



HAL
open science

Stochastic proximal subgradient descent oscillates in the vicinity of its accumulation set

Sholom Schechtman

► **To cite this version:**

Sholom Schechtman. Stochastic proximal subgradient descent oscillates in the vicinity of its accumulation set. Optimization Letters, 2022, 10.1007/s11590-022-01884-8 . hal-03676675

HAL Id: hal-03676675

<https://hal.science/hal-03676675v1>

Submitted on 24 May 2022

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

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

Stochastic Proximal Subgradient Descent Oscillates in the Vicinity of its Accumulation Set

Sholom Schechtman

March 2021

Abstract

We analyze the stochastic proximal subgradient descent in the case where the objective functions are path differentiable and verify a Sard-type condition. While the accumulation set may not be reduced to unique point, we show that the time spent by the iterates to move from one accumulation point to another goes to infinity. An oscillation-type behavior of the drift is established. These results show a strong stability property of the proximal subgradient descent. Using the theory of closed measures, Bolte, Pauwels and Ríos-Zertuche [6] established this type of behavior for the deterministic subgradient descent. Our technique of proof relies on the classical works on stochastic approximation of differential inclusions, which allows us to extend results in the deterministic case to a stochastic and proximal setting, as well as to treat these different cases in a unified manner.

1 Introduction

Let d be a positive integer, let \mathcal{X} be a nonempty, closed and convex set and let $f, g : \mathbb{R}^d \rightarrow \mathbb{R}$ be two locally Lipschitz functions. In this note, we study the behavior of the stochastic proximal subgradient descent (SPGD):

$$x_{n+1} \in \text{prox}_{g, \mathcal{X}}^{\gamma_n}(x_n - \gamma_n v_n + \gamma_n \eta_{n+1}), \quad (1)$$

where $\text{prox}_{g, \mathcal{X}}^{\gamma_n}$ is the proximal operator for the function g on \mathcal{X} (see Eq. (10) for a definition), (γ_n) is a sequence of stepsizes, (η_n) is a noise sequence and for each $n \in \mathbb{N}$, v_n is in the set $\partial f(x_n)$ of Clarke's subgradients of f at x_n .

Let $\mathcal{N}_{\mathcal{X}}(x)$ be the normal cone of \mathcal{X} at x . It is known (see [10], [13]) that, under mild conditions on f , g and (η_n) , every limit point of (x_n) is included in the set $\mathcal{Z} := \{x : 0 \in \partial f(x) + \partial g(x) + \mathcal{N}_{\mathcal{X}}(x)\}$. The proof leans on the seminal paper of Benaïm, Hofbauer and Sorin [3] (see also [2], [11]), which analyzes Eq. (1) as an Euler-like discretization of the differential inclusion (DI):

$$\dot{x}(t) \in -\partial f(x(t)) - \partial g(x(t)) - \mathcal{N}_{\mathcal{X}}(x(t)). \quad (2)$$

While the sequence (x_n) is known to converge to \mathcal{Z} , recent work [14] shows that in principle, it might not converge to a unique point. In [14, Section 2] Ríos-Zertuche considers the deterministic subgradient descent (that is to say $g = 0$, $\mathcal{X} = \mathbb{R}^d$, $\eta_n = 0$) and constructs f , which verifies main assumptions of nonsmooth optimization (such as Whitney stratifiability or Kurdyka-Łojasiewicz inequality) but the limit set of (x_n) is equal to $\mathcal{Z} = \{x : \|x\| = 1\}$. This encourages a more precise study of Eq. (1).

In [6] the authors, using the theory of closed measures, show that in the case of the deterministic subgradient descent the convergence to \mathcal{Z} arises in a structured manner. First, they prove that if x, y are two distinct accumulation points of (x_n) , then the time that the iterates spend to get from a neighborhood of x to a neighborhood of y goes to infinity. Second, in a first approximation their results imply that if x is an accumulation point of (x_n) , then

$$\frac{\sum_{i=1}^n \gamma_i v_i \mathbb{1}_{x_i \in B(x, \delta)}}{\sum_{i=1}^n \gamma_i \mathbb{1}_{x_i \in B(x, \delta)}} \xrightarrow{n \rightarrow +\infty} 0,$$

(see [6, Th. 7] or Section 3 for a precise statement). Intuitively speaking, this means that even if $x_n - x_0 = \sum_{i=0}^n \gamma_i v_i$ does not converge, on average, the drift coming from the subgradients compensate itself and vanishes at infinity. This behavior captures an oscillation phenomenon of the iterates around the critical set. Results of this type show a strong stability property of the deterministic subgradient descent.

In practical settings, when the function f is either unknown or computation of its gradient is expensive, the deterministic gradient descent is often replaced by its stochastic version, in many cases, this may lead to a faster convergence (see e.g. [7]). Proximal methods, on the other hand, along with the regularizer function g , are widely used to regularize the initial problem of minimizing f . Depending on the choice of g , we can, for instance, preserve the boundedness of the iterates [11] or promote the sparsity of solutions [16]. It is therefore interesting to establish stability results of the type [6] for the SPGD.

