HAL open science

A Kurdyka-Lojasiewicz property for stochastic optimization algorithms in a non-convex setting

Emilie Chouzenoux, Jean-Baptiste Fest, Audrey Repetti

▶ **To cite this version:**

Emilie Chouzenoux, Jean-Baptiste Fest, Audrey Repetti. A Kurdyka-Lojasiewicz property for stochastic optimization algorithms in a non-convex setting. 2023. hal-03992655

HAL Id: hal-03992655

https://inria.hal.science/hal-03992655v1

Preprint submitted on 16 Feb 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.

arXiv:2302.06447v1 [math.OC] 13 Feb 2023

A Kurdyka-Łojasiewicz property for stochastic optimization algorithms in a non-convex setting*

Emilie Chouzenoux*, Jean-Baptiste Fest*, and Audrey Repetti†‡

* CVN, CentraleSupélec, Inria, Université Paris-Saclay, 9 rue Joliot Curie, Gif-sur-Yvette, France
† Institute of Sensors, Signals and Systems, Heriot-Watt University, Edinburgh, UK
‡ Maxwell Institute for Mathematical Sciences, Edinburgh, UK
E-mail: emilie.chouzenoux@inria.fr, jean-baptiste.fest@centralesupelec.fr, a.repetti@hw.ac.uk

Abstract

Stochastic differentiable approximation schemes are widely used for solving high dimensional problems. Most of existing methods satisfy some desirable properties, including conditional descent inequalities [44, 27], and almost sure (a.s.) convergence guarantees on the objective function, or on the involved gradient [47, 26]. However, for non-convex objective functions, a.s. convergence of the iterates, i.e., the stochastic process, to a critical point is usually not guaranteed, and remains an important challenge. In this article, we develop a framework to bridge the gap between descent-type inequalities and a.s. convergence of the associated stochastic process. Leveraging a novel Kurdyka-Łojasiewicz property, we show convergence guarantees of stochastic processes under mild assumptions on the objective function. We also provide examples of stochastic algorithms benefiting from the proposed framework and derive a.s. convergence guarantees on the iterates.

Keywords. Stochastic processes, Kurdyka-Łojasiewicz properties, stochastic gradient descent, proximal algorithms

MSC. 65K05, 90C26, 90C15, 90C53.

1 Introduction

In this work, we aim at solving problem

$$\underset{\boldsymbol{x}\in\mathcal{H}}{\text{minimize}} \quad F(\boldsymbol{x}), \tag{1.1}$$

*This work was supported by the European Research Council Starting Grant MAJORIS ERC-2019-STG-850925 and by the Royal Society of Edinburgh.

where $F : \mathcal{H} \to \mathbb{R}$ is a continuously differentiable function defined on a finite-dimensional real Hilbert space $\mathcal{H}$. We consider a generic stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to solve (1.1) on a probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$. Generating $(\mathbf{x}_k)_{k\in\mathbb{N}}$ typically results from a stochastic approximation scheme [43, 20]. Celebrated examples are often based on stochastic gradient descent (SGD) algorithms, where $(\mathbf{x}_k)_{k\in\mathbb{N}}$ arises from a gradient descent scheme with the gradient of F being replaced by stochastic approximations.

The objective of this work is to develop a new framework, based on Kurdyka-Łojasiewicz (KL) theory [30, 6], to derive almost sure (a.s.) convergence guarantees of the stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$, when F is not convex.

When (1.1) is solved using a deterministic scheme, the main advantage of KL condition lies in its ability to promote interesting asymptotic behavior when F is non-necessarily convex. In this context, KL has been used to prove convergence of proximal point algorithms [1, 2], of simple splitting algorithms such as the forward-backward algorithm and its variants [2, 16, 7, 22, 12, 42, 8, 9], as well as other algorithms based on the majorization-minimization principle [15, 13, 11]. A natural question is to investigate the transfer of the proof techniques from the deterministic setting, to the stochastic setting, for asymptotic analysis including a.s. convergence of stochastic processes. Such an extension is quite challenging, mainly due to the dynamics of the functions involved in KL conditions, that cannot be controlled in a stochastic environment. Although KL inequality has already been invoked for L^1 or L^2 convergences of some stochastic schemes [24], to the best of our knowledge, no generalized KL inequality has been developed yet in this context. Hence, no asymptotic analysis including a.s. convergence of the stochastic processes has been properly formalized and completed. In particular, the authors of [35] mention that a stochastic formulation of the KL inequality would enable to study the accumulation points of the process, without however providing any theoretical results.

In this paper, we investigate the conditions that a stochastic process $(\mathbf{x}_k)_{k\in\mathbb{N}}$ minimizing F must satisfy to ensure its a.s. convergence to a critical point, under the KL inequality. To this aim, we design a new KL condition for differentiable functions, that can be used in a stochastic setting. Using this new framework, we furthermore derive conditions to ensure the convergence of processes generated by a few generic stochastic schemes based on SGD and proximal iterations. In particular, we show that our conditions enable to ensure the convergence of some state-of-the-art SGD algorithms [50, 17, 36], stochastic proximal-gradient algorithms (also known as forward-backward) [17, 25, 34], and stochastic proximal algorithms [48].

The remainder of the paper is organized as follows. Our general stochastic framework is described in section 2. In this section, we introduce general assumptions, and give preliminary results, including a first convergence result in a simplified setting. In section 3 we provide a deterministic extension of the uniformized KL property introduced in [1, Lemma 6]. We then use this extension to design a KL condition that can be used in a stochastic setting. In section 4, using the stochastic KL inequality introduced in section 3, we show the a.s. convergence of a family of stochastic processes $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a critical point. Finally, in section 5 and section 6, we apply the results of section 4 to exhibit conditions for convergence of SGD and stochastic proximal-based

algorithms, respectively, in a non-convex context.

Notation Let $(\mathcal{H}, \langle\cdot|\cdot\rangle)$ be a finite dimensional Hilbert space. $\|\cdot\|$ denotes the canonical norm associated with $\mathcal{H}$. For any subset $E \subset \mathcal{H}$, the distance function to this set is denoted by $\mathrm{dist}(\cdot, E) = \inf_{\boldsymbol{x}\in E} \|\cdot - \boldsymbol{x}\|$. Bold letters as $\boldsymbol{x}$ are used for deterministic vectors, while straight bold letters as $\mathbf{x}$ are used for stochastic vectors. Similarly, x is used for denoting a deterministic scalar variable, while straight x denotes a stochastic scalar variable. Upper-case letters are used for functions. The variable $F(\mathbf{x})$ denotes the stochastic value of function F evaluated at the stochastic variable $\mathbf{x}$. We work on the probability space $(\Omega, \mathcal{F}, P)$. We remind that a condition holds *almost surely* (a.s.) if it holds on a probability-one event of $\mathcal{F}$. We denote by $\mathbb{E}[\cdot]$ the expectation operator, and $\mathbb{E}[\cdot|\mathcal{G}]$ the conditional expectation operator regarding a generic sub sigma-algebra $\mathcal{G} \subset \mathcal{F}$. Let ∇F be the gradient of F. Then we denote by $\mathrm{zer}\,\nabla F$ the set of zeros of ∇F, i.e., the set of critical points of F. Finally, we introduce χ^∞, the set of accumulation points of sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$. So χ^∞ is a random variable from Ω to the set of subsets of $\mathcal{H}$.

2 General assumptions and preliminary results

In this section, we introduce the probabilistic setting and assumptions used in the remainder of this work.

2.1 Assumptions

The following generic assumptions will guide us throughout the remainder of this work.

Assumption 2.1 F is coercive and continuously differentiable.

Assumption 2.2 Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a stochastic process.

(i) The sequence $(F(\mathbf{x}_k)_{k\in\mathbb{N}})_{k\in\mathbb{N}}$ converges a.s. to a random variable F_∞. In addition, F_∞ is finite a.s.

(ii) $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ almost surely converges to $\mathbf{0}_N$.

These assumptions are very common, and are satisfied by a wide range of stochastic algorithms. The first assumption is standard in the field of differentiable optimization and notably ensures that F admits at least one global minimizer [4]. Assumption 2.2 is usually satisfied by classical schemes. Indeed, convergence of $(F(\mathbf{x}_k)_{k\in\mathbb{N}})_{k\in\mathbb{N}}$ and $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ are often obtained easily in both deterministic [40], and stochastic [44, 20] settