Journal of Machine Learning Research 23 (2022) 1-54 Submitted 12/20; Revised 7/22; Published 8/22

Asymptotic Study of Stochastic Adaptive Algorithms in Non-convex Landscape

Sébastien Gadat SEBASTIEN.GADAT@TSE-FR.EU
Toulouse School of Economics
Université Toulouse I Capitole
Esplanade de l'Université, 31080 Toulouse, France
Institut Universitaire de France

Ioana Gavra IOANA.GAVRA@UNIV-RENNES2.FR
IRMAR, Université de Rennes 2
Place du recteur Henri Le Moal
35043 Rennes, France

Editor: Karthik Sridharan

Abstract

This paper studies some asymptotic properties of adaptive algorithms widely used in optimization and machine learning, and among them Adagrad and Rmsprop, which are involved in most of the blackbox deep learning algorithms. Our setup is the non-convex landscape optimization point of view, we consider a one time scale parametrization and the situation where these algorithms may or may not be used with mini-batches. We adopt the point of view of stochastic algorithms and establish the almost sure convergence of these methods when using a decreasing step-size towards the set of critical points of the target function. With a mild extra assumption on the noise, we also obtain the convergence towards the set of minimizers of the function. Along our study, we also obtain a "convergence rate" of the methods, in the vein of the works of Ghadimi and Lan (2013).

Keywords: Stochastic optimization; Stochastic adaptive algorithm; Convergence of random variables.

1. Introduction

Minimizing a differentiable non-convex function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ when f is defined through an expected loss in a statistical model is a common way of estimating from an empirical set of observations in machine learning problems. In particular, some difficult optimization is generally involved in neural networks learning, we refer to Bottou et al. (2018), where the major challenge of a such problem is the large scale statistical settings (large number of observations n involved in the definition of f and large dimension of the ambient space d) and the non-convex landscape property when using a cascade of logistic regressions. We consider in this work the generic formulation:

$$\forall \theta \in \mathbb{R}^d : \qquad f(\theta) = \mathbb{E}_{X\sim\mathbb{P}}[\tilde{f}(\theta, X)],$$

where X is a random variable sampled according to an *unknown* distribution $\mathbb{P}$. To perform the optimization of f under the uncertainty on $\mathbb{P}$, we assume that we can compute all along

the process of our algorithm some noisy but unbiased approximations of the gradient of f computed at the current point of the algorithm. One typical example of a such algorithm is the so-called Stochastic Gradient Descent (SGD) introduced in the famous work of Robbins and Monro (1951), which is governed by the stochastic evolution:

$$\theta_{k+1} = \theta_k - \gamma_{k+1} \nabla_\theta \tilde{f}(\theta_k, X_{k+1}),$$

where $(\gamma_k)_{k\geq 1}$ is a well chosen step sequence and the observations $(X_k)_{k\geq 1}$ are some random realizations identically distributed according to the distribution $\mathbb{P}$. The early success of this algorithm in the sixties has been at least rejuvenated if not resurrected with the development of massive learning problems, in the last fifteen years. We refer among others to Bottou and Bousquet (2008); Moulines and Bach (2011) or to Bach (2014) and the references therein for various applications in machine learning. Although being one of the state-of-the-art methods to handle massive datasets, SGD suffers from several issues: difficulty to tune the step-size sequence or dependence on the gradient flow that may be lazy in flat areas, which is especially the case when looking at non-convex neural network problems.

Some popular improvements are commonly patched to SGD, and among others we refer to the popular acceleration obtained with the Polyak-Ruppert averaging studied by Polyak and Juditsky (1992); Ruppert (1988); Moulines and Bach (2011); Cardot et al. (2017); Gadat and Panloup (2020), and to the variance reduction with mini-batch strategies in Le Roux et al. (2012); Johnson and Zhang (2013).

While these two last improvements do not modify the underlying gradient flow, other strategies rely on a modification of the dynamical system exploiting acceleration brought by momentum with additional second order terms. The first historical example is the Heavy Ball with Friction optimization based on the contribution Polyak (1964) and then translated into a stochastic framework in Gadat et al. (2018); Sebbouh et al. (2020); Loizou and Richtárik (2020). Another example is the Nesterov Accelerated Gradient Descent of Nesterov (1983) studied in the noisy situation recently in a large number of works, among others by Ghadimi and Lan (2016); Jin et al. (2018).

Alternative recent strategies used to improve the behaviour of stochastic algorithms rely on adaptive methods: they consist in tuning the step-size sequence either with a per-coordinate strategy or with a matricial inversion in front of the gradient $\nabla_\theta \tilde{f}(\theta_k, X_{k+1})$. Among others, Adagrad introduced by Duchi et al. (2011) (with a long range memory of past gradients) and Rmsprop (with an exponential moving average) taught by Hinton et al. (2012) are typical examples of step-size adaptations with second-order moments learned on-line and these two algorithms are at the core of our work. Another state-of-the-art algorithm is ADAM introduced by Kingma and Ba (2015) and used for example in GAN optimization in Goodfellow et al. (2014). These algorithms are referred to as *adaptive methods* and have encountered a striking raise of attention these recent years in machine learning (see *e.g.* Ward et al. (2019) Zou et al. (2019)). In the statistical community, stochastic Newton and stochastic Gauss-Newton methods may also be seen as adaptive algorithms with a direct matricial inversion and multiplication: these methods have shown both good theoretical and numerical abilities for regressions (Cénac et al., 2020), logistic regression (Bercu et al., 2020b), average consensus research (Loizou and Richtárik, 2020) or optimal transport problems (Bercu et al., 2022).

To the best of our knowledge, there are little convergence mathematical results on adaptive algorithms: Belotto da Silva and Gazeau (2020) studied the deterministic dynamical system behind adaptive algorithms and obtained long-time behaviour of the trajectories of the value function following the ideas of Cabot et al. (2009a); Su et al. (2016). More recently, Barakat and Bianchi (2020) (that is more closely related to our present work) obtain the almost sure convergence of their algorithms towards critical points with a parametrization that is different from ours but they leave as an open problem the important question of the convergence towards a *minimizer* of f.[1] A version of Adagrad is also studied by Li and Orabona (2019), where the authors prove an almost sure convergence of the gradients to zero under stronger assumptions on the noise sequence (bounded support when the cost function is non-convex). Finally, some recent contributions in machine learning (Ward et al., 2019; Zou et al., 2019; Défossez et al., 2020) address some "convergence" questions for adaptive algorithms with constant step-size. They provide a non-asymptotic study with a step-size that is tuned according to the finite horizon of simulation. Even though these results are of major interest from a numerical point of view, they do not really answer the question of convergence from a trajectorial point of view (see Section 2.2 below). The objective of this work may be seen as modest at the moment: we aim to study the asymptotic behaviour of Adagrad and other related methods, *i.e.* we aim to show the almost sure convergence towards a local minimizer of the objective function f. However limited at first sight, we will see that the convergence of the trajectories outside local traps is already challenging, especially when a mini-batch strategy is used.

2. Adaptive Algorithms and Main Results

The algorithms we consider in this paper use the vectorial division/multiplication notations introduced in Adagrad (see Duchi et al. (2011)) and now widely used in machine learning. The vectorial division $\frac{u}{v}$ and multiplication $u \cdot v$ are the coordinate per coordinate operations defined by:

$$\left(\frac{u}{v}\right)_i = \frac{u_i}{v_i}, \quad \text{and} \quad (u \cdot v)_i = u_i v_i \quad \forall i \in \{1, \dots, d\}$$

In the meantime, the notation $u^{\odot 2}$ corresponds to the coordinate per coordinate square:

$$\forall i \in \{1, \dots, d\} \qquad \{u^{\odot 2}\}_i = u_i^2,$$

whereas $\sqrt{u}$ denotes the coordinate per coordinate square root:

$$\left(\sqrt{u}\right)_i = \sqrt{u}_i, \quad \forall i \in \{1, \dots, d\}$$

Finally, the sum of a vector $u \in \mathbb{R}^d$ and a scalar $\varepsilon \in \mathbb{R}$ is given by:

$$(u + \varepsilon)_i = u_i + \varepsilon \quad \forall i \in \{1, \dots, d\}.$$

1. The same week we sent our paper on Arxiv, Barakat et al. (2020) also published some results on the trap avoidance of adaptive algorithms but do not consider the mini-batch effect that is known to be a crucial ingredient for the efficiency of adaptive methods. We refer to Theorem 1 and 3 below for the conditions we obtained on the mini-batch sequence.

2.1 Definition of the Methods

Following the recent work of Belotto da Silva and Gazeau (2020), we consider the joint evolution of $(\theta_n, w_n)_{n\geq 1}$ in $\mathbb{R}^d \times \mathbb{R}^d$ of a stochastic algorithm defined by:

$$\begin{cases} \theta_{n+1} = \theta_n - \gamma_{n+1} \dfrac{g_{n+1}}{\sqrt{w_n + \varepsilon}} \\ w_{n+1} = w_n + \gamma_{n+1}(p_n g_{n+1}^{\odot 2} - q_n w_n) \end{cases}, \tag{1}$$

where $\varepsilon > 0$, $(\gamma_n)_{n\geq 1}$, $(p_n)_{n\geq 1}$, $(q_n)_{n\geq 1}$ are some deterministic positive sequences and $(g_n)_{n\geq 1}$ corresponds to a noisy stochastic evaluation of the gradient of the function f, corrupted by an additive noise sequence $(\xi_{n+1})_{n\geq 1}$:

$$g_{n+1} = \nabla f(\theta_n) + \xi_{n+1}.$$

Following the initial vectorial notations, we emphasize that (1) means that for any coordinate $i \in \{1, \ldots, d\}$, $(\theta_n)_{n\geq 1}$ and $(w_n)_{n\geq 1}$ are updated according to:

$$\begin{cases} \theta_{n+1}^i = \theta_n^i - \gamma_{n+1} \dfrac{g_{n+1}^i}{\sqrt{w_n^i + \varepsilon}} \\ w_{n+1}^i = w_n^i + \gamma_{n+1}(p_n (g_{n+1}^i)^2 - q_n w_n^i) \end{cases}.$$

We could also have chosen to write the previous parametrization with $\tilde{\gamma}_{n+1} = \gamma_{n+1} p_n$ or $\tilde{\gamma}_{n+1} = \gamma_{n+1} q_n$, which corresponds to the natural time scale on the w coordinate. We have finally used the initial one of Equation (1) to obtain a more convenient description of the possible behaviour of the algorithm according to $(p_n)_{n\geq 1}$ and $(q_n)_{n\geq 1}$.

2.2 Link with Other Parametrizations

We discuss here our choice of the Adagrad/Rmsprop parametrization (1) using the one of Belotto da Silva and Gazeau (2020), and its link with the standard parametrization introduced in (Duchi et al., 2011; Hinton et al., 2012) and used in later works for ADAM (Défossez et al., 2020; Zou et al., 2019) and Adagrad (Ward et al., 2019).

2.2.1 Historical Parametrization

We have chosen to use formulation (1), which is inspired from the limiting O.D.E. of the continuous time adaptive gradient system following previous works on accelerated or second order dynamics and among other we refer to Memory gradient diffusion (Gadat and Panloup, 2014), Ruppert-Polyak averaging (Gadat and Panloup, 2020), Heavy Ball systems (Attouch et al., 2000; Cabot et al., 2009a,b; Gadat et al., 2018) or more generally Nesterov acceleration (Nesterov, 2004; Su et al., 2016; Attouch et al., 2019) and dissipative systems (Haraux, 1991; Alvarez et al., 2002).

The pioneering works of Duchi et al. (2011) and of Hinton et al. (2012) use the following parametrization that is at first sight different:

$$\begin{cases} \widetilde{\theta}_{n+1}^{Imp} = \widetilde{\theta}_n^{Imp} - \alpha \dfrac{g_{n+1}}{\sqrt{v_{n+1} + \widetilde{\varepsilon}}} \\ v_{n+1} = \beta_2 v_n + g_{n+1}^{\odot 2}, \end{cases} \tag{2}$$

for $\beta_2 \in (0,1]$ when no heavy ball momentum (Polyak, 1964) is used in the algorithm (which is also the case we are considering in this work). We observe that Algorithm (2) uses v_{n+1} instead of v_n to calculate the new position $\widetilde{\theta}_{n+1}$. Even if it has been observed that this "implicit" dependency with v_{n+1} instead of v_n improves the numerical stability of the algorithm, we will consider instead the "explicit" counterpart where $\widetilde{\theta}_{n+1}$ depends on v_n and not on v_{n+1}, which certainly does not really modify our (asymptotic) theoretical conclusions, but that permits to strongly limit the supplementary technical difficulties due to the implicit dependency of $\widetilde{\theta}_{n+1}$ with v_{n+1}. The sequence $(\widetilde{\theta}_n)_{n\geq 1}$ is defined as:

$$\begin{cases} \widetilde{\theta}_{n+1} = \widetilde{\theta}_n - \alpha \dfrac{g_{n+1}}{\sqrt{v_n + \widetilde{\varepsilon}}} \\ v_{n+1} = \beta_2 v_n + g_{n+1}^{\odot 2} \end{cases} . \tag{3}$$

The use of an implicit parametrization may help to improve the dependency with ε of the non-asymptotic upper bounds derived in Theorem 2 below.

Introducing the natural normalizing sequence $(S_n)_{n\geq 1}$, defined by $S_0 = 1$ and the following recursion:

$$S_{n+1} = \beta_2 S_n + 1,$$

we then define $\widetilde{w}_n = v_n/S_n$ and $\widetilde{\varepsilon}_n = \widetilde{\varepsilon}/S_n$ and observe that Equation (3) yields:

$$\widetilde{\theta}_{n+1} = \widetilde{\theta}_n - \frac{\alpha}{\sqrt{S_n}} \frac{g_{n+1}}{\sqrt{\widetilde{w}_n + \widetilde{\varepsilon}_n}},$$

whereas the second coordinate evolves according to:

$$\widetilde{w}_{n+1} = \frac{\beta_2 v_n + g_{n+1}^{\odot 2}}{S_{n+1}} = \frac{g_{n+1}^{\odot 2}}{S_{n+1}} + \widetilde{w}_n \frac{\beta_2 S_n}{S_{n+1}} = \widetilde{w}_n + \frac{1}{S_{n+1}} \left[g_{n+1}^{\odot 2} - \widetilde{w}_n \right].$$

Following the recommendation of Ward et al. (2019); Défossez et al. (2020), in particular (Défossez et al., 2020, Equation (2.4)), we can introduce a new step-size sequence $(\gamma_n)_{n\geq 1}$ such that $\alpha = \gamma_{n+1}\sqrt{S_n}$ and we recover in this case a joint evolution:

$$\begin{cases} \widetilde{\theta}_{n+1} = \widetilde{\theta}_n - \gamma_{n+1} \frac{g_{n+1}}{\sqrt{\widetilde{w}_n + \widetilde{\varepsilon}_n}} \\ \widetilde{w}_{n+1} = \widetilde{w}_n + \gamma_{n+1} \left[\frac{\sqrt{S_n}}{\alpha S_{n+1}} g_{n+1}^{\odot 2} - \frac{\sqrt{S_n}}{\alpha S_{n+1}} \widetilde{w}_n \right] \end{cases} . \tag{4}$$

We then deduce that Rmsprop and Adagrad with a step-size α and an hyperparameter β_2 may be embedded in the framework of Equation (1) according to the following association:

$$\gamma_{n+1} = \frac{\alpha}{\sqrt{S_n}}, \quad S_{n+1} = \beta_2 S_n + 1, \quad \epsilon_n = \frac{\epsilon}{S_n} \quad \text{and} \quad p_n = q_n = \frac{\sqrt{S_n}}{\alpha S_{n+1}}.$$

• Case of constant $\beta_2 = 1$ - Adagrad of Duchi et al. (2011). This case is certainly the easiest to understand since the natural rescaling S_n of the sequence $(v_n)_{n\geq 1}$ is $S_n = n$. In this case, we recover a joint evolution:

$$\begin{cases} \widetilde{\theta}_{n+1} = \widetilde{\theta}_n - \frac{\alpha}{\sqrt{n}} \frac{g_{n+1}}{\sqrt{\widetilde{w}_n + \widetilde{\varepsilon}_n}} \\ \widetilde{w}_{n+1} = \widetilde{w}_n + \frac{1}{n+1} [g_{n+1}^{\odot 2} - \widetilde{w}_n] \end{cases} ,$$

which entails $\gamma_{n+1} = \frac{\alpha}{\sqrt{n}}$. The evolution of $(\theta_n, \widetilde{w}_n)_{n\geq 1}$ follows a *two-time scale stochastic dynamic*, with a decaying learning rate proportional to $n^{-1/2}$ on the location and a decaying learning rate proportional to n^{-1} on $(\widetilde{w}_n)$. Nevertheless, a decreasing value of $(\alpha_n)_{n\geq 1} \propto (n^{-1/2})$ which is standard for non-convex stochastic optimization procedures then induces a unique time scale on $(\widetilde{\theta}_n, \widetilde{w}_n)_{n\geq 1}$.

• Case of constant $\beta_2 \in (0,1)$ - Rmsprop of Hinton et al. (2012) or Adam of Kingma and Ba (2015) with no momentum. Since S_n converges exponentially fast towards $(1-\beta_2)^{-1}$, the system is close to:

$$\begin{cases} \widetilde{\theta}_{n+1} = \widetilde{\theta}_n - \sqrt{1-\beta_2}\alpha \frac{g_{n+1}}{\sqrt{\widetilde{w}_n + \epsilon(1-\beta_2)}} \\ \widetilde{w}_{n+1} = \widetilde{w}_n + (1-\beta_2)[g_{n+1}^{\odot 2} - \widetilde{w}_n] \end{cases}, \tag{5}$$

which entails a constant step-size evolution whose values are $\alpha\sqrt{1-\beta_2}$ and $1-\beta_2$.

2.2.2 Other Two-time Scale Possible Parametrization

We finally consider the situation where β_2 may depend on the current iteration n, while keeping close to 1, and that may be calculated as $\beta_2(n) = 1 - bn^{-\beta}$ with $b \in (0,1)$. This last case where the sequence goes to 1 with n corresponds to an intermediary situation between $\beta_2 = 1$ and $\beta_2 < 1$, this transition being parametrized by $\beta \in [0, +\infty]$. We shall introduce the sequence of products:

$$\pi_k = \beta_2(1) \dots \beta_2(k) \qquad \text{with} \qquad \pi_0 = 1,$$

and a straightforward computation yields

$$S_{n+1} = \sum_{k=0}^{n} \pi_n \pi_k^{-1}.$$

When $\beta = 1$, using (Bercu et al., 2020a, Lemma 5.2), we have:

$$\lim_{n \longrightarrow +\infty} n^{-b} \pi_n^{-1} = \Gamma(1-b),$$

whereas when $\beta \neq 1$

$$\lim_{n \longrightarrow +\infty} \exp(b(1-\beta)^{-1} n^{1-\beta}) \pi_n^{-1} = \exp(\Lambda)$$

where Λ can be made explicit in terms of the Riemann zeta function. We then conclude the following behaviour of $(S_n)_{n\geq 1}$ using (Gadat et al., 2018, Appendix B):

$$S_n \sim c_{b,\beta}^{-1} n^{\beta \wedge 1},$$

which implies that the joint evolution shall be written as a two-time scale evolution:

$$\begin{cases} \widetilde{\theta}_{n+1} = \widetilde{\theta}_n - \alpha n^{-\frac{\beta\wedge 1}{2}} \frac{g_{n+1}}{\sqrt{\widetilde{w}_n + \widetilde{\varepsilon}_n}} \\ \widetilde{w}_{n+1} = \widetilde{w}_n + c_{b,\beta} n^{-\frac{\beta\wedge 1}{2}} [n^{-\frac{\beta\wedge 1}{2}} g_{n+1}^{\odot 2} - n^{-\frac{\beta\wedge 1}{2}} \widetilde{w}_n] \end{cases}.$$

Adagrad $\beta_2 = 1$, α constant	Two-time scale $\propto (\alpha/\sqrt{n}, 1/n)^2$
Adagrad $\beta_2 = 1$, $\alpha \propto 1/\sqrt{n}$	Unique time scale $\propto (1/n)$
Rmsprop $\beta_2 < 1$, α constant	Unique scale constant step-size $(\sqrt{1-\beta_2}\alpha, 1-\beta_2)$
$\beta_2 = 1 - cn^{-\beta}$, α constant	Two-time scale $\propto \left(n^{-\frac{\beta\wedge 1}{2}}, n^{-\beta\wedge 1}\right)$

Table 1: Left: Standard parametrization. Right: Algorithm described by Equation (1)

We then observe a continuum of possible time scales when β varies between 0 and $+\infty$, which corresponds to the limiting situation of Adagrad and Rmsprop. Conversely, a straightforward argument shows that if $\gamma_n \propto n^{-\beta}$ and $p_n = q_n \propto n^{-r}$ with $\beta > r$ in (1), then we recover an adaptive algorithm tuned with $1 - \beta_2(n) \propto n^{-(\beta+r)}$ and $\alpha_n \propto n^{-\frac{\beta-r}{2}}$.

We provide in Table 1 a summary of the previous conclusions about the link between the parametrizations of (1) and the corresponding standard adaptive algorithms.

2.2.3 Final Remark on Step-size Sequences

We emphasize that when $(\gamma_n, p_n)_{n\geq 1}$ is chosen as a constant sequence (which is the case for the Adam and Rmsprop algorithms), the sequence $(\theta_n)_{n\geq 1}$ evolves as an ergodic Markov chain and therefore the trajectory cannot converge towards a minimizer of f (indeed it cannot converge anywhere). Hence, RMSProp and ADAM are not encompassed in our work, even with a time varying sequence $\alpha_n \longrightarrow 0$, as it would generate some unbounded sequences $(p_n)_{n\geq 1}$ and $(q_n)_{n\geq 1}$. Nevertheless, using a finite time horizon strategy with a small enough value of the step-size, Zou et al. (2019); Défossez et al. (2020) derive some theoretical guarantees on $\mathbb{E}[\|\nabla f(\widetilde{\theta}_n)\|^2]$.

In this work, we have chosen to state asymptotic results within a standard framework of stochastic algorithms:

$$\sum_{n\geq 1} \gamma_{n+1} = +\infty \qquad \text{and} \qquad \lim_{n\to+\infty} \gamma_n = 0.$$

We observe that the time-scale of the algorithm is driven by $(\gamma_{n+1})_{n\geq 1}$ on the θ coordinate, and by $(\gamma_{n+1}p_n)_{n\geq 1}$ on the w coordinate, which leads to a possibly two time-scale algorithm according to the behaviour of $(p_n)_{n\geq 1}$. We leave serious two-time scale considerations for future investigations as we are essentially interested in the $(\gamma_{n+1})_{n\geq 1}$ component, and refer to Borkar (1997); Mokkadem and Pelletier (2006); Bercu et al. (2020a); Costa and Gadat (2020) for other examples of such stochastic algorithms in various (but simpler) situations.

2.3 Assumptions and Convenient Notations

We introduce $(\mathcal{F}_n)_{n\geq 0}$, the canonical filtration associated to our random sequence $\mathcal{F}_n = \sigma\left((\theta_k, w_k)_{1\leq k\leq n}\right)$ and list below the main assumptions used in our work.

We use the symbols $\lesssim_d, \gtrsim_d$ to refer to inequalities up to a multiplicative constant that are independent from the dimension d: for two positive sequences $(u_n)_{n\geq 0}$ and $(v_n)_{n\geq 0}$, we

2. This can be achieved by setting $\gamma_n = p_n = q_n = 1/\sqrt{n}$

write

$$u_n \lesssim_d v_n \text{ if there exists } C > 0 \text{ such that } u_n \leq C v_n, \ \forall n \in \mathbb{N},$$

and the constant C is independent from the dimension of the ambient space d. We will also use the notation $u_n = \mathcal{O}_d(v_n)$ when $u_n \lesssim_d v_n$. We use the symbol $\lesssim$ that refers to an inequality up to a multiplicative constant that can depend on d, mainly in the proof of the local trap avoidance since for this result we are not interested in a quantitative effect of the dimension.