In this work we investigate further the questions of oscillations of the SPGD. Our contributions are threefold. First, we show that the time spent by the SPGD to move from one accumulation point to another goes to infinity. Second, we establish an oscillation-type behavior of the drift. These two results extend [6, Th. 7.] to a stochastic and a proximal setting. Finally, our technique of proof doesn't rely on the theory of closed measures used in [6] but is build upon the classical works on stochastic approximation of differential inclusions ([10], [11], [3]). We feel that this approach gives a simpler proof and allows us to treat the deterministic, the stochastic and the proximal cases in a unified manner.

Paper organisation. In Section 2, we recall some known facts about Clarke subgradient, path differentiable functions and differential inclusions. Our main results are given in Section 3. Section 4 is devoted to proofs.

2 Preliminaries

2.1 Notations

For $S \subset \mathbb{R}^d$, we denote $\text{cl } S$ its closure and $\text{conv } S$ its closed convex hull. For a function $F : \mathbb{R}^d \rightarrow \mathbb{R}$, we denote ∇F its gradient. Constants will usually be denoted as $C, C_1, C_2 \dots$, they can change from line to line. For a sequence (x_n) , we denote $\text{acc}\{x_n\}$ its set of accumulation points. The space of continuous functions from \mathbb{R}_+ to \mathbb{R}^d will be denoted as $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$, we endow this set with \mathbf{d} the metric of uniform convergence on compact intervals, that is to say

$$\mathbf{d}(x_n, y) \xrightarrow{n \rightarrow +\infty} 0 \iff \forall T > 0, \sup_{h \in [0, T]} \|x_n(h) - y(h)\| \xrightarrow{n \rightarrow +\infty} 0. \quad (3)$$

2.2 Perturbed solutions of differential inclusions

We say that $H : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is a set valued map if for each $x \in \mathbb{R}^d$ we have that $H(x)$ is a subset of \mathbb{R}^d . Consider the DI:

$$\dot{x}(t) \in H(x(t)). \quad (4)$$

We say that an absolutely continuous curve (a.c.) $x : \mathbb{R}_+ \rightarrow \mathbb{R}^d$ is a solution to (4) starting at $x \in \mathbb{R}^d$, if $x(0) = x$ and Eq. (4) holds for almost every $t \geq 0$, we denote $S_{H(x)}$ the set of these solutions and $S_H = \cup_{x \in \mathbb{R}^d} S_H(x)$. We say that H is *upper semi continuous at a point* $x \in \mathbb{R}^d$ if for every U a neighborhood of $H(x)$, there is $\delta > 0$ such that $\|y - x\| \leq \delta \implies H(y) \subset U$. We say that H is *upper semi continuous* (usc) if it is upper semicontinuous at every point. We have the following existence result.

Theorem 1 ([1]). *Assume that, for each x in \mathbb{R}^d , $H(x)$ is nonempty, convex and compact, and there is a constant $C \geq 0$ s.t. $\sup\{\|v\| : v \in H(x)\} \leq C(1 + \|x\|)$. Assume that H is usc, then for every $x \in \mathbb{R}^d$, the set $S_{H(x)}$ is nonempty.*

For a set-valued map H and $\delta > 0$, we denote $H^\delta(x) = \{v \in H(y) : \|y - x\| \leq \delta\}$. In this work we will be interested in perturbed solutions of the DI associated to H .

Definition 1 ([3]). *Assume that $H = \sum_{i=1}^l H_i$, where each H_i is a set valued map from \mathbb{R}^d to \mathbb{R}^d . We say that an a.c. curve $X \in C(\mathbb{R}_+, \mathbb{R}^d)$ is a perturbed solution of the DI associated to H if the following holds.*

- i) *There is a function $b : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ and a locally integrable function $\rho : \mathbb{R}_+ \rightarrow \mathbb{R}^d$ s.t. for almost every $t \geq 0$, we have:*

$$\dot{X}(t) - \rho(t) \in \sum_{i=1}^l H_i^{b(t)}(X(t)).$$

- ii) $\lim_{t \rightarrow +\infty} b(t) = 0$.

- iii) *For every $T > 0$, we have:*

$$\lim_{t \rightarrow +\infty} \sup_{0 \leq h \leq T} \left\| \int_t^{t+h} \rho(u) du \right\| = 0.$$

The following theorem states that, in some sense, when t goes to infinity a perturbed solution shadows a solution of the corresponding DI. Its proof can be found in [3, Proof of Th. 4.2].

Theorem 2. *Let X be a perturbed solution associated with $H = \sum_{i=1}^l H_i$ and assume that each of the H_i satisfies the assumptions of Th. 1. Then the family $\{X(t + \cdot) : t \in \mathbb{R}_+\}$ is relatively compact (in $C(\mathbb{R}_+, \mathbb{R}^d)$) and we have:*

$$\lim_{t \rightarrow +\infty} d(X(t + \cdot), S_H) = 0$$

Remark 1. *Strictly speaking, in [3], a perturbed solution to the DI $\dot{x}(t) \in H(x(t))$ was required to satisfy $\dot{X}(t) - \rho(t) \in H^{b(t)}(X(t))$, where $H = \sum_{i=1}^l H_i$. Nevertheless, the proof of [3, Th. 4.2] goes through with our definition.*

Given a set \mathcal{A} , we say that a continuous function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ is a *strict Lyapunov function* for the DI (4), if for every $t > 0$ and $\mathbf{x} \in \mathbf{S}_{\mathbf{H}(x)}$, we have that $\varphi(\mathbf{x}(t)) < \varphi(x)$ if $x \notin \mathcal{A}$ and $\varphi(\mathbf{x}(t)) \leq \varphi(x)$ otherwise. If such a φ exists, then more can be said about the behavior of a perturbed solution.

