Some Problems in Nonconvex Stochastic Optimization

Sholom Schechtman

Quelques problèmes en optimisation non convexe et stochastique

Some Problems in Nonconvex Stochastic Optimization

Thèse de doctorat de l'Université Gustave Eiffel

École doctorale n° 532, Mathématiques et sciences et technologies de l'information et de la communication (MSTIC)
Spécialité de doctorat: Mathématiques appliquées
Unité de recherche : LIGM

Thèse présentée et soutenue à l'Université Gustave Eiffel, le 14/12/2021, par

Sholom SCHECHTMAN

Composition du Jury

Aris DANIILIDIS Professeur, Universidad de Chile	Rapporteur
Jérôme BOLTE Professeur, Toulouse Capitole, TSE	Rapporteur
Gersende FORT Directrice de recherche, CNRS, IMT	Examinatrice
Eric MOULINES Professeur, École Polytechnique	Examinateur
Michel BENAÏM Professeur, Université de Neuchâtel	Examinateur

Encadrement de la thèse

Walid HACHEM Directeur de recherche, CNRS, Université Gustave Eiffel	Directeur de thèse
Pascal BIANCHI Professeur, Télécom Paris	Co-directeur de thèse

Love requires an Object,
But this varies so much,

W. H. Auden

Remerciements

Je tenais tout d'abord à remercier mes directeurs de thèse, Walid et Pascal, de m'avoir pris en thèse et soutenu pendant ces trois ans. Vous m'avez beaucoup apporté par vos conseils et soutiens, que ce soit sur le plan professionnel ou sur le plan humain. Je vous remercie tout particulièrement pour votre confiance dans ma manière de travailler (je sais que je n'étais pas l'élève le plus assidu) ainsi que dans mes choix des sujets de recherche. Cette confiance m'a vraiment permis de me sentir libre pendant ces trois années, ça a énormément compté pour moi; j'espère que ça a donné quelque chose de bien.

Je remercie vivement Jérôme Bolte et Aris Daniilidis pour avoir accepté de rapporter cette thèse. Il m'est dur de sous-estimer l'influence qu'ont eue vos travaux concernant l'optimisation semi-algébrique/o-minimale sur ma recherche et c'est un véritable honneur de pouvoir vous présenter mes (quelques) résultats.

Je voudrais remercier Michel Benaïm, Éric Moulines et Gersende Fort d'avoir accepté de participer à ma soutenance. Vos travaux ont eu une influence considérable sur la partie probabiliste de ma thèse. En particulier la méthode de l'ODE/inclusion différentielle développée par Michel Benaïm est à la base de tous(!) mes papiers publiés durant ces trois années. Éric et Gersende, on commence déjà à travailler ensemble et je suis impatient de collaborer avec vous tout prochainement.

Je remercie les gens de Télécom et de Marnes que j'ai croisés tout au long de ma thèse. J'ai toujours trouvé dans ces labos éloignés un accueil chaleureux. Merci Anas mon co-thésard/co-bureau/(co-)docteur pour ces trois (co-)années et tes qualités humaines. Je n'ai aucun doute que nos chemins se recroiseront encore.

Je ne listerai pas ici tous mes amis car un ordre ne leur ferait pas justice. Néanmoins, vous savez tous que votre présence, dans des moments heureux et moins, est ce qui me permet d'avancer. Les amitiés que j'ai forgées font partie des choses qui m'ont le mieux réussi dans cette vie. Cette formule, maladroite mais sincère (ou l'inverse), est ce que j'ai trouvé de mieux pour exprimer mes sentiments.

Je remercie mes professeurs de prépa Jean-Christophe Feauveau et Éric Desmeules. Vous étiez les premiers à me montrer comment se construit un raisonnement scientifique, l'importance de l'intuition géométrique et, probablement, étiez les meilleurs professeurs que je n'ai jamais eus. Si aujourd'hui j'entame une carrière de chercheur c'est sans doute grâce à vous.

Je remercie Клойчик qui est toute tchik tchik, qui m'accepte comme je suis et que j'aime beaucoup beaucoup.

Last but not least, je remercie ma famille pour leur amour, leur présence et leur soutien inconditionnel à tout moment.

Quelques problèmes en optimisation non convexe et stochastique

Résumé: Le sujet de cette thèse est l'analyse de divers algorithmes stochastiques visant à résoudre un problème d'optimisation non convexe.
Nous commençons par un problème d'optimisation lisse en analysant une famille d'algorithmes adaptatifs avec moments qui comprend entre autres ADAM et la descente de gradient accélérée de Nesterov. La convergence et la fluctuation des itérés sont établies. Un résultat général d'évitement des pièges pour les algorithmes stochastiques sous-tendus par une équation différentielle non autonome est présenté. Il est appliqué pour établir la non-convergence des itérés aux points-selles.
La suite du manuscrit est consacrée au cas où la fonction que l'on cherche à minimiser est non lisse. La plupart de nos résultats dans cette partie s'appliquent aux fonctions définissables dans une structure o-minimale. Tout d'abord, nous analysons la version à pas constants de la descente de sous-gradient stochastique (SGD) et montrons que ses itérés convergent en grande probabilité vers l'ensemble des points critiques. Deuxièmement, nous montrons que chaque point critique d'une fonction Lipschitz, définissable, générique se trouve sur une variété active, satisfaisant une condition de Verdier et d'angle et est soit un minimum local, un point selle actif ou un point critique fortement répulsif. Nous montrons, sous des conditions légères sur les perturbations, que le SGD évite les deux derniers types de points. Une amélioration de la formule de projection pour les fonctions définissables, donnant une condition de type Lipschitz sur ses sous-gradients de Clarke, est présentée. Enfin, nous établissons un phénomène d'oscillation des itérés du SGD et de ses extensions proximales.
Mots clés: optimisation stochastique, évitement des pièges, optimisation non lisse, semi-algébrique, o-minimalité, stratifications, descente de sous-gradient stochastique, ADAM, algorithmes adaptatifs avec moments

Some Problems in Nonconvex Stochastic Optimization

Abstract: The subject of this thesis is the analysis of several stochastic algorithms in a nonconvex setting. The aim is to prove and characterize their convergence.
First, we study a smooth optimization problem, analyzing a family of adaptive algorithms with momentum which includes the widely used ADAM and Nesterov's accelerated gradient descent. Convergence and fluctuation of the iterates are established. A general avoidance of traps result for stochastic algorithms underlined by a nonautonomous differential equation is presented and applied to establish the nonconvergence of the iterates to saddle points.
The rest of the manuscript is devoted to the case where the function that we seek to minimize is nonsmooth. Most of our results in this part apply to functions definable in an o-minimal structure. Firstly, we analyze the constant step version of the stochastic subgradient descent (SGD) and show that the iterates converge with high probability to the set of critical points. Secondly, we show that every critical point of a generic, definable, locally Lipschitz continuous function is lying on an active manifold, satisfying a Verdier and an angle condition and is either a local minimum, an active strict saddle or a sharply repulsive critical point. We show that, under mild conditions on the perturbation sequence, the SGD escapes active strict saddles and sharply repulsive critical points. An improvement of the projection formula for definable functions, giving a Lipschitz-like condition on its Clarke subgradients is presented and is of independent interest. Finally, we establish an oscillation phenomena of the iterates of the SGD and its proximal extensions.
Keywords: stochastic approximation, avoidance of traps, nonsmooth optimization, semialgebraic, o-minimality, stratifications, stochastic subgradient descent, ADAM, Nesterov's accelerated gradient descent, adaptive algorithms with momentum

Résumé substantiel en français

L'objet de cette thèse est l'étude de divers algorithmes stochastiques visant à résoudre des problèmes d'optimisation non convexe. L'objectif dans chacun des cas est de démontrer et de caractériser la convergence de l'algorithme vers l'ensemble des points critiques de la fonction à minimiser.

Le chapitre 3, le seul à aborder un problème d'optimisation lisse, analyse une famille d'algorithmes adaptatifs et à moment qui comprend entre autres ADAM et la descente de gradient accélérée de Nesterov. En appliquant la méthode de l'ODE, qui consiste à voir ces algorithmes comme une discrétisation d'Euler d'une équation différentielle (ED), nous montrons la convergence des itérés envers les points critiques de la fonction à minimiser. La difficulté principale de l'analyse est que l'équation différentielle mentionnée est non autonome. Nous établissons dans certains cas un phénomène de fluctuation des itérés sous forme d'un théorème central limite. Enfin nous abordons la question d'évitement des pièges. Cette question est importante car l'ensemble des points critiques d'une fonction est, dans le cas non-convexe, généralement strictement plus large que l'ensemble des minimiseurs (locals) de la fonction. Cette question avait été abordée auparavant pour des algorithmes sous-tendus par une ED autonome, l'approche étant basée sur l'application du théorème de la variété invariante de Poincaré. En utilisant la version non autonome de ce théorème nous établissons un résultat d'évitement de piège général pour tout algorithme discrétisation d'une ED non autonome. Enfin, nous appliquons ce résultat aux algorithmes étudiés pour montrer la non-convergence presque sûre des itérés envers les points selles.

Le reste du manuscrit se concentre sur le cas où la fonction à minimiser est non différentiable. La plupart de nos résultats dans ce cas s'appliquent aux fonctions semi-algébrique ou, plus généralement, aux fonctions définissables dans une structure o-minimale. Cette classe de fonction, popularisée en optimisation par les travaux de Bolte, Lewis, Daniilidis et Shiota, comprend la grande majorité des fonctions étudiées en optimisation, statistiques et traitement de signal.

Le chapitre 4 analyse la version à pas constant de l'algorithme de la descente de sous-gradient stochastique (SGD). A pas décroissants cet algorithme a déjà été étudié dans la littérature en supposant l'existence en chaque point d'un estimateur "oracle" tel que son espérance est égal au sous-gradient de Clarke de la fonction à minimiser. L'existence d'un tel oracle dans des cas pratiques étant rarement vérifiée, nous montrons sous des conditions légères, que pour presque tout point d'initialisation l'existence d'un tel oracle n'est pas nécessaire. Dans un second temps, nous montrons que quand le pas tend vers zéro, l'interpolation affine des itérés converge vers l'ensemble des solutions du flot de sous-gradients (au sens de la convergence uniforme sur les compacts). Enfin, en analysant le SGD à pas fixé comme une chaîne de Markov, nous montrons que quand le pas tend vers zéro, la mesure invariante de cette chaîne de Markov tend faiblement vers l'ensemble des mesures invariantes du flot de sous-gradient. Ce résultat nous permet de montrer que, quand le pas est petit, les itérés du SGD au pas constant convergent vers les

points critiques de la fonction en grande probabilité.

Dans le chapitre 5 nous établissons un résultat d'évitement par le SGD des points selles actifs. Ces points critiques sont d'une grande importance car, comme montré par les travaux de Davis et Drusvyatskiy, les seuls points critiques que possède une fonction définissable, faiblement convexe, générique, sont des minima locaux ou des points selles actifs. Par définition ces points selles actifs se trouvent sur une variété active telle que la fonction est différentiable sur cette variété et "change rapidement" en dehors de cette variété. Afin d'étudier le SGD au voisinage de ces points nous introduisons deux conditions supplémentaires sur la variété active: la condition de Verdier et la condition de l'angle. La première permet d'avoir une condition de type Lipschitz entre le "gradient riemannien" de la fonction sur la variété et ses sous-gradients de Clarke alors que la condition de l'angle permet de montrer que le SGD converge rapidement vers la variété active. A l'aide de ces deux conditions, sous des conditions d'isotropie sur les perturbations similaires à celles qui sont nécessaires dans le cas lisse, nous montrons que le SGD évite un point selle actif avec probabilité un. De manière indépendante nous établissons une version renforcée de la formule de projection de Bolte et al. en donnant une condition de type Lipschitz sur les sous-gradients d'une fonction définissable et Lipschitz. Nous pensons que ce type de résultat peut être important pour l'étude des problèmes d'optimisation non lisse et définissable dans une structure o-minimale. En particulier, ce résultat nous permet de démontrer la généricité de nos deux conditions: les points selles actifs d'une fonction définissable, faiblement convexe, générique se trouvent sur une variété active vérifiant les conditions de Verdier et de l'angle. Ainsi, une interprétation possible des résultats de ce chapitre est que le SGD sur une fonction générique, définissable et faiblement convexe converge vers un minimum local.

Naturellement, au vu des résultats énoncés ci-dessus on voudrait savoir quels sont les points critiques d'une fonction définissable, générique sans l'hypothèse de faible convexité. Nous montrons dans le chapitre 6 l'émergence dans la classe des fonctions définissables et localement Lipschitz d'un troisième type de point: un point critique fortement répulsif. Un tel point se trouve sur une variété active telle qu'au voisinage de cette variété il existe une région répulsive ou les sous-gradients de la fonction sont dirigés vers la variété. Le premier résultat du chapitre 6 est que tous les points critiques d'une fonction définissable, localement Lipschitz, générique se trouvent sur une variété active et sont soit des minima locaux, des points selles actifs ou des points critiques fortement répulsifs. De plus, les variétés actives correspondantes vérifient toujours les conditions de Verdier et de l'angle introduites précédemment. La question d'évitement d'un point selle actif étant abordée dans le chapitre précédent, nous montrons que, sous une condition de densité sur la loi des perturbations, les points critiques fortement répulsifs sont évités par le SGD avec probabilité un. Ainsi, une interprétation possible des résultats de ce chapitre est que le SGD sur une fonction générique, définissable et localement Lipschitz converge vers un minimum local.

Le chapitre 7 donne une caractérisation de la convergence du SGD et de ses versions proximales vers l'ensemble des points critiques. Alors que ces algorithmes

peuvent avoir plusieurs points d'accumulation, nous montrons que le temps mis par les itérés de passer d'un voisinage d'un de ces points vers un autre tend vers l'infini. De plus, un phénomène d'oscillations des itérés est établi. Ce type de résultat pour la descente de gradient déterministe avait été établi auparavant par Bolte, Pauwels et Ríos-Zertuche en utilisant la théorie des mesures fermées. Dans le chapitre 7 nous établissons nos résultats sur la base de la théorie sur l'approximation stochastique et les inclusions différentielles de Benaïm, Hofbauer et Sorin ce qui permet de traiter les cas déterministe, stochastique et proximal avec une approche unifiée.

Contents

Chapter 1

Introduction

Various problems that arise in machine learning, signal processing and high dimensional statistics can be formulated as an optimization problem consisting into finding a minimum of a real valued function $F : \mathbb{R}^d \to \mathbb{R}$.

In practical settings, this function might be unknown or the computation of its different characteristics (such as the gradient or the Hessian) might be expensive. In this case, stochastic approximation algorithms are particularly interesting since at each iteration they require only an estimator of F. When the function of interest is convex such algorithms can be studied through the tools of convex analysis (see e.g. [Bottou *et al.* 2018]). However, in many applications the convexity assumption fails. In such a case the standard way to analyze the convergence of a stochastic approximation algorithm is to view it as an Euler-like discretization of its continuous counterpart: an ordinary differential equation (ODE).

If the corresponding ODE is well behaved then one will usually be able to prove the convergence of the iterates to the set $\mathcal{Z} = \{x \in \mathbb{R}^d : \nabla F(x) = 0\}$ of critical points of F (see [Benaïm 1999, Kushner & Yin 2003, Borkar 2008]). This, however, gives only a partial answer to the question of the convergence to the set of (local) minimizers of F. Indeed, without the convexity assumption, $\mathcal{Z}$ is usually strictly larger than the latter and contains all kinds of spurious points such as local maxima or saddle points. The nonconvergence of stochastic approximation algorithms to such points was analyzed in the literature under the name of “avoidance of traps". (see e.g. [Brandière & Duflo 1996, Pemantle 1990, Benaïm 1999]).

The aim of this thesis is to analyze the convergence of stochastic approximation algorithms to the set of critical points, when the continuous counterpart is no longer a simple ODE. We will give various characterizations of this convergence and establish, among other things, several avoidance of traps results. With the exception of Chapter 3, which analyzes a family of algorithms driven by a nonautonomous differential equation, we will focus on the case where the function that we seek to minimize is nonsmooth. The latter being especially important for various applications, the most notable one being the training of a neural networks with ReLU activation functions.

To better understand our approach we start by an illustrative example - the stochastic gradient descent (SGD). The SGD, an archetype of stochastic approximations algorithms, dates back to Robbins and Monro [Robbins & Monro 1951] and is written as follows:

$$x_{n+1} = x_n - \gamma_n \nabla F(x_n) + \gamma_n \eta_{n+1} \,, \tag{1.1}$$

where (γ_n) is a sequence of positive real numbers decreasing to zero and (η_{n+1}) is a sequence of random perturbations (usually with zero mean) that modelizes our partial knowledge of F. One can view the SGD as an Euler-like discretization of the gradient flow:

$$\dot{\mathsf{x}}(t) = -\nabla F(\mathsf{x}(t))\,. \tag{1.2}$$

The so-called ODE method [Benaïm 1999, Borkar 2008, Kushner & Yin 2003] allows to rigorously compare the path taken by the iterates of Equation (1.1) to the solutions to the differential equation (1.2) and to establish that the iterates converge to the set $\mathcal{Z} = \{x \in \mathbb{R}^d : 0 \in \nabla F(x)\}$ of critical points of F.

In the case of the SGD, the question of the avoidance of saddle points was first addressed by [Brandière & Duflo 1996] and [Pemantle 1990]. Their technique of proof is build upon the Poincaré invariant manifold theorem, which states that the set of points from which ODE (1.2) converges to a saddle point is a manifold of a dimension strictly smaller than d. The idea of [Brandière & Duflo 1996, Pemantle 1990] is then to show that, under an isotropic condition on the sequence (η_n), the iterates of the SGD will be driven away of this invariant manifold.

Throughout this thesis we will study different generalizations of the ODE method and the ideas of [Brandière & Duflo 1996, Pemantle 1990]. In particular, with the exception of Chapter 3, which analyzes a family of algorithms driven by a nonautonomous differential equation, we will focus on the case where the function that we seek to minimize is nonsmooth.

If F is merely locally Lipschitz continuous, then a natural generalization of the SGD is the stochastic subgradient descent, which reads as follows:

$$x_{n+1} \in x_n - \gamma_n \partial F(x_n) + \gamma_n \eta_{n+1}\,, \tag{1.3}$$

where $\partial F(x_n)$ is the set of Clarke subgradients of F at x_n, a notion that generalizes the one of the gradient. Algorithm (1.3) being a generalization of (1.1) we will still refer to it as the SGD. The set of critical points for the Clarke subgradient is now $\mathcal{Z} = \{x \in \mathbb{R}^d : 0 \in \partial F(x)\}$, which still contains the (local) minima of F. Following the work of [Benaïm *et al.* 2005], the continuous counterpart of this algorithm is no longer an ODE but a differential inclusion (DI):

$$\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t))\,. \tag{1.4}$$

To obtain (and characterize) the convergence of Equation (1.3) to $\mathcal{Z}$ we first need to restrict the class of functions that we analyze. Indeed, in full generality the Clarke subgradient might not even be the right operator to consider. For instance, Rockafellar in [Rockafellar 1981] constructs a Lipschitz function $F : \mathbb{R}^d \to \mathbb{R}$ such that for every $x \in \mathbb{R}^d$, we have $\partial F(x) = [-1, 1]^d$. Results of [Borwein & Wang 1998] show that this example is actually typical, e.g. almost every continuous function $F : [0, 1] \to \mathbb{R}$ has $\partial F(x) = [0, 1]$ for all $x \in [0, 1]$ [1]. This implies that, generally, ∂F gives us no information about the behavior of F. In fact, [Daniilidis & Drusvyatskiy 2019] shows

[1] Here almost every refers to the fact that this set is open and dense in the Baire's topology. A topology-independent result of this type was established in [Daniilidis & Flores 2019].

that pathological dynamics can be exhibited such that (x_n) is not even converging to $\mathcal{Z}$.

While these counterexamples may seem, at first sight, discouraging, in practical settings such pathological behavior is rare. Indeed, the vast majority of functions encountered in optimization are functions definable in an o-minimal structure, a notion popularized in the optimization literature by the work of [Bolte *et al.* 2007]. The family of definable functions is broad: every semialgebraic function is definable, the exponential and the logarithm are definable. Moreover, the notion of definableness is stable by many of the elementary operations such as composition, sum, multiplication and taking the inverse. While definable functions may be nonsmooth, the nonsmoothness here appear in a very structured manner. Example given, in [Bolte *et al.* 2007] the authors have established the so-called projection formula, which gives a description of the Clarke subgradients of a definable function. More precisely, the authors of [Bolte *et al.* 2007] have shown that, given a definable function F and p an integer, there exists (X_i) a finite partition of the domain of F into C^p manifolds such that F is C^p smooth on X_i and, moreover, if $y \in X_i$, then we have:

$$P_{T_y X_i}(\partial F(y)) = \{\nabla_{X_i} F(y)\}, \tag{1.5}$$

where $\nabla_{X_i} F(y)$ is the Riemannian gradient of F restricted to X_i at y and $P_{T_y X_i}$ denotes the orthogonal projection onto the the tangent space at y of X_i. Equation (1.5) is the starting point in the proof of various properties of definable functions such as the nonsmooth Kurdyka-Łojasiewicz inequality or the path differentiability [Bolte & Pauwels 2019].

In recent years, this implicit smooth structure has allowed a thorough analysis of algorithms operating on definable functions [Attouch *et al.* 2011, Bolte *et al.* 2009, Davis & Drusvyatskiy 2021]. In particular, the work of [Davis *et al.* 2020] shows that under mild conditions on the sequence (η_n), the iterates of Equation (1.3) converge to $\mathcal{Z}$. In this thesis we will give different characterizations of this convergence, with a particular focus in Chapters 5 and 6 on the question of the avoidance of traps in a nonsmooth setting.

With Chapter 2 being dedicated to mathematical tools that will be used throughout this thesis, we finish the present chapter by a detailed descriptions of the main obtained results.

1.1 Stochastic algorithms with momentum

In Chapter 3, which is based on the publication {5}, we study a class of stochastic algorithms which admits as a continuous counterpart the following ODE, introduced in [Belotto da Silva & Gazeau 2018]. Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function to minimize and let $S : \mathbb{R}^d \to \mathbb{R}^d_+$ be a continuous function. Let $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q} : (0, +\infty) \to \mathbb{R}_+$ be continuous functions and let $\varepsilon > 0$. Starting from a point $(\mathsf{v}(0), \mathsf{m}(0), \mathsf{x}(0)) \in$

$\mathbb{R}_+^d \times \mathbb{R}^d \times \mathbb{R}^d$, the differential equation is written as follows:

$$\begin{cases} \dot{\mathsf{v}}(t) &= \mathsf{p}(t)S(\mathsf{x}(t)) - \mathsf{q}(t)\mathsf{v}(t) \\ \dot{\mathsf{m}}(t) &= \mathsf{h}(t)\nabla F(\mathsf{x}(t)) - \mathsf{r}(t)\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\mathsf{m}(t)/\sqrt{\mathsf{v}(t)+\varepsilon} \end{cases}, \tag{1.6}$$

where for two vectors $x, y \in \mathbb{R}^d$, x/y denotes the vector $(x_1/y_1, \ldots, x_d/y_d) \in \mathbb{R}^d$. The main challenge in the analysis of the algorithms that are underlined by this ODE is the fact that ODE (1.6) is nonautonomous. ODE (1.6) generalizes various differential equations encountered in stochastic approximation. For instance, its particular case is

$$\begin{aligned} \dot{\mathsf{m}}(t) &= \nabla F(\mathsf{x}(t)) - \mathsf{r}(t)\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\mathsf{m}(t) \end{aligned}$$

which can be rewritten as

$$\ddot{\mathsf{x}}(t) + \mathsf{r}(t)\dot{\mathsf{x}}(t) + \nabla F(\mathsf{x}(t)) = 0.$$

If we choose $\mathsf{r}(t) \equiv \alpha > 0$, then we obtain the well-known Heavy-Ball with friction algorithm [Attouch *et al.* 2000, Gadat *et al.* 2018]. Choosing $\mathsf{r}(t) = \alpha/t$, with $\alpha > 0$, gives us the Nesterov's accelerated gradient algorithm which was studied from this ODE in [Su *et al.* 2016a].

Going back to ODE (1.6) and choosing this time $\mathsf{h}(t) = \mathsf{r}(t) = a(t, \lambda, \alpha_1)$, $\mathsf{p}(t) = \mathsf{q}(t) = a(t, \lambda, \alpha_2)$ for $a(t, \lambda, \alpha) = \lambda^{-1}(1 - \exp(-\lambda\alpha))/(1 - \exp(-\alpha t))$, $\lambda, \alpha_1, \alpha_2 > 0$ and $S = \nabla F^{\odot 2}$, we recover the widely used ADAM algorithm [Kingma & Ba 2015] (see also [Belotto da Silva & Gazeau 2020, Sections 2.4-4.2] and [Barakat & Bianchi 2021] for the stochastic version of this algorithm).

In this chapter we establish the convergence of the stochastic algorithms driven by ODE (1.6) to the set of critical points of F. In this level of generality, the presented results are new. Convergence rates in the form of a central limit theorem are given. Last but not least, an avoidance of traps result is established. This result extends previous works of [Gadat *et al.* 2018] obtained in the context of SHB. This result not only allows to study a broader class of algorithms but also significantly weakens the assumptions. In particular, [Gadat *et al.* 2018] uses a sub-Gaussian assumption on the noise and a rather stringent assumption on the stepsizes. The main difficulty in the approach of [Gadat *et al.* 2018] lies in the use of the classical autonomous version of Poincaré's invariant manifold theorem. The key ingredient of our proof is a general avoidance of traps result, adapted to nonautonomous settings, which we believe to be of independent interest. It extends usual avoidance of traps results to a nonautonomous setting, by making use of a nonautonomous version of Poincaré's theorem [Dalec′kiĭ & Kreĭn 1974, Kloeden & Rasmussen 2011].

Contributions.

- First, we analyze ODE (1.6) by showing the existence and the uniqueness of its solutions. Convergence of these solutions to the set of critical points of F is established. In particular, no convexity assumption is made and, to the best of our knowledge, the convergence statement for the ODE that underlines the Nesterov's accelerated gradient descent is new.

- Second, we analyze a class of stochastic approximation algorithms that are Euler-like discretizations of this ODE. Examples of these are ADAM, AdaGrad, Heavy-Ball and Nesterov's accelerated gradient descent. Boundedness and convergence to the set of critical points of F is established. Under additional assumptions, convergence rates in the form of a central limit theorem are given. These results extend the works of [Gadat *et al.* 2018, Barakat & Bianchi 2021] to a more general setting. In particular, we highlight the almost sure convergence result for the (stochastic) Nesterov's accelerated gradient descent in a nonconvex setting, which is, to the best of our knowledge, new.

- Finally, a general avoidance of traps result is established for algorithms underlined by a nonautonomous ODE. An application of this result to the algorithms that we analyze is given by establishing that, under assumptions on the perturbation sequence similar to [Brandière & Duflo 1996], the iterates avoid saddle points with probability one.

1.2 Convergence of the stochastic subgradient descent with a constant stepsize

Chapter 4, based on the publication {4}, analyzes the constant step version of the SGD (1.3). While from a theoretical point of view, the vanishing step size is convenient to show the convergence of the algorithm to $\mathcal{Z}$, in practical applications such as the training of a neural net, a vanishing step size is rarely used because of slow convergence issues. In most computational frameworks, a possibly small but nevertheless constant step size is used by default. The price to pay is that the iterates are no longer expected to converge almost surely to the set $\mathcal{Z}$ but to fluctuate in the vicinity of $\mathcal{Z}$ as n is large. Therefore, in this chapter we aim to establish a result of the type

$$\forall \varepsilon > 0, \quad \limsup_{n\to\infty} \mathbb{P}(\operatorname{dist}(x_n, \mathcal{Z}) > \varepsilon) \underset{\gamma \downarrow 0}{\longrightarrow} 0\,. \tag{1.7}$$

Although this result is weaker than in the vanishing step case, constant step stochastic algorithms can reach a neighborhood of $\mathcal{Z}$ faster than their decreasing step analogues, which is an important advantage in the applications where the accuracy of the estimates is not essential. Moreover, in practice they are able to cope with non stationary or slowly changing environments which are frequently encountered in signal processing, and possibly track a changing set of solutions [Benveniste *et al.* 1990, Kushner & Yin 2003].

Before proving this convergence result we address the question of the pertinence of the algorithm (1.3) from a practical standpoint. Indeed, in stochastic approximation the designer has usually no access to the function F but rather to a sequence of i.i.d random variables (ξ_n), following a law μ, and to a function $f(x,\xi)$ such that $F(x) = \mathbb{E}_{\xi\sim\mu}[f(x,\xi)]$. In this case, the natural algorithm that comes to mind is:

$$x_{n+1} = x_n - \gamma v_{n+1}\,, \tag{1.8}$$

where v_{n+1} is a selection of the Clarke subgradient $\partial_x f(x,\xi)$ (taken relatively to the first variable) at the point (x_n, ξ_{n+1}). Denoting $\mathscr{F}_n$ the sigma algebra generated by $(x_0,\dots,x_n)$, we can rewrite this algorithm as:

$$x_{n+1} = x_n - \gamma\mathbb{E}[v_{n+1}|\mathscr{F}_n] + \gamma\eta_{n+1}\,, \tag{1.9}$$

where $\eta_{n+1} = -v_{n+1} + \mathbb{E}[v_{n+1}|\mathscr{F}_n]$ is a martingale increment. The issue that arise in this case is that the continuous counterpart of this equation is now

$$\dot{\mathsf{x}}(t) \in -\mathbb{E}_{\xi\sim\mu}[\partial_x f(\mathsf{x}(t),\xi)]\,. \tag{1.10}$$

This differential inclusion is not necessarily an instance of the DI (1.4), because we do not generally have that $\mathbb{E}[v_{n+1}|\mathscr{F}_n] \in \partial F(x_n)$. Indeed, the interchange $\mathbb{E} \leftrightarrow \partial$ holds for convex or smooth functions but fails in general. In [Majewski *et al.* 2018] the authors restrict their analysis to Clarke regular functions [Clarke *et al.* 1998, §2.4], for which the interchange of the expectation and the subdifferentiation applies. However, this assumption can be restrictive, since a function as simple as $-|x|$ is not regular at the critical point zero.

In Chapter 4 we consider a slightly more general version of algorithm (1.8), which includes the case where v_{n+1} is a selection of a so-called conservative field. This notion, introduced in [Bolte & Pauwels 2019], modelizes the output of the celebrated backpropagation algorithm used in numerical libraries such as PyTorch or Tensorflow [Paszke *et al.* 2017]. A similar issue arise in this case since the interchange between the expectation and a conservative field might not hold.

Our first result is that, under some mild conditions on the functions F and $f(\cdot,\xi)$ (for instance if they are definable), for almost every initialization point and for every $n \in \mathbb{N}$, x_n almost never hits a nondifferentiable point of $f(\cdot,\xi_{n+1})$. As a consequence, algorithm (1.3) can be rewritten as:

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

and its continuous counterpart is indeed the DI (1.4). Using this result, we show that the continuous process obtained by a piecewise affine interpolation of (x_n) is a *weak asymptotic pseudotrajectory* of the DI (1.4). In other words, the interpolated process converges in probability to the set of solutions to the DI, as $\gamma \to 0$, for the metric of uniform convergence on compact intervals.

The proof technique to establish the convergence (1.7) is then rather standard and consists to view (1.9) as a Markov process. For each $\gamma > 0$, under a drift

assumption on its kernel, this Markov process admits an invariant distribution μ_γ. Every accumulation point of (μ_γ), in the sense of the weak convergence and when $\gamma \to 0$, is then shown to be supported on $\mathcal{Z}$, which in turn implies (1.7).

Contributions.

- We analyze the SGD algorithm with a constant step size in the non-smooth, non-convex setting. Under mild conditions, we prove that when the initialization x_0 is chosen randomly x_n almost never hits a non-differentiable point of $f(\cdot, \xi_{n+1})$ and
$$\frac{x_{n+1} - x_n}{\gamma} = -\nabla F(x_n) + \eta_{n+1}\,,$$
where (η_n) is a martingale difference sequence, and $\nabla F(x_n)$ is the true gradient of F at x_n. This argument allows to bypass the oracle assumption of [Majewski *et al.* 2018, Davis *et al.* 2020].

- We establish that the continuous process obtained by a piecewise affine interpolation of (x_n) is a *weak asymptotic pseudotrajectory* of the DI (1.4). In other words, the interpolated process converges in probability to the set of solutions to the DI, as $\gamma \to 0$, for the metric of uniform convergence on compact intervals.

- We establish the long run convergence of the iterates x_n to the set $\mathcal{Z}$ of Clarke critical points of F, in the sense of Equation (1.7). This result holds under two main assumptions. First, it is assumed that F admits a chain rule, which is satisfied for instance if F is a definable function. Second, we assume a standard drift condition on the Markov chain (1.9). Finally, we provide verifiable conditions of the functions $f(\cdot, s)$ under which the drift condition holds.

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

1.3 SGD escapes active strict saddles

In Chapter 5, which is based on the publication {2}, we address the question of the avoidance of traps in a nonsmooth setting. The traps that are considered here are the active strict saddles, a notion that was recently introduced in [Davis & Drusvyatskiy 2021]. Formally, a Clarke critical point is an active strict saddle if it lies on an active manifold M such that the Riemannian Hessian of F on M has at least one negative eigenvalue. An active manifold in this setting is a manifold M such that *i)* F varies sharply outside of M, *ii)* F_M, the restriction of F to M, is smooth. For instance, the function $(y, z) \mapsto |z| - y^2$ admits the point $(0, 0)$ as an active strict saddle (see Figure 1.1).

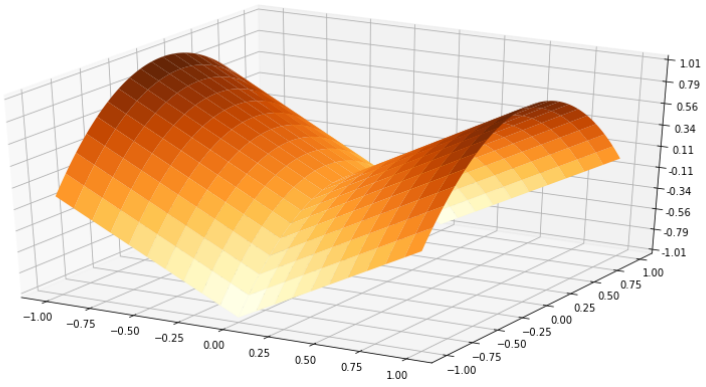

Figure 1.1: The point $(0,0)$ is an active strict saddle of $F(y,z) = |z| - y^2$. The corresponding active manifold is $M = \mathbb{R} \times \{(0,0)\}$.

The importance of this notion comes from the fact, proved in [Drusvyatskiy *et al.* 2016, Davis & Drusvyatskiy 2021], that, given a weakly convex, definable function $F : \mathbb{R}^d \to \mathbb{R}$, for almost every $u \in \mathbb{R}^d$, each of the Clarke critical points of the linearly perturbated function $F_u(x) \triangleq F(x) - \langle u, x \rangle$ is either a local minimum or an active strict saddle. In that sense, active strict saddles are generic in the class of weakly convex, definable functions.

In [Davis & Drusvyatskiy 2021] the authors have proven that, upon a random initialization, proximal methods escape active strict saddles with probability one. Such a result is possible due to the fact that proximal methods implicitly run a gradient descent on a smoothened version of F - the Moreau envelope. On the contrary, in Chapter 5 we analyze algorithm (1.3) which is inherently nonsmoooth. The aim is to establish that the SGD (with decreasing stepsizes) escapes active strict saddles with probability one.

The intuition behind our approach could be grasped by looking at the stochastic subgradient descent on the function from Figure 1.1. In this case, it is natural to write down the iterates (x_n) as (y_n, z_n) and to notice that (y_n) follows an SGD dynamic on a smoooth function $y \mapsto -y^2$. Thus, applying the results of [Brandière & Duflo 1996, Pemantle 1990] to the sequence (y_n), we obtain that, under similar assumptions on the perturbation sequence, $\mathbb{P}(y_n \to 0) = 0$. This implies that the stochastic subgradient descent avoids $(0,0)$ with probability one. On an independent note, observe that in this example the sequence (z_n) converges to zero in a very fast manner.

To formalize this type of behavior in a more general setting we have introduced two additional conditions on the active manifold. The first one, the Verdier condition, states that for x close to M:

$$\forall v \in \partial F(x), \quad v_M \approx \nabla_M F(P_M(x)) + O(\mathrm{dist}(x, M)), \tag{1.11}$$

where $P_M(x)$ is the projection of x onto M, $\nabla_M F$ is the "Riemannian gradient" of F_M and v_M is the projection of v along the tangent space of M (see Section 5.3.2

for a precise statement). A consequence of this condition is that, writing down $(y_n) = (P_M(x_n))$, a simple application of Taylor's formula gives us:

$$y_{n+1} \approx y_n - \gamma_n \nabla_M F(y_n) + \gamma_n \eta_{n+1}^M + \gamma_n O(\mathrm{dist}(x_n, M)) + O(\gamma_n^2)\,, \tag{1.12}$$

where η_{n+1}^M is the projection of η_{n+1} on the tangent space of M at y_n. That is to say, up to a residual error term, (y_n) follows an SGD dynamic on the (smooth) function F_M.

The purpose of the angle condition is to control this residual term. First, a following observation is made. Let x^* be a Clarke critical point of F lying on an active manifold M. Then, on the event $[x_n \to x^*]$, for n large enough, we have:

$$F(x_n) - F(P_M(x_n)) \gtrsim \|x_n - P_M(x_n)\|\,. \tag{1.13}$$

The angle condition then states that close to M we have:

$$F(x) - F(P_M(x)) \gtrsim \|x - P_M(x)\| \implies \langle v, x - P_M(x)\rangle \gtrsim \|x - P_M(x)\|\,, \quad \forall v \in \partial F(x)\,. \tag{1.14}$$

Combining (1.13) with (1.14), we obtain that, for n large enough, the angle between the set $\partial F(x_n)$ and the normal direction to M is lower bounded. This is used to show that $\mathrm{dist}(x_n, M)$ converges to zero in a very fast manner and thus allows to control the residual term in Equation (1.12).

The angle and the Verdier conditions provide a general way to analyze the stochastic subgradient descent in a neighborhood of an active manifold by decomposing the iterates (x_n) into a sum of two sequences: $(y_n) = (P_M(x_n))$ and $(z_n) = (x_n - y_n)$. The angle condition ensures the fact that $\|z_n\| = \mathrm{dist}(x_n, M) \to 0$ (and hence $x_n \to M$) fast enough. Combining this fact with the Verdier condition, this implies that (y_n), up to a residual term, follows an SGD dynamic on the smooth function F_M. In Chapter 5 we illustrate this proof technique by showing that, under conditions on (η_{n+1}) similar to the ones obtained in a smooth setting by [Brandière & Duflo 1996], the stochastic subgradient descent avoids active strict saddles with probability one.

An important contribution of this chapter is an improvement of the projection formula (1.5). One of the consequence of Equation (1.5) is that if we have a sequence (x_n, v_n), with $x_n \to y \in X_i$ and $v_n \in \partial F(x_n)$, then we always have $P_{T_y X_i}(v_n) \to \nabla_{X_i} F(y)$. In Theorem 5.2.1 of Chapter 5 we reinforce this result by showing that if F is locally Lipschitz continuous, then there exists (a perhaps finer) finite partition (X_i) such that for any $y \in X_i$, there is $C > 0$ such that for any $y' \in X_i$ and $x \in \mathbb{R}^d$ that are close enough to y, we have:

$$\forall v \in \partial F(x)\,, \quad \left\| P_{T_{y'} X_i}(v) - \nabla_{X_i} F(y') \right\| \leqslant C \left\| x - y' \right\|\,. \tag{1.15}$$

The Verdier condition (1.11) thus merely states that M is one of the element of this partition. The projection formula was initially proved in [Bolte *et al.* 2007] using the fact that the graph of a definable function admits a so-called Whitney-(a) stratification. Our proof of the reinforced projection formula is based on the

fact, well known literature on o-minimal theory (see [Loi 1998]), that the graph of a definable function admits a so-called Verdier stratification. Formula (1.15) gives us a Lipschitz-like condition on the subgradient operator ∂F. We believe that such a result is of independent interest and might be important in the analysis of nonsmooth algorithms operating in a definable setting.

One of the consequences of the reinforced projection formula that we prove is that the Verdier and the angle conditions are generic in the class of weakly convex functions. That is to say, given $F : \mathbb{R}^d \to \mathbb{R}$ a weakly convex and definable function, for almost every $u \in \mathbb{R}^d$, every critical point of F_u is either a local minimum or an active strict saddle, with the corresponding active manifold satisfying the Verdier and the angle conditions. Therefore, a possible interpretation of the results of this chapter is that the stochastic subgradient descent on a generic, weakly convex function converges to a local minimum.

Contributions.

- 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 [Bolte *et al.* 2007]. While such stratifications are well-known in the literature on o-minimal structures [Loi 1998], 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 to the case of definable, locally Lipschitz continuous functions by establishing a Lipschitz-like condition on the (Riemannian) gradients of two adjacent stratas.

- Secondly, we introduce two additional assumptions on an active manifold: the Verdier and the angle conditions. We prove that a generic active strict saddle of a definable and weakly convex function is lying on an active manifold satisfying both of these conditions.

- Finally, 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. We emphasize here that, while our genericity result holds under a weak convexity assumption, no weak convexity is assumed for our avoidance of traps result.

1.4 SGD on a generic definable function converges to a minimizer

In Chapter 5 we have established that, given $F : \mathbb{R}^d \to \mathbb{R}$ a weakly convex, definable function, for almost every $u \in \mathbb{R}^d$, the critical points of F_u are either local minima or active strict saddles lying on active manifolds satisfying the Verdier and the

angle conditions. Naturally, one might want to know what happens when the weak convexity assumption fails. The first part of Chapter 6, which is based on the publication {1}, addresses this question by classifying the generic Clarke critical points of locally Lipschitz continuous, definable functions. Specifically, given such a function $F : \mathbb{R}^d \to \mathbb{R}$ it analyzes the type of points that might appear in:

$$\{x \in \mathbb{R}^d : 0 \in \partial F_u(x)\}, \tag{$\mathcal{Z}_u$}$$

for a non Lebesgue-null set of $u \in \mathbb{R}^d$.

First, we must notice that such a simple classification in a weakly convex setting comes from the fact that, from a minimization perspective, the local behavior of F on an active manifold M dictates its shape outside of M. In particular, if F is weakly convex and x^* is a local minimum of F_M, then x^* is a local minimum of F (see [Drusvyatskiy & Lewis 2014]). Looking at an example as simple as $(y, z) \mapsto y^2 - |z|$ (see Figure 1.2), we see that this is no longer true when the weak convexity assumption fails.

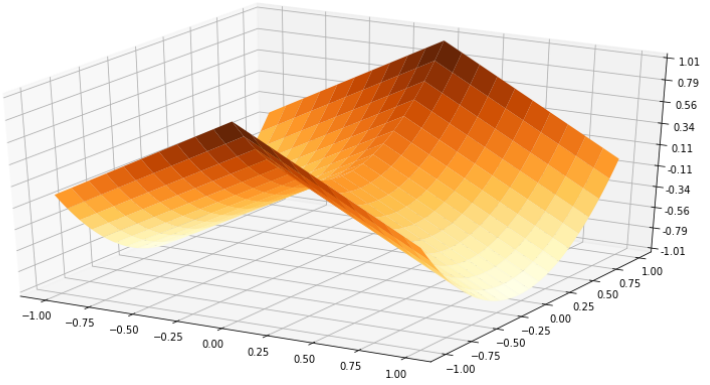

Figure 1.2: The point $(0, 0)$ is a sharply repulsive critical point of $F(y, z) = -|z| + y^2$. The corresponding active manifold is $M = \mathbb{R} \times \{(0, 0)\}$.

This motivates the introduction of a third type of points: a sharply repulsive critical point. We say that a Clarke critical point $x^* \in M$ is a sharply repulsive critical point if *i)* M is an active manifold (for F and x^*), *ii)* x^* is a local minimum of F_M, *iii)* $0 \in \partial_L F(x^*) \backslash \partial F(x^*)$, where $\partial_L F(x)$ denotes the limiting subgradient of F at x^*. Intuitively speaking, if $x^* \in M$ is a sharply repulsive critical point, then there is a large region neighboring M on which the subgradients of F are pointing towards M. A typical example of this situation is illustrated on Figure 1.2.

In Chapter 6 we show that such an example is generic. More precisely, in Theorem 6.2.5 we establish that for almost every $u \in \mathbb{R}^d$, every point in (1.4) is lying on an active manifold and is either a local minimum, an active strict saddle or a sharply repulsive critical point (for the function F_u). Moreover, the corresponding active manifolds satisfy the Verdier and the angle conditions. In this sense, a generic trap of a definable, locally Lipschitz continuous function is either an active strict saddle

or a sharply repulsive critical point. We must notice here that almost all of the points of Theorem 6.2.5 readily follow from the work of [Drusvyatskiy *et al.* 2016]. However, the question of the genericity of the angle condition seems to be delicate and our proof of this point is based on some deep results of o-minimal theory.

Since the question of the avoidance of active strict saddles was treated in Chapter 5, the rest of Chapter 6 is devoted to the question of the avoidance of sharply repulsive critical points. Our first result shows that if x^* is such a point, then, on the event $[x_n \to x^*]$, for n large enough, we have:

$$F(x_n) \geqslant F(x^*)\,.$$

While the proof of this inequality readily follows from the observations of Chapter 5, this result reveals to be interesting. Indeed, it implies that while the iterates (x_n) may in theory converge to x^* this happens only if the SGD fails to explore the repulsive region near x^*. In some sense, the algorithm perceive the function F as if x^* was indeed its local minimum.

In a second time, we show that a density-like assumption on (η_n) forces the SGD to visit the repulsive region near M and will imply the nonconvergence of the SGD to a sharply repulsive critical point. We must notice here the difference with the proof of Chapter 5 on the avoidance of active strict saddles. Indeed, if x^* is a sharply repulsive critical point, then asymptotically the sequence $(y_n) = (P_M(x_n))$ still follows an SGD dynamic on a smooth function F_M. However, since in this case x^* is a local minimum of F_M this is not sufficient to prove the nonconvergence of (x_n) to M. Therefore, in the setting of Chapter 6 the avoidance of traps result is established by using a density-like condition on (η_n).

The final Section 6.3.3 shows that, while such a density-like assumption on (η_n) might not hold, a way to ensure it in a standard stochastic approximation model is to add a small perturbation (e.g. a nondegenerate Gaussian) at each iteration of (6.1). This fact, combined with the results of Chapter 5 on the avoidance of active strict saddles, provides a practical way to avoid generic traps of definable functions, and, therefore, ensure the convergence of the SGD to a local minimum.

We must mention here that shortly after the publication {2} and just before the submission of {1} a concurrent work [Davis *et al.* 2021] has appeared. The latter, sharing a lot of similarities with Chapter 5, analyzes the SGD (and its proximal versions) in a neighborhood of an active manifold. An avoidance of active strict saddles result was obtained as well as (local) rates of convergence and asymptotic normality of the iterates were established. These results support our claim on the importance of the Verdier and the angle conditions. A major difference with our work is that their proximal aiming condition assume (close to the active manifold) the left hand side of formula (1.14). Such an assumption rules out functions with downward cusps such as $(y, z) \mapsto \pm y^2 - |z|$, which are treated in Chapters 5 and 6. As a consequence, the question of genericity in [Davis *et al.* 2021] is addressed only for the class of Clarke regular functions in which sharply repulsive critical points do not exist. In particular, we believe that convergence rates of a similar kind could

be obtained upon replacing the proximal aiming condition of [Davis *et al.* 2021] by ours angle condition.

Contributions.

- We introduce the concept of a sharply repulsive critical point. Given $F : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous function that is definable in an o-minimal structure, we show that for a full-measure set of $u \in \mathbb{R}^d$, each of the critical points of the linearly perturbated function F_u is lying on an active manifold satisfying a Verdier and an angle condition and is either a local minimum, an active strict saddle or a sharply repulsive critical point.

- We show that if x^* is a sharply repulsive critical point, then on the event $[x_n \to x^*]$, for n large enough, we have $F(x_n) \geqslant F(x^*)$. Furthermore, if the corresponding active manifold satisfies an angle condition and under a density-like assumption on the perturbation sequence (η_n) we show that the iterates of the SGD will avoid a sharply repulsive critical point with probability one. Finally, in a standard stochastic approximation model, we show that such an assumption can be ensured by adding at each iteration a small perturbation with a density. The latter, combined with the results of Chapter 5, gives a practical way to ensure the avoidance of the generic traps by the SGD.

1.5 Oscillations of the SGD and its proximal extensions

The purpose of Chapter 7, which is based on the publication {3}, is to give some characterizations of the convergence of the algorithm (1.3) and its proximal extensions. Given $F, g : \mathbb{R}^d \to \mathbb{R}$ two locally Lipschitz continuous functions and $\mathcal{X}$ a closed convex set, we are seeking to minimize $F + g$ over $\mathcal{X}$. A popular choice of algorithm in this case is the stochastic proximal subgradient descent (SPGD), which reads as follows:

$$x_{n+1} \in \operatorname{prox}_{g,\mathcal{X}}^{\gamma_n}(x_n - \gamma_n v_n + \gamma_n \eta_{n+1}) \,, \tag{1.16}$$

where $\operatorname{prox}_{g,\mathcal{X}}^{\gamma_n}$ is the proximal operator for the function g on $\mathcal{X}$ and v_n is in the set $\partial F(x_n)$.

From the work of [Davis *et al.* 2020] it is known that in this case the iterates (x_n) will converge to the set of composite critical points $\mathcal{Z} := \{x : 0 \in \partial F(x) + \partial g(x) + \mathcal{N}_{\mathcal{X}}(x)\}$, where $\mathcal{N}_{\mathcal{X}}(x)$ is the normal cone of $\mathcal{X}$ at x. However, the iterates (x_n) might not converge to a unique point. Indeed, in [Rios-Zertuche 2020, Section 2] Ríos-Zertuche considers the deterministic subgradient descent (that is to say $g = 0$, $\eta_n \equiv 0$ and $\mathcal{X} = \mathbb{R}^d$) and constructs F, which verifies main assumptions of nonsmooth optimization (such as Whitney stratifiability of its graph or Kurdyka-Łojasiewicz inequality) but the limit set of (x_n) is equal to $\mathcal{Z} = \{x : \|x\| = 1\}$. This encourages a more precise study of Equation (1.16).

In Chapter 7 we establish two additional results on the convergence of Equation (1.16). First, we show that if $x, y \in \mathcal{Z}$ are two distinct accumulation points, then

the time that the iterates spend to get from a neighborhood of x to a neighborhood of y goes to infinity. Secondly, we rewrite Equation (1.16) as:

$$x_{n+1} = x_n - \gamma_n(v_n + v_n^g + v_n^{\mathcal{X}}) + \gamma_n \eta_{n+1}\,, \tag{1.17}$$

where $v_n^g \in \partial g(x_{n+1})$ and $v_n^{\mathcal{X}} \in \mathcal{N}_{\mathcal{X}}(x_{n+1})$ and establish an oscillation-type phenomena. In a first approximation our results imply that, given $\delta > 0$ and any accumulation point x, we have:

$$\frac{\sum_{i=1}^n \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}} + \eta_{i+1}) \mathbb{1}_{B(x,\delta)}(x_i)}{\sum_{i=1}^n \gamma_i \mathbb{1}_{B(x,\delta)}(x_i)} \xrightarrow[n\to+\infty]{} 0\,. \tag{1.18}$$

Intuitively speaking, this type of behavior shows that even if $x_n - x_0 = \sum_{i=1}^n \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}} + \eta_{i+1})$ might not converge, the drift coming from the sugbradients $v_n, v_n^g, v_n^{\mathcal{X}}$ and the perturbation sequence η_{n+1}, on average, compensate itself. This suggests that the subgradient descent and its stochastic and proximal versions oscillates around its accumulation set, with the center of these oscillations moving in a vanishing speed.

This type of results was obtained by [Bolte *et al.* 2020b] for the deterministic gradient descent using the theory of closed measures. A nice feature of this chapter is that all of our results are proved using the theory of [Benaïm *et al.* 2005]. We feel that this approach gives a simpler proof of the convergence and the oscillation phenomena of the subgradient descent and its stochastic/proximal extensions.

Contributions.

- We show that the time spent by the SPGD to move from one accumulation point to another goes to infinity and establish an oscillation-type behavior of the drift. These two results extend [Bolte *et al.* 2020b, Theorem 7.] to a stochastic and a proximal setting. Our technique of proof doesn't rely on the theory of closed measures used in [Bolte *et al.* 2020b] but is build upon the classical work of Benaïm, Hofbauer and Sorin [Benaïm *et al.* 2005]. We feel that this approach gives a simpler proof and allows us to treat the deterministic, the stochastic and the proximal cases in a unified manner.

1.6 Publications

Preprints

{1} Sh. Schechtman. Stochastic subgradient descent on a generic definable function converges to a minimizer.

{2} P. Bianchi, W. Hachem, and Sh. Schechtman. Stochastic subgradient descent escapes active strict saddles.

{3} Sh. Schechtman. Stochastic proximal subgradient descent oscillates in the vicinity of its accumulation set.

{4} P. Bianchi, W. Hachem, and Sh. Schechtman, Convergence of constant step stochastic gradient descent for nonsmooth nonconvex functions.

Published

{5} A. Barakat, P. Bianchi, W. Hachem, and Sh. Schechtman. Stochastic optimization with momentum: convergence, fluctuations, and traps avoidance. *Electronic Journal of Statistics*, 2021, Vol. 15, No. 2, 3892-3947.

Chapter 2

Mathematical preliminaries

We gather in this chapter some standard results of variational analysis, differential geometry and o-minimal theory that will be used throughout this thesis.

2.1 Subgradients

Most of the results of this section can be found in classical monographs on variational analysis such as [Rockafellar & Wets 1998, Clarke *et al.* 1998].

When the function of interest is nonsmooth various notions of subgradients generalize the one of the gradient.

Definition 2.1.1 (Frechet subgradient)**.** *Consider $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous function and $x \in \mathbb{R}^d$. The set $\partial_F f(x) \subset \mathbb{R}^d$ of* Frechet subgradients of f at x *is the set of $v \in \mathbb{R}^d$ for which:*

$$\liminf_{x' \to x} \frac{f(x') - f(x) - \langle v, x' - x \rangle}{\|x' - x\|} \geqslant 0\,.$$

The set $\partial_F f(x)$ can be empty (e.g. for the function $f(x) = -\|x\|$ at 0). This motivates the following definitions.

Definition 2.1.2. *Consider $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous function and $x \in \mathbb{R}^d$. The set $\partial_L f(x)$ of* limiting subgradients of f at x *is the set of $v \in \mathbb{R}^d$ for which there is a sequence $(x_n, v_n) \to (x, v)$, with $v_n \in \partial_F f(x_n)$ for all $n \in \mathbb{N}$.*

Definition 2.1.3. *Consider $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous function and $x \in \mathbb{R}^d$. The set $\partial f(x)$ of* Clarke subgradients of f at x *is defined as follows*

$$\partial f(x) = \mathrm{conv}\{\partial_L f(x)\}\,,$$

where conv *denotes the convex hull.*

If f is C^1 around x, then $\partial f(x) = \partial_L f(x) = \partial_F f(x) = \{\nabla f(x)\}$.

Definition 2.1.4 (Clarke critical points)**.** *We say that $x \in \mathbb{R}^d$ is a Clarke critical point of f if $0 \in \partial f(x)$.*

Similarly to the smooth setting, the set of Clarke critical points of f contains local maxima and minima of f.

Since f is locally Lipschitz continuous, by Rademacher's theorem, f is differentiable almost everywhere. The following proposition describes an alternative characterization of ∂f.

Proposition 2.1.1 ([Clarke *et al.* 1998, Theorem 8.1]). *Assume that $f : \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz continuous. Denote $D \subset \mathbb{R}^d$ the set of points at which f is differentiable and let $A \subset \mathbb{R}^d$ be any Lebesgue-null set. Then*

$$\partial f(x) = \mathrm{conv}\{v : \text{ there is a sequence } x_n \to x \text{ s.t. } x_n \in D \cap A^c,\ \nabla f(x_n) \to v\}\,.$$

The key notion in the analysis of the stochastic subgradient descent is the one of path-differentiability.

Definition 2.1.5 (Path-differentiability [Bolte & Pauwels 2019]). *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 $t \in (0,1)$,*

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

Examples of path-differentiable functions include convex, concave, semialgebraic and more generally definable (see Section 2.4) functions [Bolte & Pauwels 2019]. In nonsmooth optimization, the path-differentiability condition is often a crucial hypothesis in order to obtain relevant results *e.g.*, on the convergence of iterates [Bolte *et al.* 2007, Davis *et al.* 2020, Bolte & Pauwels 2019]. In particular, as we will see in Section 2.2.2 path-differentiability of a function will ensure that it is a Lyapounov function for the subgradient flow.

2.2 Asymptotic pseudotrajectories and differential inclusions

2.2.1 ODE method

The ODE method analyzes the convergence properties of a stochastic approximation algorithm by studying its continuous counterpart: an ordinary differential equation.

Setting up the stage, let $G : \mathbb{R}^d \to \mathbb{R}^d$ be a continuous function and consider the following ODE:

$$\dot{\mathsf{x}}(t) = G(\mathsf{x}(t))\,. \tag{2.1}$$

For $x_0 \in \mathbb{R}^d$, we denote $\Phi(x_0) : \mathbb{R}_+ \to \mathbb{R}^d$ the solution to this ODE starting at x_0. The key notion of the ODE method is the one of the asymptotic pseudotrajectory (APT).

Definition 2.2.1 (Asymptotic pseudotrajectory (APT) [Benaïm 1999]). *We say that a continuous function $\mathsf{X} : \mathbb{R}_+ \to \mathbb{R}^d$ is an asymptotic pseudotrajectory for the ODE (2.1) if for all $T > 0$, we have:*

$$\sup_{h \in [0,T]} \|\mathsf{X}(t+h) - \Phi_h(\mathsf{X}(t))\| \xrightarrow[t \to +\infty]{} 0\,.$$

We endow $C(\mathbb{R}_+, \mathbb{R}^d)$, the space of continuous functions from $\mathbb{R}_+$ to $\mathbb{R}^d$, with the metric of uniform convergence on compact intervals of $\mathbb{R}_+$:

$$\boldsymbol{d}_C(\mathsf{x}, \mathsf{y}) = \sum_{n\in\mathbb{N}} 2^{-n} \left(1 \wedge \sup_{t\in[0,n]} \|\mathsf{x}(t) - \mathsf{y}(t)\| \right) . \tag{2.2}$$

Equivalently, X is an APT for the ODE (2.1) if

$$\boldsymbol{d}_C(\mathsf{X}(t+\cdot), \Phi(\mathsf{X}(t))) \xrightarrow[t\to+\infty]{} 0 .$$

As the following example shows, the notion of an APT naturally arise in the study of stochastic approximation algorithms.

Example 2.2.1. *Consider an $\mathbb{R}^d$-valued sequence (x_n) satisfying the following equation:*

$$x_{n+1} = x_n + \gamma_n G(x_n) + \gamma_n \eta_{n+1} , \tag{2.3}$$

where (γ_n) is a positive sequence and (η_{n+1}) are $\mathbb{R}^d$-valued. One can view Equation (2.3) as an Euler-like discretization of (2.1). Assume the following.

- *(γ_n) is such that $\sum_{i=0}^n \gamma_i \to +\infty$.*
- *The sequence (x_n) is bounded.*
- *For every $T > 0$, we have:*

$$\left\| \sum_{i=n}^{N(T,n)} \gamma_i \eta_{i+1} \right\| \xrightarrow[n\to+\infty]{} 0 ,$$

where for $n \in \mathbb{N}$ and $T > 0$, $N(T, n) = \sup\{k \geqslant n : \sum_{i=n}^k \gamma_i \leqslant T\}$.

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

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

where $\tau_n = \sum_{i=0}^n \gamma_i$, is an APT of the ODE (2.1) (see [Benaïm 1999, Proposition 4.1]).

Remark 1. *A typical situation when the assumption on (η_n) in Example 2.2.1 is verified is when $\sum_{i=0}^n \gamma_i \eta_{i+1}$ converges. For instance, this is the case when (η_n) is a sequence of martingale increments relatively to some filtration, $\lim_{n\to+\infty} \sum_{i=0}^n \gamma_i^2 < +\infty$ and $\sup_{n\in\mathbb{N}} \mathbb{E}[\|\eta_n\|^2] < +\infty$.*

To further characterize the convergence of a stochastic algorithm we need the notion of a Lyapounov function.

Definition 2.2.2 (Lyapounov function)**.** *Let $\mathcal{A}$ be a set in $\mathbb{R}^d$. A continuous function $V : \mathbb{R}^d \to \mathbb{R}$ is a Lyapounov function for $\mathcal{A}$ and the ODE (2.1) if for all $x \in \mathbb{R}^d$, the function $t \mapsto V(\Phi_t(x))$ is decreasing and is strictly decreasing as soon as $x \notin \mathcal{A}$.*

Example 2.2.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be C^1 and consider the ODE (2.1), with $G = -\nabla f$. Then f is a Lyapounov function for its set of critical points. Indeed,*

$$f(\Phi_t(x)) = f(x) - \int_0^t \|\nabla f(\Phi_u(x))\|^2 \, \mathrm{d}u \, .$$

When a process is an APT of an ODE with a Lyapounov function more can be said about its convergence properties.

Proposition 2.2.1 ([Benaïm 1999, Proposition 5.7 and 6.4])**.** *Let X be a bounded APT of the ODE (2.1), let $\mathcal{A} \subset \mathbb{R}^d$ and let V be a Lyapounov function for $\mathcal{A}$. Denote*

$$L_{\mathsf{X}} = \bigcap_{t \in \mathbb{R}_+} \overline{\mathsf{X}([t, +\infty))}$$

the limit set of X. If $V(\mathcal{A})$ is of empty interior, then the following holds.

- *We have that $L_{\mathsf{X}} \subset \mathcal{A}$.*
- *The function V is constant on L_{X}.*

Notice that in the context of Example 2.2.1 the set L_{X} is equal to the set of accumulation points of (x_n).

Example 2.2.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be C^1 and consider the setting of Example 2.2.1, with $G = -\nabla f$. Applying Proposition 2.2.1, we obtain that $f(x_n)$ converges and every accumulation point of (x_n) is a critical point of f.*

2.2.2 Differential inclusions

In nonsmooth analysis the notion of an ODE is replaced by the one of a differential inclusion (DI).

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

$$\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t)) \, . \tag{2.4}$$

We say that an absolutely continuous curve (a.c.) $\mathsf{x} : \mathbb{R}_+ \to \mathbb{R}^d$ is a solution to (2.4) starting at $x \in \mathbb{R}^d$, if $\mathsf{x}(0) = x$ and Equation (2.4) holds for almost every $t \geqslant 0$. We denote $\mathsf{S}_{\mathsf{H}}(x)$ the set of these solutions.

Various notions of continuity exist for set valued maps. The one that will be important for us is the notion of upper semicontinuouty.

Definition 2.2.3. *We say that a set valued map $\mathsf{H} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is* upper semi continuous *at $x \in \mathbb{R}^d$ if for every U, a neighborhood of $\mathsf{H}(x)$, there is $\delta > 0$ such that*

$$\|y - x\| \leqslant \delta \implies \mathsf{H}(y) \subset U \, .$$

We say that H is upper semi continuous *(usc) if it is upper semicontinuous at every point.*

For a locally Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$, the set valued map $\partial f : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is upper semi continuous. Moreover, for $x \in \mathbb{R}^d$, the set $\partial f(x)$ is nonempty, convex and compact. For this type of maps, we have the following existence result.

Proposition 2.2.2 ([Aubin & Cellina 1984]). *Assume that for each x in $\mathbb{R}^d$, $\mathsf{H}(x)$ is nonempty, convex and compact and that there is a constant $C \geqslant 0$ s.t. $\sup\{\|v\| : v \in \mathsf{H}(x)\} \leqslant C(1 + \|x\|)$. Assume that H is usc, then for every $x \in \mathbb{R}^d$, the set $\mathsf{S}_{\mathsf{H}}(x)$ is nonempty.*

The notion of an APT generalizes to the case of differential inclusions.

Definition 2.2.4 ([Benaïm *et al.* 2005]). *We say that a continuous curve $\mathsf{X} : \mathbb{R}_+ \to \mathbb{R}^d$ is an APT of the DI (2.4) if for all $T > 0$,*

$$\sup_{h\in[0,T]} \inf_{\mathsf{x}\in\mathsf{S}_{\mathsf{H}}(\mathsf{x}(t))} \|\mathsf{X}(t+h) - \mathsf{x}(h)\| \xrightarrow[t\to+\infty]{} 0\,. \tag{2.5}$$

Example 2.2.4. *Consider an $\mathbb{R}^d$-valued sequence (x_n) satisfying the following inclusion:*

$$x_{n+1} \in x_n + \gamma_n \mathsf{H}(x_n) + \gamma_n \eta_{n+1}\,, \tag{2.6}$$

with $(\gamma_n), (\eta_n)$ satisfying the assumptions of Example 2.2.1 and H satisfying the assumptions of Proposition 2.2.2. Assume that (x_n) is bounded, then the linearly interpolated process constructed from (x_n) is an APT for the DI (2.4) (see [Benaïm et al. 2005, Theorem 4.1]).

There is a notion of a Lyapounov function in the context of differential inclusions.

Definition 2.2.5 (Lyapounov function (DI)). *Let $\mathcal{A}$ be a set in $\mathbb{R}^d$. A continuous function $V : \mathbb{R}^d \to \mathbb{R}$ is a Lyapounov function for $\mathcal{A}$ and the DI (2.4) if for all $x \in \mathbb{R}^d$, $t > 0$ and $\mathsf{x} \in \mathsf{S}_{\mathsf{H}}(x)$, we have:*

$$V(\mathsf{x}(t)) \leqslant V(x)\,,$$

with strict inequality as soon as $x \notin \mathcal{A}$.

Example 2.2.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous, path-differentiable, function. Consider the DI (2.4), with $\mathsf{H} = -\partial f$. For $x \in \mathbb{R}^d$, consider $\mathsf{x} \in \mathsf{S}_{\mathsf{H}}(x)$. By path-differentiability of f we have:*

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

In particular, f is a Lyapounov function for this DI and the set $\{x \in \mathbb{R}^d : 0 \in \partial f(x)\}$ of Clarke critical points of F.

Similarly to Proposition 2.2.1, a Lyapounov function allows to characterize the convergence properties of an APT related to a DI.

Proposition 2.2.3 ([Benaïm *et al.* 2005, Theorem 3.6 and Proposition 3.27])**.** *Assume that* H *verifies the assumptions of Proposition 2.2.2,* X *is a bounded APT of the DI (2.4) and* V *is a Lyapounov function for a set* $\mathcal{A}$*. Assume that* $V(\mathcal{A})$ *is of empty interior and denote* L_{X} *the limit set of* X*. The following holds.*

- *We have that* $L_{\mathsf{X}} \subset \mathcal{A}$*.*
- *The function* V *is constant on* L_{X}*.*

Example 2.2.6. *Let* $f : \mathbb{R}^d \to \mathbb{R}$ *be a path-differentiable function. Consider the setting of Example 2.2.4, with* $\mathsf{H} = -\partial f$*. We have that* $f(x_n)$ *converges and the accumulation points of* x_n *are in the set* $\{x \in \mathbb{R}^d : 0 \in \partial f(x)\}$ *of Clarke critical points of* f*.*

Remark 2. *In the context of the preceding example it is not necessary to have the existence of a* $C \geqslant 0$ *such that:*

$$\sup\{\|v\| : v \in \partial f(x)\} \leqslant C(1 + \|x\|)\,.$$

Indeed, since the sequence (x_n) *is bounded, it is not hard to construct* $\tilde{f}$ *that agrees with* f *on a compact set containing* (x_n) *and such that* $\partial \tilde{f}$ *satisfies all of the assumptions of Proposition 2.2.2. Such a construction is presented in Chapter 7, Section 7.4.1.*

Remark 3. *Example 2.2.6 along with the preceding remark provides a simple proof of the main result of [Davis & Drusvyatskiy 2021] on the convergence of the stochastic subgradient descent towards the set of (Clarke) critical points.*

2.3 Submanifolds

In this section we present some standard results of differential geometry. An interested reader can find more on this subject in [Lafontaine 2015, Boumal 2020].

We say that a smooth function is an immersion if its Jacobian is injective at each point.

Definition 2.3.1 (Submanifold)**.** *Consider* $p \geqslant 1$*. We say that* $M \subset \mathbb{R}^d$ *is a* C^p *submanifold of dimension* k *if for each* $y \in M$*, there is* U *a neighborhood of* y*,* $V \subset \mathbb{R}^k$ *a neighborhood of* 0 *and* $\varphi : V \to U$ *a* C^p *immersion such that* $\varphi(0) = y$*,* $\varphi(V) = U \cap M$ *and* φ *is an homeomorphism on its image.*

The function φ from the preceding proposition is called a local parametrization of M around y.

Definition 2.3.2 (Tangent and normal spaces)**.** *Consider* $p \geqslant 1$ *and let* M *be a* C^p *submanifold of dimension* k*. Consider* $y \in M$*,* φ *as in Definition 2.3.1 and denote* $J_\varphi(y)$ *the Jacobian of* φ *at* y*. The* tangent space of M at y*, denoted* T_yM*, is a vector space of dimension* k *defined as:*

$$T_yM = \operatorname{Im} J_\varphi(y)\,.$$

The normal space of M at y *is* $N_yM = (T_yM)^\perp$*.*

Remark 4. *Another characterization of the tangent space that we will use is:*

$$T_yM = \{v \in \mathbb{R}^d : \text{ there is a } C^p \text{ curve } \boldsymbol{\gamma} :]-\varepsilon, \varepsilon[\to M, \boldsymbol{\gamma}(0) = y, \dot{\boldsymbol{\gamma}}(0) = v\}\,.$$

Definition 2.3.3. *Consider $p \geqslant 1$. We say that a function $f : M \to \mathbb{R}^k$ is C^p if M is a C^p submanifold and for every $y \in M$, there is a neighborhood U of y and a C^p function $F : U \to \mathbb{R}^k$ that agrees with f on $M \cap U$. We call F a (local) smooth extension of f around y.*

Lemma 2.3.1. *Consider $p \geqslant 1$ and $M \subset \mathbb{R}^d$ a C^p manifold of dimension k. Let φ, U, V be as in Definition 2.3.1. Then the map $\varphi^{-1} : M \cap U \to V$ is C^p.*

Proof. By [Lafontaine 2015, Theorem 1.21] there is $\tilde{V} \subset \mathbb{R}^d$ a neighborhood of zero, $U' \subset U$ a neighborhood of y and $\tilde{\varphi} : V \to U'$ a C^p diffeomorphism such that $\tilde{\varphi}(x_1, \ldots, x_k, 0, \ldots, 0) = \varphi(x_1, \ldots, x_k)$. As a consequence, $\tilde{\varphi}^{-1}$ is smooth and $P_{\mathbb{R}^k} \circ \tilde{\varphi}^{-1}$ is a local smooth extension on φ^{-1} around y, where $P_{\mathbb{R}^k}$ is the projection onto the first k coordinates. □

If a function on a manifold is C^1, then as in the euclidian case we can define its gradient.

Definition 2.3.4 ([Boumal 2020, Proposition 3.53])**.** *Let $f : M \to \mathbb{R}$ be C^1, $y \in M$ and F be a local smooth extension of f around y. We define $\nabla_M f(y)$, the gradient of f at y as:*

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

where P_{T_yM} is the orthogonal projection onto T_yM. This definition is independent from our choice of F.