2.3.1 Assumptions on the Noise and Mini-batch.

We first describe our main assumption on the sequence $(g_n)_{n\geq 1}$.

• Assumption $\mathbf{H}_\sigma^p$. We assume that the sequence $(g_n)_{n\geq 1}$ used in (1) provides an unbiased estimation of the true gradient of f at position θ_n, *i.e.* we assume that:

$$\mathbb{E}[g_{n+1} \,|\mathcal{F}_n] = \nabla f(\theta_n).$$

We furthermore assume that the noise sequence $(\xi_{n+1})_{n\geq 1}$ satisfies:

$$\forall n \geq 1 \qquad \xi_{n+1} := g_{n+1} - \nabla f(\theta_n) = \sigma_{n+1}\zeta_{n+1} \quad \text{with} \quad \mathbb{E}[\|\zeta_{n+1}\|^p|\mathcal{F}_n] \leq c(d + f(\theta_n))^{p/2}, \tag{6}$$

where c is a positive constant independent from d.We assume that we use a mini-batch of size $b_n = \sigma_{n+1}^{-2}$ to estimate $\nabla f(\theta_n)$: we observe at step n a set of independent unbiased noisy gradients $(g_{n+1}^{(1)}, g_{n+1}^{(2)}, \ldots, g_{n+1}^{(b_n)})$ and we compute g_{n+1} with a simple averaging;

$$g_{n+1} = \frac{1}{b_n}\sum_{k=1}^{b_n} g_{n+1}^{(k)}.$$

Assumption $\mathbf{H}_\sigma^p$ stands for a classical framework in stochastic optimization methods: $(\sigma_n)_{n\geq 1}$ is an auxiliary sequence that translates a possible use of mini-batches when $\sigma_n \longrightarrow 0$ as $n \longrightarrow +\infty$. The moment assumption on $(\zeta_n)_{n\geq 1}$ is the convenient assumption to handle standard problems like on-line regression, logistic regression or cascade of logistic regressions used in deep learning. We emphasize that we do not make any restrictive and somewhat unrealistic boundedness assumption of the noise $(\zeta_n)_{n\geq 1}$ or of the sequence $(\theta_n)_{n\geq 1}$ itself. Below, we will use this assumption with $p = 4$ in Theorem 1. Finally, we should observe that this assumption introduces a possible linear dependency with d on the size of the variance of the noise. A such assumption is reasonable in the Gaussian case where the variance is exactly linear with d and $\mathbf{H}_\sigma^p$ may be seen as a generalization to higher moments to a larger class of noise distributions.

To derive the convergence of our algorithms towards a local minima, we will need a more stringent condition on the noise sequence. We then introduce the next assumption that will replace $\mathbf{H}_\sigma^p$ in our second main result of almost sure convergence (see Theorem 3 below).

• Assumption $\mathbf{H}_\sigma^\infty$.

- $(\mathbf{H}_\sigma^\infty - 1)$. The noise sequence $(\xi_{n+1})_{n\geq 1}$ is centered and satisfies:

$$\xi_{n+1} = \sigma_{n+1}\zeta_{n+1} \quad \text{with} \quad \mathbb{E}[\|\zeta_{n+1}\|^2|\mathcal{F}_n] \leq 1 \quad \text{and} \quad \mathbb{E}[\|\zeta_{n+1}\|^4|\mathcal{F}_n] \leq C. \tag{7}$$

- $\underline{(\mathbf{H}_\sigma^\infty - 2)}$. The noise sequence is elliptic uniformly in n:

$$\exists m > 0 \qquad \forall n \geq 1 \quad \forall u \in \mathcal{S}^{d-1} \qquad \mathbb{E}[\langle u, \zeta_{n+1} \rangle^2] \geq m > 0.$$

We stress that the upper bound $(\mathbf{H}_\sigma^\infty - 1)$ on the second order moment is not restrictive, up to a modification of the calibration of the sequence $(\sigma_n)_{n\geq 1}$. The second assumption will be of course used to exit local traps. We briefly compare this set of assumptions with the ones of the literature of trap avoidance results. First, $(\mathbf{H}_\sigma^\infty - 1)$ involves both a second and fourth order moment, when compared to Brandiere and Duflo (1996); Benaim and Hirsch (1995); Pemantle (1990) that only use a second order moment. Indeed, in our adaptive algorithm, we will need to manage the second order moment of $\xi_{n+1}^{\otimes 2}$, which induces an assumption on the fourth order moment. Second, we have chosen to use the point of view of Benaim and Hirsch (1995); Pemantle (1990) instead of the one of Brandiere and Duflo (1996), which is an argument that is conditioned to a set $\Gamma(z^*)$ of convergence towards a local trap z^*. In Brandiere and Duflo (1996), a uniform upper bound on the second order moment is assumed *conditionally* to $\Gamma(z^*)$. Since we do not make any assumption on the location of traps z^*, an assumption on $\Gamma(z^*)$ is not so different from a uniform upper bound assumption like $(\mathbf{H}_\sigma^\infty - 1)$.

2.3.2 Assumptions on the cost function f

- Assumption $\mathbf{H}_f$ We now introduce some standard assumptions on f.

  - $\underline{(\mathbf{H}_f - 1)}$. The function f is positive and coercive, *i.e.* f satisfies:

$$\lim_{\|x\| \longrightarrow +\infty} f(x) = +\infty \qquad \text{and} \qquad \min(f) > 0.$$

    Demanding the lower bound of f to be strictly positive is mostly a convenient technical constraint and not fundamentally more restrictive than the classical assumption of positivity.

  - $\underline{(\mathbf{H}_f - 2)}$. We assume that f satisfies the so-called Lipschitz continuous gradient property:

$$\exists L > 0 \quad \forall (x, y) \in \mathbb{R}^d \qquad \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

    We emphasize that this implies the famous descent inequality, $\forall h \in \mathbb{R}^d$:

$$f(x) + \langle h, \nabla f(x) \rangle - \frac{L}{2}\|h\|^2 \leq f(x + h) \leq f(x) + \langle h, \nabla f(x) \rangle + \frac{L}{2}\|h\|^2. \tag{8}$$

    This assumption is commonly used in optimization theory and statistics. Even though it is possible to address some more sophisticated situations (see *e.g.* Bauschke et al. (2017)), it is generally admitted that most of machine learning optimization problems fall into the Lipschitz continuous gradient framework.

  - $\underline{(\mathbf{H}_f - 3)}$. We also assume that another constant c_f exists such that:

$$\|\nabla f\|^2 \leq c_f f. \tag{9}$$

This last assumption prevents a too large growth of the function f and it is immediate to verify that $(\mathbf{H}_f - 3)$ implies that f has a subquadratic growth, *i.e.* $\limsup_{\|x\| \longrightarrow +\infty} \frac{f(x)}{\|x\|^2} < +\infty$. It has been widely used in the literature of stochastic algorithm, see *e.g.* (Gadat et al., 2018) and the references therein.

- $(\mathbf{H}_f - 4)$. Finally, we assume that $\forall\, x \in \mathbb{R}^d$, the set $A_x = \{\theta, f(\theta) = x\} \cap \{\theta, \nabla f(\theta) = 0\}$ is locally finite (meaning that $\forall \theta \in A_x$, there exists a neighborhood N_θ of θ such that $A \cap N_\theta$ is finite).

2.3.3 Assumptions on the step-size sequences

We finally introduce our assumptions on the step-size sequences used all along the paper that involve $(p_n)_{n\geq 1}, (q_n)_{n\geq 1}$ and $(\gamma_n)_{n\geq 1}$. To easily assess some convergence results with quantitative conditions on our gain sequences, we will consider the following situations.

• Assumption $\mathbf{H}_{\text{Steps}}$ (Case $q_\infty > 0$)

- $(\mathbf{H}_{\text{Steps}} - 1)$. The sequences $(p_n)_{n\geq 1}$ and $(q_n)_{n\geq 1}$ satisfy:

$$\exists (r, p_\infty) \in \mathbb{R}^+ \times \mathbb{R}^+ : \quad |p_n - p_\infty| \lesssim_d n^{-r} \qquad \text{and} \qquad \lim_{n \longrightarrow +\infty} q_n = q_\infty > 0.$$

and

$$\forall n \geq 1 \qquad \gamma_{n+1} q_n < 1.$$

- $(\mathbf{H}_{\text{Steps}} - 2)$. The mini-batch sequence $(\sigma_n)_{n\geq 1}$ satisfies:

$$\sigma_n = \sigma_1 n^{-s} \qquad \text{with} \qquad s \geq 0.$$

- $(\mathbf{H}_{\text{Steps}} - 3)$. As already discussed in Section 2.2.3, the sequence $(\gamma_n)_{n\geq 0}$ satisfies:

$$\sum_{n\geq 1} \gamma_{n+1} = +\infty \quad \text{and} \quad \sum_{n\geq 1} \gamma_{n+1}^2 \sigma_n^2 < +\infty \quad \text{and} \quad \sum_{n\geq 0} p_n \gamma_{n+1} \sigma_{n+1}^2 < +\infty.$$

We point out that, for this set of assumptions, $(p_n)_{n\geq 1}$ and $(\sigma_n)_{n\geq 1}$ may be (or not) some vanishing sequences (if $r > 0$ and $p_\infty = 0$ or if $s > 0$).

• Assumption $\mathbf{H}'_{\text{Steps}}$ (Case $q_\infty = 0$)

We also introduce $\mathbf{H}'_{\text{Steps}}$ which corresponds to the same set of assumptions as the ones in $\mathbf{H}_{\text{Steps}}$ while replacing $(\mathbf{H}_{\text{Steps}} - 1)$ by $(\mathbf{H}'_{\text{Steps}} - 1)$ defined as:

- $(\mathbf{H}'_{\text{Steps}} - 1)$. The sequences $(p_n)_{n\geq 1}$ and $(q_n)_{n\geq 1}$ satisfy $p_\infty = q_\infty = 0$ and

$$\exists r, \rho \in \mathbb{R}^+ \text{ such that } p_n \lesssim_d n^{-r} \quad \text{and} \quad q_n \lesssim_d n^{-\rho}$$

and

$$\forall n \geq 1 \qquad \gamma_{n+1} q_n < 1.$$

Finally, we also suppose that there exists a positive sequence $(\upsilon_n)_{n\geq 1}$ such that $\upsilon_n \to +\infty$ and

$$\frac{\upsilon_{n+1}}{\upsilon_n}(1 - q_n \gamma_{n+1}) \leq 1; \quad \sum_{n\geq 1} p_n \gamma_{n+1} \sigma_{n+1}^2 \upsilon_{n+1} < +\infty \quad \text{and} \quad p_n \upsilon_{n+1} \leq c.$$

$(p_n)_{\geq 1}$	$(q_n)_{\geq 1}$	Theorem 1	Theorem 3
$p_\infty + O(n^{-r})$	$q_\infty + O(n^{-\rho})$	$\beta + \mathbf{2s} > \mathbf{1}$	-
$O(n^{-r})$	$q_\infty + O(n^{-\rho})$	$\beta + \mathbf{r} + \mathbf{2s} > \mathbf{1}$	$\frac{1-\beta-r}{2} < s < \frac{\beta}{2} \wedge \frac{1-\beta}{2} \wedge \left((r \wedge \rho) - \frac{\beta}{2}\right)$
$O(n^{-r})$	$O(n^{-\rho})$	$\beta + \mathbf{r} + \mathbf{2s} > \mathbf{1}$ $\rho + \beta \leq 1$	$\frac{1-\beta-r}{2} < s < \frac{\beta}{2} \wedge \frac{1-\beta}{2} \wedge \left((r \wedge \rho) - \frac{\beta}{2}\right)$

Table 2: Conditions on the sequences $(p_n, q_n)_{n\geq 1}$ to verify Theorem 1 and Theorem 3 with $\gamma_{n+1} = n^{-\beta}, \beta \in (1/2, 1]$ and $\sigma_n = \sigma_1 n^{-s}$, with $s \geq 0$.

Condition $(\mathbf{H}'_{\text{Steps}} - 1)$ is not particularly restrictive. For example, if $\gamma_n = \gamma_0 n^{-\beta}$, it is sufficient to assume that $r + 2s + \beta > 1$ (which is rather close to $(\mathbf{H}_{\text{Steps}} - 3)$) and that $\rho + \beta \leq 1$. With this choice of parameters, one can show that $(\mathbf{H}'_{\text{Steps}} - 1)$ is satisfied by choosing $\upsilon_n = n^{\upsilon}$, with $\upsilon = \min(\beta + 2s + r - 1, r) > 0$. This set of assumptions is essentially used to prove that the sequence $(w_n)_{n\geq 0}$ converges a.s. to 0, even when $q_\infty = 0$.

2.4 Main Results

We now state our three main convergence results for the stochastic algorithm defined in Equation (1).

2.4.1 Almost Sure Convergence Towards a Critical Point

Theorem 1 *Assume that* $\mathbf{H}_f$*, and* $\mathbf{H}_\sigma^p$ *hold for* $p = 4$*. Under* $\mathbf{H}_{Steps}$ *or* $\mathbf{H}'_{Steps}$ *the sequence* $(\theta_n, w_n)_{n\geq 1}$ *converges almost surely towards* $(\theta_\infty, 0)$ *where* $\nabla f(\theta_\infty) = 0$*.*

A careful inspection of the previous result shows that it does not permit to address the Adagrad algorithm with a mini-batch of a fixed size, because for the Adagrad algorithm the property $\sum \gamma_{n+1}^2 < +\infty$ trivially fails (since $\gamma_{n+1} = \frac{1}{\sqrt{n+1}}$). However, it is possible to bypass this issue with the help of slowly logarithmic increasing mini-batches, $\sigma_n^{-2} \propto \log(n)^2$.

Theorem 1 is a purely asymptotic convergence result. It provides the convergence of our adaptive algorithm defined in Equation (1) towards a set of *critical points* under mild assumptions on the noise sequence and on the function f. We emphasize that this result holds for a standard setup on stochastic algorithms with a decreasing learning rate $(\gamma_n)_{n\geq 1}$. We observe that the essential condition involved in this result is the convergence of the series that depend on $(\gamma_n, p_n, \sigma_n^2)$. In particular, when $\gamma_n = \gamma_1 n^{-\beta}$, we observe that Theorem 1 holds when:

$$\beta \in (0, 1] \quad \text{and} \quad \beta + r + 2s > 1 \quad \text{and} \quad 2\beta + 2s > 1.$$

From a theoretical point of view, the less restrictive situation corresponds to the choice $\beta = 1$ since the series converges as soon as $\sigma_{n+1}^2 p_n$ decreases like $\log(n)^{-2}$. It implies that either we need to use a very lengthy decrease of the update induced by $(p_n)_{n\geq 1}$, or use a very lengthy increase of the minibatch proportional, with a batch of size $\log^2(n)$ at step n. Of course, this last condition holds as soon as $r + 2s > 0$. When β is chosen lower

than 1, the condition becomes $r + 2s > 1 - \beta$, which may lead to a larger computational cost. Oppositely, the more restrictive situation appears when $\beta < 1/2$, which asks for a polynomially increasing mini-batch all over the iterations to guarantee the almost sure convergence. The optimal tuning of the parameters may not be read in Theorem 1, as it does not provide any quantitative information on the rate of convergence of the method. Some insights are however offered in Theorem 2.

A key tool for the proof of Theorem 1 and for the analysis of the algorithm is the following Lyapunov function:

$$V_{a,b}(\theta, w) := \|\sqrt{w+\varepsilon}\|^2 + af(\theta)^2 + bf(\theta)\|(w+\varepsilon)^{1/4}\|^2,$$

where a and b are two well chosen constants (see proof of Proposition 6 for more details).

2.4.2 Rate of "Convergence"

Using the point of view introduced in Ghadimi and Lan (2013) to assess the computational cost of non-convex stochastic optimization, it is possible to derive a more quantitative result on the sequence $(\theta_n)_{n\geq 1}$. This result is stated in terms of the expected value of the gradient of f all along the algorithm. A δ-approximation computational cost is then defined as the number of samples that are necessary to obtain an average value below δ.

Theorem 2 *Assume that $\mathbf{H}_f$ and $\mathbf{H}_\sigma^p$ hold for $p = 4$ and consider an integer $N > 0$ and τ an integer sampled uniformly over $\{1, \ldots, N\}$:*

i) If $\gamma_n = \gamma = \frac{\sqrt{\varepsilon}}{\sqrt{dN}}$ and $p_n = q_n = p = q = \frac{1}{\sqrt{dN}}$ and $\sigma_n^2 = 1$, then:

$$\mathbb{E}\left[\|\nabla f(\theta_\tau)\|_1^2\right] = \mathcal{O}\left(\frac{d}{\sqrt{\varepsilon N}}\right)$$

and the computational cost to obtain a δ-approximation is $d/(\varepsilon\delta^2)$.

ii) If $\gamma_n = \gamma = \sqrt{\frac{\varepsilon}{N}}$ and $\gamma_{n+1}p_n = \gamma_{n+1}q_n = \gamma p = \gamma q = \frac{\sqrt{\varepsilon}}{n}$ and $\sigma^2 = \frac{1}{d}$, then

$$\mathbb{E}\left[\|\nabla f(\theta_\tau)\|_1^2\right] = \mathcal{O}\left((\varepsilon N)^{-1/2}\right)$$

and the computational cost to obtain a δ-approximation is of order $d/(\varepsilon\delta^2)$.

We emphasize that this last result is not a real convergence result, which is indeed impossible to derive with a constant step-size stochastic algorithm. Nevertheless, it may be seen as a benchmark result following the usages in non-convex machine learning optimization. As indicated by Ghadimi and Lan (2013), it is a convenient way to assess a mean square convergence of stochastic optimization algorithms with non-convex landscape.

We recover here a more quantitative result that translates both the linear effect of the dimension on the "convergence" rate and the dependency of the final bound in terms of $N^{-1/2}$ when the algorithm is randomly stopped uniformly between iteration 1 and N. The presence of both d and of $N^{-1/2}$ is not surprising as it already appears to be the minimax rate of convergence in stochastic optimization with weakly convex landscapes (Nemirovski and Yudin, 1983).

The optimal tuning of the algorithm seem to be the ones that are indicated in our statement, even though other ones could be possible to achieve a δ approximation: σ^2 may be chosen of the order d^{-1} and $\gamma_n \propto \varepsilon^{1/2} N^{-1/2}$, or $\sigma^2 = 1$ and $\gamma = \varepsilon^{1/2}(dN)^{-1/2}$, which leads to a $d\delta^{-2}\varepsilon^{-1}$ computational cost. As discussed in Section 3.2, with this strategy, it seems impossible to improve the $d\delta^{-2}\varepsilon^{-1}$ computational cost obtained with other choices of the parameters. The dependency with the dimension of the ambient space is also in-line with former works on stochastic algorithms and especially on adaptive algorithms. For example, Défossez et al. (2020) also obtain a linear dependency with the dimension d of the computational cost, while a former $d^{3/2}$ dependency was obtained in (Zou et al., 2019).

Even if of rather minor importance, our result is stated with the help of $\mathbb{E}\left[\|\nabla f(\theta_\tau)\|_1^2\right]$, instead of $\mathbb{E}\left[\|\nabla f(\theta_\tau)\|^2\right]$ used by Ward et al. (2019); Défossez et al. (2020) and instead of $\mathbb{E}\left[\|\nabla f(\theta_\tau)\|^{4/3}\right]^{2/3}$ used in (Zou et al., 2019). It is therefore slightly stronger since (using our vectorial notations):

$$\|\nabla f(\theta_\tau)\|_1^2 = \left\|\sqrt{|\nabla f|}\right\|^4 \geq \|\nabla f\|^2.$$

This improvement comes from a careful tuning of a Lyapunov function that is not exactly the same as the one used in these previous works. We refer to Sections 3.1 and 3.2 for further details. Finally, we also point out that when the sequence $(\gamma_n)_{n\geq 1}$ is kept fixed, as indicated in the paragraph 2.2.2 it corresponds to a choice of $\beta_2 < 1$ kept constant all over the time evolution (see Equation 5) and $p = q = \frac{1}{\sqrt{N}}$. This result, with this range of parameters appears to be similar to those of Ward et al. (2019); Défossez et al. (2020), but in our work the assumptions on the noise sequence and on the function f are significantly weaker. In particular, we emphasize that we do not make any boundedness assumption on the noisy gradients nor on the space where θ is living, which would highly simplify the analysis of the algorithm, and do not really correspond to a practical situation where these algorithms are used. Of course, the price to pay without this set of boundedness assumption is that our constants in Theorem 2 are not explicit, which is not the case in the results obtained by Défossez et al. (2020); Ward et al. (2019) that provide some explicit upper bounds.

2.4.3 Almost Sure Convergence Towards a *Minimizer*

In this paragraph, we assess the almost sure convergence of the sequence $(\theta_n)_{n\geq 1}$ towards a local minimum of f and state that the algorithm cannot converge towards a linearly unstable point of the dynamical system, *e.g.* cannot converge towards a saddle point or a local maximum of f. More precisely, we assume that all the critical points of f that are not a local minimum are linearly unstable, *i.e.* with at least one eigenvalue with a negative real part, which may lead to a local maximum or a saddle point.

Theorem 3 *Assume that* $\mathbf{H}_f$, $\mathbf{H}_{Steps}$ *or* $\mathbf{H}'_{Steps}$ *and* $\mathbf{H}_\sigma^\infty$ *hold and that* $\sigma_1 > 0$. *Assume that* f *is twice differentiable and that all the critical points of* f *that are not a local minimum are linearly unstable.*

- *Suppose* $p_\infty = 0$, $q_\infty > 0$, $|q_\infty - q_n| \lesssim n^{-\rho}$, $\gamma_n = \gamma_1 n^{-\beta}$ *and that* (β, r, ρ, s) *are such that:*

$$\left(\frac{1}{2} - \frac{\beta + r}{2}\right) \vee \left(\frac{1}{2} - \beta\right) \leq s < \frac{\beta}{2} \wedge \frac{1-\beta}{2} \wedge \left(r - \frac{\beta}{2}\right) \wedge \left(\rho - \frac{\beta}{2}\right).$$

  Then almost surely the sequence $(\theta_n)_{n\geq 1}$ *does not converge towards a local maximum or a saddle point of* f*.*

- *If* $q_\infty = 0$*, the same conclusion holds if we furthermore assume that:*

$$\beta + \rho \leq 1$$

.

Several remarks are necessary after our last theorem, that identifies not only the limit points as the critical points of f but as the local minimizers. Hence, our contribution should be understood as a new example of stochastic method that avoids local traps, and then connected to the contributions of Pemantle (1990); Brandiere and Duflo (1996); Gadat et al. (2018); Barakat et al. (2020).

• Needed properties on f Of course, this result needs some important assumptions on the function f, and especially that all the critical points t of f such that $\nabla f(t) = 0$ that are not a local minima have at least one repulsive direction, *i.e.* $D^2 f(t)$ has at least one negative eigenvalue.

The second important assumption on f that concerns the nature of the attainable equilibria is $(\mathbf{H}_f - 4)$ that prevents from the existence of a uncountable set of unstable equilibria where the algorithm may be trapped: from Theorem 1, $(\theta_n)_{n\geq 1}$ almost surely converges towards a critical point of f and since $(f(\theta_n))_{n\geq 1}$ is also an a.s. convergent sequence, we deduce from $(\mathbf{H}_f - 4)$ that the set of the possible limits for $(\theta_n)_{n\geq 1}$ is discrete, which prevents from the existence of some pathological situations highlighted by Pemantle (1990, pages 699-700): the limit points of $(\theta_n)_{n\geq 1}$ is necessarily a discrete countable set of $\mathbb{R}^d$.

• Nature of the result Moreover, we point out that our result holds for every initialization point and does not use any integration over (θ_0, w_0). Hence, the nature of our result is different from the ones obtained in recent contributions in the field of machine learning (Lee et al., 2016, 2017): we establish that our *stochastic algorithm* converges with probability 1 to a minimizer, which is different from proving that a *deterministic or randomized algorithm* randomly initialized converges to a local minimum with probability 1, this is for example the result obtained by Lee et al. (2017) for the gradient descent.