Theorem 3 ([3, Th. 3.6 and Prop. 3.27]). *Assume that we are in the context of Th. 2 and let φ be a strict Lyapunov function for a set $\mathcal{A} \subset \mathbb{R}^d$. Assume, moreover, that $\varphi(\mathcal{A})$ is of empty interior, we have:*

$$\mathbf{L}_{\mathcal{X}} := \bigcap_{t \geq 0} \text{cl} \{ \mathbf{X}(u) : u \geq t \} \subset \mathcal{A}$$

and φ is constant on $\mathbf{L}_{\mathcal{X}}$.

In this note, we will be primarily interested in two particular set valued maps. Consider $f : \mathbb{R}^d \rightarrow \mathbb{R}$ a locally Lipschitz function. By Rademacher's theorem, f is differentiable almost everywhere. The set $\partial f(x)$ of Clarke subgradients of f at x is defined as follows:

$$\partial f(x) = \text{conv} \{ v : \text{there is a sequence } x_i \rightarrow x \text{ s.t. } f \text{ is differentiable at } x_i \text{ and } \nabla f(x_i) \rightarrow v \}. \quad (5)$$

The set $\{x : 0 \in \partial f(x)\}$ of Clarke-critical points contains local extrema (see [8]). The map $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is usc and for every x in \mathbb{R}^d , $\partial f(x)$ is nonempty, compact and convex.

Given a convex set $\mathcal{X} \subset \mathbb{R}^d$, the normal cone of \mathcal{X} is a set valued map $\mathcal{N}_{\mathcal{X}} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, defined as:

$$\mathcal{N}_{\mathcal{X}}(x) = \{ v : \langle v, y - x \rangle \leq 0, \forall y \in \mathcal{X} \}. \quad (6)$$

For each $x \in \mathcal{X}$, $\mathcal{N}_{\mathcal{X}}(x)$ is a closed convex subset of \mathbb{R}^d .

2.3 Semialgebraic and definable functions

An important case to which our results apply, is when f, g and \mathcal{X} are semialgebraic, or more generally definable. We say that a set $A \subset \mathbb{R}^N$ is semialgebraic if it can be written as a finite union and intersection of sets of the form $\{x : P(x) \leq 0\}$, where $P : \mathbb{R}^N \rightarrow \mathbb{R}$ is some polynomial. A function is semialgebraic if its graph is a semialgebraic set. While they may be nonsmooth, semialgebraic functions present strong regularity properties. Among other things, they are C^k differentiable on a dense open set (for any $k \geq 0$) and stable under many elementary operations such as composition, sum, multiplication.

A generalization of this notion, which preserve the aforementioned structural properties, is the one of definableness in an o-minimal structure. While we will not mathematically define this notion here, let us mention that any semialgebraic function, as well as the exponential and the logarithm are definable (hence, also their composition). This explains their ubiquity in the optimization literature. Up to our knowledge, the first work to exploit the link between optimization and definableness was [5]. An interested reader can find more on definability and its usefulness in optimization in [12], and more details in [17], [9], [10].

2.4 Path differentiable functions

We say that a locally Lipschitz function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is *path differentiable* if for any a.c. curve $\mathbf{x} : [0, 1] \rightarrow \mathbb{R}^d$, for almost every $t \in [0, 1]$:

$$(f \circ \mathbf{x})'(t) = \langle v, \dot{\mathbf{x}}(t) \rangle \quad \forall v \in \partial f(\mathbf{x}(t)). \quad (7)$$

By [4, Proposition 2], every convex, concave, semialgebraic or definable function is path differentiable. Moreover, if another function $g: \mathbb{R}^d \rightarrow \mathbb{R}$ is path differentiable, then $f + g$ is also path differentiable [4, Corollary 4]. From a similar point of view, if \mathcal{X} is a convex set, then for any a.c. curve $x: [0, 1] \rightarrow \mathbb{R}^d$, for almost every $t \in [0, 1]$:

$$\langle v, \dot{x}(t) \rangle = 0 \quad \forall v \in \mathcal{N}_{\mathcal{X}}(x(t)). \quad (8)$$

Consider now $f, g: \mathbb{R}^d \rightarrow \mathbb{R}$ path differentiable, $\mathcal{X} \subset \mathbb{R}^d$ a convex set and x a solution to the DI (2). Using Eq. (7) and (8) and the fact that $\partial(f + g) \subset \partial f + \partial g$, we obtain

$$(f + g)(x(t)) - (f + g)(x(0)) = - \int_0^t \|\dot{x}(u)\|^2 du. \quad (9)$$

This implies that $(f + g)(x(t)) < (f + g)(x(0))$ if $x(0) \notin \mathcal{Z}$. In other words, $f + g$ is a strict Lyapunov function for the DI (2).

