HAL open science

A new non-convex framework to improve asymptotical knowledge on generic stochastic gradient descent

Jean-Baptiste Fest, Audrey Repetti, Emilie Chouzenoux

To cite this version:

Jean-Baptiste Fest, Audrey Repetti, Emilie Chouzenoux. A new non-convex framework to improve asymptotical knowledge on generic stochastic gradient descent. MLSP 2023 - IEEE International Workshop on Machine Learning for Signal Processing, Sep 2023, Rome, Italy. hal-04165342

HAL Id: hal-04165342

https://inria.hal.science/hal-04165342v1

Submitted on 18 Jul 2023

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

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

Distributed under a Creative Commons Attribution 4.0 International License

A NEW NON-CONVEX FRAMEWORK TO IMPROVE ASYMPTOTICAL KNOWLEDGE ON GENERIC STOCHASTIC GRADIENT DESCENT

Jean-Baptiste Fest(1), *Audrey Repetti*(2), *and Émilie Chouzenoux*(1)

(1) CVN, CentraleSupélec, Inria, Université Paris-Saclay, 9 rue Joliot Curie, Gif-sur-Yvette, France
(2) School of Engineering & Physical Sciences, School of Mathematical & Computer Schiences, Heriot-Watt University, EH14 4AS Edinburgh, UK

ABSTRACT

Stochastic gradient optimization methods are broadly used to minimize non-convex smooth objective functions, for instance when training deep neural networks. However, theoretical guarantees on the asymptotic behaviour of these methods remain scarce. Especially, ensuring almost-sure convergence of the iterates to a stationary point is quite challenging. In this work, we introduce a new Kurdyka-Łojasiewicz theoretical framework to analyze asymptotic behavior of stochastic gradient descent (SGD) schemes when minimizing non-convex smooth objectives. In particular, our framework provides new almost-sure convergence results, on iterates generated by any SGD method satisfying mild conditional descent conditions. We illustrate the proposed framework by means of several toy simulation examples. We illustrate the role of the considered theoretical assumptions, and investigate how SGD iterates are impacted whether these assumptions are either fully or partially satisfied.

***Index Terms*—Stochastic gradient descent, non-convex optimization, Kurdyka-Lojasiewicz, convergence analysis.**

I. INTRODUCTION

We consider the unconstrained optimization problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^N}{\text{minimize}} \quad F(\boldsymbol{x}), \tag{1}$$

where $F\colon \mathbb{R}^N \to \mathbb{R}$ is a continuously differentiable function ($N \geq 1$), that is not necessarily assumed to be convex. We focus on the challenging situation when the evaluation of the gradient of F is subject to (stochastic) errors, during the iterative resolution procedure. This typically arises in important scenarios of supervised machine learning, when F is an expectation loss to be minimized through the access of data samples [16]. In such context, the non-convexity of F results from the use of nonlinear models, such as deep neural networks, for mapping the data features [8], [21].

The most popular approach to solve (1) in this context is probably the stochastic gradient descent (SGD) [27] and its numerous variants [16], [21], [32]. SGD generates a stochastic sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ defined as

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k, \tag{2}$$

where $(\mathbf{g}_k)_{k\in\mathbb{N}}$ is typically a random process defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ aiming at approximating the true gradient $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$. Moreover, $(\alpha_k)_{k\in\mathbb{N}}$ corresponds to a positive stepsize sequence. Practical applications of such SGD schemes to supervised learning can be found for instance in [7], [20].

In general, SGD schemes satisfy conditional descent properties on the sequence $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$, with respect to the natural filtration $(\mathcal{F}_k)_{k\in\mathbb{N}} = (\sigma(\mathbf{x}_0, \ldots, \mathbf{x}_k))_{k\in\mathbb{N}}$. Such properties can be obtained under technical assumptions on the noise on the approximated gradients and on the sequence $(\alpha_k)_{k\in\mathbb{N}}$ [4], [17], [28]. In particular, for most SGD schemes, $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ satisfies an almost-supermartingale condition, and converges to a finite limit [28]. In addition, the gradient process $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ (or a sub-sequence of it) converges to zero [4].

