [26].

– 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. In particular, we provide 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 [6] on stochastic approximations. The versatility and simplicity of this method allow for a quite direct extension to other types of first-order methods as, see e.g., [11, 35, 21]. 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 of Cluckers-Miller's integration result [24] as well as avoidance result for artificial critical points. Section 5 contains the proofs of Section 2 results.

Notations For $q \in \mathbb{N}^*$, $\mathcal{B}(\mathbb{R}^q)$ denotes the Borel sigma algebra on $\mathbb{R}^q$, λ is the Lebesgue measure on $\mathbb{R}$. $\|\cdot\|$ 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 $f : \mathbb{R}^p \to \mathbb{R}$ locally Lipschitz. By Rademacher's theorem, f is differentiable on a set

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

of full measure denoted diff_f. Its Clarke subgradient is defined for $x \in \mathbb{R}^p$ as

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

This extends to $g : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$ such that for almost all $s \in \mathbb{R}^m$ $g(\cdot, s)$ is locally Lipschitz. For almost all $s \in \mathbb{R}^m$, we denote $\partial^c_w g(\cdot, s)$ the Clarke subgradient of $g(\cdot, s)$.

For $F : \mathbb{R}^p \to \mathbb{R}^q$ locally Lipschitz, we can define the Clarke Jacobian as well,

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

and for $G : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^q$ such that $G(\cdot, s)$ is locally Lipschitz for almost all $s \in \mathbb{R}^m$, $\mathrm{Jac}^c_w G(\cdot, s)$ is the Clarke Jacobian of $G(\cdot, s)$ for almost all $s \in \mathbb{R}^m$.

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 and bounded below, in particular $f(x, \cdot)$ is P-integrable for all $x \in \mathbb{R}^p$. 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 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 almost all $s \in \mathbb{R}^m$, for all $w \in \mathbb{R}^p$, $v(w, s) \in \partial^c_w f(w, s)$.

Assumption 1

1. *There exists a measurable function $\kappa : \mathbb{R}^m \to \mathbb{R}_+$ such that for all $n \in \mathbb{N}$, κ^n is P-integrable and there exists $q_0 \in \mathbb{N}$ such that for almost all $s \in \mathbb{R}^m$,*

$$\forall x, y \in \mathbb{R}^p,\ |f(x, s) - f(y, s)| \leq \kappa(s)(1 + \max\{\|x\|, \|y\|\}^{q_0})\|x - y\|.$$

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. *The integrand f and the selection v are semialgebraic,*

4. $\sup_{k\in\mathbb{N}} \|w_k\| < \infty$ *almost surely.*

Observe that the first condition implies local Lipschitz continuity of $\mathcal{J}$ hence $\partial^c \mathcal{J}$ is well defined. Furthermore, beyond semialgebraicity, all the results of this work extend to globally subanalytic functions f and selections v, which encompasses the vast majority of machine learning examples. In particular, globally subanalytic functions are polynomially bounded and, if P has compact support, they satisfy Assumption 1 (1) (see Section 4.1 and [57]).

As to ensure Assumption 1 (4), some solutions have been proposed in the literature. In [31] the iterates are projected on a compact set. The authors of [18, 45] use a reset mechanism: the sequence is reset, either at w_0 in [45] or on a sphere in [18], when the objective function is too high. Such mechanisms are however hard to apply in our case since they require computing $\mathcal{J}(w_k)$ at each iteration.

Remark 1 (Probability spaces) *In this work, the term "almost surely" can refer to different probability spaces. Assumption 1 (1) refers to a probability space $(S, \mathcal{A}, P)$ where $S = \mathbb{R}^m$ and $\mathcal{A} = \mathcal{B}(\mathbb{R}^m)$, whereas in Assumption 1 (4), it is question of a countable product of probability spaces on which is defined the whole sequence $(w_k)_{k\in\mathbb{N}}$. In the paper, randomness has to be understood regarding these different probability spaces.*

A glance at the convergence results In this general setting, 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 behavior in a weaker sense introduced in [7] and extended to set-valued flows in [11], see also [33, 17]. This relies on 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 U of $\bar{w}$ one has

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

Intuitively, non essential accumulation points are hardly never seen. This is made more precise in [11, Corollary 4.9], 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:*

1. *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).

2. *If P has a density with respect to Lebesgue measure, 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 set, and with probability 1, $(w_k)_{k\in\mathbb{N}}$ *verifies for all* $k \in \mathbb{N}$,

$$w_{k+1} = w_k - \alpha_k \nabla_w f(w_k, \xi_k) \qquad \textit{and} \qquad \int_{\mathbb{R}^m} \nabla_w f(w_k, s) dP(s) = \nabla \mathcal{J}(w_k).$$

Also, 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 2 1. (Generalized criticality) The criticality notion (5) is inherent to the very nature of subgradient sampling methods and relates to the notion of an artificial critical point as described in [16]. Gradient artifacts can be generated by sampling from discrete or continuous distributions as in the two following examples: let $f\colon (w, s) \mapsto s|w|$ and $P = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ or P is the uniform density on $[-1, 1]$, then $\partial^c \mathcal{J}(w) = 0$ for all $w \in \mathbb{R}$, but $\mathbb{E}_{\xi\sim P}[\partial^c_w f(0, \xi)] = [-1, 1]$. Due to the chain rule characterization of conservative gradients, Definition 5. These senseless outputs are however extremely rare, this is why they do not impact first-order optimization much. This is shown in *2* through (6)

2. (Extension to conservative gradient oracles) Under the same assumptions, Theorem 1 (1) holds, mutatis mutandis, replacing the Clarke subgradient by a conservative gradient. Indeed, $\partial^c_w f$ *can be replaced by a semialgebraic set-valued map* D *where for* P*-almost all* $s \in \mathbb{R}^m$, $D(\cdot, s)$ *is a conservative gradient for* $f(\cdot, s)$*. In this case, Theorem 1 (2) also holds true* as is.

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 for all accumulation points. In addition,* $\mathcal{J}(w_k)$ *converges almost surely.*

Remark 3 *1. (On proofs) In order to prove item 2, we use a property of semialgebraic functions called* stratification*, see Definition 12. The intermediary results leading to item 2 are gathered in Section 4. The proof of Theorem 2 is based on results from [6, 11] and an abstract stochastic approximation version is found in Section 3.3.*

2. (On Assumption 2) 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 $\mathcal{J}$*. 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. Beyond semialgebraicity,* P *can be assumed to be globally subanalytic, see Section 4.1 for the definition of such functions, which includes analytic densities with semialgebraic compact support, like truncated Gaussian distributions.*

Since definability and consequently Sard's property are preserved by finite sum, Theorem 1 and theorem 2 also hold for P *being the joint distribution between a discrete and a continuous random variable. This allows to consider classification tasks in online deep learning (see theorem 3). We limit ourselves to an absolutely continuous distribution for the sake of simplicity.*

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

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

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

$\mathcal{J}$ also admits a chain rule with respect to $\mathbb{E}_{\xi \sim P}[\partial^c_w f(\cdot, \xi)]$.

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 [15, Corollary 4] and extends [42, 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

Recent works [27, 38] assume the Clarke subgradient and integral operations to be interchangeable, which in practice requires regularity assumptions. Without such assumptions, as here, first-order sampling may produce absurd trajectories or lead to spurious critical points. This was first observed in [31] and rediscovered in [15, 16]. To avoid converging to these undesirable points [31, Remark 4.2] suggests perturbing iterates by random noise. Instead, our Theorem 2 ensures that "almost all" subgradient sampling sequences accumulate to Clarke critical points, justifying the correctness of the algorithm as implemented in practice. Avoidance of spurious critical points is also obtained in [10] through probabilistic and continuity arguments. The definable framework allows for a sharper description of the set of step sizes and initializations leading to spurious points. For instance, with a finite horizon K, the set of bad initializations is a finite union of manifolds of dimension strictly lower than p, while the set of bad step sizes is finite.

Due to the role of the ODE method in the analysis of stochastic optimization methods, considering risk functions having a chain rule is not new in the literature, either as an assumption [10] or using restrictive sufficient conditions [27, 38]. With Lemma 1, we provide instead simple and explicit sufficient conditions (see also Theorem 4 for a general form). Similarly, our Assumption 2 is sufficient to obtain a strong form of Sard's condition which is usually stated in a weak abstract form as a hypothesis [31, 6, 10].

Several works consider a constrained version of (1). For instance, [54] studies a proximal stochastic subgradient method and [31] considers a projection on a compact set. We decide to focus on the unconstrained setting, but a precise study of the constrained setting could be a matter of future work.

2.3 First-order sampling, backpropagation, and deep learning

Deep learning model Let P be a probability distribution on $\mathbb{R}^d \times \mathbb{R}^I$ called *population* distribution. One of the goals of machine learning is to build a predicting function $h\colon \mathbb{R}^d \mapsto \mathbb{R}^I$ such that $h(X) \simeq Y$ when (X, Y) is distributed according to P, written $(X, Y) \sim P$. 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.

In deep learning, the predictor h is a neural network defined by 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 an 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:

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

For $l = 1, \dots, L$, the function g_l can take several forms, see e.g. [37]. 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 $w_1, \dots, w_L$.

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:

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

The expectation is generally unknown and approximated through statistical sampling. Consider a sequence $(x_k, y_k)_{k\in\mathbb{N}}$ of i.i.d. samples generated from the distribution P and, let us tackle the problem (8) with the first-order sampling algorithm (2). In this setting, $(\xi_k)_{k\in\mathbb{N}}$ is $(x_k, y_k)_{k\in\mathbb{N}}$ while 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 [50]. 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:

$$\begin{aligned} \mathrm{backprop}_w f(w, x, y) \in \partial^c f_r(f_{r-1} \circ \ldots \circ f_1(w, x, y))^T \\ \times \mathrm{Jac}^c f_{r-1}(f_{r-2} \circ \ldots \circ f_1(w, x, y)) \times \ldots \times \mathrm{Jac}^c_w f_1(w, x, y). \end{aligned} \tag{9}$$

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

Algorithm 1 First-order sampling with backpropagation

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.
for $k = 0, 1, \ldots$ **do**
    $w_{k+1} = w_k - \alpha_k \operatorname{backprop}_{w_k} \ell(h(w_k, x_k), y_k)$
end for

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 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 with compact support) The joint distribution P satisfies one of the following:

− Regression: P has a semialgebraic density ϕ with compact support with respect to Lebesgue on $\mathbb{R}^d \times \mathbb{R}^I$.

− Classification: P is supported on $\mathbb{R}^d \times (e_i)_{i=1}^I$ where for $i = 1 \ldots I$, e_i is the i-th element of the canonical basis in $\mathbb{R}^I$. The distribution P factorizes as $P(X, Y) = P(Y)P(X|Y)$ and we assume P_Y is discrete over $(e_i)_{i=1}^I$ and $P(X|Y = e_i)$ has a density ϕ_i with respect to Lebesgue, that is semialgebraic and compactly supported.

In this setting, a consequence of Theorem 1 and Theorem 5 is the following:

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 $\sum_{k\in\mathbb{N}} \alpha_k = \infty$, $\alpha_k = o(1/\log(k))$. Assume furthermore that $\sup_{k\in\mathbb{N}} \|w_k\| < \infty$ almost surely.*

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

This theorem parallels [15, Theorem 9] for a general population distribution: assumptions are simple and widespread, it is fully compatible with backpropagation, and shows that

spurious outputs hardly matter for the training phase. Note however that the validity of absolutely continuous randomness is questionable in practice. In particular, the density of W, and the finiteness of Γ^c can be affected due to quantization errors. For instance, [8] empirically shows that gradient artifacts created by conservative calculus can impact the method.

Similar models have been considered in the literature, but with somehow stronger assumptions, e.g., [52] (which rules out nonsmooth activation functions). Another important contribution is given in [44], as discussed 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 nonempty 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 subsets $A \subset \mathbb{R}^p$, the upper inverse $F^u(A) := \{x \in X \mid F(x) \subset A\}$ is measurable in $(X, \mathcal{F})$.

Proposition 1 (Measurable selections, Theorem 18.13 [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, whence by Castaing's Theorem, $g \circ F$ 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

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

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

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

1. *(Graph closedness) F is* graph closed *if its graph defined by*

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

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

2. *(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 $y \in F(z)$, $\|y\| < M$.*

3. *(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 $D : \mathbb{R}^p \rightrightarrows \mathbb{R}^p$ that is nonempty valued 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 $U \subset \mathbb{R}$ open 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 is defined almost everywhere in $U \times S$. Also, for almost all $x \in U$, $\frac{\partial f}{\partial x}(x, s)$ is defined for almost all $s \in S$.*

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

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

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

Proposition 2 (Differentiation of absolutely continuous integrals) *Let $U \subset \mathbb{R}$ open 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 at almost all $x \in U$ with $g'(x) = \int_S \frac{\partial f}{\partial x}(x,s)\,\mathrm{d}\mu(s)$.

Proof. Let $f : U \times S \to \mathbb{R}$ satisfying 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 [49]) implies that

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

Under Assumption 3, by measure completeness, Fubini-Lebesgue's Theorem applies: $g(b) - g(a) = \int_a^b \int_S \frac{\partial f}{\partial t}(t,s)\,\mathrm{d}\mu(s)\,\mathrm{d}t$. 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. By the fundamental theorem of calculus, $g'(x)$ is defined for almost all $x \in U$ 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. It can be seen as interchanging 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 for all $(x, s) \in C \times S$, $\|D(x, s)\| \leq \kappa(s)$, 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. With the chain rule definition of conservativity, Definition 5, we will show that the problem reduces to the differentiation of an absolutely continuous integral and then, we shall 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 the assumptions 1 to 4 displayed above.

Following the conservative gradient definition, Definition 5, in particular we verify $\int_S D(\cdot, s)\, \mathrm{d}\mu(s)$ is graph closed, nonempty valued and locally bounded. By the measurable selection theorem, see Proposition 1, $\int_S D(\cdot, s)\, \mathrm{d}\mu(s)$ is nonempty valued. It is locally bounded by item 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 [56, Theorem 2], and since $D(\cdot, s)$ is upper semicontinuous, compact valued for all s, then $\int_S D(\cdot, s)\, \mathrm{d}\mu(s)$ 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 $(\mathbb{R} \times S, \mathcal{B}(\mathbb{R}) \otimes \mathcal{A})$. Similarly as in the proof of Lemma 2, by Fubini's Theorem, $\varphi^u(\{0\})$ 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 $a = \int_S v(s)\,\mathrm{d}s$,

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

On the other hand, $\gamma([0,1])$ is compact by continuity of γ. Let κ given by assumption 4 for the compact set $C = \gamma([0,1])$. The Cauchy-Schwarz inequality gives 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 $[0,1]$ hence the function $(t,s) \mapsto \kappa(s)\|\dot{\gamma}(t)\|$ is locally integrable jointly in (t,s), and so is $(t,s) \mapsto \frac{\mathrm{d}}{\mathrm{d}t} f(\gamma(t), s) = \langle v(s), \dot{\gamma}(t) \rangle$. Proposition 2 now applies to $(t,s) \mapsto f(\gamma(t), s)$, hence 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 $(S, \mathcal{A})$, denote $\operatorname{supp} P$ its support, and consider a jointly measurable set-valued map $D : \mathbb{R}^p \times S \rightrightarrows \mathbb{R}^p$ such that for almost all $s \in S$, $f(\cdot, s)$ is locally Lipschitz and $D(\cdot, s)$ is a convex-valued conservative gradient for $f(\cdot, s)$. We consider the sequence $(w_k)_{k \in \mathbb{N}}$ defined by (2) where v is now a measurable selection of D, i.e., for all $(w,s) \in \mathbb{R}^p \times S$, $v(w,s) \in D(w,s)$. For all $k \in \mathbb{N}^*$, the truncated random sequence $(w_0, \ldots, w_k)$ is defined on the product probability space $(S^k, \mathcal{A}^{\otimes \mathbb{N}}, P^{\otimes k})$, and the whole trajectory $(w_k)_{k \in \mathbb{N}}$ is defined on the countable product $(S^{\mathbb{N}}, \mathcal{A}^{\otimes k}, P^{\otimes \mathbb{N}})$. When there is no ambiguity, the term "almost sure" implicitly refers to these spaces.

In order to study the sequence $(w_k)_{k \in \mathbb{N}}$, we use the nonsmooth ODE methods developed in [6, 11]. To this end, we consider the set-valued map

$$D_{\mathcal{J}} : w \mapsto \mathbb{E}_{\xi \sim P}[D(w, \xi)]$$

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 $t \in \mathbb{R}_+$, $\dot{w}(t) \in -D_{\mathcal{J}}(w(t))$.