• Fixed minibatch This situation is the easiest case since $\sigma_n^2 = \sigma^2 > 0$ helps to avoid traps. When the variance of the noise sequence is kept fixed all along the iterations (no mini-batch is used, so that $s = 0$), the previous conditions on the parameters can be summarized as: $p_\infty = 0$ ($p_n \longrightarrow 0$ as $n \longrightarrow +\infty$) and $\beta \in (1/2, 1)$; $r \in (1 - \beta, \beta)$, when $\beta < 2/3$ and $r \in (\beta/2, \beta)$ when $\beta \geq 2/3$.

• Increasing minibatch Finally, we should emphasize that this last result is rather difficult to obtain when the mini-batch parameter s is chosen strictly greater than 0 since it translates a possible vanishing level of noise when the number of iterations is increasing.

Our assumption shows that the size of the mini-batch should not grow too fast (induced by the condition $s \leq (1 - \beta)/2$ for example when $\beta > 1/2$) to obtain the convergence

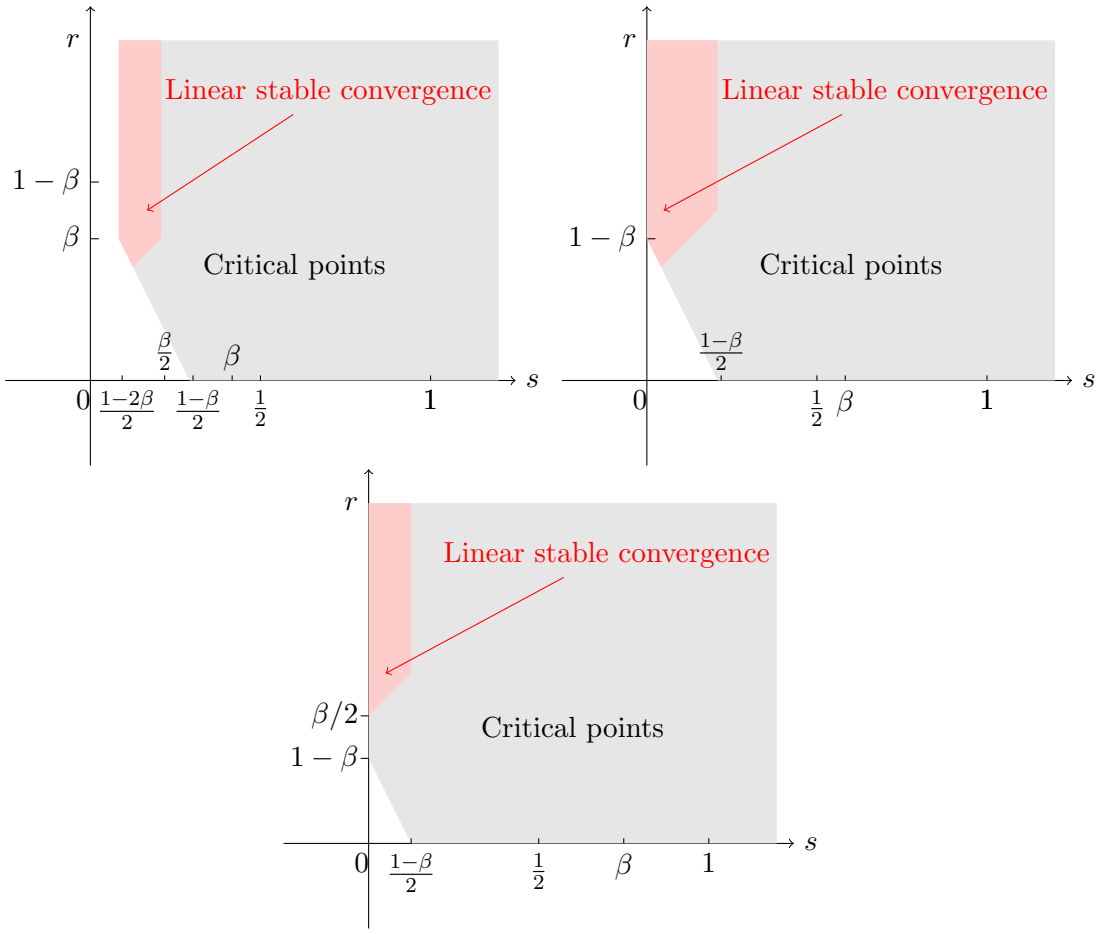

Figure 1: Nature of our convergence results when (s, r) are chosen according to the statements of Theorem 1 and Theorem 3 when $\gamma_n = \gamma_1 n^{-\beta}$. Top Left: $1/3 \leq \beta \leq 1/2$. Top Right: $1/2 \leq \beta \leq 2/3$. Bottom: $\beta \geq 2/3$. The mini-batch size is $\sigma_n^{-2} \propto n^{2s}$ and $q_\infty > 0$. To simplify the pictures we also suppose that $\rho \geq r$.

towards a local minimizer of f. Up to our knowledge, a such explicit phenomenon is new in the stochastic algorithm community. It would deserve further numerical or theoretical investigations to identify whether this condition is necessary to convert an almost sure convergence result towards critical points into a convergence result towards a stable point of the differential system.

• Adagrad case It is possible to address the specific case of Adagrad, which corresponds to $p_\infty = q_\infty = 0$, $p_n = q_n = \frac{1}{\sqrt{n}}$ and $\gamma_{n+1} = \frac{1}{\sqrt{n}}$ (see Table 1 and Section 2.2.1). We observe that Theorem 3 holds since in this case $\beta + \rho = \frac{1}{2} + \frac{1}{2} = 1$ as soon as:

$$\log^2(n) \lesssim \sigma_n^{-2} \lesssim n^{1/4}.$$

Therefore, Adagrad avoids local traps (linearly unstable critical points of f) as soon as the minibatches are at least of the order $\log(n)^2$, which is rather a slowly increasing sequence, and less than $n^{1/4}$, which is the theoretical limit we found in our analysis that generates a sufficient amount of noise to escape from traps.

• What about SGD, mini-batches and trap avoidance? The proof of Theorem 3 can also be adapted to obtain the same results for SGD as soon as $s \leq (1-\beta)/2$. Said differently, we may adapt our proof to establish that SGD avoids local traps almost surely as soon as the learning rate $\gamma_{n+1} \propto n^{-\beta}$ and the mini-batch $\sigma_n^2 \propto n^{-2s}$ satisfy:

$$\frac{1}{2} - \beta \leq s \leq \frac{\beta \wedge (1-\beta)}{2},$$

which finally yields to the same condition with the popular (popular at least theoretically) choice $\gamma_{n+1} \propto n^{-1/2}$ in the non-convex case. Up to our knowledge, the maximal size of the mini-batch that guarantees convergence to local minimizer is still an unresolved question even for the SGD. As a really ambitious question, we leave this problem for future developments.

2.5 Organization of the Paper

The rest of the paper consists in showing the proofs of the previous results. Theorem 1 and Theorem 2 are proven in Section 3. In particular, Proposition 4 studies the average one-step evolution of the algorithm through several key functions. This proposition permits to derive a Lyapunov function in Section 3.2 that translates both the average quantitative result of Theorem 2 and the asymptotic convergence result of Theorem 1. The main difficulties in these two results is to derive a mean reverting effect in terms of $\|\nabla f(\theta_n)\|^2$ without using some extra boundedness assumption, and to assess the influence of d on the quantitative result. Theorem 3 is a typical result of stochastic algorithms, and is inspired from the contributions of Pemantle (1990) and Benaim and Hirsch (1995). The proof is detailed in Section 4 and the cornerstone of this proof is the use of the stable/unstable manifold Lemma that provides an ad-hoc Lyapunov function of the dynamical system, denoted by η in Proposition 10. We also refer to the recent contribution of Barakat et al. (2020) for another typical application to stochastic algorithms. The main novelty brought in our proof is the a.s. escape of local maximum when the mini-batch has a low noise level. In particular, from a technical point of view, we take advantage of the boundedness series of Proposition 6, which is a key ingredient in the proof of Proposition 16.

3. Almost Sure Convergence to the Set of Critical Points

The purpose of this section is to prove Theorem 1 and Theorem 2. In particular, we will obtain Theorem 2 during the proof of the almost sure convergence result as a specific point of Proposition 6, *i*). The basic ingredient of our proof relies on the result of Robbins and Siegmund (1971), that will be applied with the help of an ad-hoc Lyapunov function on $(\theta_n, w_n)_{n\geq 1}$.

3.1 Preliminary Computations

Below, we will pay a specific attention to the dimension dependency in the inequalities we will obtain. We first state the following proposition that will create a mean reverting effect from iteration n to iteration $n+1$ on the pair (θ_n, w_n).

Proposition 4 *Assume that* $\|\gamma_n q_n\|_\infty < 1$, *that* $\|p_n\|_\infty < +\infty$, *that* $\mathbf{H}_f$ *and* $\mathbf{H}_\sigma^p$ *hold for* $p = 4$.

i) For any $n \geq 1$, *one has:*

$$\mathbb{E}[\|\sqrt{w_{n+1}+\varepsilon}\|^2|\mathcal{F}_n] \leq \|\sqrt{w_n+\varepsilon}\|^2 - q_n\gamma_{n+1}\|\sqrt{w_n}\|^2 + \gamma_{n+1}p_n\|\nabla f(\theta_n)\|^2 + \gamma_{n+1}\sigma_{n+1}^2 p_n(d + f(\theta_n)).$$

ii) A constant c_1 *independent from* n, d *and* ε *exists such that for any* $n \geq 1$*:*

$$\mathbb{E}[f(\theta_{n+1})\,||\mathcal{F}_n] \leq f(\theta_n)(1 + \frac{c_1\gamma_{n+1}^2}{\varepsilon}(\mathbb{1}_{\gamma_{n+1}>\frac{\sqrt{\varepsilon}}{L^2}} + \sigma_{n+1}^2)) + \frac{c_1 d}{\varepsilon}\gamma_{n+1}^2\sigma_{n+1}^2 - \frac{\gamma_{n+1}}{2}\left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^2.$$

iii) Let $k_2 = \min\left(\frac{\sqrt{\varepsilon}}{L^2}, \frac{\sqrt{\varepsilon}}{c_f}\right)$. *If we define:*

$$s_n = \mathbb{1}_{\gamma_{n+1}>k_2}\left(\frac{\gamma_{n+1}^2}{\varepsilon} + d\frac{\gamma_{n+1}^4}{\varepsilon^2}\right) + \frac{d}{\varepsilon}\sigma_{n+1}^2\gamma_{n+1}^2 + \frac{d^2}{\varepsilon}\sigma_{n+1}^4\gamma_{n+1}^4$$

then a constant c_2 *independent from* d *and* ε *exists such that:*

$$\mathbb{E}[f(\theta_{n+1})|\mathcal{F}_n] \leq f(\theta_n)^2\left[1 + c_2 s_n\right] + c_2 s_n - \frac{\gamma_{n+1}}{2}f(\theta_n)\left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^2$$

iv) If we define $k_3 = \frac{\varepsilon}{2L^2}$ *and* t_n *as:*

$$t_n := 2\gamma_{n+1}p_n\left(\frac{d}{\sqrt{\varepsilon}}\sigma_{n+1}^2 + \frac{\gamma_{n+1}^2}{\varepsilon\sqrt{\varepsilon}}(1 + \sigma_{n+1}^4 d^2)\right),$$

then a large enough constant c_3 (independent from n, d and ε) exist such that:

$$\mathbb{E}[f(\theta_{n+1})\|(w_{n+1}+\varepsilon)^{1/4}\|^2\,|\mathcal{F}_n] \leq f(\theta_n)\|(w_n+\varepsilon)^{1/4}\|^2\left(1+c_3\frac{\gamma_{n+1}^2}{\varepsilon}(\mathbb{1}_{\gamma_{n+1}>k_3}+d\sigma_{n+1}^2)\right)$$
$$-\frac{\gamma_{n+1}}{2}\left\|\sqrt{|\nabla f(\theta_n)|}\right\|^4+\gamma_{n+1}p_nf(\theta_n)\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2$$
$$+c_3\left[t_nf(\theta_n)^2+t_n\right].$$

Proof We consider each point separately.
• *Proof of i):* We observe that:

$$\mathbb{E}[\|\sqrt{w_{n+1}+\varepsilon}\|^2\,|\mathcal{F}_n] = \mathbb{E}\left[\left\|\sqrt{w_n+\gamma_{n+1}[p_ng_{n+1}^{\odot 2}-q_nw_n]+\varepsilon}\right\|^2\,|\mathcal{F}_n\right]$$
$$=\mathbb{E}\left[\sum_{i=1}^d w_n^i+\varepsilon+\gamma_{n+1}[p_n\{g_{n+1}^i\}^2-q_nw_n^i]\,|\mathcal{F}_n\right]$$
$$\leq \|\sqrt{w_n+\varepsilon}\|^2-q_n\gamma_{n+1}\|\sqrt{w_n}\|^2+\gamma_{n+1}p_n\|\nabla f(\theta_n)\|^2+\gamma_{n+1}p_n\sigma_{n+1}^2(d+f(\theta_n)),$$

where the last line comes from the definition of $\xi_{n+1}=\sigma_{n+1}\zeta_{n+1}$ and the fact that:

$$\mathbb{E}[\zeta_{n+1}\,|\mathcal{F}_n]=0 \quad \text{and} \quad \mathbb{E}[\|\zeta_{n+1}\|^2\,|\mathcal{F}_n] \lesssim_d (d+f(\theta_n)).$$

This concludes the proof. ◇
• *Proof of ii):* We develop $f(\theta_{n+1})$ using the descent inequality (8):

$$\begin{aligned} f(\theta_{n+1}) &= f\left(\theta_n-\gamma_{n+1}\frac{g_{n+1}}{\sqrt{w_n+\varepsilon}}\right)\\ &\leq f(\theta_n)-\gamma_{n+1}\langle\nabla f(\theta_n),\frac{g_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle+\frac{L^2}{2}\gamma_{n+1}^2\left\|\frac{g_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2:=m_n^+ \qquad (10)\\ &\leq f(\theta_n)-\gamma_{n+1}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2-\gamma_{n+1}\langle\nabla f(\theta_n),\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle\\ &+\frac{L^2}{2}\gamma_{n+1}^2\left\|\frac{g_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2 \end{aligned}$$

We start by computing the conditional expectation of the last term:

$$\begin{aligned} \mathbb{E}\left[\left\|\frac{g_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2\,|\mathcal{F}_n\right] &= \left\|\frac{\nabla f(\theta_n)}{\sqrt{w_n+\varepsilon}}\right\|^2+\mathbb{E}\left[\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2\,|\mathcal{F}_n\right]\\ &\leq \frac{1}{\sqrt{\varepsilon}}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2+\frac{c\sigma_{n+1}^2}{\varepsilon}(d+f(\theta_n)), \qquad (11) \end{aligned}$$

where in the last line we use the assumption $\mathbf{H}_\sigma^p$ for $p=2$. Inserting this in (10) we obtain that :

$$\mathbb{E}[f(\theta_{n+1})\,|\mathcal{F}_n] \le f(\theta_n) - \gamma_{n+1}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2 + \frac{L^2}{2}\gamma_{n+1}^2\left(\frac{1}{\sqrt{\varepsilon}}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2 + \frac{c\sigma_{n+1}^2}{\varepsilon}(d+f(\theta_n))\right)$$

Finally using the sub-quadratic growth assumption given by (9), we have that:

$$\gamma_{n+1}^2\frac{L^2}{2\sqrt{\varepsilon}}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2 \le \gamma_{n+1}^2\frac{L^2c_f}{\varepsilon}f(\theta_n)\mathbb{1}_{\gamma_{n+1}>\frac{\sqrt{\varepsilon}}{L^2}} + \frac{\gamma_{n+1}}{2}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2\mathbb{1}_{\gamma_{n+1}\le\frac{\sqrt{\varepsilon}}{L^2}}$$

Regrouping all these terms, we can conclude that a constant $c_1 = L^2(c\vee c_f)$ exists such that:

$$\mathbb{E}[f(\theta_{n+1})\,|\mathcal{F}_n] \le f(\theta_n)(1+\frac{c_1\gamma_{n+1}^2}{\varepsilon}(\mathbb{1}_{\gamma_{n+1}>\frac{\sqrt{\varepsilon}}{L^2}}+\sigma_{n+1}^2)) + \frac{c_1 d\gamma_{n+1}^2\sigma_{n+1}^2}{\varepsilon} - \frac{\gamma_{n+1}}{2}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2.$$

We emphasize that at this stage, this last inequality does not create any repelling effect on the position $(\theta_n)_{n\ge1}$ and we need to deal with the denominator $(w_n+\varepsilon)$, which is the purpose of *iii)*. ◇

• *Proof of iii):* We consider the evolution of f^2 from θ_n to θ_{n+1}. Using (10) and the positivity of f, we deduce that:

$$f(\theta_{n+1})^2 \le \{m_n^+\}^2.$$

Going back to the bound on m_n^+ we observe that:

$$\begin{aligned}
m_n^+ &\le f(\theta_n) - \gamma_{n+1}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2 - \gamma_{n+1}\langle\nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle \\
&\quad + L^2\gamma_{n+1}^2\left(\left\|\frac{\nabla f(\theta_n)}{\sqrt{w_n+\varepsilon}}\right\|^2 + \left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2\right) \\
&\le f(\theta_n) - \frac{\gamma_{n+1}}{2}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2 - \gamma_{n+1}\langle\nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle \\
&\quad + L^2\gamma_{n+1}^2\left(\left\|\frac{\nabla f(\theta_n)}{\sqrt{w_n+\varepsilon}}\right\|^2\mathbb{1}_{\gamma_{n+1}>\frac{\sqrt{\varepsilon}}{L^2}} + \left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2\right) \\
&\le f(\theta_n)\left(1+\gamma_{n+1}^2\frac{L^2}{\varepsilon}\mathbb{1}_{\gamma_{n+1}>\frac{\sqrt{\varepsilon}}{L^2}}\right) - \frac{\gamma_{n+1}}{2}\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2 \\
&\quad - \gamma_{n+1}\langle\nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle + L^2\gamma_{n+1}^2\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2
\end{aligned}$$

Denote by a_n the coefficients of $f(\theta_n)$ in the previous inequality, $a_n = 1+\gamma_{n+1}^2 \frac{L^2}{\varepsilon}\mathbb{1}_{\gamma_{n+1}>\frac{\sqrt{\varepsilon}}{L^2}}$. Expanding $\{m_n^+\}^2$ we get :

$$\begin{aligned}\{m_n^+\}^2 \leq f(\theta_n)^2 a_n^2 &+ \frac{\gamma_{n+1}^2}{4}\left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^4 + \gamma_{n+1}^2\langle \nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle^2 + \frac{L^4}{4}\gamma_{n+1}^4\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^4 \\ &- 2a_n\gamma_{n+1}f(\theta_n)\langle \nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle - \gamma_{n+1}f(\theta_n)\left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^2 \\ &+ L^2\gamma_{n+1}^2 a_n f(\theta_n)\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2 + \gamma_{n+1}^2\left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^2\langle \nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle \\ &+ 2L^2\gamma_{n+1}^3\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2\left|\langle \nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle\right| \end{aligned} \tag{12}$$

When taking the conditional expectation in the previous inequality we treat some of these terms separately. Using the Cauchy-Schwarz inequality together with assumption H_σ^2 and (9) we have that:

$$\begin{aligned}\mathbb{E}[\langle \nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle^2|\mathcal{F}_n] &\leq \|\nabla f(\theta_n)\|^2\mathbb{E}\left[\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2|\mathcal{F}_n\right] \\ &\leq c\sigma_{n+1}^2\frac{c_f}{\varepsilon}f(\theta_n)(d+f(\theta_n)) \\ &\leq 2c_f c\frac{d\sigma_{n+1}^2}{\varepsilon}(1+f(\theta_n)^2)\end{aligned}$$

The term $\frac{L^4}{4}\gamma_{n+1}^4\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^4$ can be easily bounded using H_σ^4:

$$\mathbb{E}\left[\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^4|\mathcal{F}_n\right] \leq \frac{2}{\varepsilon^2}\sigma_{n+1}^4(d^2+f(\theta_n)^2)$$

We bound the conditional expectation of the last term using the same type of arguments

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^2\left|\langle \nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle\right|\right] &\leq \|\nabla f(\theta_n)\|\mathbb{E}\left[\left\|\frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}}\right\|^3|\mathcal{F}_n\right] \\ &\leq c_f f(\theta_n)\frac{c\sigma_{n+1}^3}{\varepsilon\sqrt{\varepsilon}}(d+f(\theta_n))^{3/2} \\ &\lesssim_d \left(\frac{d}{\varepsilon}\right)^{3/2}\sigma_{n+1}^3(1+f(\theta_n)^2)\end{aligned}$$

Taking the conditional expectation in (12), using the centering of the noise sequence (ξ_n), the fact that $\sqrt{ab} \leq (a+b)/2$ and inserting these previous bounds, we have that there exists

a constant c_2 independent of d and ε such that

$$\begin{aligned}\mathbb{E}[f(\theta_{n+1})|\mathcal{F}_n] \leq f(\theta_n)^2 \left[1 + c_2 \left[\mathbb{1}_{\gamma_{n+1} > \frac{\sqrt{\varepsilon}}{L^2}} \left(\frac{\gamma_{n+1}^2}{\varepsilon} + d\frac{\gamma_{n+1}^4}{\varepsilon^2}\right) + \frac{d}{\varepsilon}\sigma_{n+1}^2\gamma_{n+1}^2 + \frac{d^2}{\varepsilon}\sigma_{n+1}^4\gamma_{n+1}^4\right]\right] \\ - \gamma_{n+1} f(\theta_n) \left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^2 + \frac{\gamma_{n+1}^2}{4}\left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^4 \\ + c_2 \left[\mathbb{1}_{\gamma_{n+1} > \frac{\sqrt{\varepsilon}}{L^2}} \left(\frac{\gamma_{n+1}^2}{\varepsilon} + d\frac{\gamma_{n+1}^4}{\varepsilon^2}\right) + \frac{d}{\varepsilon}\sigma_{n+1}^2\gamma_{n+1}^2 + \frac{d^2}{\varepsilon}\sigma_{n+1}^4\gamma_{n+1}^4\right]\end{aligned}$$

Observing that

$$\begin{aligned}\frac{\gamma_{n+1}^2}{4}\left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^4 &\leq \frac{\gamma_{n+1}^2 c_f}{4\sqrt{\varepsilon}} f(\theta_n)\left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^2 \\ &\leq \mathbb{1}_{\gamma_{n+1} > \frac{\sqrt{\varepsilon}}{c_f}} \frac{\gamma_{n+1}^2 c_f^2}{4\varepsilon} f(\theta_n)^2 + \mathbb{1}_{\gamma_{n+1} \leq \frac{\sqrt{\varepsilon}}{c_f}} \frac{\gamma_{n+1}}{2} f(\theta_n)\left|\left|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right|\right|^2\end{aligned}$$

we can conclude by taking $k_2 = \sqrt{\varepsilon}\min(1/c_f, 1/L^2)$ and re-denoting the constant c_2.

• *Proof of iv):* We observe that:

$$\begin{aligned}\|(w_{n+1}+\varepsilon)^{1/4}\|^2 &= \sum_{i=1}^d \sqrt{w_n^i + \varepsilon + \gamma_{n+1}[p_n\{g_{n+1}^i\}^2 - q_n w_n^i]} \\ &= \sum_{i=1}^d \sqrt{(w_n^i+\varepsilon)\left(1 + \gamma_{n+1}\frac{p_n\{g_{n+1}^i\}^2 - q_n w_n^i}{w_n^i+\varepsilon}\right)} \\ &= \sum_{i=1}^d \sqrt{w_n^i+\varepsilon}\left(1 + \gamma_{n+1}\frac{p_n\{g_{n+1}^i\}^2 - q_n w_n^i}{w_n^i+\varepsilon}\right)^{1/2},\end{aligned}$$

where the last line comes from the fact that $w_n^i/(w_n^i+\varepsilon) < 1$, which implies that the last term of the right hand side exists.