However, usually little is known on the asymptotic behaviour of the generated process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ itself. The machine learning literature typically focuses instead on the search for (fast) convergence rates [14], [15], [19], [29], and is often limited to the convex (or even strictly convex) setting. The lack of study of stochastic algorithms in a non-convex framework in the literature calls for developing new theory in this context. In particular, it is necessary to leverage existing deterministic approaches, typically relying on Kurdyka-Łojasiewicz (KL) properties [6], [22]. KL properties have initially been introduced to solve gradient flow problems in a continuous setting [1], [5]. Nevertheless, it has proven to be particularly efficient to improve convergences guarantees of discrete optimization schemes when no convexity assumptions are made [2], [6], [11], [12], [26].

In this work, we introduce a new theoretical framework to prove almost sure (a.s.) convergence of the stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ generated by SGD schemes of the form of (2), to a critical point of F, when F is not necessarily convex. We show that this result applies to any SGD schemes

This work is funded by the European Research Council Starting Grant MAJORIS ERC-2019-STG-850925.

979-8-3503-2411-2/23/$31.00 ©2023 IEEE

holding mild conditional descent properties, when F satisfies a Kurdyka-Łojasiewicz (KL) property [6], [22]. We further empirically investigate how the considered assumptions practically control the convergence behaviour of SGD schemes, on some toy simulation examples. The current paper relies on our recent preprint [10]. The originality of the current work is to specialize the study to SGD, and to present a comprehensive numerical illustration for the results.

The remainder of the paper is organized as follow. Section II introduces the theoretical tools required. We present our main theoretical contribution in Section III. Numerical experiments are presented in Section IV, and Section V concludes the work.

II. THEORETICAL BACKGROUND

II-A. KL property

One key mathematical tool for proving convergence of non-convex deterministic optimization schemes is the KL property. It has been initially introduced by Łojasiewicz [23] and Kurdika [22], and has been at the core of major methodological developments in non-convex optimization analysis, in the last decades, starting with the seminal papers from [3], [5].

The definition of KL property is given below.

Definition 1: (KL property) A differentiable function $F : \mathbb{R}^N \to \mathbb{R}$ satisfies the *Kurdyka-Łojasiewicz* (KL) property on $E \subset \mathcal{H}$, if for every $\widetilde{\boldsymbol{x}} \in E$, there exists a neighbourhood V of $\widetilde{\boldsymbol{x}}$, $\zeta > 0$ and $\varphi \in \Phi_\zeta$ such that

$$\|\nabla F(\boldsymbol{x})\| \varphi' \left(F(\boldsymbol{x}) - F(\widetilde{\boldsymbol{x}})\right) \geq 1,$$

for every $\boldsymbol{x} \in V$ satisfying

$$0 < F(\boldsymbol{x}) - F(\widetilde{\boldsymbol{x}}) < \zeta.$$

II-B. Convergence analysis under KL property

Definition 1 was initially motivated in a continuous setting, through the gradient flow analysis [1]. Indeed, KL property promotes finite gradient trajectories converging to the origin. As a discrete counterpart of gradient flows, gradient descent (GD) algorithms are thus expected to follow a similar behaviour.

Let us consider Problem (1), where F satisfies KL property. Let us build the sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ generated by a (deterministic) GD algorithm of the form of

$$(\forall k \in \mathbb{N}) \quad \boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla F(\boldsymbol{x}_k), \tag{3}$$

where, for every $k \in \mathbb{N}$, $\alpha_k > 0$ is a stepsize. If $(\nabla F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ and $(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)_{k\in\mathbb{N}}$ are proportional, and if $\sum_{k=0}^{+\infty} \|\nabla F(\boldsymbol{x}_k)\| < +\infty$ (e.g., if α_k is small enough and F is Lipschitz smooth), then we can deduce from KL property that

$$\sum_{k=0}^{+\infty} \|\boldsymbol{x}_{k+1} - \boldsymbol{x}_k\| < +\infty.$$

This result can then be used to deduce that $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is a convergent Cauchy sequence. Detailed examples of convergence analysis of gradient-based schemes under KL property can be found in [2].

In the non-convex case, KL property thus allows to show that the limit point of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ exists and cancels the gradient (i.e., stationary point), under mild requirements such as descent conditions, on $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$. Note that local convergence results to global solutions can also be obtained, when a good initialization (i.e., close enough to a global minimum of F) is considered [18].

II-C. Uniformized KL property

Definition 1 might be sometimes limited, as it is too 'local' to be easily manipulated. Recently, [6] introduced an alternative version of the KL property that we introduce in the following Theorem.

Definition 2: (Uniformized KL property) Let C be a compact set in $\mathcal{H}$ and $F \colon \mathcal{H} \to \mathbb{R}$ be a differentiable function constant on C, satisfying the KL property on C. Then, there exist $(\varepsilon, \zeta) \in (0, +\infty)^2$ and $\varphi \in \Phi_\zeta$ such that

$$(\forall \overline{\boldsymbol{x}} \in C)(\forall \boldsymbol{x} \in \mathcal{H}) \quad \|\nabla F(\boldsymbol{x})\| \, \varphi' \left(F(\boldsymbol{x}) - F(\overline{\boldsymbol{x}})\right) \geq 1, \tag{4}$$

when $d(\boldsymbol{x}, C) < \varepsilon$ and $0 < F(\boldsymbol{x}) - F(\overline{\boldsymbol{x}}) < \zeta$.

A typical usage of Definition 2 is when $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ is decreasing and F coercive. Then, $(F(\boldsymbol{x}_k))_{k\in\mathbb{N}}$ actually converges to a finite limit while $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ is guaranteed to be bounded. As a consequence, taking C in Definition 2 as the set of cluster points of $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ allows (4) to be verified for any iterate starting from a certain rank.

III. CONVERGENCE OF SGD FOR NON-CONVEX OBJECTIVES

III-A. Generic SGD scheme

Let us consider SGD schemes of the form of (2) for solving (1). We denote $(\mathbf{x}_k)_{k\in\mathbb{N}}$, a stochastic process generated by (2). In our study, we assume that the process satisfies the two following conditions.

Assumption 1: F is coercive and β-Lipschitz differentiable on $\mathbb{R}^N$.

Assumption 2: F satisfies the KL property on the set of critical points of F. Moreover, this set can be written as the finite union of non-empty disjoint compacts subset.

Assumption 1 is a classical hypothesis usually made in the field of differentiable optimization context [25]. In particular, it ensures the existence of a minimal value of F, denoted by $F_{\min}$. On the contrary, Assumption 2 is specific to our non-convex context, as we do not have any convexity assumption on F. Omitting some technical details here, Assumption 2 is essential for us to obtain convergence results directly on the iterates following a similar strategy as those conducted, e.g., in [2], [6], [11], but generalized to our stochastic framework. Note also that the geometry imposed on the critical set is only slightly constraining.

III-B. Gradient approximation assumptions

Before establishing our main convergence result, we first introduce technical conditions on the stochastic approximations $(\mathbf{g}_k)_{k\in\mathbb{N}}$ of the gradients $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$, involved in the SGD updates.

Assumption 3: There exists three deterministic non-negative sequences $(a_k)_{k\in\mathbb{N}}, (b_k)_{k\in\mathbb{N}}, (c_k)_{k\in\mathbb{N}}$ such that

i) $\sum_{k=0}^{+\infty} \alpha_k \left(\sqrt{a_k} + \sqrt{c_k}\right) < +\infty$,

ii) $\inf_{k\in\mathbb{N}} \alpha_k \left(1 - \frac{\alpha_k b_k \beta}{2}\right) > 0$,

and, for every $k \in \mathbb{N}$,

iii) $0 < \sqrt{\frac{b_{k+1}}{b_k}} \frac{2-\alpha_k b_k \beta}{2-\alpha_{k+1} b_{k+1}\beta} \left(1 + \frac{\alpha_{k+1}^2 a_{k+1}\beta}{2}\right) \leq 1$

iv) $\mathbb{E}_k[\mathbf{g}_k] = \nabla F(\mathbf{x}_k)$

v) $\mathbb{E}_k[\|\mathbf{g}_k\|^2] \leq a_k(F(\mathbf{x}_k) - F_{\min}) + b_k\|\nabla F(\mathbf{x}_k)\|^2 + c_k$.

