Convergence of constant step stochastic gradient descent for non-smooth non-convex functions

Pascal Bianchi, Walid Hachem, Sholom Schechtman

Convergence of constant step stochastic gradient descent for non-smooth non-convex functions

Pascal Bianchi[1], Walid Hachem[2], and Sholom Schechtman[2]

[1]LTCI, Telecom Paris, IP Paris, France.
[2]LIGM, CNRS, Univ Gustave Eiffel, ESIEE Paris, F-77454 Marne-la-Vallée, France.

January 2021

Abstract

This paper studies the asymptotic behavior of the constant step Stochastic Gradient Descent for the minimization of an unknown function, defined as the expectation of a non convex, non smooth, locally Lipschitz random function. As the gradient may not exist, it is replaced by a certain operator: a reasonable choice is to use an element of the Clarke subdifferential of the random function; another choice is the output of the celebrated backpropagation algorithm, which is popular amongst practioners, and whose properties have recently been studied by Bolte and Pauwels. Since the expectation of the chosen operator is not in general an element of the Clarke subdifferential of the mean function, it has been assumed in the literature that an oracle of the Clarke subdifferential of the mean function is available. As a first result, it is shown in this paper that such an oracle is not needed for almost all initialization points of the algorithm. Next, in the small step size regime, it is shown that the interpolated trajectory of the algorithm converges in probability (in the compact convergence sense) towards the set of solutions of a particular differential inclusion: the subgradient flow. Finally, viewing the iterates as a Markov chain whose transition kernel is indexed by the step size, it is shown that the invariant distribution of the kernel converge weakly to the set of invariant distribution of this differential inclusion as the step size tends to zero. These results show that when the step size is small, with large probability, the iterates eventually lie in a neighborhood of the critical points of the mean function.

Keywords: Clarke subdifferential, Backpropagation algorithm, Differential inclusions, Non convex and non smooth optimization, Stochastic approximation.

1 Introduction

In this work, we study the asymptotic behavior of the constant step Stochastic Gradient Descent (SGD) when the objective function is neither differentiable nor convex. Given an integer $d \geqslant 1$ and a probability space $(\Xi, \mathscr{T}, \mu)$, let $f : \mathbb{R}^d \times \Xi \to \mathbb{R}, (x, s) \mapsto f(x, s)$ be a function which is assumed to be locally Lipschitz, generally non-differentiable and non-convex in the variable x, and μ-integrable in the variable s. The goal is to find a local minimum, or at least a critical point of the function $F(x) = \int f(x, s)\, \mu(ds) = \mathbb{E} f(x, \cdot)$, *i.e.*, a point $x_\star$ such that $0 \in \partial F(x_\star)$, where ∂F is the so-called Clarke subdifferential of F. It is assumed that the

function f is available to the observer along with a sequence of independent Ξ-valued random variables $(\xi_k)_{k\in\mathbb{N}}$ on some probability space with the same probability law μ. The function F itself is assumed unknown due to, *e.g.*, the difficulty of computing the integral $\mathbb{E}f(x,\cdot)$. Such non-smooth and non-convex problems are frequently encountered in the field of statistical learning. For instance this type of problem arises in the study of neural networks when the activation function is non-smooth, which is the case of the commonly used ReLU function.

We establish the weak convergence of SGD to the set of (Clarke) critical points of F. Our main contributions are:

- We investigate the constant step size regime, whereas most works address the vanishing step size regime.
- We study an *oracle-free* version of SGD, which does not require to have access to the Clarke subgradient of the unknown function F.

To that end, our main hypotheses is that the function F is *Whitney stratifiable*. We also need to posit that the sequence of iterates is *bounded in probability*. Boundedness assumptions are quite standard in stochastic approximation, we nevertheless provide sufficient conditions: first, it holds when F is assumed coercive and smooth outside an arbitrary compact set; second, it naturally holds in the case of *projected SGD i.e.*, when the iterates are projected onto some compact set. The convergence of the projected SGD is as well addressed in the paper.

We say that a sequence of random variables $(x_n)_{n\in\mathbb{N}}$ on $\mathbb{R}^d$ is a *SGD sequence* with step size $\gamma > 0$ if, with probability one,

$$x_{n+1} = x_n - \gamma \nabla f(x_n, \xi_{n+1}) \tag{1}$$

for every n such that the function $f(\cdot, \xi_{n+1})$ is differentiable at point x_n, where $\nabla f(x_n, \xi_{n+1})$ represents the gradient w.r.t. the variable x_n. When $f(\cdot, \xi_{n+1})$ is non-differentiable at x_n, the update equation $x_n \to x_{n+1}$ is left undefined. The practioner is free to choose the value of x_{n+1} according to a predetermined selection policy. Typically, a reasonable choice is to select x_{n+1} in the set $x_n - \gamma \partial f(x_n, \xi_{n+1})$, where $\partial f(x, s)$ represents the Clarke subdifferential of the function $f(\cdot, s)$ at the point x. When such a policy is used, the resulting sequence will be referred to as a *Clarke-SGD* sequence. In fact, our study extends to the case where x_{n+1} is chosen in the set $x_n - \gamma G_{f(\cdot,\xi_{n+1})}$, where $G_{f(\cdot,\xi_{n+1})}$ is a generalized subdifferential of $f(\cdot, \xi_{n+1})$ in Norkin's sense [23] (we refer to such a sequence as a *Norkin-SGD* sequence). The Clarke subdifferential is a special case of generalized subdifferential.

An alternative used by practioners is to compute the derivative using the automatic differentiation provided in popular API's such as Tensorflow, PyTorch, etc. *i.e.*, for all n,

$$x_{n+1} = x_n - \gamma a_{f(\cdot,\xi_{n+1})}(x_n) \tag{2}$$

where a_h stands for the output of the automatic differentiation applied to a function h. We refer to such a sequence as an *autograd* sequence. This approach is useful when $f(\cdot, s)$ is a composition of matrix multiplications and non-linear activation functions, of the form

$$f(x, s) = \ell(\sigma_L(W_L \sigma_{L-1}(W_{L-1} \cdots \sigma_1(W_1 X_s))), Y_s)\,, \tag{3}$$

where $x = (W_1, \cdots, W_L)$ are the weights of the network represented by a finite sequence of L matrices, $\sigma_1, \cdots, \sigma_L$ are vector-valued functions, X_s is a feature vector, Y_s is a label and

$\ell(\cdot,\cdot)$ is some loss function. In such a case, the automatic differentiation is computed using the chain rule of function differentiation, by means of the celebrated backpropagation algorithm. When the mappings $\sigma_1, \cdots, \sigma_L, \ell(\cdot, Y_s)$ are differentiable, the chain rule indeed applies and the output coincides with the gradient. However, the chain rule fails in case of non-differentiable functions. The properties of the map a_h are studied in the recent work [8]. In general, $a_h(x)$ may not be an element of the Clarke-subdifferential $\partial h(x)$. It can even happen that $a_h(x) \neq \nabla h(x)$ at some points x where h is differentiable. However, the set of such peculiar points is proved to be Lebesgue negligible. As a consequence, if the initial point x_0 is chosen random according to some density w.r.t. the Lebesgue measure, an autograd sequence can be shown to be a SGD sequence in the sense of Equation (1) under some conditions. The aim of this paper is to analyze the asymptotic behavior of SGD sequences in the case where the step γ is constant.

About the literature. In the nonsmooth and non convex case, the convergence of SGD has been studied in [11] and [12] using the concept of generalized differentiability [23], and assuming a Sard-like condition on the critical set. More recently, using a differential inclusion (DI) approach, the papers [10] provide a similar result under the additional assumption that the objective function is Whitney-stratifiable (see also [20], in the particular case of subdifferentially regular functions). These papers make two major hypotheses on the algorithm under study, which we avoid in this paper.

The first major hypothesis in the above papers if the fact that the step size is vanishing, *i.e.*, γ is replaced with a sequence (γ_n) that tends to zero as $n \to \infty$. From a theoretical point of view, the vanishing step size is convenient because, under various assumptions, it allows to demonstrate the almost sure convergence of the iterates x_n to the set

$$\mathcal{Z} := \{x \in \mathbb{R}^d \,:\, 0 \in \partial F(x)\} \tag{4}$$

of critical points of F. However, in practical applications such as neural nets, a vanishing step size is rarely used because of slow convergence issues. In most computational frameworks, a possibly small but nevertheless constant step size is used by default. The price to pay is that the iterates are no longer expected to converge almost surely to the set $\mathcal{Z}$ but to fluctuate in the vicinity of $\mathcal{Z}$ as n is large. In this paper, we aim at establishing a result of the type

$$\forall \varepsilon > 0, \quad \limsup_{n\to\infty} \mathbb{P}(\boldsymbol{d}(x_n, \mathcal{Z}) > \varepsilon) \underset{\gamma \downarrow 0}{\longrightarrow} 0, \tag{5}$$

where $\boldsymbol{d}$ is the Euclidean distance between x_n and the set $\mathcal{Z}$. Although this result is weaker than in the vanishing step case, constant step stochastic algorithms can reach a neighborhood of $\mathcal{Z}$ faster than their decreasing step analogues, which is an important advantage in the applications where the accuracy of the estimates is not essential. Moreover, in practice they are able to cope with non stationary or slowly changing environments which are frequently encountered in signal processing, and possibly track a changing set of solutions [5, 18].

The second important difference between the present paper and the papers [20, 10] lies in the algorithm under study. In these papers, the iterates are supposed to satisfy the inclusion

$$\frac{x_{n+1} - x_n}{\gamma_{n+1}} \in -\partial F(x_n) + \eta_{n+1} \tag{6}$$

for all n, where (η_n) is a martingale increment noise w.r.t. the filtration $(\sigma(x_0, \xi_1, \ldots, \xi_n))_{n\geqslant 1}$. Under the assumption that $\gamma_n \to 0$ as $n \to \infty$, the authors of [20, 10] prove that almost surely,

the continuous time linearly interpolated process constructed from a sequence (x_n) satisfying (6) is a so-called asymptotic pseudotrajectory [4] of the Differential Inclusion (DI)

$$\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t))\,, \tag{7}$$

that will be defined on $\mathbb{R}_+ = [0, \infty)$. Heuristically, this means that a sequence (x_n) satisfying (6) shadows a solution to (7) as n tends to infinity. This result is one of the key ingredients to establish the almost sure convergence of x_n to the set $\mathcal{Z}$. Unfortunately, a SGD sequence does not satisfy the condition (6) in general (setting apart the fact that γ is constant). To be more precise, consider a Clarke-SGD sequence as defined above. For all n, $x_{n+1} = x_n - \gamma \partial f(x_n, \xi_{n+1})$, which in turn implies

$$\frac{x_{n+1} - x_n}{\gamma} \in -\mathbb{E}\partial f(x_n, \,.\,) + \eta_{n+1}\,,$$

where (η_n) is a martingale increment noise sequence, and where $\mathbb{E}\partial f(x, \,.\,)$ represents the set-valued expectation $\int \partial f(x, s) d\mu(s)$. The above inclusion is analogous to (6) in the case where $\partial F(x) = \mathbb{E}\partial f(x, \cdot)$ for all x *i.e.*, if one can interchange the expectation $\mathbb{E}$ and the Clarke subdifferential operator ∂. Although the interchange holds if *e.g.*, the functions $f(\cdot, s)$ are convex (in which case $\partial f(x, s)$ would coincide with the classical convex subdifferential), one has in general $\partial \mathbb{E} f(x, \cdot) \subset \mathbb{E}\partial f(x, \cdot)$ and the inclusion can be strict [9, Proposition 2.2.2]. As a consequence, a Clarke-SGD sequence does not admit the oracle form (6) in general. For such a sequence, the corresponding DI reads

$$\dot{\mathsf{x}}(t) \in -\mathbb{E}\partial f(\mathsf{x}(t), \,.\,)\,, \tag{8}$$

but unfortunately, the flow of this DI may contain spurious equilibria (an example is provided in the paper). In [20] the authors restrict their analysis to *regular* functions [9, §2.4], for which the interchange of the expectation and the subdifferentiation applies. However, this assumption can be restrictive, since a function as simple as $-|x|$ is not regular at the critical point zero. The issue of the absence of interchange between the expectation and the Clarke subdifferential was addressed in [12] using the notion of generalized differentiability. In this work, the convergence is established towards the set of zeroes of the generalized subdifferential of F. However, this set can be substantially larger than the set $\mathcal{Z}$ of critical points.

A second example where the oracle form of Equation (6) does not hold is given by autograd sequences. Such an example is studied in [8], assuming that the step size is vanishing and that ξ takes its values over a finite set. It is proved that the autograd sequence is an almost sure asymptotic pseudotrajectory of the DI $\dot{\mathsf{x}}(t) \in -D(\mathsf{x}(t))$, for some set-valued map D which is shown to be a *conservative* field with F as a potential. Properties of conservative fields are studied in [8]. In particular, it is proved that $D = \{\nabla f\}$ Lebesgue almost everywhere. Despite this property, the DI $\dot{\mathsf{x}}(t) \in -D(\mathsf{x}(t))$ substantially differs from (7). Again, the set of equilibria may be strictly larger than the set $\mathcal{Z}$ of critical points of F.

We finally mention the paper [26], which studies an inertial version of SGD in the vanishing step size regime. Similarly to [20, 10] and contrary to the present paper, the author assumes the oracle form of Equation (6). The almost sure convergence is established, under the rather weak assumption that F is differentiable in Norkin's generalized sense.

Contributions

- We analyze the SGD algorithm (1) in the non-smooth, non-convex setting, under realistic assumptions: the step size is assumed to be constant along the iterations, and

we neither assume the regularity of the functions involved, nor the knowledge of an oracle of ∂F as in (6). Our assumptions encompass Clarke SGD sequences, autograd and Norkin SGD sequences as special cases.

- Under mild conditions, we prove that when the initialization x_0 is randomly chosen with a density, all SGD sequences coincide almost surely, irrespective to the particular selection policy used at the points of non-differentiability. In this case, x_n almost never hits a non-differentiable point of $f(\cdot, \xi_{n+1})$ and Equation (1) actually holds for all n. Moreover, we prove that

$$\frac{x_{n+1} - x_n}{\gamma} = -\nabla F(x_n) + \eta_{n+1}\,,$$

  where (η_n) is a martingale difference sequence, and $\nabla F(x_n)$ is the true gradient of F at x_n. This argument allows to bypass the oracle assumption of [20, 10].

- We establish that the continuous process obtained by piecewise affine interpolation of (x_n) is a *weak asymptotic pseudotrajectory* of the DI (7). In other words, the interpolated process converges in probability to the set of solutions to the DI, as $\gamma \to 0$, for the metric of uniform convergence on compact intervals.

- We establish the long run convergence of the iterates x_n to the set $\mathcal{Z}$ of Clarke critical points of F, in the sense of Equation (5). This result holds under two main assumptions. First, it assumed that F admits a chain rule, which is satisfied for instance if F is a so-called tame function. Second, we assume a standard drift condition on the Markov chain (1). Finally, we provide verifiable conditions of the functions $f(\cdot, s)$ under which the drift condition holds.

- In many practical situations, the drift conditions alluded to above are not satisfied. To circumvent this issue, we analyze a projected version of the SGD algorithm, which is similar in its principle to the well-known projected gradient algorithm in the classical stochastic approximation theory.

Paper organization

Section 2 recalls some known facts about Clarke subdifferentials, conservative fields and differential inclusions. In Section 3, we study the elementary properties of almost-everywhere gradient functions, defined as the functions $\varphi(x, s)$ which coincide with $\nabla f(x, s)$ almost everywhere. Practical examples are provided. In Section 4, we study the elementary properties of SGD sequences. Section 5 establishes the convergence in probability of the interpolated process to the set of solutions to the DI. In Section 6, we establish the long run convergence of the iterates to the set of Clarke critical points. Section 7 is devoted to the projected subgradient algorithm. The proofs are found in Section 8.