Remark 5 (Solution to the differential inclusion on $\mathbb{R}_+$) *Since our analysis is based on a nonsmooth ODE method, the differential inclusion (14) must admit solutions. Available theory [34, 3] requires a convex valued conservative gradient D, which is satisfied for $D = \partial^c_w f$ or replacing D by its convex hull.*

The flow of the differential inclusion (14) is then nonempty and well defined on $\mathbb{R}_+$ as $\mathcal{J}$ in (1) is assumed to be bounded below. Indeed, consider a strict lower bound $\mathcal{J}^* \in \mathbb{R}$. Fix $T > 0$ and $w_0 \in \mathbb{R}^p$, let $\mathcal{D} = [-T, T] \times \overline{B(w_0, r)}$ with $r := 2T\sqrt{(\mathcal{J}(w_0) - \mathcal{J}^*)} > 0$, which is closed, bounded domain and contains w_0 and $t = 0$. Since $D_{\mathcal{J}}$ is a conservative gradient, it is nonempty valued, and graph closed, locally bounded hence compact valued and upper semicontinuous. It is furthermore convex valued by convexity of D since set-valued integration preserves convexity. We are in the setting of [34, Chapter 2, Section 7], and by Theorem 1 in [34, Chapter 2, Section 7], there exists $d > 0$ with $d < T$ such that (14) admits at least a solution on $[0, d]$. Let w with $w(0) = w_0$ be an arbitrary such solution. By Theorem 2 [34, Chapter 2, Section 7], w can be continued to the boundary of $\mathcal{D}$, i.e., to $t = T$ or $\|w_0 - w(t)\| = r$. Since $D_{\mathcal{J}}$ is a conservative gradient for $\mathcal{J}$, one has for all $t \in [0, d]$,

$$\left(\frac{1}{t}\int_0^t \|\dot{w}(s)\| \,\mathrm{d}s\right)^2 \leq \frac{1}{t}\int_0^t \|\dot{w}(s)\|^2 \,\mathrm{d}s = -\int_0^t \langle D_{\mathcal{J}}(w(s)), \dot{w}(s)\rangle \,\mathrm{d}s$$
$$= \mathcal{J}(w_0) - \mathcal{J}(w(t))$$

so that $\|w_0 - w(t)\| \leq \int_0^t \|\dot{w}(s)\| \,\mathrm{d}s \leq t\sqrt{(\mathcal{J}(w_0) - \mathcal{J}^*)}$. However $t\sqrt{(\mathcal{J}(w_0) - \mathcal{J}^*)} < r$, hence the solution w can be continued to $t = T$. Since w_0, T and w were arbitrary, the flow of (14) is nonempty and well defined on $\mathbb{R}_+$.

We define the *set-valued flow* Φ_t, given at all $w_0 \in \mathbb{R}^p$ and $t \in \mathbb{R}_+$ as

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

We also recall the definition of a Lyapunov function for a set-valued flow from [6]:

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

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

We define furthermore the critical set associated to $D_{\mathcal{J}}$, that is, $\operatorname{crit} D_{\mathcal{J}} := \{x \in \mathbb{R}^p \mid 0 \in D_{\mathcal{J}}(x)\}$, and show the following property:

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

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), $\dot{w}(u) \in -D_{\mathcal{J}}(w(u))$ 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 of $D_{\mathcal{J}}$, $\exists \epsilon > 0, \exists \delta > 0, \forall y \in \mathbb{R}^p$ 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 $\mathcal{J}(y) < \mathcal{J}(x)$. $\square$

Assumption 6 *Let P be a probability measure on $(S, \mathcal{A})$, consider $D : \mathbb{R}^p \times S \rightrightarrows \mathbb{R}^p$ such that for almost all $s \in S$, $f(\cdot, s)$ is locally Lipschitz continuous and $D(\cdot, s)$ is a convex-valued conservative gradient for $f(\cdot, s)$ and:*

1. *For all $w \in \mathbb{R}^p$, $f(w, \cdot)$ is integrable with respect to P.*
2. *$D : \mathbb{R}^p \times S \rightrightarrows \mathbb{R}^p$ is jointly measurable in $\mathbb{R}^p \times S$.*
3. *There exists a measurable function $\kappa : S \to \mathbb{R}_+$ such that for all $n \in \mathbb{N}$, κ^n is P-integrable . Also, there exists $q_0 \in \mathbb{N}$ such that for P-almost all $s \in S$, for all $w \in \mathbb{R}^p$, $\|D(w,s)\| \leq \kappa(s)(1 + \|w\|^{q_0})$, where $\|D(w,s)\| := \sup_{y \in D(w,s)} \|y\|$.*

Assumption 6 parallels assumptions of Theorem 4. In particular, Assumption 6 (3) implies (4) in Theorem 4. Consider furthermore the following assumption for the algorithmic recursion (2):

Assumption 7 *In addition to Assumption 6, the sequence $(\alpha_k)_{k \in \mathbb{N}}$ is strictly positive, $\sum_{k \in \mathbb{N}} \alpha_k = \infty$ and $\alpha_k \to 0$ as $k \to \infty$. v jointly measurable in $\mathcal{B}(\mathbb{R}^p) \times \mathcal{A}$ is such that for all $w \in \mathbb{R}^p$, for almost all $s \in S$, $v(w,s) \in D(w,s)$. The event $\sup_{k \in \mathbb{N}} \|w_k\| < \infty$ occurs almost surely.*

Lemma 4 (Noise extinction) *For the recursion (2) under Assumption 7, set for all $k \in \mathbb{N}$, $a_k = \mathbb{E}[v(w_k, \xi_k) | w_k]$ and $u_k = v(w_k, \xi_k) - a_k$, then*

1. *$\sup_{k \in \mathbb{N}} \mathbb{E}[\|u_k\|^2 | w_k]$ is finite almost surely.*
2. *If $\sum_{k \in \mathbb{N}} \alpha_k^2 < \infty$, then $\sum_{i=0}^k \alpha_i u_i$ converges almost surely as $k \to \infty$.*
3. *If $\phi \colon w \mapsto \sup_{s \in \operatorname{supp} P} \|v(w,s)\|$ is locally bounded and $\alpha_k = o(1/\log(k))$, then*

$$\forall T > 0, \qquad \lim_{k \to \infty} \sup_{m \geq k} \left\{ \left\| \sum_{i=k}^m \alpha_i u_i \right\|, \quad \text{s.t.} \sum_{i=k}^m \alpha_i \leq T \right\} = 0 \qquad a.s. \tag{16}$$

Proof. With these notations, (2) writes $w_{k+1} = w_k - \alpha_k(a_k + u_k)$ for all $k \in \mathbb{N}$, where we have by joint measurability of v, $a_k \in \mathbb{E}_{\xi\sim P}[D(w_k, \xi)] = D_{\mathcal{J}}(w_k)$.

As for the first statement, we have for all $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[\|u_k\|^2 \mid w_k\right] &= \mathbb{E}_{\xi\sim P}\left[\|v(w_k,\xi) - a_k\|^2\right] \leq \mathbb{E}_{\xi\sim P}\left[(\|v(w_k,\xi)\| + \|a_k\|)^2\right] \\ &\leq 4\mathbb{E}_{\xi\sim P}\left[\|D(w_k,\xi)\|^2\right] \leq 4\mathbb{E}_{\xi\sim P}\left[\kappa(\xi)^2\right](1 + \|w_k\|^{q_0})^2, \end{aligned} \tag{17}$$

so that $\sup_{k\in\mathbb{N}} \mathbb{E}[\|u_k\|^2|w_k] \leq 4\mathbb{E}_{\xi\sim P}\left[\kappa(\xi)^2\right](1 + R^{q_0})^2$ where $R := \sup_{k\in\mathbb{N}} \|w_k\|$ is almost surely finite by assumption. This proves the first statement.

Toward a proof of the second statement, set $\epsilon_k = \alpha_k u_k$ and $M_k = \sum_{i=0}^k \epsilon_i$ for all $k \in \mathbb{N}$. We will apply to $(M_k)_{k\in\mathbb{N}}$ a martingale convergence theorem. Indeed, for $k \in \mathbb{N}$, by independence of ξ_k we have $\mathbb{E}\left[u_k \mid w_k\right] = \mathbb{E}_{\xi\sim P}[v(w_k,\xi)] - a_k = 0$, hence $\mathbb{E}\left[\epsilon_k|w_k\right] = 0$, and $(M_k)_{k\in\mathbb{N}}$ is a martingale. We will verify that for all $k \in \mathbb{N}$, $\mathbb{E}[\|M_k\|^2] < \infty$ and $\sum_{k=0}^\infty \mathbb{E}\left[\|\epsilon_k\|^2 \mid w_k\right] < \infty$ almost surely so that [30, Theorem 4.5.2] provides the desired convergence almost surely.

We first show by induction that for all $k \in \mathbb{N}$, all the moments of w_k are finite. The constant w_0 is seen as a random variable whose moments are all finite. Suppose that for some $k \in \mathbb{N}$, the moments of w_k are finite. Let κ be as in Assumption 6 (3), we have

$$\begin{aligned}\|w_{k+1}\| &\leq \|w_k\| + \alpha_k\|D(w_k,\xi_k)\|] \\ &\leq \|w_k\| + \alpha_k\kappa(\xi_k)(1 + \|w_k\|^{q_0}). \end{aligned} \tag{18}$$

By the induction assumption, all the moments of w_k are finite, so does $\|w_k\|^{q_0}$. Furthermore, all the moments of $\kappa(\xi_k)$ are finite and $\kappa(\xi_k)$ and w_k are independent, hence the moments of $\kappa(\xi_k)(1 + \|w_k\|^{q_0})$ are finite. This shows by inequality (18) that w_{k+1} has all moments finite. This concludes the induction.

As a consequence, $\forall k \in \mathbb{N}$, $\|\epsilon_k\| \leq \alpha_k(\|a_k\| + \|D(w_k,\xi_k)\|) \leq \alpha_k\mathbb{E}_{\xi\sim P}[\kappa(\xi)](1 + \|w_k\|^{q_0}) + \kappa(\xi_k)(1 + \|w_k\|^{q_0})$ and the right side of the inequality has finite moments, so that ϵ_k has finite moments at all orders. Therefore, for all $k \in \mathbb{N}$, $M_k = \sum_{i=0}^k \epsilon_i$ is square integrable, as the finite sum of square integrable random variables.