3 Main results

Consider $(\Omega, \Xi, \mathbb{P})$ a probability space and (η_n) a sequence of random variables with values in \mathbb{R}^d . Define $\text{prox}_{g, \mathcal{X}}^\gamma: \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, the proximal operator for g on \mathcal{X} with a step γ :

$$\text{prox}_{g, \mathcal{X}}^\gamma(x) = \arg \min_{y \in \mathcal{X}} \{g(y) + \frac{1}{2\gamma} \|y - x\|^2\}. \quad (10)$$

We study Eq. (1) under the following assumptions.

Assumption 1.

- i) The set \mathcal{X} is a closed convex subset of \mathbb{R}^d .
- ii) The functions $f, g: \mathbb{R}^d \rightarrow \mathbb{R}$ are locally Lipschitz continuous.
- iii) There is a filtration $(\mathcal{F}_n)_{n \in \mathbb{N}}$, such that (η_n) is a martingale difference sequence adapted to it, and x_n is \mathcal{F}_n measurable for every $n \in \mathbb{N}$.
- iv) The sequence of stepsizes (γ_n) is nonnegative and such that $\sum_{i=0}^{+\infty} \gamma_i = +\infty$.

Note that if g is nonconvex, $\text{prox}_{g, \mathcal{X}}^\gamma(x)$ is a set in \mathbb{R}^d . However, as soon as x_{n+1} is chosen in a measurable manner (relatively to η_{n+1} and x_n), (x_n) will be adapted to (\mathcal{F}_n) . Such a choice is always possible (see e.g. [10]).

By [15, 10.2 and 10.10], we can rewrite Eq. (1) as:

$$x_{n+1} = x_n - \gamma_n(v_n + v_n^g + v_n^{\mathcal{X}}) + \gamma_n \eta_{n+1}, \quad (11)$$

where $v_n^g \in \partial g(x_{n+1})$ and $v_n^{\mathcal{X}} \in \mathcal{N}_{\mathcal{X}}(x_{n+1})$.

Assumption 2.

- i) Almost surely, $\sup_n \|x_n\| < +\infty$.

ii) There is $q \geq 2$ such that

$$\sum_{i=0}^{+\infty} \gamma_i^{1+q/2} < +\infty, \quad (12)$$

and, for any compact set $\mathcal{K} \subset \mathbb{R}^d$,

$$\sup_{n \in \mathbb{N}} \mathbb{E}[\|\eta_{n+1}\|^q \mathbb{1}_{x_n \in \mathcal{K}} | \mathcal{F}_n] < +\infty. \quad (13)$$

Assumptions of this type are standard in the field of stochastic approximation. Assumption 2-(i) prevent the algorithm to diverge. Note that it is superfluous if \mathcal{X} is compact. Otherwise it can be obtained by a proper choice of the regularizer g (see [11]).

Let $\tau_n = \sum_{i=1}^n \gamma_i$ be the discrete time of the algorithm. Define the linearly interpolated process $\mathbf{X} \in \mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ by:

$$\mathbf{X}(t) = x_n + \frac{t - \tau_n}{\gamma_{n+1}} (x_{n+1} - x_n) \quad \text{for } \tau_n \leq t < \tau_{n+1}.$$

Following [3] we will show that \mathbf{X} is a perturbed solution of the DI (2). The next two assumptions ensure us that $f + g$ will be a Lyapunov function for the DI (2).

Assumption 3. *The functions f and g are path differentiable.*

Assumption 4. *The set of Clarke critical values $\{f(x) + g(x) : x \in \mathcal{Z}\}$ has an empty interior.*

Assumption 4 is a classical Sard-type condition. It ensures the fact that if \mathbf{x} is a solution to the DI (2), with $\mathbf{x}(0) \in \mathcal{Z}$, then \mathbf{x} is constant. As established in [5], it is satisfied as soon as f, g and \mathcal{X} are definable.

The next two propositions are not new and can be found in one way or another in e.g. [10], [4], [13], [6]. Nevertheless, since our set of assumptions is slightly different and their proof is a simple application of Section 2.2, for completeness, we include it in Section 4.1.

Proposition 4. *Let Assumptions 1 and 2 hold, then the family $(\mathbf{X}(t + \cdot))_{t \geq 0}$ is relatively compact. Moreover, if a sequence $t_n \rightarrow +\infty$ and $\mathbf{x} \in \mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ is such that $\mathbf{d}(\mathbf{X}(t_n + \cdot), \mathbf{x}) \rightarrow 0$, then \mathbf{x} is a solution to the DI (2).*

Proposition 5. *Under Assumptions 1–4, the set $\text{acc}\{x_n\}$ is included in \mathcal{Z} and $f + g$ is constant on $\text{acc}\{x_n\}$.*

The next theorem tells us that even if $\text{acc}\{x_n\}$ is not a single point, the time that it takes to (x_n) to go from one accumulation point to another goes to infinity. This is an extension of [6, Th. 6.i), Th. 7.i)], to the best of our knowledge this result is new in a stochastic and proximal setting.