Definition 2.3.5 (critical points)**.** *Let $f : M \to \mathbb{R}$ be C^1. We say that $x^* \in M$ is a critical point of f if $\nabla_M f(x^*) = 0$.*

Every local extremum of a function defined on a submanifold is a critical point. In the euclidian setting the type of a critical point can be determined by the Hessian. A similar information is available for functions defined on a submanifold.

Definition 2.3.6. *Consider M a C^2 submanifold of $\mathbb{R}^d$ of dimension greater than 0. Let $f : M \to \mathbb{R}$ be C^2 and let $x^* \in M$ be a critical point of f. Consider φ a local parametrization around x^* and denote $\mathcal{H}$ the Hessian of $f \circ \varphi$ at $\varphi^{-1}(x^*)$.*

i) We say that x^* is a nondegenerate critical point if $\mathcal{H}$ is invertible.

ii) We say that x^* is a saddle point if $\mathcal{H}$ has at least one negative eigenvalue.

The consistency of the preceding definition comes from the fact that that if x^* is a critical point of f, then for φ_1, φ_2, any two local parametrizations of M around x^*, we have:

$$\mathcal{H}_1 = \left(J_{\varphi_2^{-1} \circ \varphi_1}(\varphi^{-1}(x^*))\right)^T \mathcal{H}_2 J_{\varphi_2^{-1} \circ \varphi_1}(\varphi^{-1}(x^*))\,,$$

where for $i \in \{1, 2\}$, $\mathcal{H}_i$ is the Hessian of $f \circ \varphi_i$ at $\varphi_i^{-1}(x^*)$ and $J_{\varphi_2^{-1} \circ \varphi_1}$ is the Jacobian of $\varphi_2^{-1} \circ \varphi_1$. [1] The proof of this result can be found in [Victor 1974, Page 42-43].

[1] By Lemma 2.3.1 the composition $\varphi_2^{-1} \circ \varphi_1$ is indeed C^2.

Remark 5. *An equivalent point of view on saddle points is given by the notion of the Riemannian Hessian. Let $f : M \to \mathbb{R}$ be C^2 and x^* be a critical point of f, we define the (Riemannian) Hessian of f at a x^* as the quadratic form $\mathcal{H}_f(x^*)$, defined on $\mathbb{R}^d \times \mathbb{R}^d$ by:*

$$\mathcal{H}_{f,M}(x^*) : \ v \mapsto v^T P_{T_{x^*}M} J_G(x^*) P_{T_{x^*}M} v \,,$$

where G is a C^1 function, defined on a neighborhood of x^*, which agrees with $\nabla_M f$ on M. This definition is independent of the choice of G (see [Boumal 2020, Section 5.5]) and a saddle point in this context is a critical point x^* such that $\mathcal{H}_{f,M}(x^*)$ has at least one negative eigenvalue.

The following lemma gathers useful properties of P_M, the projection onto M.

Lemma 2.3.2 ([Lewis & Malick 2008, Lemma 4])**.** *Consider $p \geqslant 1$ and let $M \subset \mathbb{R}^d$ be a C^p submanifold and y be in M. There is $r > 0$ such that $P_M : B(y, r) \to M$ is well defined, is C^{p-1} and the following properties hold.*

i) For $y' \in M \cap B(y, r)$, the Jacobian of P_M at y' is the projection onto $T_{y'}M$.

ii) For $x \in B(y, r)$, we have $x - P_M(x) \in N_{P_M(x)}M$.

We finish this section by a lemma that gives us a Taylor-like expansion of f around a point on a manifold.

Lemma 2.3.3. *Let $f : M \to \mathbb{R}$ be C^2 and consider $y \in M$. For $y' \in M$, we have:*

$$f(y') = f(y) + \langle \nabla_M f(y), y' - y \rangle + O(\|y' - y\|^2) \,.$$

Proof. Consider F a local smooth extension of f around y and φ a local parametrization of M around y. We have:

$$f(y') = f(y) + \langle \nabla F(y), y' - y \rangle + O(\|y' - y\|^2) \,.$$

Moreover, in the neighborhood of y:

$$\begin{aligned} y' - y &= \varphi(\varphi^{-1}(y')) - \varphi(\varphi^{-1}(y)) \\ &= J_\varphi(\varphi^{-1}(y))(\varphi^{-1}(y') - \varphi^{-1}(y)) + O(\|\varphi^{-1}(y') - \varphi^{-1}(y)\|^2) \\ &= J_\varphi(\varphi^{-1}(y))(\varphi^{-1}(y') - \varphi^{-1}(y)) + O(\|y' - y\|^2) \,, \end{aligned}$$

where the last equality comes from the fact that φ^{-1} is Lipschitz around y (since it is C^2). Moreover, $\operatorname{Im} J_\varphi(\varphi^{-1}(y)) = T_y M$. Therefore, $(y' - y) - P_{T_y M}(y' - y) = O(\|y' - y\|^2)$. This implies:

$$f(y') = f(y) + \langle \nabla_M f(y), y' - y \rangle + O(\|y' - y\|^2) \,.$$

□

2.4 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 [Bolte *et al.* 2007], where the authors analyzed the structure of the Clarke subdifferential of definable functions and extended the Kurdyka-Łojasiewicz inequality [Kurdyka 1998] to the nonsmooth setting. Nowadays, a rich body of literature enforces this link, see e.g. [Davis *et al.* 2020, Drusvyatskiy & Lewis 2010a, Bolte *et al.* 2009, Attouch *et al.* 2011, Bolte & Pauwels 2019]. A nice exposure about usefulness of o-minimal theory in optimization is [Ioffe 2009]. Results on the Verdier and Whitney stratification of definable sets can be found in [Coste 2002, van den Dries & Miller 1996, Loi 1998].

2.4.1 Definition and basic properties

Most of the results of this section can be found in [Coste 2002, van den Dries & Miller 1996].

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

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

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

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

Globally subanalytic. There is an o-minimal structure that contains, for every $n \in \mathbb{N}$, sets of the form $\{(x, t) : t = f(x)\}$, where $f : [-1, 1]^n \to \mathbb{R}$ is an analytic function that can be analytically extended in the neighborhood of the hypercube. This comes from the fact that subanalytic sets are stable by taking a

projection, which was shown by Gabrielov [Gabrielov 1968, Gabrielov 1996]. The sets belonging to this structure are called globally subanalytic (see [Bolte *et al.* 2009, Bierstone & Milman 1988] for more details).

Log-exp. There is an o-minimal structure that contains globally sub-analytic sets as well as the graph of the exponential and the logarithm (see [Wilkie 2009]). As a consequence of this result it can be shown that the loss of a neural network is a definable function [Davis *et al.* 2020].

In the following we fix some o-minimal structure $\mathcal{O}$. Definable will always means definable in $\mathcal{O}$.

An attractive property of definable sets is that they can be constructed by means of *first order formulas*. A first order formula is constructed according to the following rules.

i) If $Q : \mathbb{R}^n \to \mathbb{R}$ is a polynomial, then $Q(x) = 0$ and $Q(x) > 0$ are first order formulas.

ii) If $A \subset \mathbb{R}^n$ is definable, then $x \in A$ is a first order formula.

iii) If $\Phi(x)$ and $\Psi(x)$ are first order formulas, "$\Psi(x)$ and $\Phi(x)$", "$\Psi(x)$ or $\Phi(x)$", "not $\Phi(x)$" and "$\Psi(x) \implies \Phi(x)$" are first order formulas.

iv) If $\Phi(x, y)$ is a first order formula, where $(x, y) \in \mathbb{R}^n \times \mathbb{R}^l$, and $A \subset \mathbb{R}^n$ is definable, then "$\exists x \in A \quad \Psi(x, y)$" and "$\forall x \in A \quad \Psi(x, y)$" are first order formulas.

Proposition 2.4.1 ([Coste 2002, Theorem 1.13])**.** *If $\Phi(x)$ is a first order formula, then the set of x that satisfies $\Phi(x)$ is a definable set.*

The following lemmas show that one dimensional, definable functions behave particularly well.

Lemma 2.4.2 (Monotonicity lemma [van den Dries & Miller 1996, Theorem 4.1])**.** *Let $f : (a, b) \to \mathbb{R}$, with $-\infty \leqslant a < b \leqslant +\infty$, be a definable function and $p \geqslant 0$. There is a finite subdivision $a = a_0 < \cdots < a_k = b$ such that on each interval (a_i, a_{i+1}) f is C^p and either constant or strictly monotone.*

Lemma 2.4.3 (de l'Hôpital inverse rule [Bolte *et al.* 2009, Lemma 1])**.** *Let $\phi, \psi : [0, \varepsilon) \to \mathbb{R}$ be two definable functions that are C^1 on $(0, \varepsilon)$ and continuous at 0, with $\phi(0) = \psi(0) = 0$. Assume that $\forall t \in (0, \varepsilon)$ we have $\psi'(t) > 0$ and there is $l \in \mathbb{R}$ s.t. $\lim_{t\to 0} \frac{\phi(t)}{\psi(t)} = l$. Then $\lim_{t\to 0} \frac{\phi'(t)}{\psi'(t)} = l$.*

Lemma 2.4.4 (Definable choice)**.** *Let $A \subset \mathbb{R}^n \times \mathbb{R}^l$ be a definable set. Let π_n denote the projection on the first n coordinates. Then there is a definable function $\rho : \pi_n(A) \to \mathbb{R}^l$ s.t. for any $x \in \pi_n(A)$, $(x, \rho(x)) \in A$.*

Lemma 2.4.5 (Curve selection lemma [van den Dries & Miller 1996, Theorem 4.6], [Bolte *et al.* 2009])**.** *Let $A \subset \mathbb{R}^n$ be a definable set and $a \in \overline{A}$. For any $p > 0$, there is $\varepsilon > 0$ and a definable curve $\gamma : (-\varepsilon, 1) \to \mathbb{R}^n$ such that γ is C^p, $\gamma(0) = a$ and $\gamma((0, 1)) \subset A$.*

Every definable set can be partitioned into simpler sets called *cells*. The definition is by induction on n.

Definition 2.4.1. *A C^p cell of $\mathbb{R}$ is either a point $\{a\}$ or an open interval (a, b), with $-\infty \leqslant a < b \leqslant +\infty$. Assume that we have constructed the C^p cells of $\mathbb{R}^n$, then there are two types of C^p cells in $\mathbb{R}^{n+1}$.*
Graphs. *$D' = \{(x, \zeta_1(x)) : x \in D\}$, where D is a C^p cell of $\mathbb{R}^n$ and $\zeta_1 : D \to \mathbb{R}$ is a C^p definable function.*
Bands. *$D' = \{(x, y) : \zeta_1(x) < y < \zeta_2(x)\}$, where D is a C^p cell of $\mathbb{R}^d$ and $\zeta_1, \zeta_2 : D \to \mathbb{R}$ are C^p definable functions.*

Definition 2.4.2 (Cylindrical Definable Cell Decomposition (cdcd))**.** *A C^p cdcd of $\mathbb{R}^n$ is a finite partition of $\mathbb{R}^n$ into C^p cells. We say that a cdcd of $\mathbb{R}^n$ is compatible with a family $A_1, \ldots, A_k$, where $A_i \subset \mathbb{R}^n$ if every set of the family is a finite union of cells of cdcd.*

Proposition 2.4.6 (Cell decomposition [van den Dries & Miller 1996, 4.2])**.** *Given a finite family of definable sets $A_1, \ldots, A_k \subset \mathbb{R}^n$, there is a C^p cdcd of $\mathbb{R}^n$ compatible with $A_1, \ldots, A_k$.*

Proposition 2.4.7 (Piecewise smoothness)**.** *Let $A \subset \mathbb{R}^n$ be a definable set and $f : A \to \mathbb{R}$ be a definable function. For any $p \geqslant 0$, there is a C^p cdcd of $\mathbb{R}^n$ compatible with A such that f is continuous on any of its cell.*

To each cell we can inductively associate a dimension.

Definition 2.4.3 (Dimension of a cell)**.** *Dimension of a point is 0, $\dim(a, b) = 1$. If a cell $D' = \{(x, \zeta_1(x)) : x \in D\}$ is a graph, then $\dim D' = \dim D$. If a cell $D' = \{(x, y) : \zeta_1(x) < y < \zeta_2(x)\}$ is a band, then $\dim D' = \dim D + 1$.*

With this definition in hand, Proposition 2.4.6 allows us to define the dimension of any definable set.

Definition 2.4.4 (Dimension of a definable set)**.** *Given $A \subset \mathbb{R}^n$, choose a C^p cdcd of $\mathbb{R}^n$ compatible with A. We define $\dim A$ as the maximum dimension of a cell of this cdcd contained in A, $\dim A$ is then independent of the chosen cdcd.*

Dimension of definable sets verifies many intuitive properties.

Proposition 2.4.8 ([Coste 2002, Section 3.3])**.**

1. *Let A, B be two definable sets. Then $\dim(A \cup B) = \max(\dim A, \dim B)$.*
2. *Let A, B be definable. Then $\dim(A \times B) = \dim A + \dim B$.*
3. *If A and $f : A \to \mathbb{R}^n$ are definable, then $\dim(f(A)) \leqslant \dim A$.*
4. *Let $A \subset \mathbb{R}^{n \times l}$ be definable. For $x \in \mathbb{R}^n$, denote $A_x = \{y \in \mathbb{R}^l : (x, y) \in A\}$. Then for $d \in \mathbb{N}$, the set $A_d = \{x \in \mathbb{R}^n : \dim A_x = d\}$ is definable and $\dim(A \cap A_d \times \mathbb{R}^l) = \dim A_d + d$.*

Remark 6. *It can be established by induction that every C^p cell of dimension k is a k-dimensional submanifold (see [Coste 2002, Chapter 6]). Since the Hausdorff dimension of a k-dimensional C^p submanifold is k, this implies that the Hausdorff dimension of a definable set is equal to its "definable dimension" in the sense of Definition 2.4.4.*

We finish this section by a result that can be viewed as parametrized version of the curve selection lemma. Its proof is an adaptation of [Loi 1998, Lemma 1.7].

Lemma 2.4.9 (Wing lemma)**.** *Let V, S be definable sets such that $V \subset \overline{S}\backslash S$. Assume that $\dim V = k$, with $k > 0$, let p be an integer and denote P_V the projection onto V. There is a definable set $U \subset V$, open in V, a constant $c > 0$ and a definable C^p map $\rho\colon U \times (0, c) \to S$ such that $P_V(\rho(y, t)) = y$ and $\|P_V(\rho(y, t)) - y\| = t$.*

Proof. First, notice that by Proposition 2.4.6 and Remark 6, without loss of generality, we can assume that V is a k-dimensional manifold and, therefore, the projection on V is well defined on its neighborhood.

Let $A \subset V \times \mathbb{R} \times S$ be the following definable set:

$$A = \{(y, t, x) : y \in V, x \in S, t > 0, P_V(x) = y, \|P_V(x) - y\| = t\}\,.$$

Let $\varepsilon : V \to \mathbb{R} \cup \{+\infty\}$ be defined as $\varepsilon(y) := \inf\{t > 0 : \exists x \in S, (y, t, x) \in A\}$.
Claim: $\dim(\{y : \varepsilon(y) > 0\}) < k$. By contradiction suppose that the dimension is k. Then by Proposition 2.4.7 there is a set $B \subset V$, open in V, such that ε is continuous on B. Shrinking B, we can assume that there is a constant $c > 0$ such that $\forall y \in B$, $\varepsilon(y) > c$. This implies that $B \not\subset \overline{S}\backslash S$, a contradiction.
Therefore, there is U open in V such that for each $y \in U$, $\delta > 0$ there is $t < \delta$ and $x \in S$ such that $(y, t, x) \in A$. Fix $y \in U$, the set $\{t : \exists x \in S, (y, t, x) \in A\} \subset \mathbb{R}$ is definable and therefore it is a finite union of points and intervals. Therefore, for each $y \in U$, there is $\delta > 0$ s.t. for every $t < \delta$, there is $x \in S$ and $(y, t, x) \in A$. Let $\delta : U \to \mathbb{R}$ be a function defined as $\delta(y) = \sup\{t' : \forall t \in (0, t'), \exists x \in S, (y, t, x) \in A\}$. We know that for all $y \in U$, $\delta(y) > 0$. Moreover, upon replacing U by a smaller open set, we can assume that $\delta : U \to \mathbb{R}$ is continuous. Upon shrinking U one more time, we have the existence of $c > 0$ such that $c < \delta(y)$.
By the curve selection lemma there is $\rho : U \times (0, c) \to S$ s.t. $(y, t, \rho(y, t)) \in A$. Applying Proposition 2.4.7 to ρ, we obtain that, upon shrinking U and reducing c, ρ is C^p on $U \times (0, c)$, which finishes the proof. □

2.4.2 Stratifications

Various types of cdcd decompositions exist depending on how the neighboring cells fit together. The notion that will be important for us is the notion of stratification.

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

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

Given a family $\{A_1, \dots, A_k\}$ of subsets of 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 stratas. We say that a stratification (S_i) is definable, if every strata S_i is definable.

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

$$\boldsymbol{d}_a(E_1, E_2) = \sup_{u \in E_1, \|u\|=1} \operatorname{dist}(u, E_2). \tag{2.7}$$

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

$$\boldsymbol{d}(E_1, E_2) = \max\{\boldsymbol{d}_a(E_1, E_2), \boldsymbol{d}_a(E_2, E_1)\}\,. \tag{2.8}$$

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

Definition 2.4.5. *We say that a C^p stratification (S_i) satisfies a* Whitney-(a) property*, if for every couple of distinct stratas S_i, S_j, for each $y \in S_i \cap \overline{S_j}$ and for each sequence $(x_n)_{n\in\mathbb{N}} \in (S_j)^{\mathbb{N}}$ such that $x_n \to y$, we have:*

$$\text{w-(a)} \quad \boldsymbol{d}(T_{x_n} S_j, \tau) \to 0 \implies T_y S_i \subset \tau\,. \tag{2.9}$$

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

It is known (see [Coste 2002, van den Dries & Miller 1996]) that every definable function f admits a Whitney C^p (for any p) stratification (X_i) of its domain such that f is C^p on each strata. The following "projection formula" relates the Clarke subdifferential $\partial f(y)$ of f at y, to $\nabla_{X_i} f(y)$.

Lemma 2.4.10 (Projection formula, [Bolte *et al.* 2007, 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* Graph(f)*, such that if one denotes by X_i the projection of S_i onto its first d coordinates, the restriction $f\colon X_i \to \mathbb{R}$ is C^p and the family (X_i) is a Whitney C^p stratification of $\mathbb{R}^d$. Moreover, for any $y \in X_i$ and $v \in \partial f(y)$, we have $P_{T_y X_i}(v) = \nabla_{X_i} f(y)$.*

Lemma 2.4.10 has important consequences. One of them (see [Davis *et al.* 2020, Section 5] is that every locally Lipschitz continuous and definable function is path-differentiable.

Lemma 2.4.11 ([Davis *et al.* 2020, Theorem 5.8])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function. If* Graph(f) *admits a Whitney C^1 stratification, then f is path-differentiable.*

A Verdier stratification is a special case of Whitney stratification, which posit a stronger condition on the (asymmetric) distance between adjacent stratas. Whereas the Whitney stratification can now be considered as well known in optimization community, the Verdier stratification is comparatively less popular. We illustrate its advantage by establishing in Theorem 5.2.1 a Lipschitz-like condition in the "projection formula" (Lemma 2.4.10). We believe that this strengthened result is of independent interest.

Definition 2.4.6. *Let (S_i) be a C^p stratification of some set $A \subset \mathbb{R}^d$. We say that (S_i) satisfies a* Verdier property (v)*, if for every couple of distinct stratas S_i, S_j and for each $y \in S_i \cap \overline{S_j} \neq \varnothing$, there are two positive constants δ, C such that:*

$$(v) \quad \begin{array}{ll} y' & \in B(y,\delta) \cap S_i \\ x & \in B(y,\delta) \cap S_j \end{array} \implies \boldsymbol{d}_a(T_{y'}S_i, T_x S_j) \leqslant C \left\| y' - x \right\| . \tag{2.10}$$

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

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

Proposition 2.4.12 ([Loi 1998, 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 $\{A_1, \ldots, A_k\}$.*

In Chapter 5 this proposition will be used to prove a reinforced version of Lemma 2.4.10.

Chapter 3

Stochastic optimization with momentum: convergence, fluctuations, and traps avoidance

3.1 Introduction

Given a probability space Ξ, an integer $d > 0$, and a function $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$, consider the problem of finding a local minimum of the function $F(x) \stackrel{\Delta}{=} \mathbb{E}_\xi[f(x,\xi)]$ w.r.t. $x \in \mathbb{R}^d$, where $\mathbb{E}_\xi$ represents the expectation w.r.t. the random variable ξ on Ξ. This chapter focuses on the case where F is possibly non-convex. It is assumed that the function F is unknown to the observer, either because the distribution of ξ is unknown, or because the expectaction cannot be evaluated. Instead, a sequence $(\xi_n : n \geqslant 1)$ of i.i.d. copies of the random variable ξ is revealed online.

While the Stochastic Gradient Descent is the most classical algorithm that is used to solve such a problem, recently, several other algorithms became very popular. These include the Stochastic Heavy Ball (SHB), the stochastic version of Nesterov's Accelerated Gradient method (S-NAG) and the large class of the so-called *adaptive* gradient algorithms, among which Adam [Kingma & Ba 2015] is perhaps the most used in practice. As opposed to the vanilla Stochastic Gradient Descent, the study of such algorithms is more elaborate, for three reasons. First, the update of the iterates involves a so-called *momentum* term, or inertia, which has the effect of "smoothing" the increment between two consecutive iterates. Second, the update equation at the time index n is likely to depend on n, making these systems inherently *non-autonomous*. Third, as far as adaptive algorithms are concerned, the update also depends on some additional variable (*a.k.a.* the learning rate) computed online as a function of the history of the computed gradients.

In this chapter, we study in a unified way the asymptotic behavior of these algorithms in the situation where F is a differentiable function which is not necessarily convex, and where the stepsize of the algorithm is decreasing.

Our starting point is a generic non-autonomous Ordinary Differential Equation (ODE) introduced by Belotto da Silva and Gazeau [Belotto da Silva & Gazeau 2020] (see also [Barakat & Bianchi 2021] for Adam), depicting the continuous-time versions of the aforementioned florilegium of algorithms. The solutions to the ODE are shown to converge to the set of critical points of F. This suggests that a general provably convergent algorithm can be obtained by means of an Euler discretization of the ODE, including possible stochastic perturbations. Special cases of our general

algorithm include SHB, ADAM and S-NAG. We establish the almost sure boundedness and the convergence to critical points. Under additional assumptions, we obtain convergence rates, under the form of a central limit theorem. These results are new. They extend the works of [Gadat *et al.* 2018, Barakat & Bianchi 2021] to a general setting. In particular, we highlight the almost sure convergence result of S-NAG in a non-convex setting, which is new to the best of our knowledge.

Next, we address the question of the avoidance of "traps". In a non-convex setting, the set of critical points of a function F is generally larger than the set of local minimizers. A "trap" stands for a critical point at which the Hessian matrix of F has negative eigenvalues, namely, it is a local maximum or saddle point. We establish that the iterates cannot converge to such a point, if the noise is exciting in some directions. The result extends previous works of [Gadat *et al.* 2018] obtained in the context of SHB. This result not only allows to study a broader class of algorithms but also significantly weakens the assumptions. In particular, [Gadat *et al.* 2018] uses a sub-Gaussian assumption on the noise and a rather stringent assumption on the stepsizes. The main difficulty in the approach of [Gadat *et al.* 2018] lies in the use of the classical autonomous version of Poincaré's invariant manifold theorem. The key ingredient of our proof is a general avoidance of traps result, adapted to non-autonomous settings, which we believe to be of independent interest. It extends usual avoidance of traps results to a non-autonomous setting, by making use of a non-autonomous version of Poincaré's theorem [Dalec′kiĭ & Kreĭn 1974, Kloeden & Rasmussen 2011].

Chapter organization. In Section 3.2, we introduce and study the ODE's governing our general stochastic algorithm. We establish the existence and uniqueness of the solutions, as well as the convergence to the set of critical points. In Section 3.3, we introduce the main algorithm. We provide sufficient conditions under which the iterates are bounded and converge to the set of critical points. A central limit theorem is stated. Section 3.4 introduces a general avoidance of traps result for non-autonomous settings. Next, this result is applied to the proposed algorithm. Sections 3.5, 3.6 and 3.7 are devoted to the proofs of the results of Sections 3.2, 3.3 and 3.4, respectively.

Notations. Given an integer $d \geqslant 1$, two vectors $x, y \in \mathbb{R}^d$, and a real α, we denote by $x \odot y$, $x^{\odot\alpha}$, x/y, $|x|$, and $\sqrt{|x|}$ the vectors in $\mathbb{R}^d$ whose i-th coordinates are respectively given by $x_i y_i$, x_i^α, x_i/y_i, $|x_i|$, $\sqrt{|x_i|}$. Inequalities of the form $x \leqslant y$ are to be read componentwise. The standard Euclidean norm is denoted $\|\cdot\|$. Notation M^{T} represents the transpose of a matrix M. For $x \in \mathbb{R}^d$ and $\rho > 0$, the notation $B(x, \rho)$ stands for the open ball of $\mathbb{R}^d$ with center x and radius ρ. We also write $\mathbb{R}_+ = [0, \infty)$. If $z \in \mathbb{R}^d$ and $A \subset \mathbb{R}^d$, we write $\operatorname{dist}(z, A) \stackrel{\Delta}{=} \inf\{\|z - z'\| : z' \in A\}$. By $\mathbb{1}_A(x)$, we refer to the function that is equal to one if $x \in A$ and to zero elsewhere. The set of zeros of a function $h : \mathbb{R}^d \to \mathbb{R}^{d'}$ is $\operatorname{zer} h = \{x : h(x) = 0\}$. Let D be a domain in $\mathbb{R}^d$. Given an integer $k \geqslant 0$, the class $\mathcal{C}^k(D, \mathbb{R})$ is the class of $D \to \mathbb{R}$ maps

such that all their partial derivatives up to the order k exist and are continuous. For a function $h \in \mathcal{C}^k(D, \mathbb{R})$ and for every $i \in \{1, \dots, d\}$, we denote as $\partial_i^k h(x_1, \dots, x_d)$ the k^{th} partial derivative of the function h with respect to x_i. When $k = 1$, we just write $\partial_i h(x_1, \dots, x_d)$. The gradient of a function $F : \mathbb{R}^d \to \mathbb{R}$ at a point $x \in \mathbb{R}^d$ is denoted as $\nabla F(x)$, and its Hessian matrix at x is $\nabla^2 F(x)$ as usual. For a function $S : \mathbb{R}^d \to \mathbb{R}^d$, the notation $\nabla S(x)$ stands for the jacobian matrix of S at point x.

3.2 Ordinary Differential Equations

3.2.1 A general ODE

Our starting point will be a non-autonomous ODE which is almost identical to the one introduced in [Belotto da Silva & Gazeau 2020] and close to the one in [Barakat & Bianchi 2021]. Let F be a function in $\mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$, let S be a continuous $\mathbb{R}^d \to \mathbb{R}^d_+$ function, let $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q} : (0, \infty) \to \mathbb{R}_+$ be four continuous functions, and let $\varepsilon > 0$. Let $v_0 \in \mathbb{R}^d_+$ and $x_0, m_0 \in \mathbb{R}^d$. Starting at $\mathsf{v}(0) = v_0$, $\mathsf{m}(0) = m_0$, and $\mathsf{x}(0) = x_0$, our ODE on $\mathbb{R}_+$ with trajectories in $\mathcal{Z}_+ \triangleq \mathbb{R}^d_+ \times \mathbb{R}^d \times \mathbb{R}^d$ reads

$$\begin{cases} \dot{\mathsf{v}}(t) &= \mathsf{p}(t) S(\mathsf{x}(t)) - \mathsf{q}(t)\mathsf{v}(t) \\ \dot{\mathsf{m}}(t) &= \mathsf{h}(t) \nabla F(\mathsf{x}(t)) - \mathsf{r}(t)\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\mathsf{m}(t)/\sqrt{\mathsf{v}(t) + \varepsilon} \end{cases} \tag{ODE-1}$$

This ODE can be rewritten compactly in the following form. Write $z_0 = (v_0, m_0, x_0)$, and let $\mathsf{z}(t) = (\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t)) \in \mathcal{Z}_+$ for $t \in \mathbb{R}_+$. Let $\mathcal{Z} \triangleq \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$, and define the map $g : \mathcal{Z}_+ \times (0, \infty) \to \mathcal{Z}$ as

$$g(z, t) = \begin{bmatrix} \mathsf{p}(t) S(x) - \mathsf{q}(t) v \\ \mathsf{h}(t) \nabla F(x) - \mathsf{r}(t) m \\ -m/\sqrt{v + \varepsilon} \end{bmatrix} \tag{3.1}$$

for $z = (v, m, x) \in \mathcal{Z}_+$. With these notations, we can rewrite (ODE-1) as

$$\mathsf{z}(0) = z_0, \quad \dot{\mathsf{z}}(t) = g(\mathsf{z}(t), t) \text{ for } t > 0.$$

By setting $S(x) = \nabla F(x)^{\odot 2}$ when necessary and by properly choosing the functions h, r, p, and q, a large number of iterative algorithms used in Machine Learning can be obtained by an Euler's discretization of this ODE. For instance, choosing $\mathsf{h}(t) = \mathsf{r}(t) = a(t, \lambda, \alpha_1)$ and $\mathsf{p}(t) = \mathsf{q}(t) = a(t, \lambda, \alpha_2)$ with $a(t, \lambda, \alpha) = \lambda^{-1}(1 - \exp(-\lambda\alpha))/(1 - \exp(-\alpha t))$ and $\lambda, \alpha_1, \alpha_2 > 0$, one obtains a version of the Adam algorithm [Kingma & Ba 2015] (see [Belotto da Silva & Gazeau 2020, Sections 2.4-4.2] for details). To give another less specific example, if we set $\mathsf{p} = \mathsf{q} \equiv 0$, then the resulting ODE covers a family of algorithms to which the well-known Heavy Ball with friction algorithm [Attouch *et al.* 2000] belongs. For a comprehensive and more precise view of the deterministic algorithms that can be deduced from (ODE-1) by

an Euler's discretization, the reader is referred to [Belotto da Silva & Gazeau 2020, Table 1].

In this chapter, since we are rather interested in stochastic versions of these algorithms, Equation (ODE-1) will be the basic building block of the classical "ODE method" which is widely used in the field of stochastic approximation [Benaïm 1999]. In order to analyze the behavior of this equation in preparation of the stochastic analysis, we need the following assumptions.

Assumption 3.2.1. *The function F belongs to $\mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ and ∇F is locally Lipschitz continuous.*

Assumption 3.2.2. *F is coercive,* i.e., *$F(x) \to +\infty$ as $\|x\| \to +\infty$.*

Note that this assumption implies that the infimum $F_\star$ of F is finite, and the set $\operatorname{zer} \nabla F$ of zeros of ∇F is nonempty.

Assumption 3.2.3. *The map $S : \mathbb{R}^d \to \mathbb{R}^d_+$ is locally Lipschitz continuous.*

Assumption 3.2.4. *The continuous functions $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q} : (0, +\infty) \to \mathbb{R}_+$ satisfy:*

i) $\mathsf{h} \in \mathcal{C}^1((0, +\infty), \mathbb{R}_+)$, $\dot{\mathsf{h}}(t) \leqslant 0$ *on* $(0, +\infty)$ *and the limit* $h_\infty \triangleq \lim_{t\to\infty} \mathsf{h}(t)$ *is positive.*

ii) r *and* q *are non-increasing and* $r_\infty \triangleq \lim_{t\to\infty} \mathsf{r}(t)$, $q_\infty \triangleq \lim_{t\to\infty} \mathsf{q}(t)$ *are positive.*

iii) p *converges towards* p_∞ *as* $t \to \infty$.

iv) For all $t \in (0, +\infty)$, $\mathsf{r}(t) \geqslant \mathsf{q}(t)/4$ *and* $r_\infty > q_\infty/4$.

These assumptions are sufficient to prove the existence and the uniqueness of the solution to (ODE-1) starting at a time $t_0 > 0$. The following additional assumption extends the solution to $t_0 = 0$.

Assumption 3.2.5. *Either $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q} \in \mathcal{C}^1([0, +\infty), \mathbb{R}_+)$, or the following holds:*

i) For every $x \in \mathbb{R}^d$, we have $S(x) \geqslant \nabla F(x)^{\odot 2}$.

ii) The functions $\frac{\mathsf{h}}{\mathsf{p}}$, $\frac{\mathsf{h}}{\mathsf{q}-2\mathsf{r}}$, $t \mapsto t\mathsf{h}(t)$, $t \mapsto t\mathsf{r}(t)$, $t \mapsto t\mathsf{p}(t)$, $t \mapsto t\mathsf{q}(t)$ are bounded near zero.

iii) There exists $t_0 > 0$ such that for all $t < t_0$, $2\mathsf{r}(t) - \mathsf{q}(t) > 0$.

iv) There exists $\delta > 0$ such that $\frac{\mathsf{h}}{\mathsf{r}}, \frac{\mathsf{p}}{\mathsf{q}} \in \mathcal{C}^1([0, \delta), \mathbb{R}_+)$.

v) The initial condition $z_0 = (v_0, m_0, x_0) \in \mathcal{Z}_+$ satisfies

$$m_0 = \nabla F(x_0) \lim_{t\downarrow 0} \frac{\mathsf{h}(t)}{\mathsf{r}(t)} \quad \text{and} \quad v_0 = S(x_0) \lim_{t\downarrow 0} \frac{\mathsf{p}(t)}{\mathsf{q}(t)} .$$

Remark 7. *The functions* $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q}$ *corresponding to* ADAM *satisfy these conditions. We leave the straightforward verifications to the reader. We just observe here that the function* S *that will correspond to our stochastic algorithm in Section 3.3 below will satisfy Assumption 3.2.5–i) by an immediate application of Jensen's inequality.*

The following theorem slightly generalizes the results of [Belotto da Silva & Gazeau 2020, Theorem 3 and Theorem 5].

Theorem 3.2.1. *Let Assumptions 3.2.1 to 3.2.4 hold true. Consider* $z_0 \in \mathcal{Z}_+$ *and* $t_0 > 0$. *Then, there exists a unique global solution* $\mathsf{z} : [t_0, +\infty) \to \mathcal{Z}_+$ *to* (ODE-1) *with initial condition* $\mathsf{z}(t_0) = z_0$. *Moreover,* $\mathsf{z}([t_0, +\infty))$ *is a bounded subset of* $\mathcal{Z}_+$. *As* $t \to +\infty$, $\mathsf{z}(t)$ *converges towards the set*

$$\Upsilon \triangleq \{z_\star = (p_\infty S(x_\star)/q_\infty, 0, x_\star) \,:\, x_\star \in \operatorname{zer} \nabla F\}\,. \tag{3.2}$$

If, additionally, Assumption 3.2.5 holds, then we can take $t_0 = 0$.

Remark 8. *Theorem 3.2.1 only shows the convergence of the trajectory* $\mathsf{z}(t)$ *towards a set. Convergence of the trajectory towards a single point is not guaranteed when the set* Υ *is not countable.*

Remark 9. *A simpler version of* (ODE-1) *is obtained when omitting the momentum term. It reads:*

$$\begin{cases} \dot{\mathsf{v}}(t) &= \mathsf{p}(t)S(\mathsf{x}(t)) - \mathsf{q}(t)\mathsf{v}(t) \\ \dot{\mathsf{x}}(t) &= -\nabla F(\mathsf{x}(t))/\sqrt{\mathsf{v}(t) + \varepsilon}\,. \end{cases} \tag{ODE-1$'$}$$

This ODE encompasses the algorithms of the family of RMSPROP *[Tieleman & Hinton 2012], as shown in [Barakat & Bianchi 2021, Belotto da Silva & Gazeau 2020]. The approach for proving the previous theorem can be adapted to* (ODE-1$'$) *with only minor modifications. In the proofs below, we will point out the particularities of* (ODE-1$'$) *when necessary.*

The following paragraph is devoted to a particular case of (ODE-1), which does not satisfy Assumption 3.2.4, and which requires a more involved treatment than (ODE-1$'$).

3.2.2 The Nesterov case

The authors of [Cabot *et al.* 2009], [Su *et al.* 2016b] and others studied the ODE

$$\ddot{\mathsf{x}}(t) + \frac{\alpha}{t}\dot{\mathsf{x}}(t) + \nabla F(\mathsf{x}(t)) = 0, \quad \alpha > 0, \quad F \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R}),$$

which Euler's discretization generates the well-known Nesterov's accelerated gradient algorithm, see also [Attouch *et al.* 2018, Aujol *et al.* 2019]. This ODE can be rewritten as

$$\begin{cases} \dot{\mathsf{m}}(t) &= \nabla F(\mathsf{x}(t)) - \frac{\alpha}{t}\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\mathsf{m}(t), \end{cases} \tag{ODE-N}$$

which is formally the particular case of (ODE-1) that is taken for $\mathsf{p}(t) = \mathsf{q}(t) = 0$, $\mathsf{h}(t) = 1$, and $\mathsf{r}(t) = \alpha/t$. Obviously, this case is not covered by Assumption 3.2.4. Moreover, it turns out that, contrary to the situation described in Remark 9 above, this case cannot be dealt with by a straightforward adaptation of the proof of Theorem 3.2.1. The reason for this is as follows. Heuristically, the proof of Theorem 3.2.1 is built around the fact that the solution of (ODE-1) "shadows" for large t the solution of the autonomous ODE

$$\begin{cases} \dot{\mathsf{v}}(t) &= p_\infty S(\mathsf{x}(t)) - q_\infty \mathsf{v}(t) \\ \dot{\mathsf{m}}(t) &= h_\infty \nabla F(\mathsf{x}(t)) - r_\infty \mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\frac{\mathsf{m}(t)}{\sqrt{\mathsf{v}(t)+\varepsilon}}, \end{cases}$$

and the latter can be shown to converge to the set Υ defined in Equation (3.2), either under Assumption 3.2.4 or for the algorithms covered by Remark 9. This idea does not work anymore for (ODE-N), for its large–t autonomous counterpart

$$\begin{cases} \dot{\mathsf{m}}(t) &= \nabla F(\mathsf{x}(t)) \\ \dot{\mathsf{x}}(t) &= -\mathsf{m}(t). \end{cases}$$

can have solutions that do not converge to the critical points of F. As an example of such solutions, take $d = 1$ and $F(x) = x^2/2$. Then, $t \mapsto (\cos(t), \sin(t))$ is an oscillating solution of the latter ODE.

Yet, we have the following result. Up to our knowledge, the proof of the convergence below as $t \to +\infty$ is new.

Theorem 3.2.2. *Let Assumptions 3.2.1 and 3.2.2 hold true. Then, for each $x_0 \in \mathbb{R}^d$, there exists a unique bounded global solution $(\mathsf{m}, \mathsf{x}) : \mathbb{R}_+ \to \mathbb{R}^d \times \mathbb{R}^d$ to (ODE-N) with the initial condition $(\mathsf{m}(0), \mathsf{x}(0)) = (0, x_0)$. As $t \to +\infty$, $(\mathsf{m}(t), \mathsf{x}(t))$ converges towards the set*

$$\bar{\Upsilon} \stackrel{\Delta}{=} \{(0, x_\star) \,:\, x_\star \in \operatorname{zer} \nabla F\}. \tag{3.3}$$

3.2.3 Related works

The continuous-time dynamical system (ODE-1) we consider was first introduced in [Belotto da Silva & Gazeau 2020, Equation (2.1)] with $S = \nabla F^{\odot 2}$. Theorem 3.2.1 above is roughly the same as [Belotto da Silva & Gazeau 2020, Ths. 3 and 5], with some slight differences regarding the assumptions on the function F, or Assumption 3.2.4-*iv)*. We point out that the main focus of [Belotto da Silva & Gazeau 2020] is to study the properties of the *deterministic continous-time* dynamical system (ODE-1). In the present chapter, we highlight that the purpose of Theorem 3.2.1 is to pave the way to our analysis of the corresponding *stochastic algorithms* in Section 3.3.

Concerning Theorem 3.2.2, the existence and the uniqueness of a global solution to (ODE-N) has been previously shown in the literature, for instance in

[Cabot *et al.* 2009, Proposition 2.1] or in [Su *et al.* 2016b, Theorem 1]. The convergence statement in Theorem 3.2.2 is new to the best of our knowledge. In particular, we stress that we do not make any convexity assumption on F. The closest result we are aware of is the one of Cabot-Engler-Gadat [Cabot *et al.* 2009]. In [Cabot *et al.* 2009, Proposition 2.5], it is shown that if $\mathsf{x}(t)$ converges towards some point $\bar{x}$, then necessarily $\bar{x}$ is a critical point of F. Our result in Theorem 3.2.2 strengthens this statement, by establishing that $\mathsf{x}(t)$ actually converges to the set of critical points.

3.3 Stochastic Algorithms

In this section, we discuss the asymptotic behavior of stochastic algorithms that consist in noisy Euler's discretizations of (ODE-1) and (ODE-N) studied in the previous section.

We first set the stage. Let $(\Xi, \mathscr{T}, \mu)$ be a probability space. Denoting as $\mathscr{B}(\mathbb{R}^d)$ the Borel σ-algebra on $\mathbb{R}^d$, consider a $\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}$–measurable function $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ that satisfies the following assumption.

Assumption 3.3.1. *The following conditions hold:*

i) For every $x \in \mathbb{R}^d$, $f(x,\cdot)$ is μ–integrable.

ii) For every $s \in \Xi$, the map $f(\cdot, s)$ is differentiable. Denoting as $\nabla f(x,s)$ its gradient w.r.t. x, the function $\nabla f(x,\cdot)$ is integrable.

iii) There exists a measurable map $\kappa : \mathbb{R}^d \times \Xi \to \mathbb{R}_+$ s.t. for every $x \in \mathbb{R}^d$:

a) The map $\kappa(x,\cdot)$ is μ–integrable,

b) There exists $\varepsilon > 0$ s.t. for every $s \in \Xi$,

$$\forall\, u, v \in B(x,\varepsilon),\ \|\nabla f(u,s) - \nabla f(v,s)\| \leqslant \kappa(x,s)\|u - v\|\,.$$

Under Assumption 3.3.1, we can define the mapping $F : \mathbb{R}^d \to \mathbb{R}$ as

$$F(x) = \mathbb{E}_\xi[f(x,\xi)] \tag{3.4}$$

for all $x \in \mathbb{R}^d$, where we write $\mathbb{E}_\xi\varphi(\xi) = \int \varphi(\xi)\mu(d\xi)$. It is easy to see that the mapping F is differentiable,

$$\nabla F(x) = \mathbb{E}_\xi[\nabla f(x,\xi)]$$

for all $x \in \mathbb{R}^d$, and ∇F is locally Lipschitz.

Let $(\gamma_n)_{n\geqslant 1}$ be a sequence of positive real numbers satisfying

Assumption 3.3.2. $\gamma_{n+1}/\gamma_n \to 1$ *and* $\sum_n \gamma_n = +\infty$.

Define for every integer $n \geqslant 1$

$$\tau_n = \sum_{k=1}^{n} \gamma_k\,.$$

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space, and let $(\xi_n : n \geqslant 1)$ be a sequence of iid random variables defined from $(\Omega, \mathscr{F}, \mathbb{P})$ into $(\Xi, \mathscr{T}, \mu)$ with the distribution μ.

3.3.1 General algorithm

Our first algorithm is a discrete and noisy version of (ODE-1). Let $z_0 = (v_0, m_0, x_0) \in \mathcal{Z}_+$ and $h_0, r_0, p_0, q_0 \in (0, \infty)$. Define for every $n \geqslant 1$

$$h_n = \mathsf{h}(\tau_n),\ r_n = \mathsf{r}(\tau_n),\ p_n = \mathsf{p}(\tau_n),\ \text{and}\ q_n = \mathsf{q}(\tau_n). \tag{3.5}$$

The algorithm is written as follows.

Algorithm 1 (general algorithm)

Initialization: $z_0 \in \mathcal{Z}_+$.
for $n = 1$ **to** n_{iter} **do**
$v_{n+1} = (1 - \gamma_{n+1} q_n) v_n + \gamma_{n+1} p_n \nabla f(x_n, \xi_{n+1})^{\odot 2}$
$m_{n+1} = (1 - \gamma_{n+1} r_n) m_n + \gamma_{n+1} h_n \nabla f(x_n, \xi_{n+1})$
$x_{n+1} = x_n - \gamma_{n+1} m_{n+1} / \sqrt{v_{n+1} + \varepsilon}$.

We suppose throughout this chapter that $1 - \gamma_{n+1} q_n \geqslant 0$ for all $n \in \mathbb{N}$. This will guarantee that the quantity $\sqrt{v_n + \varepsilon}$ is always well-defined (see Algorithm 1). This mild assumption is satisfied as soon as $q_0 \leqslant \frac{1}{\gamma_1}$ since the sequence (q_n) is non-increasing and the sequence of stepsizes (γ_n) can also be supposed to be non-increasing.

Since this algorithm makes use of the function $\nabla f(x, \xi)^{\odot 2}$, a strengthening of Assumption 3.3.1 is required:

Assumption 3.3.3. *In Assumption 3.3.1, Conditions ii) and iii) are respectively replaced with the stronger conditions*

ii') For each $x \in \mathbb{R}^d$, the function $\nabla f(x, \cdot)^{\odot 2}$ is μ -integrable.

iii') There exists a measurable map $\kappa : \mathbb{R}^d \times \Xi \to \mathbb{R}_+$ s.t. for every $x \in \mathbb{R}^d$:

a) The map $\kappa(x, \cdot)$ is μ–integrable.

b) There exists $\varepsilon > 0$ s.t.

$$\forall\, u, v \in B(x, \varepsilon),\ \|\nabla f(u, s) - \nabla f(v, s)\| \vee \|\nabla f(u, s)^{\odot 2} - \nabla f(v, s)^{\odot 2}\| \leqslant \kappa(x, s)\|u - v\|.$$

Under Assumption 3.3.3, we can also define the mapping $S : \mathbb{R}^d \to \mathbb{R}^d$ as:

$$S(x) = \mathbb{E}_\xi[\nabla f(x, \xi)^{\odot 2}]$$

for all $x \in \mathbb{R}^d$. Notice that Assumptions 3.2.1 and 3.2.3 are satisfied for F and S.

Assumption 3.3.4. *Assume either of the following conditions.*

i) There exists $q \geqslant 2$ s.t. for every compact set $\mathcal{K} \subset \mathbb{R}^d$,

$$\sup_{x \in \mathcal{K}} \mathbb{E}_\xi \|\nabla f(x, \xi)\|^{2q} < \infty \quad \textit{and} \quad \sum_n \gamma_n^{1+q/2} < \infty\,.$$

ii) For every compact set $\mathcal{K} \subset \mathbb{R}^d$, there exists a real $\sigma_{\mathcal{K}} \neq 0$ s.t.

$$\mathbb{E}_\xi \exp\langle u, \nabla f(x,\xi) - \nabla F(x)\rangle \mathbb{1}_{x\in\mathcal{K}} \leqslant \exp\left(\sigma_{\mathcal{K}}^2 \|u\|^2/2\right) \text{ and}$$
$$\mathbb{E}_\xi \exp\langle u, \nabla f(x,\xi)^{\odot 2} - S(x)\rangle \mathbb{1}_{x\in\mathcal{K}} \leqslant \exp\left(\sigma_{\mathcal{K}}^2 \|u\|^2/2\right) ,$$

for every $x, u \in \mathbb{R}^d$. Moreover, for every $\alpha > 0$, $\sum_n \exp(-\alpha/\gamma_n) < \infty$.

Remark 10. *We make the following comments regarding Assumption 3.3.4.*

- *Assumption 3.3.4-i) allows to use larger stepsizes in comparison to the classical condition $\sum_n \gamma_n^2 < \infty$ which corresponds to the particular case $q = 2$.*
- *Recall that a random vector X is said to be subgaussian if there exists a real $\sigma \neq 0$ s.t. $\mathbb{E}e^{\langle u, X\rangle} \leqslant e^{\sigma^2\|u\|^2/2}$ for every constant vector $u \in \mathbb{R}^d$. In Assumption 3.3.4-ii), the subgaussian noise offers the possibility to use a sequence of stepsizes with an even slower decay rate than in Assumption 3.3.4–i).*

Assumption 3.3.5. *The set $F(\{x : \nabla F(x) = 0\})$ has an empty interior.*

Remark 11. *Assumption 3.3.5 excludes a pathological behavior of the objective function F at critical points. It is satisfied when $F \in \mathcal{C}^k(\mathbb{R}^d, \mathbb{R})$ for $k \geqslant d$. Indeed, in this case, Sard's theorem stipulates that the Lebesgue measure of $F(\{x : \nabla F(x) = 0\})$ is zero in $\mathbb{R}$.*

Theorem 3.3.1. *Let Assumptions 3.2.2, 3.2.4, and 3.3.2–3.3.5 hold true. Assume that the random sequence $(z_n = (v_n, m_n, x_n) : n \in \mathbb{N})$ given by Algorithm 1 is bounded with probability one. Then, w.p.1, the sequence (z_n) converges towards the set Υ defined in Equation (3.2). If, in addition, the set of critical points of the objective function F is finite or countable, then w.p.1, the sequence (z_n) converges to a single point of Υ.*

We now deal with the boundedness problem of the sequence (z_n). We introduce an additional assumption for this purpose.

Assumption 3.3.6. *The following conditions hold.*

i) ∇F is (globally) Lipschitz continuous.

ii) There exists $C > 0$ s.t. for all $x \in \mathbb{R}^d$, $\mathbb{E}_\xi[\|\nabla f(x,\xi)\|^2] \leqslant C(1 + F(x))$,

iii) $\sum_n \gamma_n^2 < \infty$.

Theorem 3.3.2. *Let Assumptions 3.2.2, 3.2.4, 3.3.2, 3.3.3, 3.3.4-i) (with $q = 2$) and 3.3.6 hold. Then, the sequence (v_n, m_n, x_n) given by Algorithm 1 is bounded with probability one.*

Remark 12. *The above stability result requires square summable step sizes. Showing the same boundedness result under the Assumption 3.3.4 that allows for larger step sizes is a challenging problem in the general case. In these situations, the boundedness of the iterates can be sometimes ensured by* ad hoc *means.*

Remark 13. *We can also consider the noisy discretization of* (ODE-1′) *introduced in Remark 9 above. This algorithm reads*

$$\begin{cases} v_{n+1} &= (1-\gamma_{n+1}q_n)v_n + \gamma_{n+1}p_n\nabla f(x_n,\xi_{n+1})^{\odot 2} & \text{(3.6a)} \\ x_{n+1} &= x_n - \gamma_{n+1}\nabla f(x_n,\xi_{n+1})/\sqrt{v_{n+1}+\varepsilon} & \text{(3.6b)} \end{cases}$$

for $(v_0, x_0) \in \mathbb{R}^d_+ \times \mathbb{R}^d$. *With only minor adaptations, Theorem 3.3.1 and Theorem 3.3.2 can be shown to hold as well for this algorithm. We refer to the concomitant paper [Gadat & Gavra 2020, Sec. 2.2] for the link between this algorithm and the seminal algorithms* AdaGrad *[Duchi et al. 2011] and* RMSProp *[Tieleman & Hinton 2012].*

3.3.2 Stochastic Nesterov's Accelerated Gradient (S-NAG)

S-NAG is the noisy Euler's discretization of (ODE-N). Given $\alpha > 0$, it generates the sequence (m_n, x_n) on $\mathbb{R}^d \times \mathbb{R}^d$ given by Algorithm 2.

Algorithm 2 (S-NAG with decreasing steps)

Initialization: $m_0 = 0, x_0 \in \mathbb{R}^d$.
for $n = 1$ **to** n_{iter} **do**
$m_{n+1} = (1 - \alpha\gamma_{n+1}/\tau_n)m_n + \gamma_{n+1}\nabla f(x_n, \xi_{n+1})$
$x_{n+1} = x_n - \gamma_{n+1}m_{n+1}$.

Assumption 3.3.7. *Assume either of the following conditions.*

i) There exists $q \geqslant 2$ *s.t. for every compact set* $\mathcal{K} \subset \mathbb{R}^d$,

$$\sup_{x\in\mathcal{K}} \mathbb{E}_\xi \|\nabla f(x,\xi)\|^q < \infty \quad \text{and} \quad \sum_n \gamma_n^{1+q/2} < \infty\,.$$

ii) For every compact set $\mathcal{K} \subset \mathbb{R}^d$, *there exists a real* $\sigma_{\mathcal{K}} \neq 0$ *s.t.*

$$\mathbb{E}_\xi \exp\langle u, \nabla f(x,\xi) - \nabla F(x)\rangle \mathbb{1}_{x\in\mathcal{K}} \leqslant \exp\left(\sigma_{\mathcal{K}}^2\|u\|^2/2\right)\,,$$

for every $x, u \in \mathbb{R}^d$. *Moreover, for every* $\alpha > 0$, $\sum_n \exp(-\alpha/\gamma_n) < \infty$.

Theorem 3.3.3. *Let Assumptions 3.2.2, 3.3.1, 3.3.2, 3.3.5 and 3.3.7 hold true. Assume that the random sequence* $(y_n = (m_n, x_n) : n \in \mathbb{N})$ *given by Algorithm 2 is bounded with probability one. Then, w.p.1, the sequence* (y_n) *converges towards the set* $\bar{\Upsilon}$ *defined in Equation* (3.3). *If, in addition, the set of critical points of the objective function* F *is finite or countable, then w.p.1, the sequence* (y_n) *converges to a single point of* $\bar{\Upsilon}$.

The almost sure boundedness of the sequence (y_n) is handled in what follows.

Theorem 3.3.4. *Let Assumptions 3.2.2, 3.3.1, 3.3.2 and 3.3.6 hold. Then, the sequence* $(y_n = (m_n, x_n) : n \in \mathbb{N})$ *given by Algorithm 2 is bounded with probability one.*

Remark 14. *Assumption 3.3.4-i) in Theorem 3.3.2 is not needed for Theorem 3.3.4.*

3.3.3 Central Limit Theorem

In this section, we establish a conditional central limit theorem for Algorithm 1.

Assumption 3.3.8. *Let $x_\star \in \operatorname{zer} \nabla F$. The following holds.*

i) F is twice continuously differentiable on a neighborhood of $x_\star$ and the Hessian $\nabla^2 F(x_\star)$ is positive definite.

ii) S is continuously differentiable on a neighborhood of $x_\star$.

iii) There exists $M > 0$ and $b_M > 4$ s.t.

$$\sup_{x \in B(x_\star, M)} \mathbb{E}_\xi[\|\nabla f(x, \xi)\|^{b_M}] < \infty \,. \tag{3.7}$$

Under Assumptions 3.2.4-*i)* to *iii)*, it follows from Equation (3.5) that the sequences $(h_n), (r_n), (p_n)$ and (q_n) of nonnegative reals converge respectively to $h_\infty, r_\infty, p_\infty$ and q_∞ where h_∞, r_∞ and q_∞ are supposed positive. Define $v_\star \triangleq q_\infty^{-1} p_\infty S(x_\star)$. Consider the matrix

$$V \triangleq \operatorname{diag}\left((\varepsilon + v_\star)^{\odot - \frac{1}{2}}\right) . \tag{3.8}$$

Let P be an orthogonal matrix s.t. the following spectral decomposition holds:

$$V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}} = P \operatorname{diag}(\pi_1, \cdots, \pi_d) P^{-1} \,,$$

where $\pi_1 \leqslant \cdots \leqslant \pi_d$ are the (positive) eigenvalues of $V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}}$. Define

$$\mathcal{H} \triangleq \begin{bmatrix} -r_\infty I_d & h_\infty \nabla^2 F(x_\star) \\ -V & 0 \end{bmatrix}$$

where I_d is the $d \times d$ identity matrix. Then the matrix $\mathcal{H}$ is Hurwitz. Indeed, it can be shown that the largest real part of the eigenvalues of $\mathcal{H}$ coincides with $-L$, where

$$L \triangleq \frac{r_\infty}{2} \left(1 - \sqrt{\left(1 - \frac{4 h_\infty \pi_1}{r_\infty^2}\right) \vee 0}\right) > 0 \,. \tag{3.9}$$

Assumption 3.3.9. *The sequence (γ_n) is given by $\gamma_n = \frac{\gamma_0}{n^\alpha}$ for some $\alpha \in (0, 1]$, $\gamma_0 > 0$. Moreover, if $\alpha = 1$, we assume that $\gamma_0 > \frac{1}{2(L \wedge q_\infty)}$.*

Theorem 3.3.5. *Let Assumptions 3.2.4-i) to iii), 3.3.3, 3.3.8 and 3.3.9 hold. Consider the iterates $z_n = (v_n, m_n, x_n)$ given by Algorithm 1. Set $\theta \triangleq 0$ if $\alpha < 1$ and $\theta \triangleq 1/(2\gamma_0)$ if $\alpha = 1$. Assume that the event $\{z_n \to z_\star\}$, where $z_\star = (v_\star, 0, x_\star)$, has a positive probability. Then, given that event,*

$$\frac{1}{\sqrt{\gamma_n}} \begin{bmatrix} m_n \\ x_n - x_\star \end{bmatrix} \Rightarrow \mathcal{N}(0, \Gamma) \,,$$

where $\Rightarrow$ *stands for the convergence in distribution and* $\mathcal{N}(0,\Gamma)$ *is a centered Gaussian distribution on* $\mathbb{R}^{2d}$ *with a covariance matrix* Γ *given by the unique solution to the Lyapunov equation*

$$(\mathcal{H}+\theta I_{2d})\Gamma+\Gamma(\mathcal{H}+\theta I_{2d})^T=-\begin{bmatrix}\mathrm{Cov}(h_\infty\nabla f(x_\star,\xi)) & 0\\ 0 & 0\end{bmatrix}.$$

In particular, given $\{z_n\to z_\star\}$*, the vector* $\sqrt{\gamma_n}^{-1}(x_n-x_\star)$ *converges in distribution to a centered Gaussian distribution with a covariance matrix given by:*

$$\Gamma_2=V^{\frac{1}{2}}P\left[\frac{C_{k,\ell}}{\frac{r_\infty-2\theta}{h_\infty}(\pi_k+\pi_\ell+\frac{2\theta(\theta-r_\infty)}{h_\infty})+\frac{(\pi_k-\pi_\ell)^2}{2(r_\infty-2\theta)}}\right]_{k,\ell=1\dots d}P^{-1}V^{\frac{1}{2}} \tag{3.10}$$

where $C\overset{\Delta}{=}P^{-1}V^{\frac{1}{2}}\mathbb{E}_\xi\left[\nabla f(x_\star,\xi)\nabla f(x_\star,\xi)^T\right]V^{\frac{1}{2}}P$.

A few remarks are in order.

- The matrix Γ_2 coincides with the limiting covariance matrix associated to the iterates
$$\begin{cases}m_{n+1} &= m_n+\gamma_{n+1}(h_\infty V\nabla f(x_n,\xi_{n+1})-r_\infty m_n)\\ x_{n+1} &= x_n-\gamma_{n+1}m_{n+1}\,.\end{cases}$$
This procedure can be seen as a preconditioned version of the stochastic heavy ball algorithm [Gadat *et al.* 2018] although the iterates are not implementable because of the unknown matrix V. Notice also that the limiting covariance Γ_2 depends on $v_\star$ but does not depend on the fluctuations of the sequence (v_n).

- When $h_\infty=r_\infty$ (which is the case for Adam), we recover the expression of the asymptotic covariance matrix previously provided in [Barakat & Bianchi 2021, Section 5.3] and the remarks formulated therein.

- The assumption $r_\infty>0$ is crucial to establish Theorem 3.3.5. For this reason, Theorem 3.3.5 does not generalize immediately to Algorithm 2. The study of the fluctuations of Algorithm 2 is left for future works.

3.3.4 Related works

In [Gadat *et al.* 2018], Gadat, Panloup and Saadane study the SHB algorithm, which is a noisy Euler's discretization of (ODE-1) in the situation where $\mathsf{h}=\mathsf{r}$ and $\mathsf{p}=\mathsf{q}\equiv 0$ (*i.e.*, there is no v variable). In this framework, if we set $\mathsf{h}=\mathsf{r}\equiv r>0$ in Algorithm 1 above, then Theorem 3.3.1 above recovers the analogous case in [Gadat *et al.* 2018, Theorem 2.1], which is termed as the exponential memory case. The other important case treated in [Gadat *et al.* 2018] is the case where $\mathsf{h}(t)=\mathsf{r}(t)=r/t$ for some $r>0$, referred to as the polynomially memory case. Actually, it is known that the ODE obtained for $\mathsf{h}(t)=\mathsf{r}(t)=r/t$ and $\mathsf{p}=\mathsf{q}\equiv 0$ boils down to (ODE-N) after a time variable change (see, *e.g.*, Lemma 3.5.3 below).

Nevertheless, we highlight that the stochastic algorithm that stems from this ODE and that is studied in [Gadat *et al.* 2018] is different from the S-NAG algorithm introduced above which stems from a different ODE (ODE-N). Hence, the convergence result of Theorem 3.3.3 for the S-NAG algorithm we consider is not covered by the analysis of [Gadat *et al.* 2018].

The specific case of the ADAM algorithm is analyzed in [Barakat & Bianchi 2021] in both the constant and vanishing stepsize settings (see [Barakat & Bianchi 2021, Ths. 5.2-5.4] which are the analogues of our Ths. 3.3.1-3.3.2). Note that we deal with a more general algorithm in the present chapter. Indeed, Algorithm 1 offers some freedom in the choice of the functions h, r, p, q satisfying Assumption 2.4 beyond the specific case of the ADAM algorithm studied in [Barakat & Bianchi 2021]. Apart from this generalization, we also emphasize some small improvements. Regarding Theorem 3.1, we provide noise conditions allowing to choose larger stepsizes (see Assumption 3.4 compared to [Barakat & Bianchi 2021, Assumption 4.2]). Concerning the stability result (Theorem3.3.2), we relax [Barakat & Bianchi 2021, Assumption 5.3-(iii)] which is no more needed in the present chapter (see Assumption 3.3.6) thanks to a modification of the discretized Lyapunov function used in the proof (see Section 6.4 compared to [Barakat & Bianchi 2021, Section 9.2]).

In most generality, the almost sure convergence result of the iterates of Algorithm 1 using vanishing stepsizes (Ths. 3.3.1-3.3.2) is new to the best of our knowledge. Moreover, while some recent results exist for S-NAG in the constant stepsize and for convex objective functions (see for e.g. [Assran & Rabbat 2020]), Ths. 3.3.3 and 3.3.4 which tackle the possibly non-convex setting are also new to the best of our knowledge.

In the work [Gadat & Gavra 2020] that was posted on the arXiv repository a few days after our submission, Gadat and Gavra study the specific case of the algorithm described in Equation (3.6) encompassing both ADAGRAD and RMSPROP, with the possibility to use mini-batches. For this specific algorithm, the authors establish a similar almost sure convergence result to ours [Gadat & Gavra 2020, Theorem 1] for decreasing stepsizes and derive some quantitative results bounding in expectation the gradient of the objective function along the iterations for constant stepsizes [Gadat & Gavra 2020, Theorem 2]. We highlight though that they do not consider the presence of momentum in the algorithm. Therefore, their analysis does not cover neither Algorithm 1 nor Algorithm 2.

In contrast to our analysis, some works in the literature explore the constant stepsize regime for some stochastic momentum methods either for smooth [Yan *et al.* 2018] or weakly convex objective functions [Mai & Johansson 2020]. Furthermore, concerning ADAM-like algorithms, several recent works control the minimum of the norms of the gradients of the objective function evaluated at the iterates of the algorithm over N iterations in expectation or with high probability [De *et al.* 2018, Zhou *et al.* 2018, Chen *et al.* 2018, Zou *et al.* 2019, Chen *et al.* 2019, Zaheer *et al.* 2018, Alacaoglu *et al.* 2020a, Défossez *et al.* 2020, Alacaoglu *et al.* 2020b] and establish regret bounds in the convex setting [Alacaoglu *et al.* 2020b].

Similar central limit theorems to Theorem 3.3.5 are established in the cases of

the stochastic heavy ball algorithm with exponential memory [Gadat *et al.* 2018, Theorem 2.4] and ADAM [Barakat & Bianchi 2021, Theorem 5.7]. In comparison to [Gadat *et al.* 2018], we precise that our theorem recovers their result and provides a closed formula for the asymptotic covariance matrix Γ_2. Our proof of Theorem 3.3.5 differs from the strategies adopted in [Gadat *et al.* 2018] and [Barakat & Bianchi 2021].

3.4 Avoidance of Traps

In Theorem 3.3.1 and Theorem 3.3.3 above, we established the almost sure convergence of the iterates x_n towards the set of critical points of the objective function F for both Algorithms 1 and 2. However, the landscape of F can contain what is known as "traps" for the algorithm, namely, critical points where the Hessian matrix of F has negative eigenvalues, making these critical points local maxima or saddle points. In this section, we show that the convergence of the iterates to these traps does not take place if the noise is exciting in some directions.

Starting with the contributions of Pemantle [Pemantle 1990] and Brandière and Duflo [Brandière & Duflo 1996], the numerous so-called avoidance of traps results that can be found in the literature deal with the case where the ODE that underlies the stochastic algorithm is an autonomous ODE. Obviously, this is neither the case of (ODE-1), nor of (ODE-N). To deal with this issue, we first state a general avoidance of traps result that extends [Pemantle 1990, Brandière & Duflo 1996] to a non-autonomous setting, and that has an interest of its own. We then apply this result to Algorithms 1 and 2.

3.4.1 A general avoidance-of-traps result in a non-autonomous setting

The notations in this subsection and in Sections 3.7.1–3.7.2 are independent from the rest of the chapter. We recall that for a function $h : \mathbb{R}^d \to \mathbb{R}^{d'}$, we denote by $\partial_i^k h(x_1, \ldots, x_d)$ the k^{th} partial derivative of the function h with respect to x_i.

The setting of our problem is as follows. Given an integer $d > 0$ and a continuous function $b : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$, we consider a stochastic algorithm built around the non-autonomous ODE $\dot{\mathsf{z}}(t) = b(\mathsf{z}(t), t)$. Let $z_\star \in \mathbb{R}^d$, and assume that on $\mathcal{V} \times \mathbb{R}_+$ where $\mathcal{V}$ is a certain neighborhood of $z_\star$, the function b can be developed as

$$b(z, t) = D(z - z_\star) + e(z, t), \tag{3.11}$$

where $e(z_\star, \cdot) \equiv 0$, and where the matrix $D \in \mathbb{R}^{d \times d}$ is assumed to admit the following spectral factorization: Given $0 \leqslant d^- < d$ and $0 < d^+ \leqslant d$ with $d^- + d^+ = d$, we can write

$$D = Q \Lambda Q^{-1}, \quad \Lambda = \begin{bmatrix} \Lambda^- & \\ & \Lambda^+ \end{bmatrix}, \tag{3.12}$$

where the Jordan blocks that constitute $\Lambda^- \in \mathbb{R}^{d^- \times d^-}$ (respectively $\Lambda^+ \in \mathbb{R}^{d^+ \times d^+}$) are those that contain the eigenvalues λ_i of D for which $\Re\lambda_i \leqslant 0$ (respectively

$\Re\lambda_i > 0$). Since $d^+ > 0$, the point $z_\star$ is an unstable equilibrium point of the ODE $\dot{\mathsf{z}}(t) = b(\mathsf{z}(t), t)$, in the sense that the ODE solution will only be able to converge to $z_\star$ along a specific so-called invariant manifold which precise characterization will be given in Section 3.7.1 below.

We now consider a stochastic algorithm that is built around this ODE. The condition $d^+ > 0$ makes that $z_\star$ is a trap that the algorithm should desirably avoid. The following theorem states that this will be the case if the noise term of the algorithm is omnidirectional enough. The idea is to show that the case being, the algorithm trajectories will move away from the invariant manifold mentioned above.

Theorem 3.4.1. *Given a sequence (γ_n) of nonnegative deterministic stepsizes such that $\sum_n \gamma_n = +\infty$, $\sum_n \gamma_n^2 < +\infty$, and a filtration $(\mathscr{F}_n)$, consider the stochastic approximation algorithm in $\mathbb{R}^d$*

$$z_{n+1} = z_n + \gamma_{n+1} b(z_n, \tau_n) + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1}\rho_{n+1}$$

where $\tau_n = \sum_{k=1}^n \gamma_k$. Assume that the sequences (η_n) and (ρ_n) are adapted to $\mathscr{F}_n$, and that z_0 is $\mathscr{F}_0$–measurable. Assume that there exists $z_\star \in \mathbb{R}^d$ such that Equation (3.11) holds true on $\mathcal{V} \times \mathbb{R}_+$, where $\mathcal{V}$ is a neighborhood of $z_\star$. Consider the spectral factorization (3.12), and assume that $d^+ > 0$. Assume moreover that the function e at the right hand side of Equation (3.11) satisfies the conditions:

i) $e(z_\star, \cdot) \equiv 0$.

ii) On $\mathcal{V} \times \mathbb{R}_+$, the functions $\partial_2^n \partial_1^k e(z,t)$ exist and are continuous for $0 \leqslant n < 2$ and $0 \leqslant k + n \leqslant 2$.

iii) The following convergence holds :

$$\lim_{(z,t)\to(z_\star,\infty)} \|\partial_1 e(z,t)\| = 0\,. \tag{3.13}$$

iv) There exist $t_0 > 0$ and a neighborhood $\mathcal{W} \subset \mathbb{R}^d$ of $z_\star$ s.t.

$$\sup_{z\in\mathcal{W}, t\geqslant t_0} \|\partial_2 e(z,t)\| < +\infty \quad \text{and} \quad \sup_{z\in\mathcal{W}, t\geqslant t_0} \left\|\partial_1^2 e(z,t)\right\| < +\infty\,.$$

Moreover, suppose that :

v) $\sum_n \|\rho_{n+1}\|^2 \mathbb{1}_{z_n\in\mathcal{W}} < \infty$ *almost surely.*

vi) $\limsup \mathbb{E}[\|\eta_{n+1}\|^4 \,|\, \mathscr{F}_n]\mathbb{1}_{z_n\in\mathcal{W}} < \infty$, *and* $\mathbb{E}[\eta_{n+1} \,|\, \mathscr{F}_n]\mathbb{1}_{z_n\in\mathcal{W}} = 0$.

vii) Writing $\tilde{\eta}_n = Q^{-1}\eta_n = (\tilde{\eta}_n^-, \tilde{\eta}_n^+)$ with $\tilde{\eta}_n^\pm \in \mathbb{R}^{d^\pm}$, for some $c^2 > 0$, it holds that

$$\liminf \mathbb{E}[\|\tilde{\eta}_{n+1}^+\|^2 \,|\, \mathscr{F}_n]\mathbb{1}_{z_n\in\mathcal{W}} \geqslant c^2 \mathbb{1}_{z_n\in\mathcal{W}}\,.$$

Then, $\mathbb{P}([z_n \to z_\star]) = 0$.

Remark 15. *Assumptions i) to iv) of Theorem 3.4.1 are related to the function e defined in Equation (3.11), which can be seen as a non-autonomous perturbation of the autonomous linear ODE $\dot{\mathsf{z}}(t) = D(\mathsf{z}(t) - z_\star)$. These assumptions guarantee the existence of a local (around the unstable equilibrium $z_\star$) non-autonomous invariant manifold of the non-autonomous ODE $\dot{\mathsf{z}}(t) = b(\mathsf{z}(t), t)$ with enough regularity properties, as provided by Proposition 3.7.1 and Proposition 3.7.3 below.*

3.4.2 Application to the stochastic algorithms

3.4.2.1 Trap avoidance of the general algorithm 1

In Theorem 3.3.1 above, we showed that the sequence (z_n) generated by Algorithm 1 converges almost surely towards the set Υ defined in Equation (3.2). Our purpose now is to show that the traps in Υ (to be characterized below) are avoided by the stochastic algorithm 1 under a proper omnidirectionality assumption on the noise.

