Stochastic Subgradient Descent Escapes Active Strict Saddles

Pascal Bianchi, Walid Hachem, Sholom Schechtman

Stochastic Subgradient Descent Escapes Active Strict Saddles

Pascal Bianchi, Walid Hachem, Sholom Schechtman

August 4, 2021

Abstract

In non-smooth stochastic optimization, we establish the non-convergence of the stochastic subgradient descent (SGD) to the critical points recently called active strict saddles by Davis and Drusvyatskiy. Such points lie on a manifold M where the function f has a direction of second-order negative curvature. Off this manifold, the norm of the Clarke subdifferential of f is lower-bounded. We require two conditions on f. The first assumption is a Verdier stratification condition, which is a refinement of the popular Whitney stratification. It allows us to establish a reinforced version of the projection formula of Bolte *et.al.* for Whitney stratifiable functions, and which is of independent interest. The second assumption, termed the angle condition, allows to control the distance of the iterates to M. When f is weakly convex, our assumptions are generic. Consequently, generically in the class of definable weakly convex functions, the SGD converges to a local minimizer.

Keywords. nonsmooth optimization, stochastic gradient descent, avoidance of traps, Clarke subdifferential, stratification.

1 Introduction

Stochastic approximation algorithms that operate on non-convex and non-smooth functions have recently attracted a great deal of attention, owing to their numerous applications in machine learning and in high-dimensional statistics. The archetype of such algorithms is the so-called Stochastic Subgradient Descent (SGD), which reads as follows. Given a locally Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ which is not necessarily smooth nor convex, the $\mathbb{R}^d$–valued sequence (x_n) of iterates generated by such an algorithm satisfy the inclusion

$$x_{n+1} \in x_n - \gamma_n \partial f(x_n) + \gamma_n \eta_{n+1}, \tag{1}$$

where the set–valued function ∂f is the so-called Clarke subdifferential of f, the sequence (γ_n) is a sequence of positive step sizes converging to zero, and η_{n+1} is a zero-mean random vector on $\mathbb{R}^d$ which presence is typically due to the partial knowledge of ∂f by the designer. It is desired that (x_n) converges to the set of local minimizers of the function f.

Before delving into the subject of convergence towards minimizers, let us first consider the set $\mathcal{Z} := \{x \in \mathbb{R}^d : 0 \in \partial f(x)\}$ of *Clarke critical points* of f, which is generally larger than the set of minimizers, in the non-convex case. In order to ensure the convergence of (x_n) to $\mathcal{Z}$, the sole local Lipschitz property of f is not enough (see [14] for a counterexample), and some form of structure for the function f is required. Since the work of Bolte *et.al.* [6] in

optimization theory, it is well known that the so-called *definable on an o-minimal structure* (henceforth definable) functions, which belong to the family of *Whitney stratifiable* functions (see Section 2 below), is relevant for the convergence analysis of (x_n) and beyond. This class of functions is general enough so as to contain all the functions that are practically used in machine learning, statistics, or applied optimization. In this framework, the almost sure convergence of (x_n) to $\mathcal{Z}$ was established by Davis *et.al.* in [16]. Another work in the same line is [27]. Bolte and Pauwels [7] generalize the algorithm (1) by replacing ∂f with an arbitrary so-called conservative field. The constant step size regime $\gamma_n \equiv \gamma$ is considered in [3].

Thanks to these contributions, the convergence of (x_n) to the set $\mathcal{Z}$ is now well understood. However, as said above, $\mathcal{Z}$ is in general strictly larger than the set of minimizers, and can contain "spurious" points such as local maximizers or saddle points. The issue of the *non-convergence* of the sequence given by (1) towards spurious critical points is therefore crucial. The present paper investigates this issue.

Before getting into the core of our subject, it is useful to make a quick overview of the results devoted to the avoidance of spurious critical points by the iterative algorithms. The rich literature on this subject has been almost entirely devoted to the smooth setting. In this framework, the research has followed two main axes:

- The noisy case, where the analogue of the sequence (η_n) in the smooth version of Algorithm (1) is non zero. Here, the seminal works of Pemantle [28] and Brandière and Duflo [9] allow to establish the non-convergence of the Stochastic Gradient Descent (and, more generally, of Robbins-Monro algorithms) to a certain type of spurious critical points, sometimes referred to as *traps* or *strict saddle*. A critical point of a smooth function f is called a trap if the Hessian matrix of f at this point admits at least one negative eigenvalue. With probability one, the sequence (x_n) cannot converge to a trap, provided that the projection of the random perturbation η_n onto the eigenspace of corresponding to the negative eigenvalues of the Hessian matrix (henceforth, eigenspace of negative curvature) has a non vanishing variance.

- The noiseless case where $\eta_n \equiv 0$, studied for smooth functions by [23]. Here the authors show that for Lebesgue almost all initialization points, the algorithm with constant step will avoid the traps.

While both of these approaches rely on the center-stable invariant manifold theorem which finds its roots in the work of Poincaré, they are different in spirit. Indeed, in [23] the trap avoidance is due to the random initialization of the algorithm, whereas in [9, 28], it is due to the inherent stochasticity brought by the sequence (η_n).

We now get back to the non-smooth case. Here, the only paper that tackles the problem of the spurious points avoidance is, up to our knowledge, the recent contribution [15] of Davis and Drusvyatskiy. The spurious points that were considered in this reference are the so-called *active strict saddles*. Informally, a critical point is an active strict saddle if it lies on a manifold M such that *i)* f varies sharply outside of M, *ii)* the restriction of f to M is smooth, and *iii)* the Riemannian Hessian of f on M has at least one negative eigenvalue. For instance, the function $f : \mathbb{R}^2 \to \mathbb{R}, (y, z) \mapsto |z| - y^2$ admits the point $(0, 0)$ as an active strict saddle with $M = \mathbb{R} \times \{0\}$, and the restriction of f to M is the function $f_M(y, 0) = -y^2$, which obviously has a second-order negative curvature. In this setting, and assuming that f is weakly convex, the article [15] focuses on the noiseless case, and study variants of the

(implicit) *proximal point algorithm* rather than the (explicit) subgradient descent. Similarly to [23], they show that for Lebesgue almost every initialization point, different versions of the proximal algorithm avoid active strict saddles with probability one. Such a result is possible due to the fact that proximal methods implicitly run a gradient descent on a smoothened version of f - the Moreau envelope.

Contrary to [15], the algorithm (1) studied in this paper is explicit, meaning that it does not require the computation of a proximal operator associated with the non-smooth function. In this situation, the sole randomization of the initial point is not sufficient to expect an avoidance of active strict saddles. Here, in the same line as [28, 9], our analysis strongly relies on the presence of the additive random perturbation η_n.

In the framework of definable functions, we investigate the problem of the avoidance of the active strict saddle points. Our approach goes as follows. First, we need to show that the iterates (x_n) converge sufficiently fast to M, thanks to the sharpness of f outside this manifold. To that end, we first rely on the fact that when f is definable, its graph always admits a so-called *Verdier stratification*, which is perhaps less known than the Whitney stratification, and is a refinement of the latter [26]. The key advantage of the Verdier over the Whitney stratification lies in a Lipschitz-like condition on the (Riemannian) gradients of f on two adjacent stratas, which is established in the paper. Our second tool is an assumption that we term as the *angle condition*. Roughly, this assumption provides a lower bound on the inner product between the subgradients of f at x and the normal direction from M to x when the point x is near M. The angle condition allow to control the distance between the iterate x_n of Algorithm (1) and the manifold M. As the restriction f_M of f to M is smooth, the projected iterates, using the Verdier stratification property, are shown to follow a dynamics which is similar to a (smooth) Stochastic Gradient Descent, up to a residual term induced by the projection step. In that sense, the avoidance of active strict saddles in the non-smooth setting follows from the avoidance of traps in the smooth setting, as established in [9]. We show that the strict saddle is avoided under the assumption that the (conditional) noise covariance matrix has a non zero projection on the subspace with negative curvature associated with f_M near the active strict saddle.

Before pursuing, it is important to discuss the matter of the *genericity* of the assumptions that we just outlined. First, since our avoidance results are restricted to the active strict saddles, the question of the presence of critical points that are neither local minima nor active strict saddles is immediately raised. Actually, this question was considered in [17, 15]. It is established there that if f is definable and weakly convex, then for Lebesgue almost all vectors $u \in \mathbb{R}^d$, the function $f_u(x) := f(x) - \langle u, x \rangle$ admits a finite number of Clarke critical points, and that each of these points is either an active strict saddle or a local minimizer. In that sense, in the class of definable weakly convex functions, spurious critical points generically coincide with active strict saddles. We also need to inspect the generality of the Verdier and the angle conditions. In Theorem 2 below, we show that these assumptions are automatically satisfied when f is weakly convex. From these considerations, we conclude that generically in the sense of [17, 15], the SGD algorithm (1) converges to a local minimum when f is a weakly convex function, assuming that the noise is omnidirectional enough at the strict saddles. We emphasize the fact that, while the genericity of the active strict saddles is established in the above sense for weakly convex functions, no assumption on weak convexity is made for our avoidance of traps result.

Let us summarize the contributions of this paper:

- Firstly, we bring to the fore the fact that definable functions admit stratifications of the Verdier type. These are more refined than the Whitney stratifications which were popularized in the optimization literature by [6]. While such stratifications are well-known in the literature on o-minimal structures [26], up to our knowledge, they have not been used yet in the field of non smooth optimization. To illustrate their interest in this field, we study the properties of the Verdier stratifiable functions as regards their Clarke subdifferentials. Specifically, we refine the so-called projection formula (see [6, Proposition 4] and Lemma 2 below) to the case of definable, locally Lipschitz continuous functions by establishing a Lipschitz-like condition on the (Riemannian) gradients of two adjacent stratas.
- With the help of the Verdier and the angle conditions, we show that the SGD avoids the active strict saddles if the noise η_n is omnidirectional enough.

The paper is organized as follows. Section 2 is devoted to the introduction of the mathematical tools in use in this paper. Most of the results in this section are known, except for the reinforced projection formula, which is stated in Theorem 1. In Section 3, we discuss the notion of active strict saddles. After recalling some results of [15], we introduce the Verdier and angle conditions. We also discuss the genericity of the these conditions, in the class of weakly convex functions. In Section 4, we state the main result of the paper, namely, the avoidance of active strict saddles. Section 5 is devoted to the proofs.

2 Preliminaries

Notations. Let $d \geqslant 1$ be an integer. Given a set $S \subset \mathbb{R}^d$, $\overline{S}$ denotes the closure of S, and $\mathrm{conv}(S)$ and $\overline{\mathrm{conv}}(S)$ respectively denote the convex hull and the closed convex hull of S. The distance to S is denoted as $\mathrm{dist}(x, S) := \inf\{\|y - x\| : y \in S\}$. If $E \subset \mathbb{R}^d$ is a vector space, we denote by P_E the $d \times d$ orthogonal projection matrix onto E. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is weakly convex if there is $\rho > 0$ such that the function $g(x) := f(x) + \rho \|x\|^2$ is convex. For two sequences $(a_n), (b_n)$, we write $a_n \gtrsim b_n$ if $\liminf \frac{a_n}{b_n} > 0$. With this notation $a_n \sim b_n$ means $a_n \gtrsim b_n$ and $b_n \gtrsim a_n$. For $r > 0$, $B(0, r)$ denotes the open ball of radius r.

Throughout the paper, C and C' will refer to positive constants that can change from line to line and from one statement to another.

2.1 Functions on Manifolds

We refer to [22] for a detailed introduction on differential geometry.

Let $J_g(x)$ denote the Jacobian matrix of a map g at point x. Given two integers $p \geqslant 1$ and $k \leqslant d$, a C^p map $g \colon U \to \mathbb{R}^{d-k}$ on some open set $U \subset \mathbb{R}^d$ is called a C^p submersion if the rank of $J_g(x)$ is equal to $d-k$ for every $x \in U$. We say that a set $M \subset \mathbb{R}^d$ is a C^p submanifold of dimension k, if for every $y \in M$, there is a neighborhood U of y and a C^p submersion $g \colon U \to \mathbb{R}^{d-k}$, such that $U \cap M = g^{-1}(\{0\})$. We represent the tangent space of M at y by $T_y M := \ker J_g(y)$ (*n.b.*, the definition is independent of the choice of g). Equivalently, $T_y M$ can be represented as the set of vectors $v \in \mathbb{R}^d$ such that there exists a differentiable map $c : (-\varepsilon, \varepsilon) \to \mathbb{R}^d$ such that $c((-\varepsilon, \varepsilon)) \subset M$, $c(0) = y$ and $\dot{c}(0) = v$.