Theorem 6. *Let Assumptions 1–4 hold. Let x, y be two distinct points in $\text{acc}\{x_n\}$. Consider two sequences n_i, n_j , with $n_i \leq n_j$, such that $x_{n_i} \rightarrow x$ and $x_{n_j} \rightarrow y$. Then $\tau_{n_j} - \tau_{n_i} \rightarrow +\infty$.*

Under Assumptions 1–3, the same result is true if $(f + g)(x) \leq (f + g)(y)$.

As it is shown in [14], it is possible that $\text{acc}\{x_n\}$ is not reduced to a unique point. Nevertheless, Th. 6 implies that the “nonconvergence” happens in a very slow manner. Asymptotically, the time spent by the algorithm to move from one accumulation point to another goes to infinity.

We now investigate the question of oscillations. Given U, V two open sets, such that $\text{cl}U \subset V$, we will call $I = [n_1, n_2]$ a maximal interval related to U, V if the set $X_{n_1}^{n_2} := \{x_{n_1}, x_{n_1+1}, \dots, x_{n_2}\}$ is such that $X_{n_1}^{n_2} \subset V$, $X_{n_1}^{n_2} \cap U \neq \emptyset$ and either x_{n_1-1} or x_{n_2+1} is not in V . The next two results are an extension of [6, Th. 7] to a stochastic setting.

Theorem 7 (Long intervals). *Let Assumptions 1-4 hold. Consider $x \in \text{acc}\{x_n\}$ and U, V two neighborhoods of x such that $\text{cl}U \subset V$. For $i \in \mathbb{N}$, denote $I_i = [n_{i_1}, n_{i_2}]$ a sequence of distinct maximal intervals related to U, V . Then, either one of I_i is unbounded or $\tau_{n_{i_2}} - \tau_{n_{i_1}} \rightarrow +\infty$.*

Theorem 8 (Oscillation compensation). *Let Assumptions 1-4 hold, and fix U, V and I_i as in Th. 7. Denote $A = \bigcup I_i$, then*

$$\frac{\sum_{i=1}^n \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_A(x_i)}{\sum_{i=1}^n \gamma_i \mathbb{1}_A(x_i)} \xrightarrow{n \rightarrow +\infty} 0. \quad (14)$$

Th. 8 gives an intuitive explanation of why Th. 6 holds. Indeed, while the drift coming from one iteration $v_i + v_i^g + v_i^{\mathcal{X}}$ might not go to zero (as it happens for such a simple example as $f(x) = \|x\|$, $g = 0$ and $\mathcal{X} = \mathbb{R}^d$), on average, it compensates itself. Th. 6 and 8 suggest that the algorithm oscillates around its accumulation set, while the center of these oscillations moves in $\text{acc}\{x_n\}$ with a vanishing speed.

Let us finish with a remark on the Eq. (14). At first sight, maximal intervals in Th. 8 and Th. 7 may seem artificial. A more satisfactory result would be

$$\frac{\sum_{i=1}^n \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_U(x_i)}{\sum_{i=1}^n \gamma_i \mathbb{1}_U(x_i)} \xrightarrow{n \rightarrow +\infty} 0, \quad (15)$$

where U is an open neighborhood of an accumulation point x . Looking at the proof of Th. 8, to obtain Eq. (15), we could think of defining maximal intervals as $I_i = [n_{i_1}, n_{i_2}]$ such that $\{x_{n_{i_1}}, \dots, x_{n_{i_2}}\} \subset U$ and $x_{n_{i_1}-1}, x_{n_{i_2}+1} \notin U$. Unfortunately, for this type of intervals we don't have an equivalent of Th. 7, i.e. it may very well be that the quantity $\tau_{n_{i_2}} - \tau_{n_{i_1}}$ is bounded. Actually, it is not very hard to show, that for the function from [14, Section 2], there are x, U such that Eq. (15) is false.

Nevertheless, as explained in [6], Eq. (14) is a good approximation of Eq. (15). Indeed, apply Th. 8 with U and $V = U^\delta$, where $U^\delta = \{y \in \mathbb{R}^d : \exists z \in U, \|z - y\| < \delta\}$, then, as an approximation, we have

$$\lim_{\delta \rightarrow 0} \lim_{n \rightarrow +\infty} \frac{\sum_{i=1}^n \gamma_{i+1} (v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_A(x_i)}{\sum_{i=1}^n \gamma_{i+1} \mathbb{1}_A(x_i)} \approx \lim_{n \rightarrow +\infty} \frac{\sum_{i=1}^n \gamma_{i+1} (v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_U(x_i)}{\sum_{i=1}^n \gamma_{i+1} \mathbb{1}_U(x_i)}.$$

4 Proofs

In the following we will denote $x_{n+1/2} = x_n - \gamma_n v_n + \gamma_n \eta_{n+1}$ and

$$N(T, n) = \inf\{j \geq n \text{ s.t. } \tau_j - \tau_n \geq T\}. \quad (16)$$

4.1 Proof of Prop. 4 and 5

To put ourselves in the context of Section 2.2 we need to alter the map $-\partial f - \partial g - \mathcal{N}_{\mathcal{X}}$ in a way that it verifies assumptions of Th. 2 and 3. While this section is slightly technical, conceptually, we just find a set-valued map G verifying assumptions of Th. 3 and s.t. $x_{n+1} \in G(x_n)$. A convinced reader may want to skip to Section 4.2.