Our first task is to write Algorithm 1 in a manner compatible with the statement of Theorem 3.4.1. The following decomposition holds for the sequence $(z_n = (v_n, m_n, x_n), n \in \mathbb{N})$ generated by this algorithm:

$$z_{n+1} = z_n + \gamma_{n+1} g(z_n, \tau_n) + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1}\tilde{\rho}_{n+1},$$

where $\tilde{\rho}_{n+1} = \left(0\,, 0\,, \frac{m_n}{\sqrt{v_n+\varepsilon}} - \frac{m_{n+1}}{\sqrt{v_{n+1}+\varepsilon}}\right)$, and where η_{n+1} is the martingale increment with respect to the filtration $(\mathscr{F}_n)$ which is defined by Equation (3.28).

Observe from Equation (3.2) that each $z_\star \in \Upsilon$ is written as $z_\star = (v_\star, 0, x_\star)$ where $x_\star \in \operatorname{zer} \nabla F$, and $v_\star = q_\infty^{-1} p_\infty S(x_\star)$ (in particular, $x_\star$ and $z_\star$ are in a one-to-one correspondence). We need to linearize the function $g(\cdot, t)$ around $z_\star$. The following assumptions will be required.

Assumption 3.4.1. *The functions F and S belong respectively to $\mathcal{C}^3(\mathbb{R}^d, \mathbb{R})$ and $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R}_+^d)$.*

Assumption 3.4.2. *The functions $\mathsf{h}, \mathsf{r}, \mathsf{p}, \mathsf{q}$ belong to $\mathcal{C}^1((0,\infty), \mathbb{R}_+)$ and have bounded derivatives on $[t_0, +\infty)$ for some $t_0 > 0$.*

Lemma 3.4.2. *Let Assumptions 3.2.4-i) to iii), 3.4.1 and 3.4.2 hold. Let $z_\star = (v_\star, 0, x_\star) \in \Upsilon$. Then, for every $z \in \mathcal{Z}_+$ and every $t > 0$, the following decomposition holds true:*

$$g(z,t) = D(z - z_\star) + e(z,t) + c(t),$$

$$\text{where } D = \begin{bmatrix} -q_\infty I_d & 0 & p_\infty \nabla S(x_\star) \\ 0 & -r_\infty I_d & h_\infty \nabla^2 F(x_\star) \\ 0 & -V & 0 \end{bmatrix}, \quad c(t) = \begin{bmatrix} \mathsf{p}(t) S(x_\star) - \mathsf{q}(t) v_\star \\ 0 \\ 0 \end{bmatrix},$$

and the function $e(z,t)$ (defined in Section 3.7.3.1 below for conciseness) has the same properties as its analogue in the statement of Theorem 3.4.1.

Using this lemma, the algorithm iterate z_{n+1} can be rewritten as an instance of the algorithm in the statement of Theorem 3.4.1, namely,

$$z_{n+1} = z_n + \gamma_{n+1} b(z_n, \tau_n) + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1}\rho_{n+1}, \tag{3.14}$$

where in our present setting, $b(z,t) = g(z,t) - c(t) = D(z - z_\star) + e(z,t)$ and $\rho_n = c(\tau_{n-1}) + \tilde{\rho}_n$. In the following assumption, we use the well-known fact that a symmetric matrix H has the same inertia as AHA^{T} for an arbitrary invertible matrix A.

Assumption 3.4.3. *Let $x_\star \in \operatorname{zer} \nabla F$, let $v_\star = q_\infty^{-1} p_\infty S(x_\star)$, and define the diagonal matrix $V = \operatorname{diag}((v_\star + \varepsilon)^{\odot -\frac{1}{2}})$ as in (3.8). Assume the following conditions:*

i) $\sum_n (q_\infty p_n - p_\infty q_n)^2 < \infty$,

ii) The Hessian matrix $\nabla^2 F(x_\star)$ has a negative eigenvalue.

iii) There exists $\delta > 0$ such that $\sup_{x \in B(x_\star, \delta)} \mathbb{E}_\xi[\|\nabla f(x,\xi)\|^8] < \infty$.

iv) Defining Π_u as the orthogonal projector on the eigenspace of $V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}}$ that is associated with the negative eigenvalues of this matrix, it holds that

$$\Pi_u V^{\frac{1}{2}} \mathbb{E}_\xi(\nabla f(x_\star,\xi) - \nabla F(x_\star))(\nabla f(x_\star,\xi) - \nabla F(x_\star))^T V^{\frac{1}{2}} \Pi_u \neq 0.$$

Theorem 3.4.3. *Let Assumptions 3.2.4, 3.3.3, and 3.4.1, 3.4.2 hold true. Let $z_\star \in \Upsilon$ be such that Assumption 3.4.3 holds true for this $z_\star$. Then, the eigenspace associated with the eigenvalues of D with positive real parts has the same dimension as the eigenspace of $\nabla^2 F(x_\star)$ associated with the negative eigenvalues of this matrix. Let $(z_n = (v_n, m_n, x_n) : n \in \mathbb{N})$ be the random sequence generated by Algorithm 1 with stepsizes satisfying $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < +\infty$. Then, $\mathbb{P}([z_n \to z_\star]) = 0$.*

The assumptions and the result call for some comments.

Remark 16. *The definition of a trap as regards the general algorithm in the statement of Theorem 3.4.1 is that the matrix D in Equation (3.11) has eigenvalues with positive real parts. Theorem 3.4.3 states that this condition is equivalent to $\nabla^2 F(x_\star)$ having negative eigenvalues. What's more, the dimension of the invariant subspace of D corresponding to the eigenvalues with positive real parts is equal to the dimension of the negative eigenvalue subspace of $\nabla^2 F(x_\star)$. Thus, Assumption 3.4.3–iv) provides the "largest" subspace where the noise energy must be non zero for the purpose of avoiding the trap.*

Remark 17. *Assumptions 3.4.2 and 3.4.3-i) are satisfied by many widely studied algorithms, among which* RMSProp *and* Adam.

Remark 18. *The results of Theorem 3.4.3 can be straightforwardly adapted to the case of (ODE-1′). Assumption 3.4.3-iv) on the noise is unchanged.*

In the case of the S-NAG algorithm, the assumptions become particularly simple. We state the afferent result separately.

3.4.2.2 Trap avoidance for S-NAG

Assumption 3.4.4. *Let $x_\star \in \operatorname{zer} \nabla F$ and let the following conditions hold.*

i) The Hessian matrix $\nabla^2 F(x_\star)$ has a negative eigenvalue.

ii) There exists $\delta > 0$ such that $\sup_{x \in B(x_\star, \delta)} \mathbb{E}_\xi[\|\nabla f(x, \xi)\|^4] < \infty$.

iii) $\tilde{\Pi}_u \mathbb{E}_\xi(\nabla f(x_\star, \xi) - \nabla F(x_\star))(\nabla f(x_\star, \xi) - \nabla F(x_\star))^T \tilde{\Pi}_u \neq 0$, where $\tilde{\Pi}_u$ is the orthogonal projector on the eigenspace of $\nabla^2 F(x_\star)$ associated with its negative eigenvalues.

Theorem 3.4.4. *Let Assumptions 3.2.4, 3.3.1, 3.4.1 and 3.4.4 hold. Define $y_\star = (0, x_\star)$. Let $(y_n = (m_n, x_n) : n \in \mathbb{N})$ be the random sequence given by Algorithm 2 with stepsizes satisfying $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < +\infty$. Then, $\mathbb{P}([y_n \to y_\star]) = 0$.*

3.4.3 Related works

Up to our knowledge, all the avoidance of traps results that can be found in the literature, starting from [Pemantle 1990, Brandière & Duflo 1996], refer to stochastic algorithms that are discretizations of autonomous ODE's (see for e.g., [Benaïm 1999, Sec. 9] for general Robbins Monro algorithms and [Mertikopoulos *et al.* 2020a, Sec. 4.3] for SGD). In this line of research, a powerful class of techniques relies on Poincaré's invariant manifold theorem for an autonomous ODE in a neighborhood of some unstable equilibrium point. In our work, we extend the avoidance of traps results to a non-autonomous setting, by borrowing a non-autonomous version of Poincaré's theorem from the rich literature that exists on the subject [Dalec′kiĭ & Kreĭn 1974, Kloeden & Rasmussen 2011].

In [Gadat *et al.* 2018], the authors succeeded in establishing an avoidance of traps result for their non-autonomous stochastic algorithm which is close to our S-NAG algorithm (see the discussion at the end of Section 3.3.4 above), at the expense of a sub-Gaussian assumption on the noise and a rather stringent assumption on the stepsizes. The main difficulty in the approach of [Gadat *et al.* 2018] lies in the use of the classical autonomous version of Poincaré's theorem (see [Gadat *et al.* 2018, Remark 2.1]). This kind of difficulty is avoided by our approach, which allows to obtain avoidance of traps results with close to minimal assumptions. More recently, in the contribution of [Gadat & Gavra 2020] discussed in Sec. 3.3.4, the authors establish an avoidance of traps result ([Gadat & Gavra 2020, Theorem 3]) for the algorithm described in Equation (3.6) using techniques inspired from [Pemantle 1990, Benaïm 1999]. As previously mentioned, this recent work does not handle momentum and hence neither Algorithm 1 nor Algorithm 2. Moreover, as indicated in our discussion of [Gadat *et al.* 2018], our strategy of proof is different.

Taking another point of view as concerns the trap avoidance, some recent works [Lee *et al.* 2019, Du *et al.* 2017, Jin *et al.* 2017, Panageas & Piliouras 2017, Panageas *et al.* 2019] address the problem of escaping saddle points when the algorithm is deterministic but when the initialization point is random. In contrast to this line of research, our

work considers a stochastic algorithm for which randomness enters into play at each iteration of the algorithm via noisy gradients.

3.5 Proofs for Section 3.2

3.5.1 Proof of Theorem 3.2.1

The arguments of the proof of this theorem that we provide here follow the approach of [Belotto da Silva & Gazeau 2020] with some small differences. Close arguments can be found in [Barakat & Bianchi 2021]. We provide the proof here for completeness and in preparation of the proofs that will be related with the stochastic algorithms.

3.5.1.1 Existence and uniqueness

The following lemma guarantees that the term $\sqrt{\mathsf{v}(t)+\varepsilon}$ in (ODE-1) is well-defined.

Lemma 3.5.1. *Let $t_0 \in \mathbb{R}_+$ and $T \in (t_0, \infty]$. Assume that there exists a solution $\mathsf{z}(t) = (\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t))$ to (ODE-1) on $[t_0, T)$ for which $\mathsf{v}(t_0) \geqslant 0$. Then, for all $t \in [t_0, T)$, $\mathsf{v}(t) \geqslant 0$.*

Proof. Assume that $\nu \stackrel{\Delta}{=} \inf\{t \in [t_0, T), \mathsf{v}(t) < 0\}$ satisfies $\nu < T$. If $\mathsf{v}(t_0) > 0$, Gronwall's lemma implies that $\mathsf{v}(t) \geqslant \mathsf{v}(t_0)\exp(-\int_{t_0}^{t} q(t))$ on $[t_0, \nu]$ which is in contradiction with the fact that $\mathsf{v}(\nu) = 0$. If $\mathsf{v}(t_0) = 0$, since $\nu < T$, there exists $t_1 \in (t_0, \nu)$ s.t. $\dot{\mathsf{v}}(t_1) < 0$. Hence, using the first equation from (ODE-1), we obtain $\mathsf{v}(t_1) > 0$. This brings us back to the first case, replacing t_0 by t_1. □

Recall that $F_\star = \inf F$ is finite by Assumption 3.2.2. Of prime importance in the proof will be the energy (Lyapunov) function $\mathcal{E} : \mathbb{R}_+ \times \mathcal{Z}_+ \to \mathbb{R}$, defined as

$$\mathcal{E}(h, z) = h(F(x) - F_\star) + \frac{1}{2}\left\| \frac{m}{(v+\varepsilon)^{\odot\frac{1}{4}}} \right\|^2, \tag{3.15}$$

for every $h \geqslant 0$ and every $z = (v, m, x) \in \mathcal{Z}_+$. This function is slightly different from its analogues that were used in [Alvarez 2000, Barakat & Bianchi 2021, Belotto da Silva & Gazeau 2020].

Consider $(t, z) \in (0, +\infty) \times \mathcal{Z}_+$ and set $z = (v, m, x)$. Then, using Assump-

tion 3.2.1, we can write

$$\begin{aligned}
&\partial_t \mathcal{E}(\mathsf{h}(t), z) + \langle \nabla_z \mathcal{E}(\mathsf{h}(t), z), g(z,t) \rangle \\
&= \dot{\mathsf{h}}(t)(F(x) - F_\star) - \frac{1}{4} \langle \frac{m^{\odot 2}}{(v+\varepsilon)^{\odot \frac{3}{2}}}, \mathsf{p}(t) S(x) - \mathsf{q}(t) v \rangle \\
&\quad + \langle \frac{m}{(v+\varepsilon)^{\odot \frac{1}{2}}}, \mathsf{h}(t) \nabla F(x) - \mathsf{r}(t) m \rangle - \langle \frac{m}{(v+\varepsilon)^{\odot \frac{1}{2}}}, \mathsf{h}(t) \nabla F(x) \rangle \\
&\leqslant - \left(\mathsf{r}(t) - \frac{\mathsf{q}(t)}{4} \right) \left\| \frac{m}{(v+\varepsilon)^{\odot \frac{1}{4}}} \right\|^2 + \dot{\mathsf{h}}(t)(F(x) - F_\star) - \frac{\mathsf{p}(t)}{4} \langle S(x), \frac{m^{\odot 2}}{(v+\varepsilon)^{\odot \frac{3}{2}}} \rangle .
\end{aligned} \tag{3.16}$$

With the help of this function, we can now establish the existence, the uniqueness and the boundedness of the solution of (ODE-1) on $[t_0, \infty)$ for an arbitrary $t_0 > 0$.

Lemma 3.5.2. *For each $t_0 > 0$ and $z_0 \in \mathcal{Z}_+$, (ODE-1) has a unique solution on $[t_0, \infty)$ starting at $\mathsf{z}(t_0) = z_0$. Moreover, the orbit $\{\mathsf{z}(t) : t \geqslant t_0\}$ is bounded.*

Proof. Let $t_0 > 0$, and fix $z_0 \in \mathcal{Z}_+$. On each set of the type $[t_0, t_0 + A] \times \bar{B}(z_0, R)$ where $A, R > 0$ and $\bar{B}(z_0, R) \subset (-\varepsilon, \infty)^d \times \mathbb{R}^d \times \mathbb{R}^d$, we easily obtain from our assumptions that the function g defined in (3.1) is continuous, and that $g(\cdot, t)$ is uniformly Lipschitz on $t \in [t_0, t_0 + A]$. In these conditions, Picard's theorem asserts that (ODE-1) starting from $\mathsf{z}(t_0) = z_0$ has a unique solution on a certain maximal interval $[t_0, T)$. Lemma 3.5.1 shows that $\mathsf{v}(t) \geqslant 0$ on this interval.

Let us show that $T = \infty$. Applying Inequality (3.16) with $(v, m, x) = (\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t))$ and using Assumption 3.2.4, we obtain that the function $t \mapsto \mathcal{E}(\mathsf{h}(t), \mathsf{z}(t))$ is decreasing on $[t_0, T)$. By the coercivity of F (Assumption 3.2.2) and Assumption 3.2.4–*i)*, we get that the trajectory $\{\mathsf{x}(t)\}$ is bounded. Recall the equation $\dot{\mathsf{m}}(t) = \mathsf{h}(t)\nabla F(\mathsf{x}(t)) - \mathsf{r}(t)\mathsf{m}(t)$. Using the continuity of the functions ∇F, h and r along with Gronwall's lemma, we get that $\{\mathsf{m}(t)\}$ is bounded if $T < \infty$. We can show a similar result for $\{\mathsf{v}(t)\}$. Thus, $\{\mathsf{z}(t)\}$ is bounded on $[t_0, T)$ if $T < \infty$ which is a contradiction, see, *e.g.*, [Hartman 2002, Cor.3.2].

It remains to show that the trajectory $\{\mathsf{z}(t)\}$ is bounded. To that end, let us apply the variation of constants method to the equation $\dot{\mathsf{m}}(t) = \mathsf{h}(t)\nabla F(\mathsf{x}(t)) - \mathsf{r}(t)\mathsf{m}(t)$. Writing $R(t) = \int_{t_0}^t \mathsf{r}(u)\, du$, we get that

$$\frac{d}{dt}\left(e^{R(t)} \mathsf{m}(t) \right) = e^{R(t)} \mathsf{h}(t) \nabla F(\mathsf{x}(t)).$$

Therefore, for every $t \geqslant t_0$,

$$\mathsf{m}(t) = e^{-R(t)} \mathsf{m}(t_0) + \int_{t_0}^t e^{R(u) - R(t)} \mathsf{h}(u) \nabla F(\mathsf{x}(u)) du \,.$$

Using the continuity of ∇F together with the boundedness of x, Assumption 3.2.4 and the triangle inequality, we obtain the existence of a constant $C > 0$ independent

of t s.t.

$$\|\mathsf{m}(t) - \mathsf{m}(t_0)\| - \|\mathsf{m}(t_0)\| \leqslant C\mathsf{h}(t_0) \int_{t_0}^{t} e^{-\int_u^t \mathsf{r}(s)\,ds} du \leqslant C\mathsf{h}(t_0) \int_{t_0}^{t} e^{-r_\infty (t-u)} du \leqslant \frac{C\mathsf{h}(t_0)}{r_\infty}.$$

The same reasoning applies to $\mathsf{v}(t)$ using the continuity of S and Assumption 3.2.4. This completes the proof. □

We can now extend this solution to $t_0 = 0$ along the approach of [Belotto da Silva & Gazeau 2020], where the detailed derivations can be found. The idea is to replace $\mathsf{h}(t)$ with $\mathsf{h}(\max(\eta, t))$ for some $\eta > 0$ and to do the same for p, q, and r. It is then easy to see that the ODE that is obtained by doing these replacements has a unique global solution on $\mathbb{R}_+$. By making $\eta \to 0$ and by using the Arzelà-Ascoli theorem along with Assumption 3.2.5, we obtain that (ODE-1) has a unique solution on $\mathbb{R}_+$.

3.5.1.2 Convergence

The first step in this part consists in transforming (ODE-1) into an autonomous ODE by including the time variable into the state vector. More specifically, we start with the following ODE:

$$\begin{bmatrix} \dot{\mathsf{z}}(t) \\ \dot{u}(t) \end{bmatrix} = \begin{bmatrix} g(\mathsf{z}(t), u(t)) \\ 1 \end{bmatrix} \quad \text{with} \quad \begin{bmatrix} \mathsf{z}(0) \\ u(0) \end{bmatrix} = \begin{bmatrix} z_0 \\ t_0 \end{bmatrix},$$

then, we perform the following change of variable in time

$$\begin{bmatrix} z \\ u \end{bmatrix} \mapsto \begin{bmatrix} z \\ s = 1/u \end{bmatrix}$$

allowing the solution to lie in a compact set.

We initialize the above ODE at a time instant $t_0 > 0$. Define the functions $\mathsf{H}, \mathsf{R}, \mathsf{P}, \mathsf{Q} : \mathbb{R}_+ \to \mathbb{R}_+$ by setting $\mathsf{H}(s) = \mathsf{h}(1/s)$, $\mathsf{R}(s) = \mathsf{r}(1/s)$, $\mathsf{P}(s) = \mathsf{p}(1/s)$; $\mathsf{Q}(s) = \mathsf{q}(1/s)$ for $s > 0$; $\mathsf{H}(0) = h_\infty$, $\mathsf{R}(0) = r_\infty$, $\mathsf{P}(0) = p_\infty$ and $\mathsf{Q}(0) = q_\infty$. Our autonomous dynamical system can then be described by the following system of equations:

$$\begin{cases} \dot{\mathsf{v}}(t) &= \mathsf{P}(\mathsf{s}(t))S(\mathsf{x}(t)) - \mathsf{Q}(\mathsf{s}(t))\mathsf{v}(t) \\ \dot{\mathsf{m}}(t) &= \mathsf{H}(\mathsf{s}(t))\nabla F(\mathsf{x}(t)) - \mathsf{R}(\mathsf{s}(t))\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\frac{\mathsf{m}(t)}{\sqrt{\mathsf{v}(t)+\varepsilon}} \\ \dot{\mathsf{s}}(t) &= -\mathsf{s}(t)^2 \end{cases} \tag{3.17}$$

Since the solution of the ODE $\dot{\mathsf{s}}(t) = -\mathsf{s}(t)^2$ for which $\mathsf{s}(t_0) = 1/t_0$ is $\mathsf{s}(t) = 1/t$, the trajectory $\{\mathsf{s}(t)\}$ is bounded. The three remaining equations are a reformulation of (ODE-1) for which the trajectories have already been shown to exist and to be bounded in Lemma 3.5.2. In the sequel, we denote by $\Phi : \mathcal{Z}_+ \times \mathbb{R}_+ \to \mathcal{Z}_+ \times \mathbb{R}_+$ the

semiflow induced by the autonomous ODE (3.17), *i.e.*, for every $u = (z, s) \in \mathcal{Z}_+ \times \mathbb{R}_+$, $\Phi(u, \cdot)$ is the unique global solution to the autonomous ODE (3.17) initialized at u. Observe that the orbits of this semiflow are precompact. Moreover, the function $\Phi((z, 0), \cdot)$ is perfectly defined for each $z \in \mathcal{Z}_+$ since the associated solution satisfies the ODE (3.19) defined below, which three first equations satisfy the hypotheses of Lemma 3.5.2.

Consider now a continuous function $V : \mathcal{Z}_+ \times \mathbb{R}_+ \to \mathbb{R}$ defined by:

$$V(u) = \mathcal{E}\left(\mathsf{H}(s), z\right), \quad u = (z, s) \in \mathcal{Z}_+ \times (0, \infty).$$

As for Inequality (3.16) above, we have here that

$$\frac{d}{dt} V\left(\Phi(u,t)\right) \leqslant -\left(\mathsf{r}(t) - \frac{\mathsf{q}(t)}{4}\right) \left\| \frac{\mathsf{m}(t)}{\left(\mathsf{v}(t) + \varepsilon\right)^{\odot \frac{1}{4}}} \right\|^2 + \dot{\mathsf{h}}(t)(F(\mathsf{x}(t)) - F_\star) - \frac{\mathsf{p}(t)}{4} \langle S(\mathsf{x}(t)), \frac{\mathsf{m}(t)^{\odot 2}}{\left(\mathsf{v}(t) + \varepsilon\right)^{\odot \frac{3}{2}}} \rangle$$

if $s > 0$, and the same inequality with $(\dot{\mathsf{h}}(t), \mathsf{p}(t), \mathsf{r}(t), \mathsf{q}(t))$ being replaced with $(0, p_\infty, r_\infty, q_\infty)$ otherwise.

Since $V \circ \Phi(u, \cdot)$ is non-increasing and nonnegative, we can define $V_\infty \triangleq \lim_{t\to\infty} V(\Phi(u,t))$. Let $\omega(u) \triangleq \bigcap_{s>0} \overline{\bigcup_{t\geqslant s} \Phi(u,t)}$ be the ω-limit set of the semiflow Φ issued from u. Recall that $\omega(u)$ is an invariant set for the flow $\Phi(u, \cdot)$, and that

$$\operatorname{dist}(\Phi(u,t), \omega(u)) \xrightarrow[t\to\infty]{} 0,$$

see, *e.g.*, [Haraux 1991, Theorem 1.1.8]). In order to finish the proof of Theorem 3.2.1, we need to make explicit the structure of $\omega(u)$.

We know from La Salle's invariance principle that $\omega(u) \subset V^{-1}(V_\infty)$. In particular,

$$\forall y \in \omega(u),\ \forall t \geqslant 0,\ V(\Phi(y,t)) = V(y) = V_\infty \tag{3.18}$$

by the invariance of $\omega(u)$.

From ODE (3.17), we have that any $y \in \omega(u)$ is of the form $y = (z, 0)$ since $s(t) \to 0$. As a consequence, $\Phi(y, \cdot)$ is a solution to the autonomous ODE

$$\begin{cases} \dot{\mathsf{v}}(t) &= p_\infty S(\mathsf{x}(t)) - q_\infty \mathsf{v}(t) \\ \dot{\mathsf{m}}(t) &= h_\infty \nabla F(\mathsf{x}(t)) - r_\infty \mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\frac{\mathsf{m}(t)}{\sqrt{\mathsf{v}(t)+\varepsilon}} \\ \dot{\mathsf{s}}(t) &= 0\,. \end{cases} \tag{3.19}$$

The three first equations can be written in a more compact form :

$$\dot{\mathsf{z}}(t) = g_\infty(\mathsf{z}(t)) \tag{3.20}$$

where $\mathsf{z}(t) = (\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t))$, and

$$g_\infty(z) = \lim_{t\to\infty} g(z,t) = \begin{bmatrix} p_\infty S(x) - q_\infty v \\ h_\infty \nabla F(x) - r_\infty m \\ -m/\sqrt{v+\varepsilon} \end{bmatrix}$$

for each $z \in \mathcal{Z}_+$. Consider $y = (v, m, x, 0) \in \omega(u)$. Using Equation (3.18), we obtain that $dV(\Phi(y,t))/dt = 0$, which implies that

$$\Big(r_\infty - \frac{q_\infty}{4}\Big) \left\| \frac{\mathsf{m}(t)}{(\mathsf{v}(t)+\varepsilon)^{\odot\frac{1}{4}}} \right\|^2 + \frac{p_\infty}{4} \langle S(\mathsf{x}(t)), \frac{\mathsf{m}(t)^{\odot 2}}{(\mathsf{v}(t)+\varepsilon)^{\odot\frac{3}{2}}} \rangle = 0$$

for all $(\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t), 0) = \Phi(y,t)$. As a consequence, Assumption 3.2.4-*iv)* gives $\mathsf{m}(t) = m = 0$, and then, $\mathsf{x}(t) = x$ for some x s.t. $\nabla F(x) = 0$ using ODE (3.19). We now turn to showing that $\mathsf{v}(t) = v = p_\infty S(x)/q_\infty$. We have proved so far that any element $y \in \omega(u)$ is written $y = (v, 0, x, 0)$ where $\nabla F(x) = 0$. The component $\mathsf{v}(\cdot)$ of $\Phi(y,\cdot)$ is a solution to the ODE $\dot{\mathsf{v}}(t) = p_\infty S(x) - q_\infty \mathsf{v}(t)$ and is thus written

$$\mathsf{v}(t) = \frac{p_\infty S(x)}{q_\infty} + e^{-q_\infty t}\Big(v - \frac{p_\infty S(x)}{q_\infty}\Big). \tag{3.21}$$

Fixing x, let $\mathcal{S}_x$ be the section of $\omega(u)$ defined by:

$$\mathcal{S}_x\omega(u) = \left\{ y \in \omega(u) \,:\, y = (\tilde{v}, 0, x, 0)\,,\, \tilde{v} \in \mathbb{R}^d_+ \right\}.$$

As $\omega(u)$ is invariant, we have $\mathcal{S}_x\omega(u) = \mathcal{S}_x\Phi(\omega(u), t)$ for all $t \geqslant 0$. Since the set $\{\tilde{v} \in \mathbb{R}^d_+ \text{ s.t. } (\tilde{v}, 0, x, 0) \in \mathcal{S}_x\omega(u)\}$ lies in a compact, we deduce from Equation (3.21) that this set is reduced to the singleton $\{p_\infty S(x)/q_\infty\}$ and in particular $v = p_\infty S(x)/q_\infty$. Therefore, the union of ω-limit sets of the semiflow Φ induced by ODE (3.17) coincides with the set of equilibrium points of this semiflow. The latter set itself corresponds to the set of points $(z, 0)$ s.t. $z \in \operatorname{zer} g_\infty$. It remains to notice that $\Upsilon = \operatorname{zer} g_\infty$ to finish the proof.

Remark 19. *Commenting on Remark 9, the same proof works for (ODE-1′) by using the function $F - F_\star$ as a Lyapunov function. The corresponding limit set (as $t \to +\infty$) is then of the form*

$$\{\tilde{z}_\infty = (\tilde{v}_\infty, \tilde{x}_\infty) \in \mathbb{R}^d_+ \times \mathbb{R}^d \,:\, \nabla F(\tilde{x}_\infty) = 0\,, \tilde{v}_\infty = p_\infty S(\tilde{x}_\infty)/q_\infty\}.$$

Similarly, if we set $\mathsf{p} = \mathsf{q} \equiv 0$ in (ODE-1) and we keep what remains in Assumption 3.2.4, the function $\mathsf{h}(t)(F(x) - F_\star) + \frac{1}{2}\|m\|^2$ works as a Lyapunov function, and the limit set has the form $\{(0,x) : \nabla F(x) = 0\}$.

3.5.2 Proof of Theorem 3.2.2

The existence and the uniqueness of the solution to (ODE-N) have been shown in the literature. We refer to [Cabot *et al.* 2009, Proposition 2.1-2.2.c)] for an identical

statement of this result and [Su *et al.* 2016b, Theorem 1, Appendix A] for a complete proof. The boundedness of the solution follows immediately from the coercivity of F together with the fact that the function $t \mapsto F(\mathsf{x}(t)) + \frac{1}{2}\|\mathsf{m}(t)\|^2$ is nonincreasing.

Concerning the convergence statement, our proof is based on comparing the solutions of (ODE-N) to the solutions of the ODE in [Gadat *et al.* 2018, Equation (2.3)]. We first note that under a change of variable, a solution to (ODE-N) gives a solution to [Gadat *et al.* 2018, Equation (2.3)].

Lemma 3.5.3. *Let* (m, x) *be a solution to* (ODE-N). *Define* $\mathsf{y}(t) = \frac{\kappa \mathsf{m}\left(\kappa\sqrt{t}\right)}{2\sqrt{t}}$, $\mathsf{u}(t) = \mathsf{x}\left(\kappa\sqrt{t}\right)$, *with* $\kappa = \sqrt{2\alpha+2}$ *and* $\beta = \frac{\kappa^2}{4}$. *Then,* (y, u) *verifies*

$$\begin{cases} \dot{\mathsf{y}}(t) &= \frac{\beta}{t}(\nabla F(\mathsf{u}(t))) - \mathsf{y}(t)) \\ \dot{\mathsf{u}}(t) &= -\mathsf{y}(t)\,. \end{cases} \tag{3.22}$$

Proof. By simple differentiation, we get:

$$\dot{\mathsf{y}}(t) = \frac{\beta}{t}\left[\nabla F\left(\mathsf{x}(\kappa\sqrt{t})\right) - \frac{\alpha}{\kappa\sqrt{t}}\mathsf{m}\left(\kappa\sqrt{t}\right)\right] - \frac{\kappa}{4t^{\frac{3}{2}}}\mathsf{m}\left(\kappa\sqrt{t}\right) = \frac{\beta}{t}\left(\nabla F(\mathsf{u}(t)) - \mathsf{y}(t)\right),$$

$$\dot{\mathsf{u}}(t) = -\frac{\kappa}{2\sqrt{t}}\mathsf{m}\left(\kappa\sqrt{t}\right) = -\mathsf{y}(t)\,.$$

□

Consider a solution (m, x) of (ODE-N) starting at $(m_0, x_0) \in \mathbb{R}^d \times \mathbb{R}^d$. As in Section 3.5.1.2, for every $t_0 > 0$, on $[t_0, +\infty)$, we have that $(\mathsf{m}, \mathsf{x}, \mathsf{s})$ is a solution to the autonomous ODE

$$\begin{cases} \dot{\mathsf{m}}(t) &= \nabla F(\mathsf{x}(t)) - \alpha \mathsf{s}(t)\mathsf{m}(t) \\ \dot{\mathsf{x}}(t) &= -\mathsf{m}(t) \\ \dot{\mathsf{s}}(t) &= -\mathsf{s}(t)^2\,, \end{cases} \tag{3.23}$$

starting at $(m_0, x_0, 1/t_0)$. Denote by $\Phi_N = (\Phi_N^m, \Phi_N^x, \Phi_N^s)$ the semiflow induced by ODE (3.23) and $\omega_N((m_0, x_0, 1/t_0))$ its limit set.

Define (y, u) as in Lemma 3.5.3. Starting at $(\mathsf{y}(t_0), \mathsf{u}(t_0), 1/t_0)$, we also have that $(\mathsf{y}, \mathsf{u}, \mathsf{s})$ is a solution on $[t_0, +\infty)$ to the "autonomized" Heavy-Ball ODE

$$\begin{cases} \dot{\mathsf{y}}(t) &= \beta \mathsf{s}(t)(\nabla F(\mathsf{u}(t))) - \mathsf{y}(t)) \\ \dot{\mathsf{u}}(t) &= -\mathsf{y}(t) \\ \dot{\mathsf{s}}(t) &= -\mathsf{s}(t)^2\,. \end{cases} \tag{3.24}$$

Denote by $\Phi_H = (\Phi_H^y, \Phi_H^u, \Phi_H^s)$ the semiflow induced by ODE (3.24) and $\omega_H((\mathsf{y}(t_0), \mathsf{u}(t_0), 1/t_0))$ its limit set.

Lemma 3.5.4. *For any compact set* $K \subset \mathbb{R}^{2d+1}$ *and any* $T > 0$, *the family of functions* $\left\{\Phi(z, \cdot) : [0, T] \to \mathbb{R}^{2d+1}\right\}_{z \in K}$, *where* Φ *is either* Φ_H *or* Φ_N, *is relatively compact in* $(\mathcal{C}^0([0, T], \mathbb{R}^{2d+1}), \|\cdot\|_\infty)$.

Proof. The map $\Phi : \mathbb{R}^{2d+1} \times \mathbb{R}_+ \to \mathbb{R}^{2d+1}$ is continuous, hence uniformly continuous on $K \times [0,T]$. The result follows from the application of the Arzelà-Ascoli theorem to the family $\left\{\Phi(z,\cdot) : [0,T] \to \mathbb{R}^{2d+1}\right\}_{z\in K}$. □

Let $(m,x,0) \in \omega_N((m_0,x_0,1/t_0))$. There exists a sequence (t_k) of nonnegative reals such that $(m,x,0) = \lim_{k\to\infty}(\mathsf{m}(t_k),\mathsf{x}(t_k),1/t_k)$. For any $T > 0$, using Lemma 3.5.4, up to an extraction, we can say that the sequence of functions $\{\Phi_N((\mathsf{m}(t_k),\mathsf{x}(t_k),1/t_k),\cdot)\}_k$ converges towards $(\tilde{\mathsf{m}},\tilde{\mathsf{x}},0)$ in $\mathcal{C}^0([0,T],\mathbb{R}^d)$, where $(\tilde{\mathsf{m}},\tilde{\mathsf{x}})$ is a solution to

$$\begin{cases} \dot{\tilde{\mathsf{m}}}(t) &= \nabla F(\tilde{\mathsf{x}}(t)) \\ \dot{\tilde{\mathsf{x}}}(t) &= -\tilde{\mathsf{m}}(t)\,, \end{cases} \tag{3.25}$$

with $(\tilde{\mathsf{m}}(0),\tilde{\mathsf{x}}(0)) = (m,x)$. Moreover, by Lemma 3.5.3, we also have that:

$$\begin{aligned} &\sup_{h\in[0,T^2/\kappa^2]} \left\|\tilde{\mathsf{x}}(\kappa\sqrt{h}) - \Phi_N^x((\mathsf{m}(t_k),\mathsf{x}(t_k),1/t_k),\kappa\sqrt{h})\right\| \\ &\qquad = \sup_{h\in[0,T^2/\kappa^2]} \left\|\tilde{\mathsf{x}}(\kappa\sqrt{h}) - \Phi_H^u((\mathsf{m}(t_k),\mathsf{x}(t_k),1/t_k),h)\right\| \xrightarrow[k\to+\infty]{} 0\,. \end{aligned} \tag{3.26}$$

Using Lemma 3.5.4, up to an additional extraction, we get on $\mathcal{C}^0([0,T^2/\kappa^2],\mathbb{R}^{2d+1})$ that $\{\Phi_H((\mathsf{x}(t_k),\mathsf{m}(t_k),1/t_k),\cdot)\}_k$ converges to $(\mathsf{u},\mathsf{y},0)$, where (u,y) is a solution to

$$\begin{cases} \dot{\mathsf{y}}(t) &= 0 \\ \dot{\mathsf{u}}(t) &= -\mathsf{y}(t)\,. \end{cases} \tag{3.27}$$

Therefore, $\mathsf{u}(t) = A + Bt$ for some A and B in $\mathbb{R}^d$. Imagine that $B \neq 0$. We previously proved that x (and therefore u) is bounded by some constant $C > 0$. Let $T' > \frac{C+\|A\|}{\|B\|}$. Up to an extraction, we obtain that $\{\Phi_H((\mathsf{x}(t_k),\mathsf{m}(t_k),1/t_k),\cdot)\}_k$ converges to u' on $\mathcal{C}^0([0,T'],\mathbb{R}^{2d+1})$, with $\mathsf{u}'(t) = A' + B't$ for some A' and B' in $\mathbb{R}^d$. We then have by uniqueness of the limit that $A' = A$ and $B' = B$. As a consequence, $\|\mathsf{u}'(T')\| = \|A + BT'\| > C$ and we obtain a contradiction. Hence $B = 0$.

This implies that u is constant. Then, if we go back to Eqs. (3.26) and (3.25), we get that $\tilde{\mathsf{x}}$ is constant, hence $\tilde{\mathsf{m}} \equiv 0$ and then $\nabla F(\tilde{\mathsf{x}}) \equiv 0$. In particular, this means that $m = \tilde{\mathsf{m}}(0) = 0$ and $\nabla F(x) = \nabla F(\tilde{\mathsf{x}}(0)) = 0$.

3.6 Proofs for Section 3.3

3.6.1 Preliminaries

We first recall some useful definitions and results. Let Ψ represent any semiflow on an arbitrary metric space (E,d). As in the previous section, a point $z \in E$ is called an equilibrium point of the semiflow Ψ if $\Psi(z,t) = z$ for all $t \geqslant 0$. We denote by Λ_Ψ the set of equilibrium points of Ψ. A continuous function $\mathsf{V} : E \to \mathbb{R}$ is called a Lyapunov function for the semiflow Ψ if $\mathsf{V}(\Psi(z,t)) \leqslant \mathsf{V}(z)$ for all $z \in E$ and all $t \geqslant 0$. It is called a *strict* Lyapunov function if, moreover, $\{z \in E : \forall t \geqslant 0,\ \mathsf{V}(\Psi(z,t)) = \mathsf{V}(z)\} = \Lambda_\Psi$.

If V is a strict Lyapunov function for Ψ and if $z \in E$ is a point s.t. $\{\Psi(z,t) : t \geqslant 0\}$ is relatively compact, then it holds that $\Lambda_\Psi \neq \varnothing$ and $\mathsf{d}(\Psi(z,t), \Lambda_\Psi) \to 0$, see [Haraux 1991, Theorem 2.1.7]. A continuous function $z : [0, +\infty) \to E$ is said to be an asymptotic pseudotrajectory (APT, [Benaïm & Hirsch 1996]) for the semiflow Ψ if $\lim_{t\to+\infty} \sup_{s\in[0,T]} \mathsf{d}(z(t+s), \Psi(z(t), s)) = 0$ for every $T \in (0, +\infty)$.

3.6.2 Proof of Theorem 3.3.1

Recall that Φ is the semiflow induced by the autonomous ODE (3.17) which is an "autonomized" version of our initial (ODE-1). In the remainder of this section, the proof will be divided into two main steps : (a) we show that a certain continuous-time linearly interpolated process constructed from the iterates of our algorithm 1 is an APT of Φ; (b) we exhibit a strict Lyapunov function for a restriction to a carefully chosen compact set of a well chosen semiflow related to Φ. Then, we characterize the limit set of the APT using [Benaïm 1999, Theorem 5.7] and [Benaïm 1996, Proposition 3.2]. The sequence (z_n) converges almost surely to this same limit set.

(a) APT. For every $n \geqslant 1$, define $\bar{z}_n = (v_n, m_n, x_{n-1})$ (note the shift in the index of the variable x). We have the decomposition

$$\bar{z}_{n+1} = \bar{z}_n + \gamma_{n+1} g(\bar{z}_n, \tau_n) + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1}\varsigma_{n+1}\,,$$

where g is defined in Equation (3.1),

$$\eta_{n+1} = \big(p_n(\nabla f(x_n, \xi_{n+1})^{\odot 2} - S(x_n)),\, h_n(\nabla f(x_n, \xi_{n+1}) - \nabla F(x_n)),\, 0\big)\,, \tag{3.28}$$

is a martingale increment and where we set $\varsigma_{n+1} = (\varsigma^v_{n+1}, \varsigma^m_{n+1}, \varsigma^x_{n+1})$ with the components defined by:

$$\begin{cases} \varsigma^v_{n+1} &= p_n(S(x_n) - S(x_{n-1})) \\ \varsigma^m_{n+1} &= h_n(\nabla F(x_n) - \nabla F(x_{n-1})) \\ \varsigma^x_{n+1} &= (\frac{\gamma_n}{\gamma_{n+1}} - 1)\frac{m_n}{\sqrt{v_n+\varepsilon}}\,. \end{cases}$$

We first prove that $\varsigma_n \to 0$ a.s. by considering the components separately. The components ς^m_{n+1} and ς^v_{n+1} converge a.s. to zero by using Assumptions 3.2.1, 3.2.3, together with the boundedness of the sequences (p_n) and (h_n) (which are both convergent). Indeed, since ∇F is locally Lipschitz continuous and the sequence (z_n) is supposed to be almost surely bounded, there exists a constant C s.t. $\|\nabla F(x_n) - \nabla F(x_{n-1})\| \leqslant C\|x_n - x_{n-1}\| \leqslant \frac{C}{\varepsilon}\gamma_n\|m_n\|$. The same inequality holds when replacing ∇F by S which is also locally Lipschitz continuous. The component ς^x_{n+1} also converges a.s. to zero by observing that $\|\varsigma^x_{n+1}\| \leqslant |1 - \frac{\gamma_n}{\gamma_{n+1}}|.\|m_n\|/\sqrt{\varepsilon}$ and using Assumption 3.3.2 together with the a.s. boundedness of (z_n). Now consider the martingale increment sequence (η_n), adapted to $\mathcal{F}_n$. Take $\delta > 0$. Since (z_n) is a.s bounded, there is a constant $C' > 0$ such that $\mathbb{P}(\sup \|x_n\| > C') \leqslant \delta$. Denoting

$\tilde{\eta}_n \overset{\Delta}{=} \eta_n \mathbb{1}_{\|x_n\| \leqslant C'}$ and combining Assumptions 3.2.4 with 3.3.4-i) we can show using convexity inequalities that

$$\sup_n \mathbb{E}\|\tilde{\eta}_{n+1}\|^q < \infty.$$

Then, we deduce from this result together with the corresponding stepsize assumption from 3.3.4-i) and [Benaïm 1999, Proposition 4.2] (see also [Métivier & Priouret 1987, Proposition 8]) the key property:

$$\forall T > 0\,, \quad \max\Big\{\Big\|\sum_{k=n}^{L-1} \gamma_{k+1}\tilde{\eta}_{k+1}\Big\| \;:\; L = n+1, \dots, J(\tau_n + T)\Big\} \xrightarrow[n\to\infty]{\text{a.s.}} 0 \tag{3.29}$$

where $J(t) = \max\{n \geqslant 0 \,:\, \tau_n \leqslant t\}$. Hence, for all $T > 0$, with probability at least $1 - \delta$:

$$\max\Big\{\Big\|\sum_{k=n}^{L-1} \gamma_{k+1}\eta_{k+1}\Big\| \;:\; L = n+1, \dots, J(\tau_n + T)\Big\} \xrightarrow[n\to\infty]{} 0\,. \tag{3.30}$$

Since δ can be chosen arbitrary small, Equation (3.30) remains true with probability 1. This result also holds under Assumption 3.3.4-ii) (instead of 3.3.4-i)) by applying [Benaïm 1999, Proposition 4.4].

Let $\boldsymbol{z} : [0, +\infty) \to \mathcal{Z}_+$ be the continous-time linearly interpolated process given by

$$\boldsymbol{z}(t) = \bar{z}_n + (t - \tau_n)\frac{\bar{z}_{n+1} - \bar{z}_n}{\gamma_{n+1}} \qquad (\forall n \in \mathbb{N}\,,\, \forall t \in [\tau_n, \tau_{n+1}))$$

(where $\tau_n = \sum_{k=1}^n \gamma_k$). Let $t_0 > 0$. Define $\boldsymbol{u} : [t_0, \infty) \to \mathcal{Z} \times (0, 1/t_0]$ by

$$\boldsymbol{u}(t) = \begin{bmatrix} \boldsymbol{z}(t) \\ 1/t \end{bmatrix}, \quad \text{for} \quad t \geqslant t_0 > 0.$$

Using Equation (3.30) and the almost sure boundedness of the sequence (z_n) along with the fact that ς_n converges a.s. to zero, it follows from [Benaïm 1999, Proposition 4.1, Remark 4.5] that $\boldsymbol{u}(t)$ is an APT of the already defined semiflow Φ induced by (3.17). Remark that it also holds that $\boldsymbol{z}(t)$ is an APT of the semiflow Φ^∞ induced by (3.20). As the trajectory of $\boldsymbol{u}(t)$ is precompact, the limit set

$$\boldsymbol{L}(\boldsymbol{u}) = \bigcap_{t \geqslant t_0} \overline{\boldsymbol{u}([t, \infty))}$$

is compact. Moreover, it has the form

$$\boldsymbol{L}(\boldsymbol{u}) = \begin{bmatrix} \boldsymbol{S} \\ 0 \end{bmatrix}, \quad \text{where} \quad \boldsymbol{S} \overset{\Delta}{=} \bigcap_{t \geqslant t_0} \overline{\boldsymbol{z}([t, \infty))}\,. \tag{3.31}$$

Our objective now is to prove that

$$\boldsymbol{S} \subset \Lambda_{\Phi^\infty}\,. \tag{3.32}$$

In order to establish this inclusion, we study the behavior of the restriction $\Phi|\boldsymbol{L}$ of the semiflow Φ to the set $\boldsymbol{L}$ (which is well-defined since $\boldsymbol{L}$ is Φ-invariant). Remark that

$$\Phi|\boldsymbol{L} = \begin{bmatrix} \Phi^{\infty}|\boldsymbol{S} \\ 0 \end{bmatrix},$$

where Φ^{∞} is the semiflow associated to (3.20). In the second part of the proof, we establish Equation (3.32) combining item (a) we just proved with [Benaïm 1999, Theorem 5.7] and [Benaïm 1999, Proposition 6.4]. In order to use the latter proposition, we prove a useful proposition in item (b).

(b) Strict Lyapunov function and convergence. For every $\delta > 0$ and every $z = (v, m, x) \in \mathcal{Z}_+$, define:

$$W_{\delta}(v, m, x) \triangleq \mathcal{E}_{\infty}(z) - \delta\langle \nabla F(x), m\rangle + \delta\|q_{\infty}v - p_{\infty}S(x)\|^2\,, \tag{3.33}$$

where, under Assumption 3.2.4-*i)*, the function $\mathcal{E}_{\infty}$ is defined by

$$\mathcal{E}_{\infty}(z) \triangleq \lim_{t\to+\infty} \mathcal{E}(t, z) = h_{\infty}(F(x) - F_{\star}) + \frac{1}{2}\left\| \frac{m}{(v+\varepsilon)^{\odot\frac{1}{4}}} \right\|^2 . \tag{3.34}$$

Proposition 3.6.1. *Let $t_0 > 0$ and let Assumptions 3.2.1 to 3.2.4 and 3.3.5 hold true. Let $\boldsymbol{S}$ be the limit set defined in Equation (3.31). Let $\overline{\Phi}^{\infty} : \boldsymbol{S} \times [t_0, +\infty) \to \boldsymbol{S}$ be the restriction of the semiflow Φ^{∞} to $\boldsymbol{S}$* i.e., *$\overline{\Phi}^{\infty}(z, t) = \Phi^{\infty}(z, t)$ for all $z \in \boldsymbol{S}, t \geqslant t_0$. Then,*

i) $\boldsymbol{S}$ is compact.

ii) $\overline{\Phi}^{\infty}$ is a well-defined semiflow on $\boldsymbol{S}$.

iii) The set of equilibrium points of $\overline{\Phi}^{\infty}$ is equal to $\Lambda_{\Phi^{\infty}} \cap \boldsymbol{S}$.

iv) There exists $\delta > 0$ s.t. W_{δ} is a strict Lyapunov function for the semiflow $\overline{\Phi}^{\infty}$.

Proof. The first point is a consequence of the definition of $\boldsymbol{S}$ and the boundedness of $\boldsymbol{z}$. The second point stems from the definition of Φ^{∞}. Observing that $\overline{\Phi}^{\infty}$ is valued in $\boldsymbol{S}$, the third point is immediate from the definition of $\Lambda_{\Phi^{\infty}}$. We now prove the last point. Consider $z \in \boldsymbol{S}$ and write $\overline{\Phi}^{\infty}(z, t)$ under the form $\overline{\Phi}^{\infty}(z, t) = (\mathsf{v}(t), \mathsf{m}(t), \mathsf{x}(t))$. Notice that this quantity is bounded as a function of the variable t. For *any* map $\mathsf{W} : \mathcal{Z}_+ \to \mathbb{R}$, define for all $t \geqslant t_0$, $\mathcal{L}_{\mathsf{W}}(t) \triangleq \limsup_{s\to 0} s^{-1}(\mathsf{W}(\overline{\Phi}^{\infty}(z, t+s)) - \mathsf{W}(\overline{\Phi}^{\infty}(z, t)))$. Introduce $G(z) \triangleq -\langle \nabla F(x), m\rangle$ and $H(z) \triangleq \|q_{\infty}v - p_{\infty}S(x)\|^2$ for every $z = (v, m, x) \in \mathcal{Z}_+$. Consider $\delta > 0$ (to be specified later on). We study $\mathcal{L}_{W_{\delta}} = \mathcal{L}_{\mathcal{E}_{\infty}} + \delta\mathcal{L}_G + \delta\mathcal{L}_H$. Note that $\overline{\Phi}^{\infty}(z, t) \in \boldsymbol{S} \cap \mathcal{Z}_+$ for all $t \geqslant t_0$ by an analogous result to Lemma 3.5.1 for Φ^{∞}. Thus, $t \mapsto \mathcal{E}_{\infty}(\overline{\Phi}^{\infty}(z, t))$ is differentiable at any point $t \geqslant t_0$ and $\mathcal{L}_{\mathcal{E}_{\infty}}(t) = \frac{d}{dt}\mathcal{E}_{\infty}(\overline{\Phi}^{\infty}(z, t))$. Using similar derivations to Inequality (3.16), we obtain that

$$\mathcal{L}_{\mathcal{E}_{\infty}}(t) \leqslant -\left(r_{\infty} - \frac{q_{\infty}}{4}\right)\left\| \frac{\mathsf{m}(t)}{(\mathsf{v}(t)+\varepsilon)^{\odot\frac{1}{4}}} \right\|^2 . \tag{3.35}$$

We now study $\mathcal{L}_G$. For every $t \geqslant t_0$,

$$\begin{aligned}\mathcal{L}_G(t) &= \limsup_{s\to 0} s^{-1}(-\langle \nabla F(\mathsf{x}(t+s)), \mathsf{m}(t+s)\rangle + \langle \nabla F(\mathsf{x}(t)), \mathsf{m}(t)\rangle)\\ &\leqslant \limsup_{s\to 0} s^{-1}\|\nabla F(\mathsf{x}(t)) - \nabla F(\mathsf{x}(t+s))\|\|\mathsf{m}(t+s)\| - \langle \nabla F(\mathsf{x}(t)), \dot{\mathsf{m}}(t)\rangle\,.\end{aligned}$$

Let $L_{\nabla F}$ be the Lipschitz constant of ∇F on the bounded set $\{x : (v,m,x) \in \boldsymbol{S}\}$. Define $C_1 \triangleq \sup_t \|\sqrt{\mathsf{v}(t)+\varepsilon}\|$. Then,

$$\begin{aligned}\mathcal{L}_G(t) &\leqslant L_{\nabla F}\limsup_{s\to 0} s^{-1}\|\mathsf{x}(t)-\mathsf{x}(t+s)\|\|\mathsf{m}(t+s)\| - \langle \nabla F(\mathsf{x}(t)), \dot{\mathsf{m}}(t)\rangle\\ &\leqslant L_{\nabla F}\|\dot{\mathsf{x}}(t)\|\|\mathsf{m}(t)\| - \langle \nabla F(\mathsf{x}(t)), \dot{\mathsf{m}}(t)\rangle\\ &\leqslant L_{\nabla F}\|\dot{\mathsf{x}}(t)\|\|\mathsf{m}(t)\| - h_\infty\|\nabla F(\mathsf{x}(t))\|^2 + r_\infty\langle \nabla F(\mathsf{x}(t)), \mathsf{m}(t)\rangle\\ &\leqslant \left(\frac{L_{\nabla F}C_1^{\frac{1}{2}}}{\varepsilon^{\frac{1}{4}}} + \frac{r_\infty C_1}{2u_1^2}\right)\left\|\frac{\mathsf{m}(t)}{(\mathsf{v}(t)+\varepsilon)^{\odot\frac{1}{4}}}\right\|^2 - \left(h_\infty - \frac{r_\infty u_1^2}{2}\right)\|\nabla F(\mathsf{x}(t))\|^2\end{aligned} \tag{3.36}$$

where we used the classical inequality $|\langle a,b\rangle| \leqslant \|a\|^2/(2u^2) + u^2\|b\|^2/2$ for any non-zero real u to derive the last above inequality. We now study $\mathcal{L}_H$. For every $t \geqslant t_0$,

$$\begin{aligned}\mathcal{L}_H(t) &= \limsup_{s\to 0} s^{-1}(\|q_\infty \mathsf{v}(t+s) - p_\infty S(\mathsf{x}(t+s))\|^2 - \|q_\infty \mathsf{v}(t) - p_\infty S(\mathsf{x}(t))\|^2)\\ &= \limsup_{s\to 0} s^{-1}(p_\infty^2\|S(\mathsf{x}(t)) - S(\mathsf{x}(t+s))\|^2\\ &+ 2p_\infty\langle S(\mathsf{x}(t)) - S(\mathsf{x}(t+s)), q_\infty\mathsf{v}(t+s) - p_\infty S(\mathsf{x}(t))\rangle)\\ &+ \lim_{s\to 0} s^{-1}(\|q_\infty \mathsf{v}(t+s) - p_\infty S(\mathsf{x}(t))\|^2 - \|q_\infty \mathsf{v}(t) - p_\infty S(\mathsf{x}(t))\|^2)\,.\end{aligned}$$

The second term in the righthand side coincides with $-2q_\infty\langle p_\infty S(\mathsf{x}(t)) - q_\infty\mathsf{v}(t), \dot{\mathsf{v}}(t)\rangle = -2q_\infty\|p_\infty S(\mathsf{x}(t)) - q_\infty\mathsf{v}(t)\|^2$. Denote by L_S the Lipschitz constant of S on the set $\{x : (v,m,x) \in \boldsymbol{S}\}$. Note that $s^{-1}(\|S(\mathsf{x}(t+s)) - S(\mathsf{x}(t))\|^2) \leqslant L_S^2 s\|s^{-1}(\mathsf{x}(t+s) - \mathsf{x}(t))\|^2$ which converges to zero as $s \to 0$. Thus,

$$\begin{aligned}\mathcal{L}_H(t) &= -2q_\infty\|p_\infty S(\mathsf{x}(t)) - q_\infty\mathsf{v}(t)\|^2\\ &+ \limsup_{s\to 0} 2p_\infty s^{-1}\langle S(\mathsf{x}(t)) - S(\mathsf{x}(t+s)), q_\infty\mathsf{v}(t+s) - p_\infty S(\mathsf{x}(t))\rangle\\ &\leqslant -2q_\infty\|p_\infty S(\mathsf{x}(t)) - q_\infty\mathsf{v}(t)\|^2 + 2p_\infty\|\dot{\mathsf{x}}(t)\|L_S\|q_\infty\mathsf{v}(t) - p_\infty S(x(t))\|\\ &\leqslant \frac{p_\infty}{\varepsilon^{\frac{1}{2}}u_2^2}\left\|\frac{\mathsf{m}(t)}{(\mathsf{v}(t)+\varepsilon)^{\odot\frac{1}{4}}}\right\|^2 - (2q_\infty - p_\infty u_2^2 L_S^2)\|p_\infty S(\mathsf{x}(t)) - q_\infty\mathsf{v}(t)\|^2\,.\end{aligned} \tag{3.37}$$

Recalling that $\mathcal{L}_{W_\delta} = \mathcal{L}_{\mathcal{E}_\infty} + \delta\mathcal{L}_G + \delta\mathcal{L}_H$ and combining Eqs. (3.35), (3.36) and (3.37), we obtain for every $t \geqslant t_0$,

$$\begin{aligned}\mathcal{L}_{W_\delta}(t) \leqslant -M(\delta)\left\|\frac{\mathsf{m}(t)}{(\mathsf{v}(t)+\varepsilon)^{\odot\frac{1}{4}}}\right\|^2 - \delta\left(h_\infty - \frac{r_\infty u_1^2}{2}\right)\|\nabla F(\mathsf{x}(t))\|^2\\ - \delta\left(2q_\infty - p_\infty u_2^2 L_S^2\right)\|p_\infty S(\mathsf{x}(t)) - q_\infty\mathsf{v}(t)\|^2\,.\end{aligned} \tag{3.38}$$

where $M(\delta) \triangleq r_\infty - \frac{q_\infty}{4} - \delta\left(\frac{r_\infty C_1}{2u_1^2} + \frac{L_{\nabla F} C_1^{\frac{1}{2}}}{\varepsilon^{\frac{1}{4}}} + \frac{p_\infty}{\varepsilon^{\frac{1}{2}} u_2^2}\right)$. Now select u_1, u_2 small enough s.t. $h_\infty - r_\infty u_1^2/2 > 0$ and $2q_\infty - p_\infty u_2^2 L_S^2 > 0$. Then, choose δ in such a way that $M(\delta) > 0$. Thus, there exists a constant c depending on δ s.t.

$$\forall t \geqslant t_0, \quad \mathcal{L}_{W_\delta}(t) \leqslant -c\left(\left\|\frac{\mathsf{m}(t)}{(\mathsf{v}(t)+\varepsilon)^{\odot\frac{1}{4}}}\right\|^2 + \|\nabla F(\mathsf{x}(t))\|^2 + \|p_\infty S(\mathsf{x}(t)) - q_\infty \mathsf{v}(t)\|^2\right). \tag{3.39}$$

It can easily be seen that for every $z \in \boldsymbol{S}$, $t \mapsto W_\delta(\overline{\Phi}^\infty(z,t))$ is Lipschitz continuous, hence absolutely continuous. Its derivative almost everywhere coincides with $\mathcal{L}_{W_\delta}$, which is nonpositive. Thus, W_δ is a Lyapunov function for $\overline{\Phi}^\infty$. We prove that the Lyapunov function is strict. Consider $z = (v, m, x) \in \boldsymbol{S}$ s.t. $W_\delta(\overline{\Phi}^\infty(z,t)) = W_\delta(z)$ for all $t \geqslant t_0$. The derivative almost everywhere of $t \mapsto W_\delta(\overline{\Phi}^\infty(z,t))$ is identically zero, and by Equation (3.39), this implies that

$$-c\left(\left\|\frac{\mathsf{m}(t)}{(\mathsf{v}(t)+\varepsilon)^{\odot\frac{1}{4}}}\right\|^2 + \|\nabla F(\mathsf{x}(t))\|^2 + \|p_\infty S(\mathsf{x}(t)) - q_\infty \mathsf{v}(t)\|^2\right)$$

is equal to zero for every $t \geqslant t_0$ a.e. (hence, for every $t \geqslant t_0$, by continuity of $\overline{\Phi}^\infty$). In particular for $t = t_0$, $m = \nabla F(x) = 0$ and $p_\infty S(x) - q_\infty v = 0$. Hence, $z \in \operatorname{zer} g_\infty \cap \boldsymbol{S}$. This concludes the proof since $\Lambda_{\Phi^\infty} = \operatorname{zer} g_\infty$. □

End of the Proof of Theorem 3.3.1. Finally, Assumption 3.3.5 implies that $W_\delta(\Lambda_{\Phi^\infty} \cap \boldsymbol{S})$ is of empty interior. Recall that Assumptions 3.2.1 and 3.2.3 both follow from Assumption 3.3.3 made in Theorem 3.3.1. Given Proposition 3.6.1, the proof is concluded by applying [Benaïm 1999, Proposition 6.4] to the restricted semi-flow $\bar{\Phi}^\infty$ (with $(M, \Lambda) = (\boldsymbol{S}, \Lambda_{\bar{\Phi}^\infty})$). Note that a Lyapunov function for $\Lambda_{\bar{\Phi}^\infty}$ is what is called a strict Lyapunov function. Such a function is provided by Proposition 3.6.1. We obtain as a conclusion of [Benaïm 1999, Proposition 6.4] that $\boldsymbol{S} \subset \Lambda_{\bar{\Phi}^\infty}$. This gives the desired result (Equation (3.32)) given Proposition 3.6.1-*iii)*.

The last assertion of Theorem 3.3.1 is a consequence of [Benaïm 1999, Cor. 6.6].

3.6.3 Proof of Theorem 3.3.3

We can rewrite the iterates from Algorithm 2 as follows:

$$\begin{cases} m_{n+1} &= m_n + \gamma_{n+1}(\nabla F(x_n) - \frac{\alpha}{\tau_n} m_n) + \gamma_{n+1}(\nabla f(x_n, \xi_{n+1}) - \nabla F(x_n)) \\ x_{n+1} &= x_n - \gamma_{n+1} m_{n+1}\,. \end{cases} \tag{3.40}$$

We prove that the sequence $(y_n = (m_n, x_n) : n \in \mathbb{N})$ of iterates of this algorithm converges almost surely towards the set $\bar{\Upsilon}$ defined in Equation (3.3) if it is supposed to be bounded with probability one. The proof follows a similar path to the proof in Section 3.5.2.

Indeed, denote by X and M the linearly interpolated processes constructed respectively from the sequences (x_n) and (m_n) and let $\mathsf{s}(t) = 1/t$. Recall that $\Phi_N = (\Phi_N^m, \Phi_N^x, \Phi_N^s)$ is the semiflow induced by (3.23). As in Section 3.6.2, we have that $\mathsf{Z} \triangleq (\mathsf{M}, \mathsf{X}, \mathsf{s})$ is an APT of (3.23). In particular, this means that

$$\forall T > 0\,, \quad \sup_{h\in[0,T]} \|\mathsf{X}(t+h) - \Phi_N^x(\mathsf{Z}(t), h)\| \xrightarrow[t\to\infty]{} 0\,. \tag{3.41}$$

By Lemma 3.5.3, we also have that

$$\begin{aligned}\sup_{h\in[0,T^2/\kappa^2]} \left\|\mathsf{X}(t+\kappa\sqrt{h}) - \Phi_N^x(\mathsf{Z}(t), \kappa\sqrt{h})\right\| \\ = \sup_{h\in[0,T^2/\kappa^2]} \left\|\mathsf{X}(t+\kappa\sqrt{h}) - \Phi_H^u(\mathsf{Z}(t), h)\right\| \xrightarrow[t\to\infty]{} 0\,.\end{aligned} \tag{3.42}$$

Let (m, x) be a limit point of the sequence (y_n) and let $T > 0$. Using Lemma 3.5.4, we can proceed in the same manner as in Section 3.5.2 and get a sequence (t_k) such that

$$(\mathsf{M}(t_k+\cdot), \mathsf{X}(t_k+\cdot)) \to (\mathsf{m}, \mathsf{x}) \text{ and } (\Phi_H^y(\mathsf{Z}(t_k), \cdot), \Phi_H^u(\mathsf{Z}(t_k), \cdot)) \to (\mathsf{y}, \mathsf{u})\,,$$

where $(\mathsf{m}(0), \mathsf{x}(0)) = (m, x)$, and (m, x) and (x, u) are respectively solutions to (3.25) and (3.27). As in the end of Section 3.5.2, we obtain that u and x are constant, therefore $\mathsf{m} \equiv 0$ and $\nabla F(\mathsf{x}) \equiv 0$, which finishes the proof.

3.6.4 Proof of Theorem 3.3.2

The idea of the proof is to apply Robbins-Siegmund's theorem [Robbins & Siegmund 1971] to

$$V_n = h_{n-1}F(x_n) + \frac{1}{2}\langle m_n^{\odot 2}, \frac{1}{\sqrt{v_n+\varepsilon}}\rangle$$

(note the similarity of V_n with the energy function (3.15)). Since $\inf F > -\infty$, we assume without loss of generality that $F \geqslant 0$. In this subsection, we use the notation ∇f_{n+1} as a shorthand notation for $\nabla f(x_n, \xi_{n+1})$ and C denotes some positive constant which may change from line to line. We write $\mathbb{E}_n = \mathbb{E}[\cdot \,|\, \mathscr{F}_n]$ for the conditional expectation w.r.t the σ-algebra $\mathscr{F}_n$. Define $P_n \triangleq \frac{1}{2}\langle D_n, m_n^{\odot 2}\rangle$, with $D_n \triangleq \frac{1}{\sqrt{v_n+\varepsilon}}$. We have the decomposition:

$$P_{n+1} - P_n = \frac{1}{2}\langle D_{n+1} - D_n, m_{n+1}^{\odot 2}\rangle + \frac{1}{2}\langle D_n, m_{n+1}^{\odot 2} - m_n^{\odot 2}\rangle. \tag{3.43}$$

We estimate the vector

$$D_{n+1} - D_n = \frac{\sqrt{v_n+\varepsilon} - \sqrt{v_{n+1}+\varepsilon}}{\sqrt{v_{n+1}+\varepsilon} \odot \sqrt{v_n+\varepsilon}}\,.$$

Remarking that $v_{n+1} \geqslant (1-\gamma_{n+1}q_n)v_n$ and using the update rule of v_n, we obtain for a sufficiently large n that

$$\begin{aligned}\sqrt{v_n+\varepsilon}-\sqrt{v_{n+1}+\varepsilon} &= \gamma_{n+1}\frac{q_n v_n - p_n \nabla f_{n+1}^{\odot 2}}{\sqrt{v_n+\varepsilon}+\sqrt{v_{n+1}+\varepsilon}} \\ &\leqslant \gamma_{n+1}q_n \frac{v_n}{(1+\sqrt{1-\gamma_{n+1}q_n})\sqrt{v_n+\varepsilon}} \\ &= \frac{\gamma_{n+1}q_n}{1+\sqrt{1-\gamma_{n+1}q_n}}\sqrt{v_n}\odot\frac{\sqrt{v_n}}{\sqrt{v_n+\varepsilon}} \\ &\leqslant c_{n+1}\sqrt{v_{n+1}} \text{ where } c_{n+1} \triangleq \frac{\gamma_{n+1}q_n}{\sqrt{1-\gamma_{n+1}q_n}(1+\sqrt{1-\gamma_{n+1}q_n})}\,. \end{aligned} \tag{3.44}$$

It is easy to see that $c_{n+1}/\gamma_n \to q_\infty/2$. Thus, for any $\delta > 0$, $c_{n+1} \leqslant (q_\infty+2\delta)\gamma_n/2$ for all n large enough. Using also that $\sqrt{v_{n+1}}/\sqrt{v_{n+1}+\varepsilon} \leqslant 1$, we obtain

$$D_{n+1} - D_n \leqslant \frac{q_\infty+2\delta}{2}\gamma_n D_n\,. \tag{3.45}$$

Substituting the above inequality in Equation (3.43), we obtain

$$\begin{aligned}P_{n+1}-P_n &\leqslant \left(\frac{q_\infty+2\delta}{2}\right)\frac{\gamma_n}{2}\langle D_n, m_{n+1}^{\odot 2}\rangle + \frac{1}{2}\langle D_n, m_{n+1}^{\odot 2}-m_n^{\odot 2}\rangle \\ &\leqslant \frac{q_\infty+2\delta}{2}\gamma_n P_n + \left(1+\frac{q_\infty+2\delta}{2}\gamma_n\right)\frac{1}{2}\langle D_n, m_{n+1}^{\odot 2}-m_n^{\odot 2}\rangle\,.\end{aligned}$$

Using $m_{n+1}^{\odot 2}-m_n^{\odot 2} = 2m_n\odot(m_{n+1}-m_n)+(m_{n+1}-m_n)^{\odot 2}$, and noting that $\mathbb{E}_n(m_{n+1}-m_n) = \gamma_{n+1}h_n\nabla F(x_n) - \gamma_{n+1}r_n m_n$,

$$\begin{aligned}\mathbb{E}_n\frac{1}{2}\langle D_n, m_{n+1}^{\odot 2}-m_n^{\odot 2}\rangle &= \gamma_{n+1}h_n\langle \nabla F(x_n), \frac{m_n}{\sqrt{v_n+\varepsilon}}\rangle - 2\gamma_{n+1}r_n P_n \\ &\quad + \frac{1}{2}\langle D_n, \mathbb{E}_n[(m_{n+1}-m_n)^{\odot 2}]\rangle\,.\end{aligned}$$

There exists $\delta > 0$ such that $r_\infty - \frac{q_\infty}{4} - \frac{\delta}{2} > 0$ by Assumption 3.2.4-*iv)*. As $\frac{\gamma_{n+1}}{\gamma_n}r_n - \frac{q_\infty}{4} \to r_\infty - \frac{q_\infty}{4}$, for all n large enough, $\frac{\gamma_{n+1}}{\gamma_n}r_n - \frac{q_\infty}{4} > r_\infty - \frac{q_\infty}{4} - \frac{\delta}{2} > 0$. Hence, for all n large enough,

$$\begin{aligned}\mathbb{E}_n P_{n+1} - P_n &\leqslant -2\left(r_\infty - \frac{q_\infty}{4} - \frac{\delta}{2}\right)\gamma_n P_n + \gamma_{n+1}h_n\langle \nabla F(x_n), \frac{m_n}{\sqrt{v_n+\varepsilon}}\rangle \\ &\quad + C\gamma_n^2\langle \nabla F(x_n), \frac{m_n}{\sqrt{v_n+\varepsilon}}\rangle + C\langle D_n, \mathbb{E}_n[(m_{n+1}-m_n)^{\odot 2}]\rangle\,.\end{aligned} \tag{3.46}$$

Using the inequality $\langle u, v\rangle \leqslant (\|u\|^2+\|v\|^2)/2$ and Assumption 3.3.6-*ii)*, it is easy to show the inequality $\langle \nabla F(x_n), \frac{m_n}{\sqrt{v_n+\varepsilon}}\rangle \leqslant C(1+F(x_n)+P_n)$. Moreover, using the componentwise inequality $(h_n\nabla f_{n+1} - r_n m_n)^{\odot 2} \leqslant 2h_n^2\nabla f_{n+1}^{\odot 2} + 2r_n^2 m_n^{\odot 2}$ along with Assumption 3.3.6-*ii)* and the boundedness of the sequences (h_n), (r_n) and (γ_{n+1}/γ_n), we obtain

$$\langle D_n, \mathbb{E}_n[(m_{n+1}-m_n)^{\odot 2}]\rangle \leqslant C\gamma_n^2(1+F(x_n)+P_n)\,. \tag{3.47}$$

Combining Equation (3.46) and Equation (3.47), we get

$$\mathbb{E}_n(P_{n+1} - P_n) \leqslant \gamma_{n+1}h_n\langle \nabla F(x_n), m_n \odot D_n\rangle + C\gamma_n^2(1 + F(x_n) + P_n)\,. \quad (3.48)$$

Denoting by M the Lipschitz coefficient of ∇F, we also have

$$F(x_{n+1}) \leqslant F(x_n) - \gamma_{n+1}\langle \nabla F(x_n), m_{n+1} \odot D_{n+1}\rangle + \frac{\gamma_{n+1}^2 M}{2}\left\|m_{n+1} \odot D_{n+1}\right\|^2 . \quad (3.49)$$

Using (3.45) and the update rule of m_n, we have
$\|m_{n+1} \odot D_{n+1} - m_n \odot D_n\|^2$

$$\begin{aligned}
&\leqslant C\left\|(m_{n+1} - m_n) \odot D_n\right\|^2 + C\left\|m_{n+1} \odot (D_{n+1} - D_n)\right\|^2 \\
&\leqslant C\gamma_{n+1}^2(\|\nabla f_{n+1}\|^2 + \|m_n \odot D_n\|^2) + C\gamma_{n+1}^2\left\|m_{n+1} \odot D_n\right\|^2 \\
&\leqslant C\gamma_{n+1}^2(\|m_n \odot D_n\|^2 + \|\nabla f_{n+1}\|^2)\,.
\end{aligned} \quad (3.50)$$

Finally, recalling that $V_n = h_{n-1}F(x_n) + P_n$, (h_n) is decreasing, combining Equation (3.48),(3.49),(3.50), and using Assumption 3.3.6, we have

$$\begin{aligned}
\mathbb{E}_n[V_{n+1}] &\leqslant V_n + \gamma_{n+1}h_n\langle \nabla F(x_n), \mathbb{E}_n\left[m_n \odot D_n - m_{n+1} \odot D_{n+1}\right]\rangle \\
&\quad + C\gamma_{n+1}^2\left(1 + F(x_n) + P_n + \|m_n \odot D_n\|^2\right) \\
&\quad + C\gamma_{n+1}^2\mathbb{E}_n[\|m_n \odot D_n - m_{n+1} \odot D_{n+1}\|^2] \\
&\leqslant V_n + C\gamma_n^2\left(1 + F(x_n) + P_n + \|m_n \odot D_n\|^2 + \mathbb{E}_n\left[\|\nabla f_{n+1}\|^2\right]\right) \\
&\leqslant V_n + C\gamma_n^2(1 + F(x_n) + P_n) \\
&\leqslant (1 + C\gamma_n^2)V_n + C\gamma_n^2\,,
\end{aligned}$$

where we used Cauchy-Schwarz's inequality and the fact that $\|m_n \odot D_n\|^2 \leqslant CP_n$. By the Robbins-Siegmund's theorem [Robbins & Siegmund 1971], the sequence (V_n) converges almost surely to a finite random variable $V_\infty \in \mathbb{R}^+$. Then, the coercivity of F implies that (x_n) is almost surely bounded.

We now establish the almost sure boundedness of (m_n). Assume in the sequel that n is large enough to have $(1 - \gamma_{n+1}r_n) \geqslant 0$. Consider the martingale difference sequence $\Delta_{n+1} \stackrel{\Delta}{=} \nabla f_{n+1} - \nabla F(x_n)$. We decompose $m_n = \bar{m}_n + \tilde{m}_n$ where $\bar{m}_{n+1} = (1 - \gamma_{n+1}r_n)\bar{m}_n + \gamma_{n+1}h_n\nabla F(x_n)$ and $\tilde{m}_{n+1} = (1 - \gamma_{n+1}r_n)\tilde{m}_n + \gamma_{n+1}h_n\Delta_{n+1}$, setting $\bar{m}_0 = 0$ and $\tilde{m}_0 = m_0$. We prove that both terms $\bar{m}_n$ and $\tilde{m}_n$ are bounded. Consider the first term: $\|\bar{m}_{n+1}\| \leqslant (1-\gamma_{n+1}r_n)\|\bar{m}_n\| + \gamma_{n+1}\sup_k \|h_k\nabla F(x_k)\|$, where the supremum in the above inequality is almost surely finite by continuity of ∇F. We immediately get that if $\|\bar{m}_n\| \geqslant \frac{\sup_k \|h_k\nabla F(x_k)\|}{r_\infty}$, then $\|\bar{m}_{n+1}\| \leqslant \|\bar{m}_n\|$. Thus

$$\|\bar{m}_{n+1}\| \leqslant \frac{\sup_k \|h_k\nabla F(x_k)\|}{r_\infty} + \sup_k \gamma_{k+1}\|h_k\nabla F(x_k)\|\,,$$

which implies that $\bar{m}_n$ is bounded.

Consider now the term $\tilde{m}_n$:

$$\mathbb{E}_n[\|\tilde{m}_{n+1}\|^2] = (1-\gamma_{n+1}r_n)^2\|\tilde{m}_n\|^2 + \gamma_{n+1}^2 h_n^2 \mathbb{E}_n[\|\Delta_{n+1}\|^2] \leqslant \|\tilde{m}_n\|^2 + \gamma_{n+1}^2 h_n^2 \mathbb{E}_n[\|\Delta_{n+1}\|^2]\,.$$

Then, the inequality $\mathbb{E}_n[\|\Delta_{n+1}\|^2] \leqslant \mathbb{E}_n[\|\nabla f_{n+1}\|^2]$ combined with Assumption 3.3.4-i) and the a.s. boundedness of the sequence (x_n) imply that there exists a finite random variable $C_{\mathcal{K}}$ (independent of n) s.t. $\mathbb{E}_n[\|\nabla f_{n+1}\|^2] \leqslant C_{\mathcal{K}}$. As a consequence, since $\sum_n \gamma_{n+1}^2 < \infty$ and the sequence (h_n) is bounded, we obtain that a.s.:

$$\sum_{n\geqslant 0} \gamma_{n+1}^2 h_n^2 \mathbb{E}_n[\|\Delta_{n+1}\|^2] \leqslant C C_{\mathcal{K}} \sum_{n\geqslant 0} \gamma_{n+1}^2 < +\infty\,.$$

Hence, we can apply the Robbins-Siegmund theorem to obtain that $\sup_n \|\tilde{m}_n\|^2 < \infty$ w.p.1. Finally, it can be shown that (v_n) is almost surely bounded using the same arguments, decomposing v_n into $\bar{v}_n + \tilde{v}_n$ as above. Indeed, first, we have:

$$\mathbb{E}_n[\|\tilde{v}_{n+1}\|^2] \leqslant \|\tilde{v}_n\|^2 + \gamma_{n+1}^2 p_n^2 \mathbb{E}_n[\|\nabla f_{n+1}^{\odot 2} - S(x_n)\|^2]\,.$$

Second, it also holds that:

$$\mathbb{E}_n[\|\nabla f_{n+1}^{\odot 2} - S(x_n)\|^2] \leqslant \mathbb{E}_n[\|\nabla f_{n+1}^{\odot 2}\|^2] \leqslant \mathbb{E}_n[\|\nabla f_{n+1}\|^4]\,.$$

Then, using Assumption 3.3.4-i) and the a.s. boundedness of the sequence (x_n), there exists a finite random variable $C'_{\mathcal{K}}$ (independent of n) s.t. $\mathbb{E}_n[\|\nabla f_{n+1}\|^4] \leqslant C'_{\mathcal{K}}$. Moreover, the sequence (p_n) is bounded and $\sum_n \gamma_{n+1}^2 < \infty$. As a consequence, it holds that a.s:

$$\sum_{n\geqslant 0} \gamma_{n+1}^2 p_n^2 \mathbb{E}_n[\|\nabla f_{n+1}^{\odot 2} - S(x_n)\|^2] \leqslant C C'_{\mathcal{K}} \sum_{n\geqslant 0} \gamma_{n+1}^2 < +\infty\,.$$

It follows that the Robbins-Siegmund theorem can be applied to the sequence $\|\tilde{v}_n\|^2$ as for the sequence $\|\tilde{m}_n\|^2$ to obtain that $\sup_n \|\tilde{v}_n\|^2 < \infty$ w.p.1.

3.6.5 Proof of Theorem 3.3.4

The proof of Theorem 3.3.2 easily adapts to Algorithm 2 by replacing V_n by

$$\tilde{V}_n \triangleq F(x_n) + \frac{1}{2}\|m_n\|^2\,.$$

The boundedness of (m_n) is an immediate consequence of the convergence of $\tilde{V}_n$.

3.6.6 Proof of Theorem 3.3.5

We shall use the following result.

Theorem 3.6.2 (adapted from [Pelletier 1998], Theorem 7)**.** *Let $k \geqslant 1$. On some probability space equipped with a filtration $\mathscr{F} = (\mathscr{F}_n)_{n\in\mathbb{N}}$, consider a sequence of r.v. on $\mathbb{R}^k$ given by*

$$Z_{n+1} = (I + \gamma_{n+1}\bar{H})Z_n + \gamma_{n+1}b_{n+1} + \sqrt{\gamma_{n+1}}\eta_{n+1}$$

and $\mathbb{E}[\|Z_0\|^2] < \infty$*, where* $\bar{H}$ *is a* $k \times k$ *Hurwitz matrix,* (b_n) *and* (η_n) *are random sequences, and* $\gamma_n = \gamma_0 n^{-\alpha}$ *for some* $\gamma_0 > 0$ *and* $\alpha \in (0, 1]$*. Let* $\Omega_0 \in \mathscr{F}_\infty$ *have a positive probability. Assume that the following holds almost surely on* Ω_0*:*

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

ii) There exists a constant $\bar{b} > 2$ *s.t.* $\sup_{n\geqslant 0} \mathbb{E}[\|\eta_{n+1}\|^{\bar{b}}|\mathscr{F}_n] < \infty$*.*

iii) $\mathbb{E}[\eta_{n+1}\eta_{n+1}^T|\mathscr{F}_n] = \Sigma + \Delta_n$ *where* $\mathbb{E}[\|\Delta_n\|\mathbb{1}_{\Omega_0}] \to 0$ *and* Σ *is a positive semidefinite matrix.*

iv) The sequence (b_n) *is the sum of two sequences* $(b_{n,1})$ *and* $(b_{n,2})$*, adapted to* $\mathscr{F}$*, s.t.* $\sup_{n\geqslant 0} \mathbb{E}[\|b_{n,1}\|^2] < \infty$*,* $\mathbb{E}[\|b_{n,1}\|\mathbb{1}_{\Omega_0}] \to 0$ *and* $b_{n,2} \to 0$ *a.s. on* Ω_0*.*

Then, given Ω_0*,* (Z_n) *converges in distribution to the unique stationary distribution* $\mu_\star$ *of the generalized Ornstein-Uhlenbeck process*

$$dX_t = \bar{H}X_t dt + \sqrt{\Sigma}dB_t$$

where (B_t) *is the standard Brownian motion and* $\sqrt{\Sigma}$ *is the unique positive semidefinite square root of* Σ*. The distribution* $\mu_\star$ *is the zero mean Gaussian distribution with covariance matrix* Γ *given as the solution to* $(\bar{H} + \frac{\mathbb{1}_{\alpha=1}}{2\gamma_0} I_k)\Gamma + \Gamma(\bar{H} + \frac{\mathbb{1}_{\alpha=1}}{2\gamma_0} I_k)^T = -\Sigma$*.*

Proof. The proof is identical to the proof of [Pelletier 1998, Theorem 7], only substituting the inverse of the square root of Σ by the Moore-Penrose inverse. Finally, the uniqueness of the stationary distribution $\mu_\star$ and its expression follow from [Karatzas & Shreve 1991, Theorem 6.7, p. 357] □

We define $v_n = \bar{v}_n + \delta_n$ where $\delta_0 = 0$, $\bar{v}_0 = v_0$ and

$$\begin{aligned}
\delta_{n+1} &= (1 - \gamma_{n+1}q_n)\delta_n + \gamma_{n+1}(p_n - q_n q_\infty^{-1} p_\infty)S(x_n) \\
\bar{v}_{n+1} &= (1 - \gamma_{n+1}q_n)\bar{v}_n + \gamma_{n+1}q_n q_\infty^{-1} p_\infty S(x_n) + \gamma_{n+1}p_n(\nabla f(x_n, \xi_{n+1})^{\odot 2} - S(x_n))\,.
\end{aligned}$$

For every $z = (v, m, x) \in \mathcal{Z}_+$ and $\delta \geqslant 0$, we define

$$r_n(z, \delta) \triangleq \begin{bmatrix} q_n q_\infty^{-1} p_\infty (S(x - \gamma_n \frac{m}{\sqrt{v+\delta+\varepsilon}}) - S(x)) \\ h_n(\nabla F(x - \gamma_n \frac{m}{\sqrt{v+\delta+\varepsilon}}) - \nabla F(x)) \\ \frac{\gamma_n}{\gamma_{n+1}}(\frac{1}{\sqrt{v+\varepsilon}} - \frac{1}{\sqrt{v+\delta+\varepsilon}}) \odot m \end{bmatrix}.$$

Moreover, for every $z = (v, m, x) \in \mathcal{Z}_+$ and every $n \in \mathbb{N}$, we set

$$g_n(z) = \begin{bmatrix} q_n q_\infty^{-1} p_\infty S(x) - q_n v \\ h_n \nabla F(x) - r_n m \\ -\frac{\gamma_n}{\gamma_{n+1}} \frac{m}{\sqrt{v+\varepsilon}} \end{bmatrix}.$$

Defining $\zeta_n = (\bar{v}_n, m_n, x_{n-1})$ and recalling the definition of (η_n) from Equation (3.28), we have the decomposition

$$\zeta_{n+1} = \zeta_n + \gamma_{n+1} g_n(\zeta_n) + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1} r_n(\zeta_n, \delta_n)\,.$$

Define $z_\star \triangleq (x_\star, 0, v_\star)$. Note that $g_n(z_\star) = 0$. Evaluating the Jacobian matrix G_n of g_n at $z_\star$, we obtain that there exist constants $C > 0$, $\bar{M} > 0$ and $n_0 \in \mathbb{N}$ s.t. for all $n \geqslant n_0$,

$$\|g_n(z) - G_n(z - z_\star)\| \leqslant C\|z - z_\star\|^2 \quad (\forall z \in B(z_\star, \bar{M}))\,, \tag{3.51}$$

where G_n is given by

$$G_n \triangleq \begin{bmatrix} -q_n I_d & 0 & q_n q_\infty^{-1} p_\infty \nabla S(x_\star) \\ 0 & -r_n I_d & h_n \nabla^2 F(x_\star) \\ 0 & -\frac{\gamma_n}{\gamma_{n+1}} V & 0 \end{bmatrix},$$

where ∇S is the Jacobian of S and the matrix V is defined in Equation (3.8). We define

$$G_\infty \triangleq \lim_n G_n = \begin{bmatrix} -q_\infty I_d & 0 & p_\infty \nabla S(x_\star) \\ 0 & -r_\infty I_d & h_\infty \nabla^2 F(x_\star) \\ 0 & -V & 0 \end{bmatrix}.$$

One can verify that G_∞ is Hurwitz, and that the largest real part of its eigenvalues is $-L'$, where $L' \triangleq L \wedge q_\infty$ and L is defined in Equation (3.9).

We define $\Omega^{(0)} \triangleq \{z_n \to z_\star\}$. We assume $\mathbb{P}(\Omega^{(0)}) > 0$. Using for instance [Delyon *et al.* 1999, Lemma 4 and Lemma 5], it holds that $\delta_n(\omega) \to 0$ for every $\omega \in \Omega^{(0)}$, and since $x_n(\omega) - x_{n-1}(\omega) \to 0$ on that set, we obtain that $\Omega^{(0)} = \{\zeta_n \to z_\star\}$. Let $M \in (0, \bar{M})$ be a constant, whose value will be specified later on. For every $N_0 \in \mathbb{N}$, define $\Omega^{(0)}_{N_0} \triangleq \{\zeta_n \to z_\star \text{ and } \sup_{n \geqslant N_0} \|\zeta_n - z_\star\| \leqslant M\}$. We seek to show that $\sqrt{\gamma_n}^{-1}(\zeta_n - z_\star) \Rightarrow \nu$ given $\Omega^{(0)}$, for some Gaussian measure ν, using Theorem 3.6.2. As $\Omega^{(0)}_{N_0} \uparrow \Omega^{(0)}$, it is sufficient to show that the latter convergence holds given $\Omega^{(0)}_{N_0}$, for every N_0 large enough. From now on, we consider that N_0 is fixed. We define the sequence $(\tilde{\zeta}_n)_{n \geqslant N_0}$ as $\tilde{\zeta}_{N_0} = \zeta_{N_0}$ and for every $n \geqslant N_0$,

$$\tilde{\zeta}_{n+1} = \tilde{\zeta}_n + \gamma_{n+1}\tilde{g}_n(\tilde{\zeta}_n) + \gamma_{n+1}(\eta_{n+1} + r_n(\tilde{\zeta}_n, \delta_n))\mathbb{1}_{\mathcal{A}_n}$$

where $\mathcal{A}_n$ is the event defined by

$$\mathcal{A}_n \triangleq \bigcap_{k=N_0}^{n} \{\|x_k - x_\star\| \leqslant M\} \cap \{\|\tilde{\zeta}_n - z_\star\| \leqslant M\}$$

and

$$\tilde{g}_n(z) \triangleq g_n(z)\mathbb{1}_{\|z - z_\star\| \leqslant M} - K(z - z_\star)\mathbb{1}_{\|z - z_\star\| > M}\,,$$

where $K > 0$ is a large constant which will be specified later on. The sequences $(\tilde{\zeta}_n)_{n \geqslant N_0}$ and $(\zeta_n)_{n \geqslant N_0}$ coincide on $\Omega^{(0)}_{N_0}$. Thus, it is sufficient to study the weak convergence of $(\tilde{\zeta}_n)_{n \geqslant N_0}$.

An estimate of $\|r_n(\tilde{\zeta}_n, \delta_n)\|\mathbb{1}_{\mathcal{A}_n}$. We start by studying the sequence $(\|\delta_n\|\mathbb{1}_{\mathcal{A}_n})$. Unfolding the update rule defining δ_n and using the fact that (q_n) is a sequence of

positive reals converging to $q_\infty > 0$, we obtain that

$$\begin{aligned}\|\delta_n\|\mathbb{1}_{\mathcal{A}_n} &\leqslant \sum_{k=1}^{n}\left[\prod_{j=k+1}^{n}|1-\gamma_j q_{j-1}|\right]\gamma_k|p_{k-1}-q_{k-1}q_\infty^{-1}p_\infty|\|S(x_{k-1})\|\mathbb{1}_{\mathcal{A}_n}\\ &\leqslant C\sum_{k=1}^{n}\exp\left(-\beta\sum_{j=k+1}^{n}\gamma_j\right)\gamma_k|p_{k-1}-q_{k-1}q_\infty^{-1}p_\infty| \triangleq w_n\,,\end{aligned}$$

for some $\beta > 0$. The sequence (w_n) is deterministic and converges to zero by [Delyon *et al.* 1999, Lemma 4]. There exists $n_1 \geqslant n_0$ s.t. $w_n \leqslant M$. As $v \mapsto \frac{1}{\sqrt{v+\varepsilon}}$ is Lipschitz and ∇F and S are locally Lipschitz, for every $z = (v, m, x)$ and δ s.t. $\|z - z_\star\| \leqslant M$ and $\|\delta\| \leqslant M$, we have

$$\begin{aligned}\|r_n(z,\delta)\| &\leqslant C\gamma_{n+1}\|(v+\delta+\varepsilon)^{\odot-\frac{1}{2}}\|\|m\| + C\|(v+\delta+\varepsilon)^{\odot-\frac{1}{2}} - (v+\varepsilon)^{\odot-\frac{1}{2}}\|\|m\|\\ &\leqslant C\gamma_{n+1}\|z-z_\star\| + C\|\delta\|\|z-z_\star\|\,.\end{aligned}$$

This implies that for every $n \geqslant n_1$,

$$\|r_n(\tilde\zeta_n,\delta_n)\|\mathbb{1}_{\mathcal{A}_n} \leqslant C(\gamma_{n+1}+w_n)\|\tilde\zeta_n - z_\star\|\,. \tag{3.52}$$

Tightness of $\sqrt{\gamma_n}^{-1}(\tilde\zeta_n - z_\star)$. We decompose

$$\begin{aligned}\tilde\zeta_{n+1} - z_\star &= (I_{3d}+\gamma_{n+1}G_n)(\tilde\zeta_n - z_\star) + \gamma_{n+1}\left(g_n(\tilde\zeta_n) - G_n(\tilde\zeta_n - z_\star)\right)\mathbb{1}_{\|\tilde\zeta_n - z_\star\|\leqslant M}\\ &\quad - \gamma_{n+1}(K+G_n)(\tilde\zeta_n - z_\star)\mathbb{1}_{\|\tilde\zeta_n - z_\star\|>M} + \gamma_{n+1}(\eta_{n+1} + r_n(\tilde\zeta_n,\delta_n))\mathbb{1}_{\mathcal{A}_n}\,.\end{aligned} \tag{3.53}$$

For a given $t > 0$, we write $G_\infty = B_t^{-1}G_tB_t$ the Jordan-like decomposition of G_∞, where the ones of the second diagonal of the usual Jordan decomposition are replaced by t, and where B_t is some invertible matrix. We define $S_n \triangleq B_t(\tilde\zeta_n - z_\star)$. Setting $G_n^{(t)} \triangleq B_tG_nB_t^{-1}$, we obtain

$$\begin{aligned}S_{n+1} &= (I_{3d}+\gamma_{n+1}G_n^{(t)})S_n + \gamma_{n+1}B_t\left(g_n(\tilde\zeta_n) - G_n(\tilde\zeta_n - z_\star)\right)\mathbb{1}_{\|\tilde\zeta_n - z_\star\|\leqslant M}\\ &\quad - \gamma_{n+1}(K+G_n^{(t)})S_n\mathbb{1}_{\|\tilde\zeta_n - z_\star\|>M} + \gamma_{n+1}B_t(\eta_{n+1}+r_n(\tilde\zeta_n,\delta_n))\mathbb{1}_{\mathcal{A}_n}\,.\end{aligned}$$

Choose $A \in (0, 2L')$ and $A' \in (A, 2L')$. There exists $\bar\gamma$ and $t > 0$ s.t. for every $\gamma < \bar\gamma$, $\|I + \gamma G_t\|_2 \leqslant 1 - \gamma(A' + 2L')/2$, where $\|\cdot\|_2$ is the spectral norm. As $G_n^{(t)} \to G^t$, there exists $n_2 \geqslant n_1$, such that for all $n \geqslant n_2$, $\|I + \gamma G_n^{(t)}\|_2 \leqslant 1 - \gamma A'$. Recall the notation $\mathbb{E}_n = \mathbb{E}[\cdot\,|\,\mathscr{F}_n]$. We expand $\|S_{n+1}\|^2$ and use the inequality $\left\|g_n(\tilde\zeta_n) - G_n(\tilde\zeta_n - z_\star)\right\|^2\mathbb{1}_{\|\tilde\zeta_n - z_\star\|\leqslant M} \leqslant C\|S_n\|^2$ to obtain after straightforward algebra

$$\begin{aligned}\mathbb{E}_n\|S_{n+1}\|^2 &\leqslant (1-\gamma_{n+1}A')\|S_n\|^2 + C\gamma_{n+1}^2\|S_n\|^2\\ &\quad + C\gamma_{n+1}^2(\mathbb{E}_n\|\eta_{n+1}\|^2 + \|r_n(\tilde\zeta_n,\delta_n)\|^2)\mathbb{1}_{\mathcal{A}_n}\\ &\quad + 2\gamma_{n+1}\mathrm{Re}\left(S_n^*B_t\left(g_n(\tilde\zeta_n) - G_n(\tilde\zeta_n - z_\star)\right)\right)\mathbb{1}_{\|\tilde\zeta_n - z_\star\|\leqslant M}\\ &\quad - 2\gamma_{n+1}\mathrm{Re}\left(S_n^*(K+G_n^{(t)})S_n\right)\mathbb{1}_{\|\tilde\zeta_n - z_\star\|>M} + 2\gamma_{n+1}\mathrm{Re}\left(S_n^*B_tr_n(\tilde\zeta_n,\delta_n)\right)\mathbb{1}_{\mathcal{A}_n}\,.\end{aligned}$$

Choose $c \triangleq (A' - A)/2$. If M is chosen small enough,

$$\|g_n(\tilde{\zeta}_n) - G_n(\tilde{\zeta}_n - z_\star)\| \mathbb{1}_{\|\tilde{\zeta}_n - z_\star\| \leqslant M} \leqslant \frac{c}{2} \|B_t\|^{-1} \|B_t^{-1}\| \|\tilde{\zeta}_n - z_\star\| \,.$$

Moreover, choosing $K > \sup_n \|G_n^{(t)}\|_2$, it holds that $\mathrm{Re}\left(S_n^*(K + G_n^{(t)})S_n\right) \geqslant 0$. Then,

$$\begin{aligned}\mathbb{E}_n\|S_{n+1}\|^2 &\leqslant (1 - \gamma_{n+1}(A' - c))\|S_n\|^2 + C\gamma_{n+1}^2\|S_n\|^2 \\ &\quad + C\gamma_{n+1}^2(\mathbb{E}_n\|\eta_{n+1}\|^2 + \|r_n(\tilde{\zeta}_n, \delta_n)\|^2)\mathbb{1}_{\mathcal{A}_n} + 2\gamma_{n+1}\|B_t\|\|S_n\|\|r_n(\tilde{\zeta}_n, \delta_n)\|\mathbb{1}_{\mathcal{A}_n} \,.\end{aligned}$$

Using Equation (3.52),

$$\begin{aligned}\mathbb{E}_n\|S_{n+1}\|^2 &\leqslant (1 - \gamma_{n+1}(A' - c - w_n))\|S_n\|^2 + C\gamma_{n+1}^2(1 + w_n^2)\|S_n\|^2 \\ &\qquad + C\gamma_{n+1}^2\mathbb{E}_n\|\eta_{n+1}\|^2\mathbb{1}_{\mathcal{A}_n} \,.\end{aligned}$$

Therefore, there exists $n_3 \geqslant n_2$ s.t. for all $n \geqslant n_3$,

$$\mathbb{E}\|S_{n+1}\|^2 \leqslant (1 - \gamma_{n+1}A)\mathbb{E}\|S_n\|^2 + C\gamma_{n+1}^2\mathbb{E}(\|\eta_{n+1}\|^2\mathbb{1}_{\|x_n - x_\star\| \leqslant M}) \,.$$

The second expectation in the righthand side is bounded uniformly in n by the condition (3.7). Using [Delyon *et al.* 1999, Lemma 4 and Lemma 5], we conclude that $\sup_n \gamma_n^{-1}\mathbb{E}\|S_n\|^2 < \infty$. Therefore, $\sup_n \gamma_n^{-1}\mathbb{E}\|\tilde{\zeta}_n - z_\star\|^2 < \infty$, which in turn implies $\sup_n \gamma_n^{-1}\mathbb{E}(\|\zeta_n - z_\star\|^2\mathbb{1}_{\Omega_{N_0}^{(0)}}) < \infty$.

Strongly perturbed iterations. We define $\tilde{y}_n = \sqrt{\gamma_n}^{-1}(\tilde{\zeta}_n - z_\star)$. Define

$$\bar{G}_n \triangleq \gamma_{n+1}^{-1}\left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}} - 1\right) I_{3d} + \sqrt{\frac{\gamma_n}{\gamma_{n+1}}} G_n \,.$$

The sequence $\bar{G}_n$ converges to $\bar{G}_\infty \triangleq G_\infty + \frac{\mathbb{1}_{\alpha=1}}{2\gamma_0} I_{3d}$. Recalling Equation (3.53), we can write

$$\tilde{y}_{n+1} = (I_{3d} + \gamma_{n+1}\bar{G}_\infty)\tilde{y}_n + \gamma_{n+1}\bar{r}_n + \sqrt{\gamma_{n+1}}\bar{\eta}_{n+1}$$

where $\bar{\eta}_{n+1} = \eta_{n+1}\mathbb{1}_{\mathcal{A}_n}$ and $\bar{r}_n = \bar{r}_{n,1} + \bar{r}_{n,2} + \bar{r}_{n,3}$, where

$$\begin{aligned}\bar{r}_{n,1} &\triangleq \sqrt{\gamma_{n+1}}^{-1} r_n(\tilde{\zeta}_n, \delta_n)\mathbb{1}_{\mathcal{A}_n} + (\bar{G}_n - \bar{G}_\infty)\tilde{y}_n \\ \bar{r}_{n,2} &\triangleq \sqrt{\gamma_{n+1}}^{-1}\left(g_n(\tilde{\zeta}_n) - G_n(\tilde{\zeta}_n - z_\star)\right)\mathbb{1}_{\|\tilde{\zeta}_n - z_\star\| \leqslant M} \\ \bar{r}_{n,3} &\triangleq -\sqrt{\gamma_{n+1}}^{-1}(K + G_n)(\tilde{\zeta}_n - z_\star)\mathbb{1}_{\|\tilde{\zeta}_n - z_\star\| > M} \,.\end{aligned}$$

We now check that the assumptions of Theorem 3.6.2 are fulfilled. On the event $\Omega_{N_0}^{(0)}$, we recall that $\tilde{\zeta}_n = \zeta_n$, hence $\bar{r}_{n,3}$ is identically zero. Moreover, using Equation (3.52), it holds that for all n large enough,

$$\|\bar{r}_{n,1}\| \leqslant C\left(\sqrt{\frac{\gamma_n}{\gamma_{n+1}}}(\gamma_{n+1} + w_n) + \|\bar{G}_n - \bar{G}_\infty\|\right)\|\tilde{y}_n\|$$

and therefore, $\mathbb{E}[\|\bar{r}_{n,1}\|^2] \to 0$. Now consider the term $\bar{r}_{n,2}$. By Equation (3.51),

$$\|\bar{r}_{n,2}\| \leqslant C\sqrt{\gamma_{n+1}}^{-1}\|\tilde{\zeta}_n - z_\star\|^2 \mathbb{1}_{\|\tilde{\zeta}_n - z_\star\| \leqslant M}\,.$$

Thus, $\|\bar{r}_{n,2}\|^2 \leqslant C\|\tilde{y}_n\|^2$ which implies that $\sup_{n \geqslant N_0} \mathbb{E}[\|r_{n,2}\|^2] < \infty$. Moreover, $\mathbb{E}[\|\bar{r}_{n,2}\|] \leqslant C\sqrt{\gamma_{n+1}}\mathbb{E}\|\tilde{y}_n\|^2$ tends to zero. Finally, consider $\bar{\eta}_{n+1}$. Using condition (3.7), there exist $M > 0$ and $b_M > 4$ s.t.

$$\begin{aligned}\mathbb{E}_n[\|\bar{\eta}_{n+1}\|^{b_M/2}] &\leqslant \mathbb{E}_n[\|\eta_{n+1}\|^{b_M/2}]\mathbb{1}_{\|x_n - x_\star\| \leqslant M} \\ &\leqslant C\mathbb{E}_n[\|\nabla f(x_n, \xi_{n+1})\|^{b_M}]\mathbb{1}_{\|x_n - x_\star\| \leqslant M} \leqslant C\,.\end{aligned}$$

Moreover, $\mathbb{E}_n[\bar{\eta}_{n+1}] = 0$ and finally, almost surely on $\Omega_N^{(0)}$, $\mathbb{E}_n[\bar{\eta}_{n+1}\bar{\eta}_{n+1}^{\mathrm{T}}]$ converges to

$$\Sigma \triangleq \begin{bmatrix} \mathbb{E}_\xi\left[\begin{bmatrix} p_\infty(\nabla f(x_\star,\xi)^{\odot 2} - S(x_\star)) \\ h_\infty \nabla f(x_\star, \xi)\end{bmatrix}\begin{bmatrix} p_\infty(\nabla f(x_\star,\xi)^{\odot 2} - S(x_\star)) \\ h_\infty \nabla f(x_\star, \xi)\end{bmatrix}^{\mathrm{T}}\right] & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.54}$$

Therefore, the assumptions of Theorem 3.6.2 are fulfilled for the sequence $\tilde{y}_n$. We obtain the desired result for the sequence (m_n, x_{n-1}). We now show that the same result also holds for the sequence (m_n, x_n). For this purpose, observe that

$$\frac{1}{\sqrt{\gamma_n}}\begin{bmatrix} m_n \\ x_n - x_\star\end{bmatrix} = \frac{1}{\sqrt{\gamma_n}}\begin{bmatrix} m_n \\ x_{n-1} - x_\star\end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{\sqrt{\gamma_n}}(x_n - x_{n-1})\end{bmatrix}.$$

Then, notice that $\|\frac{x_n - x_{n-1}}{\sqrt{\gamma_n}}\| = \sqrt{\gamma_n}\|\frac{m_n}{\sqrt{v_n + \varepsilon}}\| \leqslant \sqrt{\frac{\gamma_n}{\varepsilon}}\|m_n\| \to 0$ as $n \to \infty$ since it is assumed that $z_n \to z_\star$ (which implies in particular that $m_n \to 0$). Hence, it holds that $\sqrt{\gamma_n}^{-1}(x_n - x_{n-1})$ converges a.s. to 0. We conclude by invoking Slutsky's lemma.

Proof of Equation (3.10). We have the subsystem:

$$\tilde{H}\Gamma + \Gamma\tilde{H}^{\mathrm{T}} = \begin{bmatrix} -h_\infty^2\mathcal{Q} & 0 \\ 0 & 0\end{bmatrix} \qquad \text{where } \tilde{H} \triangleq \begin{bmatrix} (\theta - r_\infty)I_d & h_\infty \nabla^2 F(x_\star) \\ -V & \theta I_d\end{bmatrix} \tag{3.55}$$

and where $\mathcal{Q} \triangleq \mathrm{Cov}\,(\nabla f(x_\star, \xi))$. The next step is to triangularize the matrix $\tilde{H}$ in order to decouple the blocks of Γ. For every $k = 1, \ldots, d$, set $\nu_k^\pm \triangleq -\frac{r_\infty}{2} \pm \sqrt{r_\infty^2/4 - h_\infty \pi_k}$ with the convention that $\sqrt{-1} = \imath$ (inspecting the characteristic polynomial of H, these are the eigenvalues of H). Set $M^\pm \triangleq \mathrm{diag}\,(\nu_1^\pm, \cdots, \nu_d^\pm)$ and $R^\pm \triangleq V^{-\frac{1}{2}} P M^\pm P^{\mathrm{T}} V^{-\frac{1}{2}}$. Using the identities $M^+ + M^- = -r_\infty I_d$ and $M^+M^- = h_\infty\,\mathrm{diag}\,(\pi_1, \cdots, \pi_d)$, it can be checked that

$$\mathcal{R}\tilde{H} = \begin{bmatrix} R^- V + \theta I_d & 0 \\ -V & VR^+ + \theta I_d\end{bmatrix}\mathcal{R}, \text{ where } \mathcal{R} \triangleq \begin{bmatrix} I_d & R^+ \\ 0 & I_d\end{bmatrix}.$$

Set $\tilde{\Gamma} \triangleq \mathcal{R}\Gamma\mathcal{R}^{\mathrm{T}}$. Denote by $(\tilde{\Gamma}_{i,j})_{i,j=1,2}$ the blocks of $\tilde{\Gamma}$. Note that $\tilde{\Gamma}_{2,2} = \Gamma_{2,2}$. By left/right multiplication of Equation (3.55) respectively by $\mathcal{R}$ and $\mathcal{R}^{\mathrm{T}}$, we obtain

$$(R^- V + \theta I_d)\tilde{\Gamma}_{1,1} + \tilde{\Gamma}_{1,1}(VR^- + \theta I_d) = -h_\infty^2 \mathcal{Q} \tag{3.56}$$
$$(R^- V + \theta I_d)\tilde{\Gamma}_{1,2} + \tilde{\Gamma}_{1,2}(R^+ V + \theta I_d) = \tilde{\Gamma}_{1,1} V \tag{3.57}$$
$$(VR^+ + \theta I_d)\tilde{\Gamma}_{2,2} + \tilde{\Gamma}_{2,2}(R^+ V + \theta I_d) = V\tilde{\Gamma}_{1,2} + \tilde{\Gamma}_{1,2}^{\mathrm{T}} V\,. \tag{3.58}$$

Set $\bar{\Gamma}_{1,1} = P^{-1} V^{\frac{1}{2}} \tilde{\Gamma}_{1,1} V^{\frac{1}{2}} P$. Define $C \triangleq P^{-1} V^{\frac{1}{2}} \mathcal{Q} V^{\frac{1}{2}} P$. Equation (3.56) yields

$$(M^- + \theta I_d)\bar{\Gamma}_{1,1} + \bar{\Gamma}_{1,1}(M^- + \theta I_d) = -h_\infty^2 C\,.$$

Set $\bar{\Gamma}_{1,2} = P^{-1} V^{\frac{1}{2}} \tilde{\Gamma}_{1,2} V^{-\frac{1}{2}} P$. Equation (3.57) is rewritten $(M^- + \theta I_d)\bar{\Gamma}_{1,2} + \bar{\Gamma}_{1,2}(M^+ + \theta I_d) = \bar{\Gamma}_{1,1}$. The component (k, ℓ) is given by

$$\bar{\Gamma}_{1,2}^{k,\ell} = (\nu_k^- + \nu_\ell^+ + 2\theta)^{-1} \bar{\Gamma}_{1,1}^{k,\ell} = \frac{-h_\infty^2 C_{k,\ell}}{(\nu_k^- + \nu_\ell^+ + 2\theta)(\nu_k^- + \nu_\ell^- + 2\theta)}\,.$$

Set finally $\bar{\Gamma}_{2,2} = P^{-1} V^{-\frac{1}{2}} \Gamma_{2,2} V^{-\frac{1}{2}} P$. Equation (3.58) becomes

$$(M^+ + \theta I_d)\bar{\Gamma}_{2,2} + \bar{\Gamma}_{2,2}(M^+ + \theta I_d) = \bar{\Gamma}_{1,2} + \bar{\Gamma}_{1,2}^{\mathrm{T}}\,.$$

Thus,

$$\begin{aligned}\bar{\Gamma}_{2,2}^{k,\ell} &= \frac{\bar{\Gamma}_{1,2}^{k,\ell} + \bar{\Gamma}_{1,2}^{\ell,k}}{\nu_k^+ + \nu_\ell^+ + 2\theta} \\ &= \frac{-h_\infty^2 C_{k,\ell}}{(\nu_k^+ + \nu_\ell^+ + 2\theta)(\nu_k^- + \nu_\ell^- + 2\theta)} \left(\frac{1}{\nu_k^- + \nu_\ell^+ + 2\theta} + \frac{1}{\nu_k^+ + \nu_\ell^- + 2\theta} \right)\,.\end{aligned}$$

After tedious but straightforward computations, we obtain

$$\bar{\Gamma}_{2,2}^{k,\ell} = \frac{h_\infty^2 C_{k,\ell}}{(r_\infty - 2\theta)(h_\infty(\pi_k + \pi_\ell) + 2\theta(\theta - r_\infty)) + \frac{h_\infty^2(\pi_k - \pi_\ell)^2}{2(r_\infty - 2\theta)}}\,,$$

and the result is proved.

3.7 Proofs for Section 3.4

3.7.1 Preliminaries

Most of the avoidance of traps results in the stochastic approximation literature deal with the case where the ODE that underlies the stochastic algorithm under study is an autonomous ODE $\dot{\mathsf{z}} = h(\mathsf{z})$. In this setting, a point $z_\star \in \operatorname{zer} h$ is called a trap if $h(z)$ admits an expansion around $z_\star$ of the type $h(z) = D(z - z_\star) + o(\|z - z_\star\|)$, where the matrix D has at least one eigenvalue which real part is (strictly) positive. Initiated by Pemantle [Pemantle 1990] and by Brandière and Duflo [Brandière & Duflo 1996], the most powerful class of techniques for establishing avoidance of traps results makes use of Poincaré's invariant manifold theorem

for the ODE $\dot{\mathsf{z}} = h(\mathsf{z})$ in a neighborhood of some point $z_\star \in \operatorname{zer} h$. The idea is to show that with probability 1, the stochastic algorithm strays away from the maximal invariant manifold of the ODE where the convergence to $z_\star$ of the ODE flow can take place. As previously mentioned, since we are dealing with algorithms derived from non-autonomous ODEs, we extend the results of [Pemantle 1990, Brandière & Duflo 1996] to this setting. The proof of Theorem 3.4.1 relies on a non-autonomous version of Poincaré's theorem. We borrow this result from the rich literature that exists on the subject [Dalec′kiĭ & Kreĭn 1974, Kloeden & Rasmussen 2011].

Let us start by setting the context for the non-autonomous version that we shall need for the invariant manifold theorem. Given an integer $d > 0$ and a matrix $D \in \mathbb{R}^{d\times d}$, consider the linear autonomous differential equation

$$\dot{\mathsf{z}}(t) = D\mathsf{z}(t), \tag{3.59}$$

which solution is obviously $\mathsf{z}(t) = e^{Dt}\mathsf{z}(0)$ for $t \in \mathbb{R}$. Let us factorize D as in (3.12), and write $D = Q\Lambda Q^{-1}$ with $\Lambda = \begin{bmatrix}\Lambda^- & \\ & \Lambda^+\end{bmatrix}$ where we recall that the Jordan blocks that constitute $\Lambda^- \in \mathbb{R}^{d^-\times d^-}$ (respectively $\Lambda^+ \in \mathbb{R}^{d^+\times d^+}$) are those that contain the eigenvalues λ_i of D such that $\Re\lambda_i \leqslant 0$ (respectively $\Re\lambda_i > 0$). Let us assume here that both d^- and d^+ are positive. It will be convenient to work in the basis of the columns of Q by making the variable change

$$z \mapsto y = \begin{bmatrix}y^-\\ y^+\end{bmatrix} = Q^{-1}z,$$

where $y^\pm \in \mathbb{R}^{d^\pm}$. In this new basis, the ODE (3.59) is written as

$$\begin{bmatrix}\dot{\mathsf{y}}^-\\ \dot{\mathsf{y}}^+\end{bmatrix} = \begin{bmatrix}\Lambda^- & \\ & \Lambda^+\end{bmatrix}\begin{bmatrix}\mathsf{y}^-\\ \mathsf{y}^+\end{bmatrix}, \tag{3.60}$$

which solution is $\mathsf{y}^\pm(t) = \exp(t\Lambda^\pm)\mathsf{y}^\pm(0)$. One can readily check that for each couple of real numbers α^+ and α^- that satisfy

$$0 < \alpha^- < \alpha^+ < \min\{\Re\lambda_i \,:\, \Re\lambda_i > 0\}, \tag{3.61}$$

there exists a so-called exponential dichotomy of the ODE solutions, which amounts in our case to the existence of two constants $K^-, K^+ \geqslant 1$ such that

$$\begin{aligned}\|\exp(t\Lambda^-)\| &\leqslant K^- e^{\alpha^- t} \quad \text{for } t \geqslant 0,\\ \|\exp(t\Lambda^+)\| &\leqslant K^+ e^{\alpha^+ t} \quad \text{for } t \leqslant 0,\end{aligned}$$

see, *e.g.*, [Horn & Johnson 1994].

We now consider a non-autonomous perturbation of this ODE, which is represented in the basis of the columns of Q as

$$\dot{\mathsf{y}}(t) = h(\mathsf{y}(t), t) \quad \text{with} \quad h(y,t) = \begin{bmatrix}\Lambda^- & \\ & \Lambda^+\end{bmatrix} y + \varepsilon(y,t), \tag{3.62}$$

and $\varepsilon : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is a continuous function. In the sequel, we shall be interested in the asymptotic behavior of this equation for the large values of t, and therefore, restrict our study to the interval $\mathbb{I} = [t_0, \infty)$ for some given $t_0 \geqslant 0$ that we shall fix later. We assume that $\varepsilon(0, \cdot) = 0$ on $\mathbb{I}$. We denote as $\phi : \mathbb{I} \times \mathbb{I} \times \mathbb{R}^d \to \mathbb{R}^d$ the so-called general solution of (3.62), which is defined by the fact that $\phi(\cdot, t, x)$ is the unique noncontinuable solution of (3.62) such that $\phi(t, t, x) = x$ for $t \in \mathbb{I}$ and $x \in \mathbb{R}^d$, assuming this solution exists and is unique for each $(x, t) \in \mathbb{R}^d \times \mathbb{I}$.

In the linear autonomous case provided by the ODE (3.60), the subspace

$$\mathcal{G} = \left\{ \left(t, \begin{bmatrix} y^- \\ 0 \end{bmatrix} \right) \in \mathbb{R} \times \mathbb{R}^d \ : \ y^- \in \mathbb{R}^{d^-} \right\}$$

is trivially invariant in the sense that if $(t, y) \in \mathcal{G}$, then, $(s, \phi(s, t, y)) \in \mathcal{G}$ for each $s \in \mathbb{R}$. This concept can be generalized to the non-linear and non-autonomous case. We say that the $\mathcal{C}^1$ function $w : \mathbb{R}^{d^-} \times \mathbb{I} \to \mathbb{R}^{d^+}$ defines a global non-autonomous invariant manifold for the ODE (3.62) if $w(0, t) = 0$ for all $t \in \mathbb{I}$, and, furthermore, if for each $t \in \mathbb{I}$ and each $y^- \in \mathbb{R}^{d^-}$, writing $y = (y^-, w(y^-, t))$, the general solution $\phi(s, t, y) = (\phi^-(s, t, y), \phi^+(s, t, y))$ with $\phi^\pm(s, t, y) \in \mathbb{R}^{d^\pm}$ verifies $\phi^+(s, t, y) = w(\phi^-(s, t, y), s)$ for each $s \in \mathbb{I}$. The non-autonomous invariant manifold is the set

$$\mathcal{G} = \left\{ \left(t, \begin{bmatrix} y^- \\ w(y^-, t) \end{bmatrix} \right) \in \mathbb{I} \times \mathbb{R}^d \ : \ y^- \in \mathbb{R}^{d^-} \right\},$$

which obviously satisfies $(t, y) \in \mathcal{G} \ \Rightarrow \ (s, \phi(s, t, y)) \in \mathcal{G}$ for each $s \in \mathbb{I}$.

These invariant manifolds are described by the following proposition, which is a straightforward application of [Pötzsche & Rasmussen 2006, Theorem A.1] (see also [Kloeden & Rasmussen 2011, Theorem 6.3 p. 106, Rem. 6.6 p. 111]). It is useful to note that under the conditions provided in the statement of this proposition, the existence of the general solution ϕ of the ODE (3.62) is ensured by Picard's theorem.

Proposition 3.7.1. *Let $\mathbb{I} = [t_0, \infty)$ for some $t_0 \geqslant 0$. Assume that the function $\varepsilon(y, t)$ is such that $\varepsilon(0, \cdot) \equiv 0$ on $\mathbb{I}$, the function $\varepsilon(\cdot, t)$ is continuously differentiable for each $t \in \mathbb{I}$, and furthermore, the Jacobian matrix $\partial_1 \varepsilon(y, t)$ satisfies*

$$|\varepsilon|_1 \triangleq \sup_{(y,t) \in \mathbb{R}^d \times \mathbb{I}} \|\partial_1 \varepsilon(y, t)\| < \frac{\alpha^+ - \alpha^-}{4K} \tag{3.63}$$

with $K = K^- + K^+ + K^- K^+ (K^- \vee K^+)$ and α^-, α^+ chosen as in Equation (3.61). Then, for each $\delta \in (2K|\varepsilon|_1, (\alpha^+ - \alpha^-)/2)$ and each $\gamma \in (\alpha^- + \delta, \alpha^+ - \delta)$, the set

$$\mathcal{G} = \left\{ (t, y) \in \mathbb{I} \times \mathbb{R}^d \ : \ \sup_{s \geqslant t} \|\phi(s, t, y)\| \exp(\gamma(t - s)) < \infty \right\}$$

is nonempty, and does not depend on γ. Moreover, this set is a global invariant manifold for the ODE (3.62) that is defined by a continuously differentiable mapping

$w : \mathbb{R}^{d^-} \times \mathbb{I} \to \mathbb{R}^{d^+}$. In addition, if the partial derivatives $\partial_1^k \varepsilon : \mathbb{R}^d \times \mathbb{I}$ exist and are continuous for $k \in \{1, \dots, m\}$ with globally bounded partial derivatives

$$|\varepsilon|_k \stackrel{\Delta}{=} \sup_{(y,t)\in\mathbb{R}^d\times\mathbb{I}} \|\partial_1^k \varepsilon(y,t)\| < \infty\,, \tag{3.64}$$

under the gap condition

$$m\alpha^- < \alpha^+,\ m \in \mathbb{N}^*, \tag{3.65}$$

the partial derivatives $\partial_1^k w : \mathbb{R}^{d^-} \times \mathbb{I}$ exist and are continuous with

$$\sup_{(y^-,t)\in\mathbb{R}^{d^-}\times\mathbb{I}} \|\partial_1^k w(y^-,t)\| < \infty \quad \textit{for all } k \in \{1,\dots,m\}. \tag{3.66}$$

Finally, if $\partial_2^n \partial_1^k \varepsilon$ exist and are continuous for $0 \leqslant n < m$ and $0 \leqslant k + n \leqslant m$, then w is m-times continuously differentiable.

Let us partition the function $h(y,t)$ as

$$h(y,t) = \begin{bmatrix} h^-(y,t) \\ h^+(y,t) \end{bmatrix} = \begin{bmatrix} \Lambda^- y^- + \varepsilon^-(y,t) \\ \Lambda^+ y^+ + \varepsilon^+(y,t) \end{bmatrix}, \tag{3.67}$$

where $h^\pm : \mathbb{R}^d \times \mathbb{I} \to \mathbb{R}^{d^\pm}$, $y^\pm \in \mathbb{R}^{d^\pm}$ and $\varepsilon^\pm : \mathbb{R}^d \times \mathbb{I} \to \mathbb{R}^{d^\pm}$. With these notations, the previous proposition leads to the following lemma.

Lemma 3.7.2. *In the setting of Proposition 3.7.1, for each t in the interior of $\mathbb{I}$ and each vector $y = (y^-, y^+)$ such that $y^\pm \in \mathbb{R}^{d^\pm}$ and $y^+ = w(y^-,t)$, it holds that*

$$h^+(y,t) = \partial_1 w(y^-,t) h^-(y,t) + \partial_2 w(y^-,t)\,. \tag{3.68}$$

Assume that α^- is small enough so that Inequality (3.65) and Equation (3.64) hold true with $m = 2$. Assume in addition that $\partial_2^n \partial_1^k \varepsilon$ exists and is continuous for $0 \leqslant n < 2$ and $0 \leqslant k + n \leqslant 2$, and furthermore, that there exists a bounded neighborhood $\mathcal{V} \subset \mathbb{R}^d$ of zero such that

$$\sup_{(y,t)\in\mathcal{V}\times\mathbb{I}} \|\partial_2 \varepsilon(y,t)\| < +\infty. \tag{3.69}$$

Then, there exists a neighborhood $\mathcal{V}^- \subset \mathbb{R}^{d^-}$ of zero such that

$$\sup_{(y^-,t)\in\mathcal{V}^-\times\mathbb{I}} \left\|\partial_1\partial_2 w(y^-,t)\right\| < +\infty\,, \tag{3.70}$$

$$\sup_{(y^-,t)\in\mathcal{V}^-\times\mathbb{I}} \left\|\partial_2^2 w(y^-,t)\right\| < +\infty\,. \tag{3.71}$$

Proof. By Proposition 3.7.1, the general solution $\phi(s,t,y)$ of the ODE (3.62) can be written as $\phi(s,t,y) = (\phi^-(s,t,y), \phi^+(s,t,y))$ with $\phi^+(s,t,y) = w(\phi^-(s,t,y), s)$ for each $s \in \mathbb{I}$. Equating the derivatives with respect to s of the two members of this equation and taking $s = t$, we get the first equation.

Writing $g : \mathbb{R}^{d-} \times \mathbb{I} \to \mathbb{R}^d, (y^-, t) \mapsto (y^-, w(y^-, t))$, Equation (3.68) can be rewritten as

$$\partial_2 w(y^-, t) = h^+(g(y^-, t), t) - \partial_1 w(y^-, t) h^-(g(y^-, t), t). \tag{3.72}$$

By Proposition 3.7.1, the function w is twice differentiable, and we can write

$$\partial_2^2 w(y^-, t) = \partial_1 h^+ \partial_2 g + \partial_2 h^+ - (\partial_1 \partial_2 w) h^- - (\partial_1 w)(\partial_1 h^- \partial_2 g + \partial_2 h^-), \tag{3.73}$$

where, *e.g.*, h^+ is a shorthand notation for $h^+(g(y^-, t), t)$. It holds from Equation (3.67) and the assumptions of Proposition 3.7.1 that for each $(y, t) \in \mathbb{R}^d \times \mathbb{I}$,

$$\|\partial_1 h(y, t)\| \leqslant \|\Lambda\| + \|\partial_1 \varepsilon(y, t)\| \leqslant C, \tag{3.74}$$

where the constant $C > 0$ is independent of (y, t) and can change from an inequality to another in the remainder of the proof. By the mean value inequality and Proposition 3.7.1, we also get that

$$\|w(y^-, t)\| = \|w(y^-, t) - w(0, t)\| \leqslant \sup_{(u,s)} \|\partial_1 w(u, s)\| \, \|y^- \| \leqslant C\|y^-\|,$$

thus, $\|g(y^-, t)\| \leqslant C\|y^-\|$. By the mean value inequality again,

$$\begin{aligned} \left\|h(g(y^-, t), t)\right\| = \left\|h(g(y^-, t), t) - h(0, t)\right\| &\leqslant \sup_{(u,t)} \|\partial_1 h(u, t)\| \left\|g(y^-, t)\right\| \\ &\leqslant C \left\|g(y^-, t)\right\| \leqslant C\|y^-\|. \end{aligned}$$

By Equation (3.72) and Proposition 3.7.1, this implies that

$$\left\|\partial_2 g(y^-, t)\right\| = \left\|\partial_2 w(y^-, t)\right\| = \left\|h^+ - (\partial_1 w) h^-\right\| \leqslant C \left\|y^-\right\|, \quad \text{and} \tag{3.75}$$

$$\left\|\partial_1 \partial_2 w(y^-, t)\right\| = \left\|\partial_1 h^+ \partial_1 g - (\partial_1^2 w) h^- - (\partial_1 w)(\partial_1 h^- \partial_1 g)\right\| \leqslant C(\left\|y^-\right\| + 1). \tag{3.76}$$

Let $\mathcal{V}^- \subset \mathbb{R}^{d-}$ be a small enough neighborhood of zero so that $g(y^-, t) \in \mathcal{V}$ for each $y^- \in \mathcal{V}^-$, which is possible by the inequality $\|g(y^-, t)\| \leqslant C\|y^-\|$. By the assumption on $\|\partial_2 \varepsilon(y, t)\|$ in the statement of Lemma 3.7.2, we have

$$\forall y^- \in \mathcal{V}^-, \quad \left\|\partial_2 h(g(y^-, t), t)\right\| = \left\|\partial_2 \varepsilon(g(y^-, t), t)\right\| \leqslant C. \tag{3.77}$$

The bound (3.70) is an immediate consequence of Equation (3.76). Getting back to Equation (3.73), the bound (3.71) follows from the inequalities (3.74)–(3.77). □

Proposition 3.7.1 deals with the case where the function ε is globally Lipschitz continuous. In practical cases, such a strong assumption is not necessarily verified. In particular, for the ODEs we consider for our application, it is not satisfied (see the function e defined in Subsec. 3.7.3.1 below). Nonetheless, recall that we only need the existence of a *local* non-autonomous invariant manifold, i.e. defined in the vicinity of an arbitrary solution such as the trivial zero solution (since we suppose

here $\varepsilon(0,\cdot) = 0$) whereas the aforementioned strong assumption provides a global non-autonomous invariant manifold. Indeed, as for the avoidance of traps result we intend to show, we will only need to look at the behavior of our ODE in the neighborhood of a trap $z_\star$. Therefore, in prevision of the proof of Theorem 3.4.1, we localize the ODE (3.62) in the neighborhood of zero. This is the purpose of the next proposition.

Proposition 3.7.3. *Let $\mathbb{I} = [t_0, +\infty)$ for some $t_0 \geqslant 0$ and let $h : \mathbb{R}^d \times \mathbb{I} \to \mathbb{R}^d$ be defined as in Equation (3.62). Assume that $\varepsilon(0,\cdot) \equiv 0$ on $\mathbb{I}$, that the function $\varepsilon(\cdot,t)$ is continuously differentiable for every $t \in \mathbb{I}$ and that*

$$\lim_{(y,t)\to(0,+\infty)} \|\partial_1 \varepsilon(y,t)\| = 0\,. \tag{3.78}$$

Then, there exist $\sigma > 0, t_1 > 0$, a function $\tilde{\varepsilon} : \mathbb{R}^d \times \mathbb{I}_1 \to \mathbb{R}^d$ where $\mathbb{I}_1 \triangleq [t_1, +\infty)$ and a function $\tilde{h} : \mathbb{R}^d \times \mathbb{I}_1 \to \mathbb{R}^d$ defined for every $y \in \mathbb{R}^d, t \in \mathbb{I}_1$ by $\tilde{h}(y,t) = \Lambda y + \tilde{\varepsilon}(y,t)$ s.t. $\tilde{h}$ and $\tilde{\varepsilon}$ verify the assumptions of Proposition 3.7.1 and for every $(y,t) \in B(0,\sigma)\times\mathbb{I}_1$, we have that $\tilde{h}(y,t) = h(y,t)$ and $\tilde{\varepsilon}(y,t) = \varepsilon(y,t)$. Moreover, for any $\delta > 0$, we can choose σ, t_1 respectively small and large enough s.t. the mapping $w : \mathbb{R}^{d^-} \times \mathbb{I}_1 \to \mathbb{R}^{d^+}$ obtained from Proposition 3.7.1 (applied to $\tilde{h}$ and $\tilde{\varepsilon}$) satisfies

$$|w|_1 = \sup_{(y,t)\in\mathbb{R}^{d^-}\times\mathbb{I}_1} \|\partial_1 w(y,t)\| < \delta\,. \tag{3.79}$$

Furthermore, Equation (3.68) holds for $\tilde{h}$ and w for all $(y,t) \in B(0,\sigma)\times\mathbb{I}_1$. If, additionally, Equation (3.69) holds for ε, then there exists $\sigma_1 \leqslant \sigma$ such that

$$\sup_{(y^-,t)\in B(0,\sigma_1)\times\mathbb{I}_1} \left\|\partial_1\partial_2 w(y^-,t)\right\| < +\infty\,, \tag{3.80}$$

$$\sup_{(y^-,t)\in B(0,\sigma_1)\times\mathbb{I}_1} \left\|\partial_2^2 w(y^-,t)\right\| < +\infty\,. \tag{3.81}$$

Proof. The idea of the proof is to *localize* the function $h(y,t)$ to a neighborhood of zero in the variable y for the purpose of applying Proposition 3.7.1. This cut-off technique is known in the non-autonomous ODE literature, see, *e.g.*, [Kloeden & Rasmussen 2011, Theorem 6.10]. Let $\psi : \mathbb{R}^d \to [0,1]$ be a smooth function such that $\psi(y) = 1$ if $\|y\| \leqslant 1$, and $\psi(y) = 0$ if $\|y\| \geqslant 2$. Let $C = \max_y \|\nabla\psi(y)\|$ where $\nabla\psi$ is the Jacobian matrix of ψ. Thanks to the convergence (3.78), we can choose $t_1 > 0$ large enough and $\sigma > 0$ small enough so that

$$\sup_{(t,y)\in[t_1,\infty)\times B(0,2\sigma)} \|\partial_1\varepsilon(y,t)\| < \frac{\alpha^+ - \alpha^-}{4K(1+2C)},$$

and we set $\mathbb{I}_1 = [t_1,\infty)$. Writing $\tilde{\varepsilon}(y,t) = \psi(y/\sigma)\varepsilon(y,t)$, it holds that for each $(t,y) \in \mathbb{I}_1 \times \mathbb{R}^d$,

$$\begin{aligned}
\|\partial_1\tilde{\varepsilon}(y,t)\| &\leqslant \sigma^{-1}C\mathbb{1}_{\|y\|\leqslant 2\sigma}\|\varepsilon(y,t)\| + \mathbb{1}_{\|y\|\leqslant 2\sigma}\|\partial_1\varepsilon(y,t)\| \\
&\leqslant \left(\max_{\|y\|\leqslant 2\sigma}\|\partial_1\varepsilon(y,t)\|\right)\left(\sigma^{-1}C\|y\| + 1\right)\mathbb{1}_{\|y\|\leqslant 2\sigma} \\
&\leqslant \frac{\alpha^+ - \alpha^-}{4K},
\end{aligned}$$

where we used the mean value inequality along with $\varepsilon(0,t) = 0$ to obtain the second inequality. Thus, the function $\tilde{h}(y,t) = \Lambda y + \tilde{\varepsilon}(y,t)$ satisfies all the assumptions of Proposition 3.7.1. In addition, the function $\tilde{\varepsilon}$ coincides with the function ε on $B(0,\sigma_1) \times \mathbb{I}_1$, and so it is for the functions $\tilde{h}$ and h. Finally, it follows from [Kloeden & Rasmussen 2011, Theorem 6.3] that

$$|w|_1 \leqslant \frac{2K^2}{\alpha_+ - \alpha_- - 4K|\tilde{\varepsilon}|_1}|\tilde{\varepsilon}|_1$$

(note that L in [Kloeden & Rasmussen 2011, Theorem 6.3] corresponds to $|\tilde{\varepsilon}|_1$ with our notations). Using Equation (3.78), we can make $|\tilde{\varepsilon}|_1$ as small as needed by choosing σ, t_1 respectively small and large enough, which gives us Equation (3.79). The proof of the last two equations follows from the application of Lemma 3.7.2 to $\tilde{h}$ and w. The result is immediate after noticing that for $(y,t) \in \mathbb{R}^d \times \mathbb{I}_1$, we have $\|\partial_2\tilde{\varepsilon}(y,t)\| \leqslant \|\partial_2\varepsilon(y,t)\|$. □

3.7.2 Proof of Theorem 3.4.1

We shall rely on the following result of Brandière and Duflo. Recall that $(\Omega, \mathscr{F}, \mathbb{P})$ is a probability space equipped with a filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$.

Proposition 3.7.4. *([Brandière & Duflo 1996, Proposition 4]) Given a sequence (γ_n) of deterministic nonnegative stepsizes such that $\sum_k \gamma_k = +\infty$ and $\sum_k \gamma_k^2 < +\infty$, consider the $\mathbb{R}^d$–valued stochastic process $(z_n)_{n\in\mathbb{N}}$ given by*

$$z_{n+1} = (I + \gamma_{n+1}H_n)z_n + \gamma_{n+1}\eta_{n+1} + \gamma_{n+1}\rho_{n+1}.$$

Assume that z_0 is $\mathscr{F}_0$–measurable and that the sequences (η_n), (ρ_n) together with the sequence of random matrices (H_n) are $(\mathscr{F}_n)$–adapted. Moreover, on a given event $A \in \mathscr{F}$, assume the following facts:

i) $\sum_n \|\rho_n\|^2 < \infty$.

ii) $\limsup \mathbb{E}[\|\eta_{n+1}\|^{2+a} \,|\, \mathscr{F}_n] < \infty$ *for some* $a > 0$, *and* $\mathbb{E}[\eta_{n+1} \,|\, \mathscr{F}_n] = 0$.

iii) $\liminf \mathbb{E}[\|\eta_{n+1}\|^2 \,|\, \mathscr{F}_n] > 0$.

Let $H \in \mathbb{R}^{d\times d}$ be a deterministic matrix such that the real parts of its eigenvalues are all positive. Then,

$$\mathbb{P}\left(A \cap [z_n \to 0] \cap [H_n \to H]\right) = 0.$$

We now enter the proof of Theorem 3.4.1. Recall the development (3.11) of $b(z,t)$ near $z_\star$ and the spectral factorization (3.12) of the matrix D. To begin with, it will be convenient to make the variable change $y = Q^{-1}(z - z_\star)$, and set

$$h(y,t) = Q^{-1}b(Qy + z_\star, t) = \Lambda y + \tilde{e}(y,t),$$

with $\tilde{e}(y,t) = Q^{-1}e(Qy + z_\star, t)$, in such a way that our stochastic algorithm is rewritten as

$$y_{n+1} = y_n + \gamma_{n+1}h(y_n, \tau_n) + \gamma_{n+1}\tilde{\eta}_{n+1} + \gamma_{n+1}\tilde{\rho}_{n+1}$$

where $\tilde{\eta}_n$ is as in the statement of the theorem and $\tilde{\rho}_n = Q^{-1}\rho_n$. Observe that the assumptions on the function e in the statement of the theorem remain true for $\tilde{e}$ with $z_\star$ replaced by zero.

If the matrix Λ has only eigenvalues with (strictly) positive real parts, *i.e.*, $d^- = 0$, then we can apply Proposition 3.7.4 to the sequence (z_n). Henceforth, we deal with the more complicated case where $d^- > 0$.

Apply Proposition 3.7.3 to h to obtain $\tilde{h}$ and σ, t_1 respectively small and large enough and $w : \mathbb{R}^{d^-} \times \mathbb{I}_1 \to \mathbb{R}^{d^+}$ where $\mathbb{I}_1 := [t_1, +\infty)$. By Assumption *iv)* of Theorem 3.4.1 and Proposition 3.7.3 we can choose $\sigma_1 \leqslant \sigma$ such that Equation (3.80) and Equation (3.81) hold. Now, given $p \in \mathbb{N}$, let us define the event

$$E_p = \left[\forall n \geqslant p,\ \|y_n\| < \sigma_1, \tau_n \in \mathbb{I}_1\right].$$

On E_p, it holds that $h(y_n, \tau_n) = \tilde{h}(y_n, \tau_n)$ and

$$\begin{aligned}\forall n \geqslant p, \quad y_{n+1} &= y_n + \gamma_{n+1}h(y_n, \tau_n) + \gamma_{n+1}\tilde{\eta}_{n+1} + \gamma_{n+1}\tilde{\rho}_{n+1} \\ &= \begin{bmatrix} y_n^- \\ y_n^+ \end{bmatrix} + \gamma_{n+1}\begin{bmatrix} h^-(y_n, \tau_n) \\ h^+(y_n, \tau_n) \end{bmatrix} + \gamma_{n+1}\begin{bmatrix} \tilde{\eta}_{n+1}^- \\ \tilde{\eta}_{n+1}^+ \end{bmatrix} + \gamma_{n+1}\begin{bmatrix} \tilde{\rho}_{n+1}^- \\ \tilde{\rho}_{n+1}^+ \end{bmatrix} \end{aligned} \tag{3.82}$$

where h is partitioned as in (3.67), and where $\tilde{\eta}_n^\pm, \tilde{\rho}_n^\pm \in \mathbb{R}^{d^\pm}$. Note that, by Proposition 3.7.3 and Assumptions *vi)* and *vii)* on the sequence (η_n), we can choose σ, t_1 respectively small and large enough such that

$$\liminf \mathbb{E}[\left\|\tilde{\eta}_{n+1}^+\right\|^2 |\mathscr{F}_n]\mathbb{1}_{E_p}(y_n) - 2\limsup \mathbb{E}[\left\|\partial_1 w(y_n^-, \tau_n)\tilde{\eta}_{n+1}^-\right\|^2 |\mathscr{F}_n]\mathbb{1}_{E_p}(y_n) > \frac{c^2}{2}. \tag{3.83}$$

This inequality will be important in the end of our proof. Let t be in the interior of $\mathbb{I}_1$, and let $y = (y^-, y^+)$ be in a neighborhood of 0. Make the variable change $(y^-, y^+) \mapsto (u^-, u^+)$ with

$$\begin{aligned} u^+ &= y^+ - w(y^-, t), \\ u^- &= y^-, \end{aligned}$$

where w is the function defined in the statement of Proposition 3.7.3, and let

$$\begin{aligned} W(u^-, u^+, t) &= h^+(y,t) - \partial_1 w(y^-, t)h^-(y,t) - \partial_2 w(y^-, t) \\ &= h^+((u^-, u^+ + w(u^-, t)), t) \\ &\quad - \partial_1 w(u^-, t)h^-((u^-, u^+ + w(u^-, t)), t) - \partial_2 w(u^-, t). \end{aligned}$$

By Proposition 3.7.3 and Lemma 3.7.2, it holds that $W(u^-, 0, t) = 0$. Moreover, $W(u^-, \cdot, t) \in \mathcal{C}^1$ by the assumptions on h. Therefore, writing $y(r) = (u^-, ru^+ +$

$w(u^-,t))$ for $r\in[0,1]$, and using the decomposition (3.67), we get that

$$\begin{aligned}W(u^-,u^+,t) &= \int_0^1 \partial_2 W(u^-,ru^+,t)u^+\,dr\\ &= \Lambda^+u^+\\ &\quad+\int_0^1\left(\partial_1\varepsilon^+(y(r),t)\begin{bmatrix}0\\ I_{d^+}\end{bmatrix}-\partial_1 w(u^-,t)\partial_1\varepsilon^-(y(r),t)\begin{bmatrix}0\\ I_{d^+}\end{bmatrix}\right)u^+dr.\end{aligned}$$

We can also write $y(r)=(y^-,ry^++(1-r)w(y^-,t))$. Recalling that $w(0,t)=0$ and that $\|\partial_1 w(y^-,t)\|$ is bounded on $\mathbb{R}^{d^-}\times\mathbb{I}$, we get by the mean value inequality that $\|w(y^-,t)\|\leqslant C\|y^-\|$ where $C>0$ is a constant. Thus, $\|y(r)\|\leqslant(1+C)\|y\|$. Moreover, $\varepsilon(y,t)=Q^{-1}e(Qy,t)$ for $\|y\|<\sigma$. Thus, we get by (3.13) that $\|\partial_1\varepsilon(y(r),t)\|\to 0$ as $(y,t)\to(0,\infty)$ uniformly in $r\in[0,1]$. Using again the boundedness of $\|\partial_1 w(\cdot,\cdot)\|$, we eventually obtain that

$$W(u^-,u^+,t)=\big(\Lambda^++\Delta(y,t)\big)\,u^+,\quad\text{with}\quad\lim_{(y,t)\to(0,\infty)}\Delta(y,t)=0.$$

On the event E_p, assume that $n\geqslant p$, and write

$$u_n^+=y_n^+-w(y_n^-,\tau_n),\quad u_n^-=y_n^-,$$

(see Equation (3.82)). Choosing $\alpha_->0$ small enough so that the gap condition (3.65) is satisfied with $m=2$, we have by Taylor's expansion

$$\begin{aligned}w(y_{n+1}^-,\tau_{n+1})-w(y_n^-,\tau_n) &= w(y_{n+1}^-,\tau_{n+1})-w(y_n^-,\tau_{n+1})+w(y_n^-,\tau_{n+1})-w(y_n^-,\tau_n)\\ &= \partial_1 w(y_n^-,\tau_{n+1})(y_{n+1}^--y_n^-)+\gamma_{n+1}\partial_2 w(y_n^-,\tau_n)+\epsilon_{n+1}+\epsilon_{n+1}^\gamma\,,\end{aligned}$$

$$\begin{aligned}\text{with}\quad \|\epsilon_{n+1}\| &\leqslant \sup_{y^-\in[y_n^-,y_{n+1}^-]}\big\|\partial_1^2 w(y^-,\tau_{n+1})\big\|\,\big\|y_{n+1}^--y_n^-\big\|^2\,,\\ \text{and}\quad \big\|\epsilon_{n+1}^\gamma\big\| &\leqslant \sup_{\tau\in[\tau_n,\tau_{n+1}]}\big\|\partial_2^2 w(y_n^-,\tau)\big\|\,\gamma_{n+1}^2\,.\end{aligned}$$

Using this equation, we obtain

$$\begin{aligned}u_{n+1}^+-u_n^+ &= \gamma_{n+1}W(u_n^-,u_n^+,\tau_n)+\gamma_{n+1}\big(\tilde\eta_{n+1}^+-\partial_1 w(y_n^-,\tau_{n+1})\tilde\eta_{n+1}^-\big)\\ &\quad+\gamma_{n+1}\big(\tilde\rho_{n+1}^+-\partial_1 w(y_n^-,\tau_{n+1})\tilde\rho_{n+1}^-\big)-\epsilon_{n+1}-\epsilon_{n+1}^\gamma\\ &\quad+\gamma_{n+1}\big(\partial_1 w(y_n^-,\tau_n)-\partial_1 w(y_n^-,\tau_{n+1})\big)\,h^-(y_n,\tau_n)\,,\end{aligned}$$

which leads to

$$u_{n+1}^+=u_n^++\gamma_{n+1}\big(\Lambda^++\Delta(y_n,\tau_n)\big)\,u_n^++\gamma_{n+1}\bar\eta_{n+1}+\gamma_{n+1}\bar\rho_{n+1},\tag{3.84}$$

with $\bar\eta_{n+1}=\tilde\eta_{n+1}^+-\partial_1 w(y_n^-,\tau_n)\tilde\eta_{n+1}^-$ and

$$\begin{aligned}\bar\rho_{n+1} &= \tilde\rho_{n+1}^+-\partial_1 w(y_n^-,\tau_n)\tilde\rho_{n+1}^--\mathbb{1}_{\gamma_{n+1}>0}\frac{\epsilon_{n+1}+\epsilon_{n+1}^\gamma}{\gamma_{n+1}}\\ &\quad+\big(\partial_1 w(y_n^-,\tau_n)-\partial_1 w(y_n^-,\tau_{n+1})\big)\,h^-(y_n,\tau_n)\,.\end{aligned}\tag{3.85}$$

To finish the proof, it remains to check that the noise sequence satisfies the assumptions of Proposition 3.7.4 on the event $A_p = E_p \cap [y_n \to 0]$. In the remainder, C' will indicate some positive constant which can change from an inequality to another one.

First, we verify that $\sum_n \|\bar{\rho}_n\|^2 < \infty$ on A_p by controlling each one of the terms of $\bar{\rho}_n$. Combining the boundedness of $\partial_1 w(\cdot,\cdot)$ with the summability assumption $\sum_n \|\tilde{\rho}_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}} < +\infty$ a.s., we immediately obtain on A_p that $\sum_n \|\tilde{\rho}^+_{n+1} - \partial_1 w(y^-_n, \tau_n)\tilde{\rho}^-_{n+1}\|^2 < +\infty$ given our choice of σ. Moreover, it holds that $\left(\|\epsilon^\gamma_{n+1}\|/\gamma_{n+1}\right)^2 \leqslant C'\gamma^2_{n+1}$ by invoking Proposition 3.7.3. In addition, using the boundedness of $\partial_1^2 w(\cdot,\cdot)$, we can write

$$\mathbb{1}_{\gamma_{n+1}>0}\left\|\frac{\epsilon_{n+1}}{\gamma_{n+1}}\right\|^2 \leqslant \mathbb{1}_{\gamma_{n+1}>0}\frac{C'}{\gamma^2_{n+1}}\|y_{n+1}-y_n\|^4 \leqslant C'\gamma^2_{n+1}(\|h(y_n,\tau_n)\|^4 + \|\tilde{\eta}_{n+1}\|^4 + \|\tilde{\rho}_{n+1}\|^4)\,.$$

A coupling argument (see [Brandière & Duflo 1996, p. 401]) shows that we can simplify the condition
$\limsup \mathbb{E}[\|\eta_{n+1}\|^4 \,|\, \mathscr{F}_n]\mathbb{1}_{z_n\in\mathcal{W}} < \infty$ to $\mathbb{E}[\|\eta_{n+1}\|^4 \,|\, \mathscr{F}_n]\mathbb{1}_{z_n\in\mathcal{W}} < C'$. The latter condition implies that $\mathbb{E}[\mathbb{1}_{A_p}\sum_n \gamma^2_{n+1}\|\eta_{n+1}\|^4] \leqslant \sum_n C'\gamma^2_{n+1}$, and therefore $\sum_n \gamma^2_{n+1}\|\eta_{n+1}\|^4\,\mathbb{1}_{A_p} < +\infty$ a.s. As a consequence, noticing also the boundedness of $(h(y_n,\tau_n))$ and $(\tilde{\rho}_n)$ on A_p, we deduce that $\sum_n \mathbb{1}_{\gamma_{n+1}>0}\left\|\frac{\epsilon_{n+1}}{\gamma_{n+1}}\right\|^2 < +\infty$ on A_p. We now briefly control the last term of $\bar{\rho}_n$. By the mean value inequality, we obtain that

$$\left\|\left(\partial_1 w(y^-_n,\tau_n) - \partial_1 w(y^-_n,\tau_{n+1})\right)h^-(y_n,\tau_n)\right\| \leqslant \gamma_{n+1}\sup_{(y^-,t)}\left\|\partial_2\partial_1 w(y^-,t)\right\|\|h^-(y_n,\tau_n)\| \leqslant C'\gamma_{n+1}\,,$$

where the last inequality stems from Proposition 3.7.3-Equation (3.80) together with the boundedness of the sequence $(h(y_n,\tau_n))$. In view of Equation (3.85) and the above estimates, we deduce that $\sum_n \|\bar{\rho}_{n+1}\|^2\mathbb{1}_{A_p} < +\infty$ a.s. on A_p.

We verify the remaining conditions on the noise sequence $(\bar{\eta}_n)$. We can easily remark that $\mathbb{E}[\bar{\eta}_{n+1}|\mathscr{F}_n] = 0$ and $\|\bar{\eta}_{n+1}\| \leqslant C'\|\eta_{n+1}\|$ on A_p. Hence, $\limsup \mathbb{E}[\|\bar{\eta}_{n+1}\|^4 \,|\, \mathscr{F}_n]\mathbb{1}_{z_n\in\mathcal{W}} < \infty$. The last condition, meaning that the noise is exciting enough, stems from noting that

$$\begin{aligned}2\liminf \mathbb{E}[\|\bar{\eta}_{n+1}\|^2\,|\mathscr{F}_n]\mathbb{1}_{A_p} &\geqslant \liminf \mathbb{E}[\left\|\tilde{\eta}^+_{n+1}\right\|^2|\mathscr{F}_n]\mathbb{1}_{A_p} \\ &\quad - 2\limsup \mathbb{E}[\left\|\partial_1 w(y^-_n,\tau_n)\tilde{\eta}^-_{n+1}\right\|^2|\mathscr{F}_n]\mathbb{1}_{A_p} \\ &> \frac{c^2}{2}\,,\end{aligned}$$

where we used our choice of σ, t_1 and Equation (3.83).

Noticing that $[y_n \to 0] \subset [\Delta(y_n,\tau_n) \to 0]$, we can now apply Proposition 3.7.4 to the sequence (u^+_n) (see Equation (3.84)) with $A = A_p$ to obtain

$$\mathbb{P}\left(A_p \cap [u^+_n \to 0]\right) = \mathbb{P}\left(A_p \cap [u^+_n \to 0] \cap [\Delta(y_n,\tau_n) \to 0]\right) = 0\,.$$

We now show that $[y_n \to 0] \subset [u_n^+ \to 0]$, which amounts to prove that $w(y_n^-, \tau_n) \to 0$ given $y_n \to 0$. To that end, upon noting that $w(0, \cdot) \equiv 0$ and that $\partial_1 w(\cdot, \cdot)$ is bounded, it suffices to apply the mean value inequality, writing :

$$\|w(y_n^-, \tau_n)\| = \|w(y_n^-, \tau_n) - w(0, \tau_n)\| \leqslant \sup_{(y^-, t)} \|\partial_1 w(y^-, t)\| \, \|y_n^-\| \leqslant K \|y_n^-\| \, .$$

We have shown so far that $\mathbb{P}(A_p) = 0$. Since $y_n = Q^{-1} z_n$ and $[y_n \to 0] \subset \bigcup_{p \in \mathbb{N}} E_p$, we finally obtain that

$$\mathbb{P}[z_n \to 0] = \mathbb{P}[y_n \to 0] = \mathbb{P}\left(\bigcup_{p \in \mathbb{N}} ([y_n \to 0] \cap E_p) \right) = \mathbb{P}\left(\bigcup_{p \in \mathbb{N}} A_p \right) = 0.$$

Theorem 3.4.1 is proven.

3.7.3 Proofs for Section 3.4.2.1

3.7.3.1 Proof of Lemma 3.4.2

The matrix D coincides with $\nabla g_\infty(z_\star)$, where the function g_∞ is defined in (3.20). As such, its expression is immediate. Recalling that $p_\infty S(x_\star) - q_\infty v_\star = 0$, we get

$$\begin{aligned}
g(z,t) - D(z - z_\star) &= \begin{bmatrix} \mathsf{p}(t) S(x) - \mathsf{q}(t) v - p_\infty \nabla S(x_\star)(x - x_\star) + q_\infty (v - v_\star) \\ \mathsf{h}(t) \nabla F(x) - \mathsf{r}(t) m - h_\infty \nabla^2 F(x_\star)(x - x_\star) + r_\infty m \\ -m \left((v + \varepsilon)^{-\frac{1}{2}} - (v_\star + \varepsilon)^{-\frac{1}{2}} \right) \end{bmatrix} \\
&= \begin{bmatrix} -\mathsf{q}(t) + q_\infty & 0 & (\mathsf{p}(t) - p_\infty) \nabla S(x_\star) \\ 0 & r_\infty - \mathsf{r}(t) & (\mathsf{h}(t) - h_\infty) \nabla^2 F(x_\star) \\ \frac{m}{2(v_\star + \varepsilon)^{\frac{3}{2}}} & 0 & 0 \end{bmatrix} \begin{bmatrix} v - v_\star \\ m \\ x - x_\star \end{bmatrix} \\
&\quad + \begin{bmatrix} \mathsf{p}(t)(S(x) - S(x_\star) - \nabla S(x_\star)(x - x_\star)) \\ \mathsf{h}(t)(\nabla F(x) - \nabla^2 F(x_\star)(x - x_\star)) \\ -m \odot \left(\frac{1}{\sqrt{v + \varepsilon}} - \frac{1}{\sqrt{v_\star + \varepsilon}} + \frac{v - v_\star}{2(v_\star + \varepsilon)^{\frac{3}{2}}} \right) \end{bmatrix} + \begin{bmatrix} \mathsf{p}(t) S(x_\star) - \mathsf{q}(t) v_\star \\ 0 \\ 0 \end{bmatrix} \\
&\triangleq e(z,t) + c(t).
\end{aligned}$$

Under the assumptions made, it is easy to see that the function $e(z,t)$ has the properties required in the statement of Theorem 3.4.1.

3.7.3.2 Proof of Theorem 3.4.3

Consider the matrix D defined in the statement of Lemma 3.4.2. A spectral analysis of this matrix as regards its eigenvalues with positive real parts is done in the following lemma.

Lemma 3.7.5. *Let D be the matrix provided in the statement of Lemma 3.4.2. Each eigenvalue ζ of the matrix D such that $\Re\zeta > 0$ is real, and its algebraic and geometric multiplicities are equal. Moreover, there is a one-to-one correspondence*

φ between these eigenvalues and the negative eigenvalues of $V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$. Let d^+ be the dimension of the eigenspace of $V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$ that is associated with its negative eigenvalues, let

$$W = \begin{bmatrix} w_1 \\ \vdots \\ w_{d^+} \end{bmatrix} \in \mathbb{R}^{d^+ \times d}$$

be a matrix which rows are independent eigenvectors of $V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$ that generate this eigenspace, and denote as $\beta_k < 0$ the eigenvalue associated with w_k. Then, the rows of the rank d^+-matrix

$$A^+ = \left[0_{d^+ \times d}, \quad WV^{\frac{1}{2}}, \quad -\operatorname{diag}(r_\infty + \varphi^{-1}(\beta_k))WV^{-\frac{1}{2}}\right] \in \mathbb{R}^{d^+ \times 3d}$$

generate the left eigenspace of D, the row k being an eigenvector for the eigenvalue $\varphi^{-1}(\beta_k)$.

Proof. It is obvious that the block lower-triangular matrix D has d eigenvalues equal to $-q_\infty$ and $2d$ eigenvalues which are those of the sub-matrix

$$\widetilde{D} = \begin{bmatrix} -r_\infty I_d & h_\infty \nabla^2 F(x_\star) \\ -V & 0 \end{bmatrix}.$$

Given $\lambda \in \mathbb{C}$, we obtain by standard manipulations involving determinants that

$$\det(\widetilde{D} - \lambda) = \det(\lambda(r_\infty + \lambda) + h_\infty V \nabla^2 F(x_\star)) = \det(\lambda(r_\infty + \lambda) + h_\infty V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}}).$$

Denoting as $\{\beta_k\}_{k=1}^d$ the eigenvalues of $h_\infty V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$ counting the multiplicities, we obtain from the last equation that the eigenvalues of $\widetilde{D}$ are the solutions of the second order equations

$$\lambda^2 + r_\infty \lambda + \beta_k = 0, \quad k = 1, \dots, d.$$

The product of the roots of such an equation is β_k, and their sum is $-r_\infty \leqslant 0$. Thus, denoting as $\zeta_{k,1}$ and $\zeta_{k,2}$ these roots, it is easy to see that if $\beta_k \geqslant 0$, then $\Re\zeta_{k,1}, \Re\zeta_{k,2} \leqslant 0$, while if $\beta_k < 0$, then both $\zeta_{k,i}$ are real, and only one of them is positive. Thus, we have so far shown that the eigenvalues of D which real parts are positive are themselves real, and there is a one-to-one map φ from the set of positive eigenvalues of D to the set of negative eigenvalues of $V^{\frac{1}{2}}\nabla^2 F(x_\star)V^{\frac{1}{2}}$. Moreover, the algebraic multiplicity of the eigenvalue $\zeta > 0$ of D is equal to the multiplicity of $\varphi(\zeta)$.

Let us now turn to the left (row) eigenvectors of D that correspond to these eigenvalues. To that end, we shall solve the equation

$$uD = \zeta u \quad \text{with } u = [0, u_1, u_2], \quad u_{1,2} \in \mathbb{R}^{1 \times d}, \tag{3.86}$$

for a given eigenvalue $\zeta > 0$ of D. Developing this equation, we get

$$-r_\infty u_1 - u_2 V = \zeta u_1, \quad h_\infty u_1 \nabla^2 F(x_\star) = \zeta u_2.$$

If we now write $\tilde{u}_1 = u_1 V^{-\frac{1}{2}}$ and $\tilde{u}_2 = u_2 V^{\frac{1}{2}}$, this system becomes

$$-r_\infty \tilde{u}_1 - \tilde{u}_2 = \zeta \tilde{u}_1, \quad h_\infty \tilde{u}_1 V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}} = \zeta \tilde{u}_2,$$

or, equivalently,

$$\tilde{u}_2 = -(r_\infty + \zeta)\tilde{u}_1, \quad \tilde{u}_1 \left(\zeta^2 + r_\infty \zeta + h_\infty V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}} \right) = 0,$$

which shows that $\tilde{u}_1$ is a left eigenvector of $V^{\frac{1}{2}} \nabla^2 F(x_\star) V^{\frac{1}{2}}$ associated with the eigenvalue $\varphi(\zeta)$. What's more, assume that r is the multiplicity of $\varphi(\zeta)$, and, without generality loss, that the submatrix $W_{r,\cdot}$ made of the first r rows of W generates the left eigenspace of $\varphi(\zeta)$. Then, the matrix

$$\begin{bmatrix} 0_{r \times d} & W_{r\cdot} V^{\frac{1}{2}} & -(r_\infty + \zeta) W_{r\cdot} V^{-\frac{1}{2}} \end{bmatrix}$$

is a r-rank matrix which rows are independent left eigenvectors that generate the left eigenspace of D for the eigenvalue ζ. In particular, the algebraic and geometric multiplicities of this eigenvalue are equal. The same argument can be applied to the other positive eigenvalues of D. □

We now have all the elements to prove Theorem 3.4.3. Recall Equation (3.14):

$$z_{n+1} = z_n + \gamma_{n+1} b(z_n, \tau_n) + \gamma_{n+1} \eta_{n+1} + \gamma_{n+1} \rho_{n+1},$$

where $b(z,t) = g(z,t) - c(t) = D(z - z_\star) + e(z,t)$ and $\rho_n = c(\tau_{n-1}) + \tilde{\rho}_n$. With these same notations, we check that Assumptions *i)–vi)* in the statement of Theorem 3.4.1 are satisfied. The function $e(z,t)$ satisfies Assumptions *i)–iv)* by Lemma 3.4.2. We now verify that the sequence (ρ_n) fulfills Assumption *v)*. First, observe that $\sum_n \|c(\tau_n)\|^2 < \infty$ under Assumption 3.4.3-*i)*. Then, we control the second term $(\tilde{\rho}_n)$. After straightforward derivations, one can show the existence of a positive constant C (depending only on ε and a neighborhood $\mathcal{W}$ of $z_\star$) such that

$$\|\tilde{\rho}_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}} \leqslant C(\|m_n - m_{n+1}\|^2 + \|v_{n+1} - v_n\|^2) \mathbb{1}_{z_n \in \mathcal{W}}\,. \tag{3.87}$$

Using the boundedness of the sequences (h_n) and (r_n) together with the update rule of m_n and Assumption 3.4.3-*iii)*, there exists a positive constant C' independent of n (which may change from an inequality to another) such that

$$\mathbb{E}\left[\|m_n - m_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}}\right] \leqslant \gamma_{n+1}^2 C' \mathbb{E}\left[(1 + \mathbb{E}_\xi\left[\|\nabla f(x_n, \xi)\|^2\right]) \mathbb{1}_{z_n \in \mathcal{W}}\right] \leqslant C' \gamma_{n+1}^2\,. \tag{3.88}$$

A similar result holds for $\mathbb{E}\left[\|v_n - v_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}}\right]$ following the same arguments. In view of Eqs. (3.87)-(3.88) and the assumption $\sum_n \gamma_{n+1}^2 < +\infty$, it holds that $\mathbb{E}\left[\sum_n \|\tilde{\rho}_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}}\right] < +\infty$. Therefore, $\sum_n \|\tilde{\rho}_{n+1}\|^2 \mathbb{1}_{z_n \in \mathcal{W}} < +\infty$ a.s., which completes our verification of condition *v)* of Theorem 3.4.1. Assumption *vi)* follows from condition 3.4.3-*iii)*. Finally, let us make Assumption *vii)* of Theorem 3.4.1 more explicit. Partitioning the matrix Q^{-1} as $Q^{-1} = \begin{bmatrix} B^- \\ B^+ \end{bmatrix}$ where $B^\pm$ has $d^\pm$ rows,

Lemma 3.7.5 shows that the row spaces of B^+ and A^+ are the same, which implies that Assumption *vii)* can be rewritten equivalently as $\mathbb{E}[\|A^+\eta_{n+1}\|^2 \mid \mathscr{F}_n]\mathbb{1}_{z_n\in\mathcal{W}} \geqslant c^2\mathbb{1}_{z_n\in\mathcal{W}}$. By inspecting the form of η_n provided by Equation (3.28) (written as a column vector), one can readily check that Assumption 3.4.3-*iv)* implies Assumption *vii)* of Theorem 3.4.1 for a small enough neighborhood $\mathcal{W}$, using the continuity of the covariance matrix $V^{\frac{1}{2}}\mathbb{E}_\xi(\nabla f(x,\xi)-\nabla F(x))(\nabla f(x,\xi)-\nabla F(x))^{\mathrm{T}}V^{\frac{1}{2}}$ when x is near $x_\star$.

3.7.4 Proof of Theorem 3.4.4

As mentioned in Section 3.4.2.2, the proof of Theorem 3.4.4 is almost identical to the one of Theorem 3.4.3. We point out the main differences here. In Lemma 3.4.2, replace D by $\tilde{D} = \begin{bmatrix} 0 & h_\infty\nabla^2 F(x_\star) \\ -I_d & 0 \end{bmatrix}$ and set $c(t) = 0$. Then, in Lemma 3.7.5, replace the matrix $V^{1/2}\nabla^2 F(x_\star)V^{1/2}$ by the Hessian $\nabla^2 F(x_\star)$.

Chapter 4

Constant step stochastic approximation involving the Clarke subdifferentials of non smooth functions

4.1 Introduction

In this chapter, we study the asymptotic behavior of the constant step Stochastic Gradient Descent (SGD) when the objective function is neither differentiable nor convex. Given an integer $d \geqslant 1$ and a probability space $(\Xi, \mathscr{T}, \mu)$, let $f : \mathbb{R}^d \times \Xi \to \mathbb{R}, (x, s) \mapsto f(x, s)$ be a function which is assumed to be locally Lipschitz, generally non-differentiable and non-convex in the variable x, and μ-integrable in the variable s. The goal is to find a local minimum, or at least a critical point of the function $F(x) = \int f(x, s)\, \mu(ds) = \mathbb{E}f(x, \cdot)$, *i.e.*, a point $x_\star$ such that $0 \in \partial F(x_\star)$, where ∂F is the so-called Clarke subdifferential of F. It is assumed that the function f is available to the observer along with a sequence of independent Ξ-valued random variables $(\xi_k)_{k\in\mathbb{N}}$ on some probability space with the same probability law μ. The function F itself is assumed unknown due to, *e.g.*, the difficulty of computing the integral $\mathbb{E}f(x, \cdot)$. Such non-smooth and non-convex problems are frequently encountered in the field of statistical learning. For instance this type of problem arises in the study of neural networks when the activation function is non-smooth, which is the case of the commonly used ReLU function.

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

$$x_{n+1} = x_n - \gamma \nabla f(x_n, \xi_{n+1}) \tag{4.1}$$

for every n such that the function $f(\cdot, \xi_{n+1})$ is differentiable at point x_n, where $\nabla f(x_n, \xi_{n+1})$ represents the gradient w.r.t. the variable x_n. When $f(\cdot, \xi_{n+1})$ is non-differentiable at x_n, the update equation $x_n \to x_{n+1}$ is left undefined. The practioner is free to choose the value of x_{n+1} according to a predetermined selection policy. Typically, a reasonable choice is to select x_{n+1} in the set $x_n - \gamma \partial f(x_n, \xi_{n+1})$, where $\partial f(x, s)$ represents the Clarke subdifferential of the function $f(\cdot, s)$ at the point x. When such a policy is used, the resulting sequence will be referred to as a *Clarke-SGD* sequence. A second option used by practioners is to compute the

derivative using the automatic differentiation provided in popular API's such as Tensorflow, PyTorch, etc. *i.e.*, for all n,

$$x_{n+1} = x_n - \gamma a_{f(\cdot,\xi_{n+1})}(x_n) \tag{4.2}$$

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

$$f(x, s) = \ell(\sigma_L(W_L\sigma_{L-1}(W_{L-1}\cdots\sigma_1(W_1X_s))), Y_s)\,, \tag{4.3}$$

where $x = (W_1, \cdots, W_L)$ are the weights of the network represented by a finite sequence of L matrices, $\sigma_1, \cdots, \sigma_L$ are vector-valued functions, X_s is a feature vector, Y_s is a label and $\ell(\cdot,\cdot)$ is some loss function. In such a case, the automatic differentiation is computed using the chain rule of function differentiation, by means of the celebrated backpropagation algorithm. When the mappings $\sigma_1, \cdots, \sigma_L, \ell(\cdot, Y_s)$ are differentiable, the chain rule indeed applies and the output coincides with the gradient. However, the chain rule fails in case of non-differentiable functions. The properties of the map a_h are studied in the recent work [Bolte & Pauwels 2019]. In general, $a_h(x)$ may not be an element of the Clarke-subdifferential $\partial h(x)$. It can even happen that $a_h(x) \neq \nabla h(x)$ at some points x where h is differentiable. However, the set of such peculiar points is proved to be Lebesgue negligible. As a consequence, if the initial point x_0 is chosen random according to some density w.r.t. the Lebesgue measure, an autograd sequence can be shown to be a SGD sequence in the sense of Equation (4.1) under some conditions.

The aim of this chapter is to analyze the asymptotic behavior of SGD sequences in the case where the step γ is constant.

About the literature. In two recent papers [Majewski *et al.* 2018] and [Davis *et al.* 2020], a closely related algorithm is analyzed under the assumption that the step size is vanishing, *i.e.*, γ is replaced with a sequence (γ_n) that tends to zero as $n \to \infty$. From a theoretical point of view, the vanishing step size is convenient because, under various assumptions, it allows to demonstrate the almost sure convergence of the iterates x_n to the set

$$\mathcal{Z} \triangleq \{x \in \mathbb{R}^d \,:\, 0 \in \partial F(x)\} \tag{4.4}$$

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

$$\forall \varepsilon > 0, \quad \limsup_{n\to\infty} \mathbb{P}(\boldsymbol{d}(x_n, \mathcal{Z}) > \varepsilon) \underset{\gamma\downarrow 0}{\longrightarrow} 0, \tag{4.5}$$

where $\boldsymbol{d}$ is the Euclidean distance between x_n and the set $\mathcal{Z}$. Although this result is weaker than in the vanishing step case, constant step stochastic algorithms can reach a neighborhood of $\mathcal{Z}$ faster than their decreasing step analogues, which is an important advantage in the applications where the accuracy of the estimates is not essential. Moreover, in practice they are able to cope with non stationary or slowly changing environments which are frequently encountered in signal processing, and possibly track a changing set of solutions [Benveniste *et al.* 1990, Kushner & Yin 2003].

The second difference between the present chapter and the papers [Majewski *et al.* 2018] and [Davis *et al.* 2020] lies in the algorithm under study. In [Majewski *et al.* 2018, Davis *et al.* 2020], the iterates are supposed to satisfy the inclusion

$$\frac{x_{n+1} - x_n}{\gamma_{n+1}} \in -\partial F(x_n) + \eta_{n+1} \tag{4.6}$$

for all n, where (η_n) is a martingale increment noise w.r.t. the filtration $(\sigma(x_0, \xi_1, \dots, \xi_n))_{n \geqslant 1}$. Under the assumption that $\gamma_n \to 0$ as $n \to \infty$, the authors of [Majewski *et al.* 2018, Davis *et al.* 2020] prove that almost surely, the continuous time linearly interpolated process constructed from a sequence (x_n) satisfying (4.6) is a so-called asymptotic pseudotrajectory [Benaïm *et al.* 2005] of the Differential Inclusion (DI)

$$\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t))\,, \tag{4.7}$$

that will be defined on $\mathbb{R}_+ = [0, \infty)$. Heuristically, this means that a sequence (x_n) satisfying (4.6) shadows a solution to (4.7) as n tends to infinity. This result is one of the key ingredients to establish the almost sure convergence of x_n to the set $\mathcal{Z}$. Unfortunately, a SGD sequence does not satisfy the condition (4.6) in general (setting apart the fact that γ is constant). To be more precise, consider a Clarke-SGD sequence as defined above. For all n, $x_{n+1} = x_n - \gamma \partial f(x_n, \xi_{n+1})$, which in turn implies

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

where (η_n) is a martingale increment noise sequence, and where $\mathbb{E}\partial f(x, \,.\,)$ represents the set-valued expectation $\int \partial f(x, s) d\mu(s)$. The above inclusion is analogous to (4.6) in the case where $\partial F(x) = \mathbb{E}\partial f(x, \cdot)$ for all x *i.e.*, if one can interchange the expectation $\mathbb{E}$ and the Clarke subdifferential operator ∂. Although the interchange holds if *e.g.*, the functions $f(\cdot, s)$ are convex (in which case $\partial f(x, s)$ would coincide with the classical convex subdifferential), one has in general $\partial \mathbb{E} f(x, \cdot) \subset \mathbb{E}\partial f(x, \cdot)$ and the inclusion can be strict [Clarke *et al.* 1998, Proposition 2.2.2]. As a consequence, a Clarke-SGD sequence does not admit the oracle form (4.6) in general. For such a sequence, the corresponding DI reads

$$\dot{\mathsf{x}}(t) \in -\mathbb{E}\partial f(\mathsf{x}(t), \,.\,)\,, \tag{4.8}$$

but unfortunately, the flow of this DI may contain spurious equilibria (an example is provided in this chapter). In [Majewski *et al.* 2018] the authors restrict their

analysis to *regular* functions [Clarke *et al.* 1998, §2.4], for which the interchange of the expectation and the subdifferentiation applies. However, this assumption can be restrictive, since a function as simple as $-|x|$ is not regular at the critical point zero.

A second example where the oracle form Equation (4.6) does not hold is given by autograd sequences. Such an example is studied in [Bolte & Pauwels 2019], assuming that the step size is vanishing and that ξ takes its values over a finite set. It is proved that, the autograd sequence is an almost sure asymptotic pseudotrajectory of the DI $\dot{x}(t) \in -D(x(t))$, for some set-valued map D which is shown to be a *conservative* field with F as a potential. Properties of conservative fields are studied in [Bolte & Pauwels 2019]. In particular, it is proved that $D = \{\nabla f\}$ Lebesgue almost everywhere. Despite this property, the DI $\dot{x}(t) \in -D(x(t))$ substantially differs from (4.7). In particular, the set of equilibria may be strictly larger than the set $\mathcal{Z}$ of critical points of F.

Contributions

- We analyze the SGD algorithm (4.1) in the non-smooth, non-convex setting, under realistic assumptions: the step size is assumed to be constant along the iterations, and we neither assume the regularity of the functions involved, nor the knowledge of an oracle of ∂F as in (4.6). Our assumptions encompass Clarke SGD sequences and autograd sequences as special cases.

- Under mild conditions, we prove that when the initialization x_0 is randomly chosen with a density, all SGD sequences coincide almost surely, irrespective to the particular selection policy used at the points of non-differentiability. In this case, x_n almost never hits a non-differentiable point of $f(\cdot, \xi_{n+1})$ and Equation (4.1) actually holds for all n. Moreover, we prove that

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

  where (η_n) is a martingale difference sequence, and $\nabla F(x_n)$ is the true gradient of F at x_n. This argument allows to bypass the oracle assumption of [Majewski *et al.* 2018, Davis *et al.* 2020].

- We establish that the continuous process obtained by piecewise affine interpolation of (x_n) is a *weak asymptotic pseudotrajectory* of the DI (4.7). In other words, the interpolated process converges in probability to the set of solutions to the DI, as $\gamma \to 0$, for the metric of uniform convergence on compact intervals.

- We establish the long run convergence of the iterates x_n to the set $\mathcal{Z}$ of Clarke critical points of F, in the sense of Equation (4.5). This result holds under two main assumptions. First, it assumed that F admits a chain rule, which is satisfied for instance if F is a so-called tame function. Second, we assume

a standard drift condition on the Markov chain (4.1). Finally, we provide verifiable conditions of the functions $f(\cdot, s)$ under which the drift condition holds.

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

Chapter organization

Section 4.2 recalls some known facts about Clarke subdifferentials, conservative fields and differential inclusions. In Section 4.3, we study the elementary properties of almost-everywhere gradient functions, defined as the functions $\varphi(x, s)$ which coincide with $\nabla f(x, s)$ almost everywhere. Practical examples are provided. In Section 4.4, we study the elementary properties of SGD sequences. Section 4.5 establishes the convergence in probability of the interpolated process to the set of solutions to the DI. In Section 4.6, we establish the long run convergence of the iterates to the set of Clarke critical points. Section 4.7 is devoted to the projected subgradient algorithm. The proofs are found in Section 4.8.

4.2 Preliminaries

4.2.1 Notations

If ν, ν' are two measures on some measurable space $(\Omega, \mathcal{F})$, $\nu \ll \nu'$ means that ν is absolutely continuous w.r.t. ν. The ν-completion of $\mathscr{F}$ is defined as the sigma-algebra consisting of the sets $S \subset \Omega$ such that there exist $A, B \in \mathscr{F}$ with $A \subset S \subset B$ and $\nu(B \backslash A) = 0$. For these sets, $\nu(S) = \nu(A)$.

If E is a metric space, we denote by $\mathscr{B}(E)$ the Borel sigma field on E. Let d be an integer. We denote by $M(\mathbb{R}^d)$ the set of probability measures on $\mathscr{B}(\mathbb{R}^d)$ and by $M_1(\mathbb{R}^d) \stackrel{\Delta}{=} \{\nu \in M(\mathbb{R}^d) : \int \|x\| \nu(dx) < \infty\}$. We denote as λ^d the Lebesgue measure on $\mathbb{R}^d$. When the dimension is clear from the context, we denote as λ this Lebesgue measure. For a subset $\mathcal{K} \subset \mathbb{R}^d$, we denote by

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

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

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

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

on that set and to zero otherwise. The notation A^c represents the complementary set of a set A and $\mathrm{cl}(A)$ its closure.

4.2.2 Clarke Subdifferential and Conservative Fields

In this section we briefly review some recent results of [Bolte & Pauwels 2019]. A set-valued map $D : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is called a *conservative field*, if for each $x \in \mathbb{R}^d$, $D(x)$ is a nonempty and compact subset of $\mathbb{R}^d$, D has a closed graph, and for each absolutely continuous $a\colon [0,1] \to \mathbb{R}^d$, with $a(0) = a(1)$, it holds that:

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

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

$$F(x) = F(0) + \int_0^1 \min_{v \in D(a(t))} \langle \dot{a}(t), v \rangle \, \mathrm{d}t \,. \tag{4.9}$$

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

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

that is to say, F admits a "chain rule" [Bolte & Pauwels 2019, Lemma 2]. Moreover, by [Bolte & Pauwels 2019, Theorem 1], it holds that $D = \{\nabla F\}$ Lebesgue almost everywhere.

We say that a function F is *path differentiable* if there exists a conservative field D such that F is a potential for D. If F is path differentiable, then the Clarke subdifferential ∂F is a conservative field for the potential F [Bolte & Pauwels 2019, Corollary 2]. Another useful example of a conservative field for composite functions is the automatic differentiation field [Bolte & Pauwels 2019, Section 5]. A broad class of functions used in optimization are path differentiable, e.g. any convex, concave, regular or tame.

4.3 Almost-Everywhere Gradient Functions

4.3.1 Definition

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

The following technical lemma, which proof is provided in Section 4.8.1, is essential.

Lemma 4.3.1. *Assume that f is $\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}$-measurable and that $f(\cdot,s)$ is continuous for every $s\in\Xi$. Then $\Delta_f\in\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}$, and the function $\varphi_0:\mathbb{R}^d\times\Xi\to\mathbb{R}^d$ defined as*

$$\varphi_0(x,s)=\begin{cases}\nabla f(x,s) & \text{if } (x,s)\in\Delta_f\\ 0 & \text{otherwise,}\end{cases} \tag{4.10}$$

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

Thanks to this lemma, the following definition makes sense.

Definition 4.3.1. *Assume that $f(\cdot,s)$ is locally Lipschitz continuous for every $s\in\Xi$. A function $\varphi:\mathbb{R}^d\times\Xi\to\mathbb{R}^d$ is called an* almost everywhere (a.e.)-gradient *of f if $\varphi=\nabla f$ $\lambda\otimes\mu$-almost everywhere.*

By Lemma 4.3.1, we observe that a.e.-gradients exist, since $(\lambda\otimes\mu)(\Delta_f^c)=0$. Note that in Definition 4.3.1, we do not assume that φ is $\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}/\mathscr{B}(\mathbb{R}^d)$-measurable. The reason is that this property is not always easy to check on practical examples. However, if one denotes by $\overline{\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}}$ the $\lambda\otimes\mu$ completion of the σ-field $\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}$, an immediate consequence of Lemma 4.3.1 is that any a.e.-gradient of f is a $\overline{\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}}/\mathscr{B}(\mathbb{R}^d)$-measurable function.

4.3.2 Examples

Lazy gradient function. The function φ_0 given by Equation (4.10) is an a.e. gradient function.

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

$$\begin{cases}\varphi(x,s) & =\nabla f(x,s) \text{ if } (x,s)\in\Delta_f,\\ \varphi(x,s) & \in\partial f(x,s) \text{ otherwise.}\end{cases} \tag{4.11}$$

Note that the inclusion $\varphi(x,s)\in\partial f(x,s)$ obviously holds for *all* $(x,s)\in\mathbb{R}^d\times\Xi$, because $\nabla f(x,s)$ is an element of $\partial f(x,s)$ when the former exists. However, conversely, a function $\psi(x,s)\in\partial f(x,s)$ does not necessarily satisfy $\psi(x,s)=\nabla f(x,s)$ if $(x,s)\in\Delta_f$ (see the footnote[1]). By construction, a Clarke gradient function is an a.e. gradient function.

Selections of conservative fields.

Proposition 4.3.2. *Assume that for every $s\in\Xi$, $f(\cdot,s)$ is locally Lipschitz, path differentiable, and is a potential of some conservative field $D_s:\mathbb{R}^d\rightrightarrows\mathbb{R}^d$. Consider a function $\varphi:\mathbb{R}^d\times\Xi\to\mathbb{R}^d$ which is $\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}/\mathscr{B}(\mathbb{R}^d)$ measurable and satisfies $\varphi(x,s)\in D_s(x)$ for all $(x,s)\in\mathbb{R}^d\times\Xi$. Then, φ is an a.e. gradient function for f.*

[1] If a locally Lipschitz function g is differentiable at a point x, we have $\{\nabla g(x)\}\subset\partial g(x)$ but the inclusion could be strict (the two sets are equal if g is regular at x): for example, $g(x)=x^2\sin(1/x)$ is s.t. $\nabla g(0)=0$ and $\partial g(0)=[-1,1]$. There even exist functions for which the set of x s.t. $\{\nabla g(x)\}\subsetneq\partial g(x)$ is a set of full measure (see [Lebourg 1979, Proposition 1.9]).

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

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

where the last equality comes from the fact that for every s, $D_s = \{\nabla f(\cdot,s)\}$ λ-a.e. [Bolte & Pauwels 2019, Theorem 1]. □

We provide below an application of Proposition 4.3.2.

Autograd function. Consider Equation (4.3), which represents a loss of a neural network. Although f is just a composition of some simple functions, a direct calculation of the gradient (if it exists) may be tedious. Automatic differentiation deals with such functions by recursively applying the chain rule to the components of f. More formally consider a function $f : \mathbb{R}^d \to \mathbb{R}$ that can be written as a closed formula of simple functions, mathematically speaking this means that we can represent f by a directed graph. This graph (with $q > d$ vertices) is defined through a set-valued function $\mathbf{parents}(i) \subset \{1,\dots,i-1\}$, a directed edge in this setting will be $j \to i$ with $j \in \mathbf{parents}(i)$. Associate to each vertex a simple function $g_i \colon \mathbb{R}^{|\mathbf{parents}(i)|} \to \mathbb{R}$, given an input $x = (x_1,\dots,x_d) \in \mathbb{R}^d$ we recursively define $x_i = g_i((x_j)_{j\in\mathbf{parents}(i)})$ for $i > d$ and finally $f(x) = x_q$. For instance, if f is a cross entropy loss of a neural network, with activation functions being ReLu or sigmoid functions, then g_i are some compositions of simple functions $\mathbf{log}$, $\mathbf{exp}$, $\frac{\mathbf{1}}{\mathbf{1+x^2}}$, norms and piecewise polynomial functions, all being path differentiable [Bolte & Pauwels 2019, section 6], [Davis *et al.* 2020, Section 5.2]. Automatic differentiation libraries calculate the gradient of f by successively applying the chain rule (in the sense $(g_1 \circ g_2)' = (g_1' \circ g_2)g_2'$) to the simple functions g_i. While the chain rule is no longer valid in a nonsmooth setting (see e.g. [Kakade & Lee 2018]), it is shown in [Bolte & Pauwels 2019, Section 5] that when the simple functions are path-differentiable, the output of automatic differentiation (e.g. `autograd` in PyTorch ([Paszke *et al.* 2017])) is a selection of some conservative field D for f. We refer to [Bolte & Pauwels 2019] for a more detailed account. We denote by $a_f(x)$ the output of automatic differentiation of a function f at some point x.

Assume that $\Xi = \mathbb{N}$ and for each $s \in \Xi$, $f(\cdot,s)$ is defined through a recursive graph of path differentiable functions (in the machine learning paradigm $f(\cdot,s)$ will represent the loss related to one data point, while $F(\cdot)$ is the average loss). By Proposition 4.3.2, the map $(x,s) \mapsto a_{f(\cdot,s)}(x)$ is an a.e. gradient function for f.

4.4 SGD Sequences

4.4.1 Definition

Given a probability measure ν on $\mathscr{B}(\mathbb{R}^d)$, define the probability space $(\Omega, \mathscr{F}, \mathbb{P}^\nu)$ as $\Omega = \mathbb{R}^d \times \Xi^{\mathbb{N}}$, $\mathscr{F} = \mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}^{\otimes\mathbb{N}}$, and $\mathbb{P}^\nu = \nu \otimes \mu^{\otimes\mathbb{N}}$. We denote by $(x_0, (\xi_n)_{n\in\mathbb{N}^*})$ the canonical process on $\Omega \to \mathbb{R}^d$ *i.e.*, writing an elementary event in the space Ω as $\omega = (\omega_n)_{n\in\mathbb{N}}$, we set $x_0(\omega) = \omega_0$ and $\xi_n(\omega) = \omega_n$ for each $n \geqslant 1$. Under $\mathbb{P}^\nu$, x_0

is a $\mathbb{R}^d$-valued random variable with the probability distribution ν, and the process $(\xi_n)_{n\in\mathbb{N}*}$ is an independent and identically distributed (i.i.d.) process such that the distribution of ξ_1 is μ, and x_0 and (ξ_n) are independent. We denote by $\overline{\mathscr{F}}$ the $\lambda\otimes\mu^{\otimes\mathbb{N}}$-completion of $\mathscr{F}$.

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

Definition 4.4.1. *Assume that $f(\cdot,s)$ is locally Lipschitz continuous for every $s\in\Xi$. A sequence $(x_n)_{n\in\mathbb{N}*}$ of functions on $\Omega\to\mathbb{R}^d$ is called an SGD sequence for f with the step $\gamma>0$ if there exists an a.e.-gradient φ of f such that*

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

4.4.2 All SGD Sequences Are Almost Surely Equal

Consider the SGD sequence

$$x_{n+1}=x_n-\gamma\varphi_0(x_n,\xi_{n+1}),\tag{4.12}$$

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

$$P_\gamma g(x)=\int g(x-\gamma\varphi_0(x,s))\,\mu(ds).\tag{4.13}$$

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

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

Proposition 4.4.1. *Consider $\gamma\in\Gamma$ and $\nu\in M_{abs}(\mathbb{R}^d)$. Then, each SGD sequence (x_n) with the step γ is $\overline{\mathscr{F}}/\mathscr{B}(\mathbb{R}^d)^{\otimes\mathbb{N}}$-measurable. Moreover, for any two SGD sequences (x_n) and (x'_n) with the step γ, it holds that $\mathbb{P}^\nu\left[(x_n)\neq(x'_n)\right]=0$. Finally, the probability distribution of x_n under $\mathbb{P}^\nu$ is Lebesgue-absolutely continuous for each $n\in\mathbb{N}$.*

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

Proof. Let (x_n) be the lazy SGD sequence given by (4.12). Given an a.e. gradient φ, define the SGD sequence (z_n) as $z_0=x_0$, $z_{n+1}=z_n-\gamma\varphi(z_n,\xi_{n+1})$ for $n\geqslant 0$. The sequence (x_n) is $\mathscr{F}/\mathscr{B}(\mathbb{R}^d)^{\otimes\mathbb{N}}$-measurable thanks to Lemma 4.3.1. Moreover, applying recursively the property that $\rho P_\gamma\ll\lambda$ when $\rho\ll\lambda$, we obtain that the distribution of x_n is absolutely continuous for each $n\in\mathbb{N}$.

To establish the proposition, it suffices to show that z_n is $\overline{\mathscr{F}}/\mathscr{B}(\mathbb{R}^d)$-measurable for each $n\in\mathbb{N}$, and that $\mathbb{P}^\nu[z_n\neq x_n]=0$, which results in particular in the absolute continuity of the distribution of z_n. We shall prove these two properties by

induction on n. They are trivial for $n = 0$. Assume they are true for n. Recall that $z_{n+1} = z_n - \gamma \nabla f(z_n, \xi_{n+1})$ if $(z_n, \xi_{n+1}) \in A$, where $A \in \overline{\mathscr{B}(\mathbb{R}^d) \otimes \mathscr{T}}$ is such that $(\lambda \otimes \mu)(A^c) = 0$, and $x_{n+1} = x_n - \gamma \nabla f(x_n, \xi_{n+1}) \mathbb{1}_{(x_n, \xi_{n+1}) \in \Delta_f}$. The set $B = \{\omega \in \Omega : z_{n+1} \neq x_{n+1}\}$ satisfies $B \subset B_1 \cup B_2$, where

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

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

Proposition 4.4.1 means that the SGD sequence does not depend on the specific a.e. gradient used by the practioner, provided that the law of x_0 has a density and $\gamma \in \Gamma$. Let us make this last assumption clearer. Consider for instance $d = 1$ and suppose that $f(x, s) = 0.5x^2$ for all s. If $\gamma = 1$, the SGD sequence $x_{n+1} = x_n - \gamma x_n$ satisfies $x_1 = 0$ for any initial point and thus, does not admit a density, whereas for any other value of γ, x_n has a density for all n, provided that x_0 has a density. Otherwise stated, $\Gamma = \mathbb{R}_+ \backslash \{1\}$ in this example.

It is desirable to ensure that Γ contains almost all the points of $\mathbb{R}_+$. The next proposition shows that this will be the case under mild conditions. The proof is given in 4.8.2.

Proposition 4.4.2. *Assume that for μ–almost every $s \in \Xi$, the function $f(\cdot, s)$ satisfies the property that at λ–almost every point of $\mathbb{R}^d$, there is a neighborhood of this point on which it is C^2. Then, Γ^c is Lebesgue negligible.*

This assumption holds true as soon as for μ-almost all s, $f(\cdot, s)$ is tame, since in this case $\mathbb{R}^d$ can be partitioned in manifolds on each of which $f(\cdot, s)$ is C^2 ([Bolte *et al.* 2007]), and therefore $f(\cdot, s)$ is C^2 (in the classical sense) on the union of manifolds of full dimension, and therefore almost everywhere.

4.4.3 SGD as a Robbins-Monro Algorithm

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

Assumption 4.4.1. *i) There exists a measurable function $\kappa : \mathbb{R}^d \times \Xi \to \mathbb{R}_+$ s.t. for each $x \in \mathbb{R}^d$, $\int \kappa(x, s)\, \mu(ds) < \infty$ and there exists $\varepsilon > 0$ for which*

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

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

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

$$F : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto \int f(x, s)\, \mu(ds) \tag{4.14}$$

is locally Lipschitz on $\mathbb{R}^d$. We denote by $\mathcal{Z}$ the set of (Clarke) critical points of F, as defined in Equation (4.4).

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

Theorem 4.4.3. *Let Assumption 4.4.1 holds true. Consider $\gamma \in \Gamma$ and $\nu \in M_{abs}(\mathbb{R}^d) \cap M_1(\mathbb{R}^d)$. Let $(x_n)_{n\in\mathbb{N}^*}$ be a SGD sequence for f with the step γ. Then, for every $n \in \mathbb{N}$, it holds $\mathbb{P}^\nu$-a.e. that*

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

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

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

Theorem 4.4.3 is important because it shows that $\mathbb{P}^\nu$-a.e., the SGD sequence (x_n) verifies

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

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

4.5 Dynamical Behavior

4.5.1 Assumptions and Result

In this section we prove that the SGD sequence $(x_n)_{n\in\mathbb{N}^*}$ (which is by Theorem 4.4.3, under the stated assumptions, unique) closely follows a trajectory of a solution to the DI (4.7) as the step size γ tends to zero. To state the main result of this section, we need to strengthen Assumption 4.4.1.

Assumption 4.5.1. *The function κ of Assumption 4.4.1 satisfies:*

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

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

The first point guarantees the existence of global solutions to (4.7) starting from any initial point (see Section 2.2.2).

Assumption 4.5.2. *The closure of Γ contains 0.*

By Proposition 4.4.2, Assumption 4.5.2 is mild. It holds for instance if every $f(\cdot, s)$ is a tame function.

We recall that $\mathsf{S}_{-\partial F}(A)$ is the set of solutions to (4.7) that start from any point in the set $A \subset \mathbb{R}^d$.

Theorem 4.5.1. *Let Assumptions 4.4.1–4.5.2 hold true. Let* $\{(x_n^\gamma)_{n\in\mathbb{N}^*} : \gamma \in (0,\gamma_0]\}$ *be a collection of SGD sequences of steps* $\gamma \in (0,\gamma_0]$*. Denote by* x^γ *the piecewise affine interpolated process*

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

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

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

where the distance $\boldsymbol{d}_C$ *is defined in (2.2). Moreover, the family of distributions* $\{\mathbb{P}^\nu(\mathsf{x}^\gamma)^{-1} : \nu \in M_{abs}(\mathcal{K}), 0<\gamma<\gamma_0, \gamma\in\Gamma\}$ *is tight.*

The proof is given in Section 4.8.4.

Theorem 4.5.1 implies that the interpolated process x^γ converges in probability as $\gamma \to 0$ to the set of solutions to (4.7). Moreover, the convergence is uniform w.r.t. to the choice of the initial distribution ν in the set of absolutely continuous measures supported by a given compact set.

4.5.2 Importance of the Randomization of x_0

In this paragraph, we discuss the case where x_0 is no longer random, but set to an arbitrary point in $\mathbb{R}^d$. In this case, there is no longer any guarantee that the iterates x_n only hit the points where a gradient exist. We focus on the case where (x_n) is a Clarke-SGD sequence of the form (4.11), where the function φ is assumed $\mathscr{B}(\mathbb{R}^d)\otimes\mathscr{T}/\mathscr{B}(\mathbb{R}^d)$ measurable for simplicity. By Assumption 4.4.1, it is not difficult to see that $\varphi(x,\cdot)$ is μ-integrable for all $x \in \mathbb{R}^d$ and, denoting by $\mathbb{E}(\varphi(x,\cdot))$ the corresponding integral w.r.t. μ, we can rewrite the iterates under the form:

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

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

Example 4.5.1. *Take a finite probability space* $\Xi = \{1,2\}$ *and* $\mu(\{1\}) = \mu(\{2\}) = 1/2$*. Let* $f(x,1) = 2x\mathbb{1}_{x\leqslant 0}$ *and* $f(x,2) = 2x\mathbb{1}_{x\geqslant 0}$*. We have* $F(x) = x$*, and therefore* $\partial F(0) = \{1\}$*, whereas* $\partial f(0,1) = \partial f(0,2) = [0,2]$ *and therefore* $\int \partial f(0,s)\mu(\mathrm{d}s) = [0,1]$*. We see that* $0 \in \mathbb{E}\partial f(0,\cdot)$ *while* $0 \notin \partial F(0)$*. Furthermore, the trajectory defined on* $\mathbb{R}_+$ *as*

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

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

Example 4.5.2. *Consider the same setting as in the previous example. Consider a stochastic gradient algorithm of the form (4.1), initialized at* $x_0 = 0$ *with* φ *such that* $\varphi(0,1) = \varphi(0,2) = 0$*. Then, the iterates* x_n^γ *are identically zero. This shows that the stochastic gradient descent may converge to a non critical point of* F*. Theorem 4.5.1 may fail unless a random initial point is chosen.*

4.6 Long Run Convergence

4.6.1 Assumptions and Result

As discussed in the introduction, the SGD sequence (x_n) is not expected to converge in probability to $\mathcal{Z}$ when the step is constant. Instead, we shall establish the convergence (4.5). The "long run" convergence referred to here is understood in this sense.

In all this section, we shall focus on the lazy SGD sequences described by Equation (4.12). This incurs no loss of generality, since any two SGD sequences are equal $\mathbb{P}^\nu$-a.e. by Proposition 4.4.1 as long as $\nu \ll \lambda$. Our starting point is to see the process (x_n) and as a Markov process which kernel P_γ is defined by Equation (4.13). Our first task is to establish the ergodicity of this Markov process under the convenient assumptions. Namely, we show that P_γ has a unique invariant probability measure π_γ, *i.e.*, $\pi_\gamma P_\gamma = \pi_\gamma$, and that $\|P_\gamma^n(x,\cdot) - \pi_\gamma\|_{\mathrm{TV}} \to 0$ as $n \to \infty$ for each $x \in \mathbb{R}^d$, where $\|\cdot\|_{\mathrm{TV}}$ is the total variation norm. Further, we need to show that the family of invariant distributions $\{\pi_\gamma\}_{\gamma\in(0,\gamma_0]}$ for a certain $\gamma_0 > 0$ is tight. The long run behavior referred to above is then intimately connected with the properties of the accumulation points of this family as $\gamma \to 0$. To study these properties, we get back to the DI $\dot{\mathsf{x}} \in -\partial F(\mathsf{x})$ (we recall that a concise account of the notions relative to this dynamical system and needed in this chapter is provided in Section 2.2.2). The crucial point here is to show, with the help of Theorem 4.5.1, that the accumulation points of $\{\pi_\gamma\}$ as $\gamma \to 0$ are invariant measures for the set-valued flow induced by the DI. In its original form, this idea dates back to the work of Has'minskiĭ [Has'minskiĭ 1963]. We observe here that while the notion of invariant measure for a single-valued semiflow induced by, say, an ordinary differential equation, is classical, it is probably less known in the case of a set-valued differential inclusion. We borrow it from the work of Roth and Sandholm [Roth & Sandholm 2013].

Having shown that the accumulation points of $\{\pi_\gamma\}$ are invariant for the DI $\dot{\mathsf{x}} \in -\partial F(\mathsf{x})$, the final step of the proof is to make use of Poincaré's recurrence theorem, that asserts that the invariant measures of a semiflow are supported by the so-called Birkhoff center of this semiflow (again, a set-valued version of Poincaré's recurrence theorem is provided in [Aubin *et al.* 1991, Faure & Roth 2013]). To establish the convergence (4.5), it remains to show that the Birkhoff center of the DI $\dot{\mathsf{x}} \in -\partial F(\mathsf{x})$ coincides with $\operatorname{zer}\partial F$. The natural assumption that ensures the identity of these two sets will be that F admits a chain rule [Clarke *et al.* 1998, Bolte *et al.* 2007, Davis *et al.* 2020].

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

Assumption 4.6.1. *There exist measurable functions $V : \mathbb{R}^d \to [0, +\infty)$, $p : \mathbb{R}^d \to [0, +\infty)$, $\alpha : (0, +\infty) \to (0, +\infty)$ and a constant $C \geqslant 0$ s.t. the following holds for every $\gamma \in \Gamma \cap (0, \gamma_0]$.*

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

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

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

$$P_\gamma V(x) \leqslant V(x) - \alpha(\gamma)p(x) + C\alpha(\gamma)\mathbb{1}_{\|x\| \leqslant R}. \tag{4.15}$$

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

Assumptions of this type are frequently encountered in the field of Markov chains. Assumption 4.6.1–(i)) states that $\mathrm{cl}(B(0,R))$ is a so-called small set for the kernel P_γ, and Assumption 4.6.1–(ii)) is a standard drift assumption. Taken together, they ensure that the kernel P_γ is a so-called Harris-recurrent kernel, that it admits a unique invariant probability distribution π_γ, and finally, that this kernel is ergodic in the sense that $\|P_\gamma(x,\cdot) - \pi_\gamma\|_{\mathrm{TV}} \to 0$ as $n \to \infty$ (see [Meyn & Tweedie 2009]). The introduction of the factors $\alpha(\gamma)$ and $C\alpha(\gamma)$ in Equation (4.15) guarantees moreover the tightness of the family $\{\pi_\gamma\}_{\gamma \in (0,\gamma_0]}$.

In Section 4.6.2, we provide sufficient and verifiable conditions ensuring the validity of Assumption 4.6.1 for P_γ.

As announced above, we also need:

Assumption 4.6.2. *The function F defined by (4.14) admits a chain rule, namely, for any absolutely continuous curve $z : \mathbb{R}_+ \to \mathbb{R}^d$, for almost all $t > 0$, $\forall v \in \partial F(z(t))$, $\langle v, \dot{z}(t) \rangle = (F \circ z)'(t)$.*

Assumption 4.6.2 is satisfied as soon as F is path-differentiable, for instance when F is either convex, regular, Whitney stratifiable or tame (see [Bolte & Pauwels 2019, Proposition 1]and [Bolte *et al.* 2007, Davis *et al.* 2020]).

Theorem 4.6.1. *Let Assumptions 4.4.1-4.5.2 and 4.6.1-4.6.2 hold true. Let $\{(x_n^\gamma)_{n \in \mathbb{N}^*} : \gamma \in (0, \gamma_0]\}$ be a collection of SGD sequences of step-size γ. Then, the set $\mathcal{Z} = \{x : 0 \in \partial F(x)\}$ is nonempty and for all $\nu \in M_{abs}(\mathbb{R}^d)$ and all $\varepsilon > 0$,*

$$\limsup_{n \to \infty} \mathbb{P}^\nu \left(\boldsymbol{d}(x_n^\gamma, \mathcal{Z}) > \varepsilon \right) \xrightarrow[\substack{\gamma \to 0 \\ \gamma \in \Gamma}]{} 0. \tag{4.16}$$

4.6.2 The Validity of Assumption 4.6.1

In this paragraph, we provide sufficient conditions under which Assumption 4.6.1 hold true. A simple way to ensure the truth of Assumption 4.6.1-(i)) is to add a small random perturbation to the function $\varphi_0(x, s)$. Formally, we modify algorithms described by Equation (4.12) and (4.18), and write

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

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

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

Proposition 4.6.2. *Let Assumption 4.5.1 hold true. Then, for each $R > 0$, there exists $\varepsilon > 0$ such that*

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

Thus, Assumption 4.6.1-(i)) is satisfied for $\widetilde{P}$.

We now turn to the assumptions 4.6.1-(ii)) and 4.6.1-(iii)).

Proposition 4.6.3. *Assume that there exists $R \geqslant 0$, $C > 0$, and a measurable function $\beta : \Xi \to \mathbb{R}_+$ such that the following conditions hold:*

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

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

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

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

Then, it holds that

$$P_\gamma F(x) \leqslant F(x) - \gamma(1 - \gamma K)\mathbb{1}_{\|x\|>2R}\|\nabla F(x)\|^2 + \gamma^2 K \mathbb{1}_{\|x\|>2R} + \gamma K \mathbb{1}_{\|x\|\leqslant 2R} \quad (4.17)$$

for some constant $K > 0$. In particular, Assumptions 4.6.1-(ii)) and 4.6.1-(iii)) hold true.

We finally observe that this proposition can be easily adapted to the case where the kernel P_γ is replaced with the kernel $\widetilde{P}$ of Proposition 4.6.2.

4.7 The Projected Subgradient Algorithm

In many practical settings, the conditions of Proposition 4.6.3 that ensure the truth of Assumptions 4.6.1–(ii)) and 4.6.1–(iii)) are not satisfied. This is for instance the case when the function f is described by Equation (4.3) with the mappings σ_ℓ at the right hand side of this equation being all equal to the ReLU function. In such situations, it is often pertinent to replace the SGD sequence with a *projected* version of the algorithm. Given an a.e.-gradient φ of the function f and a non empty compact and convex set $\mathcal{K} \subset \mathbb{R}^d$, a *projected SGD sequence* $(x_n^{\gamma,\mathcal{K}})$ is given by the recursion

$$x_0^{\gamma,\mathcal{K}} = x_0, \quad x_{n+1}^{\gamma,\mathcal{K}} = \Pi_{\mathcal{K}}(x_n^{\gamma,\mathcal{K}} - \gamma\varphi(x_n^{\gamma,\mathcal{K}}, \xi_{n+1}))\,, \tag{4.18}$$

where $\Pi_{\mathcal{K}}$ stands for a Euclidean projection onto $\mathcal{K}$. Our purpose is to generalize Theorem 4.5.1 to this situation. This generalization is not immediate for several reasons. First, the projection step is likely to introduce spurious local minima. As far as the iterates (4.18) are concerned, the role of differential inclusion (4.7) is now played by the differential inclusion:

$$\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t)) - \mathcal{N}_{\mathcal{K}}(\mathsf{x}(t))\,, \tag{4.19}$$

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

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

which we shall refer to as the set of Karush-Kuhn-Tucker points. A second theoretical difficulty is related to the fact that Proposition 4.4.1 does no longer hold. Indeed, it can happen x_0 has a density, but the next iterates $x_n^{\gamma,\mathcal{K}}$ don't. The reason is that $x_n^{\gamma,\mathcal{K}}$ generally has a non zero probability to be in the (Lebesgue negligible) border of $\mathcal{K}$, that is, $\mathrm{cl}(\mathcal{K})\backslash\,\mathrm{int}(\mathcal{K})$, where $\mathrm{cl}(\mathcal{K})$ and $\mathrm{int}(\mathcal{K})$ respectively stand for the closure and the interior of $\mathcal{K}$.

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

We make the following assumption.

Assumption 4.7.1. *For every $x \in \mathbb{R}^d$, the law of $\varphi_0(x,\xi)$, where $\xi \sim \mu$, is absolutely continuous relatively to Lebesgue.*

Assumption 4.7.1 is much stronger than Assumption 4.5.2. Indeed, it implies that the distribution of $x_n^{\gamma,r} - \gamma\varphi(x_n^{\gamma,r}, \xi_{n+1})$ is always Lebesgue-absolutely continuous. It is useful to note though that Assumption 4.7.1 holds upon adding at each step a small random perturbation to φ_0 as in Section 4.6.2 above.

In order to state our first result in this framework, we need to introduce some new notations. We let $\mathbb{S}(r) := \{x : \|x\| = r, x \in \mathbb{R}^d\}$ be the sphere of radius r. By

[Folland 2013, Theorem 2.49], there is a unique measure[2] ϱ_1 on $\mathbb{S}(1)$ such that for any positive function $f : \mathbb{R}^d \to \mathbb{R}$, we have:

$$\int f(x)\lambda^d(\mathrm{d}x) = \int_0^\infty \int_{\mathbb{S}(1)} f(r\theta) r^{d-1} \varrho_1(\mathrm{d}\theta)\lambda^1(\mathrm{d}r)\,. \tag{4.20}$$

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

The next proposition, which is proven in the same way as Proposition 4.4.1, shows that for almost every $r > 0$, all projected SGD sequences are almost surely equal.

Proposition 4.7.1. *Let Assumption 4.7.1 hold true. Then, for almost every $r > 0$, $\forall \nu \in M^r$, each projected SGD sequence $(x_n^{\gamma,r})$ is $\overline{\mathscr{F}}/\mathscr{B}(\mathbb{R}^d)^{\otimes\mathbb{N}}$-measurable. Moreover, for any two projected SGD sequences $(x_n^{\gamma,r})$ and $(y_n^{\gamma,r})$, it holds that $\mathbb{P}^\nu\left[(x_n^{\gamma,r}) \neq (y_n^{\gamma,r})\right] = 0$. Finally, under $\mathbb{P}^\nu$, for every $n \in \mathbb{N}$, the probability distribution of $x_n^{\gamma,r}$ is in M^r.*

By Proposition 4.7.1 we can focus on the lazy projected SGD sequence:

$$x_{n+1}^{\gamma,r} = \Pi_r(x_n^{\gamma,r} - \gamma\varphi_0(x_n^{\gamma,r}, \xi_{n+1}))\,. \tag{4.21}$$

We define its associated kernel

$$P_\gamma^r g(x) = \int g(\Pi_r(x - \gamma\varphi_0(x, s)))\mu(\mathrm{d}s)\,. \tag{4.22}$$

The next two theorems are analogous to Theorems 4.4.3 and 4.5.1.

Theorem 4.7.2. *Let Assumptions 4.4.1 and 4.7.1 hold. Then for almost every $r > 0$, $\forall \nu \in M^r$, for every $n \in \mathbb{N}$ it holds $\mathbb{P}^\nu$-a.e.*

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

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

Theorem 4.7.3. *Let Assumptions 4.4.1–4.5.1 and 4.7.1 hold true. Denote $\mathsf{x}^{\gamma,r}$ the piecewise affine interpolated process:*

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

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

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

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

[2] As it is clear from Equation (4.20) we can see (λ^1, ϱ_1) as a polar coordinates representation of the Lebesgue measure λ^d.

We compare Theorems 4.4.3 and 4.5.1. First, because of the projection step (and with the help of Assumption 4.7.1), the law of the n-th iterate is no longer in M_{abs}, but in M^r. Second, the continuous counterpart of Equation (4.18) is now the differential inclusion (4.19) Note that, if the solutions of the DI (4.7) that start from $\mathcal{K}$ all lie in $\mathrm{cl}(B(0,r))$, then the set of these solutions coincides with the set of solutions of the DI (4.19) that start from $\mathcal{K}$.

The analysis of the convergence of the iterates in the "long run" is greatly simplified by the introduction of the projection step. Compared to Assumption 4.6.1, we only assume the existence of a small set for P_γ^r, the drift condition of the form 4.6.1-(ii))–(iii)) is then automatically satisfied, thanks to the projection step (see Section 4.8.5).

Assumption 4.7.2. *There is $R > 0$ and $\gamma_0 > 0$ such that for every $\gamma \in (0, \gamma_0]$ there is ρ_γ such that Assumption 4.6.1-(i)) hold for (R, ρ_γ) (note that R is independent of γ here).*

As shown in Section 4.6.2, Assumption 4.7.2 holds upon adding to φ_0 a small random perturbation.

Theorem 4.7.4. *Let Assumptions 4.4.1-4.5.1 and 4.6.2–4.7.2 hold. Let $\{(x_n^{\gamma,r})_{n\in\mathbb{N}^*} : \gamma \in (0,\gamma_0]\}$ be a collection of projected SGD sequences of step-size γ. Then, for almost every $0 < r \leqslant R$, the set $\mathcal{Z}_r = \{x \colon 0 \in \partial F(x) + \mathcal{N}_r(x)\}$ is nonempty and for all $\nu \in M^r$ and all $\varepsilon > 0$,*

$$\limsup_{n\to\infty} \mathbb{P}^\nu \left(\boldsymbol{d}(x_n^{\gamma,r}, \mathcal{Z}_r) > \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0. \tag{4.23}$$

Theorem 4.7.4 is analogous to Theorem 4.6.1. Notice that, since $M_{abs} \subset M^r$, x_0 can still be initialized under a Lebesgue-absolutely continuous measure. On the other hand, as explained in the beginning of this section, due to the projection step, the iterates, instead of converging to $\mathcal{Z}$, are now converging to the set of Karush-Kuhn-Tucker points related to the DI (4.19).

4.8 Proofs

4.8.1 Proof of Lemma 4.3.1

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

$$\bigcap_{\varepsilon\in\mathbb{Q}} \bigcup_{\delta\in\mathbb{Q}} \bigcap_{0<\|h\|\leqslant\delta} \left\{(y,s) : \left|\frac{f(y+h,s) - f(y,s) - \langle d_x, h\rangle}{\|h\|}\right| < \varepsilon\right\}. \tag{4.24}$$

In addition, using that $f(\cdot,s)$ is continuous, the above set is unchanged if the inner intersection over $0 < \|h\| \leqslant \delta$ is replaced by an intersection over the h s.t. $0 <$

$\|h\| \leqslant \delta$ and having *rational* coordinates *i.e.*, $h \in \mathbb{Q}^d$. Define:

$$\Delta'_f := \bigcap_{\varepsilon' \in \mathbb{Q}} \bigcup_{d \in \mathbb{Q}^d} \bigcap_{\varepsilon \in \mathbb{Q}} \bigcup_{\delta \in \mathbb{Q}} \bigcap_{\substack{0 < \|h\| \leqslant \delta \\ h \in \mathbb{Q}^d}} \left\{ (x,s) : \left| \frac{f(x+h,s) - f(x,s) - \langle d, h \rangle}{\|h\|} \right| < \varepsilon + \varepsilon' \right\} \tag{4.25}$$

By construction, Δ'_f is a measurable set. We prove that $\Delta'_f = \Delta_f$. Consider $(x,s) \in \Delta_f$ and let d_x be the gradient of $f(\cdot, s)$ at x. By (4.24) for all $\varepsilon \in \mathbb{Q}$, there is a $\delta \in \mathbb{Q}$ such that:

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

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

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

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

$$\left| \frac{f(x+h,s) - f(x,s) - \langle d_k, h \rangle}{h} \right| < \varepsilon + \varepsilon'_k \,. \tag{4.26}$$

Moreover, one may choose $\delta_k(\varepsilon) \leqslant \delta_0(\varepsilon)$. Inspecting first the inequality (4.26) for $k = 0$, we easily obtain that the quantity $\frac{f(x+h,s)-f(x,s)}{h}$ is bounded uniformly in h s.t. $0 < \|h\| \leqslant \delta_0(\varepsilon)$. Using this observation and again Equation (4.26), this in turn implies that (d_k) is a bounded sequence. There exists $d \in \mathbb{R}^d$ and s.t. $d_k \to d$ along some extracted subsequence. Now consider $\varepsilon > 0$ and choose k such that $\|d_k - d\| < \frac{\varepsilon}{2}$ and $\varepsilon'_k < \frac{\varepsilon}{2}$. For all $h \leqslant \delta_k(\varepsilon/2)$,

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

This means that d is the gradient of $f(\cdot, s)$ at x, hence $\Delta'_f \subset \Delta_f$. Hence, the first point of the Lemma 4.3.1 is proved.

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

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

and the measurability of φ_0 follows from the measurability of f and the measurability of $\mathbb{1}_{\Delta_f}$.

Finally, assume that $f(\cdot, s)$ is locally Lipschitz continuous for every $s \in \Xi$. From Rademacher's theorem [Clarke *et al.* 1998, Ch. 3], $f(\cdot, s)$ is almost everywhere differentiable, which reads $\int (1 - \mathbb{1}_{\Delta_f}(x,s)) \lambda(dx) = 0$. Using Fubini's theorem, $\int_{\mathbb{R}^d \times \Xi} (1 - \mathbb{1}_{\Delta_f}(x,s)) \; \lambda(dx) \otimes \mu(ds) = 0$, and the last point is proved.

4.8.2 Proof of Proposition 4.4.2

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

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

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

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

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

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

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

and by Fubini's theorem,

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

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

a local diffeomorphism is an open set, hence

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

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

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

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

4.8.3 Proof of Theorem 4.4.3

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

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

When Assumption 4.4.1 holds, Rademacher's theorem, lemma 4.3.1 and Fubini's theorem imply that $S_1 \in \mathscr{B}(\mathbb{R}^d)$ and $\lambda(\mathbb{R}^d \backslash S_1) = 0$. Hence, for μ-a.e. s we have $f(\cdot, s)$ differentiable at x_0, and since $\xi_1 \sim \mu$, $f(\cdot, \xi_1)$ is differentiable at x_0. Now by Rademacher's theorem again, the set $S_2 \subset \mathbb{R}^d$ where F is differentiable satisfies $\lambda(\mathbb{R}^d \backslash S_2) = 0$, therefore F is differentiable at x_0. Moreover, with probability one x_0 is in $S_1 \cap S_2$. Define $A(x) \triangleq \{s \in \Xi : (x, s) \notin \Delta_f\}$. By Assumption 4.4.1, $\|\nabla f(x, \cdot)\|$ is μ-integrable. Moreover, for all $x \in S_1 \cap S_2$ and all $v \in \mathbb{R}^d$

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

where the interchange between the limit and the integral follows from Assumption 4.4.1 and the dominated convergence theorem. Hence, $\nabla F(x) = \int \nabla f(x, s) \mathbb{1}_{\Delta_f}(x, s)\, \mu(ds)$ for all $x \in S_1 \cap S_2$. Now denote by ν_n the law of x_n. Since we assumed that $\nu_0 \ll \lambda$, it holds that $\mathbb{P}^\nu(x_0 \in S_1 \cap S_2) = 1$. Therefore, with probability one,

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

Thus, x_1 is integrable whenever x_0 is integrable, and $\mathbb{E}_0(x_1) = x_0 - \gamma \nabla F(x_0)$. Since by Assumption $\nu_1 \ll \lambda$ we can iterate our argument for x_2 and then for all x_n and the conclusions of Theorem 4.4.3 follow.

4.8.4 Proof of Theorem 4.5.1

We want to apply [Bianchi *et al.* 2019, Theorem 5.1.], and therefore verify its assumptions [Bianchi *et al.* 2019, Assumption RM]. In order to fall in its setting we first need to rewrite our kernel in a more appropriate way. As ∂F takes nonempty compact values, it admits a measurable selection $\varphi(x) \in \partial F(x)$ [Aliprantis & Border 2006, Lemma 18.2 and Corollary 18.15]. Take $\gamma \in \Gamma$, a SGD sequence (x_n^γ) and notice that by Theorem 4.4.3 it is $\mathbb{P}^\nu$ almost surely always in $\mathcal{D}_F \cap S_1$, where S_1 is the set of x where $\nabla f(x,s)$ exists for μ-a.e. s. Therefore its Markov kernel can be equivalently defined as:

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

Now we can apply [Bianchi *et al.* 2019, Theorem 5.1.] with $h_\gamma(s,x) = -(\mathbb{1}_{\mathcal{D}_F \cap S_1}(x)\nabla F(x) + \mathbb{1}_{(\mathcal{D}_F \cap S_1)^c}(x)\varphi(x))$ (note that it is independent of s) and we have $h(x,s) \in H(x,s) = H(x) \triangleq -\partial F(x)$. As we show next, [Bianchi *et al.* 2019, Assumption RM] now easily follows.
First, it is immediate from the general properties of the Clarke subdifferential that the set-valued map $-\partial F$ is proper and uppersemicontinuous with convex and compact values, hence the assumption (iii) of [Bianchi *et al.* 2019, Assumption RM]. Assumption (ii) is immediate by the uppersemicontinuity of $-\partial F$. Moreover, we obtain from Assumption 4.5.1 that there exists a constant $K \geqslant 0$ such that

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

Thus, $\mathsf{S}_{-\partial F}$ is defined on the whole $\mathbb{R}^d$, and $\mathsf{S}_{-\partial F}$ is closed in $(C(\mathbb{R}_+, \mathbb{R}^d), \boldsymbol{d})$ (see [Aubin & Cellina 1984]), hence assumption (v). Finally, assumption (vi) comes from Assumption 4.5.1.

We remark that although, [Bianchi *et al.* 2019, Theorem 5.1] deals with a family of measures $(\mathbb{P}^a)_{a \in \mathcal{K}}$, the proofs remain unchanged when we consider $(\mathbb{P}^\nu)_{\nu \in M_{abs}(\mathcal{K})}$.

4.8.5 Proof of Theorems 4.6.1 and 4.7.4

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

- Lemma 4.8.2: Q_γ has a unique invariant probability distribution π_γ, with $\pi_\gamma \in M_{abs}$ if $Q_\gamma = P_\gamma$ and $\pi_\gamma \in M^r$ otherwise, moreover Q_γ is ergodic in the sense of the Total Variation norm.

- Lemma 4.8.3: The family $\{\pi_\gamma\}_{\gamma \in (0,\gamma_0]}$ is tight.

- Proposition 4.8.4: The accumulation points of $\{\pi_\gamma\}_{\gamma \in (0,\gamma_0]}$ as $\gamma \to 0$ are invariant for the DI $\dot{\mathsf{x}} \in \mathsf{H}(\mathsf{x})$.

Before stating Lemma 4.8.2, we recall a general result on Markov processes. Let $Q : \mathbb{R}^d \times \mathscr{B}(\mathbb{R}^d) \to [0,1]$ be a Markov kernel on $\mathbb{R}^d$. A set $B \subset \mathbb{R}^d$ is said to be a small-set for the kernel Q if there exists a positive measure ρ on $\mathbb{R}^d$ such that $Q(x,A) \geqslant \rho(A)$ for each $A \in \mathscr{B}(\mathbb{R}^d)$, $x \in B$.

Proposition 4.8.1. *Assume that B is a small set for Q. Furthermore, assume that there exists a measurable function $W : \mathbb{R}^d \to [0,\infty)$ that is defined on $\mathbb{R}^d$ and bounded on B, and a real number $b \geqslant 0$, such that*

$$QW \leqslant W - 1 + b\mathbb{1}_B. \tag{4.27}$$

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

$$\forall x \in \mathbb{R}^d, \ \|Q^n(x,\cdot) - \pi\|_{TV} \xrightarrow[n\to\infty]{} 0 \tag{4.28}$$

holds true.

Indeed, by [Meyn & Tweedie 2009, Theorem 11.3.4], the kernel Q is a so-called positive Harris recurrent, meaning among others that it has a unique invariant probability distribution. Moreover, Q is aperiodic, hence the convergence (4.28), as shown by, *e.g.*, [Meyn & Tweedie 2009, Theorem 13.0.1].

Lemma 4.8.2. *Assume that either Assumptions 4.6.1-(i)) 4.6.1-(ii)) hold if $Q_\gamma = P_\gamma$ or Assumption 4.7.2 holds and $r \leqslant R$ if $Q_\gamma = P_\gamma^r$, then for every $\gamma \in (0, \gamma_0]$, the kernel Q_γ admits a unique invariant measure π_γ. Moreover,*

$$\forall x \in \mathbb{R}^d, \ \left\|Q_\gamma^n(x,\cdot) - \pi_\gamma\right\|_{TV} \xrightarrow[n\to\infty]{} 0. \tag{4.29}$$

Finally, if $Q_\gamma = P_\gamma$, assumptions of Theorem 4.4.3 hold true and $\gamma \in \Gamma$ then π_γ is absolutely continuous w.r.t. the Lebesgue measure. If $Q_\gamma = P_\gamma^r$ and assumptions of Theorem 4.7.2 hold true, then $\pi_\gamma \in M^r$.

Proof. By the inequality (4.15), the kernel P_γ satisfies an inequality of the type (4.27), namely, $P_\gamma V \leqslant V - \alpha(\gamma)\theta + C\alpha(\gamma)\mathbb{1}_{\|x\|\leqslant R}$, for some $\theta, C > 0$. Similarly, under Assumption 4.7.2 and $r \leqslant R$, we have that for every $x \in \mathrm{cl}(B(0,r))$:

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

that is to say $\mathrm{cl}(B(0,r))$ is a small set for P_γ^r. Inequality of the type Assumption 4.6.1-(ii))–(iii)) then hold for e.g. $C = r$, $\alpha(\gamma) = 1$, $V = \|x\| + r\mathbb{1}_{\|x\|>r}$ and $p = \|x\|$.

Consider the case where $Q_\gamma = P_\gamma$, to prove that π_γ is absolutely continuous w.r.t. the Lebesgue measure, consider a λ-null set A. By the convergence (4.29), we obtain that for any $x \in \mathbb{R}^d$, $P_\gamma^n(x,A) \to \pi_\gamma(A)$. Now take $\nu \ll \lambda$. By Proposition 4.4.1, we have that $\nu P_\gamma^n \ll \lambda$. Hence, by the dominated convergence theorem,

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

If $Q_\gamma = P_\gamma^r$ we obtain the same result with the help of Proposition 4.7.1. □

Lemma 4.8.3. *Let either Assumptions 4.6.1-(i)) – 4.6.1-(iii)) hold if $Q_\gamma = P_\gamma$ or Assumption 4.7.2 hold and $r \leqslant R$ if $Q_\gamma = P_\gamma^r$. Let π_γ be the invariant distribution of Q_γ. Then, the family $\{\pi_\gamma : \gamma \in (0, \gamma_0]\}$ is tight.*

Proof. If $Q_\gamma = P_\gamma^r$ then the family π_γ is supported by $\mathrm{cl}(B(0,r))$ and is, therefore, tight. Otherwise we iterate (4.15), to obtain:

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

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

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

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

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

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

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

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

The next proposition will show that any accumulation point of π_γ is an invariant measure for the set-valued flow induced by the DI $\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t))$, first we introduce some definitions. Define the shift operator $\Theta_t : C(\mathbb{R}_+, \mathbb{R}^d) \to C(\mathbb{R}_+, \mathbb{R}^d)$ by $\Theta_t(x) = x(t + \cdot)$, and the projection operator $p_0 : C(\mathbb{R}_+, \mathbb{R}^d) \to \mathbb{R}^d$ by $p_0(x) = x(0)$. Then, we have the following definition (see [Roth & Sandholm 2013] for details):

Definition 4.8.1. *We say that $\pi \in M(\mathbb{R}^d)$ is an invariant distribution for the flow induced by the DI $\dot{\mathsf{x}}(t) \in \mathsf{H}(\mathsf{x}(t))$, if there is $\nu \in M(C(\mathbb{R}_+, \mathbb{R}^d))$, such that:*

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

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

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

Proposition 4.8.4. *Let Assumptions 4.4.1–4.5.2 and 4.6.1 hold true. Denote by π_γ the unique invariant distribution of P_γ. Let (γ_n) be a sequence on $(0, \gamma_0] \cap \Gamma$ s.t. $\gamma_n \to 0$ and π_{γ_n} converges narrowly to some probability measure π. Then, π is an invariant distribution for the flow induced by $\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t))$.*

Similarly, under Assumptions 4.4.1–4.5.1 and 4.7.1–4.7.2, for $r \leqslant R$, denoting π_γ the unique invariant distribution of P^r_γ, if $\pi_{\gamma_n} \to \pi$, then π is an invariant distribution for the flow induced by $\dot{\mathsf{x}}(t) \in -\partial F(\mathsf{x}(t)) - \mathcal{N}_r(\mathsf{x}(t))$.

Proof. Consider the case where $Q_\gamma = P_\gamma$. The proof essentially follows [Bianchi *et al.* 2019, section 7.]. Fix an $\varepsilon > 0$ and write π_n instead of π_{γ_n} for simplicity. By Lemma 4.8.3 we have a compact K such that $\pi_n(K) > 1 - \varepsilon$, we thus can define the conditional measures $\pi_n^K(A) := \frac{\pi_n(A \cap K)}{\pi_n(K)}$. Moreover, we have $\pi_n^K \in M_{abs}(K)$, therefore we can apply Theorem 4.5.1 and get that there is a compact set $\mathcal{C}$ of $C(\mathbb{R}^+, \mathbb{R}^d)$ such that $\mathbb{P}^{\pi^K_{\gamma_n}, \gamma_n} \mathsf{X}^{-1}_{\gamma_n}(\mathcal{C}) \geqslant 1 - \varepsilon$. Now we have

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

hence

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

Since ε is arbitrary this proves the tightness of $v_n := \mathbb{P}^{\pi_{\gamma_n}, \gamma_n} \mathsf{X}^{-1}_{\gamma_n}$. Take $\pi_n \to \pi$ and $v_n \to v \in M(C(\mathbb{R}_+, \mathbb{R}^d))$. We now prove that v is an invariant distribution for the flow induced by the DI associated to $-\partial F$ (see Definition 4.8.1.)
We have $\pi_n = v_n p_0^{-1}$, by continuity of p_0. Thus, $\pi = v p_0^{-1}$. Therefore, we have (iii) of Definition 4.8.1. Let $\eta > 0$. By weak convergence of v_n,

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

and

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

The last term converges to $1 - \varepsilon$, by Theorem 4.5.1, and by weak convergence we have $v(\{x \in C(\mathbb{R}_+, \mathbb{R}^d) : d(x, \mathsf{S}_{-\partial F}(\mathbb{R}^d)) \geqslant \eta\}) \geqslant (1 - \varepsilon)$, now letting $\eta \to 0$, by monotone convergence we have $v(\mathsf{S}_{-\partial F}(\mathbb{R}^d))) \geqslant 1 - \varepsilon$ which proves (i) of Definition 4.8.1. Finally, the second point of Definition 4.8.1 is shown just like in [Bianchi *et al.* 2019, section 7.].

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

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

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

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

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

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

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

In [Faure & Roth 2013] (see also [Aubin *et al.* 1991]), a version of Poincaré's recurrence theorem, well-suited for our set-valued evolution systems, was provided:

Proposition 4.8.5. *Each invariant measure for* S_H *is supported by* $BC_{\mathsf{S}_\mathsf{H}}$.

With the help of Proposition 4.8.5 we can finally prove Theorem 4.6.1.

Proof. Take $\gamma \in \Gamma$, $\varepsilon > 0$ and (x_n^γ) an associated SGD sequence. We have by (4.28):

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

Now take any sequence $\gamma_i \to 0$ with $\gamma_i \in \Gamma$, and π_{γ_i} the associated invariant distribution, we know from Lemmas 4.8.3-4.8.4 that we can extract a subsequence such that $\pi_{\gamma_i} \to \pi$, with π an invariant measure for the evolution system $\mathsf{S}_{-\partial F}$. Therefore by weak convergence we have:

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

where the last line comes from the Portmanteau theorem. We show that $\operatorname{supp} \pi \subset S$, and therefore the last term is equal to zero, which concludes the proof. To that end, we make use of Proposition 4.8.5, that shows that each invariant measure of $\mathsf{S}_{-\partial F}$ is supported by $\mathrm{BC}_{\mathsf{S}_{-\partial F}}$. Thus, it remains to show that $\mathrm{BC}_{\mathsf{S}_{-\partial F}} = \mathcal{Z}$ (which at the same time will ensure us that $\mathcal{Z}$ is nonempty). It is obvious that $\mathcal{Z} \subset \mathrm{BC}_{\mathsf{S}_{-\partial F}}$. To show the reverse inclusion, take $a \in L_{\mathsf{S}_{-\partial F}(a)}$. Then, there exists a solution x to the differential inclusion such that $\mathsf{x}(0) = a$ and $a \in L_\mathsf{x}$. But under Assumption 4.6.2 it holds ([Davis *et al.* 2020, lemma 5.2]) that $\|\dot{\mathsf{x}}(t)\| = \|\partial_0 F(\mathsf{x}(t))\|$ almost everywhere, and, moreover,

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

Therefore $\mathsf{x}(t) = a$ for each $t \geqslant 0$, thus, $a \in S$. Observing that $\mathcal{Z}$ is a closed set (since ∂F is graph-closed, see [Clarke *et al.* 1998, Proposition 2.1.5]), we obtain that $\mathrm{BC}_{\mathsf{S}_{-\partial F}} = \mathcal{Z}$.

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

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

Taking $a \in L_{\mathsf{S}_{-\partial F - \mathcal{N}_r}(a)}$, and x a solution to the associated differential inclusion with $\mathsf{x}(0) = a$, we get under Assumption 4.6.2 [Davis *et al.* 2020, Lemma 6.3.] that $\|\dot{\mathsf{x}}(t)\| = \min\{\|v\| : v \in \partial F(\mathsf{x}(t)) + \mathcal{N}_r(\mathsf{x}(t))\}$, and moreover,

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

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

4.8.6 Proof of Proposition 4.6.2

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

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

Given $L > 0$, we have by Assumption 4.5.1 and Markov's inequality that there exists a constant $K > 0$ such that

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

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

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

4.8.7 Proof of Proposition 4.6.3

By Lebourg's mean value theorem [Clarke *et al.* 1998, Theorem 2.4], for each $n \in \mathbb{N}$, there exists $\alpha_n \in [0,1]$ and $\zeta_n \in \partial F(u_n)$ with $u_n = x_n - \alpha_n \gamma \nabla f(x_n, \xi_{n+1}) \mathbb{1}_{\Delta_f}(x_n, \xi_{n+1})$, such that

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

and the proof of this theorem (see [Clarke *et al.* 1998, Theorem 2.4] again) shows that u_n can be chosen measurably as a function of (x_n, ξ_{n+1}).

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

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

We shall prove that

$$\begin{aligned} \mathbb{E}_n F(x_{n+1}) &\leqslant F(x_n) - \gamma \mathbb{1}_{\|x_n\|>2R}\|\nabla F(x_n)\|^2 + \gamma K \mathbb{1}_{\|x_n\|\leqslant 2R} \\ &+ \gamma^2 K \mathbb{1}_{\|x_n\|>2R}\left((1 + \|\nabla F(x_n)\|)\Big(\int \|\nabla f(x_n, s)\|^2 \,\mu(ds)\Big)^{1/2} + \int \|\nabla f(x_n, s)\|^2 \,\mu(ds)\right) \end{aligned} \tag{4.30}$$

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

$$\begin{aligned} F(x_{n+1}) = F(x_n) &- \gamma \mathbb{1}_{\|x_n\|\leqslant 2R}\mathbb{1}_{\|u_n\|\leqslant R}\langle \zeta_n, \nabla f_{n+1}\rangle \mathbb{1}_{\Delta_f}(x_n, \xi_{n+1}) \\ &- \gamma \mathbb{1}_{\|x_n\|\leqslant 2R}\mathbb{1}_{\|u_n\|>R}\langle \zeta_n, \nabla f_{n+1}\rangle \mathbb{1}_{\Delta_f}(x_n, \xi_{n+1}) \\ &- \gamma \mathbb{1}_{\|x_n\|>2R}\mathbb{1}_{\|u_n\|\leqslant R}\langle \zeta_n - \nabla F(x_n), \nabla f_{n+1}\rangle \\ &- \gamma \mathbb{1}_{\|x_n\|>2R}\mathbb{1}_{\|u_n\|>R}\langle \nabla F(u_n) - \nabla F(x_n), \nabla f_{n+1}\rangle \\ &- \gamma \mathbb{1}_{\|x_n\|>2R}\langle \nabla F(x_n), \nabla f_{n+1}\rangle \end{aligned} \tag{4.31}$$

We start with the second term at the right hand side of this inequality. Noting from Assumption 4.5.1 that

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

we have

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

and by integrating with respect to ξ_{n+1} and using Assumption 4.5.1 again, we get that

$$\gamma \mathbb{1}_{\|x_n\|\leqslant 2R}\mathbb{E}_n[\mathbb{1}_{\|u_n\|\leqslant R}|\langle \zeta_n, \nabla f_{n+1}\rangle \mathbb{1}_{\Delta_f}(x_n, \xi_{n+1})|] \leqslant \gamma K \mathbb{1}_{\|x_n\|\leqslant 2R}. \tag{4.32}$$

Using Assumption 4.5.1, the next term at the right hand side of (4.31) can be bounded as

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

which leads to

$$\gamma \mathbb{1}_{\|x_n\| \leqslant 2R} \mathbb{E}_n[\mathbb{1}_{\|u_n\|>R} |\langle \zeta_n, \nabla f_{n+1} \rangle 1_{\Delta_f}(x_n, \xi_{n+1})|] \leqslant \gamma K \mathbb{1}_{\|x_n\| \leqslant 2R} \tag{4.33}$$

by using Assumption 4.5.1.

We tackle the next term at the right hand side of (4.31). Fix a $x_\star \notin \mathrm{cl}(B(0,R))$. By our assumptions it holds that each $x \notin \mathrm{cl}(B(0,R))$,

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

where $\beta'(\cdot)$ is square integrable thanks to Assumption 4.5.1. Since

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

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

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

Using this result, we write

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

Consequently,

$$\begin{aligned}
\gamma \mathbb{1}_{\|x_n\|>2R} \mathbb{E}_n[\mathbb{1}_{\|u_n\| \leqslant R} |\langle \zeta_n, \nabla f_{n+1} \rangle|] &\leqslant \gamma K \mathbb{1}_{\|x_n\|>2R} \Big(\int \|\nabla f(x_n, s)\|^2 \mu(ds) \Big)^{1/2} \mu[\beta'(\cdot) \geqslant K/\gamma]^{1/2} \\
&\leqslant \gamma^2 K \mathbb{1}_{\|x_n\|>2R} \Big(\int \|\nabla f(x_n, s)\|^2 \mu(ds) \Big)^{1/2}.
\end{aligned} \tag{4.34}$$

Similarly,

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

thus,

$$\gamma \mathbb{1}_{\|x_n\|>2R} \mathbb{E}_n \left[\mathbb{1}_{\|u_n\| \leqslant R} |\langle \nabla F(x_n), \nabla f_{n+1} \rangle| \right] \leqslant \gamma^2 K \mathbb{1}_{\|x_n\|>2R} \|\nabla F(x_n)\| \Big(\int \|\nabla f(x_n, s)\|^2 \mu(ds) \Big)^{1/2}. \tag{4.35}$$

We have that ∇F is Lipschitz outside $\mathrm{cl}(B(0,R))$. Thus, the next to last term at the right hand side of (4.31) satisfies

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

and we get that

$$\gamma \mathbb{1}_{\|x_n\|>2R}\mathbb{1}_{\|u_n\|>R}\mathbb{E}_n\left[|\langle \nabla F(u_n)-\nabla F(x_n), \nabla f_{n+1}\rangle|\right] \leqslant \gamma^2 K \mathbb{1}_{\|x_n\|>2R}\int \|\nabla f(x_n,s)\|^2 \mu(ds). \tag{4.36}$$

Finally, we have

$$-\gamma \mathbb{1}_{\|x_n\|>2R}\mathbb{E}_n\left[\langle \nabla F(x_n), \nabla f_{n+1}\rangle\right] = -\gamma \mathbb{1}_{\|x_n\|>2R}\|\nabla F(x_n)\|^2. \tag{4.37}$$

Inequalities (4.32)–(4.37) lead to (4.30).

Using Assumption (iii) of Proposition 4.6.3, Inequality (4.30) leads to Inequality (4.17). The validity of Assumptions 4.6.1-(ii)) and 4.6.1-(iii)) can then be checked easily.

4.8.8 Proof of Proposition 4.7.1

The next Lemma is the key ingredient in the proofs of Section 4.7.

Lemma 4.8.6. *Assume that $f(\cdot, s)$ is locally Lipschitz continuous for every $s \in \Xi$. Then for $\lambda^1 \otimes \lambda^d \otimes \mu$-almost all (r, x, s) with $r > 0$, it holds that $(\Pi_r(x), s) \in \Delta_f$. For $\lambda^1 \otimes \lambda^d$-almost all (r, x) with $r > 0$, it holds that $\Pi_r(x) \in \mathcal{D}_F$.*

Proof. Our first aim is to show that

$$\int \mathbb{1}_{\Delta_f^c}(\Pi_r(x), s)\, \lambda^1(\mathrm{d}r) \otimes \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s) = 0\,. \tag{4.38}$$

First, note by Fubini's theorem that

$$0 = \int \mathbb{1}_{\Delta_f^c}(x,s)\, \lambda^d(\mathrm{d}x) \otimes \mu(\mathrm{d}s) = \int_{\Xi\times\mathbb{R}_+}\int_{\mathbb{S}(1)} \mathbb{1}_{\Delta_f^c}(r\theta, s) r^{d-1} \varrho_1(\mathrm{d}\theta)\; \mu \otimes \lambda^1(\mathrm{d}s \times \mathrm{d}r)\,, \tag{4.39}$$

that is to say, $\varrho(\{\theta : (r\theta, s) \in \Delta_f\}) = 0$ for $\mu \otimes \lambda^1$ almost every (s, r) with $r > 0$. Decompose Equation (4.38) as

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

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

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

Similarly,

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

with the last equality coming from Equation (4.39). Hence (4.38). The second statement can be proven along similar lines. □

Consider $r > 0$ such that the conclusion of Lemma 4.8.6 hold. Then the almost sure equality of all projected SGD sequence is proven in the same way as in Proposition 4.4.1. We can therefore consider the lazy projected SGD sequence $x_{n+1}^{\gamma,r} = \Pi_r(x_n^{\gamma,r} - \gamma\varphi_0(x_n^{\gamma,r}, \xi_{n+1}))$. By Assumption 4.7.1 the law of $x_{n+1/2}^{\gamma,r} \triangleq x_n^{\gamma,r} - \gamma\varphi_0(x_n^{\gamma,r}, \xi_{n+1})$ is Lebesgue-absolutely continuous. Take A a borel set of $\mathbb{R}^d$ such that $\lambda(A) = \varrho_r(A) = 0$. Then

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

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

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

which finishes the proof.

4.8.9 Proof of Theorems 4.7.2 and 4.7.3

Noting that the law of $x_n^{\gamma,r} - \gamma\varphi_0(x_n^{\gamma,r}, \xi_{n+1})$ is Lebesgue-absolutely continuous by Assumption 4.7.1, the first point of Theorem 4.7.2 comes from Lemma 4.8.6. The second point comes upon noticing that $\Pi_r(x) - x \in -\mathcal{N}_r(\Pi_r(x))$.

Theorem 4.7.3 is proved in the same way as Theorem 4.5.1, by applying [Bianchi *et al.* 2019, Theorem 5.1.] with $h(s,x) = -\nabla F(x) - 1/\gamma(x - \gamma\nabla f(x,s) - \Pi_r(x - \gamma\nabla f(x,s))) \in -\nabla F(x) - \mathcal{N}_r(x - \gamma\nabla f(x,s))$ and $H(x) = H(s,x) = -\partial F(x) - \mathcal{N}_r(x)$.

Chapter 5

Stochastic subgradient descent escapes active strict saddles

5.1 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

$$x_{n+1} \in x_n - \gamma_n \partial f(x_n) + \gamma_n \eta_{n+1}, \tag{5.1}$$

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

Before delving into the subject of convergence towards minimizers, let us first consider the set $\mathcal{Z} := \{x \in \mathbb{R}^d : 0 \in \partial f(x)\}$ of *Clarke critical points* of f, which is generally larger than the set of minimizers, in the non-convex case. In order to ensure the convergence of (x_n) to $\mathcal{Z}$, the sole local Lipschitz property of f is not enough (see [Daniilidis & Drusvyatskiy 2019] for a counterexample), and some form of structure for the function f is required. Since the work of Bolte *et.al.* [Bolte *et al.* 2007] 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 (cf. Section 2.4), 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 [Davis *et al.* 2020]. Another work in the same line is [Majewski *et al.* 2018]. Bolte and Pauwels [Bolte & Pauwels 2019] generalize the algorithm (5.1) by replacing ∂f with an arbitrary so-called conservative field. The constant step size regime $\gamma_n \equiv \gamma$ is considered in [Bianchi *et al.* 2021a].

Thanks to these contributions, the convergence of (x_n) to the set $\mathcal{Z}$ is now well understood. However, as said above, $\mathcal{Z}$ is in general strictly larger than the set of

minimizers, and can contain "spurious" points such as local maximizers or saddle points. The issue of the *non-convergence* of the sequence given by (5.1) towards spurious critical points is therefore crucial. The present chapter 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 (5.1) is non zero. Here, the seminal works of Pemantle [Pemantle 1990] and Brandière and Duflo [Brandière & Duflo 1996] allow to establish the non-convergence of the Stochastic Gradient Descent (and, more generally, of Robbins-Monro algorithms) to a certain type of spurious critical points, sometimes referred to as *traps* or *strict saddle*. A critical point of a smooth function f is called a trap if the Hessian matrix of f at this point admits at least one negative eigenvalue. With probability one, the sequence (x_n) cannot converge to a trap, provided that the projection of the random perturbation η_n onto the eigenspace of corresponding to the negative eigenvalues of the Hessian matrix (henceforth, eigenspace of negative curvature) has a non vanishing variance.

- The noiseless case where $\eta_n \equiv 0$, studied for smooth functions by [Lee *et al.* 2016]. 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 [Lee *et al.* 2016] the trap avoidance is due to the random initialization of the algorithm, whereas in [Brandière & Duflo 1996, Pemantle 1990], 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 [Davis & Drusvyatskiy 2021] of Davis and Drusvyatskiy. The spurious points that were considered in this reference are the so-called *active strict saddles*. Informally, a critical point is an active strict saddle if it lies on a manifold M such that *i)* f varies sharply outside of M, *ii)* the restriction of f to M is smooth, and *iii)* the Riemannian Hessian of f on M has at least one negative eigenvalue. For instance, the function $f : \mathbb{R}^2 \to \mathbb{R}, (y, z) \mapsto |z| - y^2$ admits the point $(0, 0)$ as an active strict saddle with $M = \mathbb{R} \times \{0\}$, and the restriction of f to M is the function $f_M(y, 0) = -y^2$, which obviously has a second-order negative curvature. In this setting, and assuming that f is weakly convex, the article [Davis & Drusvyatskiy 2021] focuses on the noiseless case, and study variants of the (implicit) *proximal point algorithm* rather than the (explicit) subgradient descent. Similarly to [Lee *et al.* 2016], they show that for Lebesgue almost every initialization point, different versions of the

proximal algorithm avoid active strict saddles with probability one. Such a result is possible due to the fact that proximal methods implicitly run a gradient descent on a smoothened version of f - the Moreau envelope.

Contrary to [Davis & Drusvyatskiy 2021], the algorithm (5.1) studied in this chapter 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 [Pemantle 1990, Brandière & Duflo 1996], our analysis strongly relies on the presence of the additive random perturbation η_n.

In the framework of definable functions, we investigate the problem of the avoidance of the active strict saddle points. Our approach goes as follows. First, we need to show that the iterates (x_n) converge sufficiently fast to M, thanks to the sharpness of f outside this manifold. To that end, we first rely on the fact that when f is definable, its graph always admits a so-called *Verdier stratification*, which is perhaps less known than the Whitney stratification, and is a refinement of the latter [Loi 1998]. The key advantage of the Verdier over the Whitney stratification lies in a Lipschitz-like condition on the (Riemannian) gradients of f on two adjacent stratas, which is established in the chapter. Our second tool is an assumption that we term as the *angle condition*. Roughly, this assumption provides a lower bound on the inner product between the subgradients of f at x and the normal direction from M to x when the point x is near M. The angle condition allow to control the distance between the iterate x_n of Algorithm (5.1) and the manifold M. As the restriction f_M of f to M is smooth, the projected iterates, using the Verdier stratification property, are shown to follow a dynamics which is similar to a (smooth) Stochastic Gradient Descent, up to a residual term induced by the projection step. In that sense, the avoidance of active strict saddles in the non-smooth setting follows from the avoidance of traps in the smooth setting, as established in [Brandière & Duflo 1996]. 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 [Drusvyatskiy *et al.* 2016, Davis & Drusvyatskiy 2021]. 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 $f_u(x) \overset{\Delta}{=} f(x) - \langle u, x \rangle$ 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 5.3.2 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 [Drusvyatskiy *et al.* 2016, Davis & Drusvyatskiy 2021], the SGD algorithm (5.1) converges to a local minimum when f is a weakly convex function, assuming that the noise is omnidirectional enough at the strict saddles. We emphasize the fact that, while the genericity of the active strict saddles is established in the above sense for weakly convex functions, no assumption on weak convexity is made for our avoidance of traps result.

Let us summarize the contributions of this chapter:

- 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 [Bolte *et al.* 2007]. While such stratifications are well-known in the literature on o-minimal structures [Loi 1998], 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 [Bolte *et al.* 2007, Proposition 4] and Lemma 2.4.10 below) to the case of definable, locally Lipschitz continuous functions by establishing a Lipschitz-like condition on the (Riemannian) gradients of two adjacent stratas.

- With the help of the Verdier and the angle conditions, we show that the SGD avoids the active strict saddles if the noise η_n is omnidirectional enough.

The chapter is organized as follows. In Section 5.2 we fix the notations and prove the reinforced projection formula stated in Theorem 5.2.1. In Section 5.3, we discuss the notion of an active strict saddle. After recalling some results of [Davis & Drusvyatskiy 2021], 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 5.4, we state the main result of this chapter, namely, the avoidance of active strict saddles. Section 5.5 is devoted to the proofs.

5.2 Preliminaries

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

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

5.2.1 Reinforced projection formula

The following theorem, which we believe to be of independent interest, is the first main result of this chapter. It is an improvement of the projection formula of [Bolte & Pauwels 2019] (see Lemma 2.4.10) when the definable function is locally Lipschitz continuous.

Theorem 5.2.1 (Reinforced projection formula)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a definable, locally Lipschitz continuous function. Let p be a positive integer. There is (X_i), a definable Verdier C^p stratification of $\mathbb{R}^d$, such that for each $y \in X_i$ and each X_j such that $\overline{X_i} \cap X_j \neq \varnothing$, there is $C, \delta > 0$, such that for any two points $y' \in B(y,\delta) \cap X_i$, $x \in B(y,\delta) \cap X_j$,*

$$\left\| P_{T_{y'}X_i}(\nabla_{X_j} f(x)) - \nabla_{X_i} f(y') \right\| \leqslant C \left\| x - y' \right\| , \tag{5.2}$$

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

$$\left\| P_{T_{y'}X_i}(v) - \nabla_{X_i} f(y') \right\| \leqslant C \left\| x - y' \right\| . \tag{5.3}$$

Proof. In this proof $C' > 0$ will denote some constant that can change from line to line. Consider (S_i) and (X_i) as in Lemma 2.4.10. We claim that for any index j and $x \in X_j$, we have $T_{x,f(x)}S_j = \{(h, \langle \nabla_{X_i} f(x), h\rangle) : h \in T_x X_j\}$. Indeed, consider $(h_x, h_f) \in T_{x,f(x)}S_j$ and a C^p curve $c\colon (-\varepsilon,\varepsilon)$ s.t. $\dot{c}(0) = (h_x, h_f)$. Consider a C^p function F that agrees with f on X_j, then $(c_x(t), c_f(t)) = (c_x(t), F(c_x(t)))$ and we have $\dot{c}_x(0) = h_x$ and $\dot{c}_f(0) = \langle \nabla F(x), h_x\rangle = \langle \nabla_{X_j} f(x), h_x\rangle$.

Consider (S_i') a Verdier stratification of $\mathrm{Graph}(f)$ compatible with (S_i). Then the projection of S_i' onto its first d coordinates, that we denote X_i', is still a submanifold s.t. f is C^p on X_i'. Consider $(y, f(y)) \in S_i'$, S_j' a neighboring strata and C, δ as in Equation (2.10). Denote by L the Lipschitz constant of f on $B(y,\delta)$ and $\delta' = \frac{\delta}{L+1}$. Then, for every $x \in B(y,\delta')$, we have:

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

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

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

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

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

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

and

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

which proves the first statement.

Now, one can choose C, δ such that Inequality (5.2) holds uniformly on all of the stratas X'_j that are neighboring X'_i. Consider a sequence $x_n \to x$ such that (x_n) lies in the stratas of full dimension (which implies that f is differentiable at x_n) and $\nabla f(x_n) \to v$, for n large enough we will have that $x_n \in B(y, \delta)$ and, therefore, $\left\| P_{T_{y'} X_i}(\nabla f(x_n)) - \nabla_{X_i} f(y') \right\| \leqslant C \|x_n - y'\|$. Hence, passing to the limit, we have that $\left\| P_{T_{y'} X_i}(v) - \nabla_{X_i} f(y') \right\| \leqslant C \|y' - x\|$. Since any element of $\partial f(x)$ is a convex combination of such v, the second statement is proved. □

5.3 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 $\mathcal{Z} := \{x \in \mathbb{R}^d : 0 \in \partial f(x)\}$.

5.3.1 Definition and Existing Results

Let $p \geqslant 2$ be an integer.

Definition 5.3.1 (Active manifold, [Lewis 2002])**.** *Consider $x^* \in \mathcal{Z}$. A set $M \subset \mathbb{R}^d$ is called a C^p* active manifold around x^**, if there is a neighborhood U of x^* such that the following holds.*

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

ii) ***Sharpness condition:***

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

Definition 5.3.2 (Active strict saddle)**.** *We say[1] that a point $x^* \in \mathcal{Z}$ is an* active strict saddle (of order p) *if there exists a C^p active manifold M around x^*, and a vector $w \in T_{x^*} M$, such that $\nabla_M f(x^*) = 0$ and $\mathcal{H}_{f,M}(x^*)(w) < 0$.*
We say that f satisfies the active strict saddle property (of order p)*, if it has a finite number of Clarke critical points, and each of these points is either an active strict saddle of order p or a local minimizer .*

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

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

[1] The definition of active strict saddles provided in [Davis & Drusvyatskiy 2021] involves the notion of parabolic subderivatives. In this paper, we found convenient to use the equivalent Definition 5.3.2, which is closer in spirit to notions of differential geometry.

Example 5.3.1. *The point* $(0,0)$ *is an active strict saddle of the function* $f : \mathbb{R}^2 \to \mathbb{R}$ *given by* $f(y,z) = -y^2 + |z|$*. Indeed,*

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

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

While the definition of an active strict saddle might seem peculiar at first glance, the following proposition of Davis and Drusvyatskiy shows that a generic definable and weakly convex function satisfies a strict saddle property. The proof is grounded in the work of [Drusvyatskiy *et al.* 2016].

Proposition 5.3.1 ([Davis & Drusvyatskiy 2021, Theorem 2.9])**.** *Assume that* f *is definable and weakly convex. Define* $f_u(x) := f(x) - \langle u, x\rangle$*, for every* $u \in \mathbb{R}^d$*. Then, for every* $p \geqslant 2$ *and for Lebesgue-almost every* $u \in \mathbb{R}^d$*,* f_u *has the active strict saddle property of order* p*.*

It is worth noting that the result of [Davis & Drusvyatskiy 2021, Theorem 2.9] is in fact a bit stronger than Proposition 5.3.1, 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 5.3.1 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 5.3.2. *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.*

5.3.2 Verdier and Angle Conditions

On the top of the items *i-ii)* of Definition 5.3.1, we introduce the following useful conditions.

Definition 5.3.3. *Let* M *be a* C^1 *active manifold around some* $x^* \in \mathcal{Z}$*. We say that* M *satisfies the Verdier condition and the angle condition, if the following conditions hold respectively.*

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

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

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

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

Definition 5.3.4. *An active strict saddle x^* is said to satisfy the* Verdier and angle conditions*, if the active manifold M in Definition 5.3.2 satisfies the Verdier and angle conditions. The function f is said to satisfy the* active strict saddle property of order p with the Verdier and angle conditions*, if it satisfies the active strict saddle property of order p and if every active strict saddle satisfies the Verdier and angle conditions.*

The Verdier condition merely states that M is one of the stratas of the Verdier stratification of Theorem 5.2.1 The purpose of the angle condition is to relate, close to M, the linear growth of the function f and the lower boundedness of the inner product between the subgradients of f at x and the normal direction to M. The latter will allow us to prove that the iterates of SGD converge to M fast enough.

Remark 20. *Let M be an active manifold around x^*. As it will be clear from the proof of Theorem 5.3.2, when f is weakly convex, M always satisfies the angle condition. Otherwise stated, the angle condition is simply* true *in case of weakly convex functions. However, as the following example shows, one is able to find many natural examples of functions which are not weakly convex, and yet satisfy this condition.*

Example 5.3.3. *The function $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(y, z) = -y^2 - |z|$ is not weakly convex. Its unique Clarke critical point $(0, 0)$ is an active strict saddle, satisfying the Verdier and the angle conditions.*

Example 5.3.3 shows that the Verdier and angle conditions can be satisfied with no need for f to be weakly convex. Nevertheless, more can be said when this assumption holds. The following theorem strengthen the genericity result of Proposition 5.3.1 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 5.3.2. *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 5.2.1. Upon noticing that in the proof of [Drusvyatskiy *et al.* 2016, Corollary 4.8 and Theorem 4.16] the active manifold [2] can be chosen adapted to $\{X_1, \ldots, X_k\}$, the existence of an active manifold with a Verdier condition follows from [Davis & Drusvyatskiy 2021,

[2] The name *active manifold* follows the work of [Davis & Drusvyatskiy 2021], while in [Drusvyatskiy *et al.* 2016] they are called identifiable manifolds.

Theorem 2.9, Appendix A]. For the angle condition note that by weak convexity of f there is $\rho \geqslant 0$ such that:

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

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

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

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

5.4 Avoidance of Active Strict Saddles

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

$$x_{n+1} = x_n - \gamma_n v_n + \gamma_n \eta_{n+1}\,, \tag{5.4}$$

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

Assumption 5.4.1.

i) The function f is path differentiable.

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

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

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

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

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

Assumption 5.4.2. *The point x^* is an active strict saddle of order 4 satisfying the Verdier and angle conditions.*

Since $\mathcal{H}_{f,M}(x^*)$ is a quadratic form we can write down $\mathbb{R}^d = E^- \oplus E^+$, where E^- (respectively E^+) is the vector space spanned by the eigenvectors of the associated symmetric bilinear form that have negative (respectively nonnegative eigenvalues). Note that by results of Section 2.3 we have that $E^- \subset T_{x^*}M$ and by Assumption 5.4.2 we have that $\dim E^- \geqslant 1$.

Assumption 5.4.3. *The following holds almost surely on the event $[x_n \to x^*]$.*

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

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

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

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

The following theorem is the main result of this chapter.

Theorem 5.4.1. *Let Assumptions 5.4.1–5.4.3 hold. Then* $\mathbb{P}(x_n \to x^*) = 0$.

Combining Theorem 5.4.1 with the results of Section 5.3.2 we obtain that, under appropriate assumptions, the SGD on a generic definable, weakly convex function converges to a local minimizer. We state this result in the following corollary.

Corollary 5.4.2. *Let Assumptions 5.4.1 and 5.4.2 hold. Assume that* f *has the active strict saddle property of order 4 with the Verdier and angle conditions. Moreover, assume that almost surely the following holds.*

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

ii) *For every* $C > 0$,

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

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

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

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

5.5 Proof of Theorem 5.4.1

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 $\mathcal{H}_{f,M}(0)(w) < 0$.

The general idea of the proof of Theorem 5.4.1 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 [Brandière & Duflo 1996]. In this setting, the remainders z_n will be shown to be small enough so as not to alter fundamentally the approach of [Brandière & Duflo 1996].

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

$$\forall v \in \partial f(x_n), \quad \langle v, z_n\rangle \gtrsim \|z_n\|\,. \tag{5.5}$$

This will be the object of Proposition 5.5.3 below. The idea is to show that for these n, it holds that $f(x_n) - f(y_n) \gtrsim \|z_n\|$, and then, to use the angle condition (iv)) of Definition 5.3.3.

Let us temporarily assume that n_0 is deterministic, and work on $n \geqslant n_0$. Keeping Inequality (5.5) aside for further use, the next step is to make a Taylor development of $y_{n+1} = P_M(x_{n+1})$ around x_n. This leads to

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

where we used the Lipschitz continuity of the Jacobian matrix function $J_{P_M}(\cdot)$. Using Equation (5.4), we rewrite the last display as

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

Now, Lemma 2.3.2 shows that $J_{P_M}(y_n)$ coincides with the linear operator $P_{T_{y_n}M}$. Furthermore, the Verdier condition (iii)) of Definition 5.3.3 asserts that $P_{T_{y_n}M}(v_n) = \nabla_M f(y_n) + \mathcal{O}(\|z_n\|)$. Altogether, we obtain the Robbins-Monro iteration

$$y_{n+1} = y_n - \gamma_n \nabla_M f(y_n) + \gamma_n P_{T_{y_n}M}\eta_{n+1} + \gamma_n^2 \mathcal{O}(1+\|\eta_{n+1}\|^2) + \gamma_n \mathcal{O}(\|z_n\| \, (1+\|\eta_{n+1}\|)). \tag{5.6}$$

Had we not have the last term $\gamma_n \mathcal{O}(\|z_n\| \, (1 + \|\eta_{n+1}\|))$ at the right hand side, the approach of Brandière and Duflo would have been enough to obtain the nonconvergence of y_n to zero under our assumptions on the noise. The presence of this term requires us to weaken a bit their conditions. This will be done in Proposition 5.5.1. In the case of Equation (5.6), this proposition asserts that the trap avoidance remains true if

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

where

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

This is where Inequality (5.5) comes into play to establish this bound.

So far, we have assumed abusively that the moment n_0 after which $\|x_n\|$ is small and (5.5) is satisfied is deterministic. To deal with this issue, in Section 5.5.2, on an arbitrary large event A, we construct a sequence (y_n) that is (for n large enough) equal to $(P_M(x_n))$ on $A \cap [x_n \to 0]$ and satisfies an equation of the form (5.6) almost surely. Proposition 5.5.1 will allow us to prove that $\mathbb{P}([x_n \to 0] \cap A) \leqslant \mathbb{P}([y_n \to 0]) = 0$ and since the event A is arbitrary large, this will prove Theorem 5.4.1.

5.5.1 Preliminary: Avoidance of Traps in the Smooth Case

The following proposition is nearly a quote of Brandière and Duflo's theorem [Brandière & Duflo 1996, Theorem 1]. As discussed below, we alleviate some hypotheses of [Brandière & Duflo 1996].

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

Proposition 5.5.1. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(\mathscr{F}_n)$ a filtration and (γ_n) a sequence of deterministic nonnegative step sizes such that $\sum_k \gamma_k = +\infty$ and $\sum_k \gamma_k^2 < +\infty$. Let d be an integer and $D : \mathbb{R}^d \to \mathbb{R}^d$ be such that $D(0) = 0$ and there is a neighborhood of 0 such that on it D is continuously differentiable, with Lipschitz continuous Jacobian. Consider the $\mathbb{R}^d$–valued stochastic process (y_n) given by*

$$y_{n+1} = y_n - \gamma_n D(y_n) + \gamma_n \tilde{\eta}_{n+1} + \gamma_n \varrho_{n+1} + \gamma_n \tilde{\varrho}_{n+1}\,, \tag{5.7}$$

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

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

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

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

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

v) We have that:

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

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

Proposition 5.5.1 is similar to [Brandière & Duflo 1996, Theorem 1], except for the presence of the sequence $(\tilde{\varrho}_n)$. As the proof is mainly an adaptation of the proof of [Brandière & Duflo 1996, Theorem 1], we provide a sketch of proof in the appendix.

5.5.2 Application to Algorithm (5.4)

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 2.3.

Lemma 5.5.2. *Let Assumption 5.4.2 hold and let $r > 0$ be such that $P_M : B(0, r) \to M$ is well defined and is C^3 and that there is a C^4 function $F : B(0, r) \to \mathbb{R}$ that*

agrees with f on $M \cap B(0,r)$. Then, the function $F \circ P_M$ is C^3 on $B(0,r)$ and for $y \in M \cap B(0,r)$, we have:

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

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

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

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

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

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

The next proposition is a key element in our proof. To not interrupt our exposition its proof is provided in Section 5.5.3.

Proposition 5.5.3. *Let Assumptions 5.4.1–5.4.3 hold. There is $\beta, r_1 > 0$, such that for every $r < r_1$, almost surely on the event $[x_n \to 0]$, $x_n \in V_r(\beta)$ for all n large enough.*

In the remainder, we fix $\beta, r_1 > 0$ as those provided by the previous proposition. We let U be the neighborhood around zero that verify conditions of Definition 5.3.3. In the following, we choose $r \leqslant r_1$ such that P_M is C^3 on $\overline{B(0,r)}$, and $\overline{B(0,r)} \subset U$. The value of r, while always satisfying these requirements, will be adjusted in the course of the proof.

Firstly, to reduce technical issues, we notice that as in [Brandière & Duflo 1996, Section I.2] to prove Theorem 5.4.1 we can actually replace Assumption 5.4.3 by the following, more easy to handle, assumption. The notation $\mathbb{E}_n[\cdot]$ stands for $\mathbb{E}[\cdot|\mathscr{F}_n]$.

Assumption 5.5.1. *Almost surely, the sequence (η_n) is such that $\mathbb{E}_n[\eta_{n+1}] = 0$ and there is $A, B > 0$ such that for all $n \in \mathbb{N}$, we have:*

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

and

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

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

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

Note that the sequence of events $(\mathcal{A}_N)$ is increasing for the inclusion. Furthermore, Proposition 5.5.3 shows that

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

Thus,

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

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

$$\mathbb{P}\left[\left[x_n \to 0\right] \cap \mathcal{A}_{N(\delta)}\right] \geqslant \mathbb{P}\left[x_n \to 0\right] - \delta. \tag{5.8}$$

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

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

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

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

and let

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

Observe that y_n^N and z_n^N are both $\mathscr{F}_n$–measurable for all $n \geqslant N$. To establish Theorem 5.4.1, we shall show that for each $N \geqslant 0$,

$$\mathbb{P}\left[y_n^N \xrightarrow[n\to\infty]{} 0\right] = 0. \tag{5.9}$$

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,

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

Consequently, with the convergence (5.9) at hand, we get from Inequality (5.8) that $\mathbb{P}[x_n \to 0] \leqslant \delta$. Since δ is arbitrary, we obtain that $\mathbb{P}[x_n \to 0] = 0$.

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

Proposition 5.5.4. *Let Assumptions 5.4.1–5.4.2 and 5.5.1 hold. Then, the sequence $(y_n^N)_{n\geqslant N}$ satisfies the recursion:*

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

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

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

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

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

We furthermore have:

iv) The subspace E^- *defined before Assumption 5.4.3 coincides with the eigenspace of the matrix* $J_D(0)$ *corresponding to its negative eigenvalues.*

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

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

Lemma 5.5.5. *For* r *small enough, there is* $C > 0$ *such that for* $x, x' \in B(0,r)$, *we have:*

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

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

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

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

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

Proof of Proposition 5.5.4. Letting $n \geqslant N$, we write

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

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

Using Lemma 5.5.5 and Equation (5.4), we obtain

$$\begin{aligned} y^N_{n+1} &= \left(y^N_n + J_{P_M}(y^N_n)(x_{n+1} - x_n)\right)\mathbb{1}_{\tau_N > n+1} + \gamma_n \varrho^N_{n+1} + \gamma_n \zeta^N_{n+1} \\ &\quad + \left(y^N_n - \gamma_n D(y^N_n)\right)\mathbb{1}_{\tau_N \leqslant n+1} + \gamma_n\left(J_{P_M}(y^N_n)\mathbb{1}_{\tau_N = n+1} + \mathbb{1}_{\tau_N \leqslant n}\eta_{n+1}\right) \\ &= \left(y^N_n - \gamma_n J_{P_M}(y^N_n)v_n + \gamma_n J_{P_M}(y^N_n)\eta_{n+1}\right)\mathbb{1}_{\tau_N > n+1} + \gamma_n \varrho^N_{n+1} + \gamma_n \zeta^N_{n+1} \\ &\quad + \left(y^N_n - \gamma_n D(y^N_n)\right)\mathbb{1}_{\tau_N \leqslant n+1} + \gamma_n\left(J_{P_M}(y^N_n)\mathbb{1}_{\tau_N = n+1} + \mathbb{1}_{\tau_N \leqslant n}\eta_{n+1}\right), \end{aligned}$$

where ϱ^N_{n+1} and ζ^N_{n+1} are $\mathscr{F}_{n+1}$–measurable, and satisfy with the notations of Lemma 5.5.5

$$\left\|\zeta^N_{n+1}\right\| = \gamma_n^{-1}\left\|R_1(x_n, x_{n+1}, y^N_n)\right\|\mathbb{1}_{\tau_N > n+1} \leqslant C\gamma_n^{-1}\|x_{n+1} - x_n\|\left\|z^N_n\right\| \leqslant C(1 + \|\eta_{n+1}\|)\left\|z^N_n\right\|$$

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

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

Using Lemma 2.3.2 in conjunction with the Verdier condition (iii)) of Definition 5.3.3, 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

$$\begin{aligned}\tilde{\eta}_{n+1}^N &= \left(\mathbb{1}_{\tau_N>n}J_{P_M}(y_n^N) + \mathbb{1}_{\tau_N\leqslant n}\right)\eta_{n+1}, \text{ and}\\ \tilde{\varrho}_{n+1}^N &= \zeta_{n+1}^N + \tilde{\zeta}_n^N.\end{aligned} \tag{5.10}$$

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 5.5.1.

Assertion iv)) follows from Lemma 5.5.2.

To establish v)), we write

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

On the event $[y_n^N \to_n 0]$, it holds that $J_{P_M}(y_n^N) \to_n J_0$. By Lemma 2.3.2, J_0 is the orthogonal projection on T_0M, thus, $\lim_{y_n^N\to_n 0} P_{E^-}J_{P_M}(y_n^N) = P_{E^-}$. Consequently, we obtain on the event $[y_n^N \to_n 0]$:

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

and by Assumption 5.5.1. Proposition 5.5.4 is proven. □

Proposition 5.5.6. *Let Assumptions 5.4.1–5.4.2 and 5.5.1 hold true. Then, there is $C > 0$ such that*

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

Proof. We shall use the notation

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

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

We start with the development

$$\begin{aligned}\left\|z_{n+1}^N\right\|^2 &= \left\|p_{n+1}^N\right\|^2 \mathbb{1}_{n+1<\tau_N} \\ &\leqslant \left\|p_{n+1}^N\right\|^2 \mathbb{1}_{n<\tau_N} = \left\|p_{n+1}^N - p_n^N + p_n^N\right\|^2 \mathbb{1}_{n<\tau_N} \\ &= \left\|z_n^N\right\|^2 + 2\langle x_{n+1} - x_n, z_n^N\rangle - 2\langle y_{n+1}^N - y_n^N, z_n^N\rangle + \left\|p_{n+1}^N - p_n^N\right\|^2 \mathbb{1}_{n<\tau_N}.\end{aligned} \tag{5.11}$$

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

We first show that

$$\mathbb{E}_n|\langle y_{n+1}^N - y_n^N, z_n^N\rangle| \leqslant C\gamma_n \left\|z_n^N\right\|^2 + C\gamma_n^2. \tag{5.12}$$

By Proposition 5.5.4,

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

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 (5.10) 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 2.3.2, thus, $\langle \tilde{\eta}_{n+1}^N, z_n^N\rangle = 0$. As a consequence,

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

From Proposition 5.5.4 again, we have

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

and

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

Inequality (5.12) is obtained by combining these inequalities.

We next show succinctly that

$$\mathbb{E}_n \left\|p_{n+1}^N - p_n^N\right\|^2 \mathbb{1}_{n<\tau_N} \leqslant C\gamma_n^2. \tag{5.13}$$

Indeed,

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

and the result follows by standard calculations making use of the results of Proposition 5.5.4.

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

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

Getting back to Inequality (5.11), and using this result in conjunction with the inequalities (5.12) and (5.13), we obtain that

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

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

Corollary 5.5.7. *Under the assumptions of the previous proposition, there is $C > 0$ such that*

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

for $n \geqslant N$.

The proof of this corollary makes use of a technical result which is attributed to [Chung 1954]. Its proof can be found in, *e.g.*, [Bravo *et al.* 2018]:

Lemma 5.5.8 (Lemma D.2 in [Bravo *et al.* 2018])**.** *Let (a_n) be a nonnegative sequence such that for all n large enough,*

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

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

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

Proof of Corollary 5.5.7. Let $C > 0$ be the constant provided in the statement of Proposition 5.5.6. Choose $r > 0$ small enough so that $2\beta r^{-1} - C > 0$. Replacing γ_n in this statement with the bounds on this step size provided by Assumption 5.4.1–(iv)), we get from the first inequality in Proposition 5.5.6

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

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

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

Let $k > 0$ be an integer. Telescoping the second inequality stated by Proposition 5.5.6 from $n+k$ back to n, we get

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

which implies that

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

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

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

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

Theorem 5.4.1: end of the proof. We now have all the elements to establish the identity (5.9), proving Theorem 5.4.1. For this, notice that,for every $N \geqslant 0$, by Proposition 5.5.4, y_n^N satisfies an equation of the form Equation (5.7). The assumption of Proposition 5.5.1 on the sequence $(\tilde{\eta}_n)$ are satisfied by Proposition 5.5.4 and the assumptions on the sequences $(\varrho_n), (\tilde{\varrho}_n)$ follow from Assumption 5.5.1 and Corollary 5.5.7.

Hence, applying Proposition 5.5.1, we obtain that $\mathbb{P}([y_n^N \to 0]) = 0$, for all $N \geqslant 0$. As previously explained, the latter implies that $\mathbb{P}([x_n \to 0]) = 0$.

To complete the proof of Theorem 5.4.1 it remains to prove Proposition 5.5.3, which is the purpose of the next section.

5.5.3 Proof of Proposition 5.5.3

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

$$\dot{\mathsf{x}}(t) \in -\partial f(\mathsf{x}(t))\,. \tag{5.14}$$

We say that an absolutely continuous curve $\mathsf{x} : \mathbb{R}_+ \to \mathbb{R}$ is a solution of the differential inclusion (DI) (5.14) starting at $x \in \mathbb{R}^d$ if $\mathsf{x}(0) = x$ and if for almost every $t \in \mathbb{R}_+$, the inclusion (5.14) is verified. We denote $\mathsf{S}_{-\partial f}(x)$ the set of these solutions.

The idea of the proof of Proposition 5.5.3 goes as follows. For each initial point $x \in B(0, r_0)$ with $r_0 > 0$ small enough, either all the trajectories of (5.14) issued from x leave $B(0, r_0)$ in a fixed time horizon, or $f(x) - f(P_M(x)) \geqslant \alpha \|x - P_M(x)\|$. This will be the content of the next lemma. Next, we use the well-known fact that the interpolated process constructed from our iterates (x_n) is a so-called *Asymptotic Pseudo Trajectory* (APT) of the DI (5.14), as formalized in [Benaïm *et al.* 2005]

(see also, *e.g.*, [Duchi & Ruan 2018, Schechtman 2021a]). The consequence is that on the event $[x_n \to 0]$, necessarily $f(x_n) - f(P_M(x_n)) \geqslant \alpha \|x_n - P_M(x_n)\|$ after a certain finite moment. To complete the proof, it remains to make use of the angle condition (iv)) of Definition 5.3.3.

Lemma 5.5.9. *Let $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous, path differentiable function. Let M be a C^2 active manifold for f such that $0 \in M$, $f(0) = 0$, and $\nabla_M f(0) = 0$. Then, there is $\alpha, T > 0$ and $r_0 > 0$ s.t. for every $\mathsf{x} \in \mathsf{S}_{-\partial F}(x)$, with $x \in B(0, r_0)$, either $\mathsf{x}([0,T]) \not\subset B(0, r_0)$ or $f(x) - f(P_M(x)) \geqslant \alpha \|x - P_M(x)\|$.*

Proof. Let $r > 0$ be such that $B(0,r) \subset U$, where U is the neighborhood from Definition 5.3.1. Since f is C^2 on $M \cap B(0,r)$ and $\nabla_M f(0) = 0$, there is some constant C s.t. we have $\sup_{x \in B(0,r)} \|\nabla_M f(P_M(x))\| \leqslant C \|P_M(x)\|$. Denote L the Lipschitz constant of f on $B(0,r)$ and let c_m be such that $\inf\{\|v\| : v \in \partial f(x), x \in B(0,r) \cap M^c\} \geqslant c_m$. Fix $r_0 \leqslant \min(\frac{c_m^2}{2LC}, r)$ and consider $x \in B(0, r_0)$ and $\mathsf{x} \in \mathsf{S}_{-\partial F}(x)$. Denote $t_1 = \inf\{t : \mathsf{x}(t) \in M \text{ or } \mathsf{x}(t) \notin B(0, r_0)\}$. Since f is path differentiable, we have:

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

Hence, if we choose T s.t. $c_m^2 T > 2 \sup_{x' \in B(0,r_0)} |f(x')|$, we have $t_1 \leqslant T$ and either $\mathsf{x}(t_1) \notin B(0,r)$ or $\mathsf{x}(t_1) \in M$. Assume that $\mathsf{x}(t_1) \in M$ and denote $\mathsf{y}(t) = P_M(\mathsf{x}(t))$ and $\mathsf{z}(t) = \mathsf{x}(t) - \mathsf{y}(t)$. Notice that for almost every $t \geqslant 0$, we have $\|\dot{\mathsf{y}}(t)\| = \left\| P_{T_{\mathsf{y}(t)}} \dot{\mathsf{x}}(t) \right\| \leqslant L$. Moreover, by path-differentiability of f we have:

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

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

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

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

every $t \geqslant 0$:

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

Therefore,

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

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

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

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

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

It is well known that under our assumptions, on the event $[x_n \to 0]$, X is an APT for the DI (5.14), as shown in [Benaïm *et al.* 2005, Duchi & Ruan 2018, Schechtman 2021a]. Namely, for every $T > 0$,

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

Consider α, T and r_0 from Lemma 5.5.9. On the event $[x_n \to 0]$ let $\mathsf{x}_n \in \mathsf{S}_{-\partial F}(x_n)$ be such that

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

Consider $r_1 \leqslant r_0$ such that $B(0, r_1) \subset U$, where U is the neighborhood associated to α by the angle condition. If for n large enough, $\mathsf{x}_n([0,T])$ remains in $B(0, r_1)$, then by Lemma 5.5.9 we have:

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

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

$$\langle v_n, x_n - P_M(x_n)\rangle \geqslant \beta \|x_n - P_M(x_n)\|\,. \tag{5.15}$$

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

5.6 Sketch of proof of Proposition 5.5.1

We recall that $\mathbb{E}_n[\cdot]$ denotes $\mathbb{E}[\cdot|\mathscr{F}_n]$. Denote d^- the dimension of Λ^-. Using the center-stable manifold theorem, the authors of [Brandière & Duflo 1996, Page 407–409] construct a sequence (w_n)[3] in $\mathbb{R}^{d^-}$ such that

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

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

[3] U_n^+ in their notations.

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

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

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

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

The only difference with [Brandière & Duflo 1996] is in the presence of (r'_{n+1}) and the point (iv)).

Using this representation, the avoidance of traps result follows from the following proposition. The only difference with [Brandière & Duflo 1996, Proposition 4] is, once again, in the presence of the sequence (r'_n).

Proposition 5.6.1 ([Brandière & Duflo 1996, Proposition 4])**.** *Let d be an integer, $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(\mathscr{F}_n)$ a filtration on it and (w_n) be a sequence in $\mathbb{R}^d$ verifying:*
$$w_{n+1} = w_n + \gamma_n H_n + \gamma_n(r_{n+1} + r'_{n+1} + e_{n+1})\,, \tag{5.16}$$
where the sequences $(w_n), (H_n), (r_n), (r'_n)(e_n)$ are adapted to $(\mathscr{F}_n)$ and (γ_n) is a sequence of positive stepsizes s.t. $\sum_{i=0}^{+\infty} \gamma_i = +\infty$ and $\sum_{i=0}^{+\infty} \gamma_i^2 < +\infty$. Assume that on an event $\Gamma \in \mathcal{A}$ we have the following.

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

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

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

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

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

Proof. In this proof C will denote some absolute constant that can change from line to line. The proof closely follows the one of [Brandière & Duflo 1996, Proposition 4]. As in [Brandière & Duflo 1996] it is sufficient to prove the proposition in the case where there $A, B, K > 0$ such that almost surely $\mathbb{E}_n[e_{n+1}] = 0$, $A \leqslant \mathbb{E}_n[\|e_{n+1}\|] \leqslant \mathbb{E}_n[\|e_{n+1}\|^2]^{1/2} \leqslant B$ and $\sum_{i=0}^{+\infty} \|r_{i+1}\|^2 \leqslant K$.
We can rewrite Equation (5.16) as:
$$w_{n+1} = w_n + \gamma_n H w_n + \gamma_n \Delta_n w_n + \gamma_n(e_{n+1} + r_{n+1} + r'_{n+1})\,,$$

where $\Delta_n = H_n - H$. Let Q be a positive definite symmetric matrix such that $QH + H^TQ = 2\mathcal{I}$, where $\mathcal{I} \in \mathbb{R}^{d\times d}$ is the identity matrix. Denote $U_n = (w_n^TQw_n)^{1/2}$. Following the same calculations as in [Brandière & Duflo 1996], we obtain that:

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

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

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

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

$$\sum_{i=n}^{M} \gamma_i \|w_i\| \leqslant CU_{M+1} + C\left\|\sum_{i=n}^{M} \gamma_i \frac{w_i^TQ(e_{i+1} + r_{i+1} + r'_{i+1})}{U_i}\right\|.$$

Hence,

$$\begin{aligned}\left|\sum_{i=n}^{M} \gamma_i \|w_i\|\right|^2 &\leqslant C\|U_{M+1}\|^2 + C\left\|\sum_{i=n}^{M} \gamma_i \frac{w_i^TQe_{i+1}}{U_i}\right\|^2 + C\left(\sum_{i=n}^{+\infty} \gamma_i^2\right)\left(\sum_{i=n}^{+\infty} \|r_{i+1}\|^2\right) + C\left|\sum_{i=n}^{+\infty} \gamma_i \left\|r'_{i+1}\right\|\right|^2 \\ &\leqslant C\|U_{M+1}\|^2 + C\sup_{M\geqslant p}\left\|\sum_{i=n}^{M} \gamma_i \frac{w_i^TQe_{i+1}}{U_i}\right\|^2 + C\chi_n + C\left|\sum_{i=n}^{+\infty} \gamma_i \left\|r'_{i+1}\right\|\right|^2,\end{aligned}$$

where we used the fact that $\frac{\|w_n^TQ\|}{U_n}$ is bounded. On Υ_p we have that $\mathbb{E}[\|U_{M+1}\|^2] \to 0$. The sequence $(\sum_{i=n}^{M} \gamma_i \frac{w_i^TQe_{i+1}}{U_i})_{M\geqslant n}$ is a square summable martingale difference sequence. Therefore, by Doob's maximal inequality:

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

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

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

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

$$\mathbb{E}\left[\mathbb{1}_{\Upsilon_p}\left|\sum_{i=n}^{+\infty} \gamma_i \|w_i\|\right|^2\right] \leqslant C\chi_n. \tag{5.17}$$

On the other hand, following the calculations of [Brandière & Duflo 1996], on Υ_p we have:

$$-w_p = \sum_{i=p}^{+\infty} \left(R_i^1 + \gamma_i(e_{i+1} + r_{i+1} + r'_{i+1})\right) , \tag{5.18}$$

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

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

The idea of the remaining part of the proof is to apply [Brandière & Duflo 1996, Theorem A] to obtain that the left hand side of Equation 5.18 can be $\mathscr{F}_p$-measurable only with probability 0. The latter will imply $\mathbb{P}(\Upsilon_p) = 0$ and since $\Upsilon = \bigcup_{p\in\mathbb{N}} \Upsilon_p$, the proof will be finished. As in the proof [Brandière & Duflo 1996], one of the assumptions of [Brandière & Duflo 1996, Theorem A], to obtain the remaining part it suffices to have:

$$\mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \sum_{i=n}^{+\infty} \left\|R_i^1 + \gamma_i r'_{i+1}\right\|\right] = o(\sqrt{\chi_n}) \,, \tag{5.19}$$

where the difference with the proof of [Brandière & Duflo 1996, Proposition 4] is in the presence of the term r'_{i+1}. To prove Equation (5.19) we write down:

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

By Inequality (5.17) we have:

$$\begin{aligned} \mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \sup_{i\geqslant n} \|\Delta_i\| \sum_{i=n}^{+\infty} \gamma_i \|w_i\|\right] &\leqslant C\mathbb{E}[\mathbb{1}_{\Upsilon_p} \sup_{i\geqslant n} \|\Delta_i\|^2]^{1/2} \mathbb{E}\left[\left|\sum_{i=n}^{+\infty} \gamma_i \|w_i\|\right|^2\right]^{1/2} \\ &\leqslant C\mathbb{E}[\mathbb{1}_{\Upsilon_p} \sup_{i\geqslant n} \|\Delta_i\|^2]^{1/2} \sqrt{\chi_n} \\ &\leqslant o(\chi_n) \,. \end{aligned} \tag{5.20}$$

As noticed in [Brandière & Duflo 1996] we have $\sum_{i=1}^{+\infty} \left\|B_{i-1}^{-1} - B_i^{-1}\right\| < +\infty$. Therefore,

$$\mathbb{E}\left[\mathbb{1}_{\Upsilon_p} \|S_i\|\right] \sum_{i=n}^{+\infty} \left\|B_{i-1}^{-1} - B_i^{-1}\right\| \leqslant C\sqrt{\chi_n} \sum_{i=n}^{+\infty} \left\|B_{i-1}^{-1} - B_i^{-1}\right\| = o(\sqrt{\chi_n}) \,, \tag{5.21}$$

and by assumptions

$$\mathbb{E}\left[\mathbb{1}_{\Upsilon_p}\sum_{i=n}^{+\infty}\gamma_i\left\|r'_{i+1}\right\|\right] \leqslant C\chi_n = o(\sqrt{\chi_n})\,. \tag{5.22}$$

Combining (5.20), (5.21) and (5.22) we obtain Equation (5.19). Hence, we can apply [Brandière & Duflo 1996, Theorem A] to obtain that $\mathbb{P}(\Upsilon_p) = 0$. Since $\Upsilon = \bigcup_{p\in\mathbb{N}}\Upsilon_p$, the proof is finished. □

Chapter 6

Stochastic subgradient descent on a generic definable function converges to a minimizer

6.1 Introduction

Design and analysis of optimization algorithms are usually relying on some kind of optimality conditions. Canonical examples of such conditions are the second order sufficiency in nonlinear programming [Nocedal & Wright 2006] and strict complementarity in semidefinite programming [Alizadeh *et al.* 1995]. While a specific optimization problem might not verify such conditions, a standard way to justify their ubiquity is that they are in some mathematical sense generic. Formally, given a class of optimization problems (Q_u) that is parametrized by a set of vectors $u \in \mathbb{R}^d$, we say that a condition is *generic* within this class if it is satisfied for the problem Q_u, for almost every $u \in \mathbb{R}^d$. Analysis of such a kind dates back at least to the works of Simon and Saigal [Simon & Saigal 1973] and Spingarn and Rockafellar [Spingarn & Rockafellar 1979]. In the latter (Q_u) are the linear perturbations of some specific nonlinear programming problem Q and it is showed that for almost every $u \in \mathbb{R}^d$, the second order sufficiency conditions are indeed necessary in Q_u.

In the present chapter, in the spirit of [Spingarn & Rockafellar 1979], given a locally Lipschitz continuous function $f : \mathbb{R}^d \to \mathbb{R}$ that is not necessarily smooth nor convex, we analyze the following class of problems:

$$\min_{x \in \mathbb{R}^d} f_u(x)\,, \tag{Q_u}$$

where for $u, x \in \mathbb{R}^d$, we denote $f_u(x) := f(x) - \langle u, x \rangle$. In this case, the first order necessary condition for x to be a solution to (Q_u) is that $0 \in \partial f_u(x)$, where $\partial f_u(x)$ is the set of *Clarke subgradients* of f_u at x. Hence, we are interested in the generic properties of the following class of sets:

$$\{x \in \mathbb{R}^d : 0 \in \partial f_u(x)\}\,, \tag{$\mathcal{Z}_u$}$$

where for each $u \in \mathbb{R}^d$, $\mathcal{Z}_u$ (respectively $\mathcal{Z}$) denotes the set of *Clarke critical points* of f_u (respectively f).

We are specifically interested in the question of genericity from the perspective of the simplest first order algorithm - the stochastic subgradient descent (SGD):

$$x_{n+1} \in x_n - \gamma_n \partial f(x_n) + \gamma_n \eta_{n+1}\,, \tag{6.1}$$

where (γ_n) is a sequence of positive stepsizes and (η_n) is some perturbation sequence which presence is typically due to a partial knowledge of ∂f by the designer. It is known (see [Davis & Drusvyatskiy 2021, Majewski *et al.* 2018]) that, under mild conditions on the sequence (η_n) and the function f, the iterates of the SGD converge to $\mathcal{Z}$. While the set $\mathcal{Z}$ contains local minima it also contains all kinds of spurious points (e.g. local maxima and saddle points) convergence to which might be highly undesirable. We are thus interested in describing a generic set of conditions that ensures the convergence of the SGD to a local minimum.

The first important remark that we should make here is that, in the pursuit of this path, we must distinguish between the conditions that characterize a generic Clarke critical point, which are inherent to the class of problems that we analyze, and the conditions on the perturbation sequence (η_n), which depend on the practical situation and, to some degree, can be imposed by the designer.

This observation is consistent with the existing analysis of Equation (6.1) in the smooth setting. In this case, for almost every $u \in \mathbb{R}^d$ (henceforth *generic vector* $u \in \mathbb{R}^d$), every critical point of f_u is either a local minimum or a saddle point (i.e. the Hessian of f_u at this point has at least one negative eigenvalue). The nonconvergence of the SGD to a saddle point (and hence its convergence to a local minimum on a generic smooth function) was established in [Pemantle 1990, Brandière & Duflo 1996] under an assumption that, more or less, requires the lower boundedness of the (conditional) covariance of (η_n). When this type of assumption is not satisfied, as it happens for e.g. the deterministic gradient descent ($\eta_n \equiv 0$), it can indeed be guaranteed by the designer by adding a small perturbation, with lower bounded covariance, at every step.