Although Assumption 3 may seem quite demanding, it actually gathers several typical scenarios:

- Assumption 3-iv) and 3-v) are relative to the two first moments of the noise on the gradient term. Assumption 3-iv) classically requires the noise to be unbiased, which Assumption 3-v) is a mild condition on the noise variance, generalizing many encountered in the literature [7], [20], [31], [33]. A non-zero $(a_k)_{k\in\mathbb{N}}$ sequence typically models cases when F has a gradient confusion bound [30].
- Assumptions 3-i) and 3-ii) are classical summability and non-zero rules, controling the terms in Assumption 3-v). Assumption 3-ii) assumes a non-vanishing stepsize in the SGD update, and requires $(b_k)_{k\in\mathbb{N}}$ to be bounded.
- Assumption 3-iii) is a non-increasing condition which naturally appears in our convergence proof.

As shown in [10], Assumption 3 guarantees that the process satisfies some conditional descent properties. As a consequence, $(F(\mathbf{x}_k)_{k\in\mathbb{N}})_{k\in\mathbb{N}}$ converges a.s. to an a.s finite random variable F_∞, and that $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ almost surely converges to $\mathbf{0}_N$ (i.e., the zero vector of $\mathbb{R}^N$).

III-C. Proposed KL analysis for stochastic framework

In the context of stochastic optimization, the use of KL property is challenging. A first idea would be to apply the uniformized KL property (Definition 2) to any trajectory $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$, for every $\omega \in \Omega$. However, by doing so, (ϵ, ζ) and φ would be random objects whose analysis is very delicate, dependent on ω. For instance, conditional expectation operations would become tricky, and measurability of $(\mathcal{F}_k)_{k\in\mathbb{N}}$ would not be straightforward.

To overcome this challenge, we proposed in [10] a new extension of Theorem 2, better adapted to a stochastic optimization framework.

Proposition 1: Under Assumptions 1-3, there exists a bounded concave function φ and an a.s. finite positive discrete random variable K such that

$$\|\nabla F(\mathbf{x}_k)\|\varphi'(F(\mathbf{x}_k) - \mathrm{F}_\infty) \geq 1$$

a.s. for every $k > \mathrm{K}$.

The advantage of Proposition 1 (whose detailed proof is given in [10]), lies in the random variable K which concentrates all the stochastic information. As such, this new tool tends to overcome some technical obstacles raised in KL-based convergence analysis [10], and allows to build a new convergence theorem, that we present hereafter in the SGD case.

III-D. Convergence result

We now introduce our main contribution, which is the almost-sure convergence result for the generic SGD scheme (2). Let us denote by F_∞ the almost-sure finite limit of $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ and $\mathbb{E}_k[.]$ the conditional expectation operator relative to $\mathcal{F}_k$, for $k \in \mathbb{N}$ (i.e., for a given integrable or positive random variable, $\mathbb{E}_k[.]$ corresponds to its best approximation regarding all information available on the process from state 0 to state k). We introduce, for all $(k, \gamma) \in \mathbb{N} \times (0, 1)$, the event:

$$\Xi_{\gamma,k} := \Big\{ \mathrm{F}_\infty < F(\mathbf{x}_k) \text{ and } \Big|(\mathrm{F}_\infty - F_{\min}) - \frac{2}{2+\alpha_k^2 a_k \beta}\mathbb{E}_k[F(\mathbf{x}_{k+1}) - F_{\min}] + \frac{\alpha_k^2 c_k \beta}{2}\Big| \leq \gamma\big(1 - \frac{\alpha_k b_k \beta}{2}\big)\|\nabla F(\mathbf{x}_k)\|^2 \Big\}. \quad (5)$$

Theorem 1: Under Assumptions 1, 2 and 3, if there exists $\gamma_0 \in (0, 1)$ such that $\mathbb{P}(\Xi_{\gamma_0,k}) = 1$ for all $k \in \mathbb{N}$, then $(\mathbf{x}_k)_{k\in\mathbb{N}}$ almost-surely converges to a critical point $\mathbf{x}_\infty$ of F.

Equality $\mathbb{P}(\Xi_{\gamma_0,k}) = 1$ (for all $k \in \mathbb{N}$) supposes that process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is well-built enough to verify a suitable descent condition and to approach its limit F_∞ from above. Moreover, it also reflects a predictability condition; the conditional decreasing as well as the difference $\mathrm{F}_\infty - F_{\min}$ shall be controlled with respect to the evolution of the gradient norm.

The complete proof for Theorem 1 can be found in [10]. It relies on the use of Proposition 1 as a cornerstone to establish the finite length of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ almost surely. Up to our knowledge, Theorem 1 is one of the first results ensuring the almost convergence of a stochastic gradient type iterates in a non-convex setting.