We start with two technical lemmas.

Lemma 1. *Under Assumptions 1 and 2, almost surely, for every $T > 0$, we have:*

$$\lim_{n \rightarrow +\infty} \sup_{n \leq j \leq N(T,n)} \left\| \sum_{i=n}^j \gamma_i \eta_{i+1} \right\| = 0. \quad (17)$$

As a consequence, the sequence $(\|x_{n+1/2}\|)$ is almost surely bounded.

Proof. Indeed, since almost surely $\sup \|x_n\| < +\infty$, for each $\delta > 0$, there is $C > 0$ s.t. if we denote $A = \{\forall n \in \mathbb{N} \|x_n\| \leq C\}$, then $\mathbb{P}(A) > 1 - \delta$. Define $\tilde{\eta}_{m+1} = \eta_{m+1} \mathbb{1}_{\|x_m\| \leq C}$, then $\mathbb{E}[\tilde{\eta}_{m+1} | \mathcal{F}_m] = 0$ and $\sup_{n \in \mathbb{N}} \mathbb{E}[\|\tilde{\eta}_{n+1}\|^q] < +\infty$. Hence, by [2, Prop. 4.2], we have $\sup_{n \leq j \leq N(T,n)} \left\| \sum_{i=n}^j \gamma_i \tilde{\eta}_{i+1} \right\| \xrightarrow{n \rightarrow +\infty} 0$. Since δ is arbitrary, Eq. (17) follows. \square

Lemma 2. *Let Assumptions 1 and 2 hold. Let $A \in \Xi$ be a probability one set on which (x_n) and $(x_{n+1/2})$ are bounded, and let C be a random variable s.t. $\|x_n\| < C$ and C is finite valued on A . Then, for each $\omega \in A$, there are two globally Lipschitz functions $\tilde{g}, \tilde{f} : \mathbb{R}^d \rightarrow \mathbb{R}$ and a bounded set-valued map $\tilde{\mathcal{N}}_{\mathcal{X}} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ s.t. in Eq. (11) we have $v_n(w) \in \partial \tilde{f}(x_n(w))$, $v_n^g(w) \in \partial \tilde{g}(x_{n+1}(w))$ and $v_n^{\mathcal{X}}(w) \in \tilde{\mathcal{N}}_{\mathcal{X}}(x_{n+1}(w))$.*

Moreover, if x is a solution to the DI:

$$\dot{x}(t) \in -\partial \tilde{f}(x(t)) - \partial \tilde{g}(x(t)) - \tilde{\mathcal{N}}_{\mathcal{X}}(x(t)), \quad (18)$$

and that x remains in $B(0, C) \cap \mathcal{X}$, then x is a solution to the DI (2).

Finally, denoting $\tilde{\mathcal{Z}} = \{x : 0 \in \partial \tilde{f}(x) + \partial \tilde{g}(x) + \tilde{\mathcal{N}}_{\mathcal{X}}(x)\}$, we have the equality $\tilde{\mathcal{Z}} \cap B(0, C) = \mathcal{Z} \cap B(0, C)$.

Proof. Let $\Pi_{C+1} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be the projection on $B(0, C+1)$. Define $\tilde{f}(x) = f(\Pi_{C+1}(x))$, $\tilde{g}(x) = g(\Pi_{C+1}(x))$. By construction, we have that $v_n \in \partial \tilde{f}(x_n)$ and $v_n^g \in \partial g(x_{n+1})$ and that v_n, v_n^g are bounded by $L_{\tilde{f}}, L_{\tilde{g}}$ the Lipschitz constants of \tilde{f} and \tilde{g} . Hence, since $x_{n+1/2}$ is bounded, there is C_2 s.t. $\sup\{\|v_n^{\mathcal{X}}\| : n \in \mathbb{N}\} < C_2$. Defining $\tilde{\mathcal{N}}_{\mathcal{X}}(x) = \{v : \|v\| \leq \max(C_2, L_f, L_g), v \in \Pi_{\mathcal{X}}(x)\}$, where $\Pi_{\mathcal{X}}$ is a projection on \mathcal{X} , proves the first claim. The two other statements immediately follow from our construction. \square

To prove Prop. 4 it remains to show that X is a perturbed solution to the DI (18). For $t \in [\tau_n, \tau_{n+1})$, we define $\rho(t) = \eta_{n+1}$ and $\mathbf{b}(t) = \|x_{n+1} - x_n\|$. The condition on ρ immediately follows from Lemma 1. The condition on \mathbf{b} follows from the following lemma.