Following this discussion, the present chapter consists of two, largely independent, parts.

- The first part is devoted to the analysis of the generic properties of Clarke critical points. Our main result, Theorem 6.2.5, proposes a classification of the types of points that might appear in $(\mathcal{Z}_u)$ for a non Lebesgue-null set of vectors $u \in \mathbb{R}^d$. An emphasis is put on the conditions that allow the analysis of the SGD in a neighborhood of a generic critical point.

- The second part of this chapter is devoted to the analysis of the SGD in a neighborhood of a *generic trap*, i.e. a Clarke critical point that might appear in $(\mathcal{Z}_u)$ for a non Lebesgue-null set of vectors $u \in \mathbb{R}^d$ without being a local minimum. Specifically, we will present a set of conditions on the sequence (η_n) that ensure that the iterates of the SGD will avoid a generic trap.

6.1.1 Generic critical points

In our analysis of genericity we restrict ourselves to the case where f, the function of interest, is *definable in an o-minimal structure* (henceforth definable). Formally defined in Section 2.4, the class of such functions encompasses the vast majority of functions encountered in optimization. It includes every semialgebraic function, the

exponential, the logarithm as well as any of their compositions. While a definable function might be nonsmooth the nonsmoothness here appears in a very structured manner. For instance, a domain of a definable function can be partitioned into a set of manifolds called *stratas* such that on each of these stratas the function is differentiable. Starting from the seminal work of [Bolte *et al.* 2007], this implicit smooth structure has allowed a thorough analysis of optimization algorithms in a definable setting (see e.g. [Attouch *et al.* 2011, Bolte *et al.* 2009, Davis *et al.* 2020, Bolte *et al.* 2020a]).

Analysis of the generic properties of $(\mathcal{Z}_u)$ when f is definable goes back to the work of [Bolte *et al.* 2011, Drusvyatskiy *et al.* 2016] and more recently to [Davis & Drusvyatskiy 2021, Bianchi *et al.* 2021b]. The central notion in all of these works is the notion of an active manifold. Informally, M is an active manifold for a Clarke critical point $x^* \in M$ if f varies smoothly on M and sharply outside of it. The importance of this notion lies in the fact, proved in [Drusvyatskiy *et al.* 2016][1], that if f is definable, then for a generic vector $u \in \mathbb{R}^d$, the number of Clarke critical points of f_u is finite and every one of them lies on an active manifold. Recently, following the ideas of [Drusvyatskiy *et al.* 2016, Drusvyatskiy & Lewis 2014], Davis and Drusvyatskiy [Davis & Drusvyatskiy 2021] have introduced the notion of an active strict saddle: a Clarke critical point x^* of a function f, lying on an active manifold M, such that f_M, the restriction of f to M, admits a second order negative curvature at x^*. They have shown that if f is weakly convex, then for a generic vector $u \in \mathbb{R}^d$, every point in $(\mathcal{Z}_u)$ is either a local minimum or an active strict saddle. Hence, in the weakly convex case the following two examples are typical.

Local minimum. In order to be a local minimum a critical point lying on an active manifold M must be a local minimum of f_M. As an example consider $f_1 : \mathbb{R}^2 \to \mathbb{R}$ be defined as $f_1(y, z) = y^2 + |z|$. Then $x^* = (0, 0)$ is a Clarke critical point of f_1 and $M_1 = \mathbb{R} \times \{0\}$ is the corresponding active manifold.

Active strict saddle. Consider $f_2 : \mathbb{R}^2 \to \mathbb{R}$ defined as $f_2(y, z) = -y^2 + |z|$. Then $x^* = (0, 0)$ is a Clarke critical point of f_2 and $M_2 = \mathbb{R} \times \{0\}$ is the corresponding active manifold. Observe that in this case x^* is not a local minimum of f_2 due to the fact that it is not a local minimum of $f_{2|M}$.

The reason behind such a simple classification lies in the fact that, from a minimization perspective, the behavior of a weakly convex function in a neighborhood of an active manifold M is dictated by its behavior on M. Examples presented in Section 6.2 show that such a result does not hold true as soon as the weak convexity assumption fails. Indeed, in full generality, it is clear that if a critical point x^* of a function f lies on an active manifold M, then the local shape of f (and hence the type of x^*) depends both on the behavior of f_M and on the directions of the subgradients of f outside of M. Therefore, to obtain a proper classification in this general case, both of these informations must be taken into account.

This discussion motivates the introduction of a third type of a generic Clarke critical point: a *sharply repulsive critical point*. Its formal definition is given in

[1] Although this result is explicitly stated only for the limiting subgradient.

Section 6.2 but informally it is a Clarke critical point x^* of a function f, lying on an active manifold M, such that x^* is a local minimum of f_M, but there is a region close to M such that the subgradients of f point towards M. The following example is typical.

Sharply repulsive critical point. Consider $f_3 : \mathbb{R}^2 \to \mathbb{R}$ defined as $f_3(y, z) = y^2 - |z|$. Then $x^* = (0, 0)$ is a Clarke critical point of f_3 and $M_3 = \mathbb{R} \times \{0\}$ is the corresponding active manifold. In this example x^* is indeed a local minimum of $f_{3|M_3}$ but every subgradient outside of M_3 is directed towards the active manifold.

Our first result, Theorem 6.2.5, shows that for a definable, locally Lipschitz continuous function f and for a generic vector $u \in \mathbb{R}^d$, every critical point of f_u is either a local minimum, an active strict saddle or a sharply repulsive critical point. Furthermore, we establish that the corresponding active manifolds are satisfying the Verdier and the angle conditions, introduced in [Bianchi *et al.* 2021b]. Importance of these conditions in the analysis of the SGD are discussed in the next section.

6.1.2 The role of the Verdier and the angle conditions

Analyzing the iterates of the SGD in a neighborhood of an active manifold M, it might be helpful to decompose ∂f into components that are respectively tangent and normal to M. This technique of proof, developed in [Bianchi *et al.* 2021b], is natural when we think about the SGD applied to the previously presented functions f_1, f_2, f_3. In this case we can decompose the iterates (x_n) into a sum of two sequence $(y_n), (z_n)$ and notice that the sequence (y_n) (respectively (z_n)) represents the iterates of the SGD applied to the function $y \mapsto \pm y^2$ (respectively $z \mapsto \pm|z|$), where the respective signs should be obvious from the considered examples. Observe that in all of these cases (y_n) are the SGD iterates of a smooth function, while (z_n) are either converging or diverging from 0 in a very fast manner.

To formalize this type of behavior authors of [Bianchi *et al.* 2021b] have introduced two additional assumptions on the active manifold M. The first one, the Verdier condition, states that for x close to M:

$$\forall v \in \partial f(x), \quad v_M \approx \nabla_M f(P_M(x)) + O(\mathrm{dist}(x, M)) \,,$$

where $P_M(x)$ is the projection of x onto M, $\nabla_M f$ is the "Riemannian gradient" of f_M and v_M is the projection of v along the tangent space of M (see Section 6.2 for a precise statement). A consequence of this condition is that, writing down $(y_n) = (P_M(x_n))$, we obtain:

$$y_{n+1} \approx y_n - \gamma_n \nabla_M f(y_n) + \gamma_n \eta_{n+1}^M + \gamma_n O(\mathrm{dist}(x_n, M)) + O(\gamma_n^2) \,, \tag{6.2}$$

where η_{n+1}^M is the projection of η_{n+1} on the tangent space of M at y_n. That is to say, up to a residual error term, (y_n) follows an SGD dynamic on the (smooth) function f_M.

To motivate the angle condition a following observation was made in [Bianchi *et al.* 2021b]. Let x^* be a Clarke critical point of f lying on an active manifold M. Then, on the

event $[x_n \to x^*]$, for n large enough we have:

$$f(x_n) - f(P_M(x_n)) \gtrsim \|x_n - P_M(x_n)\| \,. \tag{6.3}$$

The angle condition then states that close to M we have:

$$f(x)-f(P_M(x)) \gtrsim \|x - P_M(x)\| \implies \langle v, x-P_M(x)\rangle \gtrsim \|x - P_M(x)\| \,, \quad \forall v \in \partial f(x) \,. \tag{6.4}$$

Combining (6.3) with (6.4), we obtain that for n large enough the angle between the set $\partial f(x_n)$ and the normal direction to M is lower bounded. The latter allows to control the residual term in Equation (6.2).

Both of these conditions provide a way to analyze the SGD in a neighborhood of an active manifold by decomposing the iterates (x_n) into a sum of two sequences: $(y_n) = (P_M(x_n))$ and $(z_n) = (x_n - y_n)$. The angle condition ensures the fact that $\mathrm{dist}(x_n, M) = \|z_n\| \to 0$ (and hence $x_n \to M$) fast enough. Combining this fact with the Verdier condition, this implies that (y_n), up to a residual term, follows an SGD dynamic on the smooth function f_M.

In Chapter 5 this technique of proof was used to show that, under assumptions on (η_n) similar to [Brandière & Duflo 1996], the SGD avoid active strict saddles with probability one. In this chapter we illustrate the interest of the angle condition in the analysis of the SGD in a neighborhood of a sharply repulsive critical point.

6.1.3 Avoidance of generic traps

The final part of this chapter is devoted to the analysis of the SGD in a neighborhood of a generic trap. Since the question of the nonconvergence to an active strict saddle was tackled in [Bianchi *et al.* 2021b] we focus in this part on the question of nonconvergence of the SGD to $x^* \in M$ a sharply repulsive critical point.

Our first result, which requires only very mild, moment assumptions on the sequence (η_n), is that on the event $[x_n \to x^*]$, where x^* is a sharply repulsive critical point, we have that, for n large enough,

$$f(x_n) \geqslant f(x^*) \,,$$

While the proof of this statement readily follows from Chapter 5 such a result is interesting. Indeed, it implies that while the iterates (x_n) may in theory converge to x^* this happens only if the SGD fails to explore the repulsive region near x^*. In some sense, the algorithm perceive the function f as if x^* was indeed its local minimum.

In a second time, we show that a density-like assumption on (η_n) forces the SGD to visit the repulsive region near M and will imply the nonconvergence of the SGD to a sharply repulsive critical point.

The final Section 6.3.3 shows that while such a density-like assumption on (η_n) might not hold, in a standard stochastic approximation model, a way to ensure it is to add a small perturbation (e.g. a nondegenerate Gaussian) at each iteration of (6.1). This fact, combined with the results of [Bianchi *et al.* 2021b] on the avoidance

of active strict saddles, provides a practical way to avoid generic traps of a definable function, and, therefore, ensure the convergence of the SGD to a local minimum.

6.1.4 Previous avoidance of traps results and contributions

We finish the introduction by a discussion on diverse avoidance of traps results previously stated in the literature. In the smooth setting avoidance of saddle points by the SGD (and more generally by a Robbins-Monroe algorithm) was first addressed by Brandière and Duflo [Brandière & Duflo 1996] and Pemantle [Pemantle 1990]. Under a condition that, more or less, requires a lower boundedness of the covariance of (η_n) they have established that the iterates of the SGD avoid saddle points with probability one. Later on, these results were extended in many ways, we mention here the nonconvergence to periodic hyperbolic sets by [Benaïm 1999], the nonautonomous setting [Barakat *et al.* 2021] and many others [Mertikopoulos *et al.* 2020b, Gadat & Gavra 2020]. From another perspective, the authors of [Lee *et al.* 2016] have established the nonconvergence to a saddle point of the deterministic gradient descent under a random initialization. Davis and Drusvyatskiy, in [Davis & Drusvyatskiy 2021], have presented a first nonconvergence result in the nonsmooth setting. They have introduced the concept of an active strict saddle and similarly to [Lee *et al.* 2016] have established that proximal methods avoid active strict saddle under a random initialization. As mentioned earlier, in our previous work [Bianchi *et al.* 2021b], under the same conditions on the perturbation sequence as in [Brandière & Duflo 1996], we have established the nonconvergence of the SGD to an active strict saddle lying on an active manifold that satisfies a Verdier and an angle conditions.

Finally, shortly after the publication [Bianchi *et al.* 2021b] and just before the submission of [Schechtman 2021b], on which is based the current chapter, a concurrent work [Davis *et al.* 2021] has appeared. The latter, sharing a lot of similarities with [Bianchi *et al.* 2021b], analyzes the SGD (and its proximal versions) in a neighborhood of an active manifold. An avoidance of active strict saddles result was obtained as well as (local) rates of convergence and asymptotic normality of the iterates was established. These results support our claim of the importance of the Verdier and the angle conditions. A major difference with this chapter is that their proximal aiming condition assume (close to the active manifold) the left hand side of formula (6.4). Such an assumption rules out functions with downward cusps such as $(y, z) \mapsto \pm y^2 - |z|$, which are treated in [Bianchi *et al.* 2021b] and in the present chapter. As a consequence, the question of genericity in [Davis *et al.* 2021] is addressed only for the class of Clarke regular functions in which sharply repulsive critical points do not exist. In particular, we believe that convergence rates of a similar kind could be obtained upon replacing the proximal aiming condition of [Davis *et al.* 2021] by ours angle condition.

Paper organization. Section 6.2 deals with the generic properties of Clarke critical points. In Section 6.3 we state an avoidance of traps result and discuss the convergence of the SGD to minimizers. Section 6.4 is devoted to proofs.

Notations. For $x \in \mathbb{R}^d$ and $r > 0$, we denote $B(x, r)$ the open ball centered

of radius r centered at x. Given a set $S \subset \mathbb{R}^d$, $\overline{S}$ will denote its closure and S^c its complementary. The distance to S will be denoted as $\operatorname{dist}(\cdot, S)$. We say that $V \subset S$ is *open in* S if there is an open set $U \subset \mathbb{R}^d$ such that $U \cap S = V$. We say that $\mathsf{H} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ is *a set-valued mapping* if for each $x \in \mathbb{R}^d$, $\mathsf{H}(x) \subset \mathbb{R}^d$, we denote $\operatorname{Graph}(\mathsf{H}) = \{(x, y) : x \in \operatorname{dom}(\mathsf{H}), y \in \mathsf{H}(x)\}$ its graph. We say that a property holds *locally around* x if this property holds on U an open neighborhood of x. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is weakly convex if there is $\rho > 0$ such that $f(\cdot) + \rho \|\cdot\|^2$ is convex.

Given n random variables $X_1, \ldots, X_n$ on some probability space, we denote $\sigma(X_1, \ldots, X_n)$ the sigma algebra generated by them. The set of borelians of $\mathbb{R}^d$ will de denoted as $\mathcal{B}(\mathbb{R}^d)$. Given some probability space on which we have $(\mathscr{F}_n)$ a filtration and X a random variable, we will denote $\mathbb{E}_n[X] = \mathbb{E}[X|\mathscr{F}_n]$. Given a matrix $B \in \mathbb{R}^{m \times n}$, we will denote B^T its transpose.

6.2 Generic critical points

Theorem 6.2.5 of this section classifies generic critical points of a locally Lipschitz continuous, definable function. A reader who is more interested in our avoidance of traps result can take Definitions 6.2.1–6.2.4 as granted and jump to Section 6.3. We recall that for a function $f : \mathbb{R}^d \to \mathbb{R}$ and $u, x \in \mathbb{R}^d$, we denote $f_u(x) = f(x) - \langle u, x \rangle$ and that, given a manifold M, we denote f_M the restriction of f to M.

To motivate our presentation consider first the case where $f : \mathbb{R}^d \to \mathbb{R}$ is C^2. Applying Sard's theorem to the function $x \mapsto \nabla f(x)$, we obtain that the set

$$\{u \in \mathbb{R}^d : \exists x \in \mathbb{R}^d, \nabla f(x) = u \text{ and the Hessian of } f \text{ at } x \text{ is degenerate } \}$$

is Lebesgue-null. Hence, for almost every $u \in \mathbb{R}^d$, the critical points of $f_u(x) = f(x) - \langle u, x \rangle$ are nondegenerate. This result can be extended to functions defined on a submanifold.

Proposition 6.2.1 ([Victor 1974, Chapter 7, §7]). *Consider $M \subset \mathbb{R}^d$ a C^2 submanifold of dimension greater than 0. Let $f : M \to \mathbb{R}$ be C^2. Then for almost every $u \in \mathbb{R}^d$, the critical points of $f_{u|M}$ are nondegenerate.*

Remark 21. *A function $f : M \to \mathbb{R}^d$ such that every of its critical points is nondegenerate is called a* Morse *function. Proposition 6.2.1 shows that Morse functions always exist. This result can be strengthened to the fact that the set of Morse functions is open and dense in the Whitney C^2 topology (cf. e.g. [Audin et al. 2014]). In that sense almost every smooth function on M is Morse.*

6.2.1 Active manifolds

The central notion of this work is the notion of an active manifold. It was introduced by Lewis in [Lewis 2002] and was thoroughly studied in [Drusvyatskiy & Lewis 2014, Drusvyatskiy & Lewis 2012].

Definition 6.2.1 (Active manifold)**.** *Consider $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous function and x^* such that $0 \in \partial f(x^*)$. For $p \geqslant 1$, we say that M is a C^p* active manifold around x^* *if there is a neighborhood U of x^* such that the following holds.*

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

ii) ***Sharpness.***

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

Note that in the preceding definition M can be the whole space $\mathbb{R}^d$. As a consequence any C^p function $f : \mathbb{R}^d \to \mathbb{R}$ admits $\mathbb{R}^d$ as an active manifold on any of its critical point.

The following conditions on an active manifold were introduced in [Bianchi *et al.* 2021b]. We recall that P_M denotes the projection onto M and that by Lemma 2.3.2 it is well defined in the neighborhood of M.

Definition 6.2.2 (Verdier and angle conditions)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function. Let M be a C^2 active manifold around a critical point x^*. We say that M satisfies a Verdier and an angle conditions if the following holds.*

- ***Verdier condition.*** *There is U a neighborhood of x^* and $C \geqslant 0$ such that for $y \in M \cap U$ and $x \in U$, we have:*

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

  where $P_{T_y M}$ is the orthogonal projection onto $T_y M$.

- ***Angle condition.*** *For every $\alpha > 0$, there is $\beta > 0$ and U_α a neighborhood of x^* such that for all $x \in U_\alpha$, we have:*

$$f(x) - f(P_M(x)) \geqslant \alpha \|x - P_M(x)\| \implies \langle v, x - P_M(x)\rangle \geqslant \beta \|x - P_M(x)\|\,, \quad \forall v \in \partial f(x)\,. \tag{6.5}$$

In practice an active manifold is an element of the stratification presented in Theorem 5.2.1. Hence, the Verdier condition is just a transcription of Inequality (5.3) in this setting. The importance of the angle condition can be grasped from the following observation made in [Bianchi *et al.* 2021b].

Proposition 6.2.2 ([Bianchi *et al.* 2021b, Lemma 7])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be path-differentiable and assume that there is $x^* \in \mathbb{R}^d$ a Clarke critical point lying on a C^2 active manifold M. There is $r, T, \alpha > 0$ such that for any $x \in B(x^*, r)$ and $\mathsf{x} \in \mathsf{S}_{-\partial f}(x)$, either $\mathsf{x}([0,T]) \not\subset B(x^*, r)$ or*

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

The preceding lemma shows that the set

$$\{x \in B(0,r) : f(x) < f(P_M(x)) + \alpha \|x - P_M(x)\|\}, \tag{6.6}$$

where α, r are the one of Proposition 6.2.2, can be viewed as a repulsive region for the subgradient flow. The angle condition ensures the fact that as soon as we are not in this repulsive region the negative subgradients of f are directed towards M. This information will help us to show that the iterates of the SGD converge to M fast enough.

6.2.2 Generic traps

This work focuses on the two following types of Clarke critical points.

Definition 6.2.3 (Active strict saddle [Davis & Drusvyatskiy 2021])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function. We say that a point $x^* \in \mathbb{R}^d$ is an* active strict saddle *if there is M a C^2 active manifold around x^*, of dimension greater than 0, and x^* is a saddle point for the function f_M.*

Definition 6.2.4 (Sharply repulsive critical point)**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function. We say that a Clarke critical point x^* is sharply repulsive if it lies on an active manifold M such that x^* is a local minimum of f_M and $0 \in \partial f(x^*) \backslash \partial_L f(x^*)$.*

The reason behind the chosen denomination of Definition 6.2.4 comes from the following proposition. It shows that the active manifold of a sharply repulsive critical point is always neighbored by a large repulsive region of the form (6.6).

Proposition 6.2.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous function and $x^* \in \mathbb{R}^d$ such that $0 \in \partial f(x^*) \backslash \partial_L f(x^*)$. There is $C > 0$ such that for all $\varepsilon > 0$, there is $x \in B(x^*, \varepsilon)$ such that:*

$$f(x) \leqslant f(x^*) - C \|x^* - x\| .$$

If, moreover, x^* is a sharply repulsive critical point, lying on a C^2 active manifold M, then there is $\varepsilon > 0$ such that for all $y \in B(x^*, \varepsilon) \cap M$ and for all $\varepsilon_y > 0$, there is $x \in B(y, \varepsilon_y)$ such that we have:

$$f(x) < f(P_M(x)) .$$

If f is weakly convex, then $\partial_L f = \partial f$. Hence, such a function does not have sharply repulsive critical points. As the following proposition shows, in this case, the notion of an active strict saddle is generic. In its initial version this proposition follows from the work [Drusvyatskiy *et al.* 2016] and was proved in [Davis & Drusvyatskiy 2021]. Statements concerning the Verdier and the angle conditions were proved in [Bianchi *et al.* 2021b].

Proposition 6.2.4 ([Davis & Drusvyatskiy 2021, Theorem 2.9] and [Bianchi *et al.* 2021b, Theorem 2])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be definable and weakly convex. There is $N \in \mathbb{N}$ such that for almost every $u \in \mathbb{R}^d$, the set $(\mathcal{Z}_u)$ is of cardinality less than N. Moreover, every such a point lies on an active manifold satisfying the Verdier and the angle conditions and is either a local minimum or an active strict saddle.*

The reason behind such a simple classification in Proposition 6.2.4 comes from the fact that under weak convexity a local minimum of f_M is also a local minimum of the unrestricted function f. As the following example shows this is no longer true without the weak convexity assumption.

Example 6.2.1. *Consider $f_4 : \mathbb{R}^2 \to \mathbb{R}$ defined as $f_4(y,z) = -|y| + |z|$. For $u \in B(0,1)$, the point $(0,0)$ is a sharply repulsive critical point lying on the active manifold $M_4 = \{(0,0)\}$.*

We are now ready to state the main result of this section which is a generalization of Proposition 6.2.4 to the non weakly convex case.

Theorem 6.2.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous, definable function and $p \geqslant 2$. There is $N \in \mathbb{N}$ such that for almost every $u \in \mathbb{R}^d$ the set $(\mathcal{Z}_u)$ is of cardinality less than N and every point $x_u^* \in \mathcal{Z}_u$ lies on a C^p active manifold M such that the following holds.*

i) The manifold M satisfies the Verdier and the angle conditions.

ii) If the dimension of M is greater than 0, then x_u^* is a nondegenerate critical point for the function $f_u : M \to \mathbb{R}$.

iii) The point x_u^* is either a local minimum, an active strict saddle or a sharply repulsive critical point of f_u.

6.3 Avoidance of generic traps

6.3.1 Escaping a sharply repulsive critical point

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function. Let $G : \mathbb{R}^d \to \mathbb{R}^d$ be a measurable function such that for all $x \in \mathbb{R}^d$, $G(x) \in \partial f(x)$, such a "measurable selection" always exists (cf. [Rockafellar & Wets 1998]).

On a probability space $(\Omega, \mathcal{A}, \mathbb{P})$, consider a random variable $x_0 \in \mathbb{R}^d$ and a random sequence $(\eta_n) \in (\mathbb{R}^d)^{\mathbb{N}}$. Define the iterates:

$$x_{n+1} = x_n - \gamma_n G(x_n) + \gamma_n \eta_{n+1} = x_n - \gamma_n v_n + \gamma_n \eta_{n+1}\,, \tag{6.7}$$

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

Assumption 6.3.1.

i) The function f is path-differentiable.

ii) The sequence (η_n) is adapted to $(\mathscr{F}_n)$ and x_0 is $\mathscr{F}_0$-measurable.

iii) The sequence (γ_n) is such that $\sum_{i=0}^{+\infty} \gamma_i = +\infty$, $\sum_{i=0}^{+\infty} \gamma_i^2 < +\infty$ and there is $c_1, c_2 > 0$ such that:

$$c_1 \leqslant \frac{\gamma_n^2}{\gamma_{n+1}^2} \leqslant 1 + c_2 \gamma_n\,.$$

Assumption 6.3.1 is a standard assumption in the field of stochastic approximation. We notice that the point (iii)) is satisfied by the sequences of the form $\gamma_n = \frac{1}{n^\epsilon}$ for $\epsilon \in (1/2, 1]$.

Recall that $\mathbb{E}_n[\cdot]$ denotes $\mathbb{E}[\cdot|\mathscr{F}_n]$.

Assumption 6.3.2. *The sequence (η_n) is such that $\mathbb{E}_n[\eta_{n+1}] = 0$ and for every $C > 0$, there is $K(C) > 0$ such that we have:*

$$\sup_{n\in\mathbb{N}} \mathbb{E}_n[\|\eta_{n+1}\|^2]\mathbb{1}_{\|x_n\|\leqslant C} \leqslant K(C)\,.$$

Assumption 6.3.3. *The point x^* is a sharply repulsive critical point of f such that the corresponding active manifold M is C^2.*

Our first result concerning the behavior of the SGD in the neighborhood of a sharply repulsive critical point is the following proposition. Its proof is provided in Section 6.4.2.

Proposition 6.3.1. *Let Assumptions 6.3.1–6.3.2 hold. Assume that a point $x^* \in \mathcal{Z}$ is lying on a C^2 active manifold. There is $\alpha > 0$ such that, almost surely on the event $[x_n \to x^*]$, there is $n_0 \in \mathbb{N}$ such that for $n \geqslant n_0$, we have:*

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

As a consequence, under Assumptions 6.3.1–6.3.3, for n large enough,

$$f(x_n) \geqslant f(x^*)\,.$$

A consequence of the preceding theorem is the fact that while the iterates of the SGD may in theory converge to a sharply repulsive critical point, this happens only if the sequence (x_n) fails to explore the repulsive region of the form (6.6) neighboring the active manifold. Without additional assumptions on the perturbation sequence (η_n), the following example shows that such behavior is easy to construct. Recall that $f_4 : \mathbb{R}^2 \to \mathbb{R}$ is defined as $f_4(y, z) = -|y| + |z|$.

Example 6.3.1. *Consider $z \in \mathbb{R}$ and let $x_0 = (y_0, z_0) = (0, z)$. For $n \in \mathbb{N}$, define $\eta_n = 0$ and $v_n = (0, \frac{z_n}{|z_n|})\mathbb{1}_{\|z_n\|>0}$. Then the sequence (x_n) defined by Equation (6.7) represents the iterates of the SGD applied to f_4 and $x_n \to (0, 0)$.*

The next two assumptions will force the SGD to explore the repulsive region around x^*.

Assumption 6.3.4. *For every $C > 0$, there is a continuous, positive function $h_C : \mathbb{R}^d \to \mathbb{R}$ such that for every $n \in \mathbb{N}$ and any measurable function $\psi : \mathbb{R}^d \to \mathbb{R}^d$, if $\|x_n\| \leqslant C$, then:*

$$\forall \delta > 0, \quad \mathbb{P}(\eta_{n+1} \in B(\psi(x_n), \delta)|\mathscr{F}_n) \geqslant \int \mathbb{1}_{B(\psi(x_n),\delta)}(y) h_C(y)\,\mathrm{d}y\,.$$

Assumption 6.3.5. *The active manifold from Assumption 6.3.3 satisfies an angle condition.*

Assumption 6.3.4 describes a density-like behavior of the conditional law of (η_{n+1}). Indeed, it is satisfied if the conditional laws of (η_{n+1}) are identically distributed according to some law that has a density relatively to Lebesgue which is positive at every point. As we show in Section 6.3.3 to enforce this assumption it is sufficient to add a "nondegenerate" perturbation at each step.

On the other hand, as the following proposition shows, Assumption 6.3.5 allows to control the speed of convergence of (x_n) towards a sharply repulsive critical point.

Proposition 6.3.2. *Let Assumptions 6.3.1–6.3.3 and 6.3.5 hold. There is $\kappa > 0$ such that on $[x_n \to x^*]$ the event*

$$[\operatorname{dist}(x_n, M) \leqslant \kappa\gamma_n]$$

occurs infinitely often.

Finally, with this result in hand we have that a sharply repulsive critical point is avoided by the SGD with probability one.

Theorem 6.3.3. *Let Assumptions 6.3.1–6.3.5 hold. Then, $\mathbb{P}([x_n \to x^*]) = 0$.*

The proof of Theorem 6.3.3 is slightly technical but conceptually it can be described as follows. By Proposition 6.3.2 the iterates are infinitely often located at a distance less than $\kappa\gamma_n$ from the active manifold. Since x^* is a sharply repulsive critical point, M is neighbored by a repulsive region (6.6). Assumption 6.3.4 then forces the algorithm to recur in this repulsive region, which in turn contradicts Proposition 6.3.1.

6.3.2 Convergence to minimizers

From the results of Section 6.2 we have that every Clarke critical point of a generic definable function that is not a minimizer is either a sharply repulsive critical point or an active strict saddle. Hence, to obtain the convergence of the SGD to minimizers we need to investigate the question of the avoidance of active strict saddles. As previously mentioned, this question was tackled in [Bianchi *et al.* 2021b].

Proposition 6.3.4 ([Bianchi *et al.* 2021b, Theorem 3])**.** *Let $f : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz continuous. Consider the iterates (6.7) under Assumptions 6.3.1(i))–(ii)). Assume that x^* is an active strict saddle lying on a C^4 active manifold that satisfies the Verdier and the angles conditions. Furthermore, assume that the following holds.*

i) There is $c_3, c_4 > 0$ and $\epsilon \in (1/2, 1]$ such that for all $n \in \mathbb{N}$,

$$\frac{c_3}{n^\epsilon} \leqslant \gamma_n \leqslant \frac{c_4}{n^\epsilon}\,.$$

ii) The sequence (η_{n+1}) is such that $\mathbb{E}_n[\eta_{n+1}] = 0$.

iii) For all $w \in \mathbb{R}^d \backslash \{0\}$, we have almost surely:

$$\liminf_{n\in\mathbb{N}} \mathbb{E}_n[|\langle \eta_{n+1}, w\rangle|] > 0\,,$$

and on the event $[x_n \to x^*]$:

$$\sup_{n\in\mathbb{N}} \mathbb{E}_n[\|\eta_{n+1}\|^4] < +\infty\,,$$

Then $\mathbb{P}([x_n \to x^*]) = 0$.

Notice that the assumption on (γ_n) of the preceding proposition implies Assumption 6.3.1-(iii)). Therefore, combining Proposition 6.3.4 with Theorem 6.3.3, we obtain that on a generic definable, locally Lipschitz continuous function the SGD converges to a local minimum. We state this result in the following corollary.

Corollary 6.3.5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz continuous. Assume that every of its Clarke critical points is isolated and is either a local minimum, an active strict saddle or a sharply repulsive critical point with the corresponding active manifolds being C^4-smooth and satisfying the Verdier and the angle conditions. Consider the iterates (6.7) under Assumptions 6.3.1(i))–(ii)) and 6.3.4. Moreover, assume that the following almost surely holds.*

i) For all $n \in \mathbb{N}$, $\mathbb{E}_n[\eta_{n+1}] = 0$.

ii) For every $C > 0$, there is $K(C) > 0$ such that:

$$\sup_{n\in\mathbb{N}} \mathbb{E}_n[\|\eta_{n+1}\|^4]\mathbb{1}_{\|x_n\|\leqslant C} < K(C)\,.$$

iii) There is $c_3, c_4 > 0$ and $\epsilon \in (1/2, 1]$ such that for all $n \in \mathbb{N}$,

$$\frac{c_3}{n^\epsilon} \leqslant \gamma_n \leqslant \frac{c_4}{n^\epsilon}\,.$$

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

Proof. Let x^* be one of the Clarke critical points of f. The only thing that we have to show is that under Assumption 6.3.4 we have that for all $w \in \mathbb{R}^d \backslash \{0\}$, almost surely,

$$\liminf_{n\in\mathbb{N}} \mathbb{E}_n[|\langle w, \eta_{n+1}\rangle|] > 0\,.$$

Consider $w \in \mathbb{R}^d$ and define $\delta = \frac{\|w\|}{2}$. Notice that for $x \in B(w, \delta)$ we have $|\langle w, x\rangle| \geqslant \|w\|^2 - |\langle x - w, w\rangle| \geqslant \frac{\|w\|^2}{2}$. Therefore,

$$\begin{aligned}\mathbb{E}_n[|\langle w, \eta_{n+1}\rangle|] \geqslant \mathbb{E}_n[|\langle w, \eta_{n+1}\rangle|\mathbb{1}_{\eta_{n+1}\in B(w,\delta)}] &\geqslant \frac{\|w\|^2}{2}\mathbb{P}\left(\eta_{n+1} \in B\left(w, \delta\right)|\mathscr{F}_n\right)\\ &\geqslant \frac{\|w\|^2}{2}\int_{x\in B(w,\delta)} h_{w,\delta}(x)\,\mathrm{d}x\,,\end{aligned}$$

and the right hand side of this inequality is positive by Assumption 6.3.4. □

We finish this section by a discussion on differences in the proofs of Proposition 6.3.4 and Theorem 6.3.3.

The idea of the proof of Proposition 6.3.4 can be described as follows.

- Using the angle condition, show that the iterates of the SGD converge to M fast enough.
- Using the Verdier condition, show that the sequence $(P_M(x_n))$ of the projected SGD iterates follows a gradient descent on a smooth function f_M.
- Since x^* is an active strict saddle, it is a saddle point of the function f_M and, with some minor adaptations, the nonconvergence follows from the works of [Brandière & Duflo 1996, Pemantle 1990] on avoidance of saddle points when the objective is smooth.

In [Bianchi *et al.* 2021b] the technique used to prove the first point is similar to the one used for the proof of Proposition 6.3.2. However, afterwards, the reasons for the nonconvergence to an active strict saddle are different. Indeed, in Theorem 6.3.3 the SGD avoids a sharply repulsive critical point due to the fact that the iterates (x_n) visits infinitely often a repulsive region of the form (6.6). Such repulsive region does not necessarily exist in the case of an active strict saddle (think of the function f_2 from the introduction). Hence, the proof of Proposition 6.3.4 heavily relies on the Verdier condition which is not necessary in our case.

Nevertheless, we notice that if one wants to describe the speed of convergence to x^* lying on an active manifold M such that x^* is a local minimum of f_M (e.g. a sharply repulsive critical point or a local minimum of f), then both of the Verdier and the angle conditions are useful since, as in [Bianchi *et al.* 2021b], it can be established that the iterates will converge promptly to M and the Verdier condition allows to show that, up to a manageable error term, the sequence $(P_M(x_n))$ is simply an SGD sequence applied to a smooth function f_M. It should be possible in that case to obtain rates of convergence in the spirit of [Mertikopoulos *et al.* 2020b]. We defer such considerations to future work.

6.3.3 Validity of Assumption 6.3.4

In this section we present a model that satisfies Assumptions 6.3.1–6.3.2 and show how to alter it to obtain Assumption 6.3.4. This will provide us with a practical way to ensure the convergence of the SGD to a minimizer.

We start by a motivational example.

Example 6.3.2. *In machine learning we are usually interested to optimize $f : \mathbb{R}^d \to \mathbb{R}$ written as an average of N functions:*

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

Here, N is the number of data points and f_i is the loss function related to the i-th data point. In this case a version of the SGD is obtained by choosing, at each iteration $n \in \mathbb{N}$, an index $i \in \{1, \dots, N\}$ in an uniform manner and updating:

$$x_{n+1} = x_n - \gamma_n g_i(x_n)\,, \tag{6.8}$$

where $g_i : \mathbb{R}^d \to \mathbb{R}$ are such that $\frac{1}{N}\sum_{i=1}^N g_i(x) \in \partial f(x)$. If for all $i \in \{1, \dots, N\}$ are smooth, concave or weakly convex [2], then we can choose $g_i(x) \in \partial f_i(x)$. In this case Equation (6.8) can be viewed as Equation (6.7) by putting $v_n = \frac{1}{n}\sum_{i=1}^N g_i(x_n)$ and $\eta_{n+1} = g_i(x_n) - v_n$.

In the stochastic approximation litterature (see e.g. [Borkar 2008, Kushner & Yin 2003]) this and more general settings are modelized by a probability space $(\Xi, \mathscr{T}, \mu)$ and a measurable function $g : \mathbb{R}^d \times \Xi \to \mathbb{R}^d$ such that for each $x \in \mathbb{R}^d$, the function $g(x, \cdot)$ is μ-integrable and we have the following:

$$G(x) := \int_\Xi g(x, s)\mu(\mathrm{d}s) \in \partial f(x)\,.$$

Starting from $x_0 \in \mathbb{R}^d$, at each iteration $n \in \mathbb{N}$ the practitioner samples $\xi_n \sim \mu$ in an independent way and update the iterates according to the following rule:

$$x_{n+1} = x_n - \gamma_n g(x_n, \xi_{n+1})\,.$$

We obtain Equation (6.7) by putting $v_n = G(x_n)$, $\eta_{n+1} = G(x_n) - g(x_n, \xi_{n+1})$ and $\mathscr{F}_n = \sigma(x_0, \xi_1, \dots, \xi_n)$.

In the context of this model consider a sequence (η_n^1) of i.i.d $\mathbb{R}^d$-valued random variables, with $\eta^1 \sim \nu$ s.t. the following holds.

1. For each $n \in \mathbb{N}$, η_n^1 is independent from $\mathscr{F}_n$.
2. The law ν is zero-mean with finite variance.
3. The law ν has a continuous density relatively to the Lebesgue measure on $\mathbb{R}^d$. Moreover, denoting this density $h^1 : \mathbb{R}^d \to \mathbb{R}$, we have that for each point $x \in \mathbb{R}^d$, $h^1(x) > 0$.

An example of a law that verifies the last two points is e.g. a nondegenerate gaussian.

Starting from a $\mathscr{F}_0$-measurable point $\tilde{x}_0 \in \mathbb{R}^d$, consider the following algorithm:

$$\tilde{x}_{n+1} = \tilde{x}_n - \gamma_n g(\tilde{x}_n, \xi_{n+1}) + \gamma_n \eta_{n+1}^1 = \tilde{x}_n - \gamma_n \tilde{v}_n + \gamma_n \tilde{\eta}_{n+1}\,, \tag{6.9}$$

where $\tilde{v}_n = G(\tilde{x}_n) \in \partial f(\tilde{x}_n)$ and $\tilde{\eta}_{n+1} = \tilde{v}_n - g(\tilde{x}_n, \xi_{n+1}) + \eta_{n+1}^1$.

Proposition 6.3.6. *Assume that for every $C > 0$, there is $K(C) > 0$ such that:*

$$\sup_{\|x\| \leqslant C} \int \|G(x, s) - g(x, s)\|^2 \, \mu(\mathrm{d}s) \leqslant K(C)\,.$$

Then the sequence $(\tilde{\eta}_{n+1})$ satisfies Assumptions 6.3.2 and 6.3.4 relatively to the filtration $\widetilde{\mathscr{F}_n} = \mathscr{F}_n \otimes \sigma(\eta_1^1, \dots, \eta_n^1)$.

[2] Or more generally Clarke regular. For an existence of such an oracle we invite the reader to consult [Bianchi *et al.* 2021a].

The preceding proposition provides a practical way to ensure the convergence of the SGD to a minimizer. As the following corollary states, if the objective function is generic, then it suffices to add a small random perturbation at every iteration of the algorithm.

Corollary 6.3.7. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be satisfying the assumptions of Corollary 6.3.5. Consider the iterates of Equation (6.9) and assume that the following holds.*

i) For every $C > 0$, there is $K(C) > 0$ such that:

$$\sup_{\|x\| \leqslant C} \int \|G(x) - g(x, s)\|^4 \mu(\mathrm{d}s) < K(C)\,.$$

ii) The law ν has a finite fourth order moment.

iii) There is $c_3, c_4 > 0$ and $\epsilon \in (1/2, 1]$ such that for all $n \in \mathbb{N}$:

$$\frac{c_3}{n^\epsilon} \leqslant \gamma_n \leqslant \frac{c_4}{n^\epsilon}\,.$$

Then the sequence $(\tilde{x}_n)$ is either unbounded or converges to a local minimum of f.

6.4 Proofs

We recall that most of the results on o-minimality that are used in this proof are gathered in Section 2.4. In particular, we will use the notion of first-order formula and Proposition 2.4.1 without further mentioning.

The proof of Theorem 6.2.5 is based on the following result of Drusvyatskiy, Ioffe and Lewis [Drusvyatskiy *et al.* 2016].

Proposition 6.4.1 ([Drusvyatskiy *et al.* 2016, Theorem 4.7 and Corollary 4.8])**.** *Consider $p \geqslant 2$ and $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous, definable function. There is $N > 0$ such that for almost every u, the function f_u has at most N Clarke critical points. Moreover, denoting x_u^* such a point, the following holds.*

i) There is M a C^p active manifold around x_u^* (for the function f_u).

ii) There is W, V neighborhoods of respectively u and x_u^* such that the mapping

$$w \mapsto V \cap (\partial f_w)^{-1}(0)$$

is single valued, C^p smooth on W and maps W onto M.

iii) If $0 \in \partial_L f_u(x_u^*)$ and x_u^* is a local minimum of the function $f_{u|M} : M \to \mathbb{R}$, then x_u^* is a local minimum of the unrestricted function f_u.

In [Drusvyatskiy *et al.* 2016] the preceding proposition was stated for the limiting subgradient $\partial_L f_u$ (i.e. in Definition 6.2.1 the Clarke subgradient ∂f was replaced by $\partial_L f$ and the critical point x_u^* was defined as $0 \in \partial_L f_u(x^*)$). However, the only property that was used for the proof of these points was the fact that $\dim \operatorname{Graph}(\partial_L f(x)) = d$ (here dimension is understood in the sense of Definition 2.4.4). Since by [Drusvyatskiy & Lewis 2010b, Theorem 3.5] $\dim \operatorname{Graph}(\partial f(x)) = d$, Proposition 6.4.1 remains true with our definition.

Let $\{X_1, \dots, X_k\}$ be the C^p stratification from Theorem 5.2.1. The existence of an active manifold with a Verdier condition follows from Proposition 6.4.1 upon noticing that in the proof of [Drusvyatskiy *et al.* 2016, Corollary 4.8] the active manifold can be chosen compatible with $\{X_1, \dots, X_k\}$. To deal with the angle condition and the nondegeneracy of critical points we consider separately the case where $\dim M = 0$ and $\dim M > 0$.

First case: $\dim M = 0$**.** In this case the angle condition follows from the following result of [Bolte *et al.* 2009].

Lemma 6.4.2 ([Bolte *et al.* 2009, Theorem 1 and Proposition 1])**.** *Consider $f : \mathbb{R}^d \to \mathbb{R}$ a locally Lipschitz continuous and definable function. For $(x, d) \in \mathbb{R}^d \times \mathbb{R}^d$, denote $f'(x; d) = \lim_{t \to 0} \frac{f(x+td)-f(x)}{t}$ (notice that this limit always exists by Lemma 2.4.2 and the local Lipschitz continuity of f). We have:*

$$|f(x + d) - f(x) - f'(x + d; d)| = o_x(\|d\|)\,, \tag{6.10}$$

where $o_x(\|d\|)$ means that $\frac{o_x(\|d\|)}{\|d\|} \to_{d \to 0} 0$.

Without loss of generality, assume that the critical point x_u^* is equal to zero and $f_u(0) = 0$. We have that $M = \{0\}$ and by Equation (6.10):

$$|f_u(x_u^* + x) - f_u(x_u^*) - f'(x^* + x; x)| = |f_u(x) - f'_u(x; x)| = o(\|x\|)\,.$$

For $\alpha > 0$, we have:

$$f_u(x) \geqslant \alpha \|x\| \implies f'_u(x; x) \geqslant \alpha \|x\| + o(\|x\|)\,.$$

Therefore, for x close enough to zero:

$$f'_u(x; x) \geqslant \frac{\alpha}{2} \|x\|\,.$$

Notice that if f_u is differentiable at x, then $f'_u(x; x) = \langle \nabla f_u(x), x \rangle$. Hence, the angle condition is proved on a point of differentiability.

For the general case, consider $v \in \partial f_u(x)$. By Proposition 2.1.1 there is $k \in \mathbb{N}$, a sequence $(x_1^n, \dots, x_k^n) \to (x, \dots, x)$ and a sequence $(\lambda_1, \dots, \lambda_k)$ such that $\sum_{i=1}^k \lambda_i = 1$, for each $(i, n) \in [1, \dots, k] \times \mathbb{N}$, f_u is differentiable at x_i^n and $v = \lim_{n \to +\infty} \sum_{i=1}^k \lambda_i \nabla f_u(x_i^n)$. Since at x_i^n we have $f'_u(x_i^n; x_i^n) = \langle \nabla f(x_i^n), x_i^n \rangle$, we ob-

tain:

$$\begin{aligned}\langle v, x\rangle &= \lim_{n\to+\infty} \sum_{i=1}^{k} \lambda_i \langle \nabla f_u(x_i^n), x\rangle \\ &= \lim_{n\to+\infty} \sum_{i=1}^{k} \lambda_i \langle \nabla f_u(x_i^n), x_i^n\rangle + \lim_{n\to+\infty} \sum_{i=1}^{k} \lambda_i \langle \nabla f_u(x_i^n), x - x_i^n\rangle \\ &\geqslant \frac{\alpha}{2} \lim_{n\to+\infty} \sum_{i=1}^{k} \lambda_i \left\|x_i^n\right\| + \lim_{n\to+\infty} \sum_{i=1}^{k} \lambda_i \langle \nabla f_u(x_i^n), x - x_i^n\rangle \\ &\geqslant \frac{\alpha}{2} \|x\| \,,\end{aligned}$$

where the last inequality is obtain by the triangular inequality, the fact that $x_i^n \to x$ and that $\nabla f(x_i^n)$ is bounded.

Second case: $\dim M = k > 0$**.** Let u, W, x_u^* be as in Proposition 6.4.1. In the following, without loss of generality, we will assume that $u = 0$ and W is bounded. We start by outlining the proof.

- Using Proposition 6.2.1 and the Verdier condition, we show that for almost every $w \in W$ the critical point of $f_{w|M}$ are nondegenerate.
- In Lemmas 6.4.3 and 6.4.4 we show that that the dimension of $y \in M$ such that the angle condition is verified in the neighborhood of y is equal to k.
- The preceding point along with Lemma 6.4.5 and the second point of Proposition 6.4.1 shows that for almost every perturbation $w \in W$, the angle condition is verified in a neighborhood of x_w^*.

Since $\mathbb{R}^d$ is covered by a countable union of such neighborhoods W these three points will prove Theorem 6.2.5.

To prove the first point of the outline notice that by the Verdier condition the map from Proposition 6.4.1 is actually equal to:

$$w \mapsto x_w^* \in M \qquad \text{s.t. } \nabla_M f_w(x_w^*) = 0\,.$$

Therefore, by Proposition 6.2.1 we have that for almost every $w \in W$ the critical point x_w^* is nondegenerate for the function $f_{w|M}$.

To prove the second point denote $P_{\alpha,\beta}(x)$ the first-order formula:

$$\begin{aligned}&\forall w \in W,\\ &f_w(x) \geqslant \alpha \|x - P_M(x)\| + f_w(P_M(x)) \implies \langle v_w, x - P_M(x)\rangle \geqslant \beta \|x - P_M(x)\|\,, \quad \forall v_w \in \partial f_w(x)\,,\end{aligned}$$

where, implicitly, in this formula we consider only such x for which $P_M(x)$ is well defined. The first-order formula "not $P_{\alpha,\beta}(x)$" can be written as:

$$\begin{aligned}&\exists w \in W, \exists v_w \in \partial f_w(x) \text{ such that:}\\ &\qquad f_w(x) \geqslant \alpha \|x - P_M(x)\| + f_w(P_M(x)) \text{ and } \langle v_w, x - P_M(x)\rangle < \beta \|x - P_M(x)\|\,.\end{aligned}$$

Or equivalently:

$$\begin{aligned}\exists w \in W, \exists v \in \partial f(x) \text{ such that:}\\ f(x) - f(P_M(x)) \geqslant \alpha \|x - P_M(x)\| + \langle w, x - P_M(x)\rangle\\ \langle v - w, x - P_M(x)\rangle < \beta \|x - P_M(x)\| \ .\end{aligned}$$

Fix $\alpha > 0$ and let A^α be the following definable set:

$$A^\alpha = \{y \in M : \forall \varepsilon > 0, \exists x \in B(y, \varepsilon), \text{ “not } P_{2\alpha,\alpha}(x)\text{"}\} \ . \tag{6.11}$$

The following lemma shows that for a fixed α, for almost every $y \in M$, there is a neighborhood around y such that the formula $P_{2\alpha,\alpha}$ holds.

Lemma 6.4.3. *We have that* $\dim A^\alpha < \dim M = k$.

Proof. Assume the contrary, by construction A^α lies in the boundary of the following definable set.

$$Q_\alpha := \{x \notin M : \text{ “ not } P_{2\alpha,\alpha}(x)\text{”}\} \, .$$

Applying Lemma 2.4.9, we obtain a k-dimensional definable set A', $\delta > 0$ and a definable C^1 map $\rho : A' \times (0, \delta) \to Q_\alpha$ such that $P_M(\rho(y, t)) = y$ and $\|\rho(y, t) - y\| = t$. By the definition of Q_α this means that for all $(y, t) \in A' \times (0, \delta)$ there is $v \in \partial f(\rho(y, t))$ and $w \in W$ such that

$$\langle v - w, \rho(y, t) - y\rangle < \alpha t < 2\alpha t \leqslant f(\rho(y, t)) - f(y) - \langle w, \rho(y, t) - y\rangle \, .$$

Fix $y \in A'$ and denote $\rho_y(\cdot) = \rho(y, \cdot)$. There are two definable selections $\mathsf{v}(t) \in \partial f(\rho_y(t))$ and $\mathsf{w}(t) \in W$ such that

$$\langle \mathsf{v}(t) - \mathsf{w}(t), \rho_y(t) - y\rangle < \alpha t < 2\alpha t \leqslant f(\rho_y(t)) - f(y) - \langle \mathsf{w}(t), \rho_y(t) - y\rangle \, . \tag{6.12}$$

Since f is path-differentiable, we have:

$$f(\rho(y, t)) - f(y) = \int_0^t \langle \mathsf{v}(u), \dot{\rho}_y(u)\rangle \, \mathrm{d}u \, .$$

For t' small enough, we have that the expression under the integral is continuous on $(0, t')$. Therefore, by the mean value theorem, for every $t \in (0, t')$, there is $u^t \in (0, t)$ such that

$$f(\rho_y(t)) - f(y) = t\langle \mathsf{v}(u^t), \dot{\rho}_y(u^t)\rangle \, .$$

Denote $v_y = \lim_{t\to 0} \mathsf{v}(t)$ and $w_y = \lim_{t\to 0} \mathsf{w}(t)$ (these limits exists by the monotonicity lemma and the fact that f is locally Lipschitz continuous). By Lemma 2.4.3 applied to each coordinate of $\rho_y(t) - y$ we have the existence of $R_y \in \mathbb{R}^d$ such that $R_y = \lim_{t\to 0} \frac{\rho_y(t) - y}{t} = \lim_{t\to 0} \dot{\rho}_y(t)$. Hence, from Inequality (6.12) we obtain:

$$\langle v_y - w_y, R_y\rangle \leqslant \alpha < 2\alpha \leqslant \langle v_y - w_y, R_y\rangle \, ,$$

which is a contradiction. □

Denote $\mathcal{L} \subset M$ the set:

$$\mathcal{L} = \bigcap_{\alpha>0} \bigcup_{\beta>0} \{y \in M : \exists \varepsilon > 0 \text{ s.t. } \forall x \in B(y,\varepsilon),\ P_{\alpha,\beta}(x)\}\,. \tag{6.13}$$

The second point of the outline comes from the following lemma.

Lemma 6.4.4. *The following holds.*

i) The set $\mathcal{L}$ is definable.

ii) We have that

$$\bigcap_{\alpha\in\mathbb{Q},\alpha>0} (A^\alpha)^c \cap M \subset \mathcal{L}\,. \tag{6.14}$$

iii) We have that $\dim(\mathcal{L}^c \cap M) < k$.

Proof. The first statement comes from the fact that the set $\mathcal{L}$ can be written as a first-order formula. The second statement is immediate from definitions. Finally, the last statement comes from the fact that we have:

$$\mathcal{L}^c \cap M \subset \bigcup_{\alpha\in\mathbb{Q}} A^\alpha\,.$$

By Remark 6 the "definable dimension" of a set coincides with its Hausdorff dimension and by Lemma 6.4.3 we have $\dim A^\alpha < k$. Hence, $\mathcal{L}^c \cap M$ is included in a countable union of sets of Hausdorff dimension less than k. Therefore, $\dim \mathcal{L}^c \cap M < k$. □

Define $S := \{(y,w) : y \in \mathcal{L}^c \cap M, w \in \partial f(y) \cap W\}$ and $S_w := \{w \in W : \exists y \in \mathcal{L}^c \cap M, (y,w) \in S\}$

Lemma 6.4.5. *We have that* $\dim S_w < d$.

Proof. For $y \in \mathcal{L}^c \cap M$ define the set $S_y := \{w : w \in \partial f(y)\}$. By Theorem 5.2.1 we have that $S_y \subset \nabla_M f(y) + \mathcal{N}_y M$ and therefore $\dim S_y \leqslant \dim \mathcal{N}_y M = d - k$. By Lemma 6.4.4 we have that $\dim \mathcal{L}^c \cap M < k$. Therefore, applying Proposition 2.4.8, we obtain that

$$\dim S = \dim \mathcal{L}^c \cap M + \sup_{y\in\mathcal{L}^c\cap M} \dim S_y < k + d - k = d\,.$$

Since S_w is the image of S by the projection on the last d coordinates, applying Proposition 2.4.8, we obtain $\dim S_w \leqslant \dim S < d$. □

Therefore, for almost every $w \in W$, the set $\{y \in M : w \in \partial f(y)\} = \{x_w^* \in M : \nabla_M f_w(x_w^*) = 0\}$ lies in $\mathcal{L}$. By the definition of $\mathcal{L}$ for each $y \in \mathcal{L}$ and every $\alpha > 0$, there is $\beta, \varepsilon > 0$ such that for all $x \in B(y,\varepsilon)$ and $w \in W$,

$$f_w(x) - f_w(P_M(x)) \geqslant \alpha \|x - P_M(x)\| \implies \langle v_w, x - P_M(x)\rangle \geqslant \beta \|x - P_M(x)\|\,, \quad \forall v_w \in \partial f_w(x)\,,$$

which finishes the proof.

6.4.1 Proof of Proposition 6.2.3

The first statement immediately follows from the definition of the limiting subgradient.

To prove the second statement, without loss of generality, assume that $x^* = 0$. By contradiction assume that for all $\varepsilon > 0$ there is $y_\varepsilon \in B(0,\varepsilon) \cap M$ and $\varepsilon_{y_\varepsilon} > 0$ such that for all $x \in B(y_\varepsilon, \varepsilon_{y_\varepsilon})$ we have:

$$f(x) \geqslant f(P_M(x))\,.$$

By Lemma 2.3.3 we have:

$$\begin{aligned} f(x) &\geqslant f(y_\varepsilon) + f(P_M(x)) - f(y_\varepsilon) \\ &\geqslant f(y_\varepsilon) + \langle \nabla_M f(y_\varepsilon), P_M(x) - y_\varepsilon\rangle + O(\|P_M(x) - y_\varepsilon\|^2) \\ &\geqslant f(y_\varepsilon) + \langle \nabla_M f(y_\varepsilon), x - y_\varepsilon\rangle + \langle \nabla_M f(y_\varepsilon), P_M(x) - x\rangle + O(\|P_M(x) - y_\varepsilon\|^2) \end{aligned}$$

Since f is C^2 on M, we have:

$$\begin{aligned} \langle \nabla_M f(y_\varepsilon), P_M(x) - x\rangle &\geqslant \langle \nabla_M f(P_M(x)), P_M(x) - x\rangle - \|\nabla_M f(P_M(x)) - \nabla_M f(y_\varepsilon)\|\,\|P_M(x) - x\| \\ &\geqslant \langle \nabla_M f(P_M(x)), P_M(x) - x\rangle + O(\|x - y_\varepsilon\|)\,\|P_M(x) - x\|\,. \end{aligned}$$

Notice that $\nabla_M f(P_M(x)) \in T_{P_M(x)}$ and that by Lemma 2.3.2 for x close enough to M, we have $P_M(x) - x \in \mathcal{N}_{P_M(x)}M$. Therefore, $\langle \nabla_M f(P_M(x)), P_M(x) - x\rangle = 0$ and we obtain:

$$f(x) \geqslant f(y_\varepsilon) + \langle \nabla_M f(y_\varepsilon), x - y_\varepsilon\rangle + \|x - P_M(x)\|\, O(\|y_\varepsilon - P_M(x)\|) + O(\|P_M(x) - y_\varepsilon\|^2)\,.$$

Since P_M is Lipschitz continuous, we have $O(\|P_M(x) - y_\varepsilon\|) = O(\|x - y_\varepsilon\|)$. Hence, $\nabla_M f(y_\varepsilon) \in \partial_L f(y_\varepsilon)$. Since $\nabla_M f(y_\varepsilon) \to_{y_\varepsilon \to 0} 0$, this implies that $0 \in \partial_L f(0)$, a contradiction.

6.4.2 Proof of Proposition 6.3.1

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

$$\mathsf{X}(t) = x_n + \frac{t - \sum_{i=0}^n \gamma_n}{\gamma_{n+1}}(x_{n+1}) - x_n, \qquad \text{if } t \in \left[\sum_{i=0}^{n} \gamma_i, \sum_{i=0}^{n+1} \gamma_i\right)\,.$$

By Assumption 6.3.2 and [Schechtman 2021a, Lemma 1] the sequence $(\sum_{i=0}^n \gamma_i \eta_{i+1})$ converges on the event $[x_n \to x^*]$. Hence, from the work of Benaïm, Hofbauer and Sorin [Benaïm *et al.* 2005, Proposition 1.3], on the event $[x_n \to x^*]$, X is a so-called *asymptotic pseudo trajectory* of the subgradient flow (5.14). That is to say, for every $T \geqslant 0$, by [Benaïm *et al.* 2005, Proposition 4.1] we have that:

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

Assume that the statement of the theorem is not true. Choose T, r, α as in Proposition 6.2.2 and a sequence (x_{n_k}) that converges to x^* but $f(x_{n_k}) < f(P_M(x_{n_k})) + \alpha \|x_{n_k} - P_M(x_{n_k})\|$. Denote $t_{n_k} = \sum_{i=0}^{n_k} \gamma_i$ and x_{n_k} a solution in $\mathsf{S}_{-\partial f}(x_{n_k})$ such that

$$\inf_{h \in [0,T]} \|\mathsf{x}_{n_k}(h) - \mathsf{X}(t_{n_k} + h)\| \to 0 \,.$$

By Proposition 6.2.2 for n_k large enough, we have that $\mathsf{x}_{n_k}([0,T]) \not\subset B(x^*, r)$. Therefore, we can extract a sequence $t_{n'_k} \geqslant t_{n_k}$ such that $\left\|\mathsf{X}(t_{n'_k}) - x^*\right\| \geqslant \frac{r}{2}$. Since on the event $[x_n \to x^*]$ the limit set of X is equal to x^*, this is a contradiction.

6.4.3 Proof of Proposition 6.3.2

In this proof C will denote some absolute constant that can change from line to line and from one statement to another. Without loss of generality assume that $x^* = 0$ and $f(x^*) = 0$.

To prove this proposition the following result will be needed.

Lemma 6.4.6 ([Bianchi *et al.* 2021b, Lemma 5])**.** *Let Assumption 6.3.3 hold. There is $C, r > 0$ such that the conclusions of Lemma 2.3.2, with $p = 2$, are verified on $B(0, r_1)$ and, moreover, for any $x_1, x_2 \in B(0, r)$, we have:*

$$\left\|y_2 - y_1 - P_{T_{y_1}}(x_2 - x_1)\right\| \leqslant C \|x_1 - x_2\|^2 + C \|x_1 - x_2\| \|x_1 - y_1\| \,,$$

where $y_1, y_2 = P_M(x_1), P_M(x_2)$.

Let $\alpha > 0$ be the one of Proposition 6.3.1 and let U_α, β be as in Definition 6.2.2. Consider r_1 as in the preceding lemma and let $r > 0$ be such that $B(0, r) \subset U_\alpha$ and $r \leqslant r_1$. The value of r, while always satisfying this requirement, will be adjusted in the course of the proof. Denote

$$z_n = (x_n - P_M(x_n)) \mathbb{1}_{\|x_n\| \leqslant r} \,.$$

Notice that if $\|x_n\| \leqslant r$, then $\|z_n\| = \operatorname{dist}(x_n, M)$ and $z_n \in \mathcal{N}_{P_M(x_n)} M$.

Consider $\kappa > 0$ and for $k \in \mathbb{N}$, denote

$$\tau_k(\kappa, r, \alpha) = \{\inf n \geqslant k : \operatorname{dist}(x_n, M) \leqslant \kappa\gamma_n \text{ or } \|x_n\| \geqslant r \text{ or } f(x_n) < f(P_M(x_n)) + \alpha \|x_n - P_M(x_n)\|\} \,.$$

By a slight abuse of notations we will denote $\tau_k = \tau_k(\kappa, r, \alpha)$ and $z_n^{\tau_k} = z_{n \wedge \tau_k}$.

The aim of this proof is to show that for any $k \in \mathbb{N}$, $\mathbb{P}(\tau_k = +\infty) = 0$. Since on the event $[x_n \to 0]$, for n large enough, we have $\|x_n\| \leqslant r$ and $f(x_n) < f(P_M(x_n)) + \alpha \|x_n - P_M(x_n)\|$ this will implies that $\operatorname{dist}(x_n, M) \leqslant \kappa\gamma_n$ happens infinitely often.

To establish this result we study the difference between $\left\|z_{n+1}^{\tau_k}\right\|^2$ and $\|z_n^{\tau_k}\|^2$. Using the angle condition, we show that for $\tau_k > n$, it will decrease at least at a rate $\gamma_n \|z_n^{\tau_k}\|$. Since for r small enough, $\|z_n^{\tau_k}\|$ is much larger than $\|z_n^{\tau_k}\|^2$ this will help us to conclude.

We have:

$$\begin{aligned}\left\|z_{n+1}^{\tau_k}\right\|^2 &= \|z_n^{\tau_k}\|^2 + \left(2\langle z_{n+1}-z_n, z_n\rangle + \|z_{n+1}-z_n\|^2\right)\mathbb{1}_{\tau_k>n} \\ &= \|z_n^{\tau_k}\|^2 + \left(2\langle x_{n+1}-x_n, z_n\rangle - 2\langle P_M(x_{n+1})-P_M(x_n), z_n\rangle + \|z_{n+1}-z_n\|^2\right)\mathbb{1}_{\tau_k>n}\,. \end{aligned} \tag{6.15}$$

The following lemma bound the last two quantities.

Lemma 6.4.7. *There is $d_1, d_2 > 0$ s.t. if r was chosen small enough, then:*

$$2\mathbb{E}_n[|\langle P_M(x_{n+1})-P_M(x_n), z_n\rangle|]\mathbb{1}_{\tau_k>n} \leqslant d_1(\gamma_n \|z_n\|^2 + \gamma_n^2)\mathbb{1}_{\tau_k>n}\,,$$

and

$$\mathbb{E}_n[\|z_{n+1}-z_n\|^2]\mathbb{1}_{\tau_k>n} \leqslant d_2\gamma_n^2\,.$$

Proof. To prove the first inequality apply Lemma 6.4.6. On the event $[\tau_k > n]$, noticing that z_n is orthogonal to $T_{P_M(x_n)}M$, we obtain:

$$|\langle P_M(x_{n+1})-P_M(x_n), z_n\rangle| \leqslant C\|x_{n+1}-x_n\|^2\|z_n\| + C\|x_{n+1}-x_n\|\|z_n\|^2\,.$$

Hence,

$$\begin{aligned}\mathbb{E}_n[|\langle P_M(x_{n+1})-P_M(x_n), z_n\rangle|]\mathbb{1}_{\tau_k>n} &\leqslant \left(C\mathbb{E}_n\left[\|x_{n+1}-x_n\|^2\|z_n\| + \|x_{n+1}-x_n\|\|z_n\|^2\right]\right)\mathbb{1}_{\tau_k>n} \\ &\leqslant C\left(\gamma_n\|z_n\|^2 + \gamma_n^2\|z_n\|\right)\mathbb{1}_{\tau_k>n} \\ &\leqslant C(\gamma_n\|z_n\|^2 + \gamma_n^2)\mathbb{1}_{\tau_k>n}\end{aligned}$$

The second inequality is obtained by writing down:

$$\left\|z_{n+1}^{\tau_k} - z_n^{\tau_k}\right\|^2\mathbb{1}_{\tau_k>n} \leqslant C\|x_{n+1}-x_n\|^2\mathbb{1}_{\tau_k>n} + C\|P_M(x_{n+1})-P_M(x_n)\|^2\mathbb{1}_{\tau_k>n}\,.$$

Taking the conditional expectation, we obtain the desired result using Lemma 6.4.6, Equation (4.1) and Assumption 6.3.2. □

Using the preceding lemma and the angle condition, taking the conditional expectation on Inequality (6.15), we obtain:

$$\begin{aligned}\mathbb{E}_n[\left\|z_{n+1}^{\tau_k}\right\|^2] &\leqslant \|z_n^{\tau_k}\|^2 + \left(d_1\gamma_n\|z_n\|^2 - 2\gamma_n\langle v_n, z_n\rangle + (d_1+d_2)\gamma_n^2\right)\mathbb{1}_{\tau_k>n} \\ &\leqslant \|z_n^{\tau_k}\|^2 + \left(d_1\gamma_n\|z_n\|^2 - 2\gamma_n\beta_1\|z_n\| + (d_1+d_2)\gamma_n^2\right)\mathbb{1}_{\tau_k>n}\end{aligned}$$

Denote $\theta_n = \frac{z_n}{\gamma_n}$ and $\theta_n^{\tau_k} = \theta_{n\wedge\tau_k}$.

Lemma 6.4.8. *There is $d_3, d_4 > 0$ s.t. if $r > 0$ was chosen small enough, we have:*

$$\mathbb{E}_n[\left\|\theta_{n+1}^{\tau_k}\right\|^2] \leqslant \|\theta_n^{\tau_k}\|^2 + (d_3 - d_4\|\theta_n\|)\mathbb{1}_{\tau_k>n}\,.$$

Proof.

$$\mathbb{E}_n[\left\|\theta_{n+1}^{\tau_k}\right\|^2] \leqslant \|\theta_n^{\tau_k}\|^2 + \left(-\|\theta_n\|^2 + \frac{\gamma_n^2}{\gamma_{n+1}^2}((1+d_1\gamma_n)\|\theta_n\|^2 - 2\beta_1\|\theta_n\| + C(d_1+d_2))\right)\mathbb{1}_{\tau_k>n}\,.$$

Using Assumption 6.3.1, we obtain:

$$\frac{\gamma_n^2}{\gamma_{n+1}^2}(1+d_1\gamma_n) \leqslant 1 + (c_2+d_1)\gamma_n + c_2 d_1 \gamma_n^2\,.$$

Hence,

$$\begin{aligned}\mathbb{E}_n[\left\|\theta_{n+1}^{\tau_k}\right\|^2]\mathbb{1}_{\tau_k>n} &\leqslant \|\theta_n^{\tau_k}\|^2 + \left((c_2+d_1)\gamma_n\|\theta_n\|^2 + c_2 d_1\gamma_n^2\|\theta_n\|^2 - 2c_1\beta_1\|\theta_n\| + d_1 + d_2\right)\mathbb{1}_{\tau_k>n}\\ &\leqslant \|\theta_n^{\tau_k}\|^2 + \left((c_2+d_1)\gamma_n\|\theta_n\| + c_2 d_1\gamma_n^2\|\theta_n\| - 2c_1\beta_1\right)\|\theta_n\|\,\mathbb{1}_{\tau_k>n}\\ &\quad + C(d_1+d_2)\mathbb{1}_{\tau_k>n}\,.\end{aligned}$$

Note that $\gamma_n\|\theta_n\|\,\mathbb{1}_{\tau_k>n} = \|z_n\|\,\mathbb{1}_{\tau_k>n} \leqslant r$. Hence,

$$\mathbb{E}_n[\left\|\theta_{n+1}^{\tau_k}\right\|^2\mathbb{1}_{\tau_k>n} \leqslant \|\theta_n^{\tau_k}\|^2 + ((c_2+d_1)r + c_2 d_1\gamma_n r - 2c_1\beta_1)\|\theta_n\|\,\mathbb{1}_{\tau_k>n} + C(d_1+d_2)\mathbb{1}_{\tau_k>n}\,.$$

If r was chosen such that $(c_2+d_1)r + c_2 d_1\gamma_n r - 2c_1\beta_1 < 0$, then we have:

$$\mathbb{E}_n[\left\|\theta_{n+1}^{\tau_k}\right\|^2 \leqslant \|\theta_n^{\tau}\|^2 + (d_1 + d_2 - 2c_1\beta_1\|\theta_n\|)\,\mathbb{1}_{\tau_k>n}\,,$$

which is the desired result.

□

Consider d_3, d_4 from Lemma 6.4.8. If κ was chosen greater than $\frac{2d_3}{d_4}$, then from the preceding lemma we obtain:

$$\begin{aligned}\mathbb{E}_n[\left\|\theta_{n+1}^{\tau_k}\right\|^2] &\leqslant \|\theta_n^{\tau_k}\|^2 + (d_3 - d_4\|\theta_n^{\tau_k}\|)\mathbb{1}_{\tau_k>n}\\ &\leqslant \|\theta_n^{\tau_k}\|^2 - d_3\mathbb{1}_{\tau_k>n}\end{aligned}\,.$$

Hence, for all $n \geqslant k$:

$$0 \leqslant \mathbb{E}[\|\theta_n^{\tau_k}\|^2] + (n+1-k)\mathbb{P}(\tau_k = +\infty)\,,$$

which implies that $\mathbb{P}(\tau_k = +\infty) = 0$.

Hence, for any $k \in \mathbb{N}$, τ_k is almost surely finite. Noticing that by Proposition 6.3.1, on the event $[x_n \to 0]$, the events $[\|x_n\| > r]$ and $[f(x_n) < f(P_M(x_n)) + \alpha\|x_n - P_M(x_n)\|]$ happen only a finite number of times, this implies the statement of Proposition 6.3.2.

6.4.4 Proof of Theorem 6.3.3

The proof will be done in three steps.

1. Lemma 6.4.9 shows that there is a constant C such that if $\mathrm{dist}(x_n, M)$ is of order of γ_n, then there is x'_n such that $\mathrm{dist}(x'_n, M)$ is also of order γ_n and every point $B(x'_n, C\gamma_n)$ is in the repulsive region of the form (6.6).

2. Using Assumption 6.3.4, Lemma 6.4.10 then shows that in such a case the probability of x_{n+1} visiting a repulsive region is lower bounded.

3. The preceding point along with Proposition 6.3.2 and Lemma 6.4.11 then shows that the iterates (x_n) visit a repulsive region infinitely often. The latter is impossible by Proposition 6.3.1.