For every $x \in \mathbb{R}^d$, we define

$$P_M(x) := \arg\min_{y \in M} \|y - x\| ,$$

whenever the argument of the minimum exists and is unique. The following lemma can be found in [25] (see also [22, Chap. 3, Ex. 24]).

Lemma 1 (Projection onto a manifold)**.** *Let M be a C^p submanifold, with $p \geqslant 2$. Consider $y \in M$. Then, the projection P_M is well defined in a neighborhood of y. Moreover, P_M is C^{p-1} in that neighborhood, and $J_{P_M} = P_{T_yM}$.*

We say that a function $f\colon M \to \mathbb{R}$ is C^p, if M is a C^p submanifold, and if for every $y \in M$, there is a neighborhood $U \subset \mathbb{R}^d$ of y and a C^p function $F : U \to \mathbb{R}$ that agrees with f on $M \cap U$. If $f : M \to \mathbb{R}$ is C^1, we define for every $y \in M$,

$$\nabla_M f(y) := P_{T_yM} \nabla F(y)\,,$$

where P_{T_yM} is the orthogonal projection onto T_yM, and where $F : U \to \mathbb{R}$ is any C^1 function defined in a neighborhood of y, and which agrees with f on M. The definition of $\nabla_M f(y)$ does not depend on the choice of F (see e.g. [8, Section 3.8]). We refer to $\nabla_M f(y)$ as the (Riemannian) gradient of f at M. We say that y is a *critical point* of f, if $\nabla_M f(y) = 0$.

If $f : M \to \mathbb{R}$ is C^2 and if y is a critical point of f, we define the (Riemannian) Hessian of f at a y as the quadratic form $\mathcal{H}_f(y)$, defined on $\mathbb{R}^d \to \mathbb{R}$ by:

$$\mathcal{H}_{f,M}(y) :\ v \mapsto v^T P_{T_yM} \nabla^2 F(y) P_{T_yM} v\,,$$

where F is a C^2 function defined in a neighborhood of y which agrees with f on M, and where $\nabla^2 F(y)$ is the standard Hessian matrix of F at y. The definition of $\mathcal{H}_{f,M}(y)$ does not depend on the choice of F.

The above definitions have the following consequences. For a point $y \in M$ and a vector $v \in T_yM$, consider a differentiable curve $c : (-\varepsilon, \varepsilon) \to \mathbb{R}^d$ such that $c((-\varepsilon,\varepsilon)) \subset M$, $c(0) = y$ and $\dot{c}(0) = v$. Then $(f \circ c)'(0) = \langle \nabla_M f(y), v\rangle$. If f and c are C^2, and if y is a critical point of f, then $(f \circ c)''(0) = \mathcal{H}_{f,M}(y)(v)$.

If $f : \mathbb{R}^d \to \mathbb{R}$ is a function defined on the whole space, the gradient and Hessian of the restriction of f to M are still denoted by $\nabla_M f(y)$ and $\mathcal{H}_{f,M}(y)$ respectively, when they are well defined.

2.2 Clarke subdifferential

Consider $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous function. Denote by $\mathrm{Reg}(f)$ the set of points x at which f is differentiable, and by $\nabla f(x)$ the corresponding gradient., By Rademacher's theorem, f is differentiable almost everywhere. The *Clarke subdifferential of f at x* [12] is given by:

$$\partial f(x) := \overline{\mathrm{conv}}\{v \in \mathbb{R}^d : \exists (x_n) \in \mathrm{Reg}(f)^{\mathbb{N}}, (x_n, \nabla f(x_n)) \to (x, v)\}\,.$$

That is, $\partial f(x)$ is the closed convex hull of the points of the form $\lim \nabla f(x_n)$ for some sequence (x_n) converging to x. In particular, $\partial f(x)$ simply coincides with $\{\nabla f(x)\}$ when f is continuously differentiable in a neighborhood of x. We set $\mathcal{Z} = \{x \in \mathbb{R}^d : 0 \in \partial f(x)\}$. Every point of $\mathcal{Z}$ is referred to as a Clarke critical point. In particular, $\mathcal{Z}$ includes the local minimizers and the local maximizers of f.

Definition 1 (Path-differentiability)**.** *A locally Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be* path-differentiable *if for every absolutely continuous curve $c : (0,1) \to \mathbb{R}^d$, one has for almost every $t \in (0,1)$,*

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

In non-smooth optimization, the path-differentiability condition is often a crucial hypothesis in order to obtain relevant results *e.g.*, on the convergence of iterates [6, 16, 7]. In the sequel, we will review sufficient conditions on f, which ensure its path-differentiablility. In particular, we will review the notions of *definability* w.r.t. an o-minimal structure, *Verdier stratification* and *Whitney stratification*. In a nutshell, the different notions are related by:

$$\text{definable} \implies \text{graph Verdier-stratifiable} \implies \text{graph Whitney-stratifiable} \implies \text{path-differentiable.}$$

2.3 o-minimality

An o-minimal structure can be viewed as an axiomatization of diverse properties of semialgebraic sets. In an o-minimal structure, pathological sets such as Peano curves or the graph of the function $\sin \frac{1}{x}$ do not exist. To our knowledge the first work to link ideas between optimization and o-minimal structures was [6], where the authors analyzed the structure of the Clarke subdifferential of definable function and extended the Kurdyka-Łojasiewicz inequality [21] to the nonsmooth setting. Nowadays a rich body of literature enforces this link, see e.g. [16, 18, 5, 1, 7]. A nice exposure about usefulness of o-minimal theory in optimization is [20]. Results on the Verdier and Whitney stratification of definable sets can be found in [13, 30, 26].

An *o-minimal structure* is a family $\mathcal{O} = (\mathcal{O}_n)_{n\in\mathbb{N}*}$, where $\mathcal{O}_n$ is a set of subsets of $\mathbb{R}^n$, verifying the following axioms.

1. If $Q : \mathbb{R}^n \to \mathbb{R}$ is a polynomial, then $\{Q(x) = 0\} \in \mathcal{O}_n$.
2. If A and B are in $\mathcal{O}_n$, then the same is true for $A \cap B$, $A \cup B$ and A^c.
3. If $A \in \mathcal{O}_n$ and $B \in \mathcal{O}_m$, then $A \times B \in \mathcal{O}_{n+m}$.
4. If $A \in \mathcal{O}_n$, then the projection of A on its first $(n-1)$ coordinates is in $\mathcal{O}_{n-1}$.
5. Every element of $\mathcal{O}_1$ is exactly a finite union of intervals and points.

Sets contained in $\mathcal{O}$ are called *definable*. We call a map $f : \mathbb{R}^k \to \mathbb{R}^m$ definable if its graph is definable. Definable sets and maps have remarkable stability properties, for instance, if f and A are definable, then $f(A)$ and $f^{-1}(A)$, any composition of two functions definable in the same o-minimal structure is definable, and many others. Let us look at some examples of o-minimal structures.

Semialgebraic. Semialgebraic sets form an o-minimal structure. A set $A \subset \mathbb{R}^n$ is semialgebraic if it is a finite union of intersections of sets of the form $\{Q(x) \leqslant 0\}$, where $Q : \mathbb{R}^n \to \mathbb{R}$ is some polynomial. A function is semialgebraic if its graph is a semialgebraic set. Example of such functions include any piecewise polynomial functions but also functions such as $x \mapsto x^q$, where q is any rational number. It can be shown that any o-minimal structure contains every semialgebraic set.

Globally subanalytic. There is an o-minimal structure that contains, for every $n \in \mathbb{N}$, sets of the form $\{(x, t) : t = f(x)\}$, where $f : [-1, 1]^n \to \mathbb{R}$ is an analytic function that can be analytically extended in the neighborhood of the hypercube. The sets belonging to this structure are called globally subanalytic (see [5, 4] for more details).

Log-exp. There is an o-minimal structure that contains globally sub-analytic sets as well as the graph of the exponential and the logarithm (see [31]). As a consequence of this result it can be shown that the loss of a neural network is a definable function [16].

In the following we fix an o-minimal structure $\mathcal{O}$. Definable will always mean definable $\mathcal{O}$. The most striking property of definable sets is that they can always be partitioned into a finite number of manifolds that fit well into each other. This is formalized by the concept of stratification.

2.4 Whitney Stratification

Let A be a set in $\mathbb{R}^d$, a C^p stratification of A is a finite partition of A into a family of *stratas* (S_i) such that each of the S_i is a C^p submanifold verifying

$$S_i \cap \overline{S}_j \neq \varnothing \implies S_i \subset \overline{S}_j \backslash S_j \,.$$

Given a family $\{A_1, \ldots, A_k\}$ of subsets of of A, we say that a stratification (S_i) is *compatible with* $\{A_1, \ldots, A_k\}$, if each of the A_i is a finite union of stratas. We say that a stratification (S_i) is definable, if every strata S_i is definable.

Different types of stratifications exist depending on how tangent spaces of neighboring stratas fit together. Let us first define the asymmetric distance between two vector spaces E_1, E_2:

$$\boldsymbol{d}_a(E_1, E_2) = \sup_{u \in E_1, \|u\|=1} \operatorname{dist}(u, E_2). \tag{2}$$

Note that due to the lack of symmetry $\boldsymbol{d}_a$ is not a distance. Nevertheless, we have that $\boldsymbol{d}_a(E_1, E_2) = 0 \implies E_1 \subset E_2$. A distance $\boldsymbol{d}$ between E_1 and E_2 is then classically defined as

$$\boldsymbol{d}(E_1, E_2) = \max\{\boldsymbol{d}_a(E_1, E_2), \boldsymbol{d}_a(E_2, E_1)\} \,. \tag{3}$$

This distance is equal to zero if and only if $E_1 = E_2$. For a sequence of vector spaces $(E_n)_{n\in\mathbb{N}}$, we will denote $E_n \to E$ if $\boldsymbol{d}(E_n, E) \to 0$.

Definition 2. *We say that a C^p stratification (S_i) satisfies a* Whitney-(a) property*, if for every couple of distinct stratas S_i, S_j, for each $y \in S_i \cap \overline{S_j}$ and for each sequence $(x_n)_{n\in\mathbb{N}} \in (S_j)^{\mathbb{N}}$ such that $x_n \to y$, we have:*

$$\text{w-(a)} \quad \boldsymbol{d}(T_{x_n} S_j, \tau) \to 0 \implies T_y S_i \subset \tau \,. \tag{4}$$

We will refer to (S_i) as a Whitney C^p stratification.

It is known (see [13, 30]) that every definable function f admits a Whitney C^p (for any p) stratification (X_i) of its domain such that f is C^p on each strata. The following "projection formula" relates the Clarke subdifferential $\partial f(y)$ of f at y, to $\nabla_{X_i} f(y)$.

Lemma 2 (Projection formula, [6, Lemma 8])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz, definable function and p a positive integer. There is (S_i), a definable Whitney C^p stratification of* $\operatorname{Graph}(f)$*, such that if one denotes by X_i the projection of S_i onto its first d coordinates, the restriction $f\colon X_i \to \mathbb{R}$ is C^p and the family (X_i) is a Whitney C^p stratification of $\mathbb{R}^d$. Moreover, for any $y \in X_i$ and $v \in \partial f(y)$, we have $P_{T_y X_i}(v) = \nabla_{X_i} f(y)$.*

Lemma 2 has important consequences. One of them (see [16, Section 5] is that every locally Lipschitz continuous and definable function is path-differentiable.

Lemma 3 ([16, Theorem 5.8])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function. If* $\operatorname{Graph}(f)$ *admits a Whitney C^1 stratification, then f is path-differentiable.*

2.5 Verdier Stratification

A Verdier stratification is a special case of Whitney stratification, which posit a stronger condition on the (asymmetric) distance between adjacent stratas. Whereas the Whitney stratification can now be considered as well known in optimization community, the Verdier stratification is comparatively less popular. We illustrate its advantage by establishing in Theorem 1 a Lipschitz-like condition in the "projection formula" (Lemma 2). We believe that this strengthened result is of independent interest.

Definition 3. *Let (S_i) be a C^p stratification of some set $A \subset \mathbb{R}^d$. We say that (S_i) satisfies a* Verdier property (v)*, if for every couple of distinct stratas S_i, S_j and for each $y \in S_i \cap \overline{S_j} \neq \varnothing$, there are two positive constants δ, C such that:*

$$(v) \quad \begin{array}{ll} y' & \in B(y,\delta) \cap S_i \\ x & \in B(y,\delta) \cap S_j \end{array} \implies \boldsymbol{d}_a(T_{y'}S_i, T_x S_j) \leqslant C \left\| y' - x \right\| . \tag{5}$$

We refer to (S_i) as a Verdier C^p stratification of A.