2 Preliminaries

2.1 Notations

If ν, ν' are two measures on some measurable space $(\Omega, \mathcal{F})$, $\nu \ll \nu'$ means that ν is absolutely continuous w.r.t. ν. The ν-completion of $\mathscr{F}$ is defined as the sigma-algebra consisting of the

sets $S \subset \Omega$ such that there exist $A, B \in \mathscr{F}$ with $A \subset S \subset B$ and $\nu(B \backslash A) = 0$. For these sets, $\nu(S) = \nu(A)$.

If E is a metric space, we denote by $\mathscr{B}(E)$ the Borel sigma field on E. Let d be an integer. We denote by $\mathcal{M}(\mathbb{R}^d)$ the set of probability measures on $\mathscr{B}(\mathbb{R}^d)$ and by $\mathcal{M}_1(\mathbb{R}^d) := \{\nu \in \mathcal{M}(\mathbb{R}^d) : \int \|x\| \nu(dx) < \infty\}$. We denote as λ^d the Lebesgue measure on $\mathbb{R}^d$. When the dimension is clear from the context, we denote as λ this Lebesgue measure. For a subset $\mathcal{K} \subset \mathbb{R}^d$, we denote by

$$\mathcal{M}_{abs}(\mathcal{K}) := \{\nu \in \mathcal{M}(\mathbb{R}^d) : \nu \ll \lambda \text{ and } \mathrm{supp}(\nu) \subset \mathcal{K}\},$$

where $\mathrm{supp}(\nu)$ represents the support of ν.

If P is a Markov kernel on $\mathbb{R}^d$ and $g : \mathbb{R}^d \to \mathbb{R}$ is a measurable function, Pg represents the function on $\mathbb{R}^d \to \mathbb{R}$ given by $Pg(x) = \int P(x, \mathrm{d}y) g(y)$, whenever the integral is well-defined (the integral is understood in the weak sense). For every measure $\pi \in \mathcal{M}(\mathbb{R}^d)$, we denote by πP the measure given by $\pi P = \int \pi(\mathrm{d}x) P(x, \cdot)$. We use the notation $\pi(g) = \int g \, \mathrm{d}\pi$ whenever the integral is well-defined.

For every $x \in \mathbb{R}^d$, $r > 0$, $B(x, r)$ is the open Euclidean ball with center x and radius r. The notation 1_A stands for the indicator function of a set A, equal to one on that set and to zero otherwise. The notation A^c represents the complementary set of a set A and $\mathrm{cl}(A)$ its closure.

2.2 Subdifferentials and Conservative Fields

A set valued map $H : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a map such that for each $x \in \mathbb{R}^d$, $H(x)$ is a subset of $\mathbb{R}^d$. We say that H is upper semi continuous, if its graph $\{(x, y) : y \in H(x)\}$ is closed in $\mathbb{R}^{d \times d}$. For any function $F : \mathbb{R}^d \to \mathbb{R}$, we denote by $\mathcal{D}_F$ the set of points $x \in \mathbb{R}^d$ such that F is differentiable at x. If F is locally Lipschitz continuous, it is by Rademacher's theorem almost everywhere differentiable. In this case, the Clarke's subdifferential of F coincides with the set-valued map $\partial F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ given for all $x \in \mathbb{R}^d$ by

$$\partial F(x) = \mathrm{co}\left\{ y \in \mathbb{R}^d \, : \, \exists (x_n)_{n \in \mathbb{N}} \in \mathcal{D}_F^{\mathbb{N}} \text{ s.t. } (x_n, \nabla F(x_n)) \to (x, y) \right\},$$

where co stands for the convex hull [9].

We now briefly review some recent results of [8]. A set-valued map $D : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is called a *conservative field*, if for each $x \in \mathbb{R}^d$, $D(x)$ is a nonempty and compact subset of $\mathbb{R}^d$, D has a closed graph, and for each absolutely continuous $a \colon [0, 1] \to \mathbb{R}^d$, with $a(0) = a(1)$, it holds that:

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

We say that a function $F : \mathbb{R}^d \to \mathbb{R}$ is a *potential* for the conservative field D if for every $x \in \mathbb{R}^d$ and every absolutely continuous $a \colon [0, 1] \to \mathbb{R}^d$, with $a(0) = 0$ and $a(1) = x$,

$$F(x) = F(0) + \int_0^1 \min_{v \in D(a(t))} \langle \dot{a}(t), v \rangle \, \mathrm{d}t \,. \tag{9}$$

In this case, such a function F is locally Lipschitz continuous, and for every absolutely continuous curve $a : [0, 1] \to \mathbb{R}^d$, the function $t \mapsto F(a(t))$ satisfies for almost every $t \in [0, 1]$,

$$\frac{d}{dt} F(a(t)) = \langle v, \dot{a}(t) \rangle \qquad (\forall v \in D(a(t))) \,,$$

that is to say, F admits a "chain rule" [8, Lemma 2]. Moreover, by [8, Theorem 1], it holds that $D = \{\nabla F\}$ Lebesgue almost everywhere.

We say that a function F is *path differentiable* if there exists a conservative field D such that F is a potential for D. If F is path differentiable, then the Clarke subdifferential ∂F is a conservative field for the potential F [8, Corollary 2]. Another useful example of a conservative field for composite functions is the automatic differentiation field [8, Section 5]. A broad class of functions used in optimization are path differentiable, e.g. any convex, concave, regular or tame. A tame function is a function defined in some o-minimal structure ([27]), they enjoy some nice stability properties such as any elementary operation on them remain tame (e.g. composition, sum, inverse). The domain f of a tame function admits a so-called Whitney stratification, that is to say a collection of manifolds (S_i) on each of which f is smooth with the additional property that the various gradients fit well together (see [7] for more details). The exponential and the logarithm are tame, as well as any semialgebraic function, an interested reader can find more on tameness and its usefulness in optimization in [16], and more details in [27], [7] and [10].

A similar point of view on differentiation of non-smooth functions is given by the generalized subdifferential introduced by Norkin [23]. A function $F : \mathbb{R}^d \to \mathbb{R}$ is said to be differentiable in a generalized sense if there is a set-valued map $G_F : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ such that for every x, $G_F(x)$ is nonempty, convex, compact valued, the graph of G_F is closed, and

$$F(y) = F(x) + \langle g(y), y - x\rangle + o(x, y, g)\,, \quad \text{with } g(y) \in G_F(y) \text{ and } \lim_{y\to x} \sup_{g\in G_F(y)} \frac{o(x,y,g)}{\|x-y\|} = 0\,.$$

As in the path-differentiable case, the class of such functions contains tame, regular and Whitney stratifiable functions. A nice feature of this class is that, under mild conditions, it is closed with respect to the expectation. That is to say, if $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ is such that for every $s \in \Xi$, $f(\cdot, s)$ differentiable in a generalized sense, then the same is true for $F(x) := \int f(x, s)\mu(\mathrm{d}s)$ [22]. Stochastic algorithms with decreasing steps involving the generalized subdifferential were studied in [12, 26].

2.3 Differential Inclusions

We endow the set of continuous function from $\mathbb{R}_+$ to $\mathbb{R}^d$ with the metric of uniform convergence on compact intervals of $\mathbb{R}_+$:

$$\boldsymbol{d}_C(\mathsf{x}, \mathsf{y}) = \sum_{n\in\mathbb{N}} 2^{-n} \left(1 \wedge \sup_{t\in[0,n]} \|\mathsf{x}(t) - \mathsf{y}(t)\|\right) \tag{10}$$

Given a set valued map $\mathsf{H} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, we say that $x : \mathbb{R}_+ \rightrightarrows \mathbb{R}^d$ is a solution of the differential inclusion

$$\dot{x}(t) \in \mathsf{H}(x(t)) \tag{11}$$

with initial condition $x_0 \in \mathbb{R}^d$, if x is absolutely continuous, $x(0) = x_0$ and (11) holds for almost every $t \in \mathbb{R}_+$. We denote by $\mathcal{S}_\mathsf{H} : \mathbb{R}^d \rightrightarrows C(\mathbb{R}_+, \mathbb{R}^d)$ the set-valued mapping such that for every $a \in \mathbb{R}^d$, $\mathcal{S}_\mathsf{H}(a)$ is set of solutions of (11) with $x_0 = a$. For every subset $A \subset E$, we define $\mathcal{S}_\mathsf{H}(A) = \bigcup_{a\in A} \mathcal{S}_\mathsf{H}(a)$.

If a map H has nonempty values we will say that it is upper semicontinuous if the graph of H, $\{(x, y) : y \in \mathsf{H}(x)\}$, is closed. In the case where H is upper semicontinuous with compact

and convex values and satisfies the condition

$$\exists K \geqslant 0,\ \forall x \in \mathbb{R}^d,\ \sup\{\|v\| : v \in H(x)\} \leqslant K(1 + \|x\|) \tag{12}$$

then $\mathcal{S}_{\mathsf{H}}(a)$ is non empty for each $a \in \mathbb{R}^d$, and moreover, $\mathcal{S}_{\mathsf{H}}(\mathbb{R}^d)$ is closed in the metric space $(C(\mathbb{R}_+, \mathbb{R}^d), \boldsymbol{d}_C)$ [2]. The Clarke subdifferential of a locally Lipschitz function is upper semicontinuous set valued map with nonempty compact convex values [9, Chap. 3].

3 Almost-Everywhere Gradient Functions

3.1 Definition

Let $(\Xi, \mathscr{T}, \mu)$ be a probability space, where the σ-field $\mathscr{T}$ is μ-complete. Let $d > 0$ be an integer. Consider a function $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$. We denote by $\Delta_f := \{(x, s) \in \mathbb{R}^d \times \Xi : x \in \mathcal{D}_{f(\cdot,s)}\}$ the set of points (x, s) s.t. $f(\cdot, s)$ is differentiable at x. We denote by $\nabla f(x, s)$ the gradient of $f(\cdot, s)$ at x, whenever it exists.

The following technical lemma, the proof of which is provided in Section 8.1, is essential.

Lemma 1. *Assume that f is $\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}$-measurable and that $f(\cdot, s)$ is continuous for every $s \in \Xi$. Then $\Delta_f \in \mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}$, and the function $\varphi_0 : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ defined as*

$$\varphi_0(x, s) = \begin{cases} \nabla f(x, s) & \text{if } (x, s) \in \Delta_f \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

is $\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}$-measurable. Moreover, if $f(\cdot, s)$ is locally Lipschitz continuous for every $s \in \Xi$, then $(\lambda \otimes \mu)(\Delta_f^c) = 0$.

Thanks to this lemma, the following definition makes sense.

Definition 1. *Assume that $f(\cdot, s)$ is locally Lipschitz continuous for every $s \in \Xi$. A function $\varphi : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ is called an* almost everywhere (a.e.)-gradient *of f if $\varphi = \nabla f$ $\lambda \otimes \mu$-almost everywhere.*

By Lemma 1, we observe that a.e.-gradients exist, since $(\lambda \otimes \mu)(\Delta_f^c) = 0$. Note that in Definition 1, we do not assume that φ is $\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}/\mathscr{B}(\mathbb{R}^d)$-measurable. The reason is that this property is not always easy to check on practical examples. However, if one denotes by $\overline{\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}}$ the $\lambda \otimes \mu$ completion of the σ-field $\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}$, an immediate consequence of Lemma 1 is that any a.e.-gradient of f is a $\overline{\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}}/\mathscr{B}(\mathbb{R}^d)$-measurable function.

3.2 Examples

Lazy gradient function. The function φ_0 given by Equation (13) is an a.e. gradient function.

Clarke gradient function. We shall refer to as a Clarke gradient function as any function $\varphi(x, s)$ such that

$$\begin{cases} \varphi(x, s) & = \nabla f(x, s) \text{ if } (x, s) \in \Delta_f, \\ \varphi(x, s) & \in \partial f(x, s) \text{ otherwise.} \end{cases} \tag{14}$$

Note that the inclusion $\varphi(x, s) \in \partial f(x, s)$ obviously holds for *all* $(x, s) \in \mathbb{R}^d \times \Xi$, because $\nabla f(x, s)$ is an element of $\partial f(x, s)$ when the former exists. However, conversely, a function

$\psi(x,s) \in \partial f(x,s)$ does not necessarily satisfy $\psi(x,s) = \nabla f(x,s)$ if $(x,s) \in \Delta_f$ (see the footnote[1]). By construction, a Clarke gradient function is an a.e. gradient function.

Selections of conservative fields.

Proposition 1. *Assume that for every $s \in \Xi$, $f(\cdot, s)$ is locally Lipschitz, path differentiable, and is a potential of some conservative field $D_s : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$. Consider a function $\varphi : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ which is $\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T} / \mathscr{B}(\mathbb{R}^d)$ measurable and satisfies $\varphi(x,s) \in D_s(x)$ for all $(x,s) \in \mathbb{R}^d \times \Xi$. Then, φ is an a.e. gradient function for f.*

Proof. Define $A := \{(x,s) \text{ s.t. } \varphi(x,s) \neq \nabla f(x,s)\}$. Applying Fubini's theorem we have:

$$\int 1_A(z) \lambda \otimes \mu(\mathrm{d}z) = \int \int 1_A((x,s)) \lambda(\mathrm{d}x) \mu(\mathrm{d}s) = 0\,,$$

where the last equality comes from the fact that for every s, $D_s = \{\nabla f(\cdot, s)\}$ λ-a.e. [8, Theorem 1]. □

We provide below an application of Proposition 1.

Autograd function. Consider Equation (3), which represents a loss of a neural network. Although f is just a composition of some simple functions, a direct calculation of the gradient (if it exists) may be tedious. Automatic differentiation deals with such functions by recursively applying the chain rule to the components of f. More formally consider a function $f : \mathbb{R}^d \to \mathbb{R}$ that can be written as a closed formula of simple functions, mathematically speaking this means that we can represent f by a directed graph. This graph (with $q > d$ vertices) is defined through a set-valued function $\mathbf{parents}(i) \subset \{1, \dots, i-1\}$, a directed edge in this setting will be $j \to i$ with $j \in \mathbf{parents}(i)$. Associate to each vertex a simple function $g_i \colon \mathbb{R}^{|\mathbf{parents}(i)|} \to \mathbb{R}$, given an input $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ we recursively define $x_i = g_i((x_j)_{j \in \mathbf{parents}(i)})$ for $i > d$ and finally $f(x) = x_q$. For instance, if f is a cross entropy loss of a neural network, with activation functions being ReLu or sigmoid functions, then g_i are some compositions of simple functions $\mathbf{log}$, $\mathbf{exp}$, $\frac{\mathbf{1}}{\mathbf{1+x^2}}$, norms and piecewise polynomial functions, all being path differentiable [8, section 6], [10, Section 5.2]. Automatic differentiation libraries calculate the gradient of f by successively applying the chain rule (in the sense $(g_1 \circ g_2)' = (g_1' \circ g_2) g_2'$) to the simple functions g_i. While the chain rule is no longer valid in a nonsmooth setting (see e.g. [17]), it is shown in [8, Section 5] that when the simple functions are path-differentiable, the output of automatic differentiation (e.g. `autograd` in PyTorch ([24])) is a selection of some conservative field D for f. We refer to [8] for a more detailed account. We denote by $a_f(x)$ the output of automatic differentiation of a function f at some point x.

Assume that $\Xi = \mathbb{N}$ and for each $s \in \Xi$, $f(\cdot, s)$ is defined through a recursive graph of path differentiable functions (in the machine learning paradigm $f(\cdot, s)$ will represent the loss related to one data point, while $F(\cdot)$ is the average loss). By Proposition 1, the map $(x,s) \mapsto a_{f(\cdot,s)}(x)$ is an a.e. gradient function for f.

Selections of generalized subdifferentials of Norkin. Noticing that a generalized subdifferential of a function is equal to its gradient a.e. ([22, Theorem 1.12]), the proof of the next proposition is identical to the one of Proposition 1.

[1] If a locally Lipschitz function g is differentiable at a point x, we have $\{\nabla g(x)\} \subset \partial g(x)$ but the inclusion could be strict (the two sets are equal if g is regular at x): for example, $g(x) = x^2 \sin(1/x)$ is s.t. $\nabla g(0) = 0$ and $\partial g(0) = [-1, 1]$. There even exist functions for which the set of x s.t. $\{\nabla g(x)\} \subsetneq \partial g(x)$ is a set of full measure (see [19, Proposition 1.9]).