Using $\sqrt{1+a} \leq 1 + a/2$ for any $a > -1$, we deduce that:

$$\begin{aligned}\|(w_{n+1}+\varepsilon)^{1/4}\|^2 &\leq \sum_{i=1}^d \sqrt{w_n^i+\varepsilon}\left(1 + \frac{\gamma_{n+1}}{2}\frac{p_n\{g_{n+1}^i\}^2 - q_n w_n^i}{w_n^i+\varepsilon}\right) \\ &\leq \|(w_n+\varepsilon)^{1/4}\|^2 + \frac{\gamma_{n+1}}{2}\langle\frac{1}{\sqrt{w_n+\varepsilon}}, p_n g_{n+1}^{\odot 2} - q_n w_n\rangle.\end{aligned}$$

Now observe that the second term can easily be bounded by:

$$\begin{aligned}\langle \frac{1}{\sqrt{w_n+\varepsilon}}, p_n g_{n+1}^{\odot 2} - q_n w_n\rangle &= \sum_{i=1}^{d} \frac{p_n\{g_{n+1}^i\}^2}{\sqrt{w_n^i+\varepsilon}} - \underbrace{\langle \frac{1}{\sqrt{w_n+\varepsilon}}, q_n w_n\rangle}_{\text{positive term}} \\ &\leq \sum_{i=1}^{d} \frac{p_n\{g_{n+1}^i\}^2}{\sqrt{w_n^i+\varepsilon}} \\ &\leq 2p_n \left(\left\| \frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}} \right\|^2 + \varepsilon^{-1/2}\sigma_{n+1}^2 \|\zeta_{n+1}\|^2] \right)\end{aligned}$$

We use this last inequality and $f(\theta_{n+1}) \leq m_n^+$ to conclude that:

$$f(\theta_{n+1})\|(w_{n+1}+\varepsilon)^{1/4}\|^2 \leq m_n^+ \left(\|(w_n+\varepsilon)^{1/4}\|^2 + \gamma_{n+1} p_n \left(\left\| \frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}} \right\|^2 + \frac{\sigma_{n+1}^2}{\sqrt{\varepsilon}} \|\zeta_{n+1}\|^2] \right) \right).$$

We use the following decomposition :

$$\begin{aligned} m_n^+ &\leq f(\theta_n) - \gamma_{n+1} \left\| \frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}} \right\|^2 - \gamma_{n+1} \langle \nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}} \rangle \\ &\quad + L^2 \varepsilon^{-1/2} \gamma_{n+1}^2 \left\| \frac{\nabla f(\theta_n)}{(w_n+\epsilon)^{1/4}} \right\|^2 + L^2 \varepsilon^{-1} \gamma_{n+1}^2 \sigma_{n+1}^2 \|\zeta_{n+1}\|^2 \\ &\leq f(\theta_n) - \frac{\gamma_{n+1}}{2} \left\| \frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}} \right\|^2 - \gamma_{n+1} \langle \nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}} \rangle \\ &\quad + L^2 \mathbb{1}_{\gamma_{n+1} > k_3} \varepsilon^{-1} \gamma_{n+1}^2 \left\| \nabla f(\theta_n) \right\|^2, \end{aligned}$$

where k_3 is a small enough constant ($k_3 \leq \sqrt{\varepsilon}/2L^2$). We then develop and obtain that:

$$\begin{aligned} f(\theta_{n+1})\|(w_{n+1}+\varepsilon)^{1/4}\|^2 \leq \Bigg(& f(\theta_n) - \frac{\gamma_{n+1}}{2} \left\| \frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}} \right\|^2 - \gamma_{n+1} \langle \nabla f(\theta_n), \frac{\xi_{n+1}}{\sqrt{w_n+\varepsilon}} \rangle \\ & + L^2 \varepsilon^{-1} \mathbb{1}_{\gamma_{n+1} > k_3} \gamma_{n+1}^2 \left\| \nabla f(\theta_n) \right\|^2 + L^2 \varepsilon^{-1} \gamma_{n+1}^2 \sigma_{n+1}^2 \|\zeta_{n+1}\|^2 \Bigg) \left(\|(w_n+\varepsilon)^{1/4}\|^2 \right. \\ & \left. + \gamma_{n+1} p_n \left(\left\| \frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}} \right\|^2 + \varepsilon^{-1/2} \sigma_{n+1}^2 \|\zeta_{n+1}\|^2] \right) \right) \end{aligned}$$

The baseline remark is that thanks to the Cauchy-Schwarz inequality, we have:

$$\begin{aligned} \left\| \sqrt{|\nabla f(\theta_n)|} \right\|^4 &= \left(\sum_{i=1}^{d} |\partial_i f(\theta_n)| \right)^2 \\ &\leq \sum_{i=1}^{d} \frac{\{\partial_i f(\theta_n)\}^2}{\sqrt{w_n^i+\varepsilon}} \sum_{i=1}^{d} \sqrt{w_n^i+\varepsilon} \\ &= \left\| \frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}} \right\|^2 \|(w_n+\varepsilon)^{1/4}\|^2. \end{aligned}$$

We then compute the conditional expectation of the previous terms with respect to $\mathcal{F}_n$, using the previous inequality, the centering of ξ_{n+1}, and obtain that a constant c_3 independent from d exists such that:

$$
\begin{aligned}
&\mathbb{E}[f(\theta_{n+1})\|(w_{n+1}+\varepsilon)^{1/4}\|^2\,|\mathcal{F}_n] \\
&\quad\le f(\theta_n)\|(w_n+\varepsilon)^{1/4}\|^2 + \gamma_{n+1}p_n f(\theta_n)\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2 - \frac{\gamma_{n+1}}{2}\left\|\sqrt{|\nabla f(\theta_n)|}\right\|^4 \\
&\quad + c_3\mathbb{E}\Bigg(\underbrace{\frac{\gamma_{n+1}}{\sqrt{\varepsilon}}\sigma_{n+1}^2 p_n f(\theta_n)\|\zeta_{n+1}\|^2}_{:=①} + \underbrace{\frac{\gamma_{n+1}^2}{\sqrt{\varepsilon}}p_n\sigma_{n+1}^3\left|\langle\nabla f(\theta_n),\frac{\zeta_{n+1}}{\sqrt{w_n+\varepsilon}}\rangle\right|\|\zeta_{n+1}\|^2}_{:=②} \\
&\quad + \underbrace{\mathbb{1}_{\gamma_{n+1}>k_3\varepsilon^{-1}}\gamma_{n+1}^2\|\nabla f(\theta_n)\|^2\|(w_n+\varepsilon)^{1/4}\|^2}_{:=③} + \underbrace{\frac{\gamma_{n+1}^3}{\varepsilon}p_n\|\nabla f(\theta_n)\|^2\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2}_{:=④} \\
&\quad + \underbrace{\frac{\gamma_{n+1}^3}{\varepsilon\sqrt{\varepsilon}}p_n\sigma_{n+1}^2\|\nabla f(\theta_n)\|^2\|\zeta_{n+1}\|^2}_{:=⑤} + \underbrace{\frac{\gamma_{n+1}^2}{\varepsilon}\sigma_{n+1}^2\|\zeta_{n+1}\|^2\|(w_n+\varepsilon)^{1/4}\|^2}_{:=⑥} \\
&\quad + \underbrace{\frac{\gamma_{n+1}^3}{\varepsilon}\sigma_{n+1}^2 p_n\|\zeta_{n+1}\|^2\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2}_{:=⑦} + \underbrace{\frac{\gamma_{n+1}^3}{\varepsilon\sqrt{\varepsilon}}p_n\sigma_{n+1}^4\|\zeta_{n+1}\|^4}_{:=⑧}\,\Big|\,\mathcal{F}_n\Bigg)
\end{aligned}
$$

We then study each terms in the large bracket separately. Using $f \le 1+f^2$ and $\mathbf{H}_\sigma^p$, we have:

$$
\begin{aligned}
\mathbb{E}[①\big|\mathcal{F}_n] &= \frac{\gamma_{n+1}}{\sqrt{\varepsilon}}\sigma_{n+1}^2 p_n f(\theta_n)\mathbb{E}\left(\|\zeta_{n+1}\|^2\big|\mathcal{F}_n\right) \\
&\le \frac{\gamma_{n+1}}{\sqrt{\varepsilon}}\sigma_{n+1}^2 p_n f(\theta_n)(d+f(\theta_n)) \\
&\le \frac{\gamma_{n+1}}{\sqrt{\varepsilon}}\sigma_{n+1}^2 p_n d(1+f(\theta_n))^2.
\end{aligned}
$$

The second term is handled with the help of $\mathbf{H}_\sigma^p$ (for $p=3$) and the Cauchy-Schwarz inequality:

$$
\begin{aligned}
\mathbb{E}[②\big|\mathcal{F}_n] &\le \frac{\gamma_{n+1}^2}{\varepsilon}p_n\sigma_{n+1}^3\|\nabla f(\theta_n)\|\mathbb{E}[\|\zeta_{n+1}\|^3\big|\mathcal{F}_n] \\
&\le \frac{\gamma_{n+1}^2}{\varepsilon}p_n\sigma_{n+1}^3\|\nabla f(\theta_n)\|(d+f(\theta_n))^{3/2} \\
&\le \frac{\gamma_{n+1}^2}{\varepsilon}p_n\sigma_{n+1}^3\sqrt{c_f(1+f(\theta_n))}d^{3/2}(1+f(\theta_n))^{3/2} \\
&\lesssim_d \frac{\gamma_{n+1}^2}{\varepsilon}p_n\sigma_{n+1}^3 d^{3/2}(1+f(\theta_n))^2.
\end{aligned}
$$

③ is $\mathcal{F}_n$ measurable and the subquadratic growth assumption given by (9) ensures that:

$$③ \leq \mathbb{1}_{\gamma_{n+1}>k_3} \frac{c_f \gamma_{n+1}^2}{\varepsilon} f(\theta_n) \|(w_n+\varepsilon)^{1/4}\|^2.$$

The term ④ is $\mathcal{F}_n$ measurable and we use once more that $\|\nabla f\|^2 \lesssim_d f$ to obtain that:

$$\begin{aligned} ④ &\lesssim_d \frac{\gamma_{n+1}^3}{\varepsilon} p_n f(\theta_n) \sum_{i=1}^d \frac{\{\partial_i f(\theta_n)\}^2}{\sqrt{w_n^i+\varepsilon}} \leq \frac{\gamma_{n+1}^3}{\varepsilon} p_n f(\theta_n) \varepsilon^{-1/2} \sum_{i=1}^d \{\partial_i f(\theta_n)\}^2 \\ &\lesssim_d \frac{\gamma_{n+1}^3}{\varepsilon} p_n f(\theta_n)^2. \end{aligned}$$

For ⑤ we use Assumption $\mathbf{H}_\sigma^p$ with $p=2$ and obtain that:

$$\begin{aligned} \mathbb{E}[⑤|\mathcal{F}_n] &= \frac{\gamma_{n+1}^3}{\varepsilon\sqrt{\varepsilon}} p_n \sigma_{n+1}^2 \|\nabla f(\theta_n)\|^2 \mathbb{E}[\|\zeta_{n+1}\|^2|\mathcal{F}_n] \\ &\lesssim_d \frac{\gamma_{n+1}^3}{\varepsilon\sqrt{\varepsilon}} p_n \sigma_{n+1}^2 f(\theta_n)(d+f(\theta_n)) \\ &\lesssim_d \frac{\gamma_{n+1}^3}{\varepsilon\sqrt{\varepsilon}} p_n \sigma_{n+1}^2 d(1+f(\theta_n))^2. \end{aligned}$$

⑥ is close to ③, we use $\mathbf{H}_\sigma^p$ with $p=2$ and obtain that:

$$\begin{aligned} \mathbb{E}[⑥|\mathcal{F}_n] &\leq \frac{\gamma_{n+1}^2}{\varepsilon} \sigma_{n+1}^2 \|(w_n+\varepsilon)^{1/4}\|^2 \mathbb{E}[\|\zeta_{n+1}\|^2 |\mathcal{F}_n] \lesssim_d \frac{\gamma_{n+1}^2}{\varepsilon} \sigma_{n+1}^2 (d+f(\theta_n)) \|(w_n+\varepsilon)^{1/4}\|^2 \\ &\lesssim_d \frac{\gamma_{n+1}^2}{\varepsilon} d\sigma_{n+1}^2 f(\theta_n) \|(w_n+\varepsilon)^{1/4}\|^2. \end{aligned}$$

The last line comes from the fact that as soon as f is uniformly lower bounded by a positive constant, then $\|(w_n+\varepsilon)^{1/4}\|^2 \lesssim_d f(\theta_n)\|(w_n+\varepsilon)^{1/4}\|^2$.

The term ⑦ is close to ⑤ and we immediately get :

$$\begin{aligned} \mathbb{E}[⑦|\mathcal{F}_n] &= \frac{\gamma_{n+1}^3}{\varepsilon} \sigma_{n+1}^2 p_n \left\| \frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}} \right\|^2 \mathbb{E}[\|\zeta_{n+1}\|^2|\mathcal{F}_n] \\ &\leq \frac{\gamma_{n+1}^3}{\varepsilon\sqrt{\varepsilon}} \sigma_{n+1}^2 p_n \|\nabla f(\theta_n)\|^2 \mathbb{E}[\|\zeta_{n+1}\|^2|\mathcal{F}_n] \\ &\leq \mathbb{E}[⑤|\mathcal{F}_n]. \end{aligned}$$

For ⑧ Assumption $\mathbf{H}_\sigma^p$ with $p=4$ implies that:

$$\mathbb{E}[⑧|\mathcal{F}_n] \lesssim_d \frac{\gamma_{n+1}^3}{\varepsilon\sqrt{\varepsilon}} p_n \sigma_{n+1}^4 (d^2 + f(\theta_n)^2).$$

Our bounds on ① − ⑧ and the fact that $(1+f(\theta_n))^2 \leq 2(1+f(\theta_n)^2)$ ensure that a constant c_3 exists, independent from d and ε, such that:

$$\begin{aligned}
\mathbb{E}[f(\theta_{n+1})\|(w_{n+1}+\varepsilon)^{1/4}\|^2 \,|\mathcal{F}_n] &\leq f(\theta_n)\|(w_n+\varepsilon)^{1/4}\|^2 \left(1 + c_3\frac{\gamma_{n+1}^2}{\varepsilon}(1+d\sigma_{n+1}^2)\right) \\
&- \gamma_{n+1}\left\|\sqrt{|\nabla f(\theta_n)|}\right\|^4 + \gamma_{n+1}p_n f(\theta_n)\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2 \\
&+ c_3\left[\frac{\gamma_{n+1}}{\sqrt{\varepsilon}}p_n\left(\sigma_{n+1}^2 d + \gamma_{n+1}\sigma_{n+1}^3 d^{3/2} + \frac{\gamma_{n+1}^2}{\varepsilon}(1+\sigma_{n+1}^2 d + \sigma_{n+1}^4 d^2)\right)(1+f(\theta_n)^2)\right]
\end{aligned}$$

Since $\gamma_{n+1}\sigma_{n+1}^3 d^{3/2} \leq 1/2(\sigma_{n+1}^2 d + \gamma_{n+1}^2\sigma_{n+1}^4 d^2)$, using the definition of t_n, we deduce that:

$$\begin{aligned}
\mathbb{E}[f(\theta_{n+1})\|(w_{n+1}+\varepsilon)^{1/4}\|^2 \,|\mathcal{F}_n] &\leq f(\theta_n)\|(w_n+\varepsilon)^{1/4}\|^2 \left(1 + c_3\frac{\gamma_{n+1}^2}{\varepsilon}(\mathbb{1}_{\gamma_{n+1}>k_3} + d\sigma_{n+1}^2)\right) \\
&- \frac{\gamma_{n+1}}{2}\left\|\sqrt{|\nabla f(\theta_n)|}\right\|^4 + \gamma_{n+1}p_n f(\theta_n)\left\|\frac{\nabla f(\theta_n)}{(w_n+\varepsilon)^{1/4}}\right\|^2 \\
&+ c_3\left[t_n f(\theta_n)^2 + t_n\right].
\end{aligned}$$

■

Remark 5 *Proposition 4 will permit do derive a Lyapunov function on $(\theta_n, w_n)_{n\geq 1}$ (see the next result) which implies the convergence of:*

$$\mathbb{E}\left[\sum_{n\geq 1}\gamma_{n+1}\left\|\nabla f(\theta_n)\right\|^2\right] < +\infty.$$

This kind of bound has also been obtained by Défossez et al. (2020, Theorem 4) with the help of a somewhat artificial boundedness assumption of the noisy gradients, which is not used in our work. We also point out that Zou et al. (2019) propose another function that generates a mean reverting term:

$$\sum_{n\geq 1}\gamma_{n+1}\mathbb{E}[\|\nabla f(\theta_n)\|^{4/3}] < \infty,$$

and the major difference with our result is the weaker $4/3$ instead of 2 in the series. In particular, a such $4/3$ will not allow to prove the a.s. asymptotic pseudo-trajectory result, and consequently the a.s. convergence of the trajectory towards a critical point of f.

3.2 Proof of Theorem 2

Using Proposition 4, we are now ready to state the next important result, which will be key for the almost sure convergence of $(\theta_n)_{n\geq 1}$.

Proposition 6 *Under the assumptions of Proposition 4 :*

i) Two constants $c(\theta_0, w_0)$ and κ exist such that, for all $n \geq 1$:

$$\mathbb{E}\left(\sum_{k=1}^{n} \gamma_{k+1} q_k \|\sqrt{w_k}\|^2 + \gamma_{k+1} \|\sqrt{|\nabla f(\theta_k)|}\|^4\right) \leq (1 + V_{a,b}(\theta_0, w_0)) \exp\left(\kappa \sum_{k=1}^{n} (s_k + t_k)\right),$$

where $(t_n)_{n\geq 0}$ and $(s_n)_{n\geq 0}$ are the auxiliary sequences defined in Proposition 4.

ii) If $\sum_{n\geq 1}(\gamma_{n+1}^2 \sigma_{n+1}^2 + \gamma_{n+1}\sigma_{n+1}^2 p_n) < +\infty$, then almost surely:

$$\sum_{n\geq 1} \left[\gamma_{n+1} q_n \|\sqrt{w_n}\|^2 + \gamma_{n+1} \|\sqrt{|\nabla f(\theta_n)|}\|^4\right] < +\infty. \tag{13}$$

iii) Suppose that additionally ($\mathbf{H}'_{Steps}$−1) holds. Then $\|\sqrt{w_n}\|^2 \upsilon_n$ is almost surely bounded, where υ_n is introduced in $\mathbf{H}'_{Steps}$.

Proof • Proof of *i)*. Our proof relies on a Lyapunov function defined by:

$$V_{a,b}(\theta, w) := \|\sqrt{w + \varepsilon}\|^2 + a f(\theta)^2 + b f(\theta) \|(w + \varepsilon)^{1/4}\|^2,$$

with a careful tuning of a and b.

Using *i)*, *iii)* and *iv)* of Proposition 4 and the fact that $f(\theta_n) \leq 1 + f(\theta_n)^2$, we deduce that a constant κ that depends on c_1, c_2, c_3 and of the next choice of a and b exists such that:

$$\begin{aligned}
\mathbb{E}\left[V_{a,b}(\theta_{n+1}, w_{n+1}) \,|\mathcal{F}_n\right] \leq\ & V_{a,b}(\theta_n, w_n)\left[1 + \kappa(s_n + t_n)\right] + \kappa(s_n + t_n) \\
& - q_n \gamma_{n+1} \|\sqrt{w_n}\|^2 \\
& + \left[\gamma_{n+1} p_n \|\nabla f(\theta_n)\|^2 - b\frac{\gamma_{n+1}}{2} \|\sqrt{|\nabla f(\theta_n)|}\|^4\right] \\
& + [b\gamma_{n+1} p_n - a\frac{\gamma_{n+1}}{2}] f(\theta_n) \left\|\frac{\nabla f(\theta_n)}{(w_n + \varepsilon)^{1/4}}\right\|^2.
\end{aligned}$$

We observe that for any vector u, we have $\|\sqrt{|u|}\|^4 \geq \|u\|^2$ and that $(p_n)_{n\geq 1}$ is a bounded sequence, so that we can find b large enough $b > 4 \sup_{n\geq 1} p_n$, and $a/2 \geq b p_n$, such that:

$$\begin{aligned}
\forall n \geq 1 \quad \mathbb{E}\left[V_{a,b}(\theta_{n+1}, w_{n+1}) \,|\mathcal{F}_n\right] \leq\ & V_{a,b}(\theta_n, w_n)\left[1 + \kappa(t_n + s_n)\right] + \kappa(t_n + s_n) \\
& - \gamma_{n+1}\left[q_n \|\sqrt{w_n}\|^2 + \frac{b}{4} \|\sqrt{|\nabla f(\theta_n)|}\|^4\right].
\end{aligned} \tag{14}$$

We introduce the products $(\pi_j)_{j\geq 0}$ defined by:

$$\pi_j = \prod_{u=1}^{j} (1 + \kappa(t_u + s_u)) = \pi_{j-1}\left(1 + \kappa(t_j + s_j)\right).$$

We observe that $\kappa(t_j + s_j) = \pi_{j-1}^{-1}[\pi_j - \pi_{j-1}]$. A straightforward recursion, associated with a sequence of conditional expectation argument, yields:

$$\begin{aligned}
\mathbb{E}\left[V_{a,b}(\theta_{n+1}, w_{n+1})\right] &\leq V_{a,b}(\theta_0, w_0)\pi_n + \sum_{j=1}^{n} \kappa(t_j + s_j) \prod_{u=j+1}^{n} (1 + \kappa(t_u + s_u)) \\
&\quad - \mathbb{E}\left(\sum_{j=1}^{n} \gamma_{j+1} q_j \|\sqrt{w_j}\|^2 + \frac{b}{4} \sum_{j=1}^{n} \gamma_{j+1} \|\sqrt{|\nabla f(\theta_j)|}\|^4\right) \\
&\leq V_{a,b}(\theta_0, w_0)\pi_n + \sum_{j=1}^{n} \frac{\pi_j - \pi_{j-1}}{\pi_{j-1}} \times \frac{\pi_n}{\pi_j} \\
&\quad - \mathbb{E}\left(\sum_{j=1}^{n} \gamma_{j+1} q_j \|\sqrt{w_j}\|^2 + \frac{b}{4} \sum_{j=1}^{n} \gamma_{j+1} \|\sqrt{|\nabla f(\theta_j)|}\|^4\right) \\
&\leq V_{a,b}(\theta_0, w_0)\pi_n + \pi_n \sum_{j=1}^{n} (\pi_{j-1}^{-1} - \pi_j^{-1}) \\
&\quad - \mathbb{E}\left(\sum_{j=1}^{n} \gamma_{j+1} q_j \|\sqrt{w_j}\|^2 + \frac{b}{4} \sum_{j=1}^{n} \gamma_{j+1} \|\sqrt{|\nabla f(\theta_j)|}\|^4\right)
\end{aligned}$$

A telescopic sum argument then shows that:

$$\mathbb{E}\left(\sum_{j=1}^{n} \gamma_{j+1} q_j \|\sqrt{w_j}\|^2 + \frac{b}{4} \sum_{j=1}^{n} \gamma_{j+1} \|\sqrt{|\nabla f(\theta_j)|}\|^4\right) \leq (1 + V_{a,b}(\theta_0, w_0))\pi_n.$$

We conclude using $1 + x \leq e^x$, which entails $\pi_n \leq \exp\left(\kappa \sum_{j=1}^{n} (t_j + s_j)\right)$.