It is clear from the definitions that a Verdier C^p stratification is always a Whitney C^p stratification. A fundamental result is that every definable set admits a Verdier stratification.

Proposition 1 ([26, Theorem 1.3])**.** *Let $\{A_1, \ldots, A_k\}$ be a family of definable sets of $\mathbb{R}^d$. For any $p \geqslant 1$, there is a Verdier C^p stratification of $\mathbb{R}^d$ compatible with $\{A_1, \ldots, A_k\}$.*

The following theorem, which we believe to be of independent interest, is the first main result of this paper. It is an improvement of Lemma 2.

Theorem 1 (Reinforced projection formula)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a definable, locally Lipschitz continuous function. Let p be a positive integer. There is (X_i), a definable Verdier C^p stratification of $\mathbb{R}^d$, such that for each $y \in X_i$ and each X_j such that $\overline{X_i} \cap X_j \neq \varnothing$, there is $C, \delta > 0$, such that for any two points $y' \in B(y,\delta) \cap X_i$, $x \in B(y,\delta) \cap X_j$,*

$$\left\| P_{T_{y'}X_i}(\nabla_{X_j} f(x)) - \nabla_{X_i} f(y') \right\| \leqslant C \left\| x - y' \right\| , \tag{6}$$

and, moreover, for any $x \in B(y,\delta) \cap X_i^c$ and any $v \in \partial f(x)$,

$$\left\| P_{T_{y'}X_i}(v) - \nabla_{X_i} f(y') \right\| \leqslant C \left\| x - y' \right\| . \tag{7}$$

Proof. In this proof $C' > 0$ will denote some constant that can change from line to line. Consider (S_i) and (X_i) as in Lemma 2. We claim that for any index j and $x \in X_j$, we have $T_{x,f(x)}S_j = \{(h, \langle \nabla_{X_i} f(x), h \rangle) : h \in T_x X_j\}$. Indeed, consider $(h_x, h_f) \in T_{x,f(x)}S_j$ and a C^p curve $c\colon (-\varepsilon, \varepsilon)$ s.t. $\dot{c}(0) = (h_x, h_f)$. Consider a C^p function F that agrees with f on X_j, then $(c_x(t), c_f(t)) = (c_x(t), F(c_x(t)))$ and we have $\dot{c}_x(0) = h_x$ and $\dot{c}_f(0) = \langle \nabla F(x), h_x \rangle = \langle \nabla_{X_j} f(x), h_x \rangle$.

Consider (S_i') a Verdier stratification of $\mathrm{Graph}(f)$ compatible with (S_i). Then the projection of S_i' onto its first d coordinates, that we denote X_i', is still a submanifold s.t. f is C^p on X_i'. Consider $(y, f(y)) \in S_i'$, S_j' a neighboring strata and C, δ as in Equation (5). Denote by L the Lipschitz constant of f on $B(y,\delta)$ and $\delta' = \frac{\delta}{L+1}$. Then, for every $x \in B(y,\delta')$, we have:

$$\|(y, f(y)) - (x, f(x))\| \leqslant (1 + L) \|y - x\| \leqslant \delta ,$$

that is to say $(x, f(x)) \in B((y, f(y)), \delta)$.

Consider $y' \in X'_i \cap B(y, \delta')$, $x \in X'_j \cap B(y, \delta')$ and $h_{y'} \in T_{y'} X'_i$ with $\left\|h_{y'}\right\| = 1$. We have that $(h_{y'}, \langle \nabla_{X'_i} f(y'), h_{y'} \rangle) \in T_{(y', f(y'))} S'_i$ and by the Verdier's condition there is $h_x \in T_x X'_j$ s.t.

$$\left\| \frac{1}{c_h} \left(h_{y'}, \langle \nabla_{X'_i} f(y'), h_{y'} \rangle \right) - (h_x, \langle \nabla f_{X'_j}(x), h_x \rangle) \right\| \leqslant C(L+1) \left\| x - y' \right\| ,$$

where $c_h = \left\| (h_{y'}, \langle \nabla_{X'_i} f(y'), h_{y'} \rangle) \right\| \leqslant C'$. Therefore,

$$\left\| h_{y'} - c_h h_x \right\| \leqslant C' \left\| x - y' \right\| ,$$

and

$$\begin{aligned} \left\| \langle \nabla_{X'_j} f(x) - \nabla_{X'_i} f(y'), h_{y'} \rangle \right\| &\leqslant \left\| \langle \nabla_{X'_j} f(x), h_{y'} - c_h h_x \rangle \right\| + \left\| c_h \langle \nabla_{X'_j} f(x), h_x \rangle - \langle \nabla_{X'_i} f(y'), h_{y'} \rangle \right\| \\ &\leqslant C' \left\| x - y' \right\| , \end{aligned}$$

which proves the first statement.

Now, one can choose C, δ such that Inequality (6) holds uniformly on all of the stratas X'_j that are neighboring X'_i. Consider a sequence $x_n \to x$ such that (x_n) lies in the stratas of full dimension (which implies that f is differentiable at x_n) and $\nabla f(x_n) \to v$, for n large enough we will have that $x_n \in B(y, \delta)$ and, therefore, $\left\| P_{T_{y'} X_i}(\nabla f(x_n)) - \nabla_{X_i} f(y') \right\| \leqslant C \left\| x_n - y' \right\|$. Hence, passing to the limit, we have that $\left\| P_{T_{y'} X_i}(v) - \nabla_{X_i} f(y') \right\| \leqslant C \left\| y' - x \right\|$. Since any element of $\partial f(x)$ is a convex combination of such v, the second statement is proved. □

3 Active strict saddles

In this section, $f : \mathbb{R}^d \to \mathbb{R}$ is supposed to be a locally Lipschitz continuous function. We recall the definition $\mathcal{Z} := \{x \in \mathbb{R}^d : 0 \in \partial f(x)\}$.

3.1 Definition and Existing Results

Let $p \geqslant 2$ be an integer.

Definition 4 (Active manifold, [24])**.** *Consider $x^* \in \mathcal{Z}$. A set $M \subset \mathbb{R}^d$ is called a C^p* active manifold around x^**, if there is a neighborhood U of x^* such that the following holds.*

i) ***Smoothness condition:*** *$M \cap U$ is a C^p submanifold and f is C^p on $M \cap U$.*

ii) ***Sharpness condition:***

$$\inf\{\|v\| : v \in \partial f(x), x \in U \cap M^c\} > 0 .$$

Definition 5 (Active strict saddle)**.** *We say*[1] *that a point $x^* \in \mathcal{Z}$ is an* active strict saddle (of order p) *if there exists a C^p active manifold M around x^*, and a vector $w \in T_{x^*} M$, such*

[1]The definition of active strict saddles provided in [15] involves the notion of parabolic subderivatives. In this paper, we found convenient to use the equivalent Definition 5, which is closer in spirit to notions of differential geometry.

that $\nabla_M f(x^*) = 0$ *and* $\mathcal{H}_{f,M}(x^*)(w) < 0$.
We say that f *satisfies the* active strict saddle property (of order p)*, if it has a finite number of Clarke critical points, and each of these points is either an active strict saddle of order* p *or a local minimizer .*

In the special case of a **smooth** function f, the space $M = \mathbb{R}^d$ is trivially an active manifold around any critical point x^* of f. If x^* is moreover a *trap* in the sense provided in the introduction (*i.e.*, the Hessian matrix of f at x^* admits a negative eigenvalue), then x^* is trivially an active strict saddle. Hence, the smooth setting can be handled as a special case.

The archetype of an active strict saddle is given by the following example.

Example 1. *The point* $(0,0)$ *is an active strict saddle of the function* $f : \mathbb{R}^2 \to \mathbb{R}$ *given by* $f(y,z) = -y^2 + |z|$*. Indeed,*

$$\partial f((y,z)) = \begin{cases} \{(-2y,1)\} \text{ if } z > 0\,, \\ \{(-2y,-1)\} \text{ if } z < 0\,, \\ \{-2y\} \times [-1,1] \text{ otherwise } , \end{cases}$$

and the set $M = \mathbb{R} \times \{0\}$ *is a* C^2 *active manifold. Moreover,* $\nabla_M f((y,0)) = (-2y, 0)$ *and* $\mathcal{H}_{f,M}(0)((1,0)) = -2$*, which proves the statement.*

While the definition of an active strict saddle might seem peculiar at first glance, the following proposition of Davis and Drusvyatskiy shows that a generic definable and weakly convex function satisfies a strict saddle property. The proof is grounded in the work of [17].

Proposition 2 ([15, Theorem 2.9])**.** *Assume that* f *is definable and weakly convex. Define* $f_u(x) := f(x) - \langle u, x\rangle$*, for every* $u \in \mathbb{R}^d$*. Then, for every* $p \geqslant 2$ *and for Lebesgue-almost every* $u \in \mathbb{R}^d$*,* f_u *has the active strict saddle property of order* p*.*

It is worth noting that the result of [15, Theorem 2.9] is in fact a bit stronger than Proposition 2, because it states moreover that for almost all u, the cardinality of the set of Clarke critical points of f_u is upper bounded by a finite constant which depends only on f.

One can wonder if Proposition 2 may still hold if f is definable and locally Lipschitz, but not weakly convex. The answer is negative, as shown by the following example.

Example 2. *Let* $f : \mathbb{R}^2 \to \mathbb{R}$ *be defined as* $f(y,z) = -|y| + |z|$*. Then for any* $u \in B(0,1)$*,* $(0,0)$ *is a critical point for* f_u*, but is neither a local minimum nor an active strict saddle.*

3.2 Verdier and Angle Conditions

On the top of the items *i-ii)* of Definition 4, we introduce the following useful conditions.

Definition 6. *Let* M *be a* C^1 *active manifold around some* $x^* \in \mathcal{Z}$*. We say that* M *satisfies the Verdier condition and the angle condition, if the following conditions hold respectively.*

iii) ***Verdier condition.*** *There is a neighborhood* U *of* x^* *and* $C \geqslant 0$*, such that for every* $y \in M \cap U$ *and every* $x \in U$*,*

$$\left\| P_{T_y M}(v) - \nabla_M f(y) \right\| \leqslant C \left\| x - y \right\|, \quad \forall v \in \partial f(x)\,.$$

iv) ***Angle condition.*** *For every $\alpha > 0$, there is $\beta > 0$ and a neighborhood U of x^*, such that for every $x \in U$,*

$$f(x) - f(P_M(x)) \geqslant \alpha \|x - P_M(x)\| \implies \langle v, x - P_M(x) \rangle \geqslant \beta \|x - P_M(x)\|, \quad \forall v \in \partial f(x).$$

Definition 7. *An active strict saddle x^* is said to satisfy the* Verdier and angle conditions*, if the active manifold M in Definition 5 satisfies the Verdier and angle conditions. The function f is said to satisfy the* active strict saddle property of order p with the Verdier and angle conditions*, if it satisfies the active strict saddle property of order p and if every active strict saddle satisfies the Verdier and angle conditions.*

The Verdier condition merely states that M is one of the stratas of the Verdier stratification of Theorem 1 The purpose of the angle condition is to relate, close to M, the linear growth of the function f and the lower boundedness of the inner product between the subgradients of f at x and the normal direction to M. The latter will allow us to prove that the iterates of SGD converge to M fast enough.

Remark 1. *Let M be an active manifold around x^*. As it will be clear from the proof of Theorem 2, when f is weakly convex, M always satisfies the angle condition. Otherwise stated, the angle condition is simply* true *in case of weakly convex functions. However, as the following example shows, one is able to find many natural examples of functions which are not weakly convex, and yet satisfy this condition.*

Example 3. *The function $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(y, z) = -y^2 - |z|$ is not weakly convex. Its unique Clarke critical point $(0, 0)$ is an active strict saddle, satisfying the Verdier and the angle conditions.*

Example 3 shows that the Verdier and angle conditions can be satisfied with no need for f to be weakly convex. Nevertheless, more can be said when this assumption holds. The following theorem strengthen the genericity result of Proposition 2 by establishing that the active strict saddle property with the Verdier and angle conditions is satisfied by a generic definable and weakly convex function. We recall the notation $f_u(x) = f(x) - \langle u, x \rangle$.

