Stochastic Subgradient Descent Escapes Active Strict Saddles on Weakly Convex Functions

Pascal Bianchi, Walid Hachem, Sholom Schechtman

July 8, 2023

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 strengthened 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, SGD converges to a local minimizer.

Keywords. Non-smooth optimization, stochastic gradient descent, avoidance of traps, Clarke subdifferential, stratification, weak convexity

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 toward 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 [17] for a counterexample), and

some form of structure for the function f is required. Since the work of Bolte *et al.* [7] 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 [19] (in the sense that $\limsup \mathrm{dist}(x_n, \mathcal{Z}) = 0$, the convergence of (x_n) to a unique element of $\mathcal{Z}$ is, for now, only a conjecture). Another work in the same line is [34]. Bolte and Pauwels [8] 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 [4].

Thanks to these contributions, we know that any convergent subsequence of (x_n) will have a limit in $\mathcal{Z}$. 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) toward 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 [35] and Brandière and Duflo [12] 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 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 [30]. 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 [30] the trap avoidance is due to the random initialization of the algorithm, whereas in [12, 35], 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 [18] of Davis and Drusvyatskiy. The spurious points that were considered in this reference are the so-called *active strict saddles*. Formally, a critical point is an active strict saddle if it lies on a manifold M such that *i)* f varies sharply outside 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 has a

second-order negative curvature. In this setting, and assuming that f is weakly convex, the article [18] focuses on the noiseless case, and study variants of the (implicit) *proximal point algorithm* rather than the (explicit) subgradient descent. Similarly to [30], 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 regularized version of f - the Moreau envelope (which is well-defined due to the weak-convexity of f).

Contrary to [18], 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 [35, 12], our analysis strongly relies on the presence of the additive random perturbation η_n.

In the framework of weakly convex 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, our first 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 allows to control the distance between the iterate x_n of Algorithm (1) and the manifold M. Second, we 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 [33]. 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 strata, which is established in the paper. 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 [12]. 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 [21, 18]. 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 [21, 18], 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.

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 [7]. While such stratifications are well-known in the literature on o-minimal structures [33], 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 [7, Proposition 4] and Lemma 16 below) to the case of definable, locally Lipschitz continuous functions by establishing a Lipschitz-like condition on the (Riemannian) gradients of two adjacent strata.

- 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.

Let us mention here a key difference with the first version of the paper submitted to arXiv, where the nonconvergence of SGD toward active strict saddles was stated under a weaker form of angle condition. The latter allowed us to deal with functions that are not (locally) weakly convex. Unfortunately, a serious flaw was pointed out by an anonymous reviewer, who noted that our proof strongly relied on the use of [3, Theorem 4.1], which turns out to be incorrect. We refer to Remark 4 and Appendix C for a detailed discussion on this issue.

We also point out that, shortly after the first submission of the present paper, a concurrent and an independent work [20] has appeared. In the latter, the question of the avoidance of active strict saddles by SGD was treated with very similar techniques. In particular, their proximal aiming condition coincides with our angle conditions. Other questions, such as the rate of convergence and the asymptotic normality close to the active manifolds, were also addressed. We believe that both of our works show that the Verdier and the angle condition are well-founded and might be interesting for the optimization community. They, furthermore, open the way for a more thorough investigation of the avoidance by SGD of spurious points in a non-smooth context.

The rest of 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 strengthened projection formula, which is stated in Theorem 1. In Section 3, we discuss the notion of active strict saddles. After recalling some results of [18], 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. In Section 5 we discuss possible directions for future work. Sections 6–7 are 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 $\mathrm{dist}(x, S) := \inf\{\|y - x\| : y \in S\}$. For a C^1 function $g : \mathbb{R}^p \to \mathbb{R}^d$ and $x \in \mathbb{R}^p$, we denote $J_g(x) \in \mathbb{R}^{d \times p}$ the Jacobian of g at x. 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.

$(a_n)_{n \geqslant N}$ denotes a sequence starting from $N \in \mathbb{N}$, if there is no possible confusion about the starting index it will also be denoted as (a_n). If $(\mathscr{F}_n)$ is a filtration on some probability space and η is a random variable, then for $n \in \mathbb{N}$, we denote $\mathbb{E}[\eta | \mathscr{F}_n]$ the conditional expectation of η relatively to $\mathscr{F}_n$. In the proofs, the latter will usually be denoted as $\mathbb{E}_n \eta$.

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 [28, 11] for a detailed introduction on differential geometry.

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:

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

as the (possibly empty) set of points $y^* \in M$ such that $\|y^* - x\| = \inf\{\|y - x\| : y \in M\}$. The following lemma can be found in [32] (see also [28, Chap. 3, Ex. 24]), even though the first part dates back to the 50. It states that, in the vicinity of any point of M, $P_M(x)$ is a singleton, henceforth, P_M can be identified to a function on that neighborhood. Recall that $P_{T_y M}$ is the orthogonal projection onto $T_y M$.

Lemma 1 (Projection onto a manifold). *Let M be a C^p submanifold, with $p \geqslant 2$. Consider $y \in M$. Then, there exists a neighborhood U of y, such that $P_M : U \to M$ is a single-valued map. Moreover, P_M is C^{p-1} in that neighborhood, and $J_{P_M} = P_{T_y M}$.*

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 $M \cap U$. In this case, F is called a smooth representative of f around y on M. If $f : M \to \mathbb{R}$ is C^1, we define for every $y \in M$,

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

where F is any smooth representative of f around y. The definition of $\nabla_M f(y)$ does not depend on the choice of F (see e.g. [11, Section 3.8]). We refer to $\nabla_M f(y)$ as the (Riemannian) gradient of f at M (here the Riemannian structure on M is implicitely induced from the usual Euclidian scalar product on $\mathbb{R}^d$).

If $f : M \to \mathbb{R}$ is C^p, with $p \geqslant 2$, we define for every $y \in M$, the covariant (Riemannian) Hessian of f at y as the linear operator $\mathcal{H}_{f,M}(y) : T_y M \to T_y M$ given by:

$$\mathcal{H}_{f,M}(y) : \ v \mapsto P_{T_y M} J_G(y) v,$$

where G is a C^{p-1} function defined in a neighborhood of y which agrees with $\nabla_M f$ on M, The definition of $\mathcal{H}_{f,M}(y)$ does not depend on the choice of G (see e.g. [11, Section 5.5]).

2.2 Clarke Subdifferential

Consider $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous function. Denote by $\text{Reg}(f)$ 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* [14] is given by:

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

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 subsequential convergence of iterates [7, 19, 8]. For instance, a sufficient condition on f, which ensure its path-differentiability, is that f is *definable* w.r.t. an o-minimal structure. We review the concept of o-minimality in the appendix, for the interested reader. Examples of definable functions include semialgebraic functions, analytic functions on a semialgebraic compact set, exponential and logarithm (see e.g. [6, 5, 40]). Moreover, the set of definable functions is closed w.r.t. composition. In particular, the loss of a neural network is in general a definable function [19].

2.3 Verdier 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 *strata* (S_i) such that each of the S_i is a C^p submanifold, and such that:

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

Given a family $\{A_1, \dots, A_k\}$ of subsets of A, we say that a stratification (S_i) is *compatible with* $\{A_1, \dots, A_k\}$, if each of the A_i is a finite union of strata. We say that a stratification (S_i) is definable, if every stratum S_i is definable w.r.t. some o-minimal structure (see Appendix A.1). If E_1, E_2 are two vector spaces such that $E_1 \neq \{0\}$, we define:

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

and we set $\boldsymbol{d}_a(\{0\}, E_2) = 0$. Note that $\boldsymbol{d}_a(E_1, E_2) = 0$ implies $E_1 \subset E_2$.

Definition 2. *Let (S_i) be a C^p stratification of some set $A \subset \mathbb{R}^d$. We say that (S_i) satisfies the* Verdier property-(v)*, if for every couple of distinct strata S_i, S_j such that $S_i \cap \overline{S}_j \neq \varnothing$ and for each $y \in S_i$, there are two positive constants δ, C such that:*

$$\begin{aligned} y' &\in B(y, \delta) \cap S_i \\ x &\in B(y, \delta) \cap S_j \end{aligned} \implies \boldsymbol{d}_a(T_{y'} S_i, T_x S_j) \leqslant C \|y' - x\|. \tag{3}$$

In this case, we refer to (S_i) as a Verdier C^p stratification of A.

A Verdier stratification is a special case of a Whitney stratification (we refer the reader to Appendix A.2 for a review on Whitney stratifications). Whereas the Whitney stratification can now be considered as well known in optimization community, the Verdier stratification is comparatively less popular. In the framework of nonsmooth optimization, one of the main interests of the Verdier stratification over the Whitney stratification, is given by Theorem 1 below, which is one of the contributions of this paper. Theorem 1 can be seen as a strengthening of the so-called "projection formula" for Whitney stratifiable functions, which we recall in Lemma 16 of Appendix A.2, for the sake of completeness.

Before stating this result, we make two important remarks. First, any locally Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$ whose graph admits a Verdier stratification, is path-differentiable in the sense of Definition 1. This is a consequence of Lemma 17 in Appendix A.1. Second, if f is definable w.r.t. an o-minimal structure (see Appendix A.1), then, for every $p \geqslant 1$, its graph admits a Verdier C^p stratification. This is a consequence of the following fundamental result.

Proposition 1 ([33, 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 $\{A_1, \ldots, A_k\}$.*

Finally, we state the main contribution of this section, which can be interpreted as a strengthened version of the projection formula (Lemma 16).

Theorem 1 (Strengthened 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 f is C^p on every stratum and for every couple of distinct strata X_i, X_j such that $X_i \cap \overline{X_j} \neq \varnothing$ and for every $y \in X_i$, there is $C, \delta > 0$, such that for any two points $y' \in B(y, \delta) \cap X_i$, $x \in B(y, \delta) \cap X_j$,*

$$\left\| P_{T_{y'} X_i}(\nabla_{X_j} f(x)) - \nabla_{X_i} f(y') \right\| \leqslant C \left\| x - y' \right\| , \tag{4}$$

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

$$\left\| P_{T_{y'} X_i}(v) - \nabla_{X_i} f(y') \right\| \leqslant C \left\| x - y' \right\| . \tag{5}$$

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 16. 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 $c \colon (-\varepsilon, \varepsilon) \to \mathbb{R}^d$ s.t. $c((-\varepsilon, \varepsilon)) \subset S_j$, $c(0) = (x, f(x))$ and $\dot{c}(0) = (h_x, h_f)$. 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 stratum and C, δ as in Equation (3). 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 $\|h_{y'}\| = 1$. 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.

$$\left\| \frac{1}{c_h} \left(h_{y'}, \langle \nabla_{X_i'} f(y'), h_{y'} \rangle \right) - (h_x, \langle \nabla_{X_j'} f(x), h_x \rangle) \right\| \leqslant C(L+1) \|x - y'\|,$$

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

$$\left\| h_{y'} - c_h h_x \right\| \leqslant C' \|x - y'\| \quad \text{and} \quad \left\| c_h \langle \nabla_{X_j'} f(x), h_x \rangle - \langle \nabla_{X_i'} f(y'), h_{y'} \rangle \right\| \leqslant C' \|x - y'\|.$$

Thus, it holds that:

$$\left\| \langle \nabla_{X_j'} f(x), h_{y'} - c_h h_x \rangle \right\| + \left\| c_h \langle \nabla_{X_j'} f(x), h_x \rangle - \langle \nabla_{X_i'} f(y'), h_{y'} \rangle \right\| \leqslant C' \|x - y'\|. \quad (6)$$

Noticing that by the projection formula, for all $v \in \partial f(x)$, it holds that $\langle v, h_x \rangle = \langle \nabla_{X_j'} f(x), h_x \rangle$, we also obtain for such a v,

$$\left\| \langle v, h_{y'} - c_h h_x \rangle \right\| + \left\| c_h \langle v, h_x \rangle - \langle \nabla_{X_i'} f(y'), h_{y'} \rangle \right\| \leqslant C' \|x - y'\|. \quad (7)$$

Thus, using Equation (6) and applying a triangle inequality, we obtain:

$$\left\| \langle \nabla_{X_j'} f(x) - \nabla_{X_i'} f(y'), h_{y'} \rangle \right\| \leqslant C' \|x - y'\|. \quad (8)$$

And, similarly, using Equation (7),

$$\forall v \in \partial f(x), \quad \left\| \langle v - \nabla_{X_i'} f(y'), h_{y'} \rangle \right\| \leqslant C' \|x - y'\|. \quad (9)$$

The proof is now completed by noticing that $h_{y'} \in T_{y'} X_i'$ was an arbitrary vector of unitary norm and that one can choose C', δ such that Equations (8) and (9) hold uniformly for all $x \in B(y, \delta) \cap X_j'$, where X_j' is any stratum neighboring X_i'.

$\square$

Remark 1. *Equation (5) in Theorem 1 remains true if the Clarke subdifferential ∂f is replaced by any, definable, set-valued map $D : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ which is a so-called conservative field for the potential f i.e., the condition $v \in \partial f(x)$ can be replaced by $v \in D(x)$. The concept of conservative fields was introduced in [8], in order to circumvent the fact that automatic differentiation procedures such as those used in Tensorflow, do not necessarily produce Clarke subgradients. The Clarke subdifferential ∂f is, among others, one example of a conservative field. It was shown in [8, Theorem 4] that the projection formula (Lemma 16 in this paper) still holds, if ∂f is replaced by a definable conservative field D. Using this generalization in the proof of Theorem 1 instead of Lemma 16, one conclude that the same generalization holds for our strengthened projection formula.*

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 3 (Active manifold[1], [23])**.** *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:***

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

Definition 4 (Active strict saddle)**.** *We say[2] 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 that $\nabla_M f(x^*) = 0$ and $\langle w, \mathcal{H}_{f,M}(x^*)(w) \rangle < 0$.*
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 the scalar product between $(1,0)$ and $\mathcal{H}_{f,M}(0)((1,0))$ is equal to -2.

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 [21].

Proposition 2 ([18, 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.*

---

[1]We must notice here that the notion of an active (or, as it sometimes referred to, identifiable) manifold is closely related to the notion of partial smoothness introduced in [31]. Indeed, as it was shown in [23, Proposition 8.4] both are equivalent under a non-degeneracy condition: 0 is in the relative interior of the proximal subdifferential of f at x^*.

[2]The definition of active strict saddles provided in [18] involves the notion of parabolic subderivatives. In this paper, we found convenient to use the equivalent Definition 4, which is closer in spirit to notions of differential geometry.

It is worth noting that the result of [18, 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 3, we introduce the following useful conditions.

Definition 5. *Let M be a C^2 active manifold around some $x^* \in \mathcal{Z}$. We say that M satisfies the Verdier condition and the angle condition, if there is U a neighborhood of x^* such that the following conditions hold respectively.*

iii) ***Verdier condition.*** *There is $C \geqslant 0$, 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.*** *There is $\beta > 0$, such that for every $x \in U$ and for every $v \in \partial f(x)$,*

$$\langle v, x - P_M(x) \rangle \geqslant \beta \left\| x - P_M(x) \right\|.$$

Definition 6. *An active strict saddle x^* is said to satisfy the* Verdier and angle conditions*, if the active manifold M in Definition 4 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 stratum of the Verdier stratification of Theorem 1. The purpose of the angle condition is to ensure that, close to M, the subgradients of f at x are always directed outwards of M. The latter will allow us to prove that the iterates of SGD converge to M fast enough. In the concurrent work of [20] these conditions were named as *strong (a)* and *proximal aiming* conditions.

The following theorem strengthens 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 [21, Corollary 4.8 and Theorem 4.16] the active manifold [3] can be chosen

---

[3]The name *active manifold* follows from the work of [18], while in [21] they are called identifiable manifolds. Both terminologies are usual and go back at least to [13, 42].

adapted to $\{X_1, \ldots, X_k\}$, the existence of an active manifold with a Verdier condition follows from [18, Theorem 2.9, Appendix A]. To prove the angle condition note that by weak convexity of f there is $\rho \geqslant 0$ such that:

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

Furthermore, it was noticed in [16, Theorem D.2] that weak convexity of f implies the existence of $\alpha > 0$ such that for x close enough to x^*, it holds:

$$f(x) \geqslant f(P_M(x)) + \alpha \|P_M(x) - x\| \,.$$

Combining both inequalities, we obtain:

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

Taking $U = B(x^*, r)$, with r small enough, we see that the angle condition is satisfied. $\qquad \square$

Remark 2. *Let M be an active manifold around x^*. It is clear from the proof of Theorem 2, that 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. One may wonder if there are examples of (non-weakly convex) functions that have active strict saddles without the angle condition. The following generic example exhibits one of those.*

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 condition but* not *satisfying the angle condition. Notice, furthermore, that this example is generic in the following sense: if $\|u\| < 1$, then $(0, 0)$ is also an active strict saddle of f_u, satisfying the same properties.*

4   Avoidance of Active Strict Saddles

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

$$x_{n+1} = x_n - \gamma_n v_n + \gamma_n \eta_{n+1} \,, \tag{10}$$

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

Assumption 1.

  i) *The function f is path differentiable.*

 ii) *For every $n \in \mathbb{N}$, $v_n \in \partial f(x_n)$.*

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 $n \geqslant 1$:*

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

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

Assumption 2. *The point x^* is an active strict saddle of order 5 satisfying the Verdier and angle conditions.*

The reason for which we posit a smoothness condition of order 5 (while the definition of an active strict saddle only requires the order 2) will be clear from Section 6. It is related to the fact that, in our proof, $\nabla(F \circ P_M)$ need to be C^3, where F is any smooth representative of f on M. By Lemma 1 this will be obtained as soon as M is a C^5 manifold.

We state our last assumption. Interpreting the map $\mathcal{H}_{f,M}(x^*) \circ P_{T_{x^*}M}$ as a quadratic form on $\mathbb{R}^d$, we write $\mathbb{R}^d = E^- \oplus E^+$, where E^- (respectively E^+) is the vector space spanned by the eigenvectors that have negative (respectively nonnegative) eigenvalues. By results of Section 2.1, $E^- \subset T_{x^*}M$. Moreover, by Assumption 2, we note 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:*

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

Remark 3. *We discuss Assumption 3. The first point is standard. The third point ensures that η_n explores the negative curvature space E^-, so that the iterates must eventually escape x^*, as will be explained in Section 6. The second point deserves more comments. In order to establish the convergence of SGD toward the set $\mathcal{Z}$, the assumption $\limsup \mathbb{E}[\|\eta_{n+1}\|^2 |\mathscr{F}_n] < +\infty$ is standard. However, in order to establish the avoidance of spurious critical point, this assumption should be strengthened by requiring the boundedness of the (conditional) fourth order moments. Comparatively to this paper, [35, 2] make the stronger assumption that the sequence (η_n) is by a deterministic constant, whereas the concurrent work of [20], which has been submitted shortly after the present paper, assumes a boundedness condition on eight-order moments. Finally, although [12] claims that only second order moments need to be bounded, a careful examination of their proof reveals that fourth order moments are actually needed.*

We are ready to state the following theorem, which is the main result of this paper. Its proof is devoted to Section 6.

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, 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 5 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* $n \in \mathbb{N}$.

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 $w \in \mathbb{R}^d \setminus \{0\}$,

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

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

Proof. We will first show that, almost surely, if (x_n) is bounded, then it converges to a unique point $z \in \mathcal{Z}$. Indeed, denote $A \in \mathscr{F}$ the event on which (x_n) is bounded and notice that $A = \bigcup_{C \in \mathbb{N}} A_C := \bigcup_{C \in \mathbb{N}}[\sup_{n \in \mathbb{N}} \|x_n\| \leqslant C]$. For $n \in \mathbb{N}$, define $\tilde{\eta}_{n+1} = \eta_{n+1}\mathbb{1}_{\|x_n\| \leqslant C}$ and notice that on A_C, $\eta_{n+1} = \tilde{\eta}_{n+1}$. By a standard Martingale argument $\sum_{i=0}^{+\infty} \gamma_i \tilde{\eta}_{i+1}$ converges, and $\sup_{n \in \mathbb{N}} \|v_n\| < +\infty$. Therefore, on A_C, we can apply [19, Theorem 3.2] and obtain that the limit points of (x_n) are all lying in $\mathcal{Z}$. Furthermore, on A_C, it holds that:

$$\|x_{n+1} - x_n\| \leqslant \gamma_n(L + \|\tilde{\eta}_{n+1}\|)) \,,$$

where L is the Lipschitz constant of f on $\overline{B(0,C)}$. Since the right-hand side of this inequality goes to zero, this implies that $\|x_{n+1} - x_n\| \to 0$. The latter, along with the boundedness of (x_n), implies that the set of limit points of (x_n) is a connected set. From the active strict saddle property of f we know that $\mathcal{Z}$ is finite and the only connected set of a finite set is a unique point. This implies that, on A_C, (x_n) converges to a unique point. Since $C \in \mathbb{N}$ was arbitrary, this remains true on A.

Finally, let $x^* \in \mathcal{Z}$ be an active strict saddle of f. Assumptions of the corollary immediately implies Assumption 3. Hence, $\mathbb{P}(x_n \to x^*) = 0$. Thus, as soon as (x_n) is bounded, it converges to a unique point, which is not an active strict saddle. By the active strict saddle property of f, this implies the convergence of (x_n) to a local minimum of f. $\qquad\square$

5 Further Topics

5.1 Conservative Fields and Empirical Risk Minimization

In several situations such as machine learning, one is interested in the scenario where f is defined as a finite sum of the form:

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$

for some (nondifferentiable) locally Lipschitz functions $f_1, \ldots, f_N$. In the machine learning community, this problem is referred to as *empirical risk minimization*, and the algorithm referred to as SGD, consists in drawing an independent sequence (I_n), uniformly chosen in $\{1, \ldots, N\}$, and to define the iterates:

$$\tilde{x}_{n+1} = \tilde{x}_n - \gamma_n v_{n+1} \,, \tag{11}$$

where v_{n+1} is one element of the Clarke subdifferential of $f_{I_{n+1}}$ at point x_n, or, even more generally,

$$v_{n+1} \in D_{I_{n+1}}(\tilde{x}_n) \,,$$

13

where $D_1, \ldots, D_N : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ are *conservative fields* for the potentials $f_1, \ldots, f_N$ respectively (we refer to [8] for a complete review of conservative fields). As a matter of fact, the iterates (11) do not in general satisfy the inclusion (1). Hence, our paper does not, strictly speaking, encompass the algorithm (11). As far as the generalization of our results is concerned, two approaches can be used.

A first approach follows the idea of [4, 9, 10]. It is based on the fact that the Clarke subdifferential (or more generally, any conservative field) coincides almost everywhere with the gradient (see [8]). That is, for every $i = 1, \ldots, N$ and for Lebesgue-almost every x, $D_i(x) = \{\nabla f_i(x)\}$. Under the assumption that the initialization $\tilde{x}_0$ is randomly chosen according to some probability density function on $\mathbb{R}^d$, and provided that every step size γ_n lies outside a certain Lebesgue-negligible set, every random variable $\tilde{x}_n$ can be shown to admit a density w.r.t. the Lebesgue measure. This implies that $D_i(\tilde{x}_n) = \{\nabla f_i(\tilde{x}_n)\}$ almost surely, for every i. Hence $v_{n+1} = \nabla f_{I_{n+1}}(\tilde{x}_n)$. As a consequence, the iterates $(\tilde{x}_n)$ satisfy w.p.1:

$$\tilde{x}_{n+1} = \tilde{x}_n - \gamma_n \nabla f(\tilde{x}_n) + \gamma_n \eta_{n+1},$$

where $\eta_{n+1} := \nabla f_{I_{n+1}}(\tilde{x}_n) - \nabla f(\tilde{x}_n)$ is a martingale increment sequence. Otherwise stated, under these assumptions, $\tilde{x}_n$ satisfy Equation (1) w.p.1., for every n. Our conclusions can therefore be extended to the algorithm (11), provided that the initialization is random, and under mild assumptions on the step sizes. We note that a similar conclusion can be also obtained if, instead of randomizing the initial point, one artificially adds a small random perturbation (e.g. Gaussian) to the update Equation (11), in order to ensure that every $\tilde{x}_n$ admits a density.

A second approach consists in working in the framework of conservative set-valued fields, without additional assumptions such as the above randomized initialization. As a matter of fact, the iterates (11) can be written under the form $\tilde{x}_{n+1} = \tilde{x}_n - \gamma_n D(\tilde{x}_n) + \gamma_n \eta_{n+1}$, where $D = N^{-1} \sum_{i=1}^N D_i$ is, again, a conservative field, and (η_n) is a martingale increment sequence. Previous works such as [8] consider this kind of dynamics, and establish the convergence to the set of zeroes of D. Furthermore, a strengthened projection formula such as the one of Theorem 1, also holds when substituting D with ∂f. This implies that, generically, every zero of D for definable f belongs to an active manifold satisfying a Verdier condition. Unfortunately, the main issue is that the set of zeroes of D can be substantially larger than the set of Clarke critical points. In particular, even for weakly convex functions f, it cannot be established that, generically, the spurious zeroes of D (*i.e.*, the ones which are not local minimizers of f) are active strict saddles. For instance, the following example is generic: $f : \mathbb{R} \to \mathbb{R}$, with $f(x) = x$, $D(0) = [0,1]$ and $D(x) = \{1\}$ if $x \neq 0$. A consequence is that Corollary 1 cannot be immediately generalized to conservative fields. A characterization of the spurious zeroes of D, and the proof of the fact that SGD avoids such points, constitute interesting problems for future researches.

5.2   Beyond Weak Convexity

Active strict saddles are generic in case of (definable) weakly convex functions. As a consequence, our proof of nonconvergence to active saddles (Theorem 3) allows to conclude that, in case of generic weakly convex functions, SGD converges to a local minimizer (Corollary 1). A natural question is to know whether the same conclusion holds in the absence of weak convexity.

To that end, it is mandatory to characterize, aside from active strict saddles, the generic critical points of definable but not necessarily weakly convex functions. Such a characterization was provided in the Ph.D. manuscript of the third author [37]. It is established that, generically, any Clarke critical point of a definable function is located on an active manifold M (satisfying a Verdier condition), and is either a local minimum, an active strict saddle or what the author of [37] called a *sharply repulsive critical point*, namely, a point x^* which is a local minimizer of f_M, and such that the subgradients of f in the vicinity of x^* are pointing toward M (see Example 2 for an illustration of such a point).

An important but challenging problem is to establish, in the absence of weak convexity, that both active strict saddles and sharply repulsive points are avoided.

Remark 4. *In case of generic definable but non-weakly convex functions, the nonconvergence of SGD toward active strict saddles was incorrectly stated in the first arXiv version of this paper. A serious flaw in the proof was pointed out by one the anonymous reviewers, who noted that our proof strongly relied on the use of [3, Theorem 4.1], which turns out to be incorrect. A detailed account on this issue is provided in Appendix C. The same mistake occurred in [37] which incorrectly states a nonconvergence result toward sharply repulsive points. Thus, the question of the generic convergence of SGD toward local minimizers remains open in the non-weakly convex setting.*

6 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 $w \in T_0 M$ such that $\langle w, \mathcal{H}_{f,M}(0)(w) \rangle < 0$. 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 [12]. In this setting, the remainders z_n will be shown to be small enough so as not to alter fundamentally the approach of [12].

Outline of the proof.

- *Strengthened avoidance of traps.* First, in Section 6.1 we provide a strengthened version of the avoidance of traps result of [12, 35, 38]. In a first approximation, this result, given in Proposition 3, states that if an $\mathbb{R}^d$-valued sequence (y_n) satisfying a recursion of the form

$$y_{n+1} = y_n - \gamma_n D(y_n) + \gamma_n \tilde{\eta}_{n+1} + \gamma_n \varrho_{n+1} + \gamma_n \tilde{\varrho}_{n+1} \,, \tag{12}$$

where $D : \mathbb{R}^d \to \mathbb{R}^d$ is a function such that $J_D(0)$ has at least one eigenvalue with a negative real part, the noise $\tilde{\eta}_{n+1}$ is omnidirectional in some sense, and the terms ϱ_{n+1} and $\tilde{\varrho}_{n+1}$ are small perturbation terms (see Proposition 3 for the exact assumptions), then $\mathbb{P}(y_n \to 0) = 0$. The only difference with the assumptions of [12, 35, 38] is in the presence of the term $\tilde{\varrho}_{n+1}$. To not interrupt the exposition we provide a complete proof of Proposition 3 in Section 7. In the context of our algorithm and leaving technical details aside, denoting $y_n = P_M(x_n)$, Equation (12) appears from a Taylor expansion of P_M, with $D(y_n) = \nabla_M f(y_n)$. Here, $\tilde{\varrho}_{n+1}$ appears from the Verdier condition and its norm will be controlled by $\text{dist}(x_n, M)$.

15

- *Construction of (y_n).* The goal of Section 6.2 is to construct a sequence satisfying Equation (12). A natural attempt is to define $(y_n) := (P_M(x_n))$. However, since P_M is defined only on a neighborhood of M (say U), in our construction we fix $N \in \mathbb{N}$ and for $n \geqslant N$, define y_n^N as $P_M(x_n)$ if the iterates $\{x_N, x_{N+1}, \ldots, x_n\} \subset U$ and (y_n^N) satisfies (by construction) Equation (12) otherwise. In this context, D becomes an appropriate extension of $\nabla_M f$ outside M. As explained in Section 6.2, if we establish $\mathbb{P}(y_n^N \to 0) = 0$, for all $N \in \mathbb{N}$, then this will imply that $\mathbb{P}(x_n \to 0) = 0$. Thus, the goal of the next sections will be to verify the assumptions of Proposition 3 for the sequence (y_n^N).