$\diamond$

• Proof of *ii*). This point proceeds with standard arguments: we use the Robbins-Siegmund Lemma: using our assumption on the series, we deduce that $\sum t_n$ and $\sum s_n$ are convergent and we obtain that:

1. $V_{a,b}(\theta_n, w_n) \to V_\infty$ a.s. (and in L^1) and $\sup_n \mathbb{E}[V_{a,b}(\theta_n, w_n)] < +\infty$
2. More importantly, the next series are convergent:

$$\sum_{n \geq 0} \gamma_{n+1} \|\nabla f(\theta_n)\|^2 < \sum_{n \geq 0} \gamma_{n+1} \|\sqrt{|\nabla f(\theta_n)|}\|^4 < +\infty \ a.s. \tag{15}$$

and

$$\sum_{n \geq 0} \gamma_{n+1} q_n \|\sqrt{w_n}\|^2 < +\infty \ a.s.. \tag{16}$$

This ends the proof of *ii*). $\diamond$

• Proof of *iii*). This point is meant to also encompass the particular case corresponding to Adagrad with parameters $\beta_2 = 1$ and α constant, which according to Table 1 can be

recovered with our parametrization by setting $\gamma_n = q_n = p_n = \frac{1}{\sqrt{n}}$.
We study the evolution of the sequence $\varphi_n = \|w_n\|^2 \upsilon_n$, and we will use the Robbins-Siegmund theorem in order to prove its convergence. With that in mind, we start by writing a recursive relation on φ_n:

$$\begin{aligned}
\varphi_{n+1} &= \upsilon_{n+1} \|\sqrt{w_{n+1}}\|^2 \\
&= \upsilon_{n+1} \left[\|\sqrt{w_n}\|^2 + p_n \gamma_{n+1} \|g_{n+1}\|^2 - q_n \gamma_{n+1} \|\sqrt{w_n}\|^2 \right] \\
&= \upsilon_n \|\sqrt{w_n}\|^2 \left[\frac{\upsilon_{n+1}}{\upsilon_n} (1 - q_n \gamma_{n+1}) \right] + p_n \gamma_{n+1} \upsilon_{n+1} \|g_{n+1}\|^2 \\
&= \varphi_n \left[\frac{\upsilon_{n+1}}{\upsilon_n} (1 - q_n \gamma_{n+1}) \right] + p_n \gamma_{n+1} \upsilon_{n+1} \|g_{n+1}\|^2.
\end{aligned}$$

Therefore, as soon as $\frac{\upsilon_{n+1}}{\upsilon_n}(1 - q_n \gamma_{n+1}) \leq 1$, we have that:

$$\mathbb{E}[\varphi_{n+1} | \mathcal{F}_n] \leq \varphi_n + p_n \gamma_{n+1} \upsilon_{n+1} \mathbb{E}[\|g_{n+1}\|^2 | \mathcal{F}_n]$$

In order to apply the Robbins-Siegmund theorem we study the following series:

$$\begin{aligned}
\sum_{n \geq 1} p_n \gamma_{n+1} \upsilon_{n+1} \mathbb{E}[\|g_{n+1}\|^2 | \mathcal{F}_n] &= \sum_{n \geq 1} p_n \gamma_{n+1} \upsilon_{n+1} \left(\|\nabla f(\theta_n)\|^2 + \mathbb{E}[\|\xi_{n+1}\|^2 | \mathcal{F}_n] \right) \\
&\leq \sum_{n \geq 1} p_n \upsilon_{n+1} \gamma_{n+1} \|\nabla f(\theta_n)\|^2 + \sum_{n \geq 1} p_n \gamma_{n+1} \upsilon_{n+1} \mathbb{E}[\|\xi_{n+1}\|^2 | \mathcal{F}_n].
\end{aligned}$$

As soon as there exists a constant c such that $p_n / \upsilon_{n+1} \leq c$, the first term is bounded by $c \sum_{n \geq 1} \gamma_{n+1} \|\nabla f(\theta_n)\|^2$ which, according to $ii)$, is almost surely finite. For the second term, we use the noise assumption $\mathbf{H}^2_\sigma$:

$$\begin{aligned}
\sum_{n \geq 1} p_n \gamma_{n+1} \upsilon_{n+1} \mathbb{E}[\|\xi_{n+1}\|^2 | \mathcal{F}_n] &\leq \sum_{n \geq 1} p_n \gamma_{n+1} \upsilon_{n+1} \sigma^2_{n+1} (d + f(\theta_n)) \\
&\leq (d + \sup_{n \geq 1} f(\theta_n)) \sum_{n \leq 1} p_n \gamma_{n+1} \sigma^2_{n+1} \upsilon_{n+1}
\end{aligned}$$

Since by $ii)$ $V_{a,b}(\theta_n, w_n)$ is almost surely bounded, so is $f(\theta_n)$. Thus, when the series $\sum_{n \leq 1} p_n \gamma_{n+1} \sigma^2_{n+1} \upsilon_{n+1} < +\infty$, we can conclude that

$$\sum_{n \geq 1} p_n \gamma_{n+1} \upsilon_{n+1} \mathbb{E}[\|g_{n+1}\|^2 | \mathcal{F}_n] < +\infty \quad \text{almost surely.}$$

Finally, we can apply the Robbins Siegmund theorem to obtain that $\varphi_n \to \varphi_\infty$ a.s. and that $\varphi_\infty < +\infty$ a.s. which ends the proof. ■

We emphasize that $i), ii)$ are standard consequences of the Robbins-Siegmund approach on stochastic algorithms. In particular, the use of $i)$ with the calibrations of $(\gamma_n)_{1 \leq n \leq N}$, $(\sigma_n)_{1 \leq n \leq N}$ and $(p_n)_{1 \leq n \leq N}$ that are given in the statement of Theorem 2 instantaneously

lead to the conclusion of the proof. We also point out that the basic fact to obtain this result is a tuning of the parameters that leads to:

$$\sum_{n=1}^{N} (s_n \vee t_n) \lesssim_d 1.$$

Considering constant step-size sequences, it entails the following constraints on (p, γ, σ):

$$\frac{\gamma^2}{\varepsilon} \vee \frac{\gamma p}{\sqrt{\varepsilon}} \sigma^2 d \vee \frac{\gamma^2}{\varepsilon} \sigma^2 d \lesssim_d 1,$$

whereas the size of N needed to obtain a δ approximation should verify that:

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=1}^{N} \|\sqrt{|\nabla f(\theta_k)|}\|^4\right] \leq \delta \iff \frac{1}{N\gamma}\mathbb{E}\left[\sum_{k=1}^{N} \gamma \|\sqrt{|\nabla f(\theta_k)|}\|^4\right] \leq \delta \impliedby N \geq (\delta\gamma)^{-1}.$$

The computational cost associated to a such N is then $N\sigma^{-2}$ since each iteration generates σ^{-2} computations. We then observe that the two alternative tunings of the parameters suggested in *i*) and *ii*) of Theorem 2 lead to a $d\epsilon^{-1}\delta^2$ computational cost to obtain a δ approximation.

We can finally address the bound we obtain for the Rmsprop situation, which corresponds to a constant step-size algorithm with $\gamma = \alpha\sqrt{1-\beta_2}$ and $p = \sqrt{1-\beta_2}\alpha^{-1}$. We then recover a $d\varepsilon^{-1}\delta^{-2}$ computational cost with the following tuning of the parameters: $\alpha = \sqrt{\varepsilon}$ and $1 - \beta_2 = \sqrt{\varepsilon}/(nd)$.

3.3 Proof of Theorem 1

For the sake of convenience, from now on we denote $V_n = V_{a,b}(\theta_n, w_n)$. Since we are now interested in purely asymptotic results, we omit the dependency on d in the bounds we obtain hereafter. The main difficulty here is to convert the result of Proposition 6 *ii*) into an a.s. convergence result on $(\theta_n)_{n\geq 1}$.

3.3.1 Asymptotic Pseudo-trajectory

We remind the standard definition of asymptotic pseudo-trajectories of a semiflow Φ.

Definition 7 (Pseudo-trajectory) *A continuous trajectory Z is a pseudo-trajectory of Φ if for any finite time horizon $T > 0$*

$$\lim_{t \longrightarrow +\infty} \sup_{0<u<T} |Z_{t+u} - \Phi_u(Z_t)| = 0.$$

We refer to Benaïm and Hirsch (1996); Benaïm (1999) for further details. We will use in particular the cornerstone result which is reminded below:

Theorem 8 (Theorem 3.2 of Benaïm (1999)) *If Z is an asymptotic pseudo-trajectory of Φ with a compact closure in $\mathbb{R}^d$, then every limit point of $(Z(t+.))_{t\geq 0}$ when $t \to +\infty$ (equipped with the topology of uniform convergence on compact sets) is a trajectory induced by the semiflow Φ.*

In what follows we use the results of Benaïm (1999) to show that a linear interpolation of the sequence $(Z_n)_{n\geq}$ is an asymptotic pseudo-trajectory of the flow induced by the vector field $H : \mathbb{R}^{2d} \to \mathbb{R}^{2d}$ of our adaptive algorithm defined by:

$$H(\theta, w) = \begin{pmatrix} -\dfrac{\nabla f(\theta)}{\sqrt{w+\varepsilon}} \\ p_\infty \nabla f(\theta)^{\odot 2} - q_\infty w \end{pmatrix}. \tag{17}$$

For this part, the value of q_∞ is of no importance, so we can treat at the same time the two regimes ($q_\infty > 0$ or $q_\infty = 0$). The main difference between the two, is in the convergence proof for the coordinate w_n and this will be highlighted at the appropriate moment.
Denote $\tau_0 = 0$, $\tau_n = \sum_{k=1}^n \gamma_k$ and consider $(\bar{Z}_t)_{t\geq 0}$ the continuous time process corresponding to a linear interpolation of $(Z_n)_{n\geq 0}$, given by:

$$\bar{Z}(\tau_n + s) = Z_n + s\frac{Z_{n+1} - Z_n}{\gamma_{n+1}}; \quad \forall n \in \mathbb{N};\ \forall\ 0 \leq s \leq \gamma_{n+1}.$$

The Robbins-Siegmund Lemma ensures that the sequence $(V_n)_{n\geq 1}$ converges a.s. to a finite random variable V_∞. Since all the terms of V_n are positive, $(f(\theta_n))_{n\geq 0}$ and $(\|w_n\|)_{n\geq 0}$ are a.s. bounded as well. The coercivity of f implies thus that $(Z_n)_{n\geq 0}$ is a.s. bounded.
The evolution of the sequence $(Z_n)_{n\geq 0}$ can be written as:

$$Z_{n+1} = Z_n + \gamma_{n+1}(H(Z_n) + e_n),$$

where e_n is a rest term:

$$e_n = \begin{pmatrix} \dfrac{\nabla f(\theta_n) - g_{n+1}}{\sqrt{w_n+\varepsilon}} \\ p_n g_{n+1}^{\odot 2} - p_\infty \nabla f(\theta_n)^{\odot 2} - (q_n - q_\infty) w_n \end{pmatrix}. \tag{18}$$

The next result allows us to control the rest term and ensures that the assumptions of Proposition 4.1 of Benaïm (1999) are satisfied, which in turn implies that $(\bar{Z}_t)_{t\geq 0}$ is indeed an asymptotic pseudo trajectory of the flow induced by H.

Lemma 9 *Suppose that the assumptions of Proposition 4 and* $(\mathbf{H}_{Steps} - 1)$ *hold and that* $\sum \gamma_{k+1}\sigma_{k+1}^2 p_k < +\infty$. *Let* $N(n,t) = \sup_{k\geq 0}\{t + \tau_n \geq \tau_k\}$ *and* $(e_n)_{n\geq 1}$ *defined in Equation* (18). *Then, for all* $T > 0$*:*

$$\limsup_{n\to\infty} \sup_{t\in[0,T]} \left\| \sum_{k=n+1}^{N(n,t)+1} \gamma_k e_k \right\| = 0 \ a.s.$$

Proof We consider a finite horizon $T > 0$. In order to deal with the previous sum, we write e_n as $a_n + \Delta M_{n+1} - c_n$, with $a_n = \begin{pmatrix} 0 \\ (p_n - p_\infty)\nabla f(\theta_n)^{\odot 2} - (q_n - q_\infty) w_n \end{pmatrix}$, $\Delta M_{n+1} = \begin{pmatrix} \dfrac{\nabla f(\theta_n) - g_{n+1}}{\sqrt{w_n+\varepsilon}} \\ 0 \end{pmatrix}$ and $c_n = \begin{pmatrix} 0 \\ p_n(\nabla f(\theta_n)^{\odot 2} - g_{n+1}^{\odot 2}) \end{pmatrix}$. We use the fact that:

$$\left\| \sum_{k=n+1}^{N(n,t)+1} \gamma_k e_k \right\| \leq \left\| \sum_{k=n+1}^{N(n,t)+1} \gamma_k a_k \right\| + \left\| \sum_{k=n+1}^{N(n,t)+1} \gamma_k \Delta M_{k+1} \right\| + \left\| \sum_{k=n+1}^{N(n,t)+1} \gamma_k c_k \right\|,$$

and proceed to upper bound each term of the right hand side.
• The convergence of the first term is mainly a consequence of Equations (15), (16) (convergence of the Robbins-Siegmund series), of the convergence of $(q_n)_{n\geq 1}$ towards q_∞ and the fact that $(p_n - p_\infty)_{n\geq 1}$ is a bounded sequence:

$$\begin{aligned}\left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_k a_k\right\| &\leq \left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_k\left((p_k-p_\infty)\nabla f(\theta_k)^{\odot 2}-(q_k-q_\infty)w_k\right)\right\| \\ &\leq \left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_k(p_k-p_\infty)\nabla f(\theta_k)^{\odot 2}\right\| + \left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_k(q_k-q_\infty)w_k\right\| \\ &\leq \sum_{k=n+1}^{N(n,t)+1}\gamma_k|p_k-p_\infty|\left\|\nabla f(\theta_k)^{\odot 2}\right\| + \sum_{k=n+1}^{N(n,t)+1}\gamma_k|q_k-q_\infty|\left\|w_k\right\|.\end{aligned}$$

Assumption ($\mathbf{H}_{\text{Steps}}-1$) (or ($\mathbf{H}'_{\text{Steps}}-1$)) ensures that a constant P exists such that $|p_n - p_\infty| < P$ and that for n large enough $|q_n - q_\infty| \leq q_n$. Moreover, $\forall a, b \in \mathbb{R}_+$, $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ thus $\|\nabla f(\theta_n)^{\odot 2}\| \leq \|\nabla f(\theta_n)\|^2$ and $\|w_n\| \leq \|\sqrt{w_n}\|^2$. Inserting these bounds in the previous inequality gives:

$$\left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_k a_k\right\| \leq \sum_{k=n+1}^{N(n,t)+1}P\gamma_k\left\|\nabla f(\theta_k)\right\|^2 + \sum_{k=n+1}^{N(n,t)+1}\gamma_k q_k\|\sqrt{w_k}\|^2.$$

Using the convergence of the series (15) and (16) we conclude that, $\forall t > 0$[3]:

$$\limsup_{n\to\infty}\left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_k a_k\right\| = 0.$$

• To control the second term we observe that $(\gamma_{k+1}\Delta M_{k+1})_{k\geq 0}$ is a sequence of martingale increments, since

$$\forall k \geq 1 \qquad \mathbb{E}\left[\gamma_{k+1}\frac{\xi_{k+1}}{\sqrt{w_k+\varepsilon}}\,|\mathcal{F}_k\right] = 0,$$

that the associated martingale $M_n = \sum_{k=1}^n \gamma_k \Delta M_k$ is square integrable and that furthermore:

$$\mathbb{E}[\|M_{n+1}-M_n\|^2|\mathcal{F}_n] = \mathbb{E}\left[\gamma_{n+1}^2\left\|\frac{\xi_{n+1}}{\sqrt{w_k+\varepsilon}}\right\|^2|\mathcal{F}_n\right] \leq \gamma_{n+1}^2\frac{\sigma_{n+1}^2}{\varepsilon}(d+f(\theta_n))$$

Since $\sup_n f(\theta_n)$ is almost surely bounded, $\sum_{n\geq 1}\mathbb{E}[\|M_{n+1}-M_n\|^2|\mathcal{F}_n] < +\infty$ a.s. and according to Theorem 1.3.11 of (Duflo, 1996) we can conclude that M_n converges a.s. to a finite random vector M_∞, and so $\sum_{n\geq 1}\gamma_n\Delta M_n < +\infty$ a.s.. This implies that:

$$\forall t > 0 \qquad \limsup_{n\to\infty}\left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_k\Delta M_{k+1}\right\| = 0.$$

3. This last limit holds regardless $t < T$

• The assumptions made on the noise sequence and the fact that $(\mathbb{E}[V_n])_{n\geq 1}$ is uniformly bounded ($V_n \longrightarrow V_\infty$ in L^1) imply that the last term can be handled using the same type of arguments.

We start by decomposing c_k as its expected value plus a martingale increment:

$$\left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_k c_k\right\| \leq \left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_{k+1}p_k\mathbb{E}[\xi_{k+1}^{\odot 2}\,|\mathcal{F}_k]\right\| + \left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_{k+1}p_k\left(2\xi_{k+1}\cdot\nabla f(\theta_k)+\xi_{k+1}^{\odot 2}-\mathbb{E}[\xi_{k+1}^{\odot 2}|\mathcal{F}_k]\right)\right\|.$$

Since $\sum_{n\geq 0}\gamma_{n+1}p_n\sigma_{n+1}^2 < +\infty$ the first sum converges to 0 when n goes to infinity.

The terms of the second sum are martingale increments:

$$\mathbb{E}[2\xi_{k+1}\cdot\nabla f(\theta_k)+\xi_{k+1}^{\odot 2}-\mathbb{E}[\xi_{k+1}^{\odot 2}\,|\mathcal{F}_k]\,|\mathcal{F}_k]=0.$$

Denote $\tilde{M}_{n+1} := \sum_{k=1}^{n}\gamma_{k+1}p_k\left(2\xi_{k+1}\cdot\nabla f(\theta_k)+\xi_{k+1}^{\odot 2}-\mathbb{E}[\xi_{k+1}^{\odot 2}|\mathcal{F}_k]\right)$ the associated martingale. Using the fact that $(a+b)^2\leq 2(a^2+b^2)$, we get a first bound on the conditional second order moments of the increments:

$$\mathbb{E}[\|2\xi_{k+1}\cdot\nabla f(\theta_k)+\xi_{k+1}^{\odot 2}-\mathbb{E}[\xi_{k+1}^{\odot 2}|\mathcal{F}_k]\|^2|\mathcal{F}_k]\leq 8\left(\mathbb{E}[\|\xi_{k+1}\cdot\nabla f(\theta_k)\|^2|\mathcal{F}_k]+\mathbb{E}[\|\xi_{k+1}^{\odot 2}\|^2|\mathcal{F}_k]\right).$$

Now using $\mathbf{H}_\sigma^p$ for $p=2$ and Inequality (9):

$$\begin{aligned}\mathbb{E}[\|\xi_{k+1}\cdot\nabla f(\theta_k)\|^2|\mathcal{F}_k] &= \sum_{i=1}^{d}\partial_i f(\theta_k)^2\mathbb{E}[\xi_{k+1,i}^2|\mathcal{F}_k]\leq c\sigma_{k+1}^2(d+f(\theta_k))\|\nabla f(\theta_k)\|^2\\ &\leq c_f\sigma_{k+1}^2(d+f(\theta_k))^2.\end{aligned}$$

The second term can be dealt with in a similar manner, using the assumption $\mathbf{H}_\sigma^p$ for $p=4$:

$$\mathbb{E}[\|\xi_{k+1}^{\odot 2}\|^2|\mathcal{F}_k]\leq c\sigma_{k+1}^4\mathbb{E}[\|\zeta_{k+1}\|^4|\mathcal{F}_k]\leq c\sigma_{k+1}^4(d+f(\theta_k))^2$$

Denoting by $m=\sup_{n\geq 1}(d+f(\theta_n))^2$ which is almost surely bounded, we have that :

$$\mathbb{E}[\|\Delta\tilde{M}_{n+1}\|^2|\mathcal{F}_n]\lesssim m\gamma_{n+1}^2p_n^2(\sigma_{n+1}^2+\sigma_{n+1}^4).$$

Hence, as soon as $\sum_{n\geq 1}\gamma_{n+1}^2p_n^2(\sigma_{n+1}^2+\sigma_{n+1}^4)<+\infty$ Theorem 1.3.11 of (Duflo, 1996) implies that the sequence $\tilde{M}_n$ converges almost surely to a finite random vector. In other words $\sum_{k=1}^{+\infty}\gamma_k c_k<+\infty$ almost surely and so:

$$\limsup_{n\to\infty}\left\|\sum_{k=n+1}^{N(n,t)+1}\gamma_k c_k\right\|=0,$$

which ends the proof. ■

3.3.2 Proof of the Almost Sure Convergence Towards a Critical Point

We now give the proof of the leading result of Section 3.

Proof The proof is divided into tree steps.

• Identification of the possible limit points. The a.s. boundness of $(Z_n)_{n\geq 0}$ and Lemma 9 show that the assumptions of Proposition 4.1 of Benaïm (1999) hold, which implies that $(\bar{Z}_t)_{t\geq 0}$ is an asymptotic pseudo-trajectory of the differential flow induced by H (defined in (17)) almost surely.

We shall deduce from Theorem 8 that all limit points $Z_\infty = (\theta_\infty, w_\infty)$ of $\bar{Z}_t$ are stationary points for the differential equation $\dot{z} = H(z)$ and thus that $H(Z_\infty) = 0$. Since $\|w_n\|$ is a.s. bounded, the first coordinate of the equation $H(Z_\infty) = 0$ implies that $\nabla f(\theta_\infty) = 0$.

• Convergence of $(w)_{n\geq 1}$. For this point, the two different regimes $[q_\infty > 0$ with $\mathbf{H}_{\text{Steps}}]$ and $[q_\infty = 0$ with $\mathbf{H}'_{\text{Steps}}]$, are treated separately.

First, suppose that $q_\infty = 0$. Under $\mathbf{H}'_{\text{Steps}}$, Proposition 6 *iii)* implies that $(w_n \upsilon_n)$ is almost surely bounded. Thus, since $\upsilon_n \to +\infty$, we obtain that w_n also converges to 0 almost surely.

Now, suppose that $q_\infty > 0$. Since $\nabla f(\theta_n) \to 0$ when $n \to \infty$, we observe that for any limit point Z_∞, the second coordinate of $H(Z_\infty) = 0$ also implies that $w_\infty = 0$, as soon as $q_\infty > 0$. We have shown that w_n is almost surely bounded. Moreover, $V_{a,b}$ given in Proposition 6 being a strict Lyapunov function for the O.D.E. $\dot{z} = H(z)$, we deduce from Corollary 6.6 of Benaïm (1999) that the limit set of the O.D.E. is included in $H(Z_\infty) = 0$. Hence, $(\bar{Z}(\tau_n + .))_{n\geq 0}$ being an asymptotic pseudo-trajectory of the O.D.E., Theorem 3.2 of Benaïm (1999) implies that every adherence point of the w coordinate of $(Z_n)_{n\geq 0}$ is necessarily zero, since the limit sets of the O.D.E. satisfies $w_\infty = 0$. Therefore, $(w_n)_{n\geq 0}$ is a compact sequence with 0 as a unique adherence value, which implies that $(w_n)_{n\geq 0}$ almost surely converges to 0.

• Convergence of $(\theta_n)_{n\geq 1}$.