Theorem 2. *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is a definable, weakly convex function. For every $p \geqslant 2$, and for Lebesgue-almost every $u \in \mathbb{R}^d$, f_u satisfies the active strict saddle property of order p with the Verdier and angle conditions.*

Proof. Let $\{X_1, \ldots, X_k\}$ be the C^p Verdier stratification from Theorem 1. Upon noticing that in the proof of [17, Corollary 4.8 and Theorem 4.16] the active manifold [2] can be chosen adapted to $\{X_1, \ldots, X_k\}$, the existence of an active manifold with a Verdier condition follows from [15, Theorem 2.9, Appendix A]. For the angle condition note that by weak convexity of f there is $\rho \geqslant 0$ such that:

$$f(P_M(x)) - f(x) \geqslant \langle v, P_M(x) - x \rangle - \rho \|x - P_M(x)\|^2 \quad \forall v \in \partial f(x).$$

Therefore, if $f(x) \geqslant f(P_M(x)) + \alpha \|P_M(x) - x\|$, then:

$$\forall v \in \partial f(x), \quad \langle v, x - P_M(x) \rangle \geqslant \alpha \|x - P_M(x)\| - \rho \|x - P_M(x)\|^2 .$$

Taking U a neighborhood of x^* close enough to zero, we see that the angle condition is satisfied. □

[2] The name *active manifold* follows the work of [15], while in [17] they are called identifiable manifolds.

4 Avoidance of Active Strict Saddles

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function. On a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, consider a random variable x_0 and random sequences (v_n), (η_n) on $\mathbb{R}^d$. Define the iterates:

$$x_{n+1} = x_n - \gamma_n v_n + \gamma_n \eta_{n+1}\,, \tag{8}$$

where (γ_n) is a deterministic sequence of positive numbers. Let $(\mathscr{F}_n)$ be a filtration on $(\Omega, \mathcal{A}, \mathbb{P})$.

Assumption 1.

i) The function f is path differentiable.

ii) For every n, $v_n \in \partial f(x_n)$.

iii) The sequences (v_n), (η_n) are adapted to $(\mathscr{F}_n)$, and x_0 is $\mathscr{F}_0$-measurable.

iv) There are constants $c_1, c_2 > 0$ and $\alpha \in (1/2, 1]$ s.t. for all $n \in \mathbb{N}$:

$$\frac{c_1}{n^\alpha} \leqslant \gamma_n \leqslant \frac{c_2}{n^\alpha}\,.$$

Consider a point $x^* \in \mathcal{Z}$.

Assumption 2. *The point x^* is an active strict saddle of order 4 satisfying the Verdier and angle conditions.*

Since $\mathcal{H}_{f,M}(x^*)$ is a quadratic form we can write down $\mathbb{R}^d = E^- \oplus E^+$, where E^- (respectively E^+) is the vector space spanned by the eigenvectors of the associated symmetric bilinear form that have negative (respectively nonnegative eigenvalues). Note that by results of Section 2.1 we have that $E^- \subset T_{x^*}M$ and by Assumption 2 we have that $\dim E^- \geqslant 1$.

Assumption 3. *The following holds almost surely on the event $[x_n \to x^*]$.*

i) $\mathbb{E}[\eta_{n+1}|\mathscr{F}_n] = 0$, for all n.

ii) $\limsup \mathbb{E}[\|\eta_{n+1}\|^4 \,|\mathscr{F}_n] < +\infty$.

iii) Denote η_{n+1}^- the projection of η_{n+1} onto E^-. We have:

$$\liminf \mathbb{E}[\left\|\eta_{n+1}^-\right\| \,|\mathscr{F}_n] > 0$$

The following theorem is the main result of this paper.

Theorem 3. *Let Assumptions 1–3 hold. Then $\mathbb{P}(x_n \to x^*) = 0$.*

Combining Theorem 3 with the results of Section 3.2 we obtain that, under appropriate assumptions, the SGD on a generic definable, weakly convex function converges to a local minimizer. We state this result in the following corollary.

Corollary 1. *Let Assumptions 1 and 2 hold. Assume that f has the active strict saddle property of order 4 with the Verdier and angle conditions. Moreover, assume that almost surely the following holds.*

i) $\mathbb{E}[\eta_{n+1}|\mathscr{F}_n] = 0$, *for all* n.

ii) For every $C > 0$,

$$\limsup \mathbb{E}[\|\eta_{n+1}\|^4 \,|\mathscr{F}_n]\mathbb{1}_{\|x_n\|\leqslant C} < +\infty\,.$$

iii) For all $w \in \mathbb{R}^d\backslash\{0\}$,

$$\liminf \mathbb{E}[|\langle w, \eta_{n+1}\rangle| \,|\mathscr{F}_n] > 0\,.$$

Then, almost surely, the sequence (x_n) *is either unbounded, or converges to a local minimizer of* f.

5 Proof of Theorem 3

From now on, we assume without restriction that $x^* = 0$. Thus, $\nabla_M f(0) = 0$, and there exists a vector $w \in T_0M$ such that $\mathcal{H}_{f,M}(0)(w) < 0$.

The general idea of the proof of Theorem 3 is that on the event $[x_n \to 0]$, the function P_M is defined for all large n, enabling us to write $x_n = y_n + z_n$ for these n, where $y_n = P_M(x_n)$. The iterates (y_n) can then be written under the form of a standard smooth *Robbins-Monro algorithm* for which the trap avoidance can be established by the technique of Brandière and Duflo [9]. In this setting, the remainders z_n will be shown to be small enough so as not to alter fundamentally the approach of [9].

Let us provide more details on our proof. We first show that on $[x_n \to 0]$, there is an integer n_0 such that for all all $n \geqslant n_0$, the norms $\|x_n\|$ are small, and moreover,

$$\forall v \in \partial f(x_n), \quad \langle v, z_n\rangle \gtrsim \|z_n\|\,. \tag{9}$$

This will be the object of Proposition 4 below. The idea is to show that for these n, it holds that $f(x_n) - f(y_n) \gtrsim \|z_n\|$, and then, to use the angle condition (iv) of Definition 6.

Let us temporarily assume that n_0 is deterministic, and work on $n \geqslant n_0$. Keeping Inequality (9) aside for further use, the next step is to make a Taylor development of $y_{n+1} = P_M(x_{n+1})$ around x_n. This leads to

$$\begin{aligned} P_M(x_{n+1}) &= P_M(x_n) + J_{P_M}(x_n)(x_{n+1} - x_n) + \mathcal{O}(\|x_{n+1} - x_n\|^2) \\ &= P_M(x_n) + J_{P_M}(y_n)(x_{n+1} - x_n) + \mathcal{O}(\|x_{n+1} - x_n\|^2) + \mathcal{O}(\|z_n\|\,\|x_{n+1} - x_n\|), \end{aligned}$$

where we used the Lipschitz continuity of the Jacobian matrix function $J_{P_M}(\cdot)$. Using Equation (8), we rewrite the last display as

$$y_{n+1} = y_n - \gamma_n J_{P_M}(y_n)v_n + \gamma_n J_{P_M}(y_n)\eta_{n+1} + \gamma_n^2\mathcal{O}(1 + \|\eta_{n+1}\|^2) + \gamma_n\mathcal{O}(\|z_n\|\,(1 + \|\eta_{n+1}\|)).$$

Now, lemma 1 shows that $J_{P_M}(y_n)$ coincides with the linear operator $P_{T_{y_n}M}$. Furthermore, the Verdier condition (iii) of Definition 6 asserts that $P_{T_{y_n}M}(v_n) = \nabla_M f(y_n) + \mathcal{O}(\|z_n\|)$. Altogether, we obtain the Robbins-Monro iteration

$$y_{n+1} = y_n - \gamma_n\nabla_M f(y_n) + \gamma_n P_{T_{y_n}M}\eta_{n+1} + \gamma_n^2\mathcal{O}(1 + \|\eta_{n+1}\|^2) + \gamma_n\mathcal{O}(\|z_n\|\,(1 + \|\eta_{n+1}\|)). \tag{10}$$

Had we not have the last term $\gamma_n\mathcal{O}(\|z_n\|\,(1 + \|\eta_{n+1}\|))$ at the right hand side, the approach of Brandière and Duflo would have been enough to obtain the nonconvergence of y_n to zero under our assumptions on the noise. The presence of this term requires us to weaken a bit

their conditions. This will be done in Proposition 3. In the case of Equation (10), this proposition asserts that the trap avoidance remains true if

$$\sum_{i=n}^{\infty} \gamma_i \mathbb{E}\,\|z_i\| = \mathcal{O}(\chi_n)$$

where

$$\chi_n := \sum_{i=n}^{+\infty} \gamma_i^2.$$

This is where Inequality (9) comes into play to establish this bound.

So far, we have assumed abusively that the moment n_0 after which $\|x_n\|$ is small and (9) is satisfied is deterministic. To deal with this issue, in Section 5.2, on an arbitrary large event A, we construct a sequence (y_n) that is (for n large enough) equal to $(P_M(x_n))$ on $A \cap [x_n \to 0]$ and satisfies an equation of the form (10) almost surely. Proposition 3 will allow us to prove that $\mathbb{P}([x_n \to 0] \cap A) \leqslant \mathbb{P}([y_n \to 0]) = 0$ and since the event A is arbitrary large, this will prove Theorem 3.

5.1 Preliminary: Avoidance of Traps in the Smooth Case

The following proposition is nearly a quote of Brandière and Duflo's theorem [9, Theorem 1]. As discussed below, we alleviate some hypotheses of [9].

To state this proposition recall that, by a standard result from linear algebra, for a matrix $H \in \mathbb{R}^{d\times d}$, there is a decomposition $\mathbb{R}^d = \Lambda^+ \oplus \Lambda^-$ such that Λ^+, Λ^- are stable by H and the eigenvalues of $H_{|\Lambda^-}$ (respectively $H_{|\Lambda^+}$) have eigenvalues with negative (respectively nonpositive) real parts. Recall that for a smooth map $D : \mathbb{R}^d \to \mathbb{R}^d$, we denote J_D its jacobian and that $\chi_n := \sum_{i=n}^{\infty} \gamma_i^2$.

Proposition 3. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(\mathscr{F}_n)$ a filtration and (γ_n) a sequence of deterministic nonnegative step sizes such that $\sum_k \gamma_k = +\infty$ and $\sum_k \gamma_k^2 < +\infty$. Let d be an integer and $D : \mathbb{R}^d \to \mathbb{R}^d$ be such that $D(0) = 0$ and there is a neighborhood of 0 such that on it D is continuously differentiable, with Lipschitz continuous Jacobian. Consider the $\mathbb{R}^d$–valued stochastic process (y_n) given by*

$$y_{n+1} = y_n - \gamma_n D(y_n) + \gamma_n \tilde{\eta}_{n+1} + \gamma_n \varrho_{n+1} + \gamma_n \tilde{\varrho}_{n+1}\,, \tag{11}$$

where y_0 is $\mathscr{F}_0$-measurable and the sequences $(\tilde{\eta}_n), (\varrho_n)$ and $(\tilde{\varrho}_n)$ are $(\mathscr{F}_n)$-adapted. Assume that Λ^-, the vector space associated to the eigenvectors of $J_D(0)$ that have negative real parts, is of positive dimension. Denote $\tilde{\eta}^-_{n+1}$ the projection of $\tilde{\eta}_{n+1}$ on Λ^- and assume that on the event $[y_n \to 0]$ the following almost surely holds.

i) For all n, $\mathbb{E}[\tilde{\eta}_{n+1}|\mathscr{F}_n] = 0$.

ii) $\limsup \mathbb{E}\left[\|\tilde{\eta}_{n+1}\|^4 \middle| \mathscr{F}_n\right] < +\infty$.

iii) $\liminf \mathbb{E}\left[\left\|\tilde{\eta}^-_{n+1}\right\| |\mathscr{F}_n\right] > 0\,.$

iv) $\sum_{k=0}^{+\infty} \|\varrho_{k+1}\|^2 < +\infty\,.$

v) We have that:

$$\mathbb{E}\left[\mathbb{1}_{[y_n\to 0]}\sum_{i=n}^{+\infty}\gamma_i\left\|\tilde{\varrho}_{i+1}\right\|\right]=\mathcal{O}(\chi_n).$$

Then $\mathbb{P}([y_n\to 0])=0$.

Proposition 3 is similar to [9, Theorem 1], except for the presence of the sequence $(\tilde{\varrho}_n)$. As the proof is mainly an adaptation of the proof of [9, Theorem 1], we provide a sketch of proof in the appendix.

5.2 Application to Algorithm (8)

To apply the results of the preceding section we need, first, to find a candidate for D, this is the purpose of the next lemma. Its proof readily follows from results of Section 2.

Lemma 4. *Let Assumption 2 hold and let $r>0$ be such that $P_M : B(0,r)\to M$ is well defined and is C^3 and that there is a C^4 function $F : B(0,r)\to\mathbb{R}$ that agrees with f on $M\cap B(0,r)$. Then, the function $F\circ P_M$ is C^3 on $B(0,r)$ and for $y\in M\cap B(0,r)$, we have:*

$$\nabla(F\circ P_M)(y)=\nabla_M f(y)\,.$$

Moreover, for $w\in\mathbb{R}^d$:

$$\mathcal{H}_{f,M}(0)(w)=w^T\nabla^2(F\circ P_M)w\,.$$

By Tietze's extension theorem the function $\nabla(F\circ P_M) : B(0,r)\to\mathbb{R}^d$ can be extended to a bounded continuous function $D:\mathbb{R}^d\to\mathbb{R}^d$ that we shall use in the remainder of the paper.