- *The sequence (y_n^N) can be written as a Robbins-Monro algorithm.* In Section 6.2 we show that the constructed sequence (y_n^N) satisfies indeed Equation (12), and verifies almost all the assumptions of the general Proposition 3. This will be the content of Proposition 4. In short, it follows from a Taylor expansion of P_M. Here, we put a special emphasis on the term $\tilde{\varrho}_n^N$ (analogous to $\tilde{\varrho}_n$ above), which will be shown to satisfy $\mathbb{E}[\lVert \tilde{\varrho}_{n+1}^N \rVert \,|\, \mathscr{F}_n] = \mathcal{O}(\mathrm{dist}(x_n, M))$. In the remaining sections our goal will be to control the convergence rate of (ϱ_n^N) toward zero, which will allow us to apply Proposition 3.

- *Convergence rates of (ϱ_n^N).* In Section 6.2.2 we establish the key drift inequality on $\mathrm{dist}(x_n, M)$, which allows us in Sections 6.2.3–6.2.4 establish that

$$\limsup \chi_n^{-1/2} \sum_{i=n}^{\infty} \gamma_i \mathbb{E}\left[\mathrm{dist}(x_i, M) | \mathscr{F}_n\right] = 0\,,$$

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

- *End of proof.* Finally, in Section 6.2.5 we finish the proof of Theorem 3 by applying Proposition 3 to the sequence (y_n^N) and precisely invoking the intermediate results allowing us to verify its assumptions.

6.1 Preliminary: Avoidance of Traps in the Smooth Case

The following proposition is an avoidance of traps results in the smooth case, which is tailored to our assumptions on the projected iterates on the manifold M. Its proof combines ideas from Brandière and Duflo [12] on the one hand, and from Tarrès [38] on the other. Our proposition might have some interest of its own, since it is more general than the corresponding result in [12]. We also note that its proof corrects some errors found in the latter paper[4].

Given a positive integer d, the statement of the proposition makes use of a field $D : \mathbb{R}^d \to \mathbb{R}^d$. We shall assume that D is defined on $\mathbb{R}^d$ and is C^3 in some neighborhood of 0, with $D(0) = 0$. We shall also assume that $J_D(0)$, the Jacobian of D at 0, has at least one eigenvalue with a strictly negative real part. Denote as d^- the dimension of the invariant subspace of $J_D(0)$ associated with these eigenvalues, and let $d^+ = d - d^-$. We shall need the spectral factorization

$$J_D(0) = P \begin{pmatrix} J^+ & 0 \\ 0 & J^- \end{pmatrix} P^{-1}\,,$$

of $J_D(0)$, where $J^- \in \mathbb{R}^{d^- \times d^-}$ contains those Jordan blocks of $J_D(0)$ that are associated with the eigenvalues with negative real parts.

---

[4] We would like to thank an anonymous reviewer for pointing out this fact.

We now state our proposition. In all the remainder, given a filtration $(\mathscr{F}_n)$ in a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, we shall write $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot|\mathscr{F}_n]$ and $\mathbb{P}_n(\cdot) = \mathbb{P}() \cdot |\mathscr{F}_n)$.

Proposition 3. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $(\mathscr{F}_n)$ a filtration and (γ_n) a sequence of deterministic nonnegative step sizes such that $\sum_{i=0}^{\infty} \gamma_i = +\infty$ and $\sum_{i=0}^{\infty} \gamma_i^2 < +\infty$.*
Consider the $\mathbb{R}^d$–valued stochastic process (y_n) given by

$$y_{n+1} = y_n - \gamma_n D(y_n) + \gamma_n \tilde{\eta}_{n+1} + \gamma_n \varrho_{n+1} + \gamma_n \tilde{\varrho}_{n+1}, \tag{13}$$

where y_0 is $\mathscr{F}_0$-measurable, the map D is as above, the sequences $(\tilde{\eta}_n), (\varrho_n)$, and $(\tilde{\varrho}_n)$ are $(\mathscr{F}_n)$-adapted, and both $\|\tilde{\eta}_n\|$ and $\|\tilde{\varrho}_n\|$ have finite fourth moments for each n. Write $(\tilde{\eta}_n^+, \tilde{\eta}_n^-) = P^{-1}\tilde{\eta}_n$, where $\tilde{\eta}_n^{\pm} \in \mathbb{R}^{d^{\pm}}$. Let $\Gamma \in \mathscr{F}$ be an arbitrary event and assume that on $[y_n \to 0] \cap \Gamma$ the following almost surely holds.

i) $\forall n \in \mathbb{N}, \mathbb{E}_n \tilde{\eta}_{n+1} = 0.$

ii) $\limsup \mathbb{E}_n \|\tilde{\eta}_{n+1}\|^4 < +\infty.$

iii) $\liminf \mathbb{E}_n \|\tilde{\eta}_{n+1}^-\| > 0.$

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

v) $\limsup \mathbb{E}_n \|\tilde{\varrho}_{n+1}\|^4 < +\infty.$

vi) $\lim \mathbb{E}_n \|\tilde{\varrho}_{n+1}\|^2 = 0.$

vii) $\limsup \chi_n^{-1/2} \sum_{i=n}^{\infty} \gamma_i \mathbb{E}_n \|\tilde{\varrho}_{i+1}\| = 0$, where $\chi_n = \sum_{i=n}^{\infty} \gamma_i^2$.

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

This proposition is similar to [12, Theorem 1], except for the presence of the sequence $(\tilde{\varrho}_n)$. To not interrupt the exposition we present its proof in Section 7.

6.2   Application to Algorithm (10)

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 2. *Let Assumption 2 hold and let $r > 0$ be such that $P_M : B(0, r) \to M$ is well defined and is C^4 and that there is a C^5 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^4 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 $w \in T_M 0$:

$$\langle w, \mathcal{H}_{f,M}(0)(w) \rangle = \langle w, \mathcal{H}_{F \circ P_M}(0)w \rangle,$$

where $\mathcal{H}_{F \circ P_M}(0)$ is the usual Euclidean Hessian of $F \circ P_M$ at 0.

17

Let $r_1 > 0$ be chosen in a way that the conditions of Definition 5 and Lemma 2 are satisfied on $\overline{B(0, r_1)}$. In the remainder, we fix an $r > 0$ such that $0 < r < r_1$. The value of r, while always satisfying this requirement, will be adjusted in the course of the proof.

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

Second, to reduce technical issues, we notice that as in [12, Section I.2] (see also Appendix B) to prove Theorem 3 we can actually replace Assumption 3 by the following, more easy to handle, assumption.

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:*

$$\mathbb{E}_n \left\| \eta_{n+1} \right\|^4 \leqslant B, \quad \mathbb{E}_n \left\| \eta_{n+1}^- \right\| \geqslant A.$$

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

$$\mathcal{A}_N = \left[\forall n \geqslant N, \ \|x_n\| \leqslant r \right].$$

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

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

Thus,

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

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

$$\mathbb{P}\left[[x_n \rightarrow 0] \cap \mathcal{A}_{N(\delta)}\right] \geqslant \mathbb{P}[x_n \rightarrow 0] - \delta. \tag{14}$$

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

$$\tau_N = \inf\{n \geqslant N, \ \|x_n\| > r\},$$

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: $y_{N-1}^N = 0$,

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

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 $N \geqslant 0$,

$$\mathbb{P}\left[y_n^N \xrightarrow[n \rightarrow \infty]{} 0 \right] = 0. \tag{15}$$

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,

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

Consequently, with the convergence (15) at hand, we get from Inequality (14) that $\mathbb{P}()x_n \rightarrow 0) \leqslant \delta$. Since δ is arbitrary, we obtain that $\mathbb{P}(x_n \rightarrow 0) = 0$.

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

6.2.1 Projected Iterates as a Robbins-Monro Algorithm

The next proposition shows that the sequence $(y_n^N)_{n \geq N}$ satisfies a recursion of the form (13), satisfying almost all the assumptions of Proposition 3.

Proposition 4. *Let Assumptions 1–2 and 4 hold. Then, the sequence $(y_n^N)_{n \geq N}$ 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 $(\tilde{\eta}_n^N)_{n \geq N}$, $(\varrho_n^N)_{n \geq N}$, and $(\tilde{\varrho}_n^N)_{n \geq N}$ are adapted to $(\mathscr{F}_n)$. Moreover, there is $C > 0$ such that for all $n \geq N$,

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

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

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 is included in $T_0 M$ and E_D^-, the eigenspace of the matrix $J_D(0)$ corresponding to its negative eigenvalues.*

v) *On the event $[y_n^N \to 0]$, it holds that*

$$\liminf_n \mathbb{E}_n \left\| P_{E_D^-} \tilde{\eta}_{n+1}^N \right\| \geq \liminf_n \mathbb{E}_n \left\| P_{E^-} \tilde{\eta}_{n+1}^N \right\| > 0 .$$

To prove Proposition 4 we will need the following, technical lemma.

Lemma 3. *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 $\|R_1(x, x', y)\| \leq C \|x' - x\| \|x - y\|$, and $\|R_2(x, x')\| \leq C \|x' - x\|^2$.

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 $\|R_2(x', x)\| \leq C \|x' - x\|^2$, 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 $C > 0$, $\|J_{P_M}(x) - J_{P_M}(y)\| \leq C \|x - y\|$, which finishes the proof. $\square$

Proof of Proposition 4. Letting $n \geq N$, we write

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

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 3 and Equation (10), 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 3

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

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

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

Using Lemma 1 in conjunction with the Verdier condition (iii) of Definition 5, 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

$$
\begin{aligned}
\tilde{\eta}_{n+1}^N &= \left(\mathbb{1}_{\tau_N > n}J_{P_M}(y_n^N) + \mathbb{1}_{\tau_N \leqslant n}\right)\eta_{n+1}, \text{ and} \\
\tilde{\varrho}_{n+1}^N &= \zeta_{n+1}^N + \tilde{\zeta}_n^N.
\end{aligned}
\tag{16}
$$

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 2.

To establish v), we notice that, since $E^- \subset E_D^-$,

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

On the event $[y_n^N \to 0]$, it holds that $J_{P_M}(y_n^N) \to J_0$. By Lemma 1, J_0 is the orthogonal projection on $T_0 M$, thus, $\lim_{y_n^N \to 0} P_{E^-}J_{P_M}(y_n^N) = P_{E^-}$. Consequently, we obtain on the event $[y_n^N \to 0]$:

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

and by Assumption 4. Proposition 4 is proven. $\qquad\square$

6.2.2 Drift Inequality

Proposition 4 shows that we may apply Proposition 3 to the sequence (y_n^N) as soon as we are able to control the rate of convergence of (z_n^N) toward zero. The latter will be the goal of this and subsequent sections.

The next lemma establishes the key inequality in our proof. Recall that $\beta > 0$ is the one given by the angle condition of Definition 5.

Lemma 4. *Under the assumptions of Proposition 5, there is $C > 0$ such that if $r > 0$ is chosen small enough, then*

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

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

$$
\begin{aligned}
\left\| z_{n+1}^N \right\|^2 &= \left\| p_{n+1}^N \right\|^2 \mathbb{1}_{n+1 < \tau_N} \\
&\leqslant \left\| p_{n+1}^N \right\|^2 \mathbb{1}_{n < \tau_N} = \left\| p_{n+1}^N - p_n^N + p_n^N \right\|^2 \mathbb{1}_{n < \tau_N} \\
&= \left\| z_n^N \right\|^2 + 2\langle x_{n+1} - x_n, z_n^N \rangle - 2\langle y_{n+1}^N - y_n^N, z_n^N \rangle + \left\| p_{n+1}^N - p_n^N \right\|^2 \mathbb{1}_{n < \tau_N}.
\end{aligned}
\tag{17}
$$

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

We first show that

$$\mathbb{E}_n |\langle y_{n+1}^N - y_n^N, z_n^N \rangle| \leqslant C\gamma_n \left\| z_n^N \right\|^2 + C\gamma_n^2. \tag{18}$$

By Proposition 4,

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

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 (16) 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,

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

From Proposition 4 again, we have

$$\mathbb{E}_n \left\| \varrho_{n+1}^N \right\|^2 \leqslant C\gamma_n^2 \mathbb{E}_n (1 + \|\eta_{n+1}\|^4) \mathbb{1}_{\tau_N > n+1} \leqslant C\gamma_n^2 \mathbb{E}_n (1 + \|\eta_{n+1}\|^4) \leqslant C\gamma_n,$$

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 (18) is obtained by combining these inequalities.

We next show succinctly that

$$\mathbb{E}_n \left\| p_{n+1}^N - p_n^N \right\|^2 \mathbb{1}_{n < \tau_N} \leqslant C\gamma_n^2. \tag{19}$$

Indeed,

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

and the result follows by standard calculations making use of the results of Proposition 4.

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 $\mathbb{E}_n \langle x_{n+1} - x_n, z_n^N \rangle = -\gamma_n \langle v_n, z_n^N \rangle$. Since the angle condition is satisfied on $B(0, r)$, and $\|x_n\| \leqslant r$ when $z_n^N \neq 0$, we obtain 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 (17), and using this result in conjunction with the inequalities (18) and (19), we obtain that

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

where in the last inequality we have used the fact that $\|x_n\| \leqslant r$ on the event $[n < \tau_N]$. Choosing r small enough to satisfy $Cr \leqslant \beta$, completes the proof. $\qquad \square$

6.2.3 Convergence Rates to the Manifold

Using Lemma 4 we are now able to obtain first convergence rates for (z_n^N). We recall that the exponent α defined in Assumption 1-iv) is in the interval $(1/2, 1]$.

Lemma 5. *Under the assumptions of Proposition 5, for each $a \in (0, 2\alpha - 1)$, it holds that $n^a \left\| z_n^N \right\|^2 \to_n 0$ almost surely.*

Proof. Let $a \in (0, 2\alpha - 1)$ be arbitrary. For $n \geqslant N$, denote $u_n = n^a \left\| z_n^N \right\|^2$. From Lemma 4 and Assumption 1-iv), we obtain that

$$\mathbb{E}_n u_{n+1} \leqslant \left(\frac{n+1}{n} \right)^a u_n - \beta c_1 \frac{(n+1)^a}{n^\alpha} \left\| z_n^N \right\| + Cc_2 \frac{(n+1)^a}{n^{2\alpha}}.$$

Noticing that there is $C' > 0$ such that for all $n > 1$, $(1 + 1/n)^a \leqslant 1 + C'/n$, we obtain:

$$\begin{aligned}
\mathbb{E}_n u_{n+1} &\leqslant u_n + C' n^{-1} u_n - c_1 \beta n^{a-\alpha} \left\| z_n^N \right\| + 2^a Cc_2 n^{a-2\alpha} \\
&\leqslant u_n + C' n^{a-1} \left\| z_n^N \right\|^2 - c_1 \beta n^{a-\alpha} \left\| z_n^N \right\| + 2^a Cc_2 n^{a-2\alpha}.
\end{aligned}$$

In particular, since $\alpha < 1$, if $r > 0$ is chosen small enough, then $c_1 \beta n^{a-\alpha} \left\| z_n^N \right\| \geqslant C' n^{a-1} \left\| z_n^N \right\|^2$ and we obtain:

$$\mathbb{E}_n u_{n+1} \leqslant u_n + \frac{2^a Cc_2}{n^{2\alpha - a}}.$$

Since $2\alpha - a > 1$, we obtain by Robbins-Siegmund's theorem [36] that $n^a \left\| z_n^N \right\|^2$ converges almost surely as $n \to \infty$. Since a is arbitrary in $(0, 2\alpha - 1)$, this limit is zero. $\qquad \square$

6.2.4 Control of the Weighted Sums

Finally, we can now control the weighted sum of the sequence (z_n^N). This, as it will be clear from Section 6.2.5, was the only remaining point to apply Proposition 3 to the sequence (y_n^N).