We shall deduce from this last result a convergence on the overall sequence $(\theta_n)_{n\geq 0}$.

$$\begin{aligned}\lim_{n\to\infty} V_{a,b}(\theta_n, w_n) &= \lim_{n\to\infty} \|\sqrt{w_n + \varepsilon}\|^2 + af(\theta_n)^2 + bf(\theta_n)\|(w_n + \varepsilon)^{1/4}\|^2 \\ &= \lim_{n\to\infty} af(\theta_n)^2 + bd\sqrt{\varepsilon} f(\theta_n) + d\varepsilon = V_\infty.\end{aligned}$$

Since a and b are non-negative and $f(\theta_n)$ positive, the last equality implies that a real value f_∞ exists such that the sequence $(f(\theta_n))_{n\geq 0}$ is a.s. convergent towards it:

$$\lim_{n\longrightarrow+\infty} f(\theta_n) = f_\infty \quad a.s.$$

Now, $(\theta_n)_{n\geq 0}$ is an a.s. bounded sequence and we prove that the set of possible limit points for its sub-sequences is connected. We study

$$\begin{aligned}\|\theta_{n+1} - \theta_n\|^2 &= \left\| -\gamma_{n+1} \frac{g_{n+1}}{\sqrt{w_n + \varepsilon}} \right\|^2 \\ &\leq \frac{1}{\varepsilon} \gamma_{n+1}^2 \|g_{n+1}\|^2 \\ &\leq 2\frac{\gamma_{n+1}^2}{\varepsilon} \left(\|\nabla f(\theta_n)\|^2 + \sigma_{n+1} \|\zeta_{n+1}\|^2 \right).\end{aligned}$$

We then deduce from $\mathbf{H}_\sigma^2$ and the Cauchy-Schwarz inequality that:

$$\mathbb{E}[\|\theta_{n+1}-\theta_n\|^2] \leq 2\frac{\gamma_{n+1}^2}{\varepsilon}\left(\mathbb{E}[\|\nabla f(\theta_n)\|^2] + c\sigma_{n+1}^2(d+\mathbb{E}[f(\theta_n)])\right)$$

Using the fact that $\sup_n \mathbb{E}[V_{a,b}(\theta_n,w_n)] < +\infty$, we deduce that a constant K exists such that:

$$\mathbb{E}[\|\theta_{n+1}-\theta_n\|^2] \leq K\gamma_{n+1}^2\mathbb{E}[\|\nabla f(\theta_n)\|^2] + K\gamma_{n+1}^2\sigma_{n+1}^2,$$

Another consequence of the Robbins-Siegmund Theorem in *ii*) of Proposition 4 is that $\sum_{n\geq 1}\gamma_{n+1}\mathbb{E}[\|\nabla f(\theta_n)\|^2] < +\infty$ which implies with $\mathbf{H}_{\text{Steps}}$ that:

$$\sum_{n\geq 1}\mathbb{E}[\|\theta_{n+1}-\theta_n\|^2] < +\infty.$$

The Borel-Cantelli Lemma then yields $\theta_{n+1}-\theta_n \longrightarrow 0$ almost surely, so that the set of possible limit points for its sub-sequences is connected.

We denote below by A the closed set of adherence points of $(\theta_n)_{n\geq}$:

$$A = \bigcap_{n\geq 0}\overline{\{\theta_k, k\geq n\}}.$$

If A contains two different adherence points $\theta_\infty^0 \neq \theta_\infty^1$, then a continuous path (θ_∞^s) of adherence points, from θ_∞^0 to θ_∞^1, would exist since A connected. But we also know that

$$A \subset \{\theta : f(\theta) = f_\infty\} \cap \{\theta : \nabla f(\theta) = 0\}.$$

Since $\{\theta : f(\theta) = f_\infty\} \cap \{\theta : \nabla f(\theta) = 0\}$ is locally finite, we observe that A cannot be connected, which is a contradiction. Hence, A is reduced to a singleton, and $(\theta_n)_{n\geq 1}$ is a convergent sequence:

$$\lim_{n\longrightarrow+\infty}\theta_n = \theta_\infty \qquad \text{with} \qquad \nabla f(\theta_\infty) = 0.$$

■

4. Almost Sure Convergence and Traps Avoidance

In this paragraph, we prove that the sequence $(\theta_n)_{n\geq 1}$ almost surely converges towards a critical point that is not linearly unstable. In particular, if we assume the hyperbolicity of the equilibria of the dynamical system, *i.e.* all the eigenvalues of the Hessian are non-zero around critical points of f, it is equivalent to the almost sure convergence towards a local minimum.

To make our study meaningful, we restrict our analysis to the situation where Theorem 1 holds, *i.e* we assume that either $\mathbf{H}_{\text{Steps}}$ or $\mathbf{H}'_{\text{Steps}}$ hold.

4.1 Unstable Equilibria

We begin with a simple statement that identifies the unstable hyperbolic points of the dynamical system

$$(\dot{\theta}_t, \dot{w}_t) = H(\theta_t, w_t) \quad \text{with} \quad H(\theta, w) = \left(-\frac{\nabla f(\theta)}{\sqrt{w+\varepsilon}}, p_\infty[\nabla f(\theta)^{\odot 2}] - q_\infty w \right). \tag{19}$$

These equilibria may correspond to purely repulsive equilibria or saddle points. The next proposition makes this last sentence more precise and introduces the cornerstone function η that measures in each neighborhood of an unstable (or saddle) point the distance of any point x to the local stable manifold in the expanding direction.

Proposition 10 *Consider the dynamical system* (19)*, and assume that f is twice differentiable, then:*

(i) If $q_\infty > 0$, the equilibria are $(t, 0)$ where t is a critical point of f.

(ii) If t is a local maximum of f and $D^2 f(t)$ has non-zero eigenvalues, the dynamical system is unstable near $(t, 0)$.

(iii) If t is a local minimum of f and $D^2 f(t)$ has non-zero eigenvalues, the dynamical system is stable near $(t, 0)$.

(iv) If t is a linearly unstable equilibria ($D^2 f(t)$ has at least one negative eigenvalue), then a compact neighborhood $\mathcal{N}$ of $(t, 0)$ and a $\mathcal{C}^2$ positive function η exist such that:

(a) $\forall z \in \mathcal{N} \quad \forall u \in \mathbb{R}^d \times \mathbb{R}^d \quad \eta(z+u) \geq \eta(z) + \langle \nabla \eta(z), u \rangle - \Gamma \|u\|^2$.

(b) If E_+ is the eigenspace associated to the negative eigenvalues of $D^2 f(t)$ and π_+ is the orthogonal projection on E_+, then:

$$\forall z \in \mathcal{N} \quad \forall u = (u_1, u_2) \in \mathbb{R}^d \times \mathbb{R}^d \quad \lfloor \langle \nabla \eta(z), u \rangle \rfloor_+ \geq c_1 \|\pi_+(u_1)\|.$$

(c) A constant $\kappa > 0$ exists such that:

$$\forall z \in \mathcal{N} \qquad \langle \nabla \eta(z), H(z) \rangle \geq \kappa \eta(z). \tag{20}$$

Proof • Proof of *i*). The proof of *i*) is immediate by observing that $H(\theta, w) = 0$ and $q_\infty \neq 0$ implies that $\nabla f(\theta) = 0$ and $w = 0$.
• Proof of *ii*) and *iii*). We use a linearization of the drift around an equilibria $(t, 0)$. Since t is a critical point of f, we observe that:

$$\nabla f(t+h) = D^2 f(t) h + o(\|h\|).$$

Consequently, we observe that $(\nabla f(t+h))^2 = (D^2 f(t) h)^2 = \mathcal{O}\left(\|h\|^2\right)$, which entails:

$$\begin{aligned} H(t+h, \omega) &= \left(-\frac{D^2 f(t) h}{\sqrt{\varepsilon}} + \mathcal{O}\left(\|h\| \|\omega\|\right), p_\infty \mathcal{O}(\|h\|^2) - q_\infty \omega \right) \\ &= \begin{pmatrix} -\frac{D^2 f(t)}{\sqrt{\varepsilon}} & 0 \\ 0 & -q_\infty I_d \end{pmatrix} \begin{pmatrix} h \\ \omega \end{pmatrix} + o\left(\|h\| + \|\omega\|\right). \end{aligned}$$

The conclusion follows from the spectral decomposition of $D^2 f(t)$.
• Proof of *iv*). The last point is a consequence of Benaim and Hirsch (1995, Proposition 3.1) or Benaïm (1999, Proposition 9.5). We only need to observe that the coordinates in u_2 always correspond to the attractive manifold so that $\pi_+(u) = \pi_+(u_1)$. ■

Remark 11 *This proposition holds for any $p_\infty \in \mathbb{R}_+$. In particular it does so for $p_\infty = 0$.*

4.2 Preliminary Estimates

In what follows, we establish the non-convergence towards "'traps". More specifically, we consider any position t such that $D^2 f(t)$ has a negative eigenvalue and $\mathcal{N}$ a neighborhood of $(t, 0)$ given by (iv) of Proposition 10. In particular, $(t, 0)$ corresponds to a linearly unstable point.
For a given integer n_0 when (θ_{n_0}, w_{n_0}) is in $\mathcal{N}$, we introduce the exit time of $\mathcal{N}$ defined by:

$$T_{n_0} := \inf \{ n \geq n_0 \, : \, (\theta_n, w_n) \notin \mathcal{N} \} . \tag{21}$$

We shall observe that if $\mathbb{P}(T_{n_0} < +\infty) = 1$, then (θ_n, w_n) cannot converge almost surely to the unstable point $(t, 0)$ located in $\mathcal{N}$ since the exit time is almost surely finite.
This last observation will be our keystone argument to establish Theorem 3. In particular, it encompasses the case where $q_\infty > 0$ thanks to $i) + iv)$ of Proposition 10 and the case where $q_\infty = 0$ because in the situation of $\mathbf{H}'_{\text{Steps}}$, we already know that a.s. $w_n \longrightarrow 0$.

For this purpose, we introduce the sequence of random variables $(X_n)_{n \geq n_0+1}$ defined by:

$$\forall n \geq n_0 \qquad X_{n+1} := [\eta(\theta_{n+1}, w_{n+1}) - \eta(\theta_n, w_n)] \mathbf{1}_{n < T_{n_0}} + \gamma_{n+1} \sigma_{n+1} \mathbf{1}_{n \geq T_{n_0}}, \tag{22}$$

and the associated cumulative sum:

$$\forall n \geq n_0 + 1 \qquad S_n := \eta(\theta_{n_0}, w_{n_0}) + \sum_{k=n_0+1}^{n} X_k. \tag{23}$$

We prove the following upper bound on the second order moment of $(X_n)_{n \geq n_0+1}$.

Proposition 12 *A constant $c > 0$ exists such that:*

$$\forall n \geq n_0 \qquad \mathbb{E}[X_{n+1}^2 \,|\, \mathcal{F}_n] \leq c \gamma_{n+1}^2.$$

Proof

We decompose X_{n+1} according to the position of n with respect to T_{n_0}:

$$X_{n+1} = X_{n+1} \mathbf{1}_{n < T_{n_0}} + X_{n+1} \mathbf{1}_{n \geq T_{n_0}} = X_{n+1} \mathbf{1}_{n < T_{n_0}} + \gamma_{n+1} \sigma_{n+1} \mathbf{1}_{n \geq T_{n_0}}.$$

If $n \geq T_{n_0}$, there is nothing to prove, because (σ_n) is a bounded sequence.

We then consider the case when $n < T_{n_0}$ and define $m = \sup_{z\in\mathcal{N}} \|\nabla\eta(z)\|$. A first order Taylor expansion yields:

$$\begin{aligned}
X_{n+1}^2 \mathbf{1}_{n<T_{n_0}} &= (\eta(\theta_{n+1}, w_{n+1}) - \eta(\theta_n, w_n))^2 \mathbf{1}_{n<T_{n_0}} \\
&\leq m^2[\|\theta_{n+1} - \theta_n\|^2 + \|w_{n+1} - w_n\|^2]\mathbf{1}_{n<T_{n_0}} \\
&\leq \gamma_{n+1}^2 m^2 \left(\left\| \frac{g_{n+1}}{\sqrt{w_n + \varepsilon}} \right\|^2 + \|p_n g_{n+1}^{\odot 2} - q_n w_n\|^2 \right) \mathbf{1}_{n<T_{n_0}} \\
&\leq \gamma_{n+1}^2 \frac{m^2}{\varepsilon} \left(\|g_{n+1}\|^2 + 2(p_n \|g_{n+1}^{\odot 2}\|^2 + q_n \|w_n\|^2) \right) \mathbf{1}_{n<T_{n_0}}.
\end{aligned}$$

When $n < T_{n_0}$, the process $Z_n = (\theta_n, w_n) \in \mathcal{N}$ so that $\|w_n\|^2$ is bounded. It remains to study the terms that involve $\|g_{n+1}\|^2$ and $\|g_{n+1}^{\odot 2}\|^2$. We shall observe that:

$$\begin{aligned}
\mathbf{1}_{n<T_{n_0}} \mathbb{E}[\|g_{n+1}\|^2 \mid \mathcal{F}_n] &= \mathbf{1}_{n<T_{n_0}} \left(\|\nabla f(\theta_n)\|^2 + \mathbb{E}[\|\xi_{n+1}\|^2 | \mathcal{F}_n] \right) \\
&\leq \mathbf{1}_{n<T_{n_0}} \left(\sup_{z\in\mathcal{N}} (\|\nabla f(\theta_n)\|^2 + \sigma_{n+1}^2 \right) \\
&\leq K' \mathbf{1}_{n<T_{n_0}}.
\end{aligned}$$

because $(\sigma_n)_{n\geq 1}$ is bounded by definition and ∇f is continuous and $\mathcal{N}$ is compact.

We then study $\mathbf{1}_{n<T_{n_0}} \mathbb{E}[\|g_{n+1}^{\odot 2}\|^2 | \mathcal{F}_n]$ and observe that:

$$\|g_{n+1}^{\odot 2}\|^2 = \|(\nabla f(\theta_n) + \xi_n)^{\odot 2}\|^2 \leq 4[\|\nabla f(\theta_n)^{\odot 2}\|^2 + \|\xi_{n+1}^{\odot 2}\|^2] = 4[\|\nabla f(\theta_n)^{\odot 2}\|^2 + \sigma_{n+1}^4 \|\zeta_{n+1}^{\odot 2}\|^2].$$

Thus from the assumption $\mathbf{H}_\sigma^\infty$ it follows that:

$$\mathbf{1}_{n<T_{n_0}} \mathbb{E}[\|X_{n+1}\|^2 | \mathcal{F}_n] \leq K \gamma_{n+1}^2 \mathbf{1}_{n<T_{n_0}},$$

where K is a constant that depends on $\nabla\eta$, m, ε, $\|p\|_\infty$ and $\|q\|_\infty$. This ends the proof. ■

Below, we will use the consequence of $iv) - (a)$ of Proposition 10: if $z = (\theta, w)$ is a point in $\mathcal{N}$, then a constant Γ exists such that:

$$\forall (z, u) \in \mathcal{N} \times \mathbb{R}^{2d} : \quad \eta(z + u) - \eta(z) \geq \langle \nabla\eta(z), u \rangle - \Gamma \|u\|^2 \tag{24}$$

We write the joint evolution of the algorithm $(Z_n)_{n\geq 1}$ as:

$$Z_{n+1} = Z_n + \gamma_{n+1} \left(H(Z_n) + \Delta_{n+1} + U_{n+1} \right),$$

where $(\Delta_{n+1})_{n\geq 1}$ is a sequence of martingale increment defined by:

$$\Delta_{n+1} := \begin{pmatrix} \Delta M_{n+1} \\ \Delta N_{n+1} \end{pmatrix},$$

with

- $\Delta M_{n+1} = -\frac{\xi_{n+1}}{\sqrt{w_n + \varepsilon}}$

- $\Delta N_{n+1} = p_n(\xi_{n+1}^{\odot 2} + 2\nabla f(\theta_n)\cdot\xi_{n+1} - \mathbb{E}[\xi_{n+1}^{\odot 2}|\mathcal{F}_n])$

and

$$U_{n+1} = \begin{pmatrix} 0 \\ (p_n - p_\infty)\nabla f(\theta_n)^{\odot 2} - (q_n - q_\infty)w_n + p_n\mathbb{E}[\xi_{n+1}^{\odot 2}|\mathcal{F}_n] \end{pmatrix}.$$

We introduce below the sequence $(\nu_n)_{n\geq 1}$ that stands for the convergence rate of $(p_n)_{n\geq 1}$ towards p_∞ and $(q_n)_{n\geq 1}$ towards q_∞:

$$\nu_n := |p_n - p_\infty| \vee |q_n - q_\infty|. \tag{25}$$

Proposition 13 *For a large enough non negative constant c_δ the sequence $(\delta_n)_{n\geq 1}$ defined by $\delta_n \overset{\text{def}}{=} c_\delta(\nu_n + p_n\sigma_{n+1}^2 + \gamma_{n+1})$ is such that:*

$$\mathbf{1}_{S_n\geq\delta_n}\mathbb{E}[X_{n+1}|\ \mathcal{F}_n] \geq 0.$$

Proof

We start once more from the decomposition of $(X_n)_{n\geq n_0+1}$ before or after T_{n_0} and observe that:

$$\mathbb{E}[\mathbb{1}_{n>T_{n_0}}X_{n+1}\,|\mathcal{F}_n] = \gamma_{n+1}\sigma_{n+1}\mathbb{1}_{n>T_{n_0}} \geq 0. \tag{26}$$

The definition of the stopping time T_{n_0} ensures that when $n < T_{n_0}$, Z_n belongs to the neighborhood $\mathcal{N}$ so we can apply Equation (24) and obtain the following lower bound:

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{n<T_{n_0}}X_{n+1}|\ \mathcal{F}_n] &= \mathbf{1}_{n<T_{n_0}}\mathbb{E}[\eta(Z_{n+1}) - \eta(Z_n)|\ \mathcal{F}_n] \\ &\geq \mathbb{1}_{n<T_{n_0}}\Bigg(\mathbb{E}\left[\langle\nabla\eta(Z_n), \gamma_{n+1}(H(Z_n) + \begin{pmatrix}\Delta M_{n+1}\\ \Delta N_{n+1}\end{pmatrix} + \begin{pmatrix}0\\ U_{n+1}\end{pmatrix}\rangle\ |\mathcal{F}_n\right] \\ &\quad - \Gamma\gamma_{n+1}^2\mathbb{E}\left[\left\|H(Z_n) + \begin{pmatrix}\Delta M_{n+1}\\ \Delta N_{n+1}\end{pmatrix} + \begin{pmatrix}0\\ U_{n+1}\end{pmatrix}\right\|^2\ |\,\mathcal{F}_n\right]\Bigg).\end{aligned}$$

We treat the two terms separately:

$$\begin{aligned}&\mathbb{1}_{n<T_{n_0}}\mathbb{E}\left[\langle\nabla\eta(Z_n), \gamma_{n+1}(H(Z_n) + \begin{pmatrix}\Delta M_{n+1}\\ \Delta N_{n+1}\end{pmatrix} + \begin{pmatrix}0\\ U_{n+1}\end{pmatrix}\rangle\ |\mathcal{F}_n\right] \\ &\quad = \mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\Bigg(\langle\nabla\eta(Z_n), H(Z_n)\rangle + \mathbb{E}\left[\langle\nabla\eta(Z_n), \begin{pmatrix}\Delta M_{n+1}\\ \Delta N_{n+1}\end{pmatrix}\rangle|\mathcal{F}_n\right] + \langle\nabla\eta(Z_n), \begin{pmatrix}0\\ U_{n+1}\end{pmatrix}\rangle\Bigg) \\ &\quad \geq \mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\left(k\eta(Z_n) + \langle\nabla\eta(Z_n), \begin{pmatrix}0\\ U_{n+1}\end{pmatrix}\rangle\right),\end{aligned}$$

where the last inequality comes from (iv)c of Proposition 10 (Equation (20)). Using that $\|\nabla\eta\|$ is upper bounded on $\mathcal{N}$ by m, the definition of T_{n_0} leads to:

$$\mathbb{1}_{n<T_{n_0}}\|\nabla\eta(Z_n)\| \leq m.$$

This inequality associated with the Cauchy-Schwarz inequality implies that:

$$\begin{aligned}\mathbb{1}_{n<T_{n_0}}\left|\langle\nabla\eta(Z_n),\begin{pmatrix}0\\U_{n+1}\end{pmatrix}\rangle\right| &\leq \mathbb{1}_{n<T_{n_0}}\|\nabla\eta(Z_n)\|\left\|\begin{pmatrix}0\\U_{n+1}\end{pmatrix}\right\| \\ &\leq m\mathbb{1}_{n<T_{n_0}}\|U_{n+1}\| \\ &\leq m\mathbb{1}_{n<T_{n_0}}\|\nu_n\|\nabla f(\theta_n)^{\odot 2}\| + \nu_n\|w_n\| + p_n\mathbb{E}[\xi_{n+1}^{\odot 2}|\mathcal{F}_n]\| \\ &\leq m\mathbb{1}_{n<T_{n_0}}\left(\nu_n\sup_{z\in\mathcal{N}}(\|\nabla f(\theta_n)^{\odot 2}\| + \|w_n\|) + p_n\mathbb{E}[\|\xi_{n+1}^{\odot 2}\|\ |\mathcal{F}_n]\right).\end{aligned}$$

Observing that $\|\xi_{n+1}^{\odot 2}\| = \sqrt{\sum_{i=1}^d \xi_{n+1,i}^4} \leq \sum_{i=1}^d \xi_{n+1,i}^2 = \|\xi_{n+1}\|^2$, we deduce that:

$$\mathbb{E}[\|\xi_{n+1}^{\odot 2}\|\ |\mathcal{F}_n] \leq \sigma_{n+1}^2.$$

We then define $k' = \sup_{(\theta,w)\in\mathcal{N}}\|\nabla f(\theta)^{\odot 2}\| + \|w\| < +\infty$ (because $\mathcal{N}$ is compact and ∇f is continuous). We deduce that:

$$\mathbb{1}_{n<T_{n_0}}\left|\langle\nabla\eta(Z_n),\begin{pmatrix}0\\U_{n+1}\end{pmatrix}\rangle\right| \leq k'm\mathbb{1}_{n<T_{n_0}}\left(2\nu_n + p_n\sigma_{n+1}^2\right).$$

Concerning the last term, it is sufficient to show that the expected value is bounded. We start by using the fact that $(a+b+c)^2 \leq 4(a^2+b^2+c^2)$ to split the squared norm and to obtain that:

$$\begin{aligned}&\mathbb{E}[\|H(Z_n) + \begin{pmatrix}\Delta M_{n+1}\\\Delta N_{n+1}\end{pmatrix} + \begin{pmatrix}0\\U_{n+1}\end{pmatrix}\|^2\ |\mathcal{F}_n] \\ &\qquad\leq 4\left(\|H(Z_n)\|^2 + \mathbb{E}\left[\left\|\begin{pmatrix}\Delta M_{n+1}\\\Delta N_{n+1}\end{pmatrix}\right\|^2 |\mathcal{F}_n\right] + \|U_{n+1}\|^2\right).\end{aligned}$$

Since H is continuous, $(\|H(Z_n)\|^2)_{n_0\leq n<T_{n_0}}$ is a bounded sequence. Moreover, we have seen that when $n < T_{n_0}$:

$$\|U_{n+1}\| \leq k'(2\nu_n + \sigma_{n+1}^2 p_n),$$

thus we are only left to study $\mathbb{E}\left[\left\|\begin{pmatrix}\Delta M_{n+1}\\\Delta N_{n+1}\end{pmatrix}\right\|^2 |\,\mathcal{F}_n\right]$.

$$\begin{aligned}\mathbb{E}\left[\left\|\begin{pmatrix}\Delta M_{n+1}\\\Delta N_{n+1}\end{pmatrix}\right\|^2 |\mathcal{F}_n\right] &= \mathbb{E}\left[\|\Delta M_{n+1}\|^2 + \|\Delta N_{n+1}\|^2|\mathcal{F}_n\right] \\ &\leq \frac{\sigma_{n+1}^2}{\varepsilon} + p_n^2\mathbb{E}[\|\xi_{n+1}^{\odot 2} + 2\nabla f(\theta_n)\cdot\xi_{n+1} - \mathbb{E}[\xi_{n+1}^{\odot 2}|\mathcal{F}_n]\|^2|\ \mathcal{F}_n] \\ &\leq \frac{\sigma_{n+1}^2}{\varepsilon} + 4p_n^2\left(\mathbb{E}[\|\xi_{n+1}^{\odot 2}\|^2 + 2\|\nabla f(\theta_n)\cdot\xi_{n+1}\|^2 + \|\xi_{n+1}^{\odot 2}\|^2|\mathcal{F}_n]\right) \\ &\leq \frac{\sigma_{n+1}^2}{\varepsilon} + 4p_n^2\left(2\mathbb{E}[\|\xi_{n+1}\|^4|\ \mathcal{F}_n] + 2\mathbb{E}[\|\nabla f(\theta_n)\cdot\xi_{n+1}\|^2|\ \mathcal{F}_n]\right).\end{aligned}$$

When $n < T_{n_0}$, we observe that:

$$\mathbb{E}[\|\nabla f(\theta_n)\cdot\xi_{n+1}\|^2|\ \mathcal{F}_n] \leq \sup_{z\in\mathcal{N}} \|\nabla f(\theta)^2\|_\infty \mathbb{E}[\|\xi_{n+1}\|^2|\ \mathcal{F}_n],$$

and

$$\mathbb{E}[\|\xi_{n+1}\|^4|\ \mathcal{F}_n] = \sigma_{n+1}^4 \mathbb{E}[\|\zeta_{n+1}\|^4|\ \mathcal{F}_n].$$

Again, using the compactness of $\mathcal{N}$, the continuity of ∇f and the assumption $\mathbf{H}_\sigma^\infty$ on the noise $(\zeta_n)_{n\geq 1}$, we deduct that $K_2 > 0$ exists such that:

$$\mathbb{E}\left[\|\Delta M_{n+1}\|^2 + \|\Delta N_{n+1}\|^2|\mathcal{F}_n\right] \leq K_2(\sigma_{n+1}^2 + \sigma_{n+1}^4).$$

We now gather all the terms and use the fact that the sequences $(\sigma_n)_{n\geq 0}$, $(p_n)_{n\geq 0}$ and $(\nu_n)_{n\geq 0}$ are all bounded, to conclude that a constant K_3 exists such that:

$$\mathbb{E}[\mathbf{1}_{n<T_{n_0}} X_{n+1}|\ \mathcal{F}_n] \geq \mathbf{1}_{n<T_{n_0}}\gamma_{n+1}\left(k\eta(Z_n) - k'm(2\nu_n + p_n\sigma_{n+1}^2) - K_3\gamma_{n+1}\right). \tag{27}$$

Finally, as long as $n < T_{n_0}$, $S_n = \eta(Z_n)$ and thus setting $c_\delta > \max(2k'm, K_3)/k$ and $(\delta_n)_{n\geq 1}$ as announced in the statement ends the proof. ■

We now study the evolution of S_n^2.