Lemma 3. *Under Assumptions 1 and 2, almost surely, we have that $\|x_{n+1} - x_n\| \xrightarrow{n \rightarrow +\infty} 0$.*

Proof. By Lemma 1, we have that $\|x_{n+1/2} - x_n\| \xrightarrow{n \rightarrow +\infty} 0$, moreover, we have:

$$g(x_{n+1}) + \frac{1}{2\gamma_n} \|x_{n+1} - x_{n+1/2}\|^2 \leq g(x_n) + \frac{1}{2\gamma_n} \|x_n - x_{n+1/2}\|^2.$$

Therefore,

$$\begin{aligned} \frac{1}{2\gamma_n} \|x_{n+1} - x_n\|^2 &\leq g(x_n) - g(x_{n+1}) - \frac{1}{\gamma_n} \langle x_{n+1} - x_n, x_n - x_{n+1/2} \rangle \\ &\leq \|x_{n+1} - x_n\| \left(L_g + \frac{\|x_n - x_{n+1/2}\|}{\gamma_n} \right), \end{aligned}$$

and

$$\|x_{n+1} - x_n\| \leq \gamma_n L_g + \|x_n - x_{n+1/2}\|,$$

which finishes the proof. \square

To finish the proof of Prop. 4 consider $t_n \rightarrow +\infty$ and x s.t. $\mathbf{d}(\mathbf{X}(t_n + \cdot), x) \rightarrow 0$. Then, by Th. 2, x is a solution to the DI (18), moreover, it remains in $B(0, C) \cap \mathcal{X}$, therefore, it is also a solution to the DI (2).

For the proof of Prop. 5, notice that $\tilde{f} + \tilde{g}$ is path differentiable (as a composition of path differentiable functions). Then, in the same way as in Section 2.4, we have that $\tilde{f} + \tilde{g}$ is a strict Lyapunov function for the DI (18) and for the set $\tilde{\mathcal{Z}}$. Since $\text{acc}\{x_n\} = \mathbf{L}_X \subset \text{cl } B(0, C)$, by Th. 3 we have that $\mathbf{L}_X \subset \tilde{\mathcal{Z}} \cap \text{cl } B(0, C) \subset \mathcal{Z}$, and that $f + g$ is constant on $\text{acc}\{x_n\}$.

4.2 Proof of Th. 6

Lemma 4. *Let Assumptions 1–3 hold, let τ_n be a positive sequence, with $\tau_n \rightarrow +\infty$, and x s.t. $\mathbf{X}(\tau_n + \cdot) \rightarrow x$, then*

$$(f + g)(x(h)) \leq (f + g)(x(0)), \quad \forall h \in \mathbb{R}_+. \quad (19)$$

Moreover, if for some $h \geq 0$, $(f + g)(x(h)) = (f + g)(x(0))$, then $x(h') = x(0)$ for every $h' \in [0, h]$. If additionally Assumption 4 holds, then:

$$x(h) = x(0), \quad \forall h \in \mathbb{R}_+. \quad (20)$$

Proof. By Prop. 4, x is a solution to the DI (2), and the first result follows by Eq. (9).

Under Assumption 4, we have that $x(\mathbb{R}_+) \subset \text{acc}\{x_n\} \subset \mathcal{Z}$, hence, by Prop. 5, we have that $(f + g) \circ x$ is constant. Using Assumption 3, we have for all $h \in \mathbb{R}_+$,

$$0 = (f + g)(x(h)) - (f + g)(x(0)) = - \int_0^h \|\dot{x}(u)\|^2 du. \quad (21)$$

This implies that $\int_0^h \|\dot{x}(u)\|^2 du = 0$. Hence, $\dot{x}(h) = 0$ for almost every $h \in \mathbb{R}_+$ and we obtain Eq. (20). \square

Suppose that there is $T > 0$ such that $\tau_{n_j} - \tau_{n_i} \leq T$. The sequence $\mathbf{X}(\tau_{n_i} + \cdot)$ is relatively compact, and after extraction it converges to x a solution to (2). Extract once again to have $\tau_{n_j} - \tau_{n_i} \rightarrow h$. Then

$$\mathbf{X}(\tau_{n_j}) - \mathbf{X}(\tau_{n_i}) \rightarrow x(h) - x(0) = y - x,$$

and we obtain a contradiction with Lemma 4.

4.3 Proof of Th. 7

The next lemma is the key ingredient for the proofs of Th. 7 and Th. 8.

Lemma 5. *Under Assumptions 1-4, we have*

$$\sup_{n \leq j \leq N(T, n)} \left\| \sum_{i=n}^j \gamma_i(v_i + v_i^g + v_i^{\mathcal{X}}) \right\| \xrightarrow{n \rightarrow +\infty} 0.$$

Proof. Suppose that we have $\varepsilon > 0$ and two sequences n_k and $n_k \leq j_k \leq N(T, n_k)$, such that for n_k large enough:

$$\left\| \sum_{i=n_k}^{j_k} \gamma_i(v_i + v_i^g + v_i^{\mathcal{X}}) \right\| > \varepsilon.$$

This implies:

$$\left\| x_{j_k} - x_{n_k} + \sum_{i=n_k}^{j_k} \gamma_i \eta_{i+1} \right\| > \varepsilon.$$