Proposition 5. *Let Assumptions 1–2 and 4 hold true. Then, the sequence (z_n^N) satisfies*

$$\chi_n^{-1/2} \sum_{i=n}^{\infty} \gamma_i \mathbb{E}_n \left\| z_i^N \right\| \xrightarrow[n\to\infty]{a.s.} 0.$$

Proof. Let $C > 0$ be the constant provided in the statement of Proposition 4. For each $n \geqslant N$ and $k \geqslant n$, we obtain from this proposition that

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

Taking k to infinity, we get that

$$\beta \sum_{i=n}^{\infty} \gamma_i \mathbb{E}_n \left\| z_i^N \right\| \leqslant \left\| z_n^N \right\|^2 + C\chi_n.$$

Let a be as in the statement of Lemma 5. Since $\lim n^a \left\| z_n^N \right\|^2 = 0$ a.s., and furthermore, since $\chi_n^{-1/2} \sim n^{\alpha - 1/2}$, we get that $\chi_n^{-1/2} \left\| z_n^N \right\|^2 \to_n 0$ a.s., and Proposition 5 is proven. $\qquad\square$

6.2.5 Theorem 3 - End of Proof

We are now in the position to finish the proof of Theorem 3. Indeed, recall from the discussion in the end of Section 6.2 that the proof will be finished if we prove, for every $N \in \mathbb{N}$, that $\mathbb{P}(y_n^N \to 0) = 0$. To that end, we will verify that for every $N \in \mathbb{N}$, the sequence (y_n^N) satisfies the assumptions of Proposition 3.

From Proposition 4, it is obvious that the assumptions i)–iii) in the statement of Proposition 3 are verified by $(\tilde{\eta}_n^N)$. Furthermore,

$$\sum_{i=N}^{\infty} \mathbb{E}[\left\| \varrho_i^N \right\|^2] \leqslant C \sum_{i=N}^{\infty} \gamma_i^2 \mathbb{E}[(1 + \left\| \eta_i \right\|^4)] \leqslant C\chi_N < +\infty,$$

which shows that the assumption iv) of Proposition 3 is satisfied by (ϱ_n^N). We also have from Proposition 4 that for all $n \geqslant N$,

$$\mathbb{E}_n \left\| \tilde{\varrho}_{n+1}^N \right\| \leqslant C \left\| z_n^N \right\|, \quad \mathbb{E}_n \left\| \tilde{\varrho}_{n+1}^N \right\|^2 \leqslant C \left\| z_n^N \right\|^2, \quad \text{and} \quad \mathbb{E}_n \left\| \tilde{\varrho}_{n+1}^N \right\|^4 \leqslant C \left\| z_n^N \right\|^4.$$

Thus, Assumptions v)–vi) of Proposition 3 are verified by $(\tilde{\varrho}_n^N)$ by Lemma 5. Finally, Assumption vii) of Proposition 3 is verified by Proposition 5.

As a consequence, we obtain for every $N \geqslant 0$ that $\mathbb{P}(y_n^N \to 0) = 0$ and finally that $\mathbb{P}(x_n \to 0) = 0$.

7 Proof of Proposition 3

The proof of Proposition 3 is technical and combines ideas from [12] and [38] (see also the older paper [35]).

The main idea of [12] and [35] was to observe that, due to the fact that $J_D(0)$ has at least one eigenvalue with strictly negative real part, the Ordinary Differential Equation (ODE) $\dot{\mathsf{y}}(t) = -D(\mathsf{y}(t))$ admits a so-called center-stable invariant manifold of dimension strictly less than d. This manifold satisfies the following property: if at some neighborhood of zero a solution of the ODE starts at a point outside of the manifold, then it diverges from the origin. The idea of [12] was then to construct a function that, in some sense, measures the distance of a point to the invariant manifold. The assumption on the perturbation sequence $(\tilde{\eta}_n)$, combined with some probabilistic estimates, then shows that the distance of the iterates (y_n) to this manifold never vanishes, which implies that these iterates stay away from zero.

From a technical point of view our proof combines techniques of [12] and [38]. The use of the approach of [38] in the second part of the proof circumvents an error that is found in the approach of [12].

Outline of the proof.

- *Center-stable manifold.* In Section 7.1 we begin by recalling the center-stable manifold theorem, which first version dates back to Poincaré. The main result here is that, after a linear basis change, the center-stable manifold could be represented through a function $G : \mathbb{R}^{d^+} \to \mathbb{R}^{d^-}$, where d^+ (resp. d^-) denotes the dimension of invariant subspaces of $J_D(0)$ that are associated to eigenvalues with non-negative (respectively negative) real parts. This function G is the first step of the construction of a "distance" to the center-stable manifold.

  Using the function G, we then construct from (y_n) an $\mathbb{R}^{d^-}$-valued sequence (w_n^-), that, due to the presence of a negative eigenvalue in $J_D(0)$, ought to be pushed away from the origin. In Lemma 9, we show that it satisfies the recursion:

  $$w_{n+1}^- = w_n^- + \gamma_n H_n w_n^- + \gamma_n(e_{n+1} + r_{n+1} + \tilde{r}_{n+1})\,,$$

  where the noise sequences (e_{n+1}) (r_{n+1}), and $(\tilde{r}_{n+1})$ satisfy assumptions analogous to their analogues (η_{n+1}), (ϱ_{n+1}), and $(\tilde{\varrho}_{n+1})$, and where the sequence of matrices (H_n) converges to a matrix that has only eigenvalues with positive real-parts. The presence of these matrices, combined with the noise e_n, will cause the iterates w_n^- to be pushed away from zero, and so will be the case of the y_n.

  The content of this section closely follows the work of [12], the only difference lying in the presence of the sequence $(\tilde{\rho}_{n+1})$ in Proposition 3.

- *The repulsive sequence (U_n) and its properties (Sections 7.2 and 7.3).* From w_n^- we construct a one-dimensional, positive random variable (U_n) which is just a well-chosen norm of w_n^-. The goal is, thus, to prove that $\mathbb{P}([U_n \to 0]) = 0$. The construction of U_n appears in [12, Proof of Proposition 4], however, from this point, our technique of proof starts to follow the one of [38]. The properties of (U_n) that will be required to show the non-convergence of this sequence are provided by Lemmas 10 to 12. In these lemmas,

we respectively show that

$$U_{n+1} - U_n \geqslant \gamma_n \langle a_n, e_{n+1} + r_{n+1} + \tilde{r}_{n+1} \rangle, \quad (a_n) \text{ adapted and bounded,}$$

$$\mathbb{E}_n U_{n+1}^2 - U_n^2 \gtrsim C\gamma_n^2 - \gamma_n U_n \mathbb{E}_n(\|r_{n+1}\| + \|\tilde{r}_{n+1}\|),$$

$$(U_{n+1} - U_n)^2 \lesssim \gamma_n^2 (1 + \|e_{n+1}\|^2 + \|r_{n+1}\|^2 + \|\tilde{r}_{n+1}\|^2).$$

These inequalities are then used to establish the following facts:

– For a well-chosen $L > 0$, for every $N \in \mathbb{N}$, the probability that for some $n \geqslant N$, $U_n \geqslant \sqrt{L\chi_n}$ is lower bounded (Lemma 13).

– As soon as $U_n \geqslant \sqrt{L\chi_n}$, then with positive probability $\sup_{k \geqslant n} U_k \geqslant \sqrt{L\chi_n}/2$ (Lemma 14).

Both of these lemmas are strengthened versions of [38, Lemmas 1 and 2] (to our knowledge, first these ideas were presented in [35]), where the terms $\tilde{r}_{n+1}$ are absent. As in [38], combining these two points, Lemma 15 shows that $\mathbb{P}([U_n \to 0]) = 0$ and completes the proof.

7.1 Application of the Center-stable Manifold Theorem

Consider the map $D : \mathbb{R}^d \to \mathbb{R}^d$ introduced before the statement of Proposition 3, and consider the following ODE starting in a neighborhood of zero

$$\dot{\mathsf{y}}(t) = -D(\mathsf{y}(t)).$$

Recalling the spectral factorization of the Jacobian $J_D(0)$ provided before the statement of Proposition 3 , it will be convenient to work in the basis of the columns of P by making the variable change

$$y \mapsto \mathbf{y} = \begin{bmatrix} y^+ \\ y^- \end{bmatrix} = P^{-1}y,$$

where $y^\pm \in \mathbb{R}^{d^\pm}$. With this at hand, writing $\widetilde{D}(\mathbf{y}) = P^{-1}D(P\mathbf{y})$, the former ODE becomes in the new basis

$$\underline{\dot{\mathbf{y}}}(t) = \begin{pmatrix} \dot{\mathsf{y}}^+(t) \\ \dot{\mathsf{y}}^-(t) \end{pmatrix} = -\widetilde{D}(\underline{\mathbf{y}}(t)) = -\begin{pmatrix} J^+\mathsf{y}^+(t) \\ J^-\mathsf{y}^-(t) \end{pmatrix} + R(\underline{\mathbf{y}}(t)), \tag{20}$$

where $\mathsf{y}^\pm(t) \in \mathbb{R}^{d^\pm}$, and where $R : \mathbb{R}^d \to \mathbb{R}^d$ is C^3 on a neighborhood of zero with $J_R(0) = 0$. The following classical proposition states that close to the origin, the ODE (20) admits a center-stable, invariant manifold.

Proposition 6 ([26, Theorem 1]). *There is $\mathcal{U} \subset \mathbb{R}^{d^+}$, a neighborhood of 0, and a C^2 function $G : \mathcal{U} \to \mathbb{R}^{d^-}$ such that the following holds.*

1. *It holds that $G(0) = 0$ and $J_G(0) = 0$.*

2. *The set $\mathcal{V} = \{(y^+, y^-) : y^+ \in \mathcal{U}, y^- = G(y^+)\}$ is invariant for the ODE (20). In other words, if $\mathbf{y}(t)$ is any solution to the ODE (20) that starts at $\mathcal{V}$, there is $t_0 > 0$ such that for all $t \in (-t_0, t_0)$, $\mathbf{y}(t) \in \mathcal{V}$.*

Remark 5. *Taking, if necessary, a smaller neighborhood $\mathcal{U}$ and extending G outside of it with the help of the Tietze extension theorem, we can always assume that G is defined, is C^2 on the whole space $\mathbb{R}^{d^+}$ (but $\mathcal{V}$ is still defined only for $y^+ \in \mathcal{U}$), and $\sup_{y^+ \in \mathbb{R}^{d^+}} \|J_G\| \leqslant c_G$, where $c_G > 0$ can be chosen as small as desired.*

Lemma 6. *Denoting $\widetilde{D} = (\widetilde{D}^+, \widetilde{D}^-)$ and $\mathbf{y} = (y^+, y^-)$, the second property of Proposition 6 implies that if $\mathbf{y} \in \mathcal{V}$, then $\widetilde{D}^-(\mathbf{y}) = J_G(y^+)\widetilde{D}^+(\mathbf{y})$.*

Proof. Consider a point $\mathbf{y} = (y^+, y^-) \in \mathcal{V}$ and let $\underline{\mathbf{y}}(t) = (\mathsf{y}^+(t), \mathsf{y}^-(t))$ be the solution to the ODE (20) that starts at $\mathbf{y}$. Noticing that, for t small enough, $\mathsf{y}^-(t) = G(\mathsf{y}^+(t))$, and differentiating this expression at zero completes the proof. $\qquad\square$

Let $\mathcal{U}$ and G be the ones of Proposition 6, by Remark 5 we can assume that G is C^2 on $\mathbb{R}^{d^+}$. For $\mathbf{y} = (y^+, y^-) \in \mathbb{R}^d$ such that $y^+ \in \mathcal{U}$, make a change of variable $(w^+, w^-) = (y^+, y^- - G(y^+))$. Furthermore, define $F : \mathbb{R}^d \to \mathbb{R}^{d^-}$ as

$$
\begin{aligned}
F(w) &= -\widetilde{D}^-(\mathbf{y}) + J_G(y^+)\widetilde{D}^+(\mathbf{y}) \\
&= -\widetilde{D}^-((w^+, w^- + G(w^+))) + J_G(w^+)\widetilde{D}^+((w^+, w^- + G(w^+))) \,.
\end{aligned}
$$

Lemma 7. *There is a function $\Delta : \mathbb{R}^d \to \mathbb{R}^{d^- \times d^-}$ such that $\Delta(w) = \mathcal{O}(\|w\|)$ and it holds that:*

$$
F(w) = F(w^+, w^-) = (-J^- + \Delta(w))w^- \,.
$$

Proof. Let $\mathcal{U}$ be the neighborhood of Proposition 6. By Lemma 6, for any $w = (w^+, w^-) \in \mathbb{R}^d$ such that $w^+ \in \mathcal{U}$, it holds that $F(w^+, 0) = 0$. Therefore,

$$
F(w^+, w^-) = F(w^+, w^-) - F(w^+, 0) = \int_0^1 \partial_- F(w^+, tw^-)w^- \, \mathrm{d}t \,, \tag{21}
$$

where $\partial_- F(w^+, tw^-)$ is the Jacobian of $w^- \mapsto F(w^+, w^-)$ (for fixed w^+) at tw^-. By a straightforward computation we obtain that:

$$
\partial_- F(0,0) = -J^- \,. \tag{22}
$$

Furthermore, since F is C^2, its Jacobian is Lipschitz on some neighborhood of zero, which implies that there is $C \geqslant 0$, such that for w close enough to zero and for $t \in [0, 1]$,

$$
\left\| \partial_- F(w^+, tw^-) - \partial_- F(0,0) \right\| \leqslant C \|w\| \,. \tag{23}
$$

Thus, the proof is completed by combining Equations (21)–(23). $\qquad\square$

Before using these results, it will be convenient to strengthen our assumptions on the field D and noise sequences provided in the statement of Proposition 3. These simplifications are frequently used in the field of stochastic approximation since the work of Lai and Wei [29]. The following lemma is proven in the appendix.