Now, let us verify $\sum_{k=0}^\infty \mathbb{E}\left[\|\epsilon_k\|^2 \mid w_k\right] < \infty$ almost surely. Indeed, using (17), for all $k \in \mathbb{N}$, $\mathbb{E}\left[\|\epsilon_k\|^2 \mid w_k\right] \leq 4\alpha_k^2\mathbb{E}_{\xi\sim P}\left[\kappa(\xi)^2\right](1 + \|w_k\|^{q_0})^2$. Summing from $k = 0$ to $N \in \mathbb{N}$ gives $\sum_{k=0}^N \mathbb{E}[\|\epsilon_k\|^2 \mid w_k] \leq 4(1 + \|w_k\|^{q_0})^2 \sum_{k=0}^N \alpha_k^2 \leq 4(1 + R^{q_0}) \sum_{k=0}^\infty \alpha_k^2$ where we assumed $\sum_{k=0}^\infty \alpha_k^2 < \infty$ for the second statement and $R := \sup_{k\in\mathbb{N}} \|w_k\| < \infty$ almost surely by Assumption 7. Finally, almost surely, [30, Theorem 4.5.2] applies which proves the desired convergence.

For the third statement, let for all $k \in \mathbb{N}$, $c_k = 2\max\{1, \max_{i=0,\dots,k} \phi(w_k)\}$, almost surely increasing and convergent, such that $\|u_k/c_k\| \leq 1$ almost surely. Note that u_k/c_k are martingale increments: for $k \in \mathbb{N}$, u_k/c_k is integrable and $1/c_k$ is measurable with respect to $w_0, \dots, w_k$ hence $\mathbb{E}[u_k/c_k|w_0, \dots, w_k] = \mathbb{E}[u_k|w_0, \dots, w_k]/c_k = 0$, see [30, Theorem 4.1.14]. Fix $c > 0$, we have $2\log(k) = o(c/\alpha_k)$ so that as $k \to \infty$, $\exp(-c/\alpha_k)\, k^2 = \exp(-c/\alpha_k + 2\log(k)) \to 0$ and $\exp(-c/\alpha_k) = o(1/k^2)$ is summable. We invoke [5,

Proposition 4.4]:

$$\forall T > 0 \qquad \lim_{k\to\infty} \sup_m \left\{ \left\| \sum_{i=k}^{m} \alpha_i u_i / c_i \right\|, \quad \text{s.t.} \sum_{i=k}^{m} \alpha_i \leq T \right\} = 0 \qquad a.s.$$

Note that [5, Proposition 4.4] can be applied to any subgaussian martingale difference sequence, here we apply it to $(u_k/c_k)_{k\in\mathbb{N}}$ uniformly bounded by 1, hence subgaussian. Fix $T > 0$ and set for all k, m_k the largest integer $m \geq k$ such that $\sum_{i=k}^{m} \alpha_i \leq T$. We now have for all $k \leq m \leq m_k$

$$\begin{aligned} & \left\| \sum_{i=k}^{m} \frac{\alpha_i u_i}{c_i} - \frac{1}{c_k} \sum_{i=k}^{m} \alpha_i u_i \right\| = \left\| \sum_{i=k}^{m} \left(\frac{1}{c_i} - \frac{1}{c_k} \right) \alpha_i u_i \right\| \\ \leq \quad & \left(\frac{1}{c_k} - \frac{1}{c_{m_k}} \right) \sum_{i=k}^{m} \alpha_i c_i \leq \left(\frac{1}{c_k} - \frac{1}{c_{m_k}} \right) c_{m_k} \sum_{i=k}^{m_k} \alpha_i \leq \left(\frac{1}{c_k} - \frac{1}{c_{m_k}} \right) c_{m_k} T, \end{aligned}$$

and the result follows because c_k converges almost surely. □

As in [6], let us consider a Morse-Sard assumption in order to state the central contribution of this work in terms of stochastic approximation.

Assumption 8 (Morse-Sard) $\mathcal{J}(\operatorname{crit} D_{\mathcal{J}})$ has empty interior.

Theorem 5 (Convergence of the subgradient sampling method) *Let Assumption 7 holds, then almost surely all essential accumulation points $\bar{w}$ of $(w_k)_{k\in\mathbb{N}}$ satisfy $0 \in D_{\mathcal{J}}(\bar{w})$. If in addition Assumption 8 and Lemma 4 (2) or (3) hold, then, almost surely, $\mathcal{J}(w_k)$ converges as $k \to \infty$ and all accumulation points $\bar{w}$ of $(w_k)_{k\in\mathbb{N}}$ satisfy $0 \in D_{\mathcal{J}}(\bar{w})$. Furthermore, the set of accumulation points is connected.*

Proof. The first statement is a consequence of [11, Corollary 4.9] with $D_{\mathcal{J}}$. Indeed, Assumptions 4.1 and 4.3 of [11, Corollary 4.9] obviously hold, and Assumption 4.2 is the first point of Lemma 4. Under Assumption 6, Theorem 4 applies and $D_{\mathcal{J}}$ is a conservative gradient for $\mathcal{J}$. Lemma 3 applies and $\mathcal{J}$ is Lyapunov for $\operatorname{crit} D_{\mathcal{J}}$ and the result follows. Note that it would suffice to merely ensure that $\alpha_k \to 0$ to obtain this result.

Let us now prove the second statement. In case 2 or 3 of Lemma 4, identity (16) holds which is (i) in [6, Proposition 1.3], and (ii) also holds by Assumption 7. Let $w : \mathbb{R}_+ \to \mathbb{R}^p$ be the affine interpolation of $(w_k)_{k\in\mathbb{N}}$ (see [6, Definition IV] for its formal construction, this is a central object, also used for example in [7, 33, 17, 11]). [6, Proposition 1.3] ensures that w is a perturbed solution to (14) almost surely. By [6, Theorem 4.2] w satisfies [6, Theorem 4.1 (ii)] for (14) which implies by [6, Theorem 4.3] that the limit set of w is internally chain transitive with probability 1. Furthermore, $\mathcal{J}$ is a Lyapunov function for $\operatorname{crit} D_{\mathcal{J}}$ and the differential inclusion (14). By Assumption 8, $\mathcal{J}(\operatorname{crit} D_{\mathcal{J}})$ has empty interior. All the conditions are then satisfied to apply [6, Proposition 3.27], hence almost surely the limit set of w is contained in $\operatorname{crit} D_{\mathcal{J}}$ and $\mathcal{J}$ is constant on the limit set of w. We remark that the limit set of w is equal to the set of accumulation points of $(w_k)_{k\in\mathbb{N}}$, which gives the desired result.

If Lemma 4 (2) or (3) holds, then $\|w_{k+1} - w_k\| \to 0$ almost surely as $k \to \infty$ hence, in this case, accumulation points of $(w_k)_{k\in\mathbb{N}}$ is a connected subset of $\operatorname{crit} D_{\mathcal{J}}$. □

4 On the geometry of stochastic optimization problems

4.1 Definable sets and functions

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, for all $p \in \mathbb{N}$:

1. $\mathcal{O}_p$ is stable by finite intersection, 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. If $A \in \mathcal{O}_{p+1}$ then $\pi(A) \in \mathcal{O}_p$, where π projects on the p first coordinates,.
4. $\mathcal{O}_p$ contains all sets of the form $\{x \in \mathbb{R}^p : P(x) = 0\}$, where P is a polynomial.
5. The elements of $\mathcal{O}_1$ are exactly the finite unions of intervals.

Definition 8 (Definable set, function and set-valued map) If $\mathcal{O} = (\mathcal{O}_p)_{p\in\mathbb{N}}$ is an o-minimal structure, a subset A of $\mathbb{R}^n$ with $n \geq 1$ is said to be *definable in* $\mathcal{O}$ if $A \in \mathcal{O}_n$. A function $f : \mathbb{R}^p \to \mathbb{R}^q$ and a set-valued map $F : \mathbb{R}^p \rightrightarrows \mathbb{R}^q$ are called definable (in an o-minimal structure) if their graphs are definable, as sets of $\mathbb{R}^{p+q}$.

The definition allows in particular to prove definability through the use of first-order formula (see [25] for more details). We now present the main examples of o-minimal structures we use in this paper. The following definitions apply by extension to functions through their graph as above.

The class of semialgebraic sets is the smallest o-minimal structure.

Definition 9 (Semialgebraic sets) A subset $A \subset \mathbb{R}^n$ is *semialgebraic* if there exist polynomial functions P_{ij} and Q_{ij} with $i = 1, \ldots, l$ and $j = 1, \ldots, k$ such that $A = \bigcup_{i=1}^{l} \bigcap_{j=1}^{k} \{x \in \mathbb{R}^n \mid P_{ij}(x) < 0,\ Q_{ij}(x) = 0\}$.

Globally subanalytic sets, $\mathbb{R}_{\text{an}}$ An important o-minimal structure containing analytic functions is that of *globally subanalytic sets*, denoted $\mathbb{R}_{\text{an}}$. In order to define it, we first recall the definitions of semianalytic and subanalytic sets

Definition 10 (Semianalytic sets [57]) (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 $\bigcup_{i=1}^{l} \bigcap_{j=1}^{k} \{x \in \mathbb{R}^n \mid g_{ij}(x) < 0,\ h_{ij}(x) = 0$, where the g_{ij} and h_{ij} are real analytic. (ii) (Subanalyticity) A subset A of $\mathbb{R}^n$ is *subanalytic* if there exists $m \in \mathbb{N}$ such that A is the projection of a semianalytic set $M \subset \mathbb{R}^{n+m}$ on $\mathbb{R}^n$.

Definition 11 (Globally subanalytic sets) *$A \subset \mathbb{R}^n$ is globally subanalytic if $\tau_n(A)$ is subanalytic, where* $\tau_n : x \mapsto \left(\frac{x_1}{\sqrt{1+x_1^2}}, \ldots, \frac{x_n}{\sqrt{1+x_n^2}}\right)$.

O-minimal structure $\mathbb{R}_{\text{an,exp}}$ The exponential function, yet widely used in machine learning, is not globally subanalytic. It is however definable in an o-minimal structure called $\mathbb{R}_{\text{an,exp}}$ which also contains $\mathbb{R}_{\text{an}}$, see [28].

We now recall important properties of definable sets and functions.

Proposition 3 (Definable choice [25]) *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 12 (C^r-stratification [57]) Given $A \subset \mathbb{R}^p$ definable set, $(M_i)_{i=1,\dots,n}$ is a *C^r-stratification* if $(M_i)_{i=1,\dots,n}$ is a partition of A, and for all i, j in $\{1, \dots, r\}$, $\overline{M_i} \cap M_j \neq \emptyset$ implies that M_j is included in the boundary of M_i.

Definable sets admits a C^r-stratification for all $r \in \mathbb{N}$. Take for instance the graph of the absolute value $|\cdot|$, which is a semialgebraic set. It admits as a stratification $\{M_1, M_2, M_3\}$ where $M_1 = \{(x, -x) \mid x \in]-\infty, 0[\}$, $M_2 = \{(x, x) \mid x \in]0, +\infty[\}$, and $M_3 = \{(0, 0)\}$. This stratification property has several useful consequences for our work. For a definable function $F : \mathbb{R}^p \to \mathbb{R}$, one may consider a stratification $(M'_i)_{i=1,\dots n'}$ of $\mathbb{R}^p$ in which for all $i = 1, \dots, n'$, the restriction of F to M'_i is C^r. Taking the union of the manifolds M'_i of dimension p gives a dense open subset in $\mathbb{R}^p$ on which F is C^r. We use this property with $r = 2$, for instance in the proof of Claim 1. The stratification property also implies that definable dense sets have full measure. This is a property we repeatedly use in section 4.3.

4.2 Definability and set-valued integration

The following result is a set-valued version of [24, 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. Assume $D_{\mathcal{J}} : w \mapsto \int_{\mathbb{R}^m} D(w, s)\phi(s)\,\mathrm{d}s$ is well defined, then $D_{\mathcal{J}}$ is definable in $\mathbb{R}_{\text{an,exp}}$.*

Proof. Let D be as in the theorem. We will apply [24, Theorem 1.3] to prove the definability of $D_{\mathcal{J}}$ in $\mathbb{R}_{\text{an,exp}}$. D is definable in $\mathbb{R}_{\text{an}}$ hence the set-valued map $(w, q, s) \mapsto \langle D(w, s), q\rangle$ is definable in $\mathbb{R}_{\text{an}}$ as well. Let G be its graph. The function $H : (w, q, s) \in \mathbb{R}^p \times \mathbb{R}^p \times \mathbb{R}^m \mapsto \max_{v \in D(w,s)} \langle v, q\rangle$ 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 $w \in \mathbb{R}^p$

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

For $C \subset \mathbb{R}^p$ compact convex, define the support function $h_C : q \mapsto \max_{v \in C} \langle v, q \rangle$. Then by linearity of the integral, for $(w, q) \in \mathbb{R}^p \times \mathbb{R}^p$, it holds that

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

By [2, Theorem 18.19], there exists a measurable selection $\tilde{g} : \mathbb{R}^m \to \mathbb{R}^p$ 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., $h_{D_{\mathcal{J}}(w)}(q) = \int_{\mathbb{R}^m} H(w, q, s) \phi(s) \, \mathrm{d}s$. Furthermore, by duality, for all convex compact set $C \subset \mathbb{R}^p$ it holds that $C = \{z \in \mathbb{R}^p \mid \sup_{v \in \mathbb{R}^p} \langle v, z \rangle - h_C(v) = 0\}$. Applying this property with $C = D_{\mathcal{J}}(w)$ for each $w \in \mathbb{R}^p$ gives

$$\begin{aligned} \operatorname{Graph} D_{\mathcal{J}} &= \{(w, z) \in \mathbb{R}^p \times \mathbb{R}^p \mid z \in D_{\mathcal{J}}(w)\} \\ &= \left\{ (w, z) \in \mathbb{R}^p \times \mathbb{R}^p \;\middle|\; \sup_{q \in \mathbb{R}^p} \langle q, z \rangle - h_{D_{\mathcal{J}}(w)}(q) = 0 \right\} \\ &= \left\{ (w, z) \in \mathbb{R}^p \times \mathbb{R}^p \;\middle|\; \sup_{q \in \mathbb{R}^p} \langle q, z \rangle - \int_{\mathbb{R}^m} H(w, q, s) \phi(s) \, \mathrm{d}s = 0 \right\}. \end{aligned} \tag{19}$$

By [24, Theorem 1.3], $(w, q) \in \mathbb{R}^p \times \mathbb{R}^p \mapsto \int_{\mathbb{R}^m} H(w, q, s) \phi(s) \, \mathrm{d}s$ is definable in $\mathbb{R}_{\text{an,exp}}$, hence by equality (19) $D_{\mathcal{J}}$ is also definable in $\mathbb{R}_{\text{an,exp}}$. □

4.3 Consequences in stochastic optimization

Preliminary results Before providing our stochastic results, let us establish some technical lemmas. For a definable set $L \subset \mathbb{R}^p \times \mathbb{R}^m$, we define for all $(w, s) \in \mathbb{R}^p \times \mathbb{R}^m$, $L_w := \{s \in \mathbb{R}^m \mid (w, s) \in L\}$ and $L_s := \{w \in \mathbb{R}^p \mid (w, s) \in L\}$.

Lemma 5 *Let $(r, q) \in \mathbb{N}^* \times \mathbb{N}^*$, and $L \subset \mathbb{R}^r \times \mathbb{R}^q$ be a definable set. Then L is dense if and only if there is a dense definable set $Z \subset \mathbb{R}^r$ such that for all $w \in Z$, L_w is dense in $\mathbb{R}^q$.*

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

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

By definable choice, see Proposition 3, 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, see Definition 12, reducing U if needed, z and ϵ can be chosen continuous hence L^c has nonempty interior which contradicts the density of L.

As for the reverse implication, fix any $(\bar{w}, \bar{s}) \in \mathbb{R}^r \times \mathbb{R}^q$ and $\epsilon > 0$. Since Z is dense there is $w \in Z$ such that $\|w - \bar{w}\| < \epsilon/\sqrt{2}$. Since L_w is dense, there is $s \in L_w$ such that $\|s - \bar{s}\| < \epsilon/\sqrt{2}$. Overall, we have $(w, s) \in L$ such that $\|(w, s) - (\bar{w}, \bar{s})\| < \epsilon$ which shows that L is dense as $(\bar{w}, \bar{s}) \in \mathbb{R}^r \times \mathbb{R}^q$ and $\epsilon > 0$ were arbitrary. □

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 set $L \subset \mathbb{R}^p \times \mathbb{R}^m$, 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 $(\alpha, s) \in \Delta$, L_s is dense and open. Furthermore, denoting $\Phi_{\alpha,s} = \mathrm{Id} - \alpha \nabla_w g(\cdot, s)$ from L_s dense open to $\mathbb{R}^p$, we have*

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

Proof. Denote by L a definable dense open set such that g is C^2 on L (such sets exist by stratification, see Definition 12). Let $\lambda : L \subset \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^p$ be a definable representation of the eigenvalues of $\nabla_w^2 g$, where $\nabla_w^2 g$ 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 5 such that for all $s \in S_0$ the set L_s is open dense in $\mathbb{R}^p$.