For $r>0$ such that P_M is well defined on $B(0,r)$, and for $C>0$, denote

$$V_r(C)=\{x\in B(0,r):\forall v\in\partial f(x),\langle v,x-P_M(x)\rangle\geqslant C\left\|x-P_M(x)\right\|\}\,.$$

The next proposition is a key element in our proof. To not interrupt our exposition its proof is provided in Section 5.3.

Proposition 4. *Let Assumptions 1–3 hold. There is $\beta,r_1>0$, such that for every $r<r_1$, almost surely on the event $[x_n\to 0]$, $x_n\in V_r(\beta)$ for all n large enough.*

In the remainder, we fix $\beta,r_1>0$ as those provided by the previous proposition. We let U be the neighborhood around zero that verify conditions of Definition 6. In the following, we choose $r\leqslant r_1$ such that P_M is C^3 on $\overline{B(0,r)}$, and $\overline{B(0,r)}\subset U$. The value of r, while always satisfying these requirements, will be adjusted in the course of the proof.

Firstly, to reduce technical issues, we notice that as in [9, Section I.2] to prove Theorem 3 we can actually replace Assumption 3 by the following, more easy to handle, assumption. The notation $\mathbb{E}_n[\cdot]$ stands for $\mathbb{E}[\cdot|\mathscr{F}_n]$.

Assumption 4. *Almost surely, the sequence (η_n) is such that $\mathbb{E}_n[\eta_{n+1}]=0$ and there is $A,B>0$ such that for all $n\in\mathbb{N}$, we have:*

$$\mathbb{E}_n[\|\eta_{n+1}\|^4]\leqslant B$$

and

$$\mathbb{E}_n[\left\|\eta_{n+1}^-\right\|]\geqslant A\,.$$

Given an integer $N \geqslant 0$, we define the probability event

$$\mathcal{A}_N = [\forall n \geqslant N,\ x_n \in V_r(\beta)]\,.$$

Note that the sequence of events $(\mathcal{A}_N)$ is increasing for the inclusion. Furthermore, Proposition 4 shows that

$$[x_n \to 0] \subset \bigcup_{N=0}^{\infty} \mathcal{A}_N = \lim_{N\to\infty} \mathcal{A}_N.$$

Thus,

$$\mathbb{P}\,[x_n \to 0] = \mathbb{P}\,[[x_n \to 0] \cap \lim \mathcal{A}_N] = \lim_{N\to\infty} \mathbb{P}\,[[x_n \to 0] \cap \mathcal{A}_N]\,.$$

Consequently, given an arbitrary $\delta > 0$, there is an integer $N(\delta) \geqslant 0$ such that

$$\mathbb{P}\left[[x_n \to 0] \cap \mathcal{A}_{N(\delta)}\right] \geqslant \mathbb{P}\,[x_n \to 0] - \delta. \tag{12}$$

For an integer $N \geqslant 0$, define the stopping time

$$\tau_N = \inf\{n \geqslant N,\ x_n \notin V_r(\beta)\},$$

with $\inf \varnothing = \infty$, and recall from the definition of r that for $N \leqslant n < \tau_N$, the projection $P_M(x_n)$ is well-defined. Define recursively the process $(y_n^N)_{n\geqslant N-1}$ as follows: $y_{N-1}^N = 0$,

$$y_n^N = \begin{cases} P_M(x_n) \text{ if } N \leqslant n < \tau_N, \\ y_{n-1}^N - \gamma_{n-1} D(y_{n-1}^N) + \gamma_{n-1} J_{P_M}(y_{n-1}^N)\eta_n \text{ if } n = \tau_N, \\ y_{n-1}^N - \gamma_{n-1} D(y_{n-1}^N) + \gamma_{n-1}\eta_n, \quad \text{otherwise,} \end{cases}$$

and let

$$z_n^N = (x_n - y_n^N)\mathbb{1}_{n<\tau_N} \quad \text{for } n \geqslant N.$$

Observe that y_n^N and z_n^N are both $\mathscr{F}_n$–measurable for all $n \geqslant N$. To establish Theorem 3, we shall show that for each $N \geqslant 0$,

$$\mathbb{P}\left[y_n^N \xrightarrow[n\to\infty]{} 0\right] = 0. \tag{13}$$

Indeed, on the event $\mathcal{A}_{N(\delta)}$, it holds that $y_n^{N(\delta)} = P_M(x_n)$ for $n \geqslant N(\delta)$, thus,

$$\left[[x_n \to 0] \cap \mathcal{A}_{N(\delta)}\right] \subset \left[\left[y_n^{N(\delta)} \to 0\right] \cap \mathcal{A}_{N(\delta)}\right].$$

Consequently, with the convergence (13) at hand, we get from Inequality (12) that $\mathbb{P}[x_n \to 0] \leqslant \delta$. Since δ is arbitrary, we obtain that $\mathbb{P}[x_n \to 0] = 0$.

In the remainder of this section, $N \geqslant 0$ is a fixed integer.

Proposition 5. *Let Assumptions 1–2 and 4 hold. Then, the sequence $(y_n^N)_{n\geqslant N}$ satisfies the recursion:*

$$y_{n+1}^N = y_n^N - \gamma_n D(y_n^N) + \gamma_n \tilde{\eta}_{n+1}^N + \gamma_n \varrho_{n+1}^N + \gamma_n \tilde{\varrho}_{n+1}^N\,,$$

where the random sequences $(\tilde{\eta}_n^N)_{n\geqslant N}$, $(\varrho_n^N)_{n\geqslant N}$, and $(\tilde{\varrho}_n^N)_{n\geqslant N}$ are adapted to $(\mathscr{F}_n)$. Moreover, there is $C > 0$ such that for all $n \geqslant N$,

i) $\left\|\varrho_{n+1}^N\right\| \leqslant C\gamma_n(1 + \|\eta_{n+1}\|^2)\mathbb{1}_{\tau_N > n+1}$.

ii) $\left\|\tilde{\varrho}_{n+1}^N\right\| \leqslant C \left\|z_n^N\right\| (1 + \|\eta_{n+1}\|)$.

iii) $\mathbb{E}_n \tilde{\eta}_{n+1}^N = 0$, *and* $\mathbb{E}_n \left\|\tilde{\eta}_{n+1}^N\right\|^4 < C$.

We furthermore have:

iv) The subspace E^- *defined before Assumption 3 coincides with the eigenspace of the matrix* $J_D(0)$ *corresponding to its negative eigenvalues.*

v) On the event $[y_n^N \to_n 0]$, *it holds that* $\liminf_n \mathbb{E}_n \left\|P_{E^-} \tilde{\eta}_{n+1}^N\right\| > 0$.

To prove this proposition, the following result will be needed.

Lemma 5. *For* r *small enough, there is* $C > 0$ *such that for* $x, x' \in B(0, r)$, *we have:*

$$y' - y = J_{P_M}(y)(x' - x) + R_1(x, x', y) + R_2(x, x')\,,$$

where $y', y = P_M(x'), P_M(x)$, *and where* $\|R_1(x, x', y)\| \leqslant C \|x' - x\| \|x - y\|$, *and* $\|R_2(x, x')\| \leqslant C \|x' - x\|^2$.

Proof. Since P_M is C^2 near zero, there is $\varepsilon > 0$ such that $t \mapsto P_M(x + t(x' - x))$ is C^2 on $(-\varepsilon, 1 + \varepsilon)$. Hence, by Taylor's theorem, we have

$$y' - y = J_{P_M}(x)(x' - x) + R_2(x', x)\,,$$

with $\|R_2(x', x)\| \leqslant C \|x' - x\|^2$, where C is a bound on the second derivatives of P_M. Similarly, since P_M is C^2, $x \mapsto J_{P_M}(x)$ is Lipschitz continuous. Therefore, for some $C > 0$, $\|J_{P_M}(x) - J_{P_M}(y)\| \leqslant C \|x - y\|$, which finishes the proof. □

Proof of Proposition 5. Letting $n \geqslant N$, we write

$$y_{n+1}^N = P_M(x_{n+1}) \mathbb{1}_{\tau_N > n+1} + \left(y_n^N - \gamma_n D(y_n^N)\right) \mathbb{1}_{\tau_N \leqslant n+1} + \gamma_n \left(J_{P_M}(y_n^N) \mathbb{1}_{\tau_N = n+1} + \mathbb{1}_{\tau_N \leqslant n} \eta_{n+1}\right),$$

accepting the small notational abuse in the expression $P_M(x_{n+1}) \mathbb{1}_{\tau_N > n+1}$, since the projection might not be defined when the indicator is zero. Similar abuses will also be made in the derivations below.

Using Lemma 5 and Equation (8), we obtain

$$\begin{aligned}
y_{n+1}^N &= \left(y_n^N + J_{P_M}(y_n^N)(x_{n+1} - x_n)\right) \mathbb{1}_{\tau_N > n+1} + \gamma_n \varrho_{n+1}^N + \gamma_n \zeta_{n+1}^N \\
&\quad + \left(y_n^N - \gamma_n D(y_n^N)\right) \mathbb{1}_{\tau_N \leqslant n+1} + \gamma_n \left(J_{P_M}(y_n^N) \mathbb{1}_{\tau_N = n+1} + \mathbb{1}_{\tau_N \leqslant n} \eta_{n+1}\right) \\
&= \left(y_n^N - \gamma_n J_{P_M}(y_n^N) v_n + \gamma_n J_{P_M}(y_n^N) \eta_{n+1}\right) \mathbb{1}_{\tau_N > n+1} + \gamma_n \varrho_{n+1}^N + \gamma_n \zeta_{n+1}^N \\
&\quad + \left(y_n^N - \gamma_n D(y_n^N)\right) \mathbb{1}_{\tau_N \leqslant n+1} + \gamma_n \left(J_{P_M}(y_n^N) \mathbb{1}_{\tau_N = n+1} + \mathbb{1}_{\tau_N \leqslant n} \eta_{n+1}\right),
\end{aligned}$$

where ϱ_{n+1}^N and ζ_{n+1}^N are $\mathscr{F}_{n+1}$–measurable, and satisfy with the notations of Lemma 5

$$\left\|\zeta_{n+1}^N\right\| = \gamma_n^{-1} \left\|R_1(x_n, x_{n+1}, y_n^N)\right\| \mathbb{1}_{\tau_N > n+1} \leqslant C \gamma_n^{-1} \|x_{n+1} - x_n\| \left\|z_n^N\right\| \leqslant C(1 + \|\eta_{n+1}\|) \left\|z_n^N\right\|$$

(in the last inequality, we used that $\|v_n\|$ is bounded on $[\tau_N > n]$), and

$$\begin{aligned}
\left\|\varrho_{n+1}^N\right\| &= \gamma_n^{-1} \|R_2(x_n, x_{n+1})\| \mathbb{1}_{\tau_N > n+1} \\
&\leqslant C \gamma_n^{-1} \|x_{n+1} - x_n\|^2 \mathbb{1}_{\tau_N > n+1} \\
&\leqslant C \gamma_n (1 + \|\eta_{n+1}\|^2) \mathbb{1}_{\tau_N > n+1}.
\end{aligned}$$

Using Lemma 1 in conjunction with the Verdier condition (iii) of Definition 6, we also have

$$J_{P_M}(y_n^N)v_n\mathbb{1}_{\tau_N>n+1} = P_{T_{y_n^N}M}(v_n)\mathbb{1}_{\tau_N>n+1} = \nabla_M f(y_n^N)\mathbb{1}_{\tau_N>n+1} + \tilde{\zeta}_{n+1}^N = D(y_n^N)\mathbb{1}_{\tau_N>n+1} + \tilde{\zeta}_{n+1}^N,$$

where $\tilde{\zeta}_{n+1}^N$ is $\mathscr{F}_{n+1}$–measurable, and satisfies

$$\left\|\tilde{\zeta}_{n+1}^N\right\| \leqslant C\left\|x_n - y_n^N\right\|\mathbb{1}_{\tau_N>n+1} \leqslant C\left\|z_n^N\right\|.$$

Gathering these expressions, we get

$$y_{n+1}^N = y_n^N - \gamma_n D(y_n^N) + \gamma_n\tilde{\eta}_{n+1}^N + \gamma_n\varrho_{n+1} + \gamma_n\tilde{\varrho}_{n+1},$$

where

$$\begin{aligned}\tilde{\eta}_{n+1}^N &= \left(\mathbb{1}_{\tau_N>n}J_{P_M}(y_n^N) + \mathbb{1}_{\tau_N\leqslant n}\right)\eta_{n+1}, \text{ and}\\ \tilde{\varrho}_{n+1}^N &= \zeta_{n+1}^N + \tilde{\zeta}_n^N.\end{aligned} \tag{14}$$

The assertions i) and ii) of the statement are obtained from what precedes.