Proposition 2. *Assume that for every $s \in \Xi$, $f(\cdot, s)$ is differentiable in a generalized sense, with $G_{f(\cdot,s)} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ being its generalized subdifferential. Consider a function $\varphi : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ which is $\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}/\mathscr{B}(\mathbb{R}^d)$ measurable and satisfies $\varphi(x, s) \in G_{f(\cdot,s)}(x)$ for all $(x, s) \in \mathbb{R}^d \times \Xi$. Then, φ is an a.e. gradient function for f.*

4 SGD Sequences

4.1 Definition

Given a probability measure ν on $\mathscr{B}(\mathbb{R}^d)$, define the probability space $(\Omega, \mathscr{F}, \mathbb{P}^\nu)$ as $\Omega = \mathbb{R}^d \times \Xi^{\mathbb{N}}$, $\mathscr{F} = \mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}^{\otimes\mathbb{N}}$, and $\mathbb{P}^\nu = \nu \otimes \mu^{\otimes\mathbb{N}}$. We denote by $(x_0, (\xi_n)_{n\in\mathbb{N}^*})$ the canonical process on $\Omega \to \mathbb{R}^d$ *i.e.*, writing an elementary event in the space Ω as $\omega = (\omega_n)_{n\in\mathbb{N}}$, we set $x_0(\omega) = \omega_0$ and $\xi_n(\omega) = \omega_n$ for each $n \geqslant 1$. Under $\mathbb{P}^\nu$, x_0 is a $\mathbb{R}^d$-valued random variable with the probability distribution ν, and the process $(\xi_n)_{n\in\mathbb{N}^*}$ is an independent and identically distributed (i.i.d.) process such that the distribution of ξ_1 is μ, and x_0 and (ξ_n) are independent. We denote by $\overline{\mathscr{F}}$ the $\lambda \otimes \mu^{\otimes\mathbb{N}}$-completion of $\mathscr{F}$.

Let $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ be a $\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}/\mathscr{B}(\mathbb{R})$-measurable function.

Definition 2. *Assume that $f(\cdot, s)$ is locally Lipschitz continuous for every $s \in \Xi$. A sequence $(x_n)_{n\in\mathbb{N}^*}$ of functions on $\Omega \to \mathbb{R}^d$ is called an SGD sequence for f with the step $\gamma > 0$ if there exists an a.e.-gradient φ of f such that*

$$x_{n+1} = x_n - \gamma\varphi(x_n, \xi_{n+1}) \qquad (\forall n \geqslant 0)\,.$$

4.2 All SGD Sequences Are Almost Surely Equal

Consider the SGD sequence

$$x_{n+1} = x_n - \gamma\varphi_0(x_n, \xi_{n+1}), \tag{15}$$

generated by the lazy a.e. gradient φ_0. Denote by $P_\gamma : \mathbb{R}^d \times \mathscr{B}(\mathbb{R}^d) \to [0, 1]$ the kernel of the homogeneous Markov process defined by this equation, which exists thanks to the $\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}$-measurability of φ_0. This kernel is defined by the fact that its action on a measurable function $g : \mathbb{R}^d \to \mathbb{R}_+$, denoted as $P_\gamma g(\cdot)$, is

$$P_\gamma g(x) = \int g(x - \gamma\varphi_0(x, s))\,\mu(ds). \tag{16}$$

Define Γ as the set of all steps $\gamma > 0$ such that P_γ maps $\mathcal{M}_{abs}(\mathbb{R}^d)$ into itself:

$$\Gamma := \{\gamma \in (0, +\infty) \,:\, \forall \rho \in \mathcal{M}_{abs}(\mathbb{R}^d),\ \rho P_\gamma \ll \lambda\}\,.$$

Proposition 3. *Consider $\gamma \in \Gamma$ and $\nu \in \mathcal{M}_{abs}(\mathbb{R}^d)$. Then, each SGD sequence (x_n) with the step γ is $\overline{\mathscr{F}}/\mathscr{B}(\mathbb{R}^d)^{\otimes\mathbb{N}}$-measurable. Moreover, for any two SGD sequences (x_n) and (x'_n) with the step γ, it holds that $\mathbb{P}^\nu\left[(x_n) \neq (x'_n)\right] = 0$. Finally, the probability distribution of x_n under $\mathbb{P}^\nu$ is Lebesgue-absolutely continuous for each $n \in \mathbb{N}$.*

Note that $\mathbb{P}^\nu \ll \lambda\otimes\mu^{\otimes\mathbb{N}}$ since $\nu \ll \lambda$. Thus, the probability $\mathbb{P}^\nu\left[(x_n) \neq (x'_n)\right]$ is well-defined as an integral w.r.t. $\lambda \otimes \mu^{\otimes\mathbb{N}}$.

Proof. Let (x_n) be the lazy SGD sequence given by (15). Given an a.e. gradient φ, define the SGD sequence (z_n) as $z_0 = x_0$, $z_{n+1} = z_n - \gamma\varphi(z_n, \xi_{n+1})$ for $n \geqslant 0$. The sequence (x_n) is $\mathscr{F}/\mathscr{B}(\mathbb{R}^d)^{\otimes\mathbb{N}}$-measurable thanks to Lemma 1. Moreover, applying recursively the property that $\rho P_\gamma \ll \lambda$ when $\rho \ll \lambda$, we obtain that the distribution of x_n is absolutely continuous for each $n \in \mathbb{N}$.

To establish the proposition, it suffices to show that z_n is $\overline{\mathscr{F}}/\mathscr{B}(\mathbb{R}^d)$-measurable for each $n \in \mathbb{N}$, and that $\mathbb{P}^\nu[z_n \neq x_n] = 0$, which results in particular in the absolute continuity of the distribution of z_n. We shall prove these two properties by induction on n. They are trivial for $n = 0$. Assume they are true for n. Recall that $z_{n+1} = z_n - \gamma\nabla f(z_n, \xi_{n+1})$ if $(z_n, \xi_{n+1}) \in A$, where $A \in \overline{\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}}$ is such that $(\lambda \otimes \mu)(A^c) = 0$, and $x_{n+1} = x_n - \gamma\nabla f(x_n, \xi_{n+1})1_{(x_n,\xi_{n+1})\in\Delta_f}$. The set $B = \{\omega \in \Omega : z_{n+1} \neq x_{n+1}\}$ satisfies $B \subset B_1 \cup B_2$, where

$$B_1 = \{\omega \in \Omega : z_n \neq x_n\} \quad \text{and} \quad B_2 = \{\omega \in \Omega : (z_n, \xi_{n+1}) \notin A\}.$$

By induction, $B_1 \in \overline{\mathscr{F}}$ and $\mathbb{P}^\nu(B_1) = 0$. By the aforementioned properties of A, the $\overline{\mathscr{F}}$-measurability of z_n, and the absolute continuity of its distribution, we also obtain that $B_2 \in \overline{\mathscr{F}}$ and $\mathbb{P}^\nu(B_2) = 0$. Thus, $B \in \overline{\mathscr{F}}$ and $\mathbb{P}^\nu(B) = 0$, and since x_{n+1} is $\mathscr{F}$-measurable, z_{n+1} is $\overline{\mathscr{F}}$-measurable. □

Proposition 3 means that the SGD sequence does not depend on the specific a.e. gradient used by the practioner, provided that the law of x_0 has a density and $\gamma \in \Gamma$. Let us make this last assumption clearer. Consider for instance $d = 1$ and suppose that $f(x, s) = 0.5x^2$ for all s. If $\gamma = 1$, the SGD sequence $x_{n+1} = x_n - \gamma x_n$ satisfies $x_1 = 0$ for any initial point and thus, does not admit a density, whereas for any other value of γ, x_n has a density for all n, provided that x_0 has a density. Otherwise stated, $\Gamma = \mathbb{R}_+\backslash\{1\}$ in this example.

It is desirable to ensure that Γ contains almost all the points of $\mathbb{R}_+$. The next proposition shows that this will be the case under mild conditions. The proof is given in 8.2.

Proposition 4. *Assume that for μ–almost every $s \in \Xi$, the function $f(\cdot, s)$ satisfies the property that at λ–almost every point of $\mathbb{R}^d$, there is a neighborhood of this point on which it is C^2. Then, Γ^c is Lebesgue negligible.*

This assumption holds true as soon as for μ-almost all s, $f(\cdot, s)$ is tame, since in this case $\mathbb{R}^d$ can be partitioned in manifolds on each of which $f(\cdot, s)$ is C^2 ([7]), and therefore $f(\cdot, s)$ is C^2 (in the classical sense) on the union of manifolds of full dimension, and therefore almost everywhere.

4.3 SGD as a Robbins-Monro Algorithm

We make the following assumption on the function $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$.

Assumption 1. *i) There exists a measurable function $\kappa : \mathbb{R}^d \times \Xi \to \mathbb{R}_+$ s.t. for each $x \in \mathbb{R}^d$, $\int \kappa(x, s)\,\mu(ds) < \infty$ and there exists $\varepsilon > 0$ for which*

$$\forall y, z \in B(x, \varepsilon),\ \forall s \in \Xi,\ |f(y, s) - f(z, s)| \leqslant \kappa(x, s)\|y - z\|.$$

ii) There exists $x \in \mathbb{R}^d$ such that $f(x, \cdot)$ is μ-integrable.

By this assumption, $f(x,\cdot)$ is μ-integrable for each $x \in \mathbb{R}^d$, and the function

$$F : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto \int f(x,s)\,\mu(ds) \tag{17}$$

is locally Lipschitz on $\mathbb{R}^d$. We denote by $\mathcal{Z}$ the set of (Clarke) critical points of F, as defined in Equation (4).

Let $(\mathscr{F}_n)_{n\geqslant 0}$ be the filtration $\mathscr{F}_n = \sigma(x_0, \xi_1, \dots, \xi_n)$. We denote by $\mathbb{E}_n = \mathbb{E}[\cdot|\overline{\mathscr{F}}_n]$ the conditional expectation w.r.t. $\overline{\mathscr{F}}_n$, where $\overline{\mathscr{F}}_n$, stands for the $\lambda \otimes \mu^{\mathbb{N}}$-completion of $\mathscr{F}_n$.

Theorem 1. *Let Assumption 1 holds true. Consider $\gamma \in \Gamma$ and $\nu \in \mathcal{M}_{abs}(\mathbb{R}^d) \cap \mathcal{M}_1(\mathbb{R}^d)$. Let $(x_n)_{n\in\mathbb{N}*}$ be a SGD sequence for f with the step γ. Then, for every $n \in \mathbb{N}$, it holds $\mathbb{P}^\nu$-a.e. that*

i) F, $f(\cdot, \xi_{n+1})$ and $f(\cdot, s)$ (for μ-almost every s) are differentiable at x_n.

ii) $x_{n+1} = x_n - \gamma \nabla f(x_n, \xi_{n+1})$.

iii) $\mathbb{E}_n[x_{n+1}] = x_n - \gamma \nabla F(x_n)$.

Theorem 1 is important because it shows that $\mathbb{P}^\nu$-a.e., the SGD sequence (x_n) verifies

$$x_{n+1} = x_n - \gamma \nabla F(x_n) + \gamma \eta_{n+1}$$

for some random sequence (η_n) which is a martingale difference sequence adapted to $(\overline{\mathscr{F}}_n)$.

5 Dynamical Behavior

5.1 Assumptions and Result

In this section we prove that the SGD sequence $(x_n)_{n\in\mathbb{N}*}$ (which is by Theorem 1, under the stated assumptions, unique) closely follows a trajectory of a solution to the DI (7) as the step size γ tends to zero. To state the main result of this section, we need to strengthen Assumption 1.

Assumption 2. *The function κ of Assumption 1 satisfies:*

i) There exists a constant $K \geqslant 0$ s.t. $\int \kappa(x,s)\,\mu(ds) \leqslant K(1 + \|x\|)$ for all x.

ii) For each compact set $\mathcal{K} \subset \mathbb{R}^d$, $\sup_{x\in\mathcal{K}} \int \kappa(x,s)^2 \mu(ds) < \infty$.

The first point guarantees the existence of global solutions to (7) starting from any initial point (see Section 2.3).

Assumption 3. *The closure of Γ contains 0.*

By Proposition 4, Assumption 3 is mild. It holds for instance if every $f(\cdot, s)$ is a tame function.

We recall that $\mathcal{S}_{-\partial F}(A)$ is the set of solutions to (7) that start from any point in the set $A \subset \mathbb{R}^d$.

Theorem 2. *Let Assumptions 1–3 hold true. Let* $\{(x_n^\gamma)_{n\in\mathbb{N}^*} : \gamma \in (0,\gamma_0]\}$ *be a collection of SGD sequences of steps* $\gamma \in (0,\gamma_0]$*. Denote by* x^γ *the piecewise affine interpolated process*

$$\mathsf{x}^\gamma(t) = x_n^\gamma + (t/\gamma - n)(x_{n+1}^\gamma - x_n^\gamma) \qquad (\forall t \in [n\gamma,(n+1)\gamma))\,.$$

Then, for every compact set $\mathcal{K} \subset \mathbb{R}^d$*,*

$$\forall \varepsilon > 0,\ \lim_{\substack{\gamma\to 0\\ \gamma\in\Gamma}} \left(\sup_{\nu\in\mathcal{M}_{abs}(\mathcal{K})} \mathbb{P}^\nu \left(\boldsymbol{d}_C(\mathsf{x}^\gamma, \mathcal{S}_{-\partial F}(\mathcal{K})) > \varepsilon \right) \right) = 0\,,$$

where the distance $\boldsymbol{d}_C$ *is defined in (10). Moreover, the family of distributions* $\{\mathbb{P}^\nu(\mathsf{x}^\gamma)^{-1} : \nu \in \mathcal{M}_{abs}(\mathcal{K}), 0<\gamma<\gamma_0, \gamma\in\Gamma\}$ *is tight.*

The proof is given in Section 8.4.

Theorem 2 implies that the interpolated process x^γ converges in probability as $\gamma \to 0$ to the set of solutions to (7). Moreover, the convergence is uniform w.r.t. to the choice of the initial distribution ν in the set of absolutely continuous measures supported by a given compact set.

5.2 Importance of the Randomization of x_0

In this paragraph, we discuss the case where x_0 is no longer random, but set to an arbitrary point in $\mathbb{R}^d$. In this case, there is no longer any guarantee that the iterates x_n only hit the points where a gradient exist. We focus on the case where (x_n) is a Clarke-SGD sequence of the form (14), where the function φ is assumed $\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}/\mathscr{B}(\mathbb{R}^d)$ measurable for simplicity. By Assumption 1, it is not difficult to see that $\varphi(x,\cdot)$ is μ-integrable for all $x\in\mathbb{R}^d$ and, denoting by $\mathbb{E}(\varphi(x,\cdot))$ the corresponding integral w.r.t. μ, we can rewrite the iterates under the form:

$$x_{n+1} = x_n - \gamma \mathbb{E}\varphi(x_n,\cdot) + \gamma\eta_{n+1},$$

where $\eta_{n+1} = \mathbb{E}[\varphi(x_n,\cdot)] - \varphi(x_n,\xi_{n+1})$ is a martingale difference sequence for the filtration $(\mathscr{F}_n)$. Obviously, $\mathbb{E}\varphi(x,\cdot) \in \mathbb{E}\partial f(x,\cdot)$. As said in the introduction, we need $\mathbb{E}\varphi(x,\cdot)$ to belong to $\partial F(x)$ in order to make sure that the algorithm trajectory shadows the DI $\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t))$. Unfortunately, the inclusion $\partial F(x) \subset \mathbb{E}\partial f(x,\cdot)$ can be strict, which can result in the fact that the DI $\dot{\mathsf{x}}(t) \in -\mathbb{E}\partial f(\mathsf{x}(t),\cdot)$ generates spurious trajectories that converge to spurious zeroes. The following example, which can be easily adapted to an arbitrary dimension, shows a case where this phenomenon happens.