Without loss of generality, assume that $x^* = 0$. In this section U will be a bounded neighborhood of zero such that the following holds on U.

i) There is $c_m > 0$ s.t. $\inf\{\|\partial f(x)\| : x \in U \cap M^c\} \geqslant c_m$.

ii) The function P_M is C^1 on U.

iii) The functions $f, P_M, f \circ P_M$ are Lipschitz on U, with Lipschitz constants L, L_π, L_M.

Lemma 6.4.9. *Let Assumption 6.3.3 hold. There is $r_1 > 0$ s.t. $B(0, r_1) \subset U$, and for all $y \in B(0, r_1) \cap M$, for all δ s.t. $\|y\| + \delta < r_1$, there is x such that the following holds.*

i) $\|x - y\| = \delta$.

ii) $f(x) \leqslant f(P_M(x)) - \frac{c_m^2 \delta}{4L}$.

Moreover, for every such x, denoting $\delta' = \frac{c_m^2 \delta}{8L(L+L_M)}$, for every $x' \in B(x, \delta')$, we have:

$$f(x') < f(P_M(x'))\,.$$

Proof. Consider r from Proposition 6.2.2 and let $r_1 \leqslant \frac{r}{2}$ be such that $B(0, r_1) \subset U$ and $L_\pi L \sup_{\|y\| \leqslant r_1} \|\nabla_M f(y)\| < \frac{c_m^2}{2}$. Consider y, δ as in the lemma. By Proposition 6.2.3 there is $x_0 \in B(y, \frac{\delta}{2})$ s.t. $f(x_0) < f(P_M(x_0))$. Let $\mathsf{x} : \mathbb{R}_+ \to \mathbb{R}^d$ be in $\mathsf{S}_{-\partial f}(x_0)$ and define $t_{r_1} := \inf\{t \geqslant 0 : \|\mathsf{x}(t)\| \geqslant r_1\}$. By Proposition 6.2.2 we have that $t_{r_1} < +\infty$. For $t \leqslant t_{r_1}$, we have:

$$f(\mathsf{x}(t)) \leqslant f(\mathsf{x}(0)) - c_m^2 t\,.$$

Denoting $\mathsf{y}(t) = P_M(\mathsf{x}(t))$, for $t \leqslant t_{r_1}$, by path-differentiability of f we have:

$$f(\mathsf{y}(t)) = f(\mathsf{y}(0)) - \int_0^t \langle \nabla_M f(\mathsf{y}(t)), \dot{\mathsf{y}}(t) \rangle \, \mathrm{d}t\,.$$

Since $\|\dot{\mathsf{y}}(t)\| \leqslant L_\pi L$, for $t \leqslant t_{r_1}$, we have:

$$f(\mathsf{y}(0)) \leqslant f(\mathsf{y}(t)) + LL_\pi t \sup_{\|y\| \leqslant r_1} \|\nabla_M f(y)\| \leqslant f(\mathsf{y}(t)) + \frac{c_m^2}{2} t\,.$$

Therefore, for $t \leqslant t_{r_1}$,

$$f(\mathsf{x}(t)) \leqslant f(\mathsf{x}(0)) - c_m^2 t \leqslant f(\mathsf{y}(0)) - c_m^2 t \leqslant f(\mathsf{y}(t)) - \frac{c_m^2}{2} t\,. \tag{6.16}$$

This implies that $\mathsf{y}(t_{r_1}) = P_M(\mathsf{x}(t_{r_1})) \neq \mathsf{x}(t_{r_1})$ and, therefore, $\|\mathsf{x}(t_{r_1})\| = r_1$. Hence,

$$\|x_0 - y\| \leqslant \frac{\delta}{2} < \delta < r_1 - \|y\| = \|\mathsf{x}(t_r)\| - \|y\| \leqslant \|\mathsf{x}(t_r) - y\|\,.$$

Since x is continuous, this implies that there is $t_\delta \in (0, t_r)$ s.t. $\|\mathsf{x}(t_\delta) - y\| = \delta$. Denote $x = \mathsf{x}(t_\delta)$ and notice that $\|\mathsf{x}(t_\delta) - x_0\| \leqslant Lt_\delta$. Hence,

$$\delta = \|\mathsf{x}(t_\delta) - y\| \leqslant \|x_0 - y\| + \|\mathsf{x}(t_\delta) - x_0\| \leqslant \frac{\delta}{2} + Lt_\delta\,.$$

Therefore, $t_\delta \geqslant \frac{\delta}{2L}$ and by Inequality (6.16) we have:

$$\begin{aligned} f(x) &\leqslant f(P_M(x)) - t_\delta \frac{c_m^2}{2} \\ &\leqslant f(P_M(x)) - \frac{c_m^2 \delta}{4L}\,, \end{aligned}$$

which proves the first statement.

Finally, for $x' \in B(x, \delta')$, we have:

$$\begin{aligned} f(x') &\leqslant f(x) + L\left\|x - x'\right\| \\ &\leqslant f(P_M(x)) - \frac{c_m^2 \delta}{4L} + L\delta' \\ &\leqslant f(P_M(x')) + (L + L_M)\delta' - \frac{c_m^2 \delta}{4L}\,. \end{aligned}$$

With our choice of δ' the last inequality implies that $f(x') < f(P_M(x'))$. □

Lemma 6.4.10. *Let Assumptions 6.3.1–6.3.4 hold. Consider κ from Proposition 6.3.2 and r_1 from Lemma 6.4.9. For n large enough, there is $\varrho_{\kappa, r_1} > 0$ such that:*

$$\mathbb{P}\left([f(x_{n+1}) < f(P_M(x_{n+1}))] | \mathscr{F}_n\right) \mathbb{1}_{\operatorname{dist}(x_n, M) \leqslant \kappa\gamma_n} \mathbb{1}_{\|P_M(x_n - \gamma_n v_n)\| \leqslant \frac{1}{2} r_1} \mathbb{1}_{\|x_n\| \leqslant r_1} \geqslant \varrho_{\kappa, r_1}\,.$$

Proof. The set-valued mapping $\Psi_n : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$, defined as:

$$\Psi_n(x) = \left\{ x : \|x - P_M(x_n - \gamma_n v_n)\| = \gamma_n, f(x) \leqslant f(P_M(x)) - \frac{c_m^2}{4L}\gamma_n \right\}\,,$$

is closed valued and, by Lemma 6.4.9, for n large enough, it is nonempty on $[\|P_M(x_n - \gamma_n v_n)\| \leqslant \frac{1}{2} r_1]$. Hence, by [Rockafellar & Wets 1998, Corollary 14.6], we can choose x'_n in a measurable way such that $x'_n \in \Psi(x_n)$. Define $\psi_n : \mathbb{R}^d \to \mathbb{R}$ as $\psi_n(x) = \frac{x'_n - x}{\gamma_n} + v_n$ and $\delta' = \frac{c_m^2}{4(L+L_M)L}$. By Lemma 6.4.9 we have that for all $x' \in B(x'_n, \gamma_n \delta')$:

$$f(x') < f(P_M(x'))\,.$$

Denote S^u the set $\{x : f(x) < f(P_M(x))\}$. On the event $[\operatorname{dist}(x_n, M) \leqslant \gamma_n \kappa] \cap [\|P_M(x_n - \gamma_n v_n)\| \leqslant \frac{r_1}{2}] \cap [\|x_n\| \leqslant r_1]$, using Assumption 6.3.4, we have:

$$\begin{aligned}
\mathbb{P}(x_{n+1} \in S^u | \mathscr{F}_n) &\geqslant \mathbb{P}(x_n - \gamma_n v_n + \gamma_n \eta_{n+1} \in B(x'_n, \gamma_n \delta') | \mathscr{F}_n) \\
&\geqslant \mathbb{P}(\gamma_n \eta_{n+1} \in B(x'_n - x_n + \gamma_n v_n, \gamma_n \delta') | \mathscr{F}_n) \\
&\geqslant \mathbb{P}(\eta_{n+1} \in B(\psi_n(x_n), \delta') | \mathscr{F}_n) \\
&\geqslant \int_{u \in \mathbb{R}^d} \mathbb{1}_{B(\psi_n(x_n), \delta')}(u) h_{r_1}(u)\, \mathrm{d}u
\end{aligned} \tag{6.17}$$

Denote $L' = L + 1 + L_\pi L + \kappa$. We have:

$$\begin{aligned}
\|\psi_n(x_n)\| &\leqslant \|v_n\| + \frac{\|x'_n - P_M(x_n - \gamma_n v_n)\| + \|P_M(x_n - \gamma_n v_n) - P_M(x_n)\| + \|x_n - P_M(x_n)\|}{\gamma_n} \\
&\leqslant L + 1 + L_\pi L + \kappa = L'\,.
\end{aligned}$$

Therefore, by Assumption 6.3.4, on the event $[\operatorname{dist}(x_n, M) \leqslant \gamma_n \kappa] \cap [\|P_M(x_n - \gamma_n v_n)\| \leqslant \frac{r_1}{2}] \cap [\|x_n\| \leqslant r_1]$ we have:

$$\mathbb{P}(x_{n+1} \in S^u | \mathscr{F}_n) \geqslant \inf_{\|x\| \leqslant L' + \delta'} h_{r_1}(x) \int_{\mathbb{R}^d} \mathbb{1}_{B(0,\delta')}(u)\, \mathrm{d}u\,.$$

Since h_{r_1} is positive and continuous, the infimum in the last inequality is positive, which finishes the proof. □

To finish the proof of Theorem 6.3.3 we will use the following lemma.

Lemma 6.4.11 ([Borkar 2008, Chapter 4, Lemma 14])**.** *Consider $(\Omega, \mathcal{A}, \mathbb{P})$ a probability space and $(\mathscr{F}_n)$ a filtration. Let $(F_n), (H_n)$ be two sequences of events adapted to $(\mathscr{F}_n)$ and assume that there is a constant $C > 0$ such that:*

$$\mathbb{P}(F_{n+1} | \mathscr{F}_n) \mathbb{1}_{H_n} \geqslant C\,.$$

Then

$$\mathbb{P}([F_n \text{ occurs infinitely often}]^c \cap [H_n \text{ occurs infinitely often }]) = 0\,.$$

By Proposition 6.3.2 we know that on the event $[x_n \to 0]$ the event $[\operatorname{dist}(x_n, M) \leqslant \gamma_n \kappa] \cap [\|P_M(x_n - \gamma_n v_n)\| \leqslant \frac{1}{2} r_1]$ will happen infinitely often. Therefore, by Lemmas 6.4.11 and 6.4.10 the event $[f(x_n) < f(P_M(x_n))]$ happens infinitely often. By Proposition 6.3.1 this can happen only with probability zero. Hence, $\mathbb{P}([x_n \to 0]) = 0$.

6.4.5 Proof of Proposition 6.3.6

The validity of Assumption 6.3.2 is immediate. To prove Assumption 6.3.4 denote $Q : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to \mathbb{R}_+$ the Markov kernel of η_{n+1}. For every $(x, A) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$, it is defined as:

$$Q(x, A) = \mathbb{P}(\eta_{n+1} \in A | \tilde{x}_n = x) = \int \mathbb{1}_A(G(x) - g(x, s))\mu(\mathrm{d}s)\,.$$

The Markov kernel of $\tilde{\eta}_{n+1}$, denoted $\tilde{Q}$, is then defined for every $(x, A) \in \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d)$, as:

$$\begin{aligned}\tilde{Q}(x, A) = \mathbb{P}(\tilde{\eta}_{n+1} \in A | \tilde{x}_n = x) &= \int_{z\in\mathbb{R}^d} \int_{y\in\mathbb{R}^d} \mathbb{1}_A(y + z)Q(x, \mathrm{d}y)h(z)\,\mathrm{d}z \\ &= \int_{u\in\mathbb{R}^d} \mathbb{1}_A(u) \int_{y\in\mathbb{R}^d} h^1(u - y)Q(x, \mathrm{d}y)\,\mathrm{d}u\,.\end{aligned}$$

Fix $C > 0$, notice that if $\|\tilde{x}_n\| \leqslant C$, then by Markov's inequality we have for $a \geqslant \sqrt{2C}$,

$$\int_{\|y\|\leqslant a} Q(\tilde{x}_n, \mathrm{d}y) = \mathbb{P}(\|\eta_{n+1}\| \leqslant a | \widetilde{\mathscr{F}_n}) \geqslant 1 - \frac{K(C)}{a^2} \geqslant \frac{1}{2}\,.$$

Therefore, for a measurable function $\psi : \mathbb{R}^d \to \mathbb{R}^d$ and $C, \delta > 0$, if $\|x_n\| \leqslant C$, then:

$$\mathbb{P}(\tilde{\eta}_{n+1} \in B(\psi(\tilde{x}_n), \delta) | \widetilde{\mathscr{F}_n}) \geqslant \frac{1}{2} \int_{\mathbb{R}^d} \mathbb{1}_{B(\psi(x_n),\delta)}(u) \inf_{\|y\|\leqslant\sqrt{2C}} h^1(u - y)\,\mathrm{d}u\,.$$

A simple exercise shows that the function $u \mapsto \inf_{\|y\|\leqslant\sqrt{2C}} h^1(u - y)$ is continuous and positive. Hence, Assumption 6.3.4 is verified.

Equation

Chapter 7

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

7.1 Introduction

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

$$x_{n+1} \in \operatorname{prox}_{g,\mathcal{X}}^{\gamma_n}(x_n - \gamma_n v_n + \gamma_n \eta_{n+1}), \tag{7.1}$$

where $\operatorname{prox}_{g,\mathcal{X}}^{\gamma_n}$ is the proximal operator for the function g on $\mathcal{X}$ (see Equation (7.7) for a definition), (γ_n) is a sequence of stepsizes, (η_n) is a noise sequence and for each $n \in \mathbb{N}$, v_n is in the set $\partial f(x_n)$ of Clarke's subgradients of f at x_n.

Let $\mathcal{N}_{\mathcal{X}}(x)$ be the normal cone of $\mathcal{X}$ at x. It is known (see [Davis *et al.* 2020, Majewski *et al.* 2018]) that, under mild conditions on f, g and (η_n), every limit point of (x_n) is included in the set $\mathcal{Z} := \{x : 0 \in \partial f(x) + \partial g(x) + \mathcal{N}_{\mathcal{X}}(x)\}$. The proof leans on the seminal paper of Benaïm, Hofbauer and Sorin [Benaïm *et al.* 2005] (see also [Benaïm 1999]), which analyzes Equation (7.1) as an Euler-like discretization of the differential inclusion (DI):

$$\dot{\mathsf{x}}(t) \in -\partial f(\mathsf{x}(t)) - \partial g(\mathsf{x}(t)) - \mathcal{N}_{\mathcal{X}}(\mathsf{x}(t)). \tag{7.2}$$

While the sequence (x_n) is known to converge to $\mathcal{Z}$, recent work [Rios-Zertuche 2020] shows that in principle, it might not converge to a unique point. In [Rios-Zertuche 2020, Section 2] Ríos-Zertuche considers the deterministic subgradient descent (that is to say $g = 0$, $\mathcal{X} = \mathbb{R}^d$, $\eta_n = 0$) and constructs f, which verifies main assumptions of nonsmooth optimization (such as Whitney stratifiability or Kurdyka-Łojasiewicz inequality) but the limit set of (x_n) is equal to $\mathcal{Z} = \{x : \|x\| = 1\}$. This encourages a more precise study of Equation (7.1).

In [Bolte *et al.* 2020b] the authors, using the theory of closed measures, show that in the case of the deterministic subgradient descent the convergence to $\mathcal{Z}$ arises in a structured manner. First, they prove that if x, y are two distinct accumulation points of (x_n), then the time that the iterates spend to get from a neighborhood of x to a neighborhood of y goes to infinity. Second, in a first approximation their

results imply that if x is an accumulation point of (x_n), then

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

(see [Bolte *et al.* 2020b, Theorem 7] or Section 7.3 for a precise statement). Intuitively speaking, this means that even if $x_n - x_0 = \sum_{i=0}^n \gamma_i v_i$ does not converge, on average, the drift coming from the subgradients compensate itself and vanishes at infinity. This behavior captures an oscillation phenomenon of the iterates around the critical set. Results of this type show a strong stability property of the deterministic subgradient descent.

In practical settings, when the function f is either unknown or computation of its gradient is expensive, the deterministic gradient descent is often replaced by its stochastic version, in many cases, this may lead to a faster convergence (see e.g. [Bottou *et al.* 2018]). Proximal methods, on the other hand, along with the regularizer function g, are widely used to regularize the initial problem of minimizing f. Depending on the choice of g, we can, for instance, preserve the boundedness of the iterates [Duchi & Ruan 2018] or promote the sparsity of solutions [Tibshirani 1996]. It is therefore interesting to establish stability results of the type [Bolte *et al.* 2020b] for the SPGD.

In this chapter we investigate further the questions of oscillations of the SPGD. Our contributions are threefold. First, we show that the time spent by the SPGD to move from one accumulation point to another goes to infinity. Second, we establish an oscillation-type behavior of the drift. These two results extend [Bolte *et al.* 2020b, Theorem 7.] to a stochastic and a proximal setting. Finally, our technique of proof doesn't rely on the theory of closed measures used in [Bolte *et al.* 2020b] but is build upon the classical work of Benaïm, Hofbauer and Sorin [Benaïm *et al.* 2005]. We feel that this approach gives a simpler proof and allows us to treat the deterministic, the stochastic and the proximal cases in a unified manner.

Chapter organization. In Section 7.2, we recall some known facts about the DI (7.2) and its Lyapounov function. Our main results are given in Section 7.3. Section 7.4 is devoted to proofs.

7.2 Preliminaries

7.2.1 Notations

For $S \subset \mathbb{R}^d$, we denote $\operatorname{cl} S$ its closure and $\operatorname{conv} S$ its closed convex hull. For a function $F : \mathbb{R}^d \to \mathbb{R}$, we denote ∇F its gradient. Constants will usually be denoted as $C, C_1, C_2 \ldots$, they can change from line to line. For a sequence (x_n), we denote $\operatorname{acc}\{x_n\}$ its set of accumulation points. The space of continuous functions from $\mathbb{R}_+$ to $\mathbb{R}^d$ will be denoted as $\mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$, we endow this set with $\boldsymbol{d}_C$ the metric of uniform convergence on compact intervals (see Section 2.2.1). Given a convex set $\mathcal{X} \subset \mathbb{R}^d$,

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

$$\mathcal{N}_{\mathcal{X}}(x) = \{v : \langle v, y - x\rangle \leqslant 0, \forall y \in \mathcal{X}\}\,. \tag{7.3}$$

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

7.2.2 A Lyapounov function for the differential inclusion

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

$$(f \circ \mathsf{x})'(t) = \langle v, \dot{\mathsf{x}}(t)\rangle \quad \forall v \in \partial f(\mathsf{x}(t)). \tag{7.4}$$

By [Bolte & Pauwels 2019, Proposition 2], every convex, concave, semialgebraic or definable function is path differentiable. Moreover, if another function $g\colon \mathbb{R}^d \to \mathbb{R}$ is path differentiable, then $f + g$ is also path differentiable [Bolte & Pauwels 2019, Corollary 4]. From a similar point of view, if $\mathcal{X}$ is a convex set, then for any a.c. curve $\mathsf{x} : [0,1] \to \mathbb{R}^d$, for almost every $t \in [0,1]$:

$$\langle v, \dot{\mathsf{x}}(t)\rangle = 0 \quad \forall v \in \mathcal{N}_{\mathcal{X}}(\mathsf{x}(t))\,. \tag{7.5}$$

Consider now $f, g : \mathbb{R}^d \to \mathbb{R}$ path differentiable, $\mathcal{X} \subset \mathbb{R}^d$ a convex set and x a solution to the DI (7.2). Using Equation (7.4) and (7.5) and the fact that $\partial(f+g) \subset \partial f + \partial g$, we obtain

$$(f+g)(\mathsf{x}(t)) - (f+g)(\mathsf{x}(0)) = -\int_0^t \|\dot{\mathsf{x}}(u)\|^2 \,\mathrm{d}u\,. \tag{7.6}$$

This implies that $(f+g)(\mathsf{x}(t)) < (f+g)(\mathsf{x}(0))$ if $\mathsf{x}(0) \notin \mathcal{Z}$. In other words, $f+g$ is a strict Lyapounov function for the DI (7.2).

7.3 Main results

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

$$\mathrm{prox}^{\gamma}_{g,\mathcal{X}}(x) = \arg\min_{y \in \mathcal{X}} \{g(y) + \frac{1}{2\gamma}\|y - x\|^2\}\,. \tag{7.7}$$

We study Equation (7.1) under the following assumptions.

Assumption 7.3.1.

i) *The set $\mathcal{X}$ is a closed convex subset of $\mathbb{R}^d$.*

ii) *The functions $f, g : \mathbb{R}^d \to \mathbb{R}$ are locally Lipschitz continuous.*

iii) *There is a filtration $(\mathscr{F}_n)_{n\in\mathbb{N}}$, such that (η_n) is a martingale difference sequence adapted to it, and x_n is $\mathscr{F}_n$ measurable for every $n \in \mathbb{N}$.*

iv) *The sequence of stepsizes (γ_n) is nonnegative and such that $\sum_{i=0}^{+\infty} \gamma_i = +\infty$.*

Note that if g is nonconvex, $\operatorname{prox}^{\gamma}_{g,\mathcal{X}}(x)$ is a set in $\mathbb{R}^d$. Assumption 7.3.1-(iii)) then implicitly states that x_{n+1} is chosen in a measurable manner, such a choice is always possible (see e.g. [Davis *et al.* 2020]). By [Rockafellar & Wets 1998, 10.2 and 10.10], we can rewrite Equation (7.1) as:

$$x_{n+1} = x_n - \gamma_n(v_n + v_n^g + v_n^{\mathcal{X}}) + \gamma_n \eta_{n+1}\,, \tag{7.8}$$

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

Assumption 7.3.2.

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

ii) There is $q \geqslant 2$ *such that*

$$\sum_{i=0}^{+\infty} \gamma_i^{1+q/2} < +\infty\,, \tag{7.9}$$

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

$$\sup_{n\in\mathbb{N}} \mathbb{E}[\|\eta_{n+1}\|^q \mathbb{1}_{x_n\in\mathcal{K}} | \mathscr{F}_n] < +\infty\,. \tag{7.10}$$

Assumptions of this type are standard in the field of stochastic approximation. Assumption 7.3.2-(i)) prevent the algorithm to diverge. Note that it is superfluous if $\mathcal{X}$ is compact. Otherwise it can be obtained by a proper choice of the regularizer g (see [Duchi & Ruan 2018]).

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

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

Following [Benaïm *et al.* 2005] we will show that X is an APT of the DI (7.2). The next two assumptions ensure us that $f+g$ will be a Lyapounov function for the DI (7.2).

Assumption 7.3.3. *The functions* f *and* g *are path differentiable.*

Assumption 7.3.4. *The set of Clarke critical values* $\{f(x)+g(x) : x \in \mathcal{Z}\}$ *has an empty interior.*

Assumption 7.3.4 is a classical Sard-type condition. It ensures the fact that if x is a solution to the DI (7.2), with $\mathsf{x}(0) \in \mathcal{Z}$, then x is constant. As established in [Bolte *et al.* 2007], it is satisfied as soon as f, g and $\mathcal{X}$ are definable.

The next two propositions are not new and can be found in one way or another in e.g. [Davis *et al.* 2020, Bolte & Pauwels 2019, Majewski *et al.* 2018, Bolte *et al.* 2020b]. Nevertheless, since our set of assumptions is slightly different and their proof is a simple application of Section 2.2.2, for completeness, we include it in Section 7.4.1.

Proposition 7.3.1. *Let Assumptions 7.3.1 and 7.3.2 hold, then the family* $(\mathsf{X}(t+\cdot))_{t\geqslant 0}$ *is relatively compact. Moreover, if a sequence* $t_n \to +\infty$ *and* $\mathsf{x} \in \mathcal{C}(\mathbb{R}_+, \mathbb{R}^d)$ *is such that* $\boldsymbol{d}_C(\mathsf{X}(t_n+\cdot), \mathsf{x}) \to 0$*, then* x *is a solution to the DI (7.2).*

Proposition 7.3.2. *Under Assumptions 7.3.1–7.3.4, the set* $\operatorname{acc}\{x_n\}$ *is included in* $\mathcal{Z}$ *and* $f+g$ *is constant on* $\operatorname{acc}\{x_n\}$*.*

The next theorem tells us that even if $\operatorname{acc}\{x_n\}$ is not a single point, the time that it takes to (x_n) to go from one accumulation point to another goes to infinity. This is an extension of [Bolte *et al.* 2020b, Theorem 6.i), Theorem 7.i)], to the best of our knowledge this result is new in a stochastic and proximal setting.

Theorem 7.3.3. *Let Assumptions 7.3.1–7.3.4 hold. Let* x, y *be two distinct points in* $\operatorname{acc}\{x_n\}$*. Consider two sequences* n_i, n_j*, with* $n_i \leqslant n_j$*, such that* $x_{n_i} \to x$ *and* $x_{n_j} \to y$*. Then* $\tau_{n_j} - \tau_{n_i} \to +\infty$*.*

Under Assumptions 7.3.1–7.3.3, the same result is true if $(f+g)(x) \leqslant (f+g)(y)$*.*

As it is shown in [Rios-Zertuche 2020], it is possible that $\operatorname{acc}\{x_n\}$ is not reduced to a unique point. Nevertheless, Theorem 7.3.3 implies that the "nonconvergence" happens in a very slow manner. Asymptotically, the time spent by the algorithm to move from one accumulation point to another goes to infinity.

We now investigate the question of oscillations. Given U, V two open sets, such that $\operatorname{cl} U \subset V$, we will call $I = [n_1, n_2]$ a maximal interval related to U, V if the set $X_{n_1}^{n_2} := \{x_{n_1}, x_{n_1+1}, \dots, x_{n_2}\}$ is such that $X_{n_1}^{n_2} \subset V$, $X_{n_1}^{n_2} \cap U \neq \varnothing$ and either x_{n_1-1} or x_{n_2+1} is not in V. The next two results are an extension of [Bolte *et al.* 2020b, Theorem 7] to a stochastic setting.

Theorem 7.3.4 (Long intervals)**.** *Let Assumptions 7.3.1-7.3.4 hold. Consider* $x \in \operatorname{acc}\{x_n\}$ *and* U, V *two neighborhoods of* x *such that* $\operatorname{cl} U \subset V$*. For* $i \in \mathbb{N}$*, denote* $I_i = [n_{1i}, n_{2i}]$ *a sequence of distinct maximal intervals related to* U, V*. Then, either one of* I_i *is unbounded or* $\tau_{n_{2i}} - \tau_{n_{1i}} \to +\infty$*.*

Theorem 7.3.5 (Oscillation compensation)**.** *Let Assumptions 7.3.1-7.3.4 hold, and fix* U*,* V *and* I_i *as in Theorem 7.3.4. Denote* $A = \bigcup I_i$*, then*

$$\frac{\sum_{i=1}^n \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_A(x_i)}{\sum_{i=1}^n \gamma_i \mathbb{1}_A(x_i)} \xrightarrow[n\to+\infty]{} 0\,. \tag{7.11}$$

Theorem 7.3.5 gives an intuitive explanation of why Theorem 7.3.3 holds. Indeed, while the drift coming from one iteration $v_i + v_i^g + v_i^{\mathcal{X}}$ might not go to zero (as it happens for such a simple example as $f(x) = \|x\|$, $g = 0$ and $\mathcal{X} = \mathbb{R}^d$), on average, it compensates itself. Theorem 7.3.3 and 7.3.5 suggest that the algorithm oscillates around its accumulation set, while the center of these oscillations moves in $\operatorname{acc}\{x_n\}$ with a vanishing speed.

Let us finish with a remark on the Equation (7.11). At first sight, maximal intervals in Theorem 7.3.5 and Theorem 7.3.4 may seem artificial. A more satisfactory result would be

$$\frac{\sum_{i=1}^n \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_U(x_i)}{\sum_{i=1}^n \gamma_i \mathbb{1}_U(x_i)} \xrightarrow[n\to+\infty]{} 0\,, \tag{7.12}$$

where U is an open neighborhood of an accumulation point x. Looking at the proof of Theorem 7.3.5, to obtain Equation (7.12), we could think of defining maximal intervals as $I_i = [n_{1i}, n_{2i}]$ such that $\{x_{n_{1i}}, \dots, x_{n_{2i}}\} \subset U$ and $x_{n_{1i}-1}, x_{n_{2i}+1} \notin U$. Unfortunately, for this type of intervals we dont have an equivalent of Theorem 7.3.4, i.e. it may very well be that the quantity $\tau_{n_{2i}} - \tau_{n_{1i}}$ is bounded. Actually, it is not very hard to show, that for the function from [Rios-Zertuche 2020, Section 2], there are x, U such that Equation (7.12) is false.

Nevertheless, as explained in [Bolte *et al.* 2020b], Equation (7.11) is a good approximation of Equation (7.12). Indeed, apply Theorem 7.3.5 with U and $V = U^\delta$, where $U^\delta = \{y \in \mathbb{R}^d : \exists z \in U, \|z - y\| < \delta\}$, then, as an approximation, we have

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

7.4 Proofs

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

$$N(T, n) = \inf\{j \geqslant n \text{ s.t. } \tau_j - \tau_n \geqslant T\}\,. \tag{7.13}$$

7.4.1 Proof of Proposition 7.3.1 and 7.3.2

To put ourselves in the context of Section 2.2.2 we need to alter the map $-\partial f - \partial g - \mathcal{N}_{\mathcal{X}}$ in a way that it verifies assumptions of Propositions 2.2.2 and 2.2.3. While this section is slightly technical, conceptually, we just find a set-valued map G verifying assumptions of Proposition 2.2.3 and s.t. $x_{n+1} \in G(x_n)$. This is done using the Lipschitz continuity of f, g and the boundedness of (x_n). A convinced reader may want to skip to Section 7.4.2.
We start with two technical lemmas.

Lemma 7.4.1. *Under Assumptions 7.3.1 and 7.3.2, almost surely, for every $T > 0$, we have:*

$$\lim_{n \to +\infty} \sup_{n \leqslant j \leqslant N(T,n)} \left\| \sum_{i=n}^{j} \gamma_i \eta_{i+1} \right\| = 0\,. \tag{7.14}$$

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

Proof. Indeed, since almost surely $\sup \|x_n\| < +\infty$, for each $\delta > 0$, there is $C > 0$ s.t. if we denote $A = \{\forall n \in \mathbb{N} \|x_n\| \leqslant C\}$, then $\mathbb{P}(A) > 1 - \delta$. Define $\tilde{\eta}_{n+1} = \eta_{n+1}\mathbb{1}_{\|x_n\| \leqslant C}$, then $\mathbb{E}[\tilde{\eta}_{n+1} | \mathscr{F}_n] = 0$ and $\sup_{n \in \mathbb{N}} \mathbb{E}[\|\tilde{\eta}_{n+1}\|^q] < +\infty$. Hence, by [Benaïm 1999, Proposition 4.2], we have $\sup_{n \leqslant j \leqslant N(T,n)} \left\| \sum_{i=n}^{j} \gamma_i \tilde{\eta}_{i+1} \right\| \xrightarrow[n \to +\infty]{} 0$. Since δ is arbitrary, Equation (7.14) follows. □

Lemma 7.4.2. *Let Assumptions 7.3.1 and 7.3.2 hold. Let $A \in \Xi$ be a probability one set on which (x_n) and $(x_{n+1/2})$ are bounded, and let C be a random variable s.t. $\|x_n\| < C$ and C is finite valued on A. Then for each $\omega \in A$, there are two globally*

Lipschitz functions $\tilde{g}, \tilde{f} : \mathbb{R}^d \to \mathbb{R}$ *and a bounded set-valued map* $\widetilde{\mathcal{N}}_{\mathcal{X}} : \mathbb{R}^d \rightrightarrows \mathbb{R}^d$ *s.t. in Equation* (7.8) *we have* $v_n(w) \in \partial\tilde{f}(x_n(w))$, $v_n^g(w) \in \partial\tilde{g}(x_{n+1}(w))$ *and* $v_n^{\mathcal{X}}(w) \in \widetilde{\mathcal{N}}_{\mathcal{X}}(x_{n+1}(w))$.
Moreover, if x *is a solution to the DI:*

$$\dot{\mathsf{x}}(t) \in -\partial\tilde{f}(\mathsf{x}(t)) - \partial\tilde{g}(\mathsf{x}(t)) - \widetilde{\mathcal{N}}_{\mathcal{X}}(\mathsf{x}(t))\,, \tag{7.15}$$

and that x *remains in* $B(0,C) \cap \mathcal{X}$, *then* x *is a solution to the DI* (7.2).
Finally, denoting $\widetilde{\mathcal{Z}} = \{x : 0 \in \partial\tilde{f}(x) + \partial\tilde{g}(x) + \widetilde{\mathcal{N}}_{\mathcal{X}}(x)\}$, *we have the equality* $\widetilde{\mathcal{Z}} \cap B(0,C) = \mathcal{Z} \cap B(0,C)$.

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

We say that an a.c. curve $\mathsf{x} : \mathbb{R}_+ \to \mathbb{R}^d$ is a *perturbed solution* to the DI (7.15) if there is $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ and a locally integrable function $\mathsf{b} : \mathbb{R}_+ \to \mathbb{R}^d$ s.t. for almost every $t \geqslant 0$, we have:

$$\dot{\mathsf{x}}(t) - \rho(t) \in -\partial\tilde{f}^{\mathsf{b}(t)}(\mathsf{x}(t)) - \partial\tilde{g}^{\mathsf{b}(t)}(\mathsf{x}(t)) - \widetilde{\mathcal{N}}_{\mathcal{X}}^{\mathsf{b}(t)}(\mathsf{x}(t))\,,$$

where $\mathsf{H}^{\delta}(x) = \{v \in \mathsf{H}(y) : \|y - x\| \leqslant \delta\}$, $\lim_{t\to+\infty} \mathsf{b}(t) = 0$ and for every $T > 0$, we have:

$$\lim_{t\to 0} \sup_{0\leqslant h\leqslant T} \left\| \int_t^{t+h} \rho(u) \right\| \mathrm{d}u = 0\,.$$

If x is a bounded perturbed solution to (7.15), then by [Benaïm *et al.* 2005, Theorem 4.2] it is also an APT of (7.15). Thus, to prove Proposition 7.3.1 it remains to show that X is a perturbed solution to the DI (7.15).

For $t \in [\tau(n), \tau(n+1))$, we define $\rho(t) = \eta_{n+1}$ and $\mathsf{b}(t) = \|x_{n+1} - x_n\|$. The condition on ρ immediately follows from Lemma 7.4.1. The condition on b follows from the following lemma.

Lemma 7.4.3. *Under Assumptions 7.3.1 and 7.3.2, almost surely, we have that* $\|x_{n+1} - x_n\| \xrightarrow[n\to+\infty]{} 0$.

Proof. By Lemma 7.4.1, we have that $\left\|x_{n+1/2} - x_n\right\| \xrightarrow[n\to+\infty]{} 0$. Moreover, we have:

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

Therefore,

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

and

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

which finishes the proof. □

To finish the proof of Proposition 7.3.1 consider $t_n \to +\infty$ and x s.t. $\boldsymbol{d}_C(\mathsf{X}(t_n + \cdot), \mathsf{x}) \to 0$. Then, by [Benaïm *et al.* 2005, Theorem 4.2], x is a solution to the DI (7.15), moreover, it remains in $B(0, C) \cap \mathcal{X}$, therefore, it is also a solution to the DI (7.2).

For the proof of Proposition 7.3.2, notice that $\tilde{f} + \tilde{g}$ is path differentiable (as a composition of path differentiable functions). Then, in the same way as in Section 7.2.2, we have that $\tilde{f} + \tilde{g}$ is a strict Lyapounov function for the DI (7.15) and for the set $\tilde{\mathcal{Z}}$. Since $\operatorname{acc}\{x_n\} = \mathsf{L}_{\mathsf{X}} \subset \operatorname{cl} B(0, C)$, by Proposition 2.2.3 we have that $\mathsf{L}_{\mathsf{X}} \subset \widetilde{\mathcal{Z}} \cap \operatorname{cl} B(0, C) \subset \mathcal{Z}$, and that $f + g$ is constant on $\operatorname{acc}\{x_n\}$.

Remark 22. *Strictly speaking, following [Benaïm et al. 2005], a perturbed solution to the DI is of the form* $\dot{\mathsf{x}}(t) - \rho(t) \in \mathsf{H}^{\mathsf{b}(t)}(\mathsf{x}(t))$, *where* $\mathsf{H} = -\partial\tilde{f} - \partial\tilde{g} - \tilde{\mathcal{N}}_{\mathcal{X}}$. *Nevertheless, the proof of [Benaïm et al. 2005, Theorem 4.2] goes through with our definition.*

7.4.2 Proof of Theorem 7.3.3

Lemma 7.4.4. *Let Assumptions 7.3.1– 7.3.3 hold, let* τ_n *be a positive sequence, with* $\tau_n \to +\infty$, *and* x *s.t.* $\mathsf{X}(\tau_n + \cdot) \to \mathsf{x}$, *then*

$$(f + g)(\mathsf{x}(h)) \leqslant (f + g)(\mathsf{x}(0)), \quad \forall h \in \mathbb{R}_+ . \tag{7.16}$$

Moreover, if for some $h \geqslant 0$, $(f + g)(\mathsf{x}(h)) = (f + g)(\mathsf{x}(0))$, *then* $\mathsf{x}(h') = \mathsf{x}(0)$ *for every* $h' \in [0, h]$. *If additionally Assumption 7.3.4 holds, then:*

$$\mathsf{x}(h) = \mathsf{x}(0), \quad \forall h \in \mathbb{R}_+ . \tag{7.17}$$

Proof. By Proposition 7.3.1, x is a solution to the DI (7.2), and the first result follows by Equation (7.6).
Under Assumption 7.3.4, we have that $\mathsf{x}(\mathbb{R}_+) \subset \operatorname{acc}\{x_n\} \subset \mathcal{Z}$, hence, by Proposition 7.3.2, we have that $(f + g) \circ \mathsf{x}$ is constant. Using Assumption 7.3.3, we have for all $h \in \mathbb{R}_+$,

$$0 = (f + g)(\mathsf{x}(h)) - (f + g)(\mathsf{x}(0)) = -\int_0^h \|\dot{\mathsf{x}}(u)\|^2 \, \mathrm{d}u . \tag{7.18}$$

This implies that $\int_0^h \|\dot{\mathsf{x}}(u)\|^2 \, \mathrm{d}u = 0$. Hence, $\dot{\mathsf{x}}(h) = 0$ for almost every $h \in [0, T]$ and we obtain Equation (7.17).

□

Suppose that there is $T > 0$ such that $\tau_{n_j} - \tau_{n_i} \leqslant T$. The sequence $\mathsf{X}(\tau(n_j) + \cdot)$ is relatively compact, and after extraction it converges to x a solution to (7.2). Extract once again to have $\tau_{n_j} - \tau_{n_i} \to h$. Then

$$\mathsf{X}(\tau(n_j)) - \mathsf{X}(\tau(n_i)) \to \mathsf{x}(h) - \mathsf{x}(0) = y - x ,$$

and we obtain a contradiction with Lemma 7.4.4.

7.4.3 Proof of Theorem 7.3.4

The next lemma is the key ingredient for the proofs of Theorem 7.3.4 and Theorem 7.3.5.

Lemma 7.4.5. *Under Assumptions 7.3.1–7.3.4, we have*

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

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

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

This implies:

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

Extract a sequence such that $\mathsf{X}(\tau_{n_k} + \cdot)$ converges to x and $\tau_{j_k} - \tau_{n_k} \to h$, with $h \leqslant T$. Then $x_{j_k} \to \mathsf{x}(T')$ and $x_{n_k} \to \mathsf{x}(0)$, but $\|\mathsf{x}(T') - \mathsf{x}(0)\| \geqslant \varepsilon$ which is impossible by Lemma (7.4.4). □

Suppose that no I_i is unbounded, then we can choose $n_i \in I_i = [n_{1i}, n_{2i}]$ such that $x_{n_i} \in U$. Since $x_{n_{2i+1}}$ is in V^c, after extraction $x_{n_i} \to y_1$ and $x_{n_{2i+1}} \to y_2$, with $y_2 \neq y_1$, moreover:

$$\tau_{n_{2i+1}} - \tau_{n_i} - \gamma_{n_{2i+1}} \leqslant \tau_{n_{2i}} - \tau_{n_{1i}} \,. \tag{7.19}$$

By Theorem 7.3.3, the first term of this inequality tends to infinity.

7.4.4 Proof of Theorem 7.3.5

Take I_i as in Theorem 7.3.4, and $A_N = \bigcup_{i\leqslant N} I_i$. Define

$$u_N = \frac{a_N}{b_N} = \frac{\sum_{i=0}^{+\infty} \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_{A_N}(x_i)}{\sum_{i=0}^{+\infty} \gamma_i \mathbb{1}_{A_N}(x_i)} \,.$$

Then,

$$u_{N+1} = \frac{a_n + \sum_{i=0}^{+\infty} \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}}) \mathbb{1}_{I_{N+1}}(x_i)}{b_N + \sum_{i=0}^{+\infty} \gamma_i \mathbb{1}_{I_{N+1}}(x_i)} \,. \tag{7.20}$$

Fix $\varepsilon > 0$, by Lemma 7.4.5, there is n_0 such that, for $n_k \geqslant n_0$ and $j_k \leqslant N(T, n_k)$, $\left\| \sum_{i=n_k}^{j_k} \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}}) \right\| \leqslant \varepsilon$. Decompose $I_i = [n_{1i}, n_{2i}] = \bigcup_{1\leqslant k\leqslant K_i} [a_{ki}, a_{ki+1}]$, with $a_{1i} = n_{1i}$ and $a_{ki+1} = \min\{N(T, a_{ki}), n_{2i}\}$. We obtain:

$$\begin{aligned} u_{N+1} &= \frac{a_N + \sum_{k\leqslant K_N} \sum_{i=a_{kN}}^{a_{kN+1}} \gamma_i (v_i + v_i^g + v_i^{\mathcal{X}})}{b_N + \sum_{k\leqslant K_N} \sum_{i=a_{kN}}^{a_{kN+1}} \gamma_i} \\ &\leqslant \frac{a_N + (K_N)\varepsilon}{b_N + (K_N - 1)T} \,. \end{aligned}$$

By Theorem 7.3.4, we have that $K_N \to +\infty$ and, therefore, for N large enough:

$$u_{N+1} \leqslant \frac{a_N + 2(K_N - 1)\varepsilon}{b_N + (K_N - 1)T}\,.$$

Hence, by induction:

$$u_{N+j} \leqslant \frac{a_N + 2\varepsilon \sum_{k=N}^{N+j-1}(K_i - 1)}{b_N + T\sum_{k=N}^{N+j-1}(K_N - 1)}\,.$$

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

Bibliography

[Alacaoglu *et al.* 2020a] A. Alacaoglu, Y. Malitsky and V. Cevher. *Convergence of adaptive algorithms for weakly convex constrained optimization.* arXiv preprint arXiv:2006.06650, 2020. (Cited on page 43.)

[Alacaoglu *et al.* 2020b] A. Alacaoglu, Y. Malitsky, P. Mertikopoulos and V. Cevher. *A new regret analysis for Adam-type algorithms.* In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pages 202–210. PMLR, 13–18 Jul 2020. (Cited on page 43.)

[Aliprantis & Border 2006] C. D. Aliprantis and K. C. Border. Infinite dimensional analysis: a hitchhiker's guide. Springer, Berlin; London, 2006. (Cited on page 106.)

[Alizadeh *et al.* 1995] F. Alizadeh, Jean-Pierre Haeberly and Michael Overton. *Complementarity and Nondegeneracy in Semidefinite Programming.* Mathematical Programming, vol. 77, 05 1995. (Cited on page 143.)

[Alvarez 2000] F. Alvarez. *On the minimizing property of a second order dissipative system in Hilbert spaces.* SIAM Journal on Control and Optimization, vol. 38, no. 4, pages 1102–1119, 2000. (Cited on page 49.)

[Assran & Rabbat 2020] M. Assran and M. Rabbat. *On the Convergence of Nesterov's Accelerated Gradient Method in Stochastic Settings.* In Hal Daumé III and Aarti Singh, editors, Proceedings of the 37th International Conference on Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pages 410–420, Virtual, 13–18 Jul 2020. PMLR. (Cited on page 43.)

[Attouch *et al.* 2000] H. Attouch, X. Goudou and P. Redont. *The heavy ball with friction method, I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system.* Communications in Contemporary Mathematics, vol. 2, no. 01, pages 1–34, 2000. (Cited on pages 4 and 33.)

[Attouch *et al.* 2011] Hedy Attouch, Jérôme Bolte and Benar Fux Svaiter. *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods.* Mathematical Programming, Series A, vol. 137, no. 1, pages 91–124, August 2011. (Cited on pages 3, 25 and 145.)

[Attouch *et al.* 2018] H. Attouch, Z. Chbani, J. Peypouquet and P. Redont. *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity.* Mathematical Programming, vol. 168, no. 1-2, pages 123–175, 2018. (Cited on page 35.)

[Aubin & Cellina 1984] J.-P. Aubin and A. Cellina. Differential inclusions, volume 264 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1984. Set-valued maps and viability theory. (Cited on pages 21 and 106.)

[Aubin *et al.* 1991] J.-P. Aubin, H. Frankowska and A. Lasota. *Poincaré's recurrence theorem for set-valued dynamical systems*. Ann. Polon. Math., vol. 54, no. 1, pages 85–91, 1991. (Cited on pages 97 and 110.)

[Audin *et al.* 2014] M. Audin, M. Damian and R. Erné. Morse theory and floer homology. Universitext (Berlin. Print). Springer, 2014. (Cited on page 149.)

[Aujol *et al.* 2019] J.-F. Aujol, Ch. Dossal and A. Rondepierre. *Optimal Convergence Rates for Nesterov Acceleration*. SIAM Journal on Optimization, vol. 29, no. 4, pages 3131–3153, 2019. (Cited on page 35.)

[Barakat & Bianchi 2021] A. Barakat and P. Bianchi. *Convergence and Dynamical Behavior of the ADAM Algorithm for Nonconvex Stochastic Optimization*. SIAM Journal on Optimization, vol. 31, no. 1, pages 244–274, 2021. (Cited on pages 4, 5, 31, 32, 33, 35, 42, 43, 44 and 49.)

[Barakat *et al.* 2021] Anas Barakat, Pascal Bianchi, Walid Hachem and Sholom Schechtman. *Stochastic optimization with momentum: Convergence, fluctuations, and traps avoidance*. Electron. J. Statist., vol. 15, no. 2, pages 3892–3947, 2021. (Cited on page 148.)

[Belotto da Silva & Gazeau 2018] A. Belotto da Silva and M. Gazeau. *A general system of differential equations to model first order adaptive algorithms*. arXiv, pages arXiv–1810, 2018. (Cited on page 3.)

[Belotto da Silva & Gazeau 2020] A. Belotto da Silva and M. Gazeau. *A General System of Differential Equations to Model First-Order Adaptive Algorithms*. Journal of Machine Learning Research, vol. 21, no. 129, pages 1–42, 2020. (Cited on pages 4, 31, 33, 34, 35, 36, 49 and 51.)

[Benaïm & Hirsch 1996] M. Benaïm and M. W. Hirsch. *Asymptotic pseudotrajectories and chain recurrent flows, with applications*. J. Dynam. Differential Equations, vol. 8, no. 1, pages 141–176, 1996. (Cited on page 56.)

[Benaïm *et al.* 2005] M. Benaïm, J. Hofbauer and S. Sorin. *Stochastic approximations and differential inclusions*. SIAM J. Control Optim., vol. 44, no. 1, pages 328–348 (electronic), 2005. (Cited on pages 2, 14, 21, 22, 87, 135, 137, 163, 173, 174, 176, 179 and 180.)

[Benaïm 1996] M. Benaïm. *A dynamical system approach to stochastic approximations*. SIAM J. Control Optim., vol. 34, no. 2, pages 437–472, 1996. (Cited on page 56.)

[Benaïm 1999] M. Benaïm. *Dynamics of stochastic approximation algorithms.* In Séminaire de Probabilités, XXXIII, volume 1709 of *Lecture Notes in Math.*, pages 1–68. Springer, Berlin, 1999. (Cited on pages 1, 2, 18, 19, 20, 34, 48, 56, 57, 58, 60, 148, 173 and 178.)

[Benveniste *et al.* 1990] A. Benveniste, M. Métivier and P. Priouret. Adaptive algorithms and stochastic approximations, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson. (Cited on pages 5 and 87.)

[Bianchi *et al.* 2019] P. Bianchi, W. Hachem and A. Salim. *Constant step stochastic approximations involving differential inclusions: stability, long-run convergence and applications.* Stochastics, vol. 91, no. 2, pages 288–320, 2019. (Cited on pages 106, 109 and 115.)

[Bianchi *et al.* 2021a] Pascal Bianchi, Walid Hachem and Sholom Schechtman. *Convergence of constant step stochastic gradient descent for non-smooth non-convex functions*, 2021. (Cited on pages 117 and 157.)

[Bianchi *et al.* 2021b] Pascal Bianchi, Walid Hachem and Sholom Schechtman. *Stochastic Subgradient Descent Escapes Active Strict Saddles*, 2021. (Cited on pages 145, 146, 147, 148, 150, 151, 154, 156 and 164.)

[Bierstone & Milman 1988] Edward Bierstone and Pierre Milman. *Semianalytic and subanalytic sets.* Publications Mathématiques de l'IHÉS, vol. 67, pages 5–42, 1988. (Cited on page 26.)

[Bolte & Pauwels 2019] J. Bolte and E. Pauwels. *Conservative set valued fields, automatic differentiation, stochastic gradient method and deep learning.* arXiv preprint arXiv:1909.10300, 2019. (Cited on pages 3, 6, 18, 25, 86, 88, 90, 92, 98, 117, 121, 175 and 176.)

[Bolte *et al.* 2007] J. Bolte, A. Daniilidis, A. Lewis and M. Shiota. *Clarke subgradients of stratifiable functions.* SIAM Journal on Optimization, vol. 18, no. 2, pages 556–572, 2007. (Cited on pages 3, 9, 10, 18, 25, 29, 94, 97, 98, 117, 120, 145 and 176.)

[Bolte *et al.* 2009] Jérôme Bolte, Aris Daniilidis and Adrian Lewis. *Tame functions are semismooth.* Math. Program., vol. 117, pages 5–19, 03 2009. (Cited on pages 3, 25, 26, 145 and 159.)

[Bolte *et al.* 2011] Jérôme Bolte, Aris Daniilidis and Adrian Lewis. *Generic Optimality Conditions for Semialgebraic Convex Programs.* Math. Oper. Res., vol. 36, pages 55–70, 02 2011. (Cited on page 145.)

[Bolte *et al.* 2020a] Jérôme Bolte, Lilian Glaudin, Edouard Pauwels and Mathieu Serrurier. *A Hölderian backtracking method for min-max and min-min problems.* CoRR, vol. abs/2007.08810, 2020. (Cited on page 145.)

[Bolte *et al.* 2020b] Jerome Bolte, Edouard Pauwels and Rodolfo Rios-Zertuche. *Long term dynamics of the subgradient method for Lipschitz path differentiable functions*, 2020. (Cited on pages 14, 173, 174, 176, 177 and 178.)

[Borkar 2008] V. S. Borkar. Stochastic approximation. Cambridge University Press, Cambridge; Hindustan Book Agency, New Delhi, 2008. A dynamical systems viewpoint. (Cited on pages 1, 2, 157 and 169.)

[Borwein & Wang 1998] Jonathan (Jon) Borwein and Xianfu Wang. *Lipschitz Functions with Maximal Clarke Subdifferentials Are Generic*. Proceedings of the American Mathematical Society, vol. 128, 10 1998. (Cited on page 2.)

[Bottou *et al.* 2018] Léon Bottou, Frank E. Curtis and Jorge Nocedal. *Optimization Methods for Large-Scale Machine Learning*, 2018. (Cited on pages 1 and 174.)

[Boumal 2020] Nicolas Boumal. An introduction to optimization on smooth manifolds. Available online, Nov 2020. (Cited on pages 22, 23 and 24.)

[Brandière & Duflo 1996] O. Brandière and M. Duflo. *Les algorithmes stochastiques contournent-ils les pièges?* Ann. Inst. H. Poincaré Probab. Statist., vol. 32, no. 3, pages 395–427, 1996. (Cited on pages 1, 2, 5, 8, 9, 44, 48, 70, 71, 76, 79, 118, 119, 126, 127, 128, 129, 137, 138, 139, 140, 141, 144, 147, 148 and 156.)

[Bravo *et al.* 2018] M. Bravo, D.S. Leslie and P. Mertikopoulos. *Bandit learning in concave N-person games*, 2018. (Cited on page 134.)

[Cabot *et al.* 2009] A. Cabot, H. Engler and S. Gadat. *On the long time behavior of second order differential equations with asymptotically small dissipation.* Transactions of the American Mathematical Society, vol. 361, no. 11, pages 5983–6017, 2009. (Cited on pages 35, 37 and 53.)

[Chen *et al.* 2018] J. Chen, D. Zhou, Y. Tang, Z. Yang and Q. Gu. *Closing the generalization gap of adaptive gradient methods in training deep neural networks.* arXiv preprint arXiv:1806.06763, 2018. (Cited on page 43.)

[Chen *et al.* 2019] X. Chen, S. Liu, R. Sun and M. Hong. *On the Convergence of A Class of Adam-Type Algorithms for Non-Convex Optimization.* In International Conference on Learning Representations, 2019. (Cited on page 43.)

[Chung 1954] K.-L. Chung. *On a stochastic approximation method.* The Annals of Mathematical Statistics, vol. 25, no. 3, pages 463–483, 1954. (Cited on page 134.)

[Clarke *et al.* 1998] 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. (Cited on pages 6, 17, 18, 87, 88, 97, 103, 110, 111 and 112.)

[Coste 2002] M. Coste. *AN INTRODUCTION TO O-MINIMAL GEOMETRY*. 2002. (Cited on pages 25, 26, 27, 28 and 29.)

[Dalec′kiĭ & Kreĭn 1974] Ju. L. Dalec′kiĭ and M. G. Kreĭn. Stability of solutions of differential equations in Banach space. American Mathematical Society, Providence, R.I., 1974. Translated from the Russian by S. Smith, Translations of Mathematical Monographs, Vol. 43. (Cited on pages 4, 32, 48 and 71.)

[Daniilidis & Drusvyatskiy 2019] Aris Daniilidis and Dmitriy Drusvyatskiy. *Pathological subgradient dynamics*, 2019. (Cited on pages 2 and 117.)

[Daniilidis & Flores 2019] Aris Daniilidis and Gonzalo Flores. *Linear structure of functions with maximal Clarke subdifferential*. SIAM Journal on Optimization, vol. 29, no. 1, pages 511–521, 2019. (Cited on page 2.)

[Davis & Drusvyatskiy 2021] Damek Davis and Dmitriy Drusvyatskiy. *Proximal methods avoid active strict saddles of weakly convex functions*, 2021. (Cited on pages 3, 7, 8, 22, 118, 119, 120, 122, 123, 124, 125, 144, 145, 148 and 151.)

[Davis *et al.* 2020] D. Davis, D. Drusvyatskiy, S. Kakade and J. D. Lee. *Stochastic Subgradient Method Converges on Tame Functions*. Found Comput Math, no. 20, pages 119–154, 2020. (Cited on pages 3, 7, 13, 18, 25, 26, 29, 86, 87, 88, 92, 97, 98, 110, 111, 117, 145, 173 and 176.)

[Davis *et al.* 2021] Damek Davis, Dmitriy Drusvyatskiy and Liwei Jiang. *Subgradient methods near active manifolds: saddle point avoidance, local convergence, and asymptotic normality*, 2021. (Cited on pages 12, 13 and 148.)

[De *et al.* 2018] S. De, A. Mukherjee and E. Ullah. *Convergence guarantees for RMSProp and ADAM in non-convex optimization and their comparison to Nesterov acceleration on autoencoders*. arXiv preprint arXiv:1807.06766, 2018. (Cited on page 43.)

[Défossez *et al.* 2020] A. Défossez, L. Bottou, F. Bach and N. Usunier. *A Simple Convergence Proof of Adam and Adagrad*. arXiv preprint arXiv:2003.02395, 2020. (Cited on page 43.)

[Delyon *et al.* 1999] B. Delyon, M. Lavielle and E. Moulines. *Convergence of a stochastic approximation version of the EM algorithm*. Annals of statistics, pages 94–128, 1999. (Cited on pages 66, 67 and 68.)

[Drusvyatskiy & Lewis 2010a] Dmitriy Drusvyatskiy and Adrian Lewis. *Semi-algebraic functions have small subdifferentials*. Mathematical Programming, vol. 140, 04 2010. (Cited on page 25.)

[Drusvyatskiy & Lewis 2010b] Dmitriy Drusvyatskiy and Adrian Lewis. *Semi-algebraic functions have small subdifferentials*. Mathematical Programming, vol. 140, 04 2010. (Cited on page 159.)

[Drusvyatskiy & Lewis 2012] Dmitriy Drusvyatskiy and Adrian Lewis. *Tilt Stability, Uniform Quadratic Growth, and Strong Metric Regularity of the Subdifferential*. SIAM Journal on Optimization, vol. 23, 04 2012. (Cited on page 149.)

[Drusvyatskiy & Lewis 2014] D. Drusvyatskiy and A. Lewis. *Optimality, identifiability, and sensitivity*. Mathematical Programming, vol. 147, pages 467–498, 2014. (Cited on pages 11, 145 and 149.)

[Drusvyatskiy *et al.* 2016] Dmitriy Drusvyatskiy, Alexander D. Ioffe and Adrian S. Lewis. *Generic Minimizing Behavior in Semialgebraic Optimization*. SIAM J. Optim., vol. 26, no. 1, pages 513–534, 2016. (Cited on pages 8, 12, 119, 120, 123, 124, 145, 151, 158 and 159.)

[Du *et al.* 2017] S. S. Du, C. Jin, J. D. Lee, M. I. Jordan, A. Singh and B. Poczos. *Gradient Descent Can Take Exponential Time to Escape Saddle Points*. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, editors, Advances in Neural Information Processing Systems 30, pages 1067–1077. Curran Associates, Inc., 2017. (Cited on page 48.)

[Duchi & Ruan 2018] John Duchi and Feng Ruan. *Stochastic Methods for Composite and Weakly Convex Optimization Problems*, 2018. (Cited on pages 136, 137, 174 and 176.)

[Duchi *et al.* 2011] J. Duchi, E. Hazan and Y. Singer. *Adaptive subgradient methods for online learning and stochastic optimization*. Journal of Machine Learning Research, vol. 12, no. Jul, pages 2121–2159, 2011. (Cited on page 40.)

[Faure & Roth 2013] M. Faure and G. Roth. *Ergodic properties of weak asymptotic pseudotrajectories for set-valued dynamical systems*. Stoch. Dyn., vol. 13, no. 1, pages 1250011, 23, 2013. (Cited on pages 97 and 110.)

[Folland 2013] G.B. Folland. Real analysis: Modern techniques and their applications. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013. (Cited on page 101.)

[Gabrielov 1968] Andrei M Gabrielov. *Projections of semi-analytic sets*. Functional Analysis and its applications, vol. 2, no. 4, pages 282–291, 1968. (Cited on page 26.)

[Gabrielov 1996] Andrei Gabrielov. *Complements of subanalytic sets and existential formulas for analytic functions*. Inventiones mathematicae, vol. 125, no. 1, pages 1–12, 1996. (Cited on page 26.)

[Gadat & Gavra 2020] S. Gadat and I. Gavra. *Asymptotic study of stochastic adaptive algorithm in non-convex landscape*. arXiv preprint arXiv:2012.05640, 2020. (Cited on pages 40, 43, 48 and 148.)

[Gadat *et al.* 2018] S. Gadat, F. Panloup and S. Saadane. *Stochastic heavy ball.* Electron. J. Stat., vol. 12, no. 1, pages 461–529, 2018. (Cited on pages 4, 5, 32, 42, 43, 44, 48 and 54.)

[Haraux 1991] A. Haraux. Systèmes dynamiques dissipatifs et applications, volume 17. Masson, 1991. (Cited on pages 52 and 56.)

[Hartman 2002] P. Hartman. Ordinary differential equations. Society for Industrial and Applied Mathematics, second édition, 2002. (Cited on page 50.)

[Has'minskiĭ 1963] R. Z. Has'minskiĭ. *The averaging principle for parabolic and elliptic differential equations and Markov processes with small diffusion.* Teor. Verojatnost. i Primenen., vol. 8, pages 3–25, 1963. (Cited on page 97.)

[Horn & Johnson 1994] R. A. Horn and C. R. Johnson. Topics in matrix analysis. Cambridge University Press, Cambridge, 1994. Corrected reprint of the 1991 original. (Cited on page 71.)

[Ioffe 2009] A. D. Ioffe. *An Invitation to Tame Optimization.* SIAM J. on Optimization, vol. 19, no. 4, page 1894–1917, February 2009. (Cited on page 25.)

[Jin *et al.* 2017] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade and M. I. Jordan. *How to Escape Saddle Points Efficiently.* volume 70 of *Proceedings of Machine Learning Research*, pages 1724–1732. PMLR, 2017. (Cited on page 48.)

[Kakade & Lee 2018] S. Kakade and J. D. Lee. *Provably Correct Automatic Sub-Differentiation for Qualified Programs.* In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 7125–7135. Curran Associates, Inc., 2018. (Cited on page 92.)

[Karatzas & Shreve 1991] I. Karatzas and S.E. Shreve. Brownian motion and stochastic calculus. Springer-Verlag, second édition, 1991. (Cited on page 65.)

[Kingma & Ba 2015] D. P. Kingma and J. Ba. *Adam: A method for stochastic optimization.* In International Conference on Learning Representations, 2015. (Cited on pages 4, 31 and 33.)

[Kloeden & Rasmussen 2011] P. E. Kloeden and M. Rasmussen. Nonautonomous dynamical systems, volume 176 of *Mathematical Surveys and Monographs.* American Mathematical Society, Providence, RI, 2011. (Cited on pages 4, 32, 48, 71, 72, 75 and 76.)

[Kurdyka 1998] Krzysztof Kurdyka. *On gradients of functions definable in o-minimal structures.* Annales de l'Institut Fourier, vol. 48, no. 3, pages 769–783, 1998. (Cited on page 25.)

[Kushner & Yin 2003] H. J. Kushner and G. G. Yin. Stochastic approximation and recursive algorithms and applications, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second édition, 2003. Stochastic Modelling and Applied Probability. (Cited on pages 1, 2, 5, 87 and 157.)

[Lafontaine 2015] Jacques Lafontaine. An introduction to differential manifolds. Springer International Publishing, 2015. (Cited on pages 22 and 23.)

[Lebourg 1979] G. Lebourg. *Generic Differentiability of Lipschitzian Functions*. Transactions of the American Mathematical Society, vol. 256, pages 125–144, 1979. (Cited on page 91.)

[Lee *et al.* 2016] Jason D. Lee, Max Simchowitz, Michael I. Jordan and Benjamin Recht. *Gradient Descent Only Converges to Minimizers*. In Vitaly Feldman, Alexander Rakhlin and Ohad Shamir, editors, 29th Annual Conference on Learning Theory, volume 49 of *Proceedings of Machine Learning Research*, pages 1246–1257, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. (Cited on pages 118 and 148.)

[Lee *et al.* 2019] J. D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan and B. Recht. *First-order methods almost always avoid strict saddle points*. Math. Program., vol. 176, no. 1-2, Ser. B, pages 311–337, 2019. (Cited on page 48.)

[Lewis & Malick 2008] Adrian Lewis and Jerome Malick. *Alternating Projections on Manifolds*. Mathematics of Operations Research, vol. 33, 02 2008. (Cited on page 24.)

[Lewis 2002] A. Lewis. *Active Sets, Nonsmoothness, and Sensitivity*. SIAM J. Optim., vol. 13, pages 702–725, 2002. (Cited on pages 122 and 149.)

[Loi 1998] Ta Loi. *Verdier and strict Thom stratifications in o-minimal structures*. Illinois Journal of Mathematics - ILL J MATH, vol. 42, 06 1998. (Cited on pages 10, 25, 28, 30, 119 and 120.)

[Mai & Johansson 2020] V. V. Mai and M. Johansson. *Convergence of a Stochastic Gradient Method with Momentum for Nonsmooth Nonconvex Optimization*. Proceedings of Machine Learning Research. PMLR, 2020. (Cited on page 43.)

[Majewski *et al.* 2018] S. Majewski, B. Miasojedow and E. Moulines. *Analysis of nonsmooth stochastic approximation: the differential inclusion approach*. arXiv preprint arXiv:1805.01916, 2018. (Cited on pages 6, 7, 86, 87, 88, 117, 144, 173 and 176.)

[Mertikopoulos *et al.* 2020a] P. Mertikopoulos, N. Hallak, A. Kavis and V. Cevher. *On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Problems*. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan

and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 1117–1128. Curran Associates, Inc., 2020. (Cited on page 48.)

[Mertikopoulos *et al.* 2020b] Panayotis Mertikopoulos, Nadav Hallak, Ali Kavis and Volkan Cevher. *On the Almost Sure Convergence of Stochastic Gradient Descent in Non-Convex Problems*, 2020. (Cited on pages 148 and 156.)

[Métivier & Priouret 1987] M. Métivier and P. Priouret. *Théorèmes de convergence presque sûre pour une classe d'algorithmes stochastiques à pas décroissant.* Probability Theory and Related Fields, vol. 74, no. 3, pages 403–428, Sep 1987. (Cited on page 57.)

[Meyn & Tweedie 2009] S. Meyn and R. L. Tweedie. Markov chains and stochastic stability. Cambridge University Press, New York, NY, USA, 2nd édition, 2009. (Cited on pages 98 and 107.)

[Nocedal & Wright 2006] Jorge Nocedal and Stephen J. Wright. Numerical optimization. Springer, New York, NY, USA, second édition, 2006. (Cited on page 143.)

[Panageas & Piliouras 2017] I. Panageas and G. Piliouras. *Gradient Descent Only Converges to Minimizers: Non-Isolated Critical Points and Invariant Regions.* In ITCS, 2017. (Cited on page 48.)

[Panageas *et al.* 2019] I. Panageas, G. Piliouras and X. Wang. *First-order methods almost always avoid saddle points: The case of vanishing step-sizes.* In Advances in Neural Information Processing Systems 32, pages 6474–6483, 2019. (Cited on page 48.)

[Paszke *et al.* 2017] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer. *Automatic differentiation in PyTorch.* In NIPS-W, 2017. (Cited on pages 6 and 92.)

[Pelletier 1998] M. Pelletier. *Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing.* Annals of Applied Probability, pages 10–44, 1998. (Cited on pages 64 and 65.)

[Pemantle 1990] R. Pemantle. *Nonconvergence to unstable points in urn models and stochastic approximations.* Ann. Probab., vol. 18, no. 2, pages 698–712, 1990. (Cited on pages 1, 2, 8, 44, 48, 70, 71, 118, 119, 144, 148 and 156.)

[Pötzsche & Rasmussen 2006] C. Pötzsche and M. Rasmussen. *Taylor approximation of integral manifolds.* J. Dynam. Differential Equations, vol. 18, no. 2, pages 427–460, 2006. (Cited on page 72.)

[Rios-Zertuche 2020] Rodolfo Rios-Zertuche. *Examples of pathological dynamics of the subgradient method for Lipschitz path-differentiable functions*, 2020. (Cited on pages 13, 173, 177 and 178.)

[Robbins & Monro 1951] Herbert Robbins and Sutton Monro. *A Stochastic Approximation Method*. The Annals of Mathematical Statistics, vol. 22, no. 3, pages 400 – 407, 1951. (Cited on page 1.)

[Robbins & Siegmund 1971] 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. (Cited on pages 61 and 63.)

[Rockafellar & Wets 1998] R. T. Rockafellar and R. J.-B. Wets. Variational analysis, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998. (Cited on pages 17, 152, 169 and 176.)

[Rockafellar 1981] R. T. Rockafellar. *Favorable Classes of Lipschitz Continuous Functions in Subgradient Optimization*. 1981. (Cited on page 2.)

[Roth & Sandholm 2013] G. Roth and W. H. Sandholm. *Stochastic approximations with constant step size and differential inclusions*. SIAM J. Control Optim., vol. 51, no. 1, pages 525–555, 2013. (Cited on pages 97 and 108.)

[Schechtman 2021a] Sh. Schechtman. *Stochastic proximal subgradient descent oscillates in the vicinity of its accumulation Set*, 2021. (Cited on pages 136, 137 and 163.)

[Schechtman 2021b] Sholom Schechtman. *Stochastic Subgradient Descent on a Generic Definable Function Converges to a Minimizer*, 2021. (Cited on page 148.)

[Simon & Saigal 1973] Carl Simon and Romesh Saigal. *Generic properties of the complementarity problem*. Mathematical Programming, vol. 4, 12 1973. (Cited on page 143.)

[Spingarn & Rockafellar 1979] J. E. Spingarn and R. T. Rockafellar. *The Generic Nature of Optimality Conditions in Nonlinear Programming*. Mathematics of Operations Research, vol. 4, no. 4, pages 425–430, 1979. (Cited on page 143.)

[Su *et al.* 2016a] W. Su, S. Boyd and E. J. Candès. *A Differential Equation for Modeling Nesterov's Accelerated Gradient Method: Theory and Insights*. Journal of Machine Learning Research, vol. 17, no. 153, pages 1–43, 2016. (Cited on page 4.)

[Su *et al.* 2016b] Weijie Su, Stephen Boyd and Emmanuel J. Candès. *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*. J. Mach. Learn. Res., vol. 17, pages Paper No. 153, 43, 2016. (Cited on pages 35, 37 and 54.)

[Tibshirani 1996] Robert Tibshirani. *Regression Shrinkage and Selection via the Lasso*. Journal of the Royal Statistical Society. Series B (Methodological), vol. 58, no. 1, pages 267–288, 1996. (Cited on page 174.)

[Tieleman & Hinton 2012] T. Tieleman and G. Hinton. *Lecture 6.e-rmsprop: Divide the gradient by a running average of its recent magnitude*. Coursera: Neural networks for machine learning, pages 26–31, 2012. (Cited on pages 35 and 40.)

[van den Dries & Miller 1996] L. van den Dries and C. Miller. *Geometric categories and o-minimal structures*. Duke Math. J., vol. 84, no. 2, pages 497–540, 08 1996. (Cited on pages 25, 26, 27 and 29.)

[Victor 1974] Guillemin Victor. Differential topology / victor guillemin,... alan pollack,... Prentice-Hall, Englewood Cliffs (N.J.), C 1974. (Cited on pages 23 and 149.)

[Wilkie 2009] Alex J. Wilkie. *O-minimal structures*. In Séminaire Bourbaki Volume 2007/2008 Exposés 982-996, number 326 de Astérisque. Société mathématique de France, 2009. talk:985. (Cited on page 26.)

[Yan *et al.* 2018] Y. Yan, T. Yang, Z. Li, Q. Lin and Y. Yang. *A unified analysis of stochastic momentum methods for deep learning*. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, pages 2955–2961, 2018. (Cited on page 43.)

[Zaheer *et al.* 2018] M. Zaheer, S. Reddi, D. Sachan, S. Kale and S. Kumar. *Adaptive methods for nonconvex optimization*. In Advances in Neural Information Processing Systems, pages 9793–9803, 2018. (Cited on page 43.)

[Zhou *et al.* 2018] D. Zhou, Y. Tang, Z. Yang, Y. Cao and Q. Gu. *On the Convergence of Adaptive Gradient Methods for Nonconvex Optimization*. arXiv preprint arXiv:1808.05671, 2018. (Cited on page 43.)

[Zou *et al.* 2019] F. Zou, L. Shen, Z. Jie, W. Zhang and W. Liu. *A sufficient condition for convergences of adam and rmsprop*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 11127–11135, 2019. (Cited on page 43.)