Lemma 8. *The assumptions of Proposition 3 can be replaced with the following assumptions. There exist constants $C_\times > 0$, and there exist constants $c_\times > 0$ as small as needed, such that on the whole probability space (and not only on $[y_n \to 0] \cap \Gamma$), the following holds true:*

i) The map $\widetilde{D}$ satisfies $\sup_{y \in \mathbb{R}^d} \|\widetilde{D}(y)\| \leqslant C_D$.

ii) The approximation of the function $F(w)$ provided by Lemma *7* satisfies $\sup_{w \in \mathbb{R}^d} \|\Delta(w)\| \leqslant c_F$.

iii) $\mathbb{E}_n \tilde{\eta}_{n+1} = 0$.

iv) $\sup \mathbb{E}_n \|\tilde{\eta}_{n+1}\|^4 \leqslant C_{\tilde{\eta},1}$.

v) $\inf \mathbb{E}_n \|\tilde{\eta}_{n+1}^-\| \geqslant C_{\tilde{\eta},2}$.

vi) $\sum_{i=0}^{\infty} \|\rho_{i+1}\|^2 \leqslant c_\rho$.

vii) $\sup \mathbb{E}_n \|\tilde{\rho}_{n+1}\|^4 \leqslant C_{\tilde{\rho}}$.

viii) $\sup \mathbb{E}_n \|\tilde{\rho}_{n+1}\|^2 \leqslant c_{\tilde{\rho},1}$.

ix) $\sup \chi_n^{-1/2} \mathbb{E}_n \left[\sum_{i=n}^{\infty} \gamma_i \|\tilde{\rho}_{i+1}\| \right] \leqslant c_{\tilde{\rho},2}$.

We are now in position to use the center-stable invariant manifold theorem to construct our sequence (w_n^-) that will be shown to stray away from zero. Consider $G : \mathbb{R}^{d^+} \to \mathbb{R}^{d^-}$ as in Proposition 6, with the extension provided by Remark 5. For all $n \in \mathbb{N}$, make the basis change $\mathbf{y}_n = (y_n^+, y_n^-) = P^{-1} y_n$ where (y_n) is the sequence of iterates provided in the statement of Proposition 3, and write $w_n = (w_n^+, w_n^-) = (y_n^+, y_n^- - G(y_n^+))$. It is obvious that $[y_n \to 0] \subset [w_n^- \to 0]$, which leads us to show in the remainder of the proof that $\mathbb{P}(w_n^- \to 0) = 0$. The expression of w_n^- as an iterative system is provided by the following lemma.

Lemma 9. *The adapted sequence (w_n^-) is provided by the iteration*

$$w_{n+1}^- = w_n^- + \gamma_n H_n w_n^- + \gamma_n (e_{n+1} + r_{n+1} + \tilde{r}_{n+1}),$$

where the sequences (e_n), (r_n) and $(\tilde{r}_n)$ are $\mathbb{R}^{d^-}$–valued, (H_n) is $\mathbb{R}^{d^- \times d^-}$–valued, and all are adapted to $(\mathcal{F}_n)$. Furthermore, there exists constants $C_\times > 0$ and constants $c_\times > 0$ that are as small as needed, such as the following events hold with probability one:

i) $\|H_n + J^-\| \leqslant c_H$.

ii) $\mathbb{E}_n e_{n+1} = 0$.

iii) $\sup \mathbb{E}_n \|e_{n+1}\|^4 \leqslant C_{e,1}$.

iv) $\inf \mathbb{E}_n \|e_{n+1}\| \geqslant C_{e,2}$.

v) $\sum_{i=0}^{\infty} \|r_{i+1}\|^2 \leqslant c_r$.

vi) $\sup \mathbb{E}_n \|\tilde{r}_{n+1}\|^4 \leqslant C_{\tilde{r}}$.

vii) $\sup \mathbb{E}_n \|\tilde{r}_{n+1}\|^2 \leqslant c_{\tilde{r},1}$.

viii) $\sup \chi_n^{-1/2} \mathbb{E}_n \left[\sum_{i=n}^{\infty} \gamma_i \|\tilde{r}_{i+1}\| \right] \leqslant c_{\tilde{r},2}$.

Proof. Let $(\tilde{\eta}_n^+, \tilde{\eta}_n^-) = P^{-1}\tilde{\eta}_n$, $(\varrho_n^+, \varrho_n^-) = P^{-1}\varrho_n$, and $(\tilde{\varrho}_n^+, \tilde{\varrho}_n^-) = P^{-1}\tilde{\varrho}_n$ with the obvious dimensions. It is clear that

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

Furthermore,

$$\begin{aligned}
G(y_{n+1}^+) &= G(y_n^+) + J_G(y_n^+)(y_{n+1}^+ - y_n^+) + \xi(y_n^+, y_{n+1}^+) \\
&= G(y_n^+) + \gamma_n J_G(y_n^+)(-\widetilde{D}^+(y_n) + \tilde{\eta}_{n+1}^+ + \varrho_{n+1}^+ + \tilde{\varrho}_{n+1}^+) + \xi(y_n^+, y_{n+1}^+),
\end{aligned}$$

where $\|\xi(y_n^+, y_{n+1}^+)\| \leqslant C\|y_{n+1}^+ - y_n^+\|^2$, see Remark 5. We therefore have

$$\begin{aligned}
w_{n+1}^- &= y_{n+1}^- - G(y_{n+1}^+) \\
&= w_n^- - \gamma_n \left(\widetilde{D}^-(y_n) - J_G(y_n^+)\widetilde{D}^+(y_n)\right) + \gamma_n (e_{n+1} + r_{n+1} + \tilde{r}_{n+1}) \\
&= w_n^- + \gamma_n F(w_n) + \gamma_n (e_{n+1} + r_{n+1} + \tilde{r}_{n+1}) \\
&= w_n^- + \gamma_n H_n w_n^- + \gamma_n (e_{n+1} + r_{n+1} + \tilde{r}_{n+1}),
\end{aligned} \tag{24}$$

where $H_n = -J^- + \Delta(w_n)$, and $\Delta(w)$ is the function given by Lemma 7 and controlled by Lemma 8–ii). The bound i) above follows. The other random variables at the right hand side of (24) are given as

$$\begin{aligned}
e_{n+1} &= \tilde{\eta}_{n+1}^- - J_G(y_n^+)\tilde{\eta}_{n+1}^+, \\
r_{n+1} &= \varrho_{n+1}^- - J_G(y_n^+)\varrho_{n+1}^+ + \xi(y_n^+, y_{n+1}^+)/\gamma_n, \\
\tilde{r}_{n+1} &= \tilde{\varrho}_{n+1}^- - J_G(y_n^+)\tilde{\varrho}_{n+1}^+.
\end{aligned}$$

Considering e_{n+1}, ii) and iii) follow immediately from Lemma 8–iii) and Lemma 8–iv) and the boundedness of $\|J_G\|$. To obtain iv), we write

$$\begin{aligned}
\mathbb{E}_n \|e_{n+1}\| &\geqslant \mathbb{E}_n \left\|\tilde{\eta}_{n+1}^-\right\| - \mathbb{E}_n \left\|J_G(y_n^+)\tilde{\eta}_{n+1}^+\right\| \\
&\geqslant \mathbb{E}_n \left\|\tilde{\eta}_{n+1}^-\right\| - \sup_{y^+} \|J_G(y^+)\| \, \mathbb{E}_n \left\|\tilde{\eta}_{n+1}^+\right\|
\end{aligned}$$

Since $\|J_G(y_n^+)\|$ can be taken as small as desired by Remark 5, iv) follows from Lemma 8–iv) and Lemma 8–v).

The conclusions on $(\tilde{r}_{n+1})$ follow from the analogous properties of $(\tilde{\varrho}_{n+1})$ provided by Lemma 8 and the boundedness of $\|J_G\|$.

It remains to establish the bound v) on (r_n). The contributions of the terms ρ_{n+1}^- and $J_G(y_n^+)\varrho_{n+1}^+$ to the sum in v) can be controlled by Lemma 8–vi). Regarding $\xi(y_n^+, y_{n+1}^+)/\gamma_n$, we have

$$\|\xi(y_n^+, y_{n+1}^+)\|/\gamma_n \leqslant C\gamma_n(\|\widetilde{D}^+(y_n)\|^2 + \|\tilde{\eta}_{n+1}^+\|^2 + \|\varrho_{n+1}^+\|^2 + \|\tilde{\varrho}_{n+1}^+\|^2),$$

and the contribution of this term is easily controlled by using the different bounds provided by Lemma 8. $\square$

To show that $\mathbb{P}(w_n^- \to 0) = 0$, we show the non-convergence to zero of a sequence (U_n), where U_n is a well chosen norm of w_n^-. As in [12], this construction goes as follows. Since the eigenvalues of J^- have all negative real parts, we know from a theorem of Lyapunov [24,

Cor. 2.2.4] that there exists a positive definite matrix Q such that $QJ^- + (J^-)^\top Q = -2I_{d^-}$. With this at hand, we set

$$U_n = \sqrt{\langle w_n^-, Qw_n^- \rangle},$$

in other words, U_n is the norm of w_n^- in the Euclidean space defined by the scalar product $\langle \cdot, \cdot \rangle_Q$, and is thus written $U_n = \|w_n^-\|_Q$. To deal with (U_n), we build on the technique developed by Tarrès in [38]. Preliminary inequalities involving U_n are needed.

7.2 Technical Inequalities on U_n

In the following $\lambda_{\min}$ (respectively $\lambda_{\max}$) denotes the minimal (respectively the maximal) eigenvalue of Q. In all the remainder, we shall assume that the constant c_H in Lemma 9–i) is small enough so that $\|Q(H_n + J^-)\| \leqslant 1/2$. During the proof, we shall keep track of some constants denoted as $C_\times$, where $\times$ is the number of the lemma where these constants are introduced.

Lemma 10. *There is a $(\mathscr{F}_n)$-adapted, $\mathbb{R}^{d^-}$-valued sequence (a_n) such that*

$$U_{n+1} - U_n \geqslant \gamma_n \langle a_n, e_{n+1} + r_{n+1} + \tilde{r}_{n+1} \rangle$$

with probability one. Furthermore, $\|a_n\| \leqslant \lambda_{\max}/\lambda_{\min}$.

Proof. As shown in [12, page 405], for any two vectors $a, b \in \mathbb{R}^{d^-}$, with $a \neq 0$, it holds that:

$$\|a + b\|_Q - \|a\|_Q \geqslant \frac{\langle a, b \rangle_Q}{\|a\|_Q}. \tag{25}$$

Furthermore, it is obvious that $\|b\|_Q \geqslant \langle u, b \rangle_Q$, with u being an arbitrary vector such that $\|u\|_Q = 1$. Therefore, if $U_n = 0$, then we have from the previous lemma that

$$U_{n+1} = U_{n+1} - U_n \geqslant \gamma_n \langle u, e_{n+1} + r_{n+1} + \tilde{r}_{n+1} \rangle_Q. \tag{26}$$

Otherwise, if $U_n \neq 0$, then, using Equation (25), we obtain:

$$\begin{aligned}
U_{n+1} - U_n &\geqslant U_n^{-1} \langle w_n^-, w_{n+1}^- - w_n^- \rangle_Q \\
&= \gamma_n U_n^{-1} \langle w_n^-, QH_n w_n^- \rangle + \gamma_n U_n^{-1} \langle w_n^-, e_{n+1} + r_{n+1} + \tilde{r}_{n+1} \rangle_Q.
\end{aligned}$$

Noticing from the definition of the matrix Q through Lyapounov's theorem that $\langle x, -QJ^-x \rangle = \|x\|^2$, and recalling that $\|Q(H_n + J^-)\| \leqslant 1/2$, we obtain that $\langle w_n^-, QH_n w_n^- \rangle \geqslant 0$, thus,

$$U_{n+1} - U_n \geqslant \gamma_n U_n^{-1} \langle w_n^-, e_{n+1} + r_{n+1} + \tilde{r}_{n+1} \rangle_Q.$$

The result follows by taking $a_n = U_n^{-1} Q w_n^-$ if $U_n \neq 0$ and $a_n = Qu$ otherwise. $\qquad\square$

Lemma 11. *It holds that*

$$\mathbb{E}_n U_{n+1}^2 - U_n^2 \geqslant C_{11} \gamma_n^2 - 2\lambda_{\max} \gamma_n U_n (\mathbb{E}_n \|r_{n+1}\| + \mathbb{E}_n \|\tilde{r}_{n+1}\|), \tag{27}$$

where $C_{11} = \lambda_{\min} C_{e,2}^2/2$ from Lemma 9–iv).

Proof. We have

$$\begin{aligned} U_{n+1}^2 - U_n^2 &= \|w_{n+1}^-\|_Q^2 - \|w_n^-\|_Q^2 = 2\langle w_n^-, w_{n+1}^- - w_n^-\rangle_Q + \|w_{n+1}^- - w_n^-\|_Q^2 \\ &= 2\gamma_n\langle w_n^-, QH_n w_n^-\rangle + 2\gamma_n\langle w_n^-, Qe_{n+1}\rangle + 2\gamma_n\langle w_n^-, r_{n+1} + \tilde{r}_{n+1}\rangle_Q + \|w_{n+1}^- - w_n^-\|_Q^2 \\ &\geqslant 2\gamma_n\langle w_n^-, Qe_{n+1}\rangle + 2\gamma_n\langle w_n^-, r_{n+1} + \tilde{r}_{n+1}\rangle_Q + \|w_{n+1}^- - w_n^-\|_Q^2, \end{aligned}$$

remembering that $\langle w_n^-, QH_n w_n^-\rangle \geqslant 0$ as in the previous proof. Moreover,

$$\begin{aligned} \|w_{n+1}^- - w_n^-\|_Q^2 &= \gamma_n^2\|H_n w_n^- + e_{n+1} + r_{n+1} + \tilde{r}_{n+1}\|_Q^2 \\ &= \gamma_n^2\|e_{n+1}\|_Q^2 + \gamma_n^2\|H_n w_n^- + r_{n+1} + \tilde{r}_{n+1}\|_Q^2 + 2\gamma_n^2\langle e_{n+1}, H_n w_n^- + r_{n+1} + \tilde{r}_{n+1}\rangle_Q \\ &\geqslant \gamma_n^2\|e_{n+1}\|_Q^2 + 2\gamma_n^2\langle e_{n+1}, H_n w_n^-\rangle_Q + 2\gamma_n^2\langle e_{n+1}, r_{n+1} + \tilde{r}_{n+1}\rangle_Q. \end{aligned}$$

Since $\mathbb{E}_n e_{n+1} = 0$, we obtain that

$$\mathbb{E}_n U_{n+1}^2 - U_n^2 \geqslant \gamma_n^2 \mathbb{E}_n\|e_{n+1}\|_Q^2 - 2\gamma_n^2\mathbb{E}_n\left|\langle e_{n+1}, r_{n+1} + \tilde{r}_{n+1}\rangle_Q\right| - 2\gamma_n\mathbb{E}_n\left|\langle w_n^-, r_{n+1} + \tilde{r}_{n+1}\rangle_Q\right|.$$

Using Lemma 9, we have $\mathbb{E}_n\|e_{n+1}\|_Q^2 \geqslant \lambda_{\min}\mathbb{E}_n\|e_{n+1}\|^2 \geqslant \lambda_{\min}C_{e,2}^2$. Moreover,

$$\mathbb{E}_n\left|\langle e_{n+1}, r_{n+1} + \tilde{r}_{n+1}\rangle_Q\right| \leqslant (\mathbb{E}_n\|e_{n+1}\|_Q^2)^{1/2}(\mathbb{E}_n\|r_{n+1} + \tilde{r}_{n+1}\|_Q^2)^{1/2},$$

which can be made as small as wished thanks to the bounds given by iii), v) and vii) in the statement of Lemma 9, and played against $\mathbb{E}_n\|e_{n+1}\|^2$ to provide the term $C_{11}\gamma_n^2$ at the right hand side of Inequality (27). We also have $|\langle w_n^-, r_{n+1} + \tilde{r}_{n+1}\rangle_Q| \leqslant \|w_n^-\|_Q(\|r_{n+1}\|_Q + \|\tilde{r}_{n+1}\|_Q) \leqslant \lambda_{\max}U_n(\|r_{n+1}\| + \|\tilde{r}_{n+1}\|)$, which proves the lemma. $\square$

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

Lemma 12. *There is a constant $C_{12} > 0$ such that if for $n \in \mathbb{N}$, U_n is such that $U_n^2 \leqslant L\chi_n$ for some $L > 0$, then*

$$(U_{n+1} - U_n)^2 \leqslant C_{12}\gamma_n^2(L + \|r_{n+1}\|^2 + \|r_{n+1}'\|^2 + \|e_{n+1}\|^2).$$

Proof. We have

$$\begin{aligned} (U_{n+1} - U_n)^2 &= (\|w_{n+1}^-\|_Q - \|w_n^-\|_Q)^2 \leqslant \|w_{n+1}^- - w_n^-\|_Q^2 \\ &\leqslant C\gamma_n^2(\|H_n w_n^-\|_Q^2 + \|e_{n+1}\|_Q^2 + \|r_{n+1}\|_Q^2 + \|\tilde{r}_{n+1}\|_Q^2) \\ &\leqslant C\gamma_n^2(U_n^2 + \|e_{n+1}\|^2 + \|r_{n+1}\|^2 + \|\tilde{r}_{n+1}\|) \\ &\leqslant C\gamma_n^2(L\chi_0 + \|e_{n+1}\|^2 + \|r_{n+1}\|^2 + \|\tilde{r}_{n+1}\|), \end{aligned}$$

where the last inequality comes from the fact that $U_n^2 \leqslant L\chi_n \leqslant L\chi_0$. $\square$

7.3 Proof of Proposition 3 by Proving that $\mathbb{P}(U_n \to 0) = 0$

For $N \in \mathbb{N}$ and $L > 0$, denote

$$\tau_N(L) := \inf\{k \geqslant N : U_k^2 \geqslant L\chi_k\}.$$

The following lemma is an adaptation of [38, Lemma 1] to our setting.

Lemma 13. *For $L > 0$, if the quantity $\max(c_{\tilde{r},2}, c_r^{1/2})\sqrt{L}$ is small enough, then there is a constant $C_{13} = C_{13}(L) > 0$ such that for all $N \in \mathbb{N}$, $\mathbb{P}_N(\tau_N(L) < +\infty) \geqslant C_{13}(L)$.*

Proof. Fixing N, L, with small notational abuse we will write in this proof $\tau := \tau_N(L)$. Notice that for $N \leqslant n < \tau$, we have $U_n \leqslant \sqrt{L\chi_n} \leqslant \sqrt{L\chi_N}$. Lemma 11 shows that the random process $(Z_n)_{n \geqslant N}$ defined as

$$Z_n = U_n^2 - C_{11} \sum_{i=N}^{n-1} \gamma_i^2 + 2\lambda_{\max} \sum_{i=N}^{n-1} \gamma_i U_i(\|r_{i+1}\| + \|\tilde{r}_{i+1}\|)$$

is a submartingale. Thus, the stopped process $(Z_{n \wedge \tau})$ is a $\mathscr{F}_n$–submartingale, and it holds that $\mathbb{E}_N[Z_{n \wedge \tau}] \geqslant Z_N$, which implies:

$$\mathbb{E}_N[U_{n \wedge \tau}^2] \geqslant U_N^2 + C_{11} \sum_{i=N}^{n-1} \gamma_i^2 \mathbb{P}_N(\tau > n) - 2\lambda_{\max}\sqrt{L\chi_N}\mathbb{E}_N\left[\sum_{i=N}^{+\infty} \gamma_i(\|r_{i+1}\| + \|\tilde{r}_{i+1}\|)\right].$$

By Cauchy-Schwarz's inequality and the bound of Lemma 9–v), it holds that $\sum_{i=N}^{\infty} \gamma_i\|r_{i+1}\| \leqslant \sqrt{\chi_N c_r}$. Similarly, using Lemma 9–viii), we obtain that $\mathbb{E}_N[\sum_{i=N}^{\infty} \gamma_i\|\tilde{r}_{i+1}\|] \leqslant c_{\tilde{r},2}\sqrt{\chi_N}$. Therefore,

$$\mathbb{E}_N[U_{n \wedge \tau}^2 - U_N^2] \geqslant \left(C_{11} \sum_{i=N}^{n-1} \gamma_i^2 - 2\lambda_{\max}\sqrt{Lc_r}\chi_N - 2\lambda_{\max}c_{\tilde{r},2}\sqrt{L}\chi_N\right)\mathbb{P}_N(\tau > n)$$

$$- \left(2\lambda_{\max}\sqrt{Lc_r}\chi_N + 2\lambda_{\max}c_{\tilde{r},2}\sqrt{L}\chi_N\right)\mathbb{P}_N(\tau \leqslant n).$$

Thus, if $\lambda_{\max}\max(c_{\tilde{r},2}, c_r^{1/2})\sqrt{L} \leqslant C_{11}/8$, which can be assumed due to the expression of C_{11} provided by Lemma 11, then,

$$\mathbb{E}_N[U_{n \wedge \tau}^2 - U_N^2] \geqslant C_{11}\left((\chi_N - \chi_n) - \frac{1}{2}\chi_N\right)\mathbb{P}_N(\tau > n) - \frac{C_{11}\chi_N}{2}\mathbb{P}_N(\tau \leqslant n) \tag{28}$$

$$\geqslant C_{11}\left(\frac{\chi_N - 2\chi_n}{2}\right)\mathbb{P}_N(\tau > n) - \frac{C_{11}\chi_N}{2}\mathbb{P}_N(\tau \leqslant n).$$

On the other hand, if $\tau = N$, then $U_{n \wedge \tau}^2 - U_N^2 = 0$ and if $\tau > N$, then

$$\begin{aligned}
U_{n \wedge \tau}^2 - U_N^2 &\leqslant U_{n \wedge \tau}^2 \\
&\leqslant U_n^2 \mathbb{1}_{\tau > n} + U_\tau^2 \mathbb{1}_{\tau \leqslant n} \\
&\leqslant L\chi_n \mathbb{1}_{\tau > n} + 2(U_{\tau-1}^2 + (U_\tau - U_{\tau-1})^2)\mathbb{1}_{\tau \leqslant n} \\
&\leqslant L\chi_n \mathbb{1}_{\tau > n} + 2L\chi_N \mathbb{1}_{\tau \leqslant n} + (U_\tau - U_{\tau-1})^2 \mathbb{1}_{\tau \leqslant n},
\end{aligned} \tag{29}$$

where the last inequality follows from the fact that $U_{\tau-1} \leqslant L\chi_{\tau-1} \leqslant L\chi_N$.

Using Lemma 12, we also obtain:

$$\mathbb{E}_N[(U_\tau - U_{\tau-1})^2 \mathbb{1}_{N < \tau \leqslant n}] \leqslant C_{12}\mathbb{E}_N\left[\sum_{i=N+1}^{n} \mathbb{1}_{\tau=i}\gamma_i^2\left(L + \|e_i\|^2 + \|r_i\|^2 + \|\tilde{r}_i\|^2\right)\right]$$

$$\leqslant C_{12}(L + c_r)\chi_N\mathbb{P}(\tau \leqslant n) + C_{12}\mathbb{E}_N\left[\sum_{i=N+1}^{n} \gamma_i^2 \mathbb{1}_{\tau=i}(\|e_i\|^2 + \|\tilde{r}_i\|^2)\right] \tag{30}$$

By the Cauchy-Schwarz inequality, we obtain that $\mathbb{E}_N[\mathbb{1}_{\tau=i} \|\tilde{r}_i\|^2] \leqslant \mathbb{P}_N(\tau = i)^{1/2} C_{\tilde{r}}^{1/2}$ from Lemma 9, and similarly, $\mathbb{E}_N[\mathbb{1}_{\tau=i} \|e_i\|^2] \leqslant C_{e,1}^{1/2} \mathbb{P}_N(\tau = i)^{1/2}$. Thus, combining Equations (29) and (30), we obtain:

$$
\begin{aligned}
\mathbb{E}_N[U_{n\wedge\tau}^2 - U_N^2] \leqslant{} & L\chi_n \mathbb{P}_N(\tau > n) + (2L + C_{12}L + C_{12}c_r)\chi_N \mathbb{P}_N(\tau \leqslant n) \\
& + C_{12}(C_{e,1}^{1/2} + C_{\tilde{r}}^{1/2})\chi_N \mathbb{P}_N(\tau \leqslant n)^{1/2} \,.
\end{aligned}
\tag{31}
$$

Finally, combining Equations (28) and (31), we obtain:

$$
\begin{aligned}
C_{11}\left(\frac{\chi_N - 2\chi_n}{2}\right)\mathbb{P}_N(\tau > n) \leqslant{} & L\chi_n\mathbb{P}_N(\tau > n) + (2L + C_{12}L + C_{12}c_r + C_{11}/2)\chi_N\mathbb{P}_N(\tau \leqslant n) \\
& + C_{12}(C_{e,1}^{1/2} + C_{\tilde{r}}^{1/2})\chi_N\mathbb{P}_N(\tau \leqslant n)^{1/2} \,.
\end{aligned}
$$

And letting n tend to infinity, we obtain:

$$
\frac{C_{11}}{2}\mathbb{P}_N(\tau = \infty) \leqslant \left(2L + C_{12}L + C_{12}c_r + C_{11}/2 + C_{12}(C_{e,1}^{1/2} + C_{\tilde{r}}^{1/2})\right)(1 - \mathbb{P}_N(\tau = \infty))^{1/2} \,.
$$

Similarly to [38, Proof of Lemma 1], this inequality shows the existence of a constant C_{13}, that depends only on $L, C_{11}, C_{12}, C_{e,1}, c_r, C_{\tilde{r}}$ such that $\mathbb{P}_N(\tau(N, L) = \infty) \leqslant (1 - C_{13}(L))$, which completes the proof. $\qquad\square$

The following lemma is an adaptation of [38, Lemma 2] to our setting.

Lemma 14. *If L is chosen such that*

$$
\sqrt{L} \geqslant \frac{\lambda_{\max}}{\lambda_{\min}}\left(2c_r^{1/2} + 16C_{e,1}^{1/4} + 16c_{\tilde{r},2}\right),
$$

then for all $N \in \mathbb{N}$ and $n \geqslant N$,

$$
\mathbb{P}_n(\liminf U_n > 0)\mathbb{1}_{\tau_N(L)=n} \geqslant \frac{1}{2}\mathbb{1}_{\tau_N(L)=n} \,.
$$

Proof. The idea of the proof is to use Lemma 10 and to show that there is some $0 < b < \sqrt{L}/2$ such that, with probability at least 1/2, the norm of $\sum_{i=n}^{\infty} \gamma_n\langle a_i, e_{i+1} + r_{i+1} + \tilde{r}_{i+1}\rangle$ will be less that $b\sqrt{\chi_n}$. Since, when $\tau_N(L) = n$, it holds that $U_n \geqslant \sqrt{L}$, this will show that $\liminf U_n > 0$ at least with probability 1/2. To simplify the notations we will denote in this proof

$$
E_{n+1}, R_{n+1}, \tilde{R}_{n+1} = \langle a_n, e_{n+1}\rangle, \langle a_n, r_{n+1}\rangle, \langle a_n, \tilde{r}_{n+1}\rangle \,.
$$

Fix $N \in \mathbb{N}$, $n \geqslant N$ and notice that by Doob's inequality:

$$
\mathbb{E}_n\left[\sup_{k \geqslant n}\left|\sum_{i=n}^{k} \gamma_i E_{i+1}\right|^2\right] \leqslant 4\mathbb{E}_n\left[\sum_{i=n}^{\infty} \gamma_i^2 E_{i+1}^2\right] \leqslant 4\frac{\lambda_{\max}^2}{\lambda_{\min}^2}C_{e,1}^{1/2}\chi_n \,,
$$

where we have used the fact that $\mathbb{E}_n[\|e_{n+1}\|^4] \leqslant C_{e,1}$ and $\lambda_{\min}\|a_n\| \leqslant \lambda_{\max}$. Moreover,

$$
\mathbb{E}_n\left[\sup_{k \geqslant n}\left|\sum_{i=n}^{k} \gamma_i \tilde{R}_{i+1}\right|\right] \leqslant \|a_n\|\mathbb{E}_n\left[\sum_{i=n}^{\infty} \gamma_i \|\tilde{r}_{i+1}\|\right] \leqslant \frac{c_{\tilde{r},2}\lambda_{\max}}{\lambda_{\min}}\sqrt{\chi_n} \,.
$$

Therefore, for any constant $b > 0$, applying Markov's inequality, we obtain:

$$\mathbb{P}_n \left(\inf_{k \geqslant n} \sum_{i=n}^k \gamma_i E_{i+1} < -b\sqrt{\chi_n} \right) \leqslant \mathbb{P}_n \left(\sup_{k \geqslant n} \left| \sum_{i=n}^k \gamma_i E_{i+1} \right|^2 \geqslant b^2 \chi_n \right) \leqslant \frac{4\lambda_{\max}^2 C_{e,1}^{1/2}}{\lambda_{\min}^2 b^2} ,$$

and

$$\mathbb{P}_n \left(\inf_{k \geqslant n} \sum_{i=n}^k \gamma_i \tilde{R}_{i+1} < -b\sqrt{\chi_n} \right) \leqslant \mathbb{P}_n \left(\left| \sup_{k \geqslant n} \sum_{i=n}^k \gamma_i \tilde{R}_{i+1} \right| \geqslant b\sqrt{\chi_n} \right) \leqslant \frac{c_{\tilde{r},2}\lambda_{\max}}{\lambda_{\min} b} .$$

Thus, denoting Γ_1 the event:

$$\Gamma_1 := \left[\inf_{k \geqslant n} \sum_{i=n}^k \gamma_i \tilde{R}_{i+1} < -b\sqrt{\chi_n} \right] \bigcup \left[\inf_{k \geqslant n} \sum_{i=n}^k \gamma_i E_{i+1} < -b\sqrt{\chi_n} \right]$$

and fixing $b = 4\lambda_{\max}/\lambda_{\min} \left(C_{e,1}^{1/4} + c_{\tilde{r},2} \right)$, we obtain

$$\mathbb{P}_n(\Gamma_1) \leqslant \frac{4\lambda_{\max}^2 C_{e,1}^{1/2}}{\lambda_{\min}^2 b^2} + \frac{c_{\tilde{r},2}\lambda_{\max}}{\lambda_{\min} b} \leqslant \frac{1}{2} .$$

Moreover, for such b, using Lemma 10, we obtain for $k \geqslant n$,

$$(U_k - U_n)\mathbb{1}_{\Gamma_1^c} \geqslant \mathbb{1}_{\Gamma_1^c} \sum_{i=n}^k \gamma_i(E_{i+1} + R_{i+1} + \tilde{R}_{i+1}) \geqslant \mathbb{1}_{\Gamma_1^c} \sum_{i=n}^k \gamma_i R_{i+1} - 2b\sqrt{\chi_n}\mathbb{1}_{\Gamma_1^c} .$$

Furthermore, by the Cauchy-Schwarz inequality, it also holds that:

$$\sup_{k \geqslant n} \left| \sum_{i=n}^k \gamma_i R_{i+1} \right| \leqslant \frac{\lambda_{\max}}{\lambda_{\min}} \sum_{i=n}^\infty \gamma_i \|r_{i+1}\| \leqslant \frac{\lambda_{\max}}{\lambda_{\min}} \chi_n^{1/2} c_r^{1/2} ,$$

and finally, for $k \geqslant n$, we obtain,

$$U_k \mathbb{1}_{\Gamma_1^c} \geqslant U_n \mathbb{1}_{\Gamma_1^c} - (c_r^{1/2}\lambda_{\max}/\lambda_{\min} + 2b)\sqrt{\chi_n}\mathbb{1}_{\Gamma_1^c} .$$

In Particular, if $\sqrt{L} \geqslant 2c_r^{1/2}\lambda_{\max}/\lambda_{\min} + 4b$, then on the event $\Gamma_1^c \cap [\tau_N(L) = n]$ it holds

$$U_k \geqslant U_n - \sqrt{L}/2\sqrt{\chi_n} \geqslant \left(\sqrt{L} - \sqrt{L}/2 \right) \sqrt{\chi_n} \geqslant \sqrt{L}\sqrt{\chi_n}/2 .$$

This shows that $\Gamma_1^c \cap [\tau_N(L) = n] \subset [\liminf U_n > 0] \cap [\tau_N(L) = n]$. Thus,

$$\begin{aligned}
\mathbb{P}_n([\liminf U_n > 0])\mathbb{1}_{\tau_N(L)=n} &\geqslant \mathbb{P}_n(\Gamma_1^c \cap [\tau_N(L) = n]) \\
&= \mathbb{P}_n(\Gamma_1^c)\mathbb{1}_{\tau_N(L)=n} \\
&\geqslant \frac{1}{2}\mathbb{1}_{\tau_N(L)=n} ,
\end{aligned}$$

which completes the proof. $\qquad\square$

We are now in position to complete the proof of Proposition 3 by proving that $\mathbb{P}(U_n \to 0) = 0$. This will be the content of the following lemma.

Lemma 15. $\mathbb{P}(U_n \to 0) = 0$.

Proof. We proceed as in [38]. Indeed, we can always choose $c_r, c_{\tilde{r},2}$ and L respectively small and large enough such that the prerequisites of Lemmas 13 and 14 are satisfied. Thus, by Lemma 13, there is $C_{13}(L)$ such that $\mathbb{P}(\tau_N(L) < +\infty) \geqslant C_{13}(L) > 0$. Then, applying Lemma 14, we obtain:

$$\mathbb{E}_N\big[\mathbb{1}_{\liminf U_n > 0}\big] \geqslant \sum_{i=N}^{\infty} \mathbb{E}_N\big[\mathbb{1}_{\liminf U_n > 0} \mathbb{1}_{\tau_N(L)=i}\big] \geqslant \frac{1}{2} \sum_{i=N}^{\infty} \mathbb{E}_N\big[\mathbb{1}_{\tau_N(L) = i}\big] = \frac{C_{13}(L)}{2}.$$

Since $[\liminf U_n > 0] \in \mathscr{F}_\infty = \sigma(\bigcup_{i=0}^{\infty} \mathscr{F}_i)$, we know by Lévy's zero-one law (see e.g. [41, Theorem 14.2]) that $\lim_{N \to \infty} \mathbb{E}_N\big[\mathbb{1}_{\liminf U_n > 0}\big] = \mathbb{1}_{\liminf U_n > 0}$ almost surely. Thus, almost surely, $\mathbb{1}_{\liminf U_n > 0} = 1$, which shows that $\mathbb{P}([U_n \to 0]) = 0$. $\qquad\square$

Acknowledgments

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*, 137:91–124, 2011.

[2] M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer, Berlin, 1999.

[3] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions. *SIAM Journal on Control and Optimization*, 44(1):328–348, 2005.

[4] P. Bianchi, W. Hachem, and S. Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 30:1117–1147, 2022.

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

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

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

[8] J. Bolte and E. Pauwels. Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning. *Mathematical Programming*, 188:19–51, 2021.

[9] Jérôme Bolte and Edouard Pauwels. A mathematical model for automatic differentiation in machine learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 10809–10819, 2020.

[10] Jérôme Bolte, Tam Le, and Edouard Pauwels. Subgradient sampling for nonsmooth nonconvex minimization. *arXiv preprint arxiv:2202.13744*, 2022.

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

[12] O. Brandière and M. Duflo. Les algorithmes stochastiques contournent-ils les pièges? *Annales de l'Institut Henri Poincaré. Probabilités et Statistiques*, 32(3):395–427, 1996.

[13] James V. Burke and Jorge J. Moré. On the identification of active constraints. *SIAM Journal on Numerical Analysis*, 25(5):1197–1211, 1988.

[14] 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.

[15] Michel Coste. *An introduction to o-minimal geometry*. Istituti editoriali e poligrafici internazionali Pisa, 2000.

[16] D. Drusvyatskiy D. Davis and V. Charisopoulos. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arxiv:1907.09547*, 2019.

[17] A. Daniilidis and D. Drusvyatskiy. Pathological subgradient dynamics. *SIAM Journal on Optimization*, 30(2):1327–1338, 2020.

[18] D. Davis and D. Drusvyatskiy. Proximal methods avoid active strict saddles of weakly convex functions. *Foundations of Computational Mathematics*, 22:561–606, 2022.

[19] D. Davis, D. Drusvyatskiy, S. Kakade, and J. D. Lee. Stochastic subgradient method converges on tame functions. *Foundations of Computational Mathematics*, 20:119–154, 2020.

[20] Damek Davis, Dmitriy Drusvyatskiy, and Liwei Jiang. Subgradient methods near active manifolds: saddle point avoidance, local convergence, and asymptotic normality. *arXiv preprint arxiv:2108.11832*, 2021.

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

[22] D. Drusvyatskiy and A. Lewis. Semi-algebraic functions have small subdifferentials. *Mathematical Programming*, 140:5–29, 2013.

[23] D. Drusvyatskiy and A. Lewis. Optimality, identifiability, and sensitivity. *Mathematical Programming*, 147:467–498, 2014.

[24] R. A. Horn and C. R. Johnson. *Topics in matrix analysis*. Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original.

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

[26] A. Kelley. The Stable, Center-Stable, Center, Center-Unstable, Unstable Manifolds. *Journal of Differential Equations*, 1966.

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

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

[29] T.Z. Lai and C.Z. Wei. A note on martingale difference sequences satisfying the local marcinkiewicz-zygmund condition. *Bulletin of the Institute of Mathematics, Academia Sinica*, 11:1–13, 1983.

[30] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Proceedings of the Annual Conference on Learning Theory*, volume 49, pages 1246–1257, 2016.

[31] A. Lewis. Active sets, nonsmoothness, and sensitivity. *SIAM Journal on Optimization*, 13:702–725, 2002.

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

[33] Ta Loi. Verdier and strict Thom stratifications in o-minimal structures. *Illinois Journal of Mathematics*, 42(2):347–356, 1998.

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

[35] R. Pemantle. Nonconvergence to unstable points in urn models and stochastic approximations. *The Annals of Probability*, 18(2):698–712, 1990.

[36] H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York, 1971.

[37] Sholom Schechtman. *Some Problems in Nonconvex Stochastic Optimization*. Theses, Université Gustave Eiffel, December 2021.

[38] P. Tarrès. Pièges répulsifs. *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 330(2):125–130, 2000.

[39] L. Van den Dries and C. Miller. Geometric categories and o-minimal structures. *Duke Mathematical Journal*, 84(2):497–540, 08 1996.

[40] Alex Wilkie. O-minimal structures. *Séminaire Bourbaki*, 985, 2007.

[41] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[42] S. J. Wright. Identifiable surfaces in constrained optimization. *SIAM Journal on Control and Optimization*, 31(4):1063–1079, 1993.

Appendix A   Definability and Whitney Stratifications

A.1   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 [7], where the authors analyzed the structure of the Clarke subdifferential of a definable function and extended the Kurdyka-Łojasiewicz inequality [27] to the nonsmooth setting. Nowadays a rich body of literature enforces this link, see e.g. [19, 22, 6, 1, 8]. A nice exposure about usefulness of o-minimal theory in optimization is [25]. Results on the Verdier and Whitney stratification of definable sets can be found in [15, 39, 33].

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 $\mathbb{R}^n\backslash A$.

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. These properties show that most of the functions that are used in optimization are definable. Examples of such are: semialgebraic functions, analytic functions restricted to a semialgebraic compact, exponential and logarithm (see e.g. [6, 5, 40]). In particular, it can be shown that the loss of a neural network is a definable function [19].

A.2   Whitney Stratification

From the definition of $\boldsymbol{d}_a$ in Equation (2), one can define the following distance between two vector spaces E_1, E_2:

$$\boldsymbol{d}(E_1, E_2) = \max\{\boldsymbol{d}_a(E_1, E_2), \boldsymbol{d}_a(E_2, E_1)\}. \tag{32}$$

Definition 7. *We say that a C^p stratification (S_i) satisfies a* Whitney-(a) *property, if for every couple of distinct strata S_i, S_j, for each $y \in S_i \cap \overline{S_j}$ and for each sequence (x_n) in S_j such that $x_n \to y$, it holds:*

$$\text{w-(a)} \quad \textit{There is } E \subset \mathbb{R}^d \textit{ such that} \quad \boldsymbol{d}(T_{x_n}S_j, E) \to 0 \implies T_yS_i \subset E. \tag{33}$$

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

It is known (see [15, 39]) 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 stratum. The following "projection formula" relates the Clarke subdifferential $\partial f(y)$ of f at y, to $\nabla_{X_i} f(y)$.

Lemma 16 (Projection formula, [7, 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 $\mathrm{Graph}(f)$, such that if one denotes by X_i the projection of S_i onto its first d coordinates, the restriction $f : X_i \to \mathbb{R}$ 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 16 has important consequences. One of them (see [19, Section 5]) is that every locally Lipschitz continuous and definable function is path-differentiable.

Lemma 17 ([19, Theorem 5.8]). *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function. If $\mathrm{Graph}(f)$ admits a Whitney C^1 stratification, then f is path-differentiable.*

Appendix B   Probabilistic Arguments and Proof of Lemma 8

The arguments to prove Lemma 8 are similar to the ones presented in [12, 35, 38]. For future references we have found convenient to present them for a general sequence (w_n), which might not be explicitly related to (y_n).

Let d be an integer, $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, $(\mathscr{F}_n)$ a filtration on it and (w_n) be a sequence in $\mathbb{R}^d$ verifying:

$$w_{n+1} = w_n + \gamma_n D_n(w_n) w_n + \gamma_n r_{1,n+1} + \gamma_n r_{2,n+1} + \gamma_n e_{n+1} \,, \tag{34}$$

where (γ_n) is a real-valued sequence, $(r_{1,n}), (r_{2,n}), (e_n)$ are $\mathbb{R}^d$-valued and adapted to $(\mathscr{F}_n)$ and for each $n \in \mathbb{N}$, $D_n : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ is some measurable map. Furthermore, assume that w_0 is $\mathscr{F}_0$-measurable.

Fix a deterministic, real-valued sequence (χ_n), four measurable functions $G_1, G_2, G_3, G_4 : \mathbb{R}^d \to \mathbb{R}_+$, an event $\Gamma \in \mathscr{F}$ and a measurable function $D : \mathbb{R}^d \to \mathbb{R}^d$. Assume that we want to prove the fact that $\mathbb{P}(\Gamma \cap [w_n \to 0]) = 0$ under the following set of assumptions.

Assumption 5. *On the event $\Gamma \cap [w_n \to 0]$, the following holds.*

i) For all $n \in \mathbb{N}$, $\mathbb{E}_n e_{n+1} = 0$ and

$$\limsup \mathbb{E}_n G_1(e_{n+1}) < +\infty\,, \quad \liminf \mathbb{E}_n G_2(e_{n+1}) > 0\,.$$

ii)

$$\sum_{i=0}^{\infty} \|r_{1,i+1}\|^2 < +\infty\,.$$

iii)

$$\limsup \chi_n \sum_{i=n}^{\infty} \gamma_i \mathbb{E}_n \|r_{2,i+1}\| = 0$$

and

$$\limsup \mathbb{E}_n G_3(r_{2,n+1}) < +\infty\,, \quad \lim \mathbb{E}_n G_4(r_{2,n+1}) = 0\,.$$

iv) $\limsup \|D_n(y_n) - D(y_n)\| = 0$.

Then, to prove that $\mathbb{P}([w_n \to 0] \cap \Gamma) = 0$ it is sufficient to prove it under the following, more easy to handle, assumption.

Assumption 6. *There are six, strictly positive, constants $C_{G_1}, C_{G_2}, C_{G_3}, c_{G_4}, c_{r_1}, c_{r_2}, c_D$, where $c_\times$ can be chosen as small as needed, such that almost surely the following holds.*

i) For all $n \in \mathbb{N}$, $\mathbb{E}_n e_{n+1} = 0$ and

$$\sup_{n \in \mathbb{N}} \mathbb{E}_n G_1(e_{n+1}) \leqslant C_{G_1}, \quad \inf_{n \in \mathbb{N}} \mathbb{E}_n G_2(e_{n+1}) \geqslant C_{G_2}.$$

ii)

$$\sum_{i=0}^{\infty} \|r_{1,i+1}\|^2 \leqslant c_{r_1}.$$

iii)

$$\sup_{n \in \mathbb{N}} \chi_n \sum_{i=n}^{\infty} \gamma_i \mathbb{E}_n \|r_{2,i+1}\| \leqslant c_{r_2}$$

and

$$\sup_{n \in \mathbb{N}} \mathbb{E}_n G_3(r_{2,n+1}) \leqslant C_{G_3}, \quad \sup_{n \in \mathbb{N}} \mathbb{E}_n G_4(r_{2,n+1})] \leqslant c_{G_4}.$$

iv) For all $n \in \mathbb{N}$, $\|D_n(y_n) - D(y_n)\| \leqslant c_D$.