Table I gives a few examples of state-of-the-art schemes directly verifying our specific Assumption 3. This table is not exhaustive, and Assumption 3 could be verified by other algorithms, e.g., [8], [9] (see also [10] for proximal algorithms).

IV. NUMERICAL ILLUSTRATIONS

In this section we conduct some experiments on a non-convex scalar problem, so as to illustrate the behaviour of SGD algorithm when Assumptions 1, 2 and 3 are satisfied by function F, and its moments approximation gradient sequence. To do so, we proceed by gradually increasing the complexity of noise structure.

Scheme	α_k	a_k	b_k	c_k	Ass. 3-i) and ii) ?
SGD [31]	α	0	B^2	0	Yes
SGD [13]	$\lambda_k\gamma_k$	0	$1+\tau_k$	ζ_k	$\inf_{k\in\mathbb{N}} \gamma_k\lambda_k > 0$
BFGS [34]	α_k	0	1	$\sigma^2 m_k^{-1}$	$\sum_{k=0}^{+\infty} m_k^{-1} < +\infty$
BFGS [24]	η	0	ρ	0	Yes

Table I. Examples of state-of-the-art algorithms satisfying conditions of Theorem 1, hence ensuring their convergence in a non-convex setting. For the sake of readability, we perused the same notations as the authors in their articles.

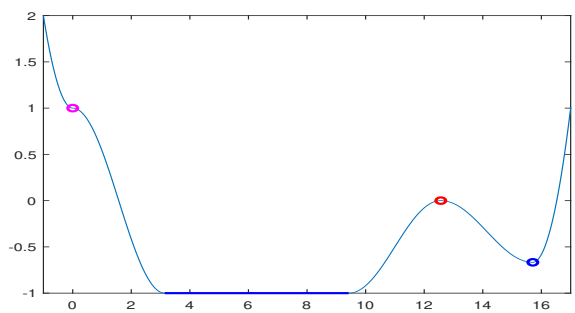

Fig. 1. Graph of function F in (6). Global minimizers correspond to $[\pi, 3\pi]$. Point $x = 5\pi$ in blue (resp. $x = 4\pi$ in in red) corresponds to a local minimizer (resp. maximizer), and $x = 0$ is a saddle point.

All along this study, we work with the following F function.

$$F : x \in \mathbb{R} \mapsto \begin{cases} x^2 + 1 \text{ if } x < 0, \\ \cos(x) \text{ if } 0 \leq x < \pi, \\ -1 \text{ if } \pi \leq x < 3\pi, \\ 1/2\cos(x) - 1/2 \text{ if } 3\pi \leq x < 4\pi, \\ 1/3\cos(x) - 1/3 \text{ if } 4\pi \leq x < 5\pi, \\ x^2 - 10\pi x + 25\pi^2 - 2/3 \text{ if } x \geq 5\pi. \end{cases} \tag{6}$$

The graph of function F in (6) is illustrated in Figure 1. This function is non-convex, but Lipschitz-differentiable with Lipschitz constant equals to $\beta = 2$. Hence Assumption 1 holds. Moreover, since the graph of F is semi-algebraic, Assumption 2 is also verified [5].

In our simulations, we first numerically verify that $(\mathrm{x}_k)_{k\in\mathbb{N}}$ converges almost surely to a stationary point of F, denoted by x_∞. Then we analyse more specifically how x_∞ is approximated by $(\mathrm{x}_k)_{k\in\mathbb{N}}$, ideally in such a way that there exists $\gamma_0 \in (0,1)$ such that $\mathbb{P}(\Xi_{\gamma_0,k}) = 1$ for all $k \in \mathbb{N}$. In particular, we run experiments considering different settings for the sequences $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$ and $(c_k)_{k\in\mathbb{N}}$ appearing in Assumption 3.

IV-A. Experiments under $a_k := 0, c_k := 0$

In this section we investigate the case where the only non-zero sequence (except for the stepsize) interfering in Assumption 3 is $(b_k)_{k\in\mathbb{N}}$. Such kind of noise assumption is generally considered as a baseline in the literature of stochastic optimization, as it is typically verified for the usual constant stepsize SGD scheme when F satisfies the Strong Growth condition [31]. Assumptions 3 i)-iii) are then verified as soon as $(b_k)_{k\in\mathbb{N}}$ is a non-increasing sequence. As the latter have to be bounded to fulfill Assumption 3-ii), it thus becomes equivalent to take $b_k \equiv b$, for every $k \in \mathbb{N}$, to also verify Assumption 3-v).

x_0	b	Nature of x_∞	$F(\mathrm{x}_k) > \mathrm{F}_\infty\ \forall k \in \mathbb{N}$?
$-1/2$	10	Saddle point	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
	10^3	Saddle point	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
1	10	Global Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
	10^3	Global Min	Yes
$4\pi + \epsilon$	10	Global or Local Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
	10^3	Local Max	No

Table II. Asymptotical behaviours of $(\mathrm{x}_k)_{k\in\mathbb{N}}$ under $\mathbb{E}_k\left[\mathrm{g}_k^2\right] \leq bF'(\mathrm{x}_k)^2$ $(k \in \mathbb{N})$

The approximation sequence is generated empirically so as to satisfy both Assumption 3-iv) and v). Specifically, we set, for every $k \in \mathbb{N}$, $\mathrm{g}_k = \mathrm{e}_k^1 \nabla F(\mathrm{x}_k)$, where e_k^1 is sampled uniformly in $\left[1 - \sqrt{3b-1}, 1 + \sqrt{3b-1}\right]$. For the stepsize, we set, for every $k \in \mathbb{N}$, $\alpha_k = \alpha$ such that $\alpha b = \beta^{-1}$. Finally, we introduce two different perturbation levels to test the SGD algorithm in particularly extreme cases: a moderate perturbation $b = 10$, and an excessively high perturbation $b = 10^3$.

Table II shows the behaviour of process $(\mathrm{x}_k)_{k\in\mathbb{N}}$ considering three different initializations $x_0 \in \{-1/2, 1, 4\pi + \epsilon\}$. The first initialization $x_0 = -1/2$ is located on the left of saddle point $x = 0$ (see graph of F in Figure 1). The second initialization $x_0 = 1$ is between the saddle point and the interval $[\pi, 3\pi]$ of global minimizers. And the last one $x_0 = 4\pi + \epsilon$ is in a small neighborhood of the local maximizer $x = 4\pi$, taking $\epsilon = 10^{-5}$. In most scenarios, $(\mathrm{x}_k)_{k\in\mathbb{N}}$ converges to a stationary point so that $F(\mathrm{x}_k)_{k\in\mathbb{N}}$ remains above its limit. The only tricky case arises for high perturbation level when $x_0 = 4\pi + \epsilon$, which seems to be too close to local maximizer $x = 4\pi$.

IV-B. Experiments under $a_k := 0$

In this section, we no longer impose $(c_k)_{k\in\mathbb{N}}$ to be equal to zero. Such a situation is regularly encountered as a first relaxed version of the noise resulting from the Strong Growth condition [31]. One typical situation is when, for $k \in \mathbb{N}$, the difference between g_k and $F'(\mathrm{x}_k)$ follows a Gaussian distribution that remains independent from $\mathcal{F}_k$. We adopt such model to conduct our investigation. More specifically, we choose, for all $k \in \mathbb{N}$, $\mathrm{g}_k = \mathrm{e}_k^1 F'(\mathrm{x}_k) + \mathrm{e}_k^2$. Here, e_k^2 is normally distributed, with zero-mean and standard deviation $\sigma_k > 0$, and does not depend on process $(\mathrm{x}_k)_{k\in\mathbb{N}}$ so as to have $c_k = 2\sigma_k^2$. Moreover, e_k^1 keeps the same properties as in Section IV-A. In order to easily verify Assumption 3 i)-

x_0	σ	Nature of x_∞	$F(\mathrm{x}_k) > \mathrm{F}_\infty\ \forall k \in \mathbb{N}$?
$-1/2$	10	Global Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
	10^2	Global Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
1	10	Global Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
	10^2	Global Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
$4\pi + \epsilon$	10	Global or Local Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
	10^2	Global or Local Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$

Table III. Asymptotical behaviours of $(\mathrm{x}_k)_{k\in\mathbb{N}}$ under $\mathbb{E}_k\left[\mathrm{g}_k^2\right] \leq 20F'(\mathrm{x}_k)^2 + 2\sigma^2/(k+1)^{(2+2\varepsilon)}$ $(k \in \mathbb{N})$