Proposition 14 *If $\delta_n = c_\delta(\nu_n + p_n\sigma_{n+1}^2 + \gamma_{n+1})$ is such that $\gamma_{n+1}\delta_n^2 = o(\gamma_{n+1}^2\sigma_{n+1}^2)$ then:*

$$\mathbb{E}[S_{n+1}^2 - S_n^2\,|\,\mathcal{F}_n] \gtrsim \gamma_{n+1}^2\sigma_{n+1}^2.$$

Proof Our starting point is:

$$\begin{aligned} S_{n+1}^2 - S_n^2 &= (S_n + X_{n+1})^2 - S_n^2 \\ &= X_{n+1}^2 + 2S_nX_{n+1} \\ &= X_{n+1}^2 + 2S_n(\mathbb{1}_{S_n\geq\delta_n}X_{n+1} + \mathbb{1}_{S_n<\delta_n}X_{n+1}). \end{aligned}$$

We then use Proposition 13 and get:

$$\mathbb{E}[S_{n+1}^2 - S_n^2\,|\,\mathcal{F}_n] \geq \mathbb{E}[X_{n+1}^2\,|\,\mathcal{F}_n] + 2S_n\mathbb{1}_{S_n<\delta_n}\mathbb{E}[X_{n+1}\,|\mathcal{F}_n]. \tag{28}$$

To derive a lower bound of $2S_n\mathbb{1}_{S_n<\delta_n}\mathbb{E}[X_{n+1}\,|\mathcal{F}_n]$, we observe that $(S_n)_{n\geq 0}$ and η are positive, so that if we use Equations (27) and (26), we obtain:

$$\begin{aligned} 2S_n\mathbb{1}_{S_n<\delta_n}\mathbb{E}[X_{n+1}\,|\mathcal{F}_n] &\geq -2S_n\mathbb{1}_{S_n<\delta_n}\mathbb{E}[\mathbb{1}_{n<T_{n_0}}|X_{n+1}|\,|\mathcal{F}_n] \\ &\geq -2\delta_n\gamma_{n+1}[k'm(2\nu_n + p_n\sigma_{n+1}^2) + K_3\gamma_{n+1}] \\ &\geq -2\delta_n^2\gamma_{n+1}. \end{aligned}$$

We are led to analyze $\mathbb{E}[X_{n+1}^2]$. According to the definition of $(X_n)_{n\geq n_0}$, to the definition of the hitting time T_{n_0} and to the construction of η, we observe that from Equation (24):

$$\begin{aligned}
\mathbb{1}_{n<T_{n_0}} X_{n+1} &= \mathbb{1}_{n<T_{n_0}}[\eta(Z_{n+1}) - \eta(Z_n)] \\
&= \mathbb{1}_{n<T_{n_0}}[\eta[Z_n + (Z_{n+1} - Z_n)] - \eta(Z_n)] \\
&\geq \mathbb{1}_{n<T_{n_0}}\left[\langle \nabla\eta(Z_n), Z_{n+1} - Z_n\rangle - \Gamma\|Z_{n+1} - Z_n\|^2\right] \\
&\geq \mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\langle \nabla\eta(Z_n), H(Z_n)\rangle + \mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\langle \nabla\eta(Z_n), \Delta_{n+1} + U_n\rangle \\
&- \Gamma\mathbb{1}_{n<T_{n_0}}\gamma_{n+1}^2\|H(Z_n) + \Delta_{n+1} + U_n\|^2 \\
&\geq \kappa\gamma_{n+1}\eta(Z_n)\mathbb{1}_{n<T_{n_0}} + \mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\langle \nabla\eta(Z_n), \Delta_{n+1}\rangle \\
&- \gamma_{n+1}\mathbb{1}_{n<T_{n_0}}\|U_n\|\|\nabla\eta(Z_n)\| - \Gamma\mathbb{1}_{n<T_{n_0}}\gamma_{n+1}^2\|H(Z_n) + \Delta_{n+1} + U_n\|^2,
\end{aligned}$$

where in the last line we used the reverting effect translated in Equation (20) and the Cauchy-Schwarz inequality. Since η is positive, we then obtain that:

$$\begin{aligned}
\mathbb{1}_{n<T_{n_0}} X_{n+1} &\geq \mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\langle \nabla\eta(Z_n), \Delta_{n+1}\rangle \\
&- \mathbb{1}_{n<T_{n_0}}\left[\gamma_{n+1}\|U_n\|\|\nabla\eta(Z_n)\| + \Gamma\gamma_{n+1}^2\|H(Z_n) + \Delta_{n+1} + U_n\|^2\right].
\end{aligned} \tag{29}$$

We denote the positive part of any real value a by $\lfloor a\rfloor_+$ and we use that $a \geq b \Longrightarrow \lfloor a\rfloor_+ \geq \lfloor b\rfloor_+$ and the inequality

$$\lfloor a - |b|\rfloor_+ \geq \lfloor a\rfloor_+ - |b|.$$

Considering Equation (29), we then observe that:

$$\begin{aligned}
\lfloor \mathbb{1}_{n<T_{n_0}} X_{n+1}\rfloor_+ \geq &\mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\left[\lfloor\langle \nabla\eta(Z_n), \Delta_{n+1}\rangle\rfloor_+\right. \\
&\left. -\|U_n\|\|\nabla\eta(Z_n)\| - \Gamma\gamma_{n+1}\|H(Z_n) + \Delta_{n+1} + U_n\|^2\right].
\end{aligned}$$

Once more the regularity of $\nabla\eta$, the compactness of $\mathcal{N}$ and the definition of U_n, guarantee that a constant $\kappa > 0$ exists such that:

$$\mathbb{1}_{n<T_{n_0}}\|U_n\|\|\nabla\eta(Z_n)\| \leq \kappa(\nu_n + p_n\sigma_{n+1}^2).$$

Computing the conditional expectation and using the arguments of (27), we have

$$\mathbb{1}_{n<T_{n_0}}\mathbb{E}[\|H(Z_n) + \Delta_{n+1} + U_n\|^2 \,|\mathcal{F}_n] < K_3,$$

so that:

$$\begin{aligned}
\mathbb{E}[\lfloor \mathbb{1}_{n<T_{n_0}} X_{n+1}\rfloor_+ \,|\mathcal{F}_n] \geq &\mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\mathbb{E}[\lfloor\langle \nabla\eta(Z_n), \Delta_{n+1}\rangle\rfloor_+ \,|\mathcal{F}_n] \\
&- \mathbb{1}_{n<T_{n_0}}\left[\kappa(\nu_n + p_n\sigma_{n+1}^2)\gamma_{n+1} + \Gamma K_3\gamma_{n+1}^2\right].
\end{aligned}$$

Using that when $n < T_{n_0}$, $Z_n \in \mathcal{N}$, we can apply $iv) - (b)$ of Proposition 10 so that:

$$\begin{aligned}
\mathbb{E}[\lfloor \mathbb{1}_{n<T_{n_0}} X_{n+1}\rfloor_+ \,|\mathcal{F}_n] \geq &\mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\left(c_1\mathbb{E}[\|\pi_+(\Delta M_{n+1})\|\,|\mathcal{F}_n]\right. \\
&\left. - \left[\kappa(\nu_n + p_n\sigma_{n+1}^2) + \Gamma K_3\gamma_{n+1}\right]\right),
\end{aligned}$$

where π_+ is the orthogonal projection on E_+, the eigenspace associated to the negative eigenvalues of $D^2 f(t)$. For n large enough, the almost sure convergence of $(w_n)_{n\geq 0}$ to 0 and our elliptic assumption $(\mathbf{H}_\sigma^\infty - 2)$ on the sequence $(\xi_{n+1})_{n\geq 0}$ yield:

$$\mathbb{E}[\|\pi_+(\Delta M_{n+1})\| \,|\mathcal{F}_n] \geq \frac{\sigma_{n+1}}{2},$$

which entails:

$$\mathbb{E}[\lfloor \mathbb{1}_{n<T_{n_0}} X_{n+1}\rfloor_+ \,|\mathcal{F}_n] \geq \mathbb{1}_{n<T_{n_0}}\gamma_{n+1}\left[c_1\frac{\sigma_{n+1}}{2} - C(\nu_n + p_n\sigma_{n+1}^2 + \gamma_{n+1}^2)\right].$$

Decomposing now $X_{n+1} = \lfloor X_{n+1}\rfloor_+ - \lfloor -X_{n+1}\rfloor_+$, we deduce that:

$$\begin{aligned}
\mathbb{E}[\mathbb{1}_{n<T_{n_0}} X_{n+1}^2 \,|\mathcal{F}_n] &= \mathbb{E}[\mathbb{1}_{n<T_{n_0}} \lfloor X_{n+1}\rfloor_+^2 \,|\mathcal{F}_n] + \mathbb{E}[\mathbb{1}_{n<T_{n_0}} \lfloor -X_{n+1}\rfloor_+^2 \,|\mathcal{F}_n] \\
&\geq \mathbb{E}[\mathbb{1}_{n<T_{n_0}} \lfloor X_{n+1}\rfloor_+^2 \,|\mathcal{F}_n] \\
&\geq \mathbb{E}[\mathbb{1}_{n<T_{n_0}} \lfloor X_{n+1}\rfloor_+ \,|\mathcal{F}_n]^2 \\
&\geq \mathbb{1}_{n<T_{n_0}}\gamma_{n+1}^2\left[c_1\frac{\sigma_{n+1}}{2} - C(\nu_n + p_n\sigma_{n+1}^2 + \gamma_{n+1}^2)\right]^2. \\
&\gtrsim \mathbb{1}_{n<T_{n_0}}\gamma_{n+1}^2\sigma_{n+1}^2.
\end{aligned}$$

The last line is justified by the assumption $\gamma_{n+1}\delta_n^2 = o(\gamma_{n+1}^2\sigma_{n+1}^2)$ which ensures that $\delta_n = o(\sigma_n)$. We shall also observe that:

$$\mathbb{E}[\mathbb{1}_{n\geq T_{n_0}} X_{n+1}^2 \,|\mathcal{F}_n] = (\gamma_{n+1}\sigma_{n+1})^2.$$

We then deduce that:

$$\mathbb{E}[X_{n+1}^2 \,|\mathcal{F}_n] \gtrsim (\gamma_{n+1}\sigma_{n+1})^2.$$

Since $\gamma_{n+1}\delta_n^2 = o((\gamma_{n+1}\sigma_{n+1})^2)$, we can now conclude by inserting the previous bounds in Inequality (28).

■

4.3 End of the Proof

We mimic the strategy of Lemma 9.6 of Benaïm (1999) and of Theorem 3.2 of Gadat et al. (2018) with the help of the two sequences defined in (23) and (22).

$$S_n = \eta(Z_{n_0}) + \sum_{k=n_0+1}^{n} X_k \quad \text{with} \quad X_{n+1} = (\eta(Z_{n+1}) - \eta(Z_n))\mathbb{1}_{n<T_{n_0}} + \gamma_{n+1}\sigma_{n+1}\mathbb{1}_{n\geq T_{n_0}}.$$

We summarize the preliminary results proven in the previous subsection as they will be used in what follows:

(I_1) Proposition 12 states that:

$$\mathbb{E}[X_{n+1}^2\,\mathcal{F}_n] \leq c(\gamma_{n+1}\sigma_{n+1})^2.$$

(I_2) Proposition 13 yields that if $\delta_n = c_\delta(\nu_n + p_n\sigma_{n+1}^2 + \gamma_{n+1})$, then

$$\mathbb{1}_{S_n \geq \delta_n} \mathbb{E}[X_{n+1} \,|\mathcal{F}_n] \geq 0.$$

(I_3) Proposition 14 yields: if $\gamma_{n+1}\delta_n^2 = o((\gamma_{n+1}\sigma_{n+1})^2)$ (which also means that $\delta_n = o(\sigma_n)$), then

$$\mathbb{E}[S_{n+1}^2 - S_n^2 \,|\mathcal{F}_n] \gtrsim (\gamma_{n+1}\sigma_{n+1})^2.$$

For any integer q, we consider an integer $n \geq q$ and introduce the two sequences $(u_n)_{n\geq q}$ and $(\bar{U}_n)_{n\geq q}$ defined by:

$$u_n = \sum_{i\geq n} \gamma_{i+1}^2\sigma_{i+1}^2 \qquad \text{and} \qquad \bar{U}_n = \sum_{i=0}^{n} \gamma_{i+1}^2\sigma_{i+1}^2.$$

For any positive real value $b > 0$ and any integer $q > 0$, we consider the sequence of stopping times $(T_b^q)_{q\geq 0}$ defined by:

$$T_b^q := \inf\left\{ i \geq q \,:\, S_i \geq \sqrt{bu_i} \right\}.$$

A stopping times T_b^q stands for the first time the sequence $(S_n)_{n\geq 1}$ becomes larger than the threshold $(\sqrt{bu_n})_{n\geq 1}$, which converges towards zero at a controlled rate. We prove the following result.

Proposition 15 *If $\gamma_n^2 = o(u_n)$, then a small enough $b > 0$ and a large enough q exist such that:*

$$\mathbb{P}(T_b^q < +\infty \,|\mathcal{F}_q) \geq \frac{1}{2}.$$

Proof Our starting point is the lower bound given by Proposition 14 and we observe that a small enough $a > 0$ exists such that:

$$\mathcal{M}_n := S_n^2 - a\bar{U}_n,$$

is a submartingale since:

$$\begin{aligned}
\mathbb{E}[\mathcal{M}_{n+1} \,|\mathcal{F}_n] &= \mathbb{E}[S_{n+1}^2 - a\bar{U}_{n+1} \,|\mathcal{F}_n] \\
&\geq S_n^2 + a(\gamma_{n+1}\sigma_{n+1})^2 - a(\bar{U}_n + (\gamma_{n+1}\sigma_{n+1})^2) \\
&= \mathcal{M}_n.
\end{aligned}$$

We consider an integer $n \geq q+1$, apply the Optional Stopping Theorem and verify that:

$$\begin{aligned}
\mathbb{E}[\mathcal{M}_{n\wedge T_b^q} - \mathcal{M}_q \,|\mathcal{F}_q] \geq 0 &\Longleftrightarrow \mathbb{E}[S_{n\wedge T_b^q}^2 - S_q^2 \,|\mathcal{F}_q] \geq a\mathbb{E}\left[\sum_{i=q}^{n\wedge T_b^q} \gamma_{i+1}^2\sigma_{i+1}^2 \,|\mathcal{F}_q \right] \\
&\Longleftrightarrow \mathbb{E}[S_{n\wedge T_b^q}^2 - S_q^2 \,|\mathcal{F}_q] \geq a\mathbb{P}\left(T_b^q > n \,|\mathcal{F}_q\right) \sum_{i=q}^{n} \gamma_{i+1}^2\sigma_{i+1}^2. \qquad (30)
\end{aligned}$$

In the meantime, we observe that:

$$\begin{aligned}\{S_{n\wedge T_b^q}\}^2 - \{S_q\}^2 &= \{S_{n\wedge T_b^q}\}^2 - \{S_{(n\wedge T_b^q)-1}\}^2 + \{S_{(n\wedge T_b^q)-1}\}^2 - \{S_q\}^2 \\ &= \{S_{(n\wedge T_b^q)-1} + X_{n\wedge T_b^q}\}^2 - \{S_{(n\wedge T_b^q)-1}\}^2 + \{S_{(n\wedge T_b^q)-1}\}^2 - \{S_q\}^2 \\ &\leq \{S_{(n\wedge T_b^q)-1}\}^2 + 2S_{(n\wedge T_b^q)-1}X_{n\wedge T_b^q} + \{X_{n\wedge T_b^q}\}^2 \\ &\leq 2\left(\{S_{(n\wedge T_b^q)-1}\}^2 + \{X_{n\wedge T_b^q}\}^2\right) \\ &\leq 2bu_{(n\wedge T_b^q)-1} + 2\{X_{n\wedge T_b^q}\}^2,\end{aligned}$$

where the last inequality is a consequence of the definition of the stopping time T_b^q. Since $(u_n)_{n\geq 0}$ is a decreasing sequence, we then have that:

$$\{S_{n\wedge T_b^q}\}^2 - \{S_q\}^2 \leq 2bu_{q-1} + 2\{X_{n\wedge T_b^q}\}^2, \tag{31}$$

and we are led to upper bound the term $\{X_{n\wedge T_b^q}\}^2$.

The definition of $(X_n)_{n\geq 1}$ yields:

$$X_{n\wedge T_b^q}^2 = \left(\eta(Z_{n\wedge T_b^q}) - \eta(Z_{(n\wedge T_b^q)-1})\right)^2 \mathbb{1}_{(n\wedge T_b^q)-1<T_{n_0}} + \gamma_{n\wedge T_b^q}^2 \sigma_{n\wedge T_b^q}^2 \mathbb{1}_{(n\wedge T_b^q)-1\geq T_{n_0}}.$$

Using a similar argument as the one used in the proof of Proposition 12, a Taylor expansion associated with the smoothness of η and f leads to:

$$\begin{aligned}\mathbb{1}_{(n\wedge T_b^q)-1<T_{n_0}}\left(\eta(Z_{n\wedge T_b^q}) - \eta(Z_{n\wedge T_b^q-1})\right)^2 &\lesssim \mathbb{1}_{n\wedge T_b^q-1<T_{n_0}}\gamma_{n\wedge T_b^q}^2\left(1 + \|\xi_{n\wedge T_b^q}\|^2 + \|\xi_{n\wedge T_b^q}^2\|^2\right) \\ &\lesssim \mathbb{1}_{n\wedge T_b^q-1<T_{n_0}}\gamma_{n\wedge T_b^q}^2\left(1 + \sigma_{n\wedge T_b^q}^2\|\zeta_{n\wedge T_b^q}\|^2 + \sigma_{n\wedge T_b^q}^4\|\zeta_{n\wedge T_b^q}\|^4\right) \\ &\lesssim \gamma_{q+1}^2 + \sum_{i\geq q}^n \gamma_{i+1}^2\sigma_{i+1}^2\|\zeta_{i+1}\|^2 + \sum_{i\geq q}^n \gamma_{i+1}^2\sigma_{i+1}^4\|\zeta_{i+1}\|^4,\end{aligned}$$

where in the last line we used that $(\gamma_n)_{n\geq q+1}$ is a decreasing sequence and a rough upper bound of $\sigma_{n\wedge T_b^q}\|\zeta_{n\wedge T_b^q}\|$. We then compute the expectation with respect to $\mathcal{F}_q$ and observe that $\mathbb{E}[\|\zeta_{i+1}\|^2 \,|\mathcal{F}_q] \leq 1$ and $\mathbb{E}[\|\zeta_{i+1}\|^4 \,|\mathcal{F}_q] \leq C$ and thus:

$$\mathbb{E}[X_{n\wedge T_b^q}^2 \,|\mathcal{F}_q] \lesssim \gamma_q^2\sigma_q^2 + \gamma_{q+1}^2 + u_q.$$

This last bound, together with Inequality (31) implies that a constant $C > 0$ exists such that:

$$\mathbb{E}[\{S_{n\wedge T_b^q}\}^2 - \{S_q\}^2|\mathcal{F}_q] \leq C\left(bu_q + \gamma_{q+1}^2\right).$$

Finally we obtain from (30) that:

$$a\mathbb{P}\left(T_b^q > n\,|\mathcal{F}_q\right)\sum_{i=q}^n \gamma_{i+1}^2\sigma_{i+1}^2 \leq C\left(bu_q + \gamma_{q+1}^2\right).$$

We therefore deduce an upper bound on the probability that the stopping time is larger than n:

$$\mathbb{P}\left(T_b^q > n \,|\mathcal{F}_q\right) \leq \frac{C\left(bu_q + \gamma_{q+1}^2\right)}{a\sum_{i=q}^n \gamma_{i+1}^2\sigma_{i+1}^2} = \frac{Cb}{a}\frac{u_q}{u_q - u_n} + C\frac{\gamma_q^2}{a(u_q - u_n)}.$$

We then take the limit $n \longrightarrow +\infty$ in the previous inequality and obtain that:

$$\mathbb{P}\left(T_b^q = +\infty \,|\mathcal{F}_q\right) \leq \frac{Cb}{a} + C\frac{\gamma_q^2}{au_q},$$

because $\lim_{n\longrightarrow+\infty} u_n = 0$. We then observe that $u_q = \sum_{i\geq q}\gamma_{i+1}^2\sigma_{i+1}^2$ and then the second term on the right hand side goes to zero as soon as:

$$\gamma_q^2 = o\left(\sum_{i\geq q}\gamma_{i+1}^2\sigma_{i+1}^2\right) = o(u_q).$$

Using our assumption on the two sequences, we can conclude the proof of the proposition by setting b small enough ($b < a/3C$ for example). ■

The next result states that S_n may remain larger than $\frac{1}{2}\sqrt{bu_q}$ with a positive probability when $S_q \geq \sqrt{bu_q}$. For this purpose, we introduce

$$\mathcal{S} := \inf\left\{n > q \,:\, S_n < \frac{1}{2}\sqrt{bu_q}\right\},$$

that stands for the first time $(S_n)_{n\geq q}$ comes back below the threshold $\sqrt{bu_q}$.

Proposition 16 *Assume that $\alpha_{n+1} = \sigma_{n+1}\gamma_{n+1}$ and $\delta_n = o(\sqrt{u_n})$ and $\gamma_n = O(u_n)$. Then there exits q large enough and a constant $c > 0$ such that:*

$$\mathbb{1}_{S_q\geq\sqrt{bu_q}}\mathbb{P}\left(\mathcal{S} = +\infty \,|\mathcal{F}_q\right) \geq \mathbb{1}_{S_q\geq\sqrt{bu_q}}\frac{b}{b+c}.$$

Proof From our assumption $\delta_n = o(\sqrt{u_n})$, we observe that for q large enough if S_q is greater than $\frac{1}{2}\sqrt{bu_q}$ then $S_q \geq \delta_q$, so Proposition 13 implies that $(S_{n\wedge\mathcal{S}})_{n\geq q}$ is a submartingale. For such a choice of n and q, we can use the Doob decomposition and write that:

$$S_{n\wedge\mathcal{S}} = W_n + I_n,$$

where I_n is an increasing $\mathcal{F}_n$ predictable process with $I_q = 0$ and W_n is a martingale. We then observe that

$$\mathbb{P}\left(\mathcal{S} = +\infty \,|\mathcal{F}_q\right) = \mathbb{P}\left(\forall n \geq q :\, S_n \geq \frac{1}{2}\sqrt{bu_q}|\mathcal{F}_q\right) \geq \mathbb{P}\left(\forall n \geq q :\, W_n \geq \frac{1}{2}\sqrt{bu_q}|\mathcal{F}_q\right).$$

Furthermore, if $S_q \geq \sqrt{bu_q}$, then $W_q \geq \sqrt{bu_q}$, which entails:

$$\mathbb{1}_{S_q\geq\sqrt{bu_q}}\mathbb{P}\left(\mathcal{S} = +\infty \,|\mathcal{F}_q\right) \geq \mathbb{1}_{S_q\geq\sqrt{bu_q}}\mathbb{P}\left(\forall n \geq q :\, W_n - W_q \geq -\frac{1}{2}\sqrt{bu_q}|\mathcal{F}_q\right). \tag{32}$$

Using the fact that $(W_n)_{n\geq q}$ is a martingale and the definition of S_n, we can use the quadratic decomposition of $(W_n - W_q)^2$ and observe that:

$$\mathbb{E}\left[(W_n - W_q)^2 \,|\mathcal{F}_q\right] = \sum_{i=q}^{n-1} Var[X_{i+1}|\mathcal{F}_q] \leq \sum_{i=q}^{n-1} \mathbb{E}[X_{i+1}^2|\mathcal{F}_q].$$

The upper bound obtained in Proposition 12 is not sharp enough to be directly applied here, so in order to deal with the term on the right hand side we return to the upper bound obtained in the proof of Proposition 12 which gives for all i:

$$X_{i+1}^2 \mathbb{1}_{i<T_{n_0}} \lesssim \gamma_{i+1}^2 \left(\|g_{i+1}\|^2 + 2(p_i \|g_{i+1}^{\odot 2}\|^2 + q_i \|w_i\|)\right) \mathbb{1}_{i<T_{n_0}}. \tag{33}$$

Now, for $i \geq q$ we have:

$$\begin{aligned}
\mathbb{E}[\|g_{i+1}\|^2|\mathcal{F}_q] &= \mathbb{E}[\mathbb{E}[\|g_{i+1}\|^2|\mathcal{F}_i]|\mathcal{F}_q] \\
&\leq 2(\|\nabla f(\theta_i)\|^2 + \mathbb{E}[\mathbb{E}[\|\xi_{i+1}\|^2|\mathcal{F}_i]|\mathcal{F}_q]) \\
&\leq 2(\|\nabla f(\theta_i)\|^2 + \sigma_{i+1}^2),
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\|g_{i+1}^{\odot 2}\|^2|\mathcal{F}_q] &\leq 8(\|\nabla f(\theta_i)^{\odot 2}\|^2 + \sigma_{i+1}^4 \mathbb{E}[\mathbb{E}[\|\zeta_{i+1}^2\|^2|\mathcal{F}_i]|\mathcal{F}_q]) \\
&\lesssim \|\nabla f(\theta_i)^{\odot 2}\|^2 + \sigma_{i+1}^4.
\end{aligned}$$

The sequences $(\sigma_i)_{i\geq 0}$ and $(p_i)_{i\geq 0}$ are bounded and Theorem 1 shows that the sequence $(\|\nabla f(\theta_i)\|)_{i\geq 0}$ converges almost surely to 0, so there exists $q > 0$ such that $\forall i \geq q$:

$$p_i \sigma_{i+1}^4 \leq \|p\|_\infty \|\sigma\|_\infty^2 \sigma_{i+1}^2 \quad \text{and} \quad \|\nabla f(\theta_i)^{\odot 2}\|^2 \leq \|\nabla f(\theta_i)\|^2.$$

Inserting this into our previous bound (33) we get that (for q large enough and $i \geq q$)

$$X_{i+1}^2 \mathbb{1}_{i<T_{n_0}} \lesssim \gamma_{i+1}^2 \left(\|\nabla f(\theta_i)\|^2 + q_i \|w_i\| + \sigma_{i+1}^2\right) \mathbb{1}_{i<T_{n_0}}.$$

Using the definition of X_n, we have that:

$$X_{i+1}^2 \lesssim \gamma_{i+1}^2 \left(\|\nabla f(\theta_i)\|^2 + q_i \|w_i\| + \sigma_{i+1}^2\right),$$

which implies that:

$$\begin{aligned}
\mathbb{E}\left[(W_n - W_q)^2 \,|\mathcal{F}_q\right] &\lesssim \sum_{i=q}^{n-1} \gamma_{i+1}^2 \sigma_{i+1}^2 + \sum_{i=q}^{n-1} \gamma_{i+1}^2 \left(\|\nabla f(\theta_i)\|^2 + q_i \|\sqrt{w_i}\|^2\right) + \sum_{i=q}^{n-1} \gamma_{i+1}^2 p_i \sigma_i^4 \\
&\lesssim u_q + \gamma_{q+1} \sum_{i=q}^{n-1} \gamma_{i+1} \left(\|\nabla f(\theta_i)\|^2 + q_i \|\sqrt{w_i}\|^2\right).
\end{aligned}$$

For the second term of the right-hand side we use the almost sure convergence of the series stated in (13) of Proposition 6. It leads to:

$$\mathbb{E}\left[(W_n - W_q)^2 \,|\mathcal{F}_q\right] \lesssim \sum_{i=q}^{n-1} \gamma_{i+1}^2 \sigma_{i+1}^2 + \gamma_{q+1}.$$

Thus there exists a constant $a > 0$ such that for all $n \geq q$:

$$\mathbb{E}\left[(W_n - W_q)^2 \,|\mathcal{F}_q\right] \leq a(u_q + \gamma_q).$$

Using a similar argument as the one of Benaim and Hirsch (1995) or Gadat et al. (2018), for all $h, t \in \mathbb{R}_+$ we deduce that:

$$\begin{aligned}\mathbb{P}\left(\inf_{q\leq i\leq n}(W_i - W_q) \leq -h \,|\mathcal{F}_q\right) &\leq \mathbb{P}\left(\sup_{q\leq i\leq n}|W_i - W_q - t| \geq h + t \,|\mathcal{F}_q\right) \\ &\leq \frac{\mathbb{E}[(W_n - W_q)^2 \,|\mathcal{F}_q] + t^2}{(h+t)^2}.\end{aligned}$$

Setting $t = a(u_q + \gamma_q)/h$ in the last term we have:

$$\begin{aligned}\frac{\mathbb{E}[(W_n - W_q)^2 \,|\mathcal{F}_q] + t^2}{(h+t)^2} &\leq \frac{a(\gamma_q + u_q) + \frac{a^2(\gamma_q+u_q)^2}{h^2}}{(h + \frac{a(\gamma_q+u_q)}{h})^2} \\ &= \frac{a(\gamma_q + u_q)h^2 + a^2(\gamma_q + u_q)^2}{(h^2 + a(\gamma_q + u_q))^2} \\ &= \frac{a(\gamma_q + u_q)}{h^2 + a(\gamma_q + u_q)}.\end{aligned}$$

Now when $h = 1/2\sqrt{bu_q}$ we obtain that:

$$\mathbb{P}\left(\inf_{q\leq i\leq n}(W_i - W_q) \leq -\frac{\sqrt{bu_q}}{2} \,|\mathcal{F}_q\right) \leq \frac{4a(\gamma_q + u_q)}{bu_q + 4a(\gamma_q + u_q)}.$$

Using the second assumption $\gamma_q \lesssim u_q$, a constant $c > 0$ exists such that for q large enough:

$$\mathbb{P}\left(\inf_{q\leq i\leq n}(W_i - W_q) \leq -\frac{\sqrt{bu_q}}{2} \,|\mathcal{F}_q\right) \leq \frac{c}{b+c}.$$

Inserting this bound in (32) ends the proof. ■

At this point, Propositions 15 and 16 enable us to prove that the sequence (S_n) does not converge to 0 a.s. using the arguments of Benaïm (1999). We are now able to conclude the proof of Theorem 3. We gather below the conditions involved by Theorem 1, and Propositions 14, 15 and 16.

- **(A)** Theorem 1 requires that $\mathbf{H}_{\text{Steps}}$ or $\mathbf{H}'_{\text{Steps}}$ hold, namely that:

$$\sum p_n\gamma_{n+1}\sigma^2_{n+1} < +\infty \quad \text{and} \quad \sum \gamma^2_{n+1}\sigma^2_{n+1} < +\infty. \tag{34}$$

- **(B)** Proposition 14 needs:

$$\delta_n = |p_n - p_\infty| + |q_n - q_\infty| + p_n\sigma_n^2 + \gamma_n = o\left(\sqrt{\gamma_n}\sigma_n\right). \tag{35}$$

We observe in particular that $p_n\sigma_n^2 = o(\sqrt{\gamma_n}\sigma_n)$ and $\gamma_n = o(\sqrt{\gamma_n}\sigma_n)$, which implies that:

$$p_\infty = 0.$$

In particular, since $(\sigma_n)_{n\geq 1}$ is a bounded sequence, we deduce that $p_n\sigma_n^2 = O(p_n)$ and therefore, (35) entails:

$$\nu_n + \gamma_n = o(\sqrt{\gamma_n}\sigma_n),$$

which yields:

$$\frac{\beta}{2} + s < r \wedge \rho \quad \text{and} \quad s < \frac{\beta}{2}.$$

- **(C)** Proposition 15 necessitates:

$$\gamma_n = o(\sqrt{u_n}). \tag{36}$$

- **(D)** Proposition 16 finally requires that:

$$\delta_n = o(\sqrt{u_n}) \quad \text{and} \quad \gamma_n = O(u_n). \tag{37}$$

We observe that condition (36) is already included in condition (37) because by definition $\delta_n \geq \gamma_n$. Therefore, we can forget **(C)**, which is contained in **(D)**. Using $u_n \sim n\gamma_n^2\sigma_n^2$, we observe that the first condition of (37) leads to $\nu_n = o(\sqrt{n}\gamma_n\sigma_n)$ and $\sigma_n\sqrt{n} \longrightarrow +\infty$. Since $\sqrt{\gamma_n}\sigma_n = o(\sqrt{n}\gamma_n\sigma_n)$, we conclude that the condition on ν_n in **(D)** is already included in **(B)**. Moreover, the term $p_n\sigma_n^2$ of δ_n is already negligible when compared to $\sqrt{n}\gamma_n\sigma_n$. Finally, the only additional constraint brought by $\delta_n = o(\sqrt{u_n})$ is $\sqrt{n}\sigma_n \longrightarrow +\infty$.
The second condition in (37) shows that $n\gamma_n\sigma_n^2 \nrightarrow 0$ which is stronger than the previous one. We then deduce that Proposition 16 finally needs:

$$s \leq \frac{1-\beta}{2}.$$

We then aggregate all the constraints on s with respect to the choice of the step-size sequence $(\gamma_n)_{n\geq 1}$ and $(p_n, q_n)_{n\geq 1}$ and observe that:

$$\left(\frac{1}{2} - \frac{\beta + r}{2}\right) \vee \left(\frac{1}{2} - \beta\right) \leq s < \frac{\beta}{2} \wedge \frac{1-\beta}{2} \wedge \left(r - \frac{\beta}{2}\right) \wedge \left(\rho - \frac{\beta}{2}\right) \tag{38}$$

while when s attains the lower bound given by the left hand side, we need to tune the mini-batch size as $\sigma_n^{-1} \gtrsim \log(n)$.

Remark 17 *When the variance of the noise sequence does not converge to* 0 *(*$s = 0$*), the previous conditions on the parameters can be summarized as:* $\beta \in [1/2, 1)$*;* $\rho > \beta/2$ *(and* $\rho \leq 1 - \beta$ *if* $q_\infty = 0$*);* $r > 1 - \beta$*, when* $\beta < 2/3$ *and* $r > \beta/2$ *for* $\beta \geq 2/3$*.*

Proof [Proof of Theorem 3] The proof is divived into two steps.
Step 1: S_n does not converge to 0 a.s. Let $\mathcal{G}$ denote the event:

$$\mathcal{G} := \left\{ \lim_{n \longrightarrow +\infty} S_n \neq 0 \right\}.$$

The definition of T_b^q implies that for any $q \in \mathbb{N}^*$ and $n \geq q$:

$$\mathbb{E}[\mathbb{1}_{\mathcal{G}} | \mathcal{F}_n] \mathbb{1}_{T_b^q = n} = \mathbb{E}[\mathbb{1}_{\mathcal{G}} | \mathcal{F}_n] \mathbb{1}_{T_b^q = n} \mathbb{1}_{S_n \geq \sqrt{bu_n}}.$$

In the meantime, if $\mathcal{S} = +\infty$ then (S_n) does not converge to 0, so $\{\mathcal{S} = +\infty\} \subset \mathcal{G}$. For q large enough, such that Proposition 16 holds, and for all $n \geq q$:

$$\mathbb{E}[\mathbb{1}_{\mathcal{G}} | \mathcal{F}_n] \mathbb{1}_{T_b^q = n} \mathbb{1}_{S_n \geq \sqrt{bu_n}} \geq \mathbb{P}(\mathcal{S} = +\infty | \mathcal{F}_n) \mathbb{1}_{T_b^q = n} \mathbb{1}_{S_n \geq \sqrt{bu_n}} \geq \frac{b}{c+b} \mathbb{1}_{T_b^q = n} \mathbb{1}_{S_n \geq \sqrt{bu_n}}.$$

Thus, if we consider all the integers n larger than q, we obtain:

$$\begin{aligned}
\mathbb{E}[1_{\mathcal{G}} \ | \mathcal{F}_q] &\geq \sum_{n \geq q} \mathbb{E}[\mathbb{1}_{\mathcal{G}} \mathbb{1}_{T_b^q = n} \ | \mathcal{F}_q] = \sum_{n \geq q} \mathbb{E}[\mathbb{E}[\mathbb{1}_{\mathcal{G}} | \ \mathcal{F}_n] \mathbb{1}_{T_b^q = n} \ | \mathcal{F}_q] \\
&\geq \sum_{n \geq q} \frac{b}{c+b} \mathbb{E}[\mathbb{1}_{T_b^q = n} \ | \mathcal{F}_q] \\
&\geq \frac{b}{c+b} \mathbb{P}(T_b^q < +\infty | \mathcal{F}_q).
\end{aligned}$$

We now apply Proposition 15 and obtain that:

$$\mathbb{E}[1_{\mathcal{G}} \ | \mathcal{F}_q] \geq \frac{b}{2(c+b)} > 0.$$

By definition, $\mathcal{G} \subset \mathcal{F}_\infty$ so $\lim_{q \to +\infty} \mathbb{E}[\mathbb{1}_{\mathcal{G}} | \mathcal{F}_q] = \mathbb{1}_{\mathcal{G}}$ and thus the previous inequality guarantees that $\mathbb{1}_{\mathcal{G}} = 1$ almost surely.
Step 2: the algorithm escapes any neighborhood of an unstable point in a finite time a.s.

As mentioned before, we shall prove that if the algorithm is at step n_0 in a neighborhood $\mathcal{N}$ of a local maximum, it escapes $\mathcal{N}$ a.s. in a finite time, meaning that $\mathbb{P}(T_{n_0} = +\infty) = 0$, where T_{n_0} is the stopping time defined by (21).
Suppose that $T_{n_0} = +\infty$. In this case, by definition, $X_{n+1} = \eta(\theta_{n+1}, w_{n+1}) - \eta(\theta_n, w_n)$ for all $n \geq n_0$ and thus:

$$S_n = \eta(\theta_n, w_n), \quad \forall n \geq n_0.$$

Theorem 1 ensures that (θ_n, w_n) converges a.s to a point $(\theta_\infty, 0)$. This together with the regularity of the function η implies that the sequence S_n goes to $\eta(\theta_\infty, 0)$ when $n \to +\infty$. Since $\mathcal{N}$ is compact, the limit point $(\theta_\infty, 0)$ belongs to $\mathcal{N}$ and according to Proposition 10 c) there exists $k > 0$ such that:

$$0 \leq k\eta(\theta_\infty, 0) \leq \langle \nabla \eta(\theta_\infty, 0), H(\theta_\infty, 0) \rangle.$$

As seen in the proof of Theorem 1, the limit point $(\theta_\infty, 0)$ is almost surely an equilibrium point for the dynamical system driven by H, so $H(\theta_\infty, 0) = 0$. As a result we have that:

$$\lim_{n \longrightarrow +\infty} S_n = \eta(\theta_\infty, 0) = 0.$$

From Step 1, we have seen that $\mathbb{P}(\lim_{n\to+\infty} S_n = 0) = 0$, which concludes the proof. ■

Acknowledgments

This work is partially supported by the Fondation Simone et Cino Del Duca through the project OpSiMorE, and by the French Agence Nationale de la Recherche (ANR), project under reference ANR-PRC-CE23 MASDOL. S. Gadat acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future program (Investissements d'Avenir, grant ANR-17-EURE-0010). S. Gadat also gratefully acknowledges the Centre Lagrange for its support, as does I. Gavra the Centre Henri Lebesgue, program ANR-11-LABX-0020-0 .

References

F. Alvarez, H. Attouch, J. Bolte, and P. Redont. A second-order gradient-like dissipative dynamical system with hessian-driven damping.: Application to optimization and mechanics. *Journal de Mathématiques Pures et Appliquées*, 81(8):747 – 779, 2002.

H. Attouch, X. Goudou, and P. Redont. The heavy ball with friction method, i. the continuous dynamical system. *Communications in Contemporary Mathematics*, 02(01):1–34, 2000.

H. Attouch, Z. Chbani, and H. Riahi. Rate of convergence of the nesterov accelerated gradient method in the subcritical case 3. *ESAIM: COCV*, 25:2, 2019. doi: 10.1051/cocv/2017083. URL `https://doi.org/10.1051/cocv/2017083`.

F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research*, 15(19):595–627, 2014. URL `http://jmlr.org/papers/v15/bach14a.html`.

A. Barakat and P. Bianchi. Convergence and dynamical behavior of the Adam algorithm for non-convex stochastic optimization. *Preprint*, 2020.

A. Barakat, P. Bianchi, W. Hachem, and Sh. Schechtman. Stochastic optimization with momentum: convergence, fluctuations, and traps avoidance. *Preprint*, 2020.

H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017. doi: 10.1287/moor.2016.0817.

A. Belotto da Silva and M. Gazeau. A general system of differential equations to model first-order adaptive algorithms. *J. Mach. Learn. Res.*, 21:1–42, 2020. ISSN 1532-4435.

M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de probabilités XXXIII*, pages 1–68. Springer, 1999.

M. Benaïm and M. Hirsch. Asymptotic pseudotrajectories and chain recurrent flows, with applications. *Journal of Dynamics and Differential Equations*, 8(1):141–176, 1996.

M. Benaim and Morris W Hirsch. Dynamics of morse-smale urn processes. *Ergodic Theory and Dynamical Systems*, 15(6):1005–1030, 1995.

B. Bercu, M. Costa, and S. Gadat. Stochastic estimation algorithm for superquantile estimation. *arxiv preprint*, 2020a.

B. Bercu, A. Godichon, and B. Portier. An efficient stochastic newton algorithm for parameter estimation in logistic regressions. *SIAM Journal on Control and Optimization*, 58(1):348–367, 2020b.

B. Bercu, J. Bigot, S. Gadat, and E. Siviero. Statistical properties of stochastic newton algorithms for regularized semi-discrete optimal transport, 2022. URL `https://doi.org/10.1093/imaiai/iaac014`.

V.S. Borkar. Stochastic approximation with two time scales. *Systems Control Lett.*, 29, 1997.

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 161–168. NIPS Foundation (http://books.nips.cc), 2008. URL `http://leon.bottou.org/papers/bottou-bousquet-2008`.

L. Bottou, F. E Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

O. Brandiere and M. Duflo. Les algorithmes stochastiques contournent-ils les pieges? In *Annales de l'IHP Probabilités et statistiques*, volume 32, pages 395–427, 1996.

A. Cabot, H. Engler, and S. Gadat. On the long time behavior of second order differential equations with asymptotically small dissipation. *Transactions of the AmE.an Mathematical Society*, 361(11):5983–6017, 2009a.

A. Cabot, H. Engler, and S. Gadat. Second-order differential equations with asymptotically small dissipation and piecewise flat potentials. *Electronic Journal of Differential Equations*, 2009:33–38, 2009b.

H. Cardot, P. Cénac, and A. Godichon-Baggioni. Online estimation of the geometric median in hilbert spaces: Nonasymptotic confidence balls. *The Annals of Statistics*, 45(2):591–614, 2017.

P. Cénac, A. Godichon-Baggioni, and B. Portier. An efficient averaged stochastic gauss-newton algorithm for estimating parameters of non linear regressions models, 2020.

M. Costa and S. Gadat. Non-asymptotic study of a recursive superquantile estimation algorithm. *arxiv preprint*, 2020.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011. ISSN 1532-4435.

M. Duflo. *Algorithmes stochastiques*, volume 23 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 1996. ISBN 3-540-60699-8.

A. Défossez, L. Bottou, F. Bach, and N. Usunier. On the convergence of Adam and Adagrad. 2020.

S. Gadat and F. Panloup. Long time behaviour and stationary regime of memory gradient diffusions. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 564–601, 2014.

S. Gadat and F. Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. 2020.

S. Gadat, F. Panloup, and S. Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, pages 461–529, 2018.

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1-2):59–99, 2016.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

A. Haraux. *Systemes dynamiques dissipatifs et applications*, volume 17. Masson, 1991.

G. Hinton, N. Srivastava, and K. Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. *Cited on*, 14(8), 2012.

C. Jin, P. Netrapalli, and M.I. Jordan. Accelerated gradient descent escapes saddle points faster than gradient descent. In *Conference On Learning Theory*, pages 1042–1085. PMLR, 2018.

R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26:315–323, 2013.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

N. Le Roux, M. Schmidt, F. R Bach, et al. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2672–2680. Citeseer, 2012.

J. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M. I. Jordan, and B. Recht. First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*, 2017.

J. D Lee, M. Simchowitz, M. I Jordan, and B. Recht. Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257, 2016.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 983–992. PMLR, 16–18 Apr 2019. URL `https://proceedings.mlr.press/v89/li19c.html`.

N. Loizou and P. Richtárik. Momentum and stochastic momentum for stochastic gradient, newton, proximal point and subspace descent methods. *Computational Optimization and Applications*, pages 1–58, 2020.

A. Mokkadem and M. Pelletier. Convergence rate and averaging of nonlinear two-time-scale stochastic approximation algorithms. *The Annals of Applied Probability*, 16, 11 2006. doi: 10.1214/105051606000000448.

E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24:451–459, 2011.

A. S. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons Inc., New York, 1983.

Y. Nesterov. *Introductory lectures on convex optimization*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, MA, 2004.

Y. E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. akad. nauk Sssr*, volume 269, pages 543–547, 1983.

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

B. Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4:1–17, 1964.

B. T. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838–855, 1992.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22, 1951.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pages 233–257. Elsevier, 1971.

D. Ruppert. Efficient estimations from a slowly convergent robbins-monro process. *Technical Report, 781, Cornell University Operations Research and Industrial Engineering*, 1988.

O. Sebbouh, R. M Gower, and A. Defazio. On the convergence of the stochastic heavy ball method. *arXiv preprint arXiv:2006.07867*, 2020.

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*, 17(153): 1–43, 2016. URL `http://jmlr.org/papers/v17/15-084.html`.

R. Ward, X. Wu, and L. Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. volume 97 of *Proceedings of Machine Learning Research*, pages 6677–6686, Long Beach, California, USA, 2019. PMLR.

F. Zou, F. Shen, Z. Jie, W. Zhang, and W. Liu. A sufficient condition for convergences of Adam and RMSProp. In *CVPR*, pages 11127–11135, 2019.