The noise $\tilde{\eta}_n^N$ is obviously $\mathscr{F}_n$–measurable. Moreover, $\mathbb{E}_n\tilde{\eta}_{n+1}^N = 0$ since $\mathbb{1}_{\tau_N>n}J_{P_M}(y_n^N) + \mathbb{1}_{\tau_N\leqslant n}$ is $\mathscr{F}_n$–measurable. The last bound in iii) follows from Assumption 4.

Assertion iv) follows from Lemma 4.

To establish v), we write

$$\begin{aligned}\left\|(\tilde{\eta}_{n+1}^N)^-\right\| &= \left\|P_{E^-}J_{P_M}(y_n^N)\eta_{n+1}\right\|\mathbb{1}_{\tau_N>n} + \left\|P_{E^-}\eta_{n+1}\right\|\mathbb{1}_{\tau_N\leqslant n}\\ &\geqslant \left\|P_{E^-}\eta_{n+1}\right\| - \left\|P_{E^-}J_{P_M}(y_n^N)\eta_{n+1} - P_{E^-}\eta_{n+1}\right\|\mathbb{1}_{\tau_N>n}.\end{aligned}$$

On the event $[y_n^N \to_n 0]$, it holds that $J_{P_M}(y_n^N) \to_n J_0$. By Lemma 1, J_0 is the orthogonal projection on T_0M, thus, $\lim_{y_n^N\to_n 0} P_{E^-}J_{P_M}(y_n^N) = P_{E^-}$. Consequently, we obtain on the event $[y_n^N \to_n 0]$:

$$\begin{aligned}\liminf_n \mathbb{E}_n\left\|(\tilde{\eta}_{n+1}^N)^-\right\| &\geqslant \liminf_n \mathbb{E}_n\left\|\eta_{n+1}^-\right\| - \limsup_n\left(\left\|P_{E^-}J_{P_M}(y_n^N) - P_{E^-}\right\|\mathbb{E}_n\left\|\eta_{n+1}\right\|\right)\\ &\geqslant \liminf_n \mathbb{E}_n\left\|\eta_{n+1}^-\right\|\\ &> 0,\end{aligned}$$

and by Assumption 4. Proposition 5 is proven. □

Proposition 6. *Let Assumptions 1–2 and 4 hold true. Then, there is $C > 0$ such that*

$$\begin{aligned}\mathbb{E}_n\left\|z_{n+1}^N\right\|^2 &\leqslant \left\|z_n^N\right\|^2 - \gamma_n\left(\frac{2\beta}{r} - C\right)\left\|z_n^N\right\|^2 + C\gamma_n^2, \text{ and}\\ \mathbb{E}_n\left\|z_{n+1}^N\right\|^2 &\leqslant \left\|z_n^N\right\|^2 - \gamma_n\left(2\beta - Cr\right)\left\|z_n^N\right\| + C\gamma_n^2.\end{aligned}$$

Proof. We shall use the notation

$$p_n^N = x_n - y_n^N,$$

which enables us to write $z_n^N = p_n^N\mathbb{1}_{n<\tau_N}$.

We start with the development

$$\begin{aligned}\left\|z_{n+1}^N\right\|^2 &= \left\|p_{n+1}^N\right\|^2 \mathbb{1}_{n+1<\tau_N} \\ &\leqslant \left\|p_{n+1}^N\right\|^2 \mathbb{1}_{n<\tau_N} = \left\|p_{n+1}^N - p_n^N + p_n^N\right\|^2 \mathbb{1}_{n<\tau_N} \\ &= \left\|z_n^N\right\|^2 + 2\langle x_{n+1} - x_n, z_n^N\rangle - 2\langle y_{n+1}^N - y_n^N, z_n^N\rangle + \left\|p_{n+1}^N - p_n^N\right\|^2 \mathbb{1}_{n<\tau_N}. \end{aligned} \tag{15}$$

We now deal separately with each of the three rightmost terms in the last expression.

We first show that

$$\mathbb{E}_n|\langle y_{n+1}^N - y_n^N, z_n^N\rangle| \leqslant C\gamma_n \left\|z_n^N\right\|^2 + C\gamma_n^2. \tag{16}$$

By Proposition 5,

$$\langle y_{n+1}^N - y_n^N, z_n^N\rangle = \gamma_n\langle -D(y_n^N) + \tilde{\eta}_{n+1}^N + \varrho_{n+1}^N + \tilde{\varrho}_{n+1}^N, z_n^N\rangle.$$

We have $\langle D(y_n^N), z_n^N\rangle = \langle \nabla_M f(y_n^N), z_n^N\rangle = 0$ since $\nabla_M f(y_n^N) \in T_{y_n^N}M$. Furthermore, we get from Equation (14) that

$$\mathbb{1}_{n<\tau_N}\tilde{\eta}_{n+1}^N = \mathbb{1}_{n<\tau_N} J_{P_M}(y_n^N)\eta_{n+1} = \mathbb{1}_{n<\tau_N} P_{T_{y_n^N}M}(\eta_{n+1})$$

by Lemma 1, thus, $\langle \tilde{\eta}_{n+1}^N, z_n^N\rangle = 0$. As a consequence,

$$|\langle y_{n+1}^N - y_n^N, z_n^N\rangle| \leqslant \gamma_n(\left\|z_n^N\right\|^2 + \left\|\varrho_{n+1}^N + \tilde{\varrho}_{n+1}^N\right\|^2) \leqslant \gamma_n \left\|z_n^N\right\|^2 + 2\gamma_n(\left\|\varrho_{n+1}^N\right\|^2 + \left\|\tilde{\varrho}_{n+1}^N\right\|^2).$$

From Proposition 5 again, we have

$$\mathbb{E}_n \left\|\varrho_{n+1}^N\right\|^2 \leqslant C\gamma_n\mathbb{E}_n(1 + \|\eta_{n+1}\|^2)\mathbb{1}_{\tau_N>n+1} \leqslant C\gamma_n\mathbb{E}_n(1 + \|\eta_{n+1}\|^2) \leqslant C\gamma_n\,,$$

and

$$\mathbb{E}_n \left\|\tilde{\varrho}_{n+1}^N\right\|^2 \leqslant C \left\|z_n^N\right\|^2 (1 + \mathbb{E}_n \|\eta_{n+1}\|^2) \leqslant C \left\|z_n^N\right\|^2.$$

Inequality (16) is obtained by combining these inequalities.

We next show succinctly that

$$\mathbb{E}_n \left\|p_{n+1}^N - p_n^N\right\|^2 \mathbb{1}_{n<\tau_N} \leqslant C\gamma_n^2. \tag{17}$$

Indeed,

$$\begin{aligned}\left\|p_{n+1}^N - p_n^N\right\|^2 \mathbb{1}_{n<\tau_N} &= \left\|x_{n+1} - x_n - (y_{n+1}^N - y_n^N)\right\|^2 \mathbb{1}_{n<\tau_N} \\ &\leqslant C\gamma_n^2 \left(\|v_n\|^2 + \|\eta_{n+1}\|^2 + \left\|D(y_n^N)\right\|^2 + \left\|\tilde{\eta}_{n+1}^N\right\|^2 + \left\|\varrho_{n+1}^N\right\|^2 + \left\|\tilde{\varrho}_{n+1}^N\right\|^2\right) \mathbb{1}_{n<\tau_N},\end{aligned}$$

and the result follows by standard calculations making use of the results of Proposition 5.

We finally deal with the term $\langle x_{n+1} - x_n, z_n^N\rangle$. Since $\mathbb{E}_n\eta_{n+1} = 0$, we have $\mathbb{E}_n\langle x_{n+1} - x_n, z_n^N\rangle = -\gamma_n\langle v_n, z_n^N\rangle$. Observing that $x_n \in V_r(\beta)$ when $z_n^N \neq 0$, we obtain from the very definition of the set $V_r(\beta)$ that

$$\mathbb{E}_n\langle x_{n+1} - x_n, z_n^N\rangle \leqslant -\gamma_n\beta \left\|z_n^N\right\|.$$

Getting back to Inequality (15), and using this result in conjunction with the inequalities (16) and (17), we obtain that

$$\mathbb{E}_n \left\|z_{n+1}^N\right\|^2 \leqslant \left\|z_n^N\right\|^2 + C\gamma_n \left\|z_n^N\right\|^2 - 2\gamma_n\beta \left\|z_n^N\right\| + C\gamma_n^2.$$

Since $x_n \in B(0, r)$ on the event $[n < \tau_N]$, it holds that $\left\|z_n^N\right\| \leqslant r$ and thus, $\left\|z_n^N\right\|^2 \leqslant r \left\|z_n^N\right\|$. This leads at once to the inequalities in the statement of the proposition. □

Corollary 2. *Under the assumptions of the previous proposition, there is $C > 0$ such that*

$$\sum_{i=n}^{\infty} \gamma_i \mathbb{E} \left\| z_i^N \right\| \leqslant C\chi_n$$

for $n \geqslant N$.

The proof of this corollary makes use of a technical result which is attributed to [11]. Its proof can be found in, *e.g.*, [10]:

Lemma 6 (Lemma D.2 in [10])**.** *Let (a_n) be a nonnegative sequence such that for all n large enough,*

$$a_{n+1} \leqslant a_n \left(1 - \frac{P}{n^p}\right) + \frac{Q}{n^{p+q}},$$

where $p \in (0, 1]$, $q > 0$, and $P, Q > 0$. It is further assumed that $P > q$ if $p = 1$. Then, there exists $C > 0$ such that

$$a_n \leqslant \frac{C}{n^q}.$$

Proof of Corollary 2. Let $C > 0$ be the constant provided in the statement of Proposition 6. Choose $r > 0$ small enough so that $2\beta r^{-1} - C > 0$. Replacing γ_n in this statement with the bounds on this step size provided by Assumption 1–(iv), we get from the first inequality in Proposition 6

$$\mathbb{E} \left\| z_{n+1}^N \right\|^2 \leqslant \left(1 - \frac{c_1}{n^\alpha} \left(\frac{2\beta}{r} - C\right)\right) \mathbb{E} \left\| z_n^N \right\|^2 + \frac{c_2 C}{n^{2\alpha}}.$$

We apply the previous lemma with $a_n = \mathbb{E} \left\| z_n^N \right\|^2$, after adjusting $r > 0$ when needed in order that all the conditions in the statement of this lemma are satisfied. We get that there exists a constant $C' > 0$ such that

$$\mathbb{E} \left\| z_n^N \right\|^2 \leqslant \frac{C'}{n^\alpha}.$$

Let $k > 0$ be an integer. Telescoping the second inequality stated by Proposition 6 from $n + k$ back to n, we get

$$\mathbb{E} \left\| z_{n+k}^N \right\|^2 \leqslant \mathbb{E} \left\| z_n^N \right\|^2 - (2\beta - Cr) \sum_{i=n}^{n+k-1} \gamma_i \mathbb{E} \left\| z_i^N \right\| + C \sum_{i=n}^{n+k-1} \gamma_i^2,$$

which implies that

$$(2\beta - Cr) \sum_{i=n}^{n+k-1} \gamma_i \mathbb{E} \left\| z_i^N \right\| \leqslant \mathbb{E} \left\| z_n^N \right\|^2 + C \sum_{i=n}^{n+k-1} \gamma_i^2 \leqslant \frac{C'}{n^\alpha} + C \sum_{i=n}^{n+k-1} \gamma_i^2.$$

Making $k \to \infty$, we obtain that

$$(2\beta - Cr) \sum_{i=n}^{\infty} \gamma_i \mathbb{E} \left\| z_i^N \right\| \leqslant \frac{C'}{n^\alpha} + C\chi_n.$$

To complete the proof, it remains to notice that since $\gamma_n \sim n^{-\alpha}$ with $\alpha \in (1/2, 1]$, it holds that $\chi_n \sim n^{1-2\alpha} \gtrsim n^{-\alpha}$. □

Theorem 3: end of the proof. We now have all the elements to establish the identity (13), proving Theorem 3. For this, notice that,for every $N \geqslant 0$, by Proposition 5, y_n^N satisfies an equation of the form Equation (11). The assumption of Proposition 3 on the sequence $(\tilde{\eta}_n)$ are satisfied by Proposition 5 and the assumptions on the sequences $(\varrho_n), (\tilde{\varrho}_n)$ follow from Assumption 4 and Corollary 2.

Hence, applying Proposition 3, we obtain that $\mathbb{P}([y_n^N \to 0]) = 0$, for all $N \geqslant 0$. As previously explained, the latter implies that $\mathbb{P}([x_n \to 0]) = 0$.

To complete the proof of Theorem 3 it remains to prove Proposition 4, which is the purpose of the next section.

5.3 Proof of Proposition 4

The standard way to analyze the convergence of the SGD to the set of Clarke critical points is by studying its continuous counterpart - the subgradient flow:

$$\dot{\mathsf{x}}(t) \in -\partial f(\mathsf{x}(t))\,. \tag{18}$$

We say that an absolutely continuous curve $\mathsf{x} : \mathbb{R}_+ \to \mathbb{R}$ is a solution of the differential inclusion (DI) (18) starting at $x \in \mathbb{R}^d$ if $\mathsf{x}(0) = x$ and if for almost every $t \in \mathbb{R}_+$, the inclusion (18) is verified. We denote $\mathsf{S}_{-\partial f}(x)$ the set of these solutions.

The idea of the proof of Proposition 4 goes as follows. For each initial point $x \in B(0, r_0)$ with $r_0 > 0$ small enough, either all the trajectories of (18) issued from x leave $B(0, r_0)$ in a fixed time horizon, or $f(x) - f(P_M(x)) \geqslant \alpha \|x - P_M(x)\|$. This will be the content of the next lemma. Next, we use the well-known fact that the interpolated process constructed from our iterates (x_n) is a so-called *Asymptotic Pseudo Trajectory* (APT) of the DI (18), as formalized in [2] (see also, *e.g.*, [19, 29]). The consequence is that on the event $[x_n \to 0]$, necessarily $f(x_n) - f(P_M(x_n)) \geqslant \alpha \|x_n - P_M(x_n)\|$ after a certain finite moment. To complete the proof, it remains to make use of the angle condition (iv) of Definition 6.

Lemma 7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous, path differentiable function. Let M be a C^2 active manifold for f such that $0 \in M$, $f(0) = 0$, and $\nabla_M f(0) = 0$. Then, there is $\alpha, T > 0$ and $r_0 > 0$ s.t. for every $\mathsf{x} \in \mathsf{S}_{-\partial F}(x)$, with $x \in B(0, r_0)$, either $\mathsf{x}([0, T]) \not\subset B(0, r_0)$ or $f(x) - f(P_M(x)) \geqslant \alpha \|x - P_M(x)\|$.*

Proof. Let $r > 0$ be such that $B(0, r) \subset U$, where U is the neighborhood from Definition 4. Since f is C^2 on $M \cap B(0, r)$ and $\nabla_M f(0) = 0$, there is some constant C s.t. we have $\sup_{x \in B(0,r)} \|\nabla_M f(P_M(x))\| \leqslant C \|P_M(x)\|$. Denote L the Lipschitz constant of f on $B(0, r)$ and let c_m be such that $\inf\{\|v\| : v \in \partial f(x), x \in B(0, r) \cap M^c\} \geqslant c_m$. Fix $r_0 \leqslant \min(\frac{c_m^2}{2LC}, r)$ and consider $x \in B(0, r_0)$ and $\mathsf{x} \in \mathsf{S}_{-\partial F}(x)$. Denote $t_1 = \inf\{t : \mathsf{x}(t) \in M \text{ or } \mathsf{x}(t) \notin B(0, r_0)\}$. Since f is path differentiable, we have:

$$\inf_{x' \in B(0,r_0)} f(x') \leqslant f(\mathsf{x}(t)) = f(x) - \int_0^t \|\dot{\mathsf{x}}(u)\| \,\mathrm{d}u \leqslant f(x) - c_m^2 t_1 \leqslant \sup_{x' \in B(0,r_0)} f(x') - c_m^2 t_1\,.$$

Hence, if we choose T s.t. $c_m^2 T > 2 \sup_{x' \in B(0,r_0)} |f(x')|$, we have $t_1 \leqslant T$ and either $\mathsf{x}(t_1) \notin B(0, r)$ or $\mathsf{x}(t_1) \in M$. Assume that $\mathsf{x}(t_1) \in M$ and denote $\mathsf{y}(t) = P_M(\mathsf{x}(t))$ and $\mathsf{z}(t) = \mathsf{x}(t) - \mathsf{y}(t)$.

Notice that for almost every $t \geqslant 0$, we have $\|\dot{\mathsf{y}}(t)\| = \left\| P_{T_{\mathsf{y}(t)}} \dot{\mathsf{x}}(t) \right\| \leqslant L$. Moreover, by path-differentiability of f we have:

$$\begin{aligned} |f(\mathsf{y}(t_1)) - f(\mathsf{y}(0))| &\leqslant \int_0^{t_1} |\langle \nabla_M f(\mathsf{y}(u)), \dot{\mathsf{y}}(u) \rangle| \, \mathrm{d}u \\ &\leqslant \int_0^{t_1} \|\nabla_M f(\mathsf{y}(u))\| \, \|\dot{\mathsf{y}}(u)\| \, \mathrm{d}u \\ &\leqslant C \int_0^{t_1} \|\mathsf{y}(u)\| \, \|\dot{\mathsf{y}}(u)\| \, \mathrm{d}u \\ &\leqslant LCr_0 t_1 \leqslant \frac{1}{2} c_m^2 t_1 \, . \end{aligned}$$

Where the first inequality comes from the fact that f is path differentiable and that for all $u \in [0,T]$, $\dot{\mathsf{y}}(u) \in T_{\mathsf{y}(u)} M$. Denote $\alpha = \frac{c_m^2}{4L}$ and assume by contradiction that $f(x) - f(P_M(x)) \leqslant \alpha \, \|x - P_M(x)\|$. We have:

$$\begin{aligned} 0 = f(\mathsf{x}(t_1)) - f(\mathsf{y}(t_1)) &\leqslant f(x) - c_m^2 t_1 - f(\mathsf{y}(t_1)) \\ &\leqslant f(x) - f(\mathsf{y}(0)) + \frac{c_m^2}{2} t_1 - c_m^2 t_1 \\ &\leqslant \alpha \, \|x - P_M(x)\| - \frac{c_m^2}{2} t_1 \, . \end{aligned}$$

Which implies that $\|x - P_M(x)\| \geqslant \frac{c_m^2}{2\alpha} t_1 \geqslant 2Lt_1$. On the other hand, we have that $\|\mathsf{z}(t)\| = \mathrm{dist}(\mathsf{x}(t), M)$. Since the distance function is 1-Lipschitz, we have for almost every $t \geqslant 0$:

$$\left| \frac{d}{dt} \|\mathsf{z}(t)\| \right| \leqslant \|\dot{\mathsf{x}}(t)\| \leqslant L \, .$$

Therefore,

$$0 = \|\mathsf{z}(t_1)\| \geqslant \|\mathsf{z}(0)\| - Lt_1 = \|x - P_M(x)\| - Lt_1 \, ,$$

which implies that $\|x - P_M(x)\| \leqslant Lt_1$, a contradiction. □

Let $\mathsf{X} : \mathbb{R}_+ \to \mathbb{R}^d$ be the linearly interpolated process defined as:

$$\mathsf{X}(t) = x_n + \frac{t - \sum_{i=0}^n \gamma_i}{\gamma_{n+1}} (x_{n+1} - x_n), \quad \text{if } t \in [\tau_n, \tau_{n+1}) \, ,$$

where $\tau_n = \sum_{i=0}^n \gamma_i$.

It is well known that under our assumptions, on the event $[x_n \to 0]$, X is an APT for the DI (18), as shown in [2, 19, 29]. Namely, for every $T > 0$,

$$\sup_{h \in [0,T]} \inf_{\mathsf{x} \in \mathsf{S}_{-\partial f}(\mathsf{X}(t))} \|\mathsf{X}(t+h) - \mathsf{x}(h)\| \xrightarrow[t \to +\infty]{} 0 \, .$$

Consider α, T and r_0 from Lemma 7. On the event $[x_n \to 0]$ let $\mathsf{x}_n \in \mathsf{S}_{-\partial F}(x_n)$ be such that

$$\sup_{h \in [0,T]} \|\mathsf{X}(\tau_n + h) - \mathsf{x}_n(h)\| \xrightarrow[n \to +\infty]{} 0 \, .$$

Consider $r_1 \leqslant r_0$ such that $B(0, r_1) \subset U$, where U is the neighborhood associated to α by the angle condition. If for n large enough, $\mathsf{x}_n([0,T])$ remains in $B(0, r_1)$, then by Lemma 7 we have:

$$f(x_n) \geqslant \alpha \left\| x_n - P_M(x_n) \right\| + f(P_M(x_n)) \,,$$

which, by the angle condition, implies that there is $\beta > 0$

$$\langle v_n, x_n - P_M(x_n) \rangle \geqslant \beta \left\| x_n - P_M(x_n) \right\| \,. \tag{19}$$

Otherwise, on the event $[x_n \to 0]$, there is $h_n \in [0,T]$ such that after an extraction $\mathsf{X}(\tau_n + h_n) \to x$, with $x \notin B(0, r_1)$. Since the limit points of X are the accumulation points of the sequence (x_n), this contradicts the fact that $x_n \to 0$.

Appendix A Sketch of proof of Proposition 3

We recall that $\mathbb{E}_n[\cdot]$ denotes $\mathbb{E}[\cdot|\mathscr{F}_n]$. Denote d^- the dimension of Λ^-. Using the center-stable manifold theorem, the authors of [9, Page 407–409] construct a sequence (w_n)[3] in $\mathbb{R}^{d^-}$ such that

$$w_n = w_n + \gamma_n H_n w_n + \gamma_n (r_{n+1} + r'_{n+1} + e_{n+1}) \,,$$

where the sequences $(w_n), (r_n), (r'_n), (e_n)$ are adapted to $(\mathscr{F}_n)$ and we have the inclusion $[y_n \to 0] \subset [w_n \to 0]$. Moreover, on the event $[y_n \to 0]$, the following almost surely holds.

i) There is H an invertible matrix such that all of the real parts of its eigenvalues are positive and
$$H_n \to H \,.$$

ii) The sequence (e_n) is such that $\mathbb{E}_n[e_{n+1}] = 0$ and
$$0 < \liminf \mathbb{E}_n[\|e_{n+1}\|^2] \leqslant \limsup \mathbb{E}_n[\|e_{n+1}\|^2] < +\infty \,.$$

iii) The sequence (r_n) is such that $\sum_{i=0}^{+\infty} \|r_{i+1}\|^2 < +\infty$.

iv) The sequence (r'_n) is such that $\mathbb{E}[\mathbb{1}_\Gamma \sum_{i=n}^{+\infty} \gamma_i \left\| r'_{i+1} \right\|] = \mathcal{O}(\chi_n)$.

The only difference with [9] is in the presence of (r'_{n+1}) and the point (iv).

Using this representation, the avoidance of traps result follows from the following proposition. The only difference with [9, Proposition 4] is, once again, in the presence of the sequence (r'_n).

Proposition 7 ([9, Proposition 4])**.** *Let d be an integer, $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(\mathscr{F}_n)$ a filtration on it and (w_n) be a sequence in $\mathbb{R}^d$ verifying:*

$$w_{n+1} = w_n + \gamma_n H_n + \gamma_n (r_{n+1} + r'_{n+1} + e_{n+1}) \,, \tag{20}$$

where the sequences $(w_n), (H_n), (r_n), (r'_n)(e_n)$ are adapted to $(\mathscr{F}_n)$ and (γ_n) is a sequence of positive stepsizes s.t. $\sum_{i=0}^{+\infty} \gamma_i = +\infty$ and $\sum_{i=0}^{+\infty} \gamma_i^2 < +\infty$. Assume that on an event $\Gamma \in \mathcal{A}$ we have the following.

[3] U_n^+ in their notations.

i) *The sequence* (γ_n) *is such that* $\sum_{i=0}^{+\infty} \gamma_i = +\infty$ *and* $\sum_{i=0}^{+\infty} \gamma_i^2 < +\infty$.

ii) *The sequence* (e_n) *is such that* $\mathbb{E}_n[e_{n+1}] = 0$ *and*

$$0 < \liminf \mathbb{E}_n[\|e_{n+1}\|] \leqslant \limsup \mathbb{E}_n[\|e_{n+1}\|^2]^{1/2} < +\infty\,.$$

iii) *The sequence* (r_n) *is such that* $\sum_{i=0}^{+\infty} \|r_{i+1}\|^2 < +\infty$.

iv) *The sequence* (r'_n) *is such that* $\mathbb{E}\left[\mathbb{1}_\Gamma \sum_{i=n}^{+\infty} \gamma_i \left\|r'_{i+1}\right\|\right] = \mathcal{O}(\chi_n)$.

Let $H \in \mathbb{R}^{d\times d}$ *be a matrix such that all of the real parts of its eigenvalues are positive. Then, denoting* $\Upsilon = \Gamma \cap [w_n \to 0] \cap [H_n \to H]$, *we have* $\mathbb{P}(\Upsilon) = 0$.

Proof. In this proof C will denote some absolute constant that can change from line to line. The proof closely follows the one of [9, Proposition 4]. As in [9] it is sufficient to prove the proposition in the case where there $A, B, K > 0$ such that almost surely $\mathbb{E}_n[e_{n+1}] = 0$, $A \leqslant \mathbb{E}_n[\|e_{n+1}\|] \leqslant \mathbb{E}_n[\|e_{n+1}\|^2]^{1/2} \leqslant B$ and $\sum_{i=0}^{+\infty} \|r_{i+1}\|^2 \leqslant K$.
We can rewrite Equation (20) as:

$$w_{n+1} = w_n + \gamma_n H w_n + \gamma_n \Delta_n w_n + \gamma_n (e_{n+1} + r_{n+1} + r'_{n+1})\,,$$

where $\Delta_n = H_n - H$. Let Q be a positive definite symmetric matrix such that $QH + H^T Q = 2\mathcal{I}$, where $\mathcal{I} \in \mathbb{R}^{d\times d}$ is the identity matrix. Denote $U_n = (w_n^T Q w_n)^{1/2}$. Following the same calculations as in [9], we obtain that:

$$\begin{aligned}
(U_{n+1} - U_n) &\geqslant \frac{1}{U_n} w_{n+1}^T Q w_n \\
&\geqslant \frac{\gamma_n}{U_n}\left(\|w_n\|^2 + w_n^T Q \Delta_n w_n + w_n^T Q (e_{n+1} + r_{n+1} + r'_{n+1})\right) \\
&\geqslant \gamma_n \|w_n\| \left(\frac{1}{\lambda_{max}^{1/2}} - \frac{\|Q\Delta_n\|}{\lambda_{min}^{1/2}}\right) + \frac{\gamma_n w_n^T Q (e_{n+1} + r_{n+1} + r'_{n+1})}{U_n}\,,
\end{aligned}$$

where $\lambda_{max}, \lambda_{min}$ are respectively the maximal and the minimal eigenvalue of Q. The event Υ is included in a union of events Υ_p defined as:

$$\Upsilon_p = \Upsilon \cap \left[\forall n \geqslant p, \frac{1}{\lambda_{max}^{1/2}} - \frac{\|Q\Delta_n\|}{\lambda_{min}^{1/2}} \geqslant \frac{1}{2\lambda_{max}^{1/2}}\right] \cap \left[\sup_{n\geqslant p} \|w_n\| \leqslant 1\right] \cap \left[\sum_{i=p}^{+\infty} \gamma_i \left\|r'_{i+1}\right\| < 1\right].$$

Therefore, on Υ_p, there is $C > 0$ such that for $M \geqslant n \geqslant p$, we have:

$$\sum_{i=n}^{M} \gamma_i \|w_i\| \leqslant C U_{M+1} + C \left\|\sum_{i=n}^{M} \gamma_i \frac{w_i^T Q (e_{i+1} + r_{i+1} + r'_{i+1})}{U_i}\right\|.$$

Hence,

$$\begin{aligned}
\left|\sum_{i=n}^{M} \gamma_i \|w_i\|\right|^2 &\leqslant C \|U_{M+1}\|^2 + C \left\|\sum_{i=n}^{M} \gamma_i \frac{w_i^T Q e_{i+1}}{U_i}\right\|^2 + C \left(\sum_{i=n}^{+\infty} \gamma_i^2\right)\left(\sum_{i=n}^{+\infty} \|r_{i+1}\|^2\right) + C \left|\sum_{i=n}^{+\infty} \gamma_i \left\|r'_{i+1}\right\|\right|^2 \\
&\leqslant C \|U_{M+1}\|^2 + C \sup_{M\geqslant p} \left\|\sum_{i=n}^{M} \gamma_i \frac{w_i^T Q e_{i+1}}{U_i}\right\|^2 + C\chi_n + C \left|\sum_{i=n}^{+\infty} \gamma_i \left\|r'_{i+1}\right\|\right|^2,
\end{aligned}$$

where we used the fact that $\frac{\|w_n^T Q\|}{U_n}$ is bounded. On Υ_p we have that $\mathbb{E}[\|U_{M+1}\|^2] \to 0$. The sequence $(\sum_{i=n}^{M} \gamma_i \frac{w_i^T Q e_{i+1}}{U_i})_{M \geqslant n}$ is a square summable martingale difference sequence. Therefore, by Doob's maximal inequality:

$$\mathbb{E}\left[\mathbb{1}_\Gamma \sup_{M\in\mathbb{N}} \left|\sum_{i=n}^{M} \gamma_i \frac{w_i^T Q e_{i+1}}{U_i}\right|^2\right] \leqslant C\mathbb{E}\left[\sum_{i=n}^{+\infty} \gamma_i^2 \|e_{i+1}\|^2\right] \leqslant C\chi_n\,.$$

Finally, on Υ_p we have $\sum_{i=n}^{+\infty} \gamma_i \left\|r'_{i+1}\right\| < 1$. Therefore, by assumptions:

$$\mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \left|\sum_{i=n}^{+\infty} \gamma_i \left\|r'_{i+1}\right\|\right|^2\right] \leqslant \mathbb{E}\left[\mathbb{1}_\Upsilon \sum_{i=n}^{+\infty} \gamma_i \left\|r'_{i+1}\right\|\right] \leqslant C\chi_n$$

Hence, there is $C > 0$ such that:

$$\mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \left|\sum_{i=n}^{+\infty} \gamma_i \|w_i\|\right|^2\right] \leqslant C\chi_n\,. \tag{21}$$

On the other hand, following the calculations of [9], on Υ_p we have:

$$-w_p = \sum_{i=p}^{+\infty} \left(R_i^1 + \gamma_i(e_{i+1} + r_{i+1} + r'_{i+1})\right)\,, \tag{22}$$

where we denote $R_n = \Delta_n w_n$ and for $n \geqslant p$:

$$\begin{aligned}
R_n^1 &= \gamma_n R_n - (B_{n-1}^{-1} - B_n^{-1}) S_n\,,\\
S_n &= \sum_{i=n}^{+\infty} \gamma_i (R_i + e_{i+1} + r_{i+1} + r'_{i+1})\,,\\
B_n &= \prod_{i=p}^{n} (1 + \gamma_i H)\,.
\end{aligned}$$

The idea of the remaining part of the proof is to apply [9, Theorem A] to obtain that the left hand side of Equation 22 can be $\mathscr{F}_p$-measurable only with probability 0. The latter will imply $\mathbb{P}(\Upsilon_p) = 0$ and since $\Upsilon = \bigcup_{p\in\mathbb{N}} \Upsilon_p$, the proof will be finished. As in the proof [9], one of the assumptions of [9, Theorem A], to obtain the remaining part it suffices to have:

$$\mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \sum_{i=n}^{+\infty} \left\|R_i^1 + \gamma_i r'_{i+1}\right\|\right] = o(\sqrt{\chi_n})\,, \tag{23}$$

where the difference with the proof of [9, Proposition 4] is in the presence of the term r'_{i+1}. To prove Equation (23) we write down:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \sum_{i=n}^{+\infty} \left\|R_i^1 + \gamma_i r'_{i+1}\right\|\right] \leqslant\; & C\mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \sup_{i\geqslant n} \|\Delta_i\| \sum_{i=n}^{+\infty} \gamma_i \|w_i\|\right] + C\mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \sum_{i=n}^{+\infty} \left\|B_{i-1}^{-1} - B_i^{-1}\right\| \|S_i\|\right]\\
& + \mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \sum_{i=n}^{+\infty} \gamma_i \left\|r'_{i+1}\right\|\right]
\end{aligned}$$

By Inequality (21) we have:

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{\Upsilon_p}\sup_{i\geqslant n}\|\Delta_i\|\sum_{i=n}^{+\infty}\gamma_i\|w_i\|\right] &\leqslant C\mathbb{E}[\mathbb{1}_{\Upsilon_p}\sup_{i\geqslant n}\|\Delta_i\|^2]^{1/2}\mathbb{E}\left[\left|\sum_{i=n}^{+\infty}\gamma_i\|w_i\|\right|^2\right]^{1/2}\\ &\leqslant C\mathbb{E}[\mathbb{1}_{\Upsilon_p}\sup_{i\geqslant n}\|\Delta_i\|^2]^{1/2}\sqrt{\chi_n}\\ &\leqslant o(\chi_n)\,.\end{aligned} \tag{24}$$

As noticed in [9] we have $\sum_{i=1}^{+\infty}\left\|B_{i-1}^{-1}-B_i^{-1}\right\|<+\infty$. Therefore,

$$\mathbb{E}\left[\mathbb{1}_{\Upsilon_p}\|S_i\|\right]\sum_{i=n}^{+\infty}\left\|B_{i-1}^{-1}-B_i^{-1}\right\|\leqslant C\sqrt{\chi_n}\sum_{i=n}^{+\infty}\left\|B_{i-1}^{-1}-B_i^{-1}\right\|=o(\sqrt{\chi_n})\,, \tag{25}$$

and by assumptions

$$\mathbb{E}\left[\mathbb{1}_{\Upsilon_p}\sum_{i=n}^{+\infty}\gamma_i\left\|r'_{i+1}\right\|\right]\leqslant C\chi_n=o(\sqrt{\chi_n})\,. \tag{26}$$

Combining (24), (25) and (26) we obtain Equation (23). Hence, we can apply [9, Theorem A] to obtain that $\mathbb{P}(\Upsilon_p)=0$. Since $\Upsilon=\bigcup_{p\in\mathbb{N}}\Upsilon_p$, the proof is finished. □

References

[1] H. Attouch, J. Bolte, and B.F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming, Series A*, 137(1):91–124, August 2011.

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

[3] P. Bianchi, W. Hachem, and Sh. Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *arXiv preprint arXiv:2005.08513*, 2020.

[4] E. Bierstone and P. Milman. Semianalytic and subanalytic sets. *Publications Mathématiques de l'IHÉS*, 67:5–42, 1988.

[5] J. Bolte, A. Daniilidis, and A. Lewis. Tame functions are semismooth. *Math. Program.*, 117:5–19, 03 2009.

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

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

[8] N. Boumal. An introduction to optimization on smooth manifolds. Available online, Nov 2020.

[9] O. Brandière and M. Duflo. Les algorithmes stochastiques contournent-ils les pièges? *Ann. Inst. H. Poincaré Probab. Statist.*, 32(3):395–427, 1996.

[10] M. Bravo, D.S. Leslie, and P. Mertikopoulos. Bandit learning in concave n-person games, 2018.

[11] K.-L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, 25(3):463–483, 1954.

[12] 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.

[13] M. Coste. An introduction to o-minimal geometry. 2002.

[14] A. Daniilidis and D. Drusvyatskiy. Pathological subgradient dynamics, 2019.

[15] D. Davis and D. Drusvyatskiy. Proximal methods avoid active strict saddles of weakly convex functions, 2021.

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

[17] D. Drusvyatskiy, A.D. Ioffe, and A.S. Lewis. Generic minimizing behavior in semialgebraic optimization. *SIAM J. Optim.*, 26(1):513–534, 2016.

[18] D. Drusvyatskiy and A. Lewis. Semi-algebraic functions have small subdifferentials. *Mathematical Programming*, 140, 04 2010.

[19] J. Duchi and F. Ruan. Stochastic methods for composite and weakly convex optimization problems, 2018.

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

[21] K. Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'Institut Fourier*, 48(3):769–783, 1998.

[22] J. Lafontaine. *An Introduction to Differential Manifolds*. Springer International Publishing, 2015.

[23] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir, editors, *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR.

[24] A. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM J. Optim.*, 13:702–725, 2002.

[25] A. Lewis and J. Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33, 02 2008.

[26] Ta Loi. Verdier and strict Thom stratifications in o-minimal structures. *Illinois Journal of Mathematics - ILL J MATH*, 42, 06 1998.

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

[28] R. Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *Ann. Probab.*, 18(2):698–712, 1990.

[29] Sh. Schechtman. Stochastic proximal subgradient descent oscillates in the vicinity of its accumulation set, 2021.

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

[31] A.J. Wilkie. O-minimal structures. In *Séminaire Bourbaki Volume 2007/2008 Exposés 982-996*, number 326 in Astérisque. Société mathématique de France, 2009. talk:985.