Let F be the complement of the critical values of the function λ_i, for $i = 1, \ldots, p$ on L. The set of critical values F^c is finite by the definable Sard's theorem [13]. 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 definable choice, Proposition 3, there exists $\tilde{w} : E_{\alpha,i} \to \mathbb{R}^p$ definable, such that $\forall s \in E_{\alpha,i}$, $\alpha \lambda_i(\tilde{w}(s), s) = 1$ and $\nabla_w \lambda_i(\tilde{w}(s), s) = 0$. Assume for the sake of contradiction that there exists a nonempty open subset $U \subset E_{\alpha,i}$. By definability of $\tilde{w}$ and stratification, 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

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

hence we have for 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 $\alpha \in \Gamma$. 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 5 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

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

where I_p is the identity matrix of size p. Diagonalizing $\nabla_w^2 g(w, s)$, 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 $\dim Z \leq p-1$. Assume for the sake of contradiction that there exists a nonempty open set $V \subset \Phi^{-1}_{\alpha,s}(Z)$. The intersection $V \cap K^c_{\alpha,s}$ is open and nonempty because $K^c_{\alpha,s}$ is dense and both sets are open. Since $\Phi_{\alpha,s}$ is a local diffeomorphism on $K^c_{\alpha,s}$, the image $\Phi_{\alpha,s}(V \cap K^c_{\alpha,s})$ 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 5.

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 $\nabla_w g = v$ 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 $w \in Z$, $\nabla_w g(w,s) = v(w,s)$ for all s in a definable dense open set in $\mathbb{R}^m$.*

Theorem 7 (Genericity of gradient sequences) *Let $g\colon \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$ and $v : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}^p$ definable functions and $C \subset \mathbb{R}^p \times \mathbb{R}^m$ be a definable dense open set, such that for all $(w,s) \in C$, $\nabla_w g(w,s) = v(w,s)$. Given an arbitrary sequence $(s_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^m$ and an arbitrary $w_0 \in \mathbb{R}^p$, consider the recursion*

$$w_{k+1} = w_k - \alpha_k v(w_k, s_k) \text{ for all } k \in \mathbb{N}. \tag{20}$$

Then there is $\Gamma \subset \mathbb{R}$ which complement is finite such that if $\{\alpha_k\}_{k\in\mathbb{N}} \cap \Gamma = \emptyset$, then for each $k \in \mathbb{N}$, there exists a dense definable subset $\Sigma_k \subset \mathbb{R}^p \times (\mathbb{R}^m)^k$ such that for all $(w_0, s_0, \ldots, s_{k-1}) \in \Sigma_k$, it holds that $w_i \in R$, for all $i = 0, \ldots, k$, where $R \subset \mathbb{R}^p$ is the definable set such that for all $w \in R$,

- *$v(w,s) = \nabla_w g(w,s)$ for all s in a dense definable set.*
- *$g(\cdot, s)$ is C^2 in a neighborhood of w, for all s in a dense definable set.*

In particular there is a full measure residual set, $W \subset \mathbb{R}^p$ such that for each $w_0 \in W$ and $k \in \mathbb{N}$, for all $(s_0, \ldots, s_{k-1})$ in a definable dense set, $w_i \in R$, for all $i = 0, \ldots, k$

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 and open, and for every $\alpha \in \Gamma$, the definable set $\{s \in \mathbb{R}^m \mid (\alpha, s) \in \Delta\}$ is dense open in $\mathbb{R}^m$.

Recall that $L \cap C \subset \mathbb{R}^p \times \mathbb{R}^m$ is definable dense. By Lemma 5 there exists a definable dense set $\Sigma_0 \subset \mathbb{R}^p$ such that for all $w \in \Sigma_0$, the set $\{s \in \mathbb{R}^m | (w,s) \in L \cap C\}$ is dense. It satisfies the desired property, that is $\Sigma_0 \subset R$. Indeed for any $w \in \Sigma_0$, the set $\{s \in \mathbb{R}^m | (w,s) \in L \cap C\}$ is dense and for each such s, $(w,s) \in L \cap C$, that is g is C^2 at (w,s) and $v(w,s) = \nabla_w g(w,s)$ so that $w \in R$.

We set for all k, $\Delta_k = \{s \in \mathbb{R}^m \mid (\alpha_k, s) \in \Delta\}$. By Claim 1, for any $(\alpha, s) \in \Delta$, $\Phi_{\alpha,s} := \mathrm{Id}_p - \alpha\nabla_w g(\cdot, s)$ from $\{w \in \mathbb{R}^p \mid (w,s) \in L\}$, dense and open, to $\mathbb{R}^p$ verifies

$$\forall Z \subset \mathbb{R}^p \text{ definable, } \dim Z \leq p-1 \implies \dim \Phi^{-1}_{\alpha,s}(Z) \leq p-1. \tag{21}$$

Remark that $w_{k+1} = \Phi_{\alpha_k,s_k} \circ \ldots \circ \Phi_{\alpha_0,s_0}(w_0)$, as long as $s_i \in \Delta_i$ for $i = 0, \ldots, k$.

Let us proceed by induction, fix $k \in \mathbb{N}$ and assume that we have $\Sigma_k \subset R \times \Delta_0 \times \ldots \times \Delta_{k-1}$, definable dense, such that for all $(w_0, s_0, \ldots, s_{k-1}) \in \Sigma_k$, $w_i \in R$ for all $i = 0, \ldots, k$. Note that for $k = 0$, Σ_0 constructed above satisfies the desired hypothesis with the convention that the product set from 0 to -1 is empty.

Let us construct Σ_{k+1}. Remark that $s_i \in \Delta_i$ for $i = 0, \ldots, k$ so that $w_{k+1} = \Phi_{\alpha_k,s_k} \circ \ldots \circ \Phi_{\alpha_0,s_0}(w_0)$, as long as $(w_0, s_0, \ldots, s_k) \in \Sigma_k \times \Delta_k$.

Consider the set-valued map $N_k \colon s_k \rightrightarrows \Phi^{-1}_{\alpha_k,s_k}(R)$ and by backward recursion for $i = k-1, \ldots, 0$, $N_i \colon (s_i, \ldots, s_k) \rightrightarrows \Phi^{-1}_{\alpha_i,s_i}(N_{i+1}(s_{i+1}, \ldots, s_k))$. Set

$$\Sigma_{k+1} = \{(w, s_0, \ldots, s_k) \in \Sigma_k \times \Delta_k | w \in N_0(s_0, \ldots, s_k)\}.$$

Let us verify that Σ_{k+1} satisfies the desired properties. We have $\Sigma_{k+1} \subset \Sigma_k \times \Delta_k$ so that for any $(w_0, s_0, \ldots, s_k) \in \Sigma_{k+1}$, $(w_0, s_0, \ldots, s_{k-1}) \in \Sigma_k$ and $w_i \in R$ for $i = 0, \ldots, k$ by induction hypothesis. Furthermore, $s_i \in \Delta_i$ for all $i = 0, \ldots, k$ so that $w_{k+1} = \Phi_{\alpha_k,s_k} \circ \ldots \circ \Phi_{\alpha_0,s_0}(w_0)$. Note that by construction, $w \in N_0(s_0, \ldots, s_k)$ if and only if $\Phi_{\alpha_k,s_k} \circ \ldots \circ \Phi_{\alpha_0,s_0}(w_0) \in R$ which is the desired property. It remains to show that Σ_{k+1} is dense, and the induction will be complete. Note that $\Sigma_{k+1} = \Sigma_k \times \Delta_k \cap \{(w, s_0, \ldots, s_k) | ((s_0, \ldots, s_k), w) \in \mathrm{Graph}(N_0)\}$. Since $\Sigma_k \times \Delta_k$ is definable dense and $\mathrm{Graph}(N_0)$ is definable, it suffices to check that $\mathrm{Graph}(N_0)$ is dense. This is done by backward induction.

Let us first check that $\mathrm{Graph}(N_k)$ is dense. We have $(s_k, w_k) \notin \mathrm{Graph}(N_k)$ if and only if $(w_k, s_k) \notin L$ (Φ_{α_k,s_k} is not defined at w_k), or $w_k \in \Phi^{-1}_{\alpha_k,s_k}(R^c)$ so that $\mathrm{Graph}(N_k)^c = L^c \cup \{(s_k, w_k) | (w_k, s_k) \in L,\ w_k \in \Phi^{-1}_{\alpha_k,s_k}(R^c)\}$. Recall that R^c is definable and is the complement of a dense set, therefore it has at most dimension $p-1$ so that if $s_k \in \Delta_k$, $\Phi^{-1}_{\alpha_k,s_k}(R^c)$ also has dimension at most $p-1$. On the other hand, the set $\{w_k \in \mathbb{R}^p | (w_k, s_k) \in L^c\}$ is the complement of L_{s_k} (with the notation of Claim 1) definable and dense for $s_k \in \Delta_k$. Therefore for all $s_k \in \Delta_k$, the set $\{w_k \in \mathbb{R}^p | (s_k, w_k) \notin \mathrm{Graph}(N_k)\}$ has dimension at most $p-1$ and the set $\{w_k \in \mathbb{R}^p | (s_k, w_k) \in \mathrm{Graph}(N_k)\}$ is dense so that $\mathrm{Graph}(N_k)$ is dense by Lemma 5.