Example 1. *Take a finite probability space* $\Xi = \{1,2\}$ *and* $\mu(\{1\}) = \mu(\{2\}) = 1/2$*. Let* $f(x,1) = 2x1_{x\leqslant 0}$ *and* $f(x,2) = 2x1_{x\geqslant 0}$*. We have* $F(x) = x$*, and therefore* $\partial F(0) = \{1\}$*, whereas* $\partial f(0,1) = \partial f(0,2) = [0,2]$ *and therefore* $\int \partial f(0,s)\mu(\mathrm{d}s) = [0,1]$*. We see that* $0 \in \mathbb{E}\partial f(0,\cdot)$ *while* $0 \notin \partial F(0)$*. Furthermore, the trajectory defined on* $\mathbb{R}_+$ *as*

$$\mathsf{x}(t) = \begin{cases} 1-t & \text{for } t\in[0,1] \\ 0 & \text{for } t>1 \end{cases}, \quad \mathsf{x}(0) = 1,$$

is a solution to the DI $\dot{\mathsf{x}}(t) \in -\mathbb{E}\partial f(\mathsf{x}(t),\cdot)$*, but not to the DI* $\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t))$*.*

Example 2. *Consider the same setting as in the previous example. Consider a stochastic gradient algorithm of the form (1), initialized at* $x_0 = 0$ *with* φ *such that* $\varphi(0,1) = \varphi(0,2) = 0$*. Then, the iterates* x_n^γ *are identically zero. This shows that the stochastic gradient descent may converge to a non critical point of* F*. Theorem 2 may fail unless a random initial point is chosen.*

6 Long Run Convergence

6.1 Assumptions and Result

As discussed in the introduction, the SGD sequence (x_n) is not expected to converge in probability to $\mathcal{Z}$ when the step is constant. Instead, we shall establish the convergence (5). The "long run" convergence referred to here is understood in this sense.

In all this section, we shall focus on the lazy SGD sequences described by Equation (15). This incurs no loss of generality, since any two SGD sequences are equal $\mathbb{P}^\nu$-a.e. by Proposition 3 as long as $\nu \ll \lambda$. Our starting point is to see the process (x_n) as a Markov process which kernel P_γ is defined by Equation (16). Our first task is to establish the ergodicity of this Markov process under the convenient assumptions. Namely, we show that P_γ has a unique invariant probability measure π_γ, *i.e.*, $\pi_\gamma P_\gamma = \pi_\gamma$, and that $\|P_\gamma^n(x,\cdot) - \pi_\gamma\|_{\mathrm{TV}} \to 0$ as $n \to \infty$ for each $x \in \mathbb{R}^d$, where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm. Further, we need to show that the family of invariant distributions $\{\pi_\gamma\}_{\gamma\in(0,\gamma_0]}$ for a certain $\gamma_0 > 0$ is tight. The long run behavior referred to above is then intimately connected with the properties of the accumulation points of this family as $\gamma \to 0$. To study these properties, we get back to the DI $\dot{\mathsf{x}} \in -\partial F(\mathsf{x})$ (we recall that a concise account of the notions relative to this dynamical system and needed in this paper is provided in Section 2.3). The crucial point here is to show, with the help of Theorem 2, that the accumulation points of $\{\pi_\gamma\}$ as $\gamma \to 0$ are invariant measures for the set-valued flow induced by the DI. In its original form, this idea dates back to the work of Has'minskiĭ [15]. We observe here that while the notion of invariant measure for a single-valued semiflow induced by, say, an ordinary differential equation, is classical, it is probably less known in the case of a set-valued differential inclusion. We borrow it from the work of Roth and Sandholm [25].

Having shown that the accumulation points of $\{\pi_\gamma\}$ are invariant for the DI $\dot{\mathsf{x}} \in -\partial F(\mathsf{x})$, the final step of the proof is to make use of Poincaré's recurrence theorem, that asserts that the invariant measures of a semiflow are supported by the so-called Birkhoff center of this semiflow (again, a set-valued version of Poincaré's recurrence theorem is provided in [3, 13]). To establish the convergence (5), it remains to show that the Birkhoff center of the DI $\dot{\mathsf{x}} \in -\partial F(\mathsf{x})$ coincides with $\operatorname{zer}\partial F$. The natural assumption that ensures the identity of these two sets will be that F admits a chain rule [9, 7, 10].

Our assumption regarding the behavior of the Markov kernel P_γ reads as follows.

Assumption 4. *There exist measurable functions $V : \mathbb{R}^d \to [0,+\infty)$, $p : \mathbb{R}^d \to [0,+\infty)$, $\alpha : (0,+\infty) \to (0,+\infty)$ and a constant $C \geqslant 0$ s.t. the following holds for every $\gamma \in \Gamma \cap (0,\gamma_0]$.*

i) There exists $R > 0$ and a positive Borel measure ρ on $\mathbb{R}^d$ (R, ρ possibly depending on γ) such that

$$\forall x \in \mathrm{cl}(B(0,R)),\ \forall A \in \mathscr{B}(\mathbb{R}^d),\ P_\gamma(x,A) \geqslant \rho(A).$$

ii) $\sup_{\mathrm{cl}(B(0,R))} V < \infty$ and $\inf_{B(0,R)^c} p > 0$. Moreover, for every $x \in \mathbb{R}^d$,

$$P_\gamma V(x) \leqslant V(x) - \alpha(\gamma)p(x) + C\alpha(\gamma)1_{\|x\|\leqslant R}. \tag{18}$$

iii) The function $p(x)$ diverges to infinity as $\|x\| \to \infty$.

Assumptions of this type are frequently encountered in the field of Markov chains. Assumption 4–(i) states that $\mathrm{cl}(B(0,R))$ is a so-called small set for the kernel P_γ, and Assumption 4–(ii) is a standard drift assumption. Taken together, they ensure that the kernel P_γ is a so-called Harris-recurrent kernel, that it admits a unique invariant probability distribution π_γ, and finally, that this kernel is ergodic in the sense that $\|P_\gamma(x,\cdot) - \pi_\gamma\|_{\mathrm{TV}} \to 0$ as $n \to \infty$ (see [21]). The introduction of the factors $\alpha(\gamma)$ and $C\alpha(\gamma)$ in Equation (18) guarantees moreover the tightness of the family $\{\pi_\gamma\}_{\gamma\in(0,\gamma_0]}$.

In Section 6.2, we provide sufficient and verifiable conditions ensuring the validity of Assumption 4 for P_γ.

As announced above, we also need:

Assumption 5. *The function F defined by (17) admits a chain rule, namely, for any absolutely continuous curve $z : \mathbb{R}_+ \to \mathbb{R}^d$, for almost all $t > 0$, $\forall v \in \partial F(z(t))$, $\langle v, \dot{z}(t)\rangle = (F \circ z)'(t)$.*

Assumption 5 is satisfied as soon as F is path-differentiable, for instance when F is either convex, regular, Whitney stratifiable or tame (see [8, Proposition 1] and [7, 10]).

Since Assumption 3 is satisfied as soon as $f(\cdot, s)$ is tame for every $s \in \Xi$, one can wonder if it can be somehow coordinated with Assumption 5. Unfortunately, F is not necessarily tame even if $f(\cdot, s)$ is tame for every $s \in \Xi$. Nonetheless, one can hope that the practical situations where $f(\cdot, s)$ is tame and F is not are rare. In particular, F will be tame as soon as Ξ is finite (hence the expectation is just a finite sum), which is the case in many machine learning models.

Theorem 3. *Let Assumptions 1-3 and 4-5 hold true. Let $\{(x_n^\gamma)_{n\in\mathbb{N}^*} : \gamma \in (0,\gamma_0]\}$ be a collection of SGD sequences of step-size γ. Then, the set $\mathcal{Z} = \{x\colon 0 \in \partial F(x)\}$ is nonempty and for all $\nu \in \mathcal{M}_{abs}(\mathbb{R}^d)$ and all $\varepsilon > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}^\nu \left(\boldsymbol{d}(x_n^\gamma, \mathcal{Z}) > \varepsilon\right) \xrightarrow[\substack{\gamma\to 0\\ \gamma\in\Gamma}]{} 0. \tag{19}$$

6.2 On Assumption 4

In this paragraph, we provide sufficient conditions under which Assumption 4 hold true. A simple way to ensure the truth of Assumption 4-(i) is to add a small random perturbation to the function $\varphi_0(x,s)$. Formally, we modify algorithms described by Equation (15) and (21), and write

$$x_{n+1} = x_n - \gamma\varphi_0(x_n, \xi_{n+1}) + \gamma\epsilon_{n+1}$$

where (ϵ_n) is a sequence of centered i.i.d. random variables of law μ^d, independent from $\{x_0, (\xi_n)\}$, and such that the distribution of $\epsilon_1 \sim \mu^d$ has a continuous and positive density on $\mathbb{R}^d$. The Gaussian case $\epsilon_1 \sim \mathcal{N}(0, aI_d)$ where $a > 0$ is some small variance is of course a typical example of such a perturbation.

Consider now a fixed γ and denote by $\widetilde{P}$ the Markov kernel induced by the modified equation.

Proposition 5. *Let Assumption 2 hold true. Then, for each $R > 0$, there exists $\varepsilon > 0$ such that*

$$\forall x \in \mathrm{cl}(B(0,R)),\ \forall A \in \mathscr{B}(\mathbb{R}^d),\ \widetilde{P}(x,A) \geqslant \varepsilon\,\lambda(A \cap \mathrm{cl}(B(0,1))),$$

Thus, Assumption 4-(i) is satisfied for $\widetilde{P}$.

We now turn to the assumptions 4-(ii) and 4-(iii).

Proposition 6. *Assume that there exists $R \geqslant 0$, $C > 0$, and a measurable function $\beta : \Xi \to \mathbb{R}_+$ such that the following conditions hold:*

i) For every $s \in \Xi$, the function $f(\cdot, s)$ is differentiable outside the ball $\mathrm{cl}(B(0,R))$. Moreover, for each $x, x' \notin \mathrm{cl}(B(0,R))$, $\|\nabla f(x,s) - \nabla f(x',s)\| \leqslant \beta(s)\|x - x'\|$ and $\int \beta^2 d\mu < \infty$.

ii) For all $x \notin \mathrm{cl}(B(0,R))$, $\int \|\nabla f(x,s)\|^2 \mu(ds) \leqslant C(1 + \|\nabla F(x)\|^2)$.

iii) $\lim_{\|x\|\to\infty} \|\nabla F(x)\| = +\infty$.

iv) Function F is lower bounded i.e., *$\inf F > -\infty$.*

Then, it holds that

$$P_\gamma F(x) \leqslant F(x) - \gamma(1 - \gamma K) 1_{\|x\|>2R} \|\nabla F(x)\|^2 + \gamma^2 K 1_{\|x\|>2R} + \gamma K 1_{\|x\|\leqslant 2R} \tag{20}$$

for some constant $K > 0$. In particular, Assumptions 4-(ii) and 4-(iii) hold true.

We finally observe that this proposition can be easily adapted to the case where the kernel P_γ is replaced with the kernel $\widetilde{P}$ of Proposition 5.

7 The Projected Subgradient Algorithm

In many practical settings, the conditions of Proposition 6 that ensure the truth of Assumptions 4–(ii) and 4–(iii) are not satisfied. This is for instance the case when the function f is described by Equation (3) with the mappings σ_ℓ at the right hand side of this equation being all equal to the ReLU function. In such situations, it is often pertinent to replace the SGD sequence with a *projected* version of the algorithm. Given an a.e.-gradient φ of the function f and a non empty compact and convex set $\mathcal{K} \subset \mathbb{R}^d$, a *projected SGD sequence* $(x_n^{\gamma,\mathcal{K}})$ is given by the recursion

$$x_0^{\gamma,\mathcal{K}} = x_0, \quad x_{n+1}^{\gamma,\mathcal{K}} = \Pi_{\mathcal{K}}(x_n^{\gamma,\mathcal{K}} - \gamma\varphi(x_n^{\gamma,\mathcal{K}}, \xi_{n+1}))\,, \tag{21}$$

where $\Pi_{\mathcal{K}}$ stands for a Euclidean projection onto $\mathcal{K}$. Our purpose is to generalize Theorem 2 to this situation. This generalization is not immediate for several reasons. First, the projection step is likely to introduce spurious local minima. As far as the iterates (21) are concerned, the role of differential inclusion (7) is now played by the differential inclusion:

$$\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t)) - \mathcal{N}_{\mathcal{K}}(\mathsf{x}(t))\,, \tag{22}$$

where $\mathcal{N}_{\mathcal{K}}(x)$ stands the normal cone of $\mathcal{K}$ at point x. The set of equilibria of the above differential inclusion coincides with the set

$$\mathcal{Z}_{\mathcal{K}} := \{x \in \mathbb{R}^d \,:\, 0 \in -\partial F(x) - \mathcal{N}_{\mathcal{K}}(x)\}\,,$$

which we shall refer to as the set of Karush-Kuhn-Tucker points. A second theoretical difficulty is related to the fact that Proposition 3 does no longer hold. Indeed, it can happen x_0 has a density, but the next iterates $x_n^{\gamma,\mathcal{K}}$ don't. The reason is that $x_n^{\gamma,\mathcal{K}}$ generally has a non zero probability to be in the (Lebesgue negligible) border of $\mathcal{K}$, that is, $\mathrm{cl}(\mathcal{K})\backslash\,\mathrm{int}(\mathcal{K})$, where $\mathrm{cl}(\mathcal{K})$ and $\mathrm{int}(\mathcal{K})$ respectively stand for the closure and the interior of $\mathcal{K}$.

We shall focus here on the case where $\mathcal{K} = \mathrm{cl}(B(0,r))$ with $r > 0$. We shall use Π_r, $x_n^{\gamma,r}$, $\mathcal{N}_r$ as shorthand notations for $\Pi_{\mathrm{cl}(B(0,r))}$, $x_n^{\gamma,\mathrm{cl}(B(0,r))}$, and $\mathcal{N}_{\mathrm{cl}(B(0,r))}$ respectively. In this case $\mathcal{N}_r(x) = \{0\}$ if $\|x\| < r$, $\mathcal{N}_r(x) = \{\lambda x : \lambda \geqslant 0\}$ if $\|x\| = r$ and $\mathcal{N}_r(x) = \varnothing$ otherwise.

We make the following assumption.

Assumption 6. *For every $x \in \mathbb{R}^d$, the law of $\varphi_0(x,\xi)$, where $\xi \sim \mu$, is absolutely continuous relatively to Lebesgue.*

Assumption 6 is much stronger than Assumption 3. Indeed, it implies that the distribution of $x_n^{\gamma,r} - \gamma\varphi(x_n^{\gamma,r}, \xi_{n+1})$ is always Lebesgue-absolutely continuous. It is useful to note though that Assumption 6 holds upon adding at each step a small random perturbation to φ_0 as in Section 6.2 above.

In order to state our first result in this framework, we need to introduce some new notations. We let $\mathbb{S}(r) := \{x : \|x\| = r, x \in \mathbb{R}^d\}$ be the sphere of radius r. By [14, Theorem 2.49], there is a unique measure[2] ϱ_1 on $\mathbb{S}(1)$ such that for any positive function $f : \mathbb{R}^d \to \mathbb{R}$, we have:

$$\int f(x)\lambda^d(\mathrm{d}x) = \int_0^\infty \int_{\mathbb{S}(1)} f(r\theta) r^{d-1} \varrho_1(\mathrm{d}\theta)\lambda^1(\mathrm{d}r)\,. \tag{23}$$

We define the measure ϱ_r on $\mathbb{S}(r)$ as $\varrho_r(A) = \varrho_1(A/r)$ for each Borel set $A \subset \mathbb{S}(r)$. We denote as $\mathcal{M}^r$ the set of measures $\nu = \nu_1 + \nu_2$, where $\nu_1 \in \mathcal{M}_{abs}$ and $\nu_2 \ll \varrho_r$. For a set $\mathcal{C} \subset \mathbb{R}^d$ we define $\mathcal{M}^r(\mathcal{C})$ as the measures in $\mathcal{M}^r$ that are supported on $\mathcal{C}$. Notice that $\mathcal{M}_{abs}(\mathcal{C}) \subset \mathcal{M}^r(\mathcal{C})$.