To show that proving $\mathbb{P}([w_n \to 0]) = 0$ under Assumption 6 is sufficient we, for an arbitrary large event $A \in \mathscr{F}$, construct a sequence $(\tilde{w}_n)$ that is equal to (w_n) on $\Gamma \cap [w_n \to 0]$ but satisfies Assumption 6 almost surely.

1) *Equivalence between Assumption 5-(ii) and Assumption 6-(ii).* Notice that

$$\left[\sum_{i=0}^{\infty} \|r_{1,i+1}^2\| < +\infty \right] \subset \bigcap_{C \in \mathbb{Q}} \bigcup_{n_0 \in \mathbb{N}} \left[\sum_{i=n_0}^{\infty} \|r_{1,i+1}\|^2 \leqslant C \right].$$

Therefore, for any $C \in \mathbb{Q}$ and $\delta > 0$, there is $n_0 \in \mathbb{N}$ such that

$$\mathbb{P}(\Gamma \cap [w_n \to 0]) \leqslant \mathbb{P}\left(\Gamma \cap [w_n \to 0] \cap \left[\sum_{i=n_0}^{\infty} \|r_{1,i+1}\|^2 \leqslant C \right] \right) + \delta.$$

For $n \geqslant n_0$, denote $(\tilde{r}_{1,n+1})$ the sequence defined as $\tilde{r}_{1,n+1} = r_{1,n+1} \mathbb{1}_{\sum_{i=n_0}^{n+1} \|r_{1,i+1}\|^2 \leqslant C}$ and $(\tilde{w}_n)$ the sequence defined as:

$$\tilde{w}_{n+1} = \tilde{w}_n + \gamma_n D_n(\tilde{w}_n)\tilde{w}_n + \gamma_n \tilde{r}_{1,n+1} + \gamma_n r_{2,n+1} + \gamma_n e_{n+1}.$$

Shifting the sequences $(\tilde{w}_n), (\gamma_n), (D_n), (\tilde{r}_n), (r_n), (e_n)$ by n_0, we obtain that the shifted sequences satisfy Assumption 5, with Assumption 5-(ii) replaced by Assumption 6-(ii). Furthermore, by construction if we prove that $\mathbb{P}([\tilde{w}_n \to 0]) = 0$, then:

$$\mathbb{P}(\Gamma \cap [w_n \to 0]) \leqslant \mathbb{P}\left(\Gamma \cap [w_n \to 0] \cap \left[\sum_{i=n_0}^{\infty} \|r_{1,i+1}\|^2 \leqslant C\right]\right) + \delta$$

$$= \mathbb{P}\left(\Gamma \cap [\tilde{w}_n \to 0] \cap \left[\sum_{i=n_0}^{\infty} \|r_{1,i+1}\|^2 \leqslant C\right]\right) + \delta$$

$$\leqslant \mathbb{P}([\tilde{w}_n \to 0]) + \delta = \delta \,.$$

Since δ is arbitrary, this will show that $\mathbb{P}([w_n \to 0]) = 0$.

2) Equivalence between Assumption 5-(iii) and Assumption 6-(iii). In a similar manner, for any $\delta, c_{r_2}, c_{G_4} > 0$, there is $n_0 \in \mathbb{N}$ and $C_{G_3} > 0$ such that, up to an event of a probability less than δ, Assumption 6-(iii) holds for $n \geqslant n_0$. Thus, for $n \geqslant n_0$, denote $A_n \in \mathscr{F}_n$ the event

$$A_n = \left[\sup_{n_0 \leqslant k \leqslant n} \chi_k \sum_{i=k}^{\infty} \gamma_i \mathbb{E}_k \|r_{2,i+1}\| \leqslant c_{r_2} \,, \mathbb{E}_n G_3(r_{2,n+1}) \leqslant C_{G_3} \,, \mathbb{E}_n G_4(r_{2,n+1}) \leqslant c_{G_4}\right] \,.$$

Then, define $(\tilde{w}_n)$ with the same recursion than (w_n), but where $r_{2,n+1}$ is replaced by $r_{2,n+1} \mathbb{1}_{A_n}$. By construction Assumption 6-(iii) is satisfied and proving $\mathbb{P}([\tilde{w}_n \to 0]) = 0$ will imply $\mathbb{P}(\Gamma \cap [w_n \to 0]) \leqslant \delta$, which, since δ is arbitrary will show that $\mathbb{P}([w_n \to 0]) = 0$.

3) Equivalence betweeen Assumption 5-(iv) and Assumption 6-(iv). As previously, for any $\delta, c_D > 0$, there is $n_0 \in \mathbb{N}$ such that for $n \geqslant n_0$, Assumption 6-(iv) is satisfied on $\Gamma \cap [w_n \to 0]$ up to an event of probability less than δ. Thus, defining, for $n \geqslant n_0$, $(\tilde{w}_n)$ with the same equation as (w_n), but replacing $D_n(y_n)$ by $D_n(y_n)\mathbb{1}_{\|D_n(y_n)-D(y_n)\| \leqslant c_D} + D(y_n)\mathbb{1}_{\|D_n(y_n)-D(y_n)\| > c_D}$, and shifting all of the sequences by n_0, we obtain that $(\tilde{w}_n)$ satisfy Assumption 6-(iv) and that

$$\mathbb{P}([w_n \to 0] \cap \Gamma) \leqslant \mathbb{P}([\tilde{w}_n \to 0]) + \delta \,.$$

4) Equivalence between Assumption 5-(i) and Assumption 6-(i). Here, since we want to preserve the zero-mean assumption on (e_n), the construction is slightly different and goes back to [29]. Let (e'_n) be a sequence of bounded, zero-mean and i.i.d. random variables defined on an auxiliary probability space $(\Omega', \mathscr{F}', \mathbb{P}')$. On the probability space $(\Omega \times \Omega', \mathscr{F} \otimes \mathscr{F}', \mathbb{P} \otimes \mathbb{P}')$ define a filtration $(\tilde{\mathscr{F}}_n)$ as $\tilde{\mathscr{F}}_n = \mathscr{F}_n \otimes \sigma(\{e'_j : j \leqslant n\})$, where σ denotes the smallest sigma-algebra generated by a sequence of randoms variables.

As previously, for any $\delta > 0$, there is some $n_0 \in \mathbb{N}$ and some constants C_{G_1}, C_{G_2} such that, for $n \geqslant n_0$, Assumption 6-(i) is satisfied on $\Gamma \cap [w_n \to 0]$ up to an event with probability less than δ. For $n \geqslant n_0$, define $A_n \in \mathscr{F}_n$ as:

$$A_n = [\mathbb{E}_n e_{n+1} = 0, \mathbb{E}_n G_1(e_{n+1}) \leqslant C_{G_1}, \mathbb{E}_n G_2(e_{n+1}) \geqslant C_{G_2}] \,.$$

Then, define a sequence of $(\tilde{\mathscr{F}}_n)$ adapted random variables $(\tilde{e}_n)$ as:

$$\forall (a_1, a_2) \in \Omega \times \Omega', \quad \tilde{e}_n(a_1, a_2) = e_n(a_1)\mathbb{1}_{A_n}(a_1) + e'_n(a_2)\mathbb{1}_{A_n^c}(a_1) \,.$$

Let us define $(\tilde{w}_n)$ in the same manner as (w_n) but replacing (e_n) by $(\tilde{e}_n)$ and as previously shift all of the sequences by n_0. By construction such a sequence satisfies Assumption 6-(i) and as previously:

$$\mathbb{P}(\Gamma \cap [w_n \to 0]) \leqslant \mathbb{P}([\tilde{w}_n \to 0]) + \delta \,.$$

Thus, proving that $\mathbb{P}([\tilde{w}_n \to 0]) = 0$ will be sufficient to conclude that $\mathbb{P}(\Gamma \cap [w_n \to 0]) = 0$.

Appendix C   Issues with Asymptotic Pseudotrajectories

In the first version of this paper, Theorem 3 was stated in the case where M satisfied the following, weaker version, of angle condition.

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

$$f(x) - f(P_M(x)) \geqslant \alpha \|x - P_M(x)\| \implies \langle v, x - P_M(x) \rangle \geqslant \beta \|x - P_M(x)\|. \qquad (35)$$

The advantage of this version over the one used in this work is that Equation (35) is verified by a much larger family of functions than weakly convex ones. Typically, the function $(y, z) \mapsto -y^2 - |z|$ satisfies (35) but not the equation of Definition 5.

The main idea behind the initial (incorrect) proof was to establish the existence of $\alpha > 0$, such that, on the event $[x_n \to x_*]$, for n large enough, the iterates must satisfy the left-hand side of (35), almost surely. The (weak version) of angle condition then implied that they satisfy the right-hand side of (35), which is exactly what we need for the proof of Theorem 3 (more precisely, to establish Lemmas 4–5 and Proposition 5).

Unfortunately, the proof of the first claim was using [3, Theorem 4.1], which as noted by an anonymous reviewer was incorrect. To avoid future errors of a similar kind we provide in this section a detailed description of our mistake.

Before going further, let us emphasize, that up to our knowledge, most of the results based on the work of Benaïm, Hofbauer and Sorin are using [3, Theorem 4.2], which, appropriately restated (see Proposition 7), remains valid. In particular, we are not aware of any other works, where a mistake is due to an incorrect use of [3, Theorem 4.1].

Differential inclusions. In the following, we fix a set-valued map $\mathsf{H} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ (i.e. $\forall x \in \mathbb{R}^d$, $\mathsf{H}(x) \subset \mathbb{R}^d$) that satisfies assumptions of [3]. The relevant example for our work is the case where $\mathsf{H} = -\partial f$. We say that an absolutely continuous curve $\mathsf{x} : \mathbb{R}_+ \rightrightarrows \mathbb{R}^d$ is a solution of the differential inclusion (DI)

$$\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t)), \qquad (36)$$

with initial condition $x_0 \in \mathbb{R}^d$, if $\mathsf{x}(0) = x_0$ and (36) holds for almost every $t \in \mathbb{R}_+$. We denote $S_{\mathsf{H}}(x_0)$ the set of these solutions.

Stochastic approximation of differential inclusions. Let (γ_n) be a sequence of decreasing step-sizes and (x_n) a recursive algorithm satisfying

$$x_{n+1} \in x_n + \gamma_n \mathsf{H}(x_n) + \gamma_n U_{n+1}, \qquad (37)$$

where (U_n) is an $\mathbb{R}^d$-valued sequence of perturbations. One of the purposes of the work of Benaïm, Höfbauer and Sorin ([3]) was to show that under mild conditions on (γ_n) and (U_n) the iterates produced by Equation (37) closely follow a solution of the DI (36).

Indeed, let us define the linearly interpolated process $\mathsf{X} : \mathbb{R}_+ \to \mathbb{R}^d$ as:

$$\mathsf{X}(t) = x_n + \frac{t - \sum_{i=0}^n \gamma_i}{\gamma_{n+1}}(x_{n+1} - x_n), \quad \text{if } \sum_{i=0}^n \gamma_i \leqslant t \leqslant \sum_{i=0}^{n+1} \gamma_i. \qquad (38)$$

Then [3, Proof of Theorem 4.2] shows the following result.

Proposition 7. *Assume the following.* i) *The iterates (x_n) are bounded.* ii) *The step-sizes satisfy $\gamma_n \to 0$ and $\sum_{i=0}^{\infty} \gamma_i = +\infty$.* iii) *For all $T > 0$ it holds that*

$$\sup_{n \leqslant k \leqslant \tau(n,T)} \left\| \sum_{i=n}^{k-1} \gamma_i U_{i+1} \right\| \xrightarrow[n \to \infty]{} 0, \quad \text{where } \tau(n,T) = \inf \left\{ k \geqslant n : \sum_{i=n}^{k} \gamma_i \geqslant T \right\}.$$

Then, the family $(\mathsf{X}(t + \cdot))_{t \in \mathbb{R}_+}$ is relatively compact (in the space of bounded, continuous functions with the norm of uniform convergence on compact sets) and for any $t_n \to +\infty$ and $\mathsf{z} : \mathbb{R}_+ \to \mathbb{R}^d$ such that $\mathsf{X}(t_n + \cdot) \to \mathsf{z}$, it holds that z is a solution to (36).

Asymptotic pseudotrajectory. The asymptotic result of the previous paragraph is quite strong and is actually sufficient for most of the results of [3] about accumulation points of iterates given by a stochastic approximation recursion (in particular they are ICT sets). However, the linearly interpolated process (38) was claimed to possess a stronger property.

Following [3], a continuous curve $\mathsf{X} : \mathbb{R}_+ \to \mathbb{R}^d$ is said to be an asymptotic pseudotrajectory (APT) of (36) if for every $T > 0$, it holds that

$$\lim_{t \to +\infty} \sup_{h \in [0,T]} \inf_{\mathsf{x} \in S_{\mathsf{H}}(\mathsf{X}(t))} \| X(t + h) - \mathsf{x}(h) \| = 0. \tag{39}$$

The following result was claimed to be true.

Proposition 8. *[3, Theorem 4.1] Assume that $\mathsf{z} : \mathbb{R}_+ \to \mathbb{R}^d$ is bounded. Then there is equivalence between.*

1. *z is an APT.*

2. *z is uniformly continuous and if there is $t_n \to +\infty$ and $\tilde{\mathsf{z}} : \mathbb{R}_+ \to \mathbb{R}^d$ such that $\mathsf{z}(t_n + \cdot) \to \tilde{\mathsf{z}}$ (in the sense of uniform convergence on compact intervals), then $\tilde{\mathsf{z}}$ is a solution to the DI* (36).

In the first version of this work, the implication 2 $\implies$ 1 was used to obtain the fact that asymptotically the iterates of SGD satisfy the left-hand side of (35) (and thus, by the weak angle condition, the right-hand side). Following that, the proof was identical to the one of Theorem 3. Unfortunately, while 1 $\implies$ 2, we actually do not have the reverse implication. Indeed, the following example was provided by an anonymous reviewer.

Consider $\mathsf{z}(t) = \frac{1}{t+1}$ and $\mathsf{H} = \partial | \cdot |$. Thus, $\mathsf{H}(x) = \text{sign}(x)$ if $x \neq 0$ and $\mathsf{H}(0) = [-1, 1]$. The curve z is uniformly continuous, and for any $T > 0$ and $t_n \to +\infty$,

$$\sup_{h \in [0,T]} \| \mathsf{z}(t_n + h) \| \to 0.$$

Since $\tilde{\mathsf{z}} \equiv 0$ is a solution to the DI associated to H we indeed have 2. Nevertheless, for any $n \in \mathbb{N}$ and x a solution to the DI (36) starting at $\mathsf{z}(t_n)$, we have $\mathsf{x}(T) = \mathsf{z}(t_n) + T$. Thus,

$$\lim_{t_n \to +\infty} \sup_{h \in [0,T]} \inf_{\mathsf{x} \in S_{\mathsf{H}}(\mathsf{z}(t_n))} \| \mathsf{z}(t_n + h) - \mathsf{x}(h) \| \geqslant T, \tag{40}$$

and the first point of Proposition 8 is not satisfied.