This extends by backward induction. Assume that $\mathrm{Graph}(N_i)$ is dense for some $i \in \{1, \ldots, k\}$. We have $(s_i, \ldots, s_k, w_i) \notin \mathrm{Graph}(N_i)$ if and only if $(w_i, s_i) \notin L$ (Φ_{α_i,s_i} is not defined at w_i) or $w_i \in \Phi^{-1}_{\alpha_i,s_i}(N_{i+1}(s_{i+1}, \ldots, s_k)^c)$. As N_{i+1} has a dense graph, by Lemma 5, for all $(s_{i+1}, \ldots, s_k)$ in a dense definable set R_i, $N_{i+1}(s_{i+1}, \ldots, s_k)$ is dense, and $N_{i+1}(s_{i+1}, \ldots, s_k)^c$ has dimension at most $p-1$. Therefore, similarly as above, for all $s_i \in \Delta_i$ and $(s_{i+1}, \ldots, s_k) \in R_i$, the set $\{w_i | (s_i, \ldots, s_k, w_i) \in \mathrm{Graph}(N_i)\}$ is dense and $\mathrm{Graph}(N_i)$ is dense. By induction, $\mathrm{Graph}(N_0)$ is dense and this shows that Σ_{k+1} has the correct property.

This proves the first statement. Now by Lemma 5, for each $k \in \mathbb{N}$, there is $W_k \subset \mathbb{R}^p$ definable dense such that for each $w_0 \in W_k$, for all $(s_0, \ldots, s_{k-1})$ in a dense definable set, $w_i \in R$ for all $i = 0, \ldots, k$. We set $W = \cap_{k \in \mathbb{N}} W_k$, W is a residual set by countable intersection of residual sets (with dense interior), and it has full measure as a countable intersection of full measure sets. □

Remark 6 *With the notation of Theorem 7, if $(s_i)_{i \in \mathbb{N}}$ are independant and identically*

distributed with a density with respect to Lebesgue measure, $w_0 \notin W$ and $\alpha_k \in \Gamma$ for all $k \in \mathbb{N}$, then almost surely, $w_k \in R$ for all $k \in \mathbb{N}$.

5 Proofs of Section 2

Proof of Lemma 1. Let us show that Theorem 4 applies. 1: measurability of f follows from semialgebraicity in Assumption 1 (3) and integrability follows from Assumption 1 (1). 2: these also entail that for all $s \in \mathbb{R}^m$, $f(\cdot, s)$ is locally Lipschitz and path-differentiable [15, Proposition 2] with conservative gradient $\partial^c_w f(\cdot, s)$. 3: by semialgebraicity of f, the set-valued map $\partial^c_w f$ is semialgebraic, hence jointly measurable. Toward 4, let $K \subset \mathbb{R}^p$ compact and $M > 0$ such that $K \subset B(0, M)$. Let κ be given by Assumption 1 (1), we have $\|\partial^c_w f(x, s)\| \leq \kappa(s)(1 + M^{q_0})$ for all $x \in K$, for almost all $s \in \mathbb{R}^m$ which implies 4 because κ is integrable.

By Theorem 4, $\mathbb{E}_{\xi\sim P}[\partial^c_w f(\cdot, \xi)]$ is a conservative gradient for $\mathcal{J}$ and $\mathcal{J}$ is path-differentiable and in particular admits a chain rule with respect to $\partial^c \mathcal{J}$. □

Proof of Theorem 1.

We first show 1 which is an application of the first part of Theorem 5. We need to verify that Assumption 1 is sufficient for Assumption 7 to hold. We let $D_{\mathcal{J}}(\cdot) = \mathbb{E}_{\xi\sim P}[\partial^c_w f(\cdot, \xi)]$ which is convex valued because integration in Definition 3 preserves convexity. Furthermore, for all $s \in \mathbb{R}^m$, $\partial^c_w f(\cdot, s)$ is a conservative gradient for $f(\cdot, s)$, because f is semialgebraic, hence path differentiable. We are therefore in the setting of Assumption 7. 1: measurability of f follows from semialgebraicity in Assumption 1 (3) and integrability follows from Assumption 1 (1). 2: by semialgebraicity of f, $(w, s) \mapsto \partial^c_w f(w, s)$ is also semialgebraic and therefore jointly measurable. Assumption 1 implies that for all s and all w, $f(\cdot, s)$ is $\kappa(s)(1+\|w\|^{q_0})$ Lipschitz for on $B(0, \|w\|)$, so that $\partial^c_w f(w, s)$ satisfies the same bound which is Assumption 6.3. The rest of Assumption 7 follows from Assumption 1 and Item 1 follows from Theorem 5.

We now turn to item 2. The arguments are the same, except that a smaller set-valued field drives the dynamics thanks to Theorem 7 and an interchanging of integral and derivative. As above, the function f is semialgebraic. Let $C := \{(w, s) \in \mathbb{R}^p \times \mathbb{R}^m \mid \partial^c_w f(w, s) = \{\nabla_w f(w, s)\}\}$. By Assumption 1, for almost all $s \in \mathbb{R}^m$, $\partial^c_w f(\cdot, s)$ is a conservative gradient for $f(\cdot, s)$, so that $\{w \in \mathbb{R}^p | (w, s) \in C\}$ is a full measure definable set [15, Theorem 1] hence dense definable. By Lemma 5, C is also dense. Since for all (w, s), $v(w, s) \in \partial^c f(w, s)$, for all $(w, s) \in C \subset \mathbb{R}^p \times \mathbb{R}^m$, $\nabla_w f(w, s) = v(w, s)$ and we can apply Theorem 7 and Remark 6. There is a full measure residual set $W \subset \mathbb{R}^p$, and $\Gamma \subset \mathbb{R}$ whose complement is finite such that if $w_0 \in W$ and $\alpha_k \in \Gamma$ for all $k \in \mathbb{N}$, almost surely, for all $k \in \mathbb{N}$, we have $v(w_k, s_k) = \nabla_w f(w_k, s_k)$ and $w_k \in R$, where $R \subset \mathbb{R}^p$ is a dense definable set such that for all $w \in R$:

- $v(w, s) = \nabla_w g(w, s)$ for all s in a dense definable set.
- $f(\cdot, s)$ is C^2 in a neighborhood of w, for all s in a dense definable set.

Fix $w \in R$ and let κ be as in Assumption 1. For all $a \in [-1, 1]$ and $h \in \mathbb{R}^p$, $|f(w +$

$ah, s) - f(w, s)| \leq \kappa(s)(1 + (\|w\| + \|h\|)^{q_0})\|h\|$ for P-almost all $s \in \mathbb{R}^m$. The right-hand side is integrable, and we may apply the dominated convergence theorem,

$$\begin{aligned}
\lim_{a\to 0} \frac{\mathcal{J}(w+ah) - \mathcal{J}(w)}{a} &= \lim_{a\to 0} \frac{1}{a} \int_{\mathbb{R}^m} f(w+ah, s) - f(w, s)\, \mathrm{d}P(s) \\
= \int_{\mathbb{R}^m} \lim_{a\to 0} \frac{f(w+ah, s) - f(w, s)}{a}\, \mathrm{d}P(s) &= \int_{\mathbb{R}^m} \langle \nabla_w f(w, s), h\rangle\, \mathrm{d}P(s) \\
= \left\langle \int_{\mathbb{R}^m} \nabla_w f(w, s)\, \mathrm{d}P(s), h \right\rangle,
\end{aligned}$$

where the limit under the integral is because for all s in a dense definable set (hence almost all s), $f(\cdot, s)$ is C^2 in a neighborhood of w. This shows, in particular, that $\mathcal{J}$ is differentiable at w.

We set $\tilde{D}_{\mathcal{J}} = \partial^c \mathcal{J}$, $\bar{v}$ a measurable selection in $\tilde{D}_{\mathcal{J}}$ such that $v(w) = \nabla \mathcal{J}(w)$ whenever $\mathcal{J}$ is differentiable at w. This means that for all w, there is $v_w \colon \mathbb{R}^m \to \mathbb{R}^p$ integrable such that $\int v_w(s) dP(s) = \bar{v}(w)$ and for all s, $v_w(s) = \nabla \mathcal{J}(w) \in \partial^c_w f(w, s)$. Set $\tilde{v}(w, s) = v_w(s)$ for all (w, s) and let L be as in Claim 1 definable, open and dense. We have that for all $(w, s) \in L$, $g(\cdot, s)$ is C^2 in a neighborhood of w so that $\tilde{v}(w, s) = \nabla_w f(w, s)$. Since $\tilde{v}$ agrees with $\nabla_w f(w, s)$ (definable, measurable) on the full measure set L (definable, dense), it is therefore jointly measurable [49, Proposition 3, Section 18.1] and satisfy for all w, $\int \tilde{v}(w, s) dP(s) \in \tilde{D}_{\mathcal{J}}(w)$.