The next proposition, which is proven in the same way as Proposition 3, shows that for almost every $r > 0$, all projected SGD sequences are almost surely equal.

Proposition 7. *Let Assumption 6 hold true. Then, for almost every $r > 0$, $\forall \nu \in \mathcal{M}^r$, each projected SGD sequence $(x_n^{\gamma,r})$ is $\overline{\mathscr{F}}/\mathscr{B}(\mathbb{R}^d)^{\otimes\mathbb{N}}$-measurable. Moreover, for any two projected SGD sequences $(x_n^{\gamma,r})$ and $(y_n^{\gamma,r})$, it holds that $\mathbb{P}^\nu\left[(x_n^{\gamma,r}) \neq (y_n^{\gamma,r})\right] = 0$. Finally, under $\mathbb{P}^\nu$, for every $n \in \mathbb{N}$, the probability distribution of $x_n^{\gamma,r}$ is in $\mathcal{M}^r$.*

By Proposition 7 we can focus on the lazy projected SGD sequence:

$$x_{n+1}^{\gamma,r} = \Pi_r(x_n^{\gamma,r} - \gamma\varphi_0(x_n^{\gamma,r}, \xi_{n+1}))\,. \tag{24}$$

We define its associated kernel

$$P_\gamma^r g(x) = \int g(\Pi_r(x - \gamma\varphi_0(x,s)))\mu(\mathrm{d}s)\,. \tag{25}$$

The next two theorems are analogous to Theorems 1 and 2.

Theorem 4. *Let Assumptions 1 and 6 hold. Then for almost every $r > 0$, $\forall \nu \in \mathcal{M}^r$, for every $n \in \mathbb{N}$ it holds $\mathbb{P}^\nu$-a.e.*

i) F, $f(\cdot, \xi_{n+1})$ and $f(\cdot, s)$ (for μ-a.e. s) are differentiable at $x_n^{\gamma,r}$.

ii) $x_{n+1}^{\gamma,r} \in x_n^{\gamma,r} - \gamma\nabla f(x_n^{\gamma,r}, \xi_{n+1}) - \gamma\mathcal{N}_r(\Pi_r(x_n^{\gamma,r} - \gamma\nabla f(x_n^{\gamma,r}, \xi_{n+1})))$.

[2] As it is clear from Equation (23) we can see (λ^1, ϱ_1) as a polar coordinates representation of the Lebesgue measure λ^d.

Theorem 5. *Let Assumptions 1–2 and 6 hold true. Denote* $\mathsf{x}^{\gamma,r}$ *the piecewise affine interpolated process:*

$$\mathsf{x}^{\gamma,r}(t) = x_n^{\gamma,r} + (t/\gamma - n)(x_{n+1}^{\gamma,r} - x_n^{\gamma,r}) \qquad (\forall t \in [n\gamma, (n+1)\gamma)) \,.$$

Then, for almost every $r > 0$*, for every compact set* $\mathcal{K} \subset \mathrm{cl}(B(0,r))$*,*

$$\forall \varepsilon > 0, \ \lim_{\gamma \to 0} \left(\sup_{\nu \in \mathcal{M}^r(\mathcal{K})} \mathbb{P}^\nu \left(\boldsymbol{d}_C(\mathsf{x}^{\gamma,r}, \mathcal{S}_{-\partial F - \mathcal{N}_r}(\mathcal{K})) > \varepsilon \right) \right) = 0 \,.$$

Moreover, for any $\gamma_0 > 0$*, the family of distributions* $\{\mathbb{P}^\nu(\mathsf{x}^{\gamma,r})^{-1} : \nu \in \mathcal{M}^r(\mathcal{K}), 0 < \gamma < \gamma_0\}$ *is tight.*

We compare Theorems 1 and 2. First, because of the projection step (and with the help of Assumption 6), the law of the n-th iterate is no longer in $\mathcal{M}_{abs}$, but in $\mathcal{M}^r$. Second, the continuous counterpart of Equation (21) is now the differential inclusion (22) Note that, if the solutions of the DI (7) that start from $\mathcal{K}$ all lie in $\mathrm{cl}(B(0,r))$, then the set of these solutions coincides with the set of solutions of the DI (22) that start from $\mathcal{K}$.

The analysis of the convergence of the iterates in the "long run" is greatly simplified by the introduction of the projection step. Compared to Assumption 4, we only assume the existence of a small set for P_γ^r, the drift condition of the form 4-(ii)–(iii) is then automatically satisfied, thanks to the projection step (see Section 8.5).

Assumption 7. *There is* $R > 0$ *and* $\gamma_0 > 0$ *such that for every* $\gamma \in (0, \gamma_0]$ *there is* ρ_γ *such that Assumption 4-(i) hold for* (R, ρ_γ) *(note that* R *is independent of* γ *here).*

As shown in Section 6.2, Assumption 7 holds upon adding to φ_0 a small random perturbation.

Theorem 6. *Let Assumptions 1-2 and 5–7 hold. Let* $\{(x_n^{\gamma,r})_{n\in\mathbb{N}^*} : \gamma \in (0, \gamma_0]\}$ *be a collection of projected SGD sequences of step-size* γ*. Then, for almost every* $0 < r \leqslant R$*, the set* $\mathcal{Z}_r = \{x : 0 \in \partial F(x) + \mathcal{N}_r(x)\}$ *is nonempty and for all* $\nu \in \mathcal{M}^r$ *and all* $\varepsilon > 0$*,*

$$\limsup_{n\to\infty} \mathbb{P}^\nu \left(\boldsymbol{d}(x_n^{\gamma,r}, \mathcal{Z}_r) > \varepsilon \right) \xrightarrow[\gamma\to 0]{} 0. \tag{26}$$

Theorem 6 is analogous to Theorem 3. Notice that, since $\mathcal{M}_{abs} \subset \mathcal{M}^r$, x_0 can still be initialized under a Lebesgue-absolutely continuous measure. On the other hand, as explained in the beginning of this section, due to the projection step, the iterates, instead of converging to $\mathcal{Z}$, are now converging to the set of Karush-Kuhn-Tucker points related to the DI (22).

8 Proofs

8.1 Proof of Lemma 1

By definition, $(x,s) \in \Delta_f$ means that there exists $d_x \in \mathbb{R}^d$ (the gradient) s.t. $f(x+h,s) = f(x,s) + \langle d_x, h\rangle + o(\|h\|)$. That is to say (x,s) belongs to the set:

$$\bigcap_{\varepsilon\in\mathbb{Q}} \bigcup_{\delta\in\mathbb{Q}} \bigcap_{0<\|h\|\leqslant\delta} \left\{ (y,s) : \left| \frac{f(y+h,s) - f(y,s) - \langle d_x, h\rangle}{\|h\|} \right| < \varepsilon \right\} . \tag{27}$$

In addition, using that $f(\cdot, s)$ is continuous, the above set is unchanged if the inner intersection over $0 < \|h\| \leqslant \delta$ is replaced by an intersection over the h s.t. $0 < \|h\| \leqslant \delta$ and having *rational* coordinates *i.e.*, $h \in \mathbb{Q}^d$. Define:

$$\Delta'_f := \bigcap_{\varepsilon' \in \mathbb{Q}} \bigcup_{d \in \mathbb{Q}^d} \bigcap_{\varepsilon \in \mathbb{Q}} \bigcup_{\delta \in \mathbb{Q}} \bigcap_{\substack{0 < \|h\| \leqslant \delta \\ h \in \mathbb{Q}^d}} \left\{ (x,s) : \left| \frac{f(x+h,s) - f(x,s) - \langle d, h \rangle}{\|h\|} \right| < \varepsilon + \varepsilon' \right\} \tag{28}$$

By construction, Δ'_f is a measurable set. We prove that $\Delta'_f = \Delta_f$. Consider $(x,s) \in \Delta_f$ and let d_x be the gradient of $f(\cdot, s)$ at x. By (27) for all $\varepsilon \in \mathbb{Q}$, there is a $\delta \in \mathbb{Q}$ such that:

$$(x,s) \in \bigcap_{h \leqslant \delta, h \in \mathbb{Q}^d} \left\{ \left| \frac{f(x+h,s) - f(x,s) - \langle d_x, h \rangle}{h} \right| < \varepsilon \right\}$$

For any $\varepsilon' > 0$, choose $d' \in \mathbb{Q}^d$ such that $\|d' - d_x\| \leqslant \varepsilon'$. Using the previous inclusion, for all ε, there exists therefore $\delta \in \mathbb{Q}$ s.t.

$$(x,s) \in \bigcap_{h \leqslant \delta, h \in \mathbb{Q}^d} \left\{ \left| \frac{f(x+h,s) - f(x,s) - \langle d_q, h \rangle}{h} \right| < \varepsilon + \varepsilon' \right\}$$

which means $\Delta_f \subset \Delta'_f$. To show the converse, consider $(x,s) \in \Delta'_f$. Let (ε'_k) be a positive sequence of rationals converging to zero. By definition, for every k, there exists $d_k \in \mathbb{Q}^d$ s.t. for all ε, there exists $\delta_k(\varepsilon)$, s.t. for all (rational) $h \leqslant \delta_k(\varepsilon)$,

$$\left| \frac{f(x+h,s) - f(x,s) - \langle d_k, h \rangle}{h} \right| < \varepsilon + \varepsilon'_k \,. \tag{29}$$

Moreover, one may choose $\delta_k(\varepsilon) \leqslant \delta_0(\varepsilon)$. Inspecting first the inequality (29) for $k = 0$, we easily obtain that the quantity $\frac{f(x+h,s)-f(x,s)}{h}$ is bounded uniformly in h s.t. $0 < \|h\| \leqslant \delta_0(\varepsilon)$. Using this observation and again Equation (29), this in turn implies that (d_k) is a bounded sequence. There exists $d \in \mathbb{R}^d$ and s.t. $d_k \to d$ along some extracted subsequence. Now consider $\varepsilon > 0$ and choose k such that $\|d_k - d\| < \frac{\varepsilon}{2}$ and $\varepsilon'_k < \frac{\varepsilon}{2}$. For all $h \leqslant \delta_k(\varepsilon/2)$,

$$\left| \frac{f(x+h,s) - f(x,s) - \langle d, h \rangle}{h} \right| \leqslant \left| \frac{f(x+h,s) - f(x,s) - \langle d_k, h \rangle}{h} \right| + \|d - d_k\| < \varepsilon$$

This means that d is the gradient of $f(\cdot, s)$ at x, hence $\Delta'_f \subset \Delta_f$. Hence, the first point of the Lemma 1 is proved.

Denoting as e_i the i^{th} canonical vector of $\mathbb{R}^d$, the i^{th}-component $[\varphi_0]_i$ in $\mathbb{R}^d$ of the function φ_0 is given as

$$[\varphi_0(x,s)]_i = \lim_{t \to 0} \frac{f(x + te_i, s) - f(x,s)}{t} 1_{\Delta_f}(x,s),$$

and the measurability of φ_0 follows from the measurability of f and the measurability of 1_{Δ_f}.

Finally, assume that $f(\cdot, s)$ is locally Lipschitz continuous for every $s \in \Xi$. From Rademacher's theorem [9, Ch. 3], $f(\cdot, s)$ is almost everywhere differentiable, which reads $\int (1 - 1_{\Delta_f}(x,s)) \lambda(dx) = 0$. Using Fubini's theorem, $\int_{\mathbb{R}^d \times \Xi} (1 - 1_{\Delta_f}(x,s)) \, \lambda(dx) \otimes \mu(ds) = 0$, and the last point is proved.

8.2 Proof of Proposition 4

The idea of the proof is to show that for almost every γ and s we have that $g_{s,\gamma}(x) := (x - \gamma\nabla f(x,s))1_{\Delta_f}(x,s)$ is almost everywhere a local diffeomorphism.

In order to prove that we define for each $(x,s) \in \mathbb{R}^d \times \Xi$ the pseudo-hessian $\mathcal{H}(x,s) \in \mathbb{R}^{d\times d}$ as

$$\mathcal{H}(x,s)_{i,j} = \limsup_{t\to 0} \frac{\langle \nabla f(x+te_j,s)1_{\Delta_f}(x+te_j,s) - \nabla f(x,s), e_i\rangle}{t} 1_{\Delta_f}(x,s)\,.$$

Since it is a limit of measurable functions, $\mathcal{H}$ is $\mathcal{B}(\mathbb{R}^d)\otimes\mathscr{T}$ measurable, and if $f(\cdot,s)$ is two times differentiable at x then $\mathcal{H}(x,s)$ is just the ordinary hessian. Now we define $l(x,s,\gamma) = \det(\gamma\mathcal{H}(x,s) - \mathrm{Id})$ if every entry in $\mathcal{H}(x,s)$ is finite, and $l(x,s,\gamma) = 1$ otherwise, it is a $\mathcal{B}(\mathbb{R}^d)\otimes\mathscr{T}\otimes\mathcal{B}(\mathbb{R}_+)$ measurable function (as a sum of two measurable functions). By the inverse function theorem we have that if $f(\cdot,s)$ is C^2 at x and if $\det(\gamma\mathcal{H}(x,s) - \mathrm{Id}) \neq 0$, then $g_{s,\gamma}(\cdot)$ is a local diffeomorphism at x. Therefore $l(x,s,\gamma) \neq 0$ implies either the latter or $f(\cdot,s)$ is not C^2 at x (or both).
Let λ^d, λ^1 denote Lebesgue measures respectively on $\mathbb{R}^d$ and $\mathbb{R}_+$, we have by Fubini's theorem:

$$\begin{aligned}\int 1_{l(x,s,\gamma)=0}\lambda^d(\mathrm{d}x)\otimes\mu(\mathrm{d}s)\otimes\lambda^1(\mathrm{d}\gamma) &= \int \lambda^d\otimes\mu(\{(x,s) : l(x,s,\gamma)=0\})\lambda^1(\mathrm{d}\gamma)\\ &= \int\int\int 1_{l(x,s,\gamma)=0}\lambda^1(\mathrm{d}\gamma)\lambda^d(\mathrm{d}x)\mu(\mathrm{d}s)\\ &= 0\,,\end{aligned}$$

where the last equality comes from the fact that for (x,s) fixed $l(x,s,\gamma) = 0$ only if $1/\gamma$ is in the spectrum of $\mathcal{H}(x,s)$ which is finite. Therefore we have a Γ a set of full measure in $\mathbb{R}_+$ such that for $\gamma\in\Gamma$ we have $\lambda^d\otimes\mu(\{(x,s) : l(x,s,\gamma)=0\}) = 0$. Once again applying Fubini's theorem we get that for almost every $s\in\Xi$ we have $\{x : g_{s,\gamma}(\cdot)$ is a local diffeomorphism at $x\}$) is of λ^d-full measure (since for each s, $f(\cdot,x)$ is almost everywhere C^2). Finally, for $A\subset\mathbb{R}^d$, $\gamma\in\Gamma$ and $\nu\in\mathcal{M}_{abs}(\mathbb{R}^d)$, we have

$$\nu P_\gamma(A) = \nu\otimes\mu(\{(x,s) : g_{s,\gamma}(x)\in A\}) \leqslant \lambda^d\otimes\mu(\{(x,s) : g_{s,\gamma}(x)\in A\})\,,$$

and by Fubini's theorem,

$$\begin{aligned}\lambda^d\otimes\mu(\{(x,s) : g_{s,\gamma}(x)\in A\}) &= \int \lambda^d(\{x : g_{s,\gamma}(x)\in A\})\mu(\mathrm{d}s)\\ &= \int \lambda^d(\{x : g_{s,\gamma}(x)\in A \text{ and } f(\cdot,s) \text{ is } C^2 \text{ at } x\})\mu(\mathrm{d}s)\\ &= \int \lambda^d(\{x : g_{s,\gamma}(x)\in A \text{ and } g_{s,\gamma}(\cdot) \text{ is a local diffeomorphism at } x\})\mu(\mathrm{d}s)\,.\end{aligned}$$

Now by separability of $\mathbb{R}^d$ there is a countable family of open neighborhoods $(V_i)_{i\in\mathbb{N}}$ such that for any open set O we have $O = \bigcup_{j\in J} V_j$. The set of x where $g(\cdot,s,\gamma)$ is a local diffeomorphism is an open set, hence

$$\{x : g_{s,\gamma}(x)\in A \text{ and } g_{s,\gamma}(\cdot) \text{ is a local diffeomorphism at } x\} = \bigcup_{i\in I} V_i \cap \{x : g_{s,\gamma}(x)\in A\}\,.$$

Since an image of a null set by a diffeomorphism is a null set we have

$$\lambda^d(\{x : g_{s,\gamma}(x)\in A\}\cap V_i) = 0\,.$$

Hence, $\nu P_\gamma(A) = 0$, which proves our claim.

8.3 Proof of Theorem 1

Take $\nu \ll \lambda$ and a SGD sequence $(x_n)_{n\in\mathbb{N}}$, let $S_1 \subset \mathbb{R}^d$ be the set of x for which $\nabla f(x,s)$ exists for μ- almost every s, *i.e.*,

$$S_1 := \left\{ x \in \mathbb{R}^d \,:\, \int_\Xi (1 - 1_{\Delta_f}(x,s))\ \mu(ds) = 0 \right\}.$$

When Assumption 1 holds, Rademacher's theorem, lemma 1 and Fubini's theorem imply that $S_1 \in \mathscr{B}(\mathbb{R}^d)$ and $\lambda(\mathbb{R}^d \backslash S_1) = 0$. Hence, for μ-a.e. s we have $f(\cdot, s)$ differentiable at x_0, and since $\xi_1 \sim \mu$, $f(\cdot, \xi_1)$ is differentiable at x_0. Now by Rademacher's theorem again, the set $S_2 \subset \mathbb{R}^d$ where F is differentiable satisfies $\lambda(\mathbb{R}^d \backslash S_2) = 0$, therefore F is differentiable at x_0. Moreover, with probability one x_0 is in $S_1 \cap S_2$. Define $A(x) := \{s \in \Xi : (x,s) \notin \Delta_f\}$. By Assumption 1, $\|\nabla f(x,\cdot)\|$ is μ-integrable. Moreover, for all $x \in S_1 \cap S_2$ and all $v \in \mathbb{R}^d$

$$\begin{aligned}
\Big\langle \int \nabla f(x,s) 1_{\Delta_f}(x,s)\, \mu(ds), v \Big\rangle &= \int_{\Xi\backslash A(x)} \langle \nabla f(x,s), v\rangle\, \mu(ds) \\
&= \int_{\Xi\backslash A(x)} \lim_{t\in\mathbb{R}^*\to 0} \frac{f(x+tv,s) - f(x,s)}{t}\, \mu(ds) \\
&= \lim_{t\in\mathbb{R}^*\to 0} \int_\Xi \frac{f(x+tv,s) - f(x,s)}{t}\, \mu(ds) \\
&= \lim_{t\in\mathbb{R}^*\to 0} \frac{F(x+tv) - F(x)}{t} \quad = \langle \nabla F(x), v\rangle
\end{aligned}$$

where the interchange between the limit and the integral follows from Assumption 1 and the dominated convergence theorem. Hence, $\nabla F(x) = \int \nabla f(x,s) 1_{\Delta_f}(x,s)\, \mu(ds)$ for all $x \in S_1 \cap S_2$. Now denote by ν_n the law of x_n. Since we assumed that $\nu_0 \ll \lambda$, it holds that $\mathbb{P}^\nu(x_0 \in S_1 \cap S_2) = 1$. Therefore, with probability one,

$$x_1 = x_1 1_{S_1\cap S_2}(x_0) = (x_0 - \gamma \nabla f(x_0, \xi_1)) 1_{S_1\cap S_2}(x_0) = x_0 - \gamma \nabla f(x_0, \xi_1)\,.$$

Thus, x_1 is integrable whenever x_0 is integrable, and $\mathbb{E}_0(x_1) = x_0 - \gamma \nabla F(x_0)$. Since by Assumption $\nu_1 \ll \lambda$ we can iterate our argument for x_2 and then for all x_n and the conclusions of Theorem 1 follow.

8.4 Proof of Theorem 2

We want to apply [6, Theorem 5.1.], and therefore verify its assumptions [6, Assumption RM]. In order to fall in its setting we first need to rewrite our kernel in a more appropriate way. As ∂F takes nonempty compact values, it admits a measurable selection $\varphi(x) \in \partial F(x)$ [1, Lemma 18.2 and Corollary 18.15]. Take $\gamma \in \Gamma$, a SGD sequence (x_n^γ) and notice that by Theorem 1 it is $\mathbb{P}^\nu$ almost surely always in $\mathcal{D}_F \cap S_1$, where S_1 is the set of x where $\nabla f(x,s)$ exists for μ-a.e. s. Therefore its Markov kernel can be equivalently defined as:

$$P'_\gamma(x,g) := 1_{\mathcal{D}_F\cap S_1}(x) P_\gamma(x,g) + 1_{(\mathcal{D}_F\cap S_1)^c}(x) g(x - \gamma\varphi(x))\,.$$

Now we can apply [6, Theorem 5.1.] with $h_\gamma(s,x) = -(1_{\mathcal{D}_F\cap S_1}(x)\nabla F(x) + 1_{(\mathcal{D}_F\cap S_1)^c}(x)\varphi(x))$ (note that it is independent of s) and we have $h(x,s) \in H(x,s) = H(x) := -\partial F(x)$. As we show next, [6, Assumption RM] now easily follows.

First, it is immediate from the general properties of the Clarke subdifferential that the set-valued map $-\partial F$ is proper and uppersemicontinuous with convex and compact values, hence the assumption (iii) of [6, Assumption RM]. Assumption (ii) is immediate by the uppersemicontinuity of $-\partial F$. Moreover, we obtain from Assumption 2 that there exists a constant $K \geqslant 0$ such that

$$\|\partial F(x)\| \leqslant K(1 + \|x\|).$$

Thus, $\mathcal{S}_{-\partial F}$ is defined on the whole $\mathbb{R}^d$, and $\mathcal{S}_{-\partial F}$ is closed in $(C(\mathbb{R}_+, \mathbb{R}^d), \boldsymbol{d})$ (see [2]), hence assumption (v). Finally, assumption (vi) comes from Assumption 2.

We remark that although, [6, Theorem 5.1] deals with a family of measures $(\mathbb{P}^a)_{a\in\mathcal{K}}$, the proofs remain unchanged when we consider $(\mathbb{P}^\nu)_{\nu\in\mathcal{M}_{abs}(\mathcal{K})}$.

8.5 Proof of Theorems 3 and 6

Both theorems are proved in the same way. In the following Q_γ will denote either P_γ and in this case H will denote $-\partial F$, or $Q_\gamma = P^r_\gamma$ and $\mathsf{H} = -\partial F - \mathcal{N}_r$. The proof will be done in three steps:

- Lemma 2: Q_γ has a unique invariant probability distribution π_γ, with $\pi_\gamma \in \mathcal{M}_{abs}$ if $Q_\gamma = P_\gamma$ and $\pi_\gamma \in \mathcal{M}^r$ otherwise, moreover Q_γ is ergodic in the sense of the Total Variation norm.
- Lemma 3: The family $\{\pi_\gamma\}_{\gamma\in(0,\gamma_0]}$ is tight.
- Proposition 9: The accumulation points of $\{\pi_\gamma\}_{\gamma\in(0,\gamma_0]}$ as $\gamma \to 0$ are invariant for the DI $\dot{\mathsf{x}} \in \mathsf{H}(\mathsf{x})$.

Before stating Lemma 2, we recall a general result on Markov processes. Let $Q : \mathbb{R}^d \times \mathscr{B}(\mathbb{R}^d) \to [0, 1]$ be a Markov kernel on $\mathbb{R}^d$. A set $B \subset \mathbb{R}^d$ is said to be a small-set for the kernel Q if there exists a positive measure ρ on $\mathbb{R}^d$ such that $Q(x, A) \geqslant \rho(A)$ for each $A \in \mathscr{B}(\mathbb{R}^d)$, $x \in B$.

Proposition 8. *Assume that B is a small set for Q. Furthermore, assume that there exists a measurable function $W : \mathbb{R}^d \to [0, \infty)$ that is defined on $\mathbb{R}^d$ and bounded on B, and a real number $b \geqslant 0$, such that*

$$QW \leqslant W - 1 + b1_B. \tag{30}$$

Then, Q admits a unique invariant probability distribution π, and moreover, the ergodicity result

$$\forall x \in \mathbb{R}^d, \ \|Q^n(x, \cdot) - \pi\|_{TV} \xrightarrow[n\to\infty]{} 0 \tag{31}$$

holds true.

Indeed, by [21, Theorem 11.3.4], the kernel Q is a so-called positive Harris recurrent, meaning among others that it has a unique invariant probability distribution. Moreover, Q is aperiodic, hence the convergence (31), as shown by, *e.g.*, [21, Theorem 13.0.1].

Lemma 2. *Assume that either Assumptions 4-(i) 4-(ii) hold if $Q_\gamma = P_\gamma$ or Assumption 7 holds and $r \leqslant R$ if $Q_\gamma = P^r_\gamma$, then for every $\gamma \in (0, \gamma_0]$, the kernel Q_γ admits a unique invariant measure π_γ. Moreover,*

$$\forall x \in \mathbb{R}^d, \ \left\|Q^n_\gamma(x, \cdot) - \pi_\gamma\right\|_{TV} \xrightarrow[n\to\infty]{} 0. \tag{32}$$

Finally, if $Q_\gamma = P_\gamma$, assumptions of Theorem 1 hold true and $\gamma \in \Gamma$ then π_γ is absolutely continuous w.r.t. the Lebesgue measure. If $Q_\gamma = P_\gamma^r$ and assumptions of Theorem 4 hold true, then $\pi_\gamma \in \mathcal{M}^r$.

Proof. By the inequality (18), the kernel P_γ satisfies an inequality of the type (30), namely, $P_\gamma V \leqslant V - \alpha(\gamma)\theta + C\alpha(\gamma)1_{\|x\|\leqslant R}$, for some $\theta, C > 0$. Similarly, under Assumption 7 and $r \leqslant R$, we have that for every $x \in \mathrm{cl}(B(0,r))$:

$$P_\gamma^r(x, A) = P_\gamma(x, \Pi_r^{-1}(A)) \geqslant \rho_\gamma(\Pi_r^{-1}(A))\,,$$

that is to say $\mathrm{cl}(B(0,r))$ is a small set for P_γ^r. Inequality of the type Assumption 4-(ii)–(iii) then hold for e.g. $C = r$, $\alpha(\gamma) = 1$, $V = \|x\| + r1_{\|x\|>r}$ and $p = \|x\|$.

Consider the case where $Q_\gamma = P_\gamma$, to prove that π_γ is absolutely continuous w.r.t. the Lebesgue measure, consider a λ-null set A. By the convergence (32), we obtain that for any $x \in \mathbb{R}^d$, $P_\gamma^n(x, A) \to \pi_\gamma(A)$. Now take $\nu \ll \lambda$. By Proposition 3, we have that $\nu P_\gamma^n \ll \lambda$. Hence, by the dominated convergence theorem,

$$0 = \nu P_\gamma^n(A) = \int P_\gamma^n(x, A)\nu(\mathrm{d}x) \to \int \pi_\gamma(A)\nu(\mathrm{d}x) = \pi_\gamma(A)\,.$$

If $Q_\gamma = P_\gamma^r$ we obtain the same result with the help of Proposition 7. □

Lemma 3. *Let either Assumptions 4-(i) – 4-(iii) hold if $Q_\gamma = P_\gamma$ or Assumption 7 hold and $r \leqslant R$ if $Q_\gamma = P_\gamma^r$. Let π_γ be the invariant distribution of Q_γ. Then, the family $\{\pi_\gamma : \gamma \in (0, \gamma_0]\}$ is tight.*

Proof. If $Q_\gamma = P_\gamma^r$ then the family π_γ is supported by $\mathrm{cl}(B(0,r))$ and is, therefore, tight. Otherwise we iterate (18), to obtain:

$$\sum_{k=0}^{n} Q_\gamma^{k+1} V \leqslant \sum_{k=0}^{n} Q_\gamma^k V - \alpha(\gamma) \sum_{k=0}^{n} Q_\gamma^k p + C(n+1)\alpha(\gamma)\,.$$

Therefore, since $0 \leqslant Q_\gamma^k V < +\infty$ we have:

$$\alpha(\gamma) \sum_{k=0}^{n} Q_\gamma^k p \leqslant V + C(n+1)\alpha(\gamma)\,.$$

For a fixed $M > 0$ we will bound now $\pi_\gamma(p \wedge M)$. Since π_γ is an invariant distribution for Q_γ, we have $\pi_\gamma P_\gamma^k = \pi_\gamma$. Hence, we have:

$$\begin{aligned}\pi_\gamma(p \wedge M) &= \frac{1}{n+1} \sum_{k=0}^{n} \pi_\gamma Q_\gamma^k(p \wedge M) \leqslant \frac{1}{n+1} \sum_{k=0}^{n} \pi_\gamma(Q_\gamma^k p \wedge M) \\ &\leqslant \pi_\gamma\left(\left[\frac{V}{(n+1)\alpha(\gamma)} + C\right] \wedge M\right)\,.\end{aligned}$$

Letting $n \to +\infty$, by the dominated convergence theorem we obtain $\pi_\gamma(p \wedge M) \leqslant \pi_\gamma(C \wedge M)$. And therefore by monotone convergence theorem $\pi_\gamma(p) \leqslant C$.
Fix now $\varepsilon > 0$, there is a $K > 0$ such that $\frac{C}{K} \leqslant \varepsilon$, and by coercivity of p there is $r > 0$ such that:

$$\pi_\gamma(\|x\| > r) \leqslant \pi_\gamma(p > K) \leqslant \frac{C}{K}$$

where the last bound comes from Markov's inequality. This concludes the proof. □

The next proposition will show that any accumulation point of π_γ is an invariant measure for the set-valued flow induced by the DI $\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t))$, first we introduce some definitions. Define the shift operator $\Theta_t : C(\mathbb{R}_+, \mathbb{R}^d) \to C(\mathbb{R}_+, \mathbb{R}^d)$ by $\Theta_t(x) = x(t + \cdot)$, and the projection operator $p_0 : C(\mathbb{R}_+, \mathbb{R}^d) \to \mathbb{R}^d$ by $p_0(x) = x(0)$. Then, we have the following definition (see [25] for details):

Definition 3. *We say that $\pi \in \mathcal{M}(\mathbb{R}^d)$ is an invariant distribution for the flow induced by the DI $\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t))$, if there is $\nu \in \mathcal{M}(C(\mathbb{R}_+, \mathbb{R}^d))$, such that:*

i) $\operatorname{supp} \nu \in \overline{\mathcal{S}_{\mathsf{H}}(\mathbb{R}^d)}$,

ii) $\nu\Theta_t^{-1} = \nu$,

iii) $\nu p_0^{-1} = \pi$.

Proposition 9. *Let Assumptions 1–3 and 4 hold true. Denote by π_γ the unique invariant distribution of P_γ. Let (γ_n) be a sequence on $(0, \gamma_0] \cap \Gamma$ s.t. $\gamma_n \to 0$ and π_{γ_n} converges narrowly to some probability measure π. Then, π is an invariant distribution for the flow induced by $\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t))$.*

Similarly, under Assumptions 1–2 and 6–7, for $r \leqslant R$, denoting π_γ the unique invariant distribution of P_γ^r, if $\pi_{\gamma_n} \to \pi$, then π is an invariant distribution for the flow induced by $\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t)) - \mathcal{N}_r(\mathsf{x}(t))$.