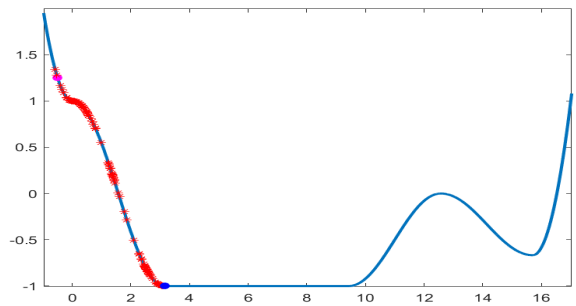

Fig. 2. Example of a typical situation where process $(\mathrm{x}_k)_{k\in\mathbb{N}}$ escapes from saddle point $x = 0$. Pink point: starting point $x = -1/2$; Blue point: End point belonging to $[\pi, 3\pi]$.

iii) , we set constant sequences $(\alpha_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$, equal to α and b, respectively. In practice, we choose $b = 10$ with $\alpha b = \beta^{-1}$, and $\sigma_k := \sigma/(k+1)^{(1+\varepsilon)}$ for $\sigma \in \{10, 10^2\}$.

In our simulations, we choose $b = 10$ as a moderate level of multiplicative noise, and add $(c_k)_{k\in\mathbb{N}}$ as an additive one. Despite the higher complexity of the uncertainty model, we obtain slightly better results as the process is able to escape from saddle or local minimizer in all runs (see Figure 2).

A summary of the behaviour of process $(\mathrm{x}_k)_{k\in\mathbb{N}}$ under these conditions is reported in Table III.

IV-C. Experiments for non-zero $(a_k, b_k, c_k)_{k\in\mathbb{N}}$

Here we consider the more generic case where none of the sequences $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$ nor $(c_k)_{k\in\mathbb{N}}$ is equal to 0.

We suppose that, for every $k \in \mathbb{N}$, the second order moment of g_k also possesses a non-zero component relatively to $F(\mathrm{x}_k) - F_{\min}$. The most difficult assumption to verify is then Ass. 3-iii), which cannot hold for $(\alpha_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ constant sequences. One can show that choosing, for every $k \in \mathbb{N}$, $\alpha_k = \alpha$, $b_k = 3b/(k+1)^2$, and $c_k = 3\sigma^2/(k+1)^{2+2\epsilon}$, for $\alpha > 0$, $b > 0$, $\sigma > 0$, and $a_k = 2(\alpha^2\beta k^{2+\epsilon})^{-1}$, is sufficient to fulfill Assumptions 3-i)-iii).

We simulate the approximation sequence $\mathrm{g}_k = \mathrm{e}_k^3\sqrt{F(\mathrm{x}_k) - F_{\min}} + \mathrm{e}_k^1 F'(\mathrm{x}_k) + \mathrm{e}_k^2$, where both e_k^1 and e_k^3 follows a 1-mean uniform distribution between $[1 - \sqrt{3b-1}, 1 + \sqrt{3b-1}]$ and $\left[-\sqrt{3}/\alpha(k+1)^{-(1+\epsilon)}, \sqrt{3}/\alpha(k+1)^{-(1+\epsilon)}\right]$, respectively, and $\mathrm{e}_k^2 \sim \mathcal{N}(0, \sigma_k^2)$ with $\sigma_k = \sigma/(k+1)^{(1+\varepsilon)}$.

A summary of the behaviour of process $(\mathrm{x}_k)_{k\in\mathbb{N}}$ under these conditions is reported in Table IV. The only significant difference of behavior compared with the previous subsection is observed for $(\sigma, x_0) = (10^2, 1)$. In this particular case, numerical errors prevent the algorithm from converging.

x_0	σ	Nature of x_∞	$F(\mathrm{x}_k) > \mathrm{F}_\infty\ \forall k \in \mathbb{N}$?
$-1/2$	10	Global Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
	10^2	Global Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
1	10	Global Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
	10^2	Non-convergence	Undefined
$4\pi + \epsilon$	10	Global or Local Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$
	10^2	Global or Local Min	Yes if $F(\mathrm{x}_k) \neq F(\mathrm{x}_\infty)$

Table IV. Asymptotical behaviours of $(\mathrm{x}_k)_{k\in\mathbb{N}}$ under $\mathbb{E}_k\left[\mathrm{g}_k^2\right] \leq \frac{8}{\alpha^2\beta}(F(\mathrm{x}_k) - F_{\min})^2 + 40F'(\mathrm{x}_k)^2 + 4\sigma^2/(k+1)^{(2+2\varepsilon)}$ $(k \in \mathbb{N})$.