Extract a sequence such that $X(\tau_{n_k} + \cdot)$ converges to x and $\tau_{j_k} - \tau_{n_k} \rightarrow h$, with $h \leq T$. Then $x_{j_k} \rightarrow x(h)$ and $x_{n_k} \rightarrow x(0)$, but $\|x(h) - x(0)\| \geq \varepsilon$ which is impossible by Lemma 4. \square

Suppose that no I_i is unbounded, then we can choose $n_i \in I_i = [n_{i_1}, n_{i_2}]$ such that $x_{n_i} \in U$. Since $x_{n_{i_2+1}}$ is in V^c , after extraction $x_{n_i} \rightarrow y_1$ and $x_{n_{i_2+1}} \rightarrow y_2$, with $y_2 \neq y_1$, moreover:

$$\tau_{n_{i_2+1}} - \tau_{n_i} - \gamma_{n_{i_2+1}} \leq \tau_{n_{i_2}} - \tau_{n_{i_1}}. \quad (22)$$

By Th. 6, the first term of this inequality tends to infinity.

4.4 Proof of Th. 8

Take I_i as in Th. 7, and $A_n = \bigcup_{i \leq n} I_i$. Define

$$u_n = \frac{a_n}{b_n} = \frac{\sum \gamma_i(v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_{A_n}(x_i)}{\sum \gamma_i \mathbb{1}_{A_n}(x_i)}.$$

Then,

$$u_{n+1} = \frac{a_n + \sum \gamma_i(v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_{I_{n+1}}(x_i)}{b_n + \sum \gamma_i \mathbb{1}_{I_{n+1}}(x_i)}. \quad (23)$$

Fix $\varepsilon > 0$, by Lemma 5, there is n_0 such that, for $n \geq n_0$, $\left\| \sum_{i=n_k}^{j_k} \gamma_i(v_i + v_i^g + v_i^{\mathcal{X}}) \right\| \leq \varepsilon$. Decompose $I_i = [n_{i_1}, n_{i_2}] = \bigcup_{1 \leq k \leq K_i} [a_{i,k}, a_{i,k+1}]$, with $a_{i,1} = n_{i_1}$ and $a_{i,k+1} = \min\{N(T, a_{i,k}), n_{i_2}\}$. We obtain:

$$\begin{aligned} u_{n+1} &= \frac{a_n + \sum_{k \leq K_n} \sum_{i=a_{n,k}}^{a_{n,k+1}} \gamma_i(v_i + v_i^g + v_i^{\mathcal{X}})}{b_n + \sum_{k \leq K_n} \sum_{i=a_{n,k}}^{a_{n,k+1}} \gamma_i} \\ &\leq \frac{a_n + (K_n)\varepsilon}{b_n + (K_n - 1)T}. \end{aligned}$$

By Th. 7, we have that $K_n \rightarrow +\infty$ and, therefore, for n large enough:

$$u_{n+1} \leq \frac{a_n + 2(K_n - 1)\varepsilon}{b_n + (K_n - 1)T}.$$

Hence, by induction:

$$u_{n+j} \leq \frac{a_n + 2\varepsilon \sum_{k=n}^{n+j-1} (K_i - 1)}{b_n + T \sum_{k=n}^{n+j-1} (K_i - 1)}.$$

Therefore, $\lim u_n \leq \frac{2\varepsilon}{T}$. Since ε is arbitrary, this finishes the proof.

Fundings

This work was supported by Région Île-de-France.

References

- [1] 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.
- [2] M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Math.*, pages 1–68. Springer, Berlin, 1999.
- [3] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM J. Control Optim.*, 44(1):328–348 (electronic), 2005.
- [4] J. Bolte and E. Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning. *arXiv preprint arXiv:1909.10300*, 2019.
- [5] Jérôme Bolte, Aris Daniilidis, Adrian Lewis, and Masahiro Shiota. Clarke subgradients of stratifiable functions. *SIAM Journal on Optimization*, 18(2):556–572, 2007.
- [6] Jerome Bolte, Edouard Pauwels, and Rodolfo Rios-Zertuche. Long term dynamics of the subgradient method for lipschitz path differentiable functions, 2020.
- [7] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning, 2018.
- [8] 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.
- [9] M. Coste. *An Introduction to O-minimal Geometry*. Dottorato di ricerca in matematica / Università di Pisa, Dipartimento di Matematica. Istituti editoriali e poligrafici internazionali, 2000.
- [10] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, Jan 2019.
- [11] John Duchi and Feng Ruan. Stochastic methods for composite and weakly convex optimization problems, 2018.
- [12] A. D. Ioffe. An invitation to tame optimization. *SIAM J. on Optimization*, 19(4):1894–1917, February 2009.

- [13] S. Majewski, B. Miasojedow, and E. Moulines. Analysis of nonsmooth stochastic approximation: the differential inclusion approach. *arXiv preprint arXiv:1805.01916*, 2018.
- [14] Rodolfo Rios-Zertuche. Examples of pathological dynamics of the subgradient method for lipschitz path-differentiable functions, 2020.
- [15] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.
- [16] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.
- [17] L. van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Math. J.*, 84(2):497–540, 08 1996.