Proof. Consider the case where $Q_\gamma = P_\gamma$. The proof essentially follows [6, section 7.]. Fix an $\varepsilon > 0$ and write π_n instead of π_{γ_n} for simplicity. By Lemma 3 we have a compact K such that $\pi_n(K) > 1 - \varepsilon$, we thus can define the conditional measures $\pi_n^K(A) := \frac{\pi_n(A \cap K)}{\pi_n(K)}$. Moreover, we have $\pi_n^K \in \mathcal{M}_{abs}(K)$, therefore we can apply Theorem 2 and get that there is a compact set $\mathcal{C}$ of $C(\mathbb{R}^+, \mathbb{R}^d)$ such that $\mathbb{P}^{\pi_{\gamma_n}^K, \gamma_n} \mathsf{X}_{\gamma_n}^{-1}(\mathcal{C}) \geqslant 1 - \varepsilon$. Now we have

$$\mathbb{P}^{\pi_n, \gamma_n}(\cdot) = \int_{\mathbb{R}^d} \mathbb{P}^{a, \gamma_n}(\cdot) \pi_n(\mathrm{d}a) \geqslant \int_K \mathbb{P}^{a, \gamma_n}(\cdot) \pi_n(\mathrm{d}a) \geqslant \pi_n(K) \mathbb{P}^{\pi_n^K, \gamma_n}(\cdot)\,,$$

hence

$$\mathbb{P}^{\pi_{\gamma_n}, \gamma_n} \mathsf{X}_{\gamma_n}^{-1}(\mathcal{C}) \geqslant \pi_n(K) \mathbb{P}^{\pi_{\gamma_n}^K, \gamma_n} \mathsf{X}_{\gamma_n}^{-1}(\mathcal{C}) \geqslant (1 - \varepsilon)^2\,.$$

Since ε is arbitrary this proves the tightness of $v_n := \mathbb{P}^{\pi_{\gamma_n}, \gamma_n} \mathsf{X}_{\gamma_n}^{-1}$. Take $\pi_n \to \pi$ and $v_n \to v \in \mathcal{M}(C(\mathbb{R}_+, \mathbb{R}^d))$. We now prove that v is an invariant distribution for the flow induced by the DI associated to $-\partial F$ (see Definition 3.)
We have $\pi_n = v_n p_0^{-1}$, by continuity of p_0. Thus, $\pi = v p_0^{-1}$. Therefore, we have (iii) of Definition 3. Let $\eta > 0$. By weak convergence of v_n,

$$v(\{x \in C(\mathbb{R}_+, \mathbb{R}^d) : d(x, \mathcal{S}_{-\partial F}(\mathbb{R}^d)) \leqslant \eta\}) \geqslant \limsup_n v_n(\{x \in C(\mathbb{R}_+, \mathbb{R}^d) : d(x, \mathcal{S}_{-\partial F}(\mathbb{R}^d)) \leqslant \eta\})$$

and

$$\begin{aligned} v_n(\{x \in C(\mathbb{R}_+, \mathbb{R}^d) : d(x, \mathcal{S}_{-\partial F}(\mathbb{R}^d)) \leqslant \eta\}) &\geqslant v_n(\{x \in C(\mathbb{R}_+, \mathbb{R}^d) : d(x, \mathcal{S}_{-\partial F}(K)) < \eta\}) \\ &\geqslant \pi_n(K) \mathbb{P}^{\pi_{\gamma_n}^K, \gamma_n}(d(\mathsf{X}^{\gamma_n}, \mathcal{S}_{-\partial F}(K)) < \eta) \\ &\geqslant (1 - \varepsilon) \mathbb{P}^{\pi_{\gamma_n}^K, \gamma_n}(d(\mathsf{X}^{\gamma_n}, \mathcal{S}_{-\partial F}(K)) < \eta)\,. \end{aligned}$$

The last term converges to $1 - \varepsilon$, by Theorem 2, and by weak convergence we have $v(\{x \in C(\mathbb{R}_+, \mathbb{R}^d) : d(x, \mathcal{S}_{-\partial F}(\mathbb{R}^d)) \geqslant \eta\}) \geqslant (1 - \varepsilon)$, now letting $\eta \to 0$, by monotone convergence

we have $v(\mathcal{S}_{-\partial F}(\mathbb{R}^d))) \geqslant 1-\varepsilon$ which proves (i) of Definition 3. Finally, the second point of Definition 3 is shown just like in [6, section 7.].

The proof of the case $Q_\gamma = P_\gamma^r$ is substantially the same under straightforward adaptations. □

After some definitions we recall an important result about the support of a flow-invariant measure. The limit set L_f of a function $f \in C(\mathbb{R}_+, \mathbb{R}^d)$ is

$$L_f = \bigcap_{t \geqslant 0} \overline{f([t,\infty))},$$

and the limit set $L_{\mathcal{S}_\mathsf{H}(a)}$ of a point $a \in \mathbb{R}^d$ for $\mathcal{S}_\mathsf{H}$ is

$$L_{\mathcal{S}_\mathsf{H}(a)} = \bigcup_{\mathsf{x} \in \mathcal{S}_\mathsf{H}(a)} L_\mathsf{x}.$$

A point $a \in \mathbb{R}^d$ is said $\mathcal{S}_\mathsf{H}$-recurrent if $a \in L_{\mathcal{S}_\mathsf{H}(a)}$. The Birkhoff center $\mathrm{BC}_{\mathcal{S}_\mathsf{H}}$ of $\mathcal{S}_\mathsf{H}$ is the closure of the set of its recurrent points:

$$\mathrm{BC}_{\mathcal{S}_\mathsf{H}} = \overline{\left\{ a \in \mathbb{R}^d \ : \ a \in L_{\mathcal{S}_\mathsf{H}(a)} \right\}}.$$

In [13] (see also [3]), a version of Poincaré's recurrence theorem, well-suited for our set-valued evolution systems, was provided:

Proposition 10. *Each invariant measure for $\mathcal{S}_\mathsf{H}$ is supported by $BC_{\mathcal{S}_\mathsf{H}}$.*

With the help of Proposition 10 we can finally prove Theorem 3.

Proof. Take $\gamma \in \Gamma$, $\varepsilon > 0$ and (x_n^γ) an associated SGD sequence. We have by (31):

$$\limsup_{n \to \infty} \mathbb{P}^\nu \left[\mathrm{dist}(x_n^\gamma, \mathcal{Z}) > \varepsilon \right] = \pi_\gamma(\{x \in \mathbb{R}^d : d(x, \mathcal{Z}) > \varepsilon\}).$$

Now take any sequence $\gamma_i \to 0$ with $\gamma_i \in \Gamma$, and π_{γ_i} the associated invariant distribution, we know from Lemmas 3-9 that we can extract a subsequence such that $\pi_{\gamma_i} \to \pi$, with π an invariant measure for the evolution system $\mathcal{S}_{-\partial F}$. Therefore by weak convergence we have:

$$\begin{aligned} \lim_{i \to +\infty} \pi_{\gamma_i}(\{x \in \mathbb{R}^d : d(x, \mathcal{Z}) > 2\varepsilon\}) &\leqslant \lim_{i \to +\infty} \pi_{\gamma_i}(\{x \in \mathbb{R}^d : d(x, \mathcal{Z}) \geqslant \varepsilon\}) \\ &\leqslant \pi(\{x \in \mathbb{R}^d : d(x, \mathcal{Z}) \geqslant \varepsilon\}), \end{aligned}$$

where the last line comes from the Portmanteau theorem. We show that $\operatorname{supp} \pi \subset S$, and therefore the last term is equal to zero, which concludes the proof. To that end, we make use of Proposition 10, that shows that each invariant measure of $\mathcal{S}_{-\partial F}$ is supported by $\mathrm{BC}_{\mathcal{S}_{-\partial F}}$. Thus, it remains to show that $\mathrm{BC}_{\mathcal{S}_{-\partial F}} = \mathcal{Z}$ (which at the same time will ensure us that $\mathcal{Z}$ is nonempty). It is obvious that $\mathcal{Z} \subset \mathrm{BC}_{\mathcal{S}_{-\partial F}}$. To show the reverse inclusion, take $a \in L_{\mathcal{S}_{-\partial F}(a)}$. Then, there exists a solution x to the differential inclusion such that $\mathsf{x}(0) = a$ and $a \in L_\mathsf{x}$. But under Assumption 5 it holds ([10, lemma 5.2]) that $\|\dot{\mathsf{x}}(t)\| = \|\partial_0 F(\mathsf{x}(t))\|$ almost everywhere, and, moreover,

$$\forall t \geqslant 0, \quad F(\mathsf{x}(t)) - F(\mathsf{x}(0)) = -\int_0^t \|\partial_0 F(\mathsf{x}(u))\|^2 du.$$

Therefore $\mathsf{x}(t) = a$ for each $t \geqslant 0$, thus, $a \in S$. Observing that $\mathcal{Z}$ is a closed set (since ∂F is graph-closed, see [9, Proposition 2.1.5]), we obtain that $\mathrm{BC}_{\mathcal{S}_{-\partial F}} = \mathcal{Z}$.

Similarly, take $\gamma_i \to 0$ and and $(x_n^{\gamma_i, r})$ the associated projected SGD sequences. After an extraction we get that $\pi_{\gamma_i} \to \pi$, with π an invariant measure for the flow $\mathcal{S}_{-\partial F - \mathcal{N}_r}$ and:

$$\lim_{\gamma_i \to 0} \limsup_{n \to \infty} \mathbb{P}^{\nu} \left[\mathrm{dist}(x_n^{\gamma_i, r}, \mathcal{Z}_r) > 2\varepsilon \right] \leqslant \pi(\{x \in \mathbb{R}^d : d(x, \mathcal{Z}_r) > \varepsilon\}) \,.$$

Taking $a \in L_{\mathcal{S}_{-\partial F - \mathcal{N}_r}(a)}$, and x a solution to the associated differential inclusion with $\mathsf{x}(0) = a$, we get under Assumption 5 [10, Lemma 6.3.] that $\|\dot{\mathsf{x}}(t)\| = \min\{\|v\| : v \in \partial F(\mathsf{x}(t)) + \mathcal{N}_r(\mathsf{x}(t))\}$, and moreover,

$$\forall t \geqslant 0, \quad F(\mathsf{x}(t)) - F(\mathsf{x}(0)) = -\int_0^t \|\dot{\mathsf{x}}(u)\|^2 \, \mathrm{d}u \,.$$

That is to say $\mathsf{x}(t) = a$ and $a \in \mathcal{Z}_r$, which finishes the proof. □

8.6 Proof of Proposition 5

Denote as ρ the probability distribution of the random variable $\gamma\epsilon_1$. By assumption, ρ has a continuous density that is positive at each point of $\mathbb{R}^d$. We denote as f this density. Let θ_x be the probability distribution of the random variable $Z = x - \gamma\varphi_0(x, \xi_1)$, which is the image of μ by the function $x - \gamma\varphi_0(x, \cdot)$. Our purpose is to show that

$$\exists \varepsilon > 0, \ \forall x \in \mathrm{cl}(B(0,R)), \ \forall A \in \mathscr{B}(\mathbb{R}^d), \ (\theta_x \otimes \rho)\left[Z + \gamma\eta_1 \in A\right] \geqslant \varepsilon \, \lambda(A \cap \mathrm{cl}(B(0,1))).$$

Given $L > 0$, we have by Assumption 2 and Markov's inequality that there exists a constant $K > 0$ such that

$$\theta_x \left[Z \notin \mathrm{cl}(B(0,L))\right] \leqslant \frac{K}{L}(1 + \|x\|).$$

Thus, taking L large enough, we obtain that $\forall x \in \mathrm{cl}(B(0,R)), \ \theta_x \left[Z \notin \mathrm{cl}(B(0,L))\right] < 1/2$. Moreover, we can always choose $\varepsilon > 0$ is such a way that $f(u) \geqslant 2\varepsilon$ for $u \in \mathrm{cl}(B(0, L+1))$, by the continuity and the positivity of f on the compact $\mathrm{cl}(B(0, L+1))$. Thus,

$$\begin{aligned}
(\theta_x \otimes \rho)\left[Z + \gamma\eta_1 \in A\right] &= \int_A du \int_{\mathbb{R}^d} \theta_x(dv) \, f(u - v) \\
&\geqslant \int_{A \cap \mathrm{cl}(B(0,1))} du \int_{\mathrm{cl}(B(0,L))} \theta_x(dv) \, f(u - v) \\
&\geqslant 2\varepsilon \int_{A \cap \mathrm{cl}(B(0,1))} du \int_{\mathrm{cl}(B(0,L))} \theta_x(dv) \\
&\geqslant \varepsilon \, \lambda(A \cap \mathrm{cl}(B(0,1))).
\end{aligned}$$

8.7 Proof of Proposition 6

By Lebourg's mean value theorem [9, Theorem 2.4], for each $n \in \mathbb{N}$, there exists $\alpha_n \in [0,1]$ and $\zeta_n \in \partial F(u_n)$ with $u_n = x_n - \alpha_n \gamma \nabla f(x_n, \xi_{n+1}) 1_{\Delta_f}(x_n, \xi_{n+1})$, such that

$$F(x_{n+1}) = F(x_n) - \gamma \langle \zeta_n, \nabla f(x_n, \xi_{n+1}) \rangle 1_{\Delta_f}(x_n, \xi_{n+1}),$$

and the proof of this theorem (see [9, Theorem 2.4] again) shows that u_n can be chosen measurably as a function of (x_n, ξ_{n+1}).

In the following, for the ease of readability, we make use of shorthand (and abusive) notations of the type $1_{\|x\|>2R}\langle \nabla F(x), \cdots \rangle$ to refer to $\langle \nabla F(x), \cdots \rangle$ if $\|x\| > 2R$ and to zero if not. We also denote $\nabla f(x_n, \xi_{n+1})$ as ∇f_{n+1} to shorten the equations. We write

$$\begin{aligned} F(x_{n+1}) = F(x_n) &- \gamma 1_{\|x_n\|\leqslant 2R}\langle \zeta_n, \nabla f_{n+1}\rangle 1_{\Delta_f}(x_n, \xi_{n+1}) \\ &- \gamma 1_{\|x_n\|>2R}\langle \zeta_n - \nabla F(x_n), \nabla f_{n+1}\rangle - \gamma 1_{\|x_n\|>2R}\langle \nabla F(x_n), \nabla f_{n+1}\rangle. \end{aligned}$$

We shall prove that

$$\begin{aligned} \mathbb{E}_n F(x_{n+1}) \leqslant F(x_n) &- \gamma 1_{\|x_n\|>2R}\|\nabla F(x_n)\|^2 + \gamma K 1_{\|x_n\|\leqslant 2R} \\ &+ \gamma^2 K 1_{\|x_n\|>2R}\left((1+\|\nabla F(x_n)\|)\Big(\int \|\nabla f(x_n, s)\|^2\,\mu(ds)\Big)^{1/2} + \int \|\nabla f(x_n, s)\|^2\,\mu(ds)\right) \end{aligned} \tag{33}$$

where the constant $K > 0$ is an absolute finite constant that can change from line to line in the derivations below. To that end, we write

$$\begin{aligned} F(x_{n+1}) = F(x_n) &- \gamma 1_{\|x_n\|\leqslant 2R} 1_{\|u_n\|\leqslant R}\langle \zeta_n, \nabla f_{n+1}\rangle 1_{\Delta_f}(x_n, \xi_{n+1}) \\ &- \gamma 1_{\|x_n\|\leqslant 2R} 1_{\|u_n\|>R}\langle \zeta_n, \nabla f_{n+1}\rangle 1_{\Delta_f}(x_n, \xi_{n+1}) \\ &- \gamma 1_{\|x_n\|>2R} 1_{\|u_n\|\leqslant R}\langle \zeta_n - \nabla F(x_n), \nabla f_{n+1}\rangle \\ &- \gamma 1_{\|x_n\|>2R} 1_{\|u_n\|>R}\langle \nabla F(u_n) - \nabla F(x_n), \nabla f_{n+1}\rangle \\ &- \gamma 1_{\|x_n\|>2R}\langle \nabla F(x_n), \nabla f_{n+1}\rangle \end{aligned} \tag{34}$$

We start with the second term at the right hand side of this inequality. Noting from Assumption 2 that

$$1_{\|u_n\|\leqslant R}\|\zeta_n\| \leqslant \sup_{\|x\|\leqslant R}\|\partial F(x)\| \leqslant \sup_{\|x\|\leqslant R}\int \|\partial f(x,s)\|\,\mu(ds) \leqslant \sup_{\|x\|\leqslant R}\int \kappa(x,s)\,\mu(ds) \leqslant K,$$

we have

$$\gamma 1_{\|x_n\|\leqslant 2R} 1_{\|u_n\|\leqslant R}|\langle \zeta_n, \nabla f(x_n, \xi_{n+1})\rangle| \leqslant \gamma K 1_{\|x_n\|\leqslant 2R}\|\nabla f_{n+1}\|,$$

and by integrating with respect to ξ_{n+1} and using Assumption 2 again, we get that

$$\gamma 1_{\|x_n\|\leqslant 2R}\mathbb{E}_n[1_{\|u_n\|\leqslant R}|\langle \zeta_n, \nabla f_{n+1}\rangle 1_{\Delta_f}(x_n, \xi_{n+1})|] \leqslant \gamma K 1_{\|x_n\|\leqslant 2R}. \tag{35}$$

Using Assumption 2, the next term at the right hand side of (34) can be bounded as

$$\begin{aligned} &\gamma 1_{\|x_n\|\leqslant 2R} 1_{\|u_n\|>R}|\langle \zeta_n, \nabla f_{n+1}\rangle 1_{\Delta_f}(x_n, \xi_{n+1})| \\ &\leqslant \gamma 1_{\|x_n\|\leqslant 2R} 1_{\|u_n\|>R}\|\nabla F(u_n)\|\,\|\nabla f_{n+1}\| \\ &\leqslant \gamma 1_{\|x_n\|\leqslant 2R} K\left(1 + \|x_n\| + \gamma\|\nabla f_{n+1}\|\right)\|\nabla f_{n+1}\| \\ &\leqslant \gamma K 1_{\|x_n\|\leqslant 2R}\left(1 + \|\nabla f_{n+1}\| + \gamma\|\nabla f_{n+1}\|^2\right), \end{aligned}$$

which leads to

$$\gamma 1_{\|x_n\|\leqslant 2R}\mathbb{E}_n[1_{\|u_n\|>R}|\langle \zeta_n, \nabla f_{n+1}\rangle 1_{\Delta_f}(x_n, \xi_{n+1})|] \leqslant \gamma K 1_{\|x_n\|\leqslant 2R} \tag{36}$$

by using Assumption 2.

We tackle the next term at the right hand side of (34). Fix a $x_\star \notin \mathrm{cl}(B(0,R))$. By our assumptions it holds that each $x \notin \mathrm{cl}(B(0,R))$,

$$\|\nabla f(x,s)\| \leqslant \|\nabla f(x_\star,s)\| + \beta(s)\|x - x_\star\| \leqslant \beta'(s)(1+\|x\|),$$

where $\beta'(\cdot)$ is square integrable thanks to Assumption 2. Since

$$\int \beta'(s)^2 \mu(ds) = \int_0^\infty \mu[\beta'(\cdot) \geqslant \sqrt{t}\ ]\ dt < \infty,$$

it holds that $\mu[\beta'(\cdot) \geqslant 1/t\ ] = o_{t\to 0}(t^2)$. Using triangle inequality, we get that

$$\begin{aligned} 1_{\|x_n\|>2R} 1_{\|u_n\|\leqslant R} &= 1_{\|x_n\|>2R} 1_{\|x_n - \alpha_n\gamma\nabla f_{n+1}\|\leqslant R} \leqslant 1_{\|x_n\|>2R} 1_{\|\nabla f_{n+1}\|\geqslant (\|x_n\|-R)/\gamma} \\ &\leqslant 1_{\|x_n\|>2R} 1_{\beta'(\xi_{n+1})\geqslant \frac{\|x_n\|-R}{\gamma(1+\|x_n\|)}} \leqslant 1_{\|x_n\|>2R} 1_{\beta'(\xi_{n+1})\geqslant \frac{R}{\gamma(1+2R)}}. \end{aligned}$$

Using this result, we write

$$\begin{aligned} \gamma 1_{\|x_n\|>2R} 1_{\|u_n\|\leqslant R} |\langle \zeta_n, \nabla f_{n+1}\rangle| &\leqslant K\gamma 1_{\|x_n\|>2R} 1_{\|u_n\|\leqslant R} \|\nabla f_{n+1}\| \\ &\leqslant K\gamma 1_{\|x_n\|>2R} \|\nabla f_{n+1}\| 1_{\beta'(\xi_{n+1})\geqslant \frac{R}{\gamma(1+2R)}} \end{aligned}$$

Consequently,

$$\begin{aligned} \gamma 1_{\|x_n\|>2R} \mathbb{E}_n[1_{\|u_n\|\leqslant R} |\langle \zeta_n, \nabla f_{n+1}\rangle|] &\leqslant \gamma K 1_{\|x_n\|>2R} \Big(\int \|\nabla f(x_n,s)\|^2 \mu(ds)\Big)^{1/2} \mu[\beta'(\cdot) \geqslant K/\gamma]^{1/2} \\ &\leqslant \gamma^2 K 1_{\|x_n\|>2R} \Big(\int \|\nabla f(x_n,s)\|^2 \mu(ds)\Big)^{1/2}. \end{aligned} \tag{37}$$

Similarly,

$$\gamma 1_{\|x_n\|>2R} 1_{\|u_n\|\leqslant R} |\langle \nabla F(x_n), \nabla f_{n+1}\rangle| \leqslant \gamma K 1_{\|x_n\|>2R} \|\nabla F(x_n)\| \, \|\nabla f_{n+1}\| 1_{\beta'(\xi_{n+1})\geqslant \frac{R}{\gamma(1+2R)}},$$

thus,

$$\gamma 1_{\|x_n\|>2R} \mathbb{E}_n \left[1_{\|u_n\|\leqslant R} |\langle \nabla F(x_n), \nabla f_{n+1}\rangle|\right] \leqslant \gamma^2 K 1_{\|x_n\|>2R} \|\nabla F(x_n)\| \Big(\int \|\nabla f(x_n,s)\|^2 \mu(ds)\Big)^{1/2}. \tag{38}$$

We have that ∇F is Lipschitz outside $\mathrm{cl}(B(0,R))$. Thus, the next to last term at the right hand side of (34) satisfies

$$\gamma 1_{\|x_n\|>2R} 1_{\|u_n\|>R} |\langle \nabla F(u_n) - \nabla F(x_n), \nabla f_{n+1}\rangle| \leqslant \gamma^2 K 1_{\|x_n\|>2R} \|\nabla f_{n+1}\|^2,$$

and we get that

$$\gamma 1_{\|x_n\|>2R} 1_{\|u_n\|>R} \mathbb{E}_n \left[|\langle \nabla F(u_n) - \nabla F(x_n), \nabla f_{n+1}\rangle|\right] \leqslant \gamma^2 K 1_{\|x_n\|>2R} \int \|\nabla f(x_n,s)\|^2 \mu(ds). \tag{39}$$

Finally, we have

$$-\gamma 1_{\|x_n\|>2R} \mathbb{E}_n \left[\langle \nabla F(x_n), \nabla f_{n+1}\rangle\right] = -\gamma 1_{\|x_n\|>2R} \|\nabla F(x_n)\|^2. \tag{40}$$

Inequalities (35)–(40) lead to (33).

Using Assumption (iii) of Proposition 6, Inequality (33) leads to Inequality (20). The validity of Assumptions 4-(ii) and 4-(iii) can then be checked easily.

8.8 Proof of Proposition 7

The next Lemma is the key ingredient in the proofs of Section 7.

Lemma 4. *Assume that $f(\cdot, s)$ is locally Lipschitz continuous for every $s \in \Xi$. Then for $\lambda^1 \otimes \lambda^d \otimes \mu$-almost all (r, x, s) with $r > 0$, it holds that $(\Pi_r(x), s) \in \Delta_f$. For $\lambda^1 \otimes \lambda^d$-almost all (r, x) with $r > 0$, it holds that $\Pi_r(x) \in \mathcal{D}_F$.*

Proof. Our first aim is to show that

$$\int 1_{\Delta_f^c}(\Pi_r(x), s)\, \lambda^1(\mathrm{d}r) \otimes \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s) = 0\,. \tag{41}$$

First, note by Fubini's theorem that

$$0 = \int 1_{\Delta_f^c}(x, s)\, \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s) = \int_{\Xi \times \mathbb{R}_+} \int_{\mathbb{S}(1)} 1_{\Delta_f^c}(r\theta, s) r^{d-1} \varrho_1(\mathrm{d}\theta)\ \mu \otimes \lambda^1(\mathrm{d}s \times \mathrm{d}r)\,, \tag{42}$$

that is to say, $\varrho(\{\theta : (r\theta, s) \in \Delta_f\}) = 0$ for $\mu \otimes \lambda^1$ almost every (s, r) with $r > 0$. Decompose Equation (41) as

$$\begin{aligned}
&\int 1_{\Delta_f^c}(\Pi_r(x), s)\ \lambda^1(\mathrm{d}r) \otimes \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s) \\
&= \int 1_{\|x\| \geqslant r} 1_{\Delta_f^c}(\Pi_r(x), s)\ \lambda^1(\mathrm{d}r) \otimes \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s) + \int 1_{\|x\| < r} 1_{\Delta_f^c}(x, s)\ \lambda^1(\mathrm{d}r) \otimes \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s).
\end{aligned}$$

Since for each s, $f(\cdot, s)$ is differentiable almost everywhere, we have by Fubini's theorem:

$$\int 1_{\|x\| < r} 1_{\Delta_f^c}(x, s)\ \lambda^1(\mathrm{d}r) \otimes \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s) = 0.$$

Similarly,

$$\begin{aligned}
&\int 1_{\|x\| \geqslant r} 1_{\Delta_f^c}(\Pi_r(x), s)\ \lambda^1(\mathrm{d}r) \otimes \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s) \\
&= \int 1_{\|x\| \geqslant r} 1_{\Delta_f^c}\left(\frac{rx}{\|x\|}, s\right)\ \lambda^1(\mathrm{d}r) \otimes \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s) \\
&= \int_{\mathbb{R}_+} \int_{\Xi \times \mathbb{R}_+} \int_{\mathbb{S}(1)} 1_{r' \geqslant r} 1_{\Delta_f^c}(r'\theta, s)(r')^{d-1} \varrho(\mathrm{d}\theta)\ \mu \otimes \lambda^1(\mathrm{d}s \times \mathrm{d}r)\ \lambda^1(\mathrm{d}r') \\
&= 0\,,
\end{aligned}$$

with the last equality coming from Equation (42). Hence (41). The second statement can be proven along similar lines. □

Consider $r > 0$ such that the conclusion of Lemma 4 hold. Then the almost sure equality of all projected SGD sequence is proven in the same way as in Proposition 3. We can therefore consider the lazy projected SGD sequence $x_{n+1}^{\gamma,r} = \Pi_r(x_n^{\gamma,r} - \gamma\varphi_0(x_n^{\gamma,r}, \xi_{n+1}))$. By Assumption 6 the law of $x_{n+1/2}^{\gamma,r} := x_n^{\gamma,r} - \gamma\varphi_0(x_n^{\gamma,r}, \xi_{n+1})$ is Lebesgue-absolutely continuous. Take A a borel set of $\mathbb{R}^d$ such that $\lambda(A) = \varrho_r(A) = 0$. Then

$$\mathbb{P}(x_{n+1}^{\gamma,r} \in A) \leqslant \mathbb{P}(x_{n+1/2}^{\gamma,r} \in A) + \mathbb{P}\left(r \frac{x_{n+1/2}^{\gamma,r}}{\left\| x_{n+1/2}^{\gamma,r} \right\|} \in A \right).$$

The first term is equal to zero by Lebesgue-absolutely continuity of the law of $x_{n+1/2}^{\gamma,r}$. For the second term we write:

$$\mathbb{P}\left(r \frac{x_{n+1/2}^{\gamma,r}}{\left\| x_{n+1/2}^{\gamma,r} \right\|} \in A \right) = \int (r')^{d-1} 1_A(r\theta) \varrho(\mathrm{d}\theta) \lambda^1(\mathrm{d}r') = \int (r')^{d-1} \varrho_r(A) \lambda^1(\mathrm{d}r') = 0\,,$$

which finishes the proof.

8.9 Proof of Theorems 4 and 5

Noting that the law of $x_n^{\gamma,r} - \gamma\varphi_0(x_n^{\gamma,r}, \xi_{n+1})$ is Lebesgue-absolutely continuous by Assumption 6, the first point of Theorem 4 comes from Lemma 4. The second point comes upon noticing that $\Pi_r(x) - x \in -\mathcal{N}_r(\Pi_r(x))$.

Theorem 5 is proved in the same way as Theorem 2, by applying [6, Theorem 5.1.] with $h(s,x) = -\nabla F(x) - 1/\gamma(x - \gamma\nabla f(x,s) - \Pi_r(x - \gamma\nabla f(x,s))) \in -\nabla F(x) - \mathcal{N}_r(x - \gamma\nabla f(x,s))$ and $H(x) = H(s,x) = -\partial F(x) - \mathcal{N}_r(x)$.

Acknowledgements

The authors wish to thank Jérôme Bolte and Edouard Pauwels for their inspiring remarks. This work is partially supported by the Région Ile-de-France.

References

[1] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: a Hitchhiker's Guide*. Springer, Berlin; London, 2006.

[2] J.-P. Aubin and A. Cellina. *Differential inclusions*, volume 264 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1984. Set-valued maps and viability theory.

[3] J.-P. Aubin, H. Frankowska, and A. Lasota. Poincaré's recurrence theorem for set-valued dynamical systems. *Ann. Polon. Math.*, 54(1):85–91, 1991.

[4] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM J. Control Optim.*, 44(1):328–348 (electronic), 2005.

[5] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.

[6] P. Bianchi, W. Hachem, and A. Salim. Constant step stochastic approximations involving differential inclusions: stability, long-run convergence and applications. *Stochastics*, 91(2):288–320, 2019.

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

[8] J. Bolte and E. Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning. *arXiv preprint arXiv:1909.10300*, 2019.

[9] F. H. Clarke, Yu. S. Ledyaev, R. J. Stern, and P. R. Wolenski. *Nonsmooth analysis and control theory*, volume 178 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1998.

[10] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Found Comput Math*, (20):119–154, 2020.

[11] Y. M. Ermoliev and VI Norkin. Solution of nonconvex nonsmooth stochastic optimization problems. *Cybernetics and Systems Analysis*, 39(5):701–715, 2003.

[12] Y.M. Ermoliev and V.I. Norkin. Stochastic generalized gradient method for solving nonconvex nonsmooth stochastic optimization problems. *Cybernetics and Systems Analysis*, 34(2):196–215, June 1998.

[13] M. Faure and G. Roth. Ergodic properties of weak asymptotic pseudotrajectories for set-valued dynamical systems. *Stoch. Dyn.*, 13(1):1250011, 23, 2013.

[14] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

[15] R. Z. Has'minskiĭ. The averaging principle for parabolic and elliptic differential equations and Markov processes with small diffusion. *Teor. Verojatnost. i Primenen.*, 8:3–25, 1963.

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

[17] S. Kakade and J. D. Lee. Provably correct automatic sub-differentiation for qualified programs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7125–7135. Curran Associates, Inc., 2018.

[18] H. J. Kushner and G. G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

[19] G. Lebourg. Generic differentiability of Lipschitzian functions. *Transactions of the American Mathematical Society*, 256:125–144, 1979.

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

[21] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

[22] VS Mikhalevich, AM Gupal, and VI Norkin. Methods of nonconvex optimization. *Nauka*, 1987.

[23] V.I. Norkin. Generalized-differentiable functions. *Cybernetics and Systems Analysis*, 16:10–12, 01 1980.

[24] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017.

[25] G. Roth and W. H. Sandholm. Stochastic approximations with constant step size and differential inclusions. *SIAM J. Control Optim.*, 51(1):525–555, 2013.

[26] Andrzej Ruszczyński. Convergence of a stochastic subgradient method with averaging for nonsmooth nonconvex constrained optimization. *Optimization Letters*, 14, 10 2020.

[27] L. van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Math. J.*, 84(2):497–540, 08 1996.