V. CONCLUSION

In this article we introduce a new theoretical framework to study almost sure convergence of SGD schemes, in a non-convex context. We further give numerical illustrations to investigate the behaviour of SGD processes, and the relevancy of the different assumptions necessary to ensure their almost sure convergence. This work is base on the theoretical work we initially conducted in [10], where we introduced a new KL framework to investigate almost sure convergence of stochastic processes, in a smooth but non-convex context.

VI. REFERENCES

[1] P.-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[2] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116:5–16, 2009.

[3] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[4] D. P. Bertsekas and J. N. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

[5] J. Bolte, A. Daniilidis, and A. Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[6] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

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

[8] C. Castera, J. Bolte, C. Févotte, and E. Pauwels. Second-order step-size tuning of SGD for non-convex optimization. *Neural Processing Letters*, 54(3):1727–1752, 2022.

[9] E. Chouzenoux and J.-B. Fest. SABRINA: A stochastic subspace majorization-minimization algorithm. *Journal of Optimization Theory and Applications*, 195:919–952, 2022.

[10] E. Chouzenoux, J.-B. Fest, and A. Repetti. A kurdyka-lojasiewicz property for stochastic optimization algorithms in a non-convex setting. Technical report, 2023. https://arxiv.org/abs/2302.06447.

[11] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. Variable metric forward–backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132, 2014.

[12] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. A block coordinate variable metric forward–backward algorithm. *Journal of Global Optimization*, 66(3):457–485, 2016.

[13] P. L. Combettes and J.-C. Pesquet. Stochastic forward-backward and primal-dual approximation algorithms with application to online image restoration. In *Proceedings of the 24th European Signal Processing Conference (EUSIPCO 2016)*, pages 1813–1817, 2016.

[14] A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27, 2014.

[15] A. Defazio, J. Domke, et al. Finito: A faster, permutable incremental gradient method for big data problems. In *Proceedings of the International Conference on Machine Learning (ICML 2014)*, pages 1125–1133, 2014.

[16] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

[17] S. Gadat. Stochastic optimization algorithms, non asymptotic and asymptotic behaviour. *Lecture Notes, University of Toulouse*, 2017. https://perso.math.univ-toulouse.fr/m2r/files/2016/02/B5-2016-2017-new.pdf.

[18] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan. How to escape saddle points efficiently. In *Proceedings of the International Conference on Machine Learning (ICML 2017)*, pages 1724–1732. PMLR, 2017.

[19] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

[20] A. Khaled and P. Richtárik. Better theory for SGD in the nonconvex world. Technical report, 2020. https://arxiv.org/abs/2002.03329.

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. Technical report, 2014. https://arxiv.org/abs/1412.6980.

[22] K. Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'Institut Fourier*, volume 48, pages 769–783, 1998.

[23] S. Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[24] S. Y. Meng. *Stochastic second-order optimization for over-parameterized machine learning models*. PhD thesis, University of British Columbia, 2020.

[25] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 1999.

[26] A. Repetti and Y. Wiaux. Variable metric forward-backward algorithm for composite minimization problems. *SIAM Journal on Optimization*, 31(2):1215–1241, 2021.

[27] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

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

[29] N. Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. *Advances in Neural Information Processing Systems*, 25, 2012.

[30] K. A. Sankararaman, S. De, Z. Xu, W. R. Huang, and T. Goldstein. The impact of neural network over-parameterization on gradient confusion and stochastic gradient descent. In *Proceedings of the International Conference on Machine Learning (ICML 2020)*, pages 8469–8479. PMLR, 2020.

[31] M. Schmidt and N. L. Roux. Fast convergence of stochastic gradient descent under a strong growth condition. Technical report, 2013. https://arxiv.org/abs/1308.6370.

[32] T. Tieleman, G. Hinton, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2):26–31, 2012.

[33] S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)*, pages 1195–1204. PMLR, 2019.

[34] X. Wang, S. Ma, D. Goldfarb, and W. Liu. Stochastic quasi-Newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017.