Now, since for all $k \in \mathbb{N}$, $w_k \in R$ almost surely, it holds that almost surely, $\mathcal{J}$ is differentiable at w_k with $\nabla \mathcal{J}(w_k) = \mathbb{E}\left[\nabla_w f(w_k, \xi_k) | w_k\right] = \mathbb{E}\left[v(w_k, \xi_k) | w_k\right]$ and $v(w_k, s_k) = \nabla_w f(w_k, s_k) = \tilde{v}(w_k, s_k)$. We actually have for all $k \in \mathbb{N}$

$$\begin{aligned}
&w_{k+1} = w_k - \alpha_k \tilde{v}(w_k, s_k), \\
&\int_{\mathbb{R}^m} \tilde{v}(w_k, s)\, \mathrm{d}P(s) = \nabla \mathcal{J}(w_k), \text{ and} \qquad \int_{\mathbb{R}^m} \tilde{v}(w, s)\, \mathrm{d}P(s) \in \partial^c \mathcal{J}(w),\ \forall w \in \mathbb{R}^p
\end{aligned}$$

which is equivalent to the recursion given in Item 2. To obtain the desired convergence result we may apply Theorem 5 similarly as above, with $\tilde{D}_{\mathcal{J}} = \partial^c \mathcal{J}$ in place of $D_{\mathcal{J}}$ and $\tilde{v}$ in place of v and the result follows. □

Proof of Theorem 2. We have already shown that Assumption 1 is sufficient to ensure that Assumption 7 holds so that Theorem 5 applies. The claimed statement is indeed the second statement of Theorem 5 which holds Assumption 3. So we simply need to check that Assumption 8 holds.

As for the first statement, under Assumption 2, the function $\mathcal{J}$ is definable using [24, Theorem 1.3]. From Theorem 6 the set-valued map $D_{\mathcal{J}} = \mathbb{E}_{\xi\sim P}\left[\partial^c f(\cdot, \xi)\right]$ is definable and it is a conservative gradient for $\mathcal{J}$ by Lemma 1. By definable conservative Sard's theorem [15], the set of $D_{\mathcal{J}}$-critical values of $\mathcal{J}$ which is $\mathcal{J}(\operatorname{crit} D_{\mathcal{J}})$ is finite, hence Assumption 8 is satisfied. The result is then a consequence of the second statement of Theorem 5 under hypothesis 2 of Lemma 4.

Regarding the second statement, Theorem 1 gives us $\Gamma \subset \mathbb{R}$ whose complement is finite, $W \subset \mathbb{R}^p$ 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 we have with probability 1 the recursion holds with $\tilde{v} = \nabla_w f$

in place of v and $\tilde{D}_{\mathcal{J}} := \partial^c \mathcal{J}$ in place of $D_{\mathcal{J}}$. Hence, almost surely, one has an actual stochastic subgradient sequence so that the result follows as above from Theorem 5 applied to $\tilde{v}$ and $\tilde{D}_{\mathcal{J}}$ which are both semialgebraic. □

Proof of Theorem 3. We prove the regression case, i.e., when P has a compactly supported semialgebraic density ϕ with respect to the Lebesgue measure. The classification case uses similar arguments, see Remark 3 (2) We apply Theorem 1, which holds replacing the Clarke subgradient by a general conservative gradient D, see Remark 2 (2) and Theorem 5 similarly as above. We verify that in the setting of section 2.3, Assumptions 1, 2 and hypothesis 3 of Lemma 4 are satisfied. Under Assumption 4, let $D : \mathbb{R}^p \times \mathbb{R}^d \times \mathbb{R}^I \rightrightarrows \mathbb{R}^p$ be the product of Clarke Jacobians in (9), it is semialgebraic, for almost all $s \in \mathbb{R}^d \times \mathbb{R}^I$, $\mathrm{backprop}_w f(\cdot, s)$ is a semialgebraic selection of $D(\cdot, s)$, and $D(\cdot, s)$ is a conservative gradient for $f(\cdot, s)$. As D is semialgebraic, it is polynomially bounded hence there exist $K > 0$, $p_0 \in \mathbb{N}$, $q_0 \in \mathbb{N}$ such that for all $w \in \mathbb{R}^p$ and almost all $s \in \mathbb{R}^d \times \mathbb{R}^I$, $\|D(w, s)\| \leq K(1 + \|s\|^{p_0})(1 + \|w\|^{q_0})$. Let $\kappa(s) := K(1 + \|s\|^{p_0})$ for $s \in \mathbb{R}^d \times \mathbb{R}^I$. Since P has compact support, then for all $n \in \mathbb{N}$, the function κ^n is integrable with respect to P. Assumption 1 (1) is satisfied. f and $\mathrm{backprop}_w$ are semialgebraic by assumption, hence Assumption 1 (3) is satisfied. The other assumptions are directly satisfied by assumptions of Theorem 3. Note that for any w, $\sup_{s \in \mathrm{supp}\, P} \|\, \mathrm{backprop}_w f(w, s)\| \leq \tilde{K}(1 + \|w\|^{q_0})$ which is locally bounded so that we may apply Theorem 5 under the hypothesis 3. of Lemma 4 similarly as in the proof of Theorem 2.

□

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 [40, 41, 42].

Definition 13 (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 are defined in Definition 5. In both cases, the corresponding set-valued gradient map is a singleton almost everywhere and contains the Clarke subgradient of f everywhere [40, 15]. Functions with generalized gradient are called *differentiable in the generalized sense*, and those with conservative gradients are called path-differentiable.

6.2 Relations between the two notions

The following strengthens the chain rule along semismooth curves given in [51, Theorem 1].

Proposition 4 *If D is a generalized gradient of f in the sense of Definition 13, then it is a conservative gradient of f.*

Proof. We shall use the chain rule characterization of conservative gradients, Definition 5. Let D be a generalized gradient of f as in Definition 13, 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 measure 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 [15]), $g\colon [0,1] \to \mathbb{R}^p$, defined for all $t \in R$ by $g(t) = \arg\max_{v \in D(\gamma(t))} \langle \dot\gamma(t), v\rangle$ 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., [49, Section 3.3]) and 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., [32, Theorem 1.35]) ensures that almost all $t \in E_3$ have density 1, that is, $\lambda([t-\delta, t+\delta] \cap E_3)/\lambda([t-\delta, t+\delta]) \to 1$ as $\delta \to 0$. 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

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

where the two identities on the first line follow from the differentiability of $f \circ \gamma$ and γ at $\bar{t} \in E_3$, the third one stems from the semismooth property of generalized gradients (Definition 13) 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. □

Functions differentiable in the generalized sense are path-differentiable. In the semialgebraic case, both notions coincide [26], but the inclusion is strict in general.

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 F to C is path-differentiable but not differentiable in the generalized sense.*

Proof. First let us recall a substitution formula for absolutely continuous function [55, 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 concluding 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. □

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. Benaïm, *Dynamics of stochastic approximation algorithms*, in Seminaire de probabilites XXXIII, Springer, 1999, pp. 1–68.

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

[7] 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.

[8] D. Bertoin, J. Bolte, S. Gerchinovitz, and E. Pauwels, *Numerical influence of relu'(0) on backpropagation*, in Advances in Neural Information Processing Systems, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds., vol. 34, Curran Associates, Inc., 2021, pp. 468–479.

[9] 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.

[10] P. Bianchi, W. Hachem, and S. Schechtman, *Convergence of constant step stochastic gradient descent for non-smooth non-convex functions*, Set-Valued and Variational Analysis, (2022), https://doi.org/10.1007/s11228-022-00638-z.

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

[12] 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).

[13] 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.

[14] 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).

[15] J. Bolte and E. Pauwels, *Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning*, Mathematical Programming, (2020), https://doi.org/10.1007/s10107-020-01501-5.

[16] J. Bolte and E. Pauwels, *A mathematical model for automatic differentiation in machine learning*, in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 10809–10819.

[17] 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).

[18] V. Borkar, *Stochastic approximation: a dynamical systems viewpoint*, vol. 48, 2009.

[19] 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.

[20] 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.

[21] 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.

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

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

[24] 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.

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

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

[27] 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.

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

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

[30] R. Durrett, *Probability: Theory and Examples*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2010, https://books.google.fr/books?id=evbGTPhuvSoC.

[31] 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.

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

[33] M. Faure and G. Roth, *Ergodic properties of weak asymptotic pseudotrajectories for set-valued dynamical systems*, Stochastics and Dynamics, 13 (2013), p. 1250011.

[34] A. F. Filippov, *Differential equations with discontinuous righthand sides*, in Mathematics and Its Applications, 1988.

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

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

[37] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444, https://doi.org/10.1038/nature14539, https://doi.org/10.1038/nature14539.

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

[39] S. Marx and E. Pauwels, *Path differentiability of ode flows*, Journal of Differential Equations, 338 (2022), pp. 321–351, https://doi.org/10.1016/j.jde.2022.07.038.

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

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] E. Nurminski, *Minimization of nondifferentiable functions in the presence of noise*, Cybernetics and Systems Analysis - CYBERN SYST ANAL-ENGL TR, 10 (1974), pp. 619–621, https://doi.org/10.1007/BF01071541.

[46] 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.

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

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

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

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

[51] 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.

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

[53] D. Sahoo, Q. Pham, J. Lu, and S. C. H. Hoi, *Online deep learning: Learning deep neural networks on the fly*, in IJCAI, 2018.

[54] S. Schechtman, *Stochastic proximal subgradient descent oscillates in the vicinity of its accumulation set*, Optimization Letters, (2022), https://doi.org/10.1007/s11590-022-01884-8.

[55] 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.

[56] 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.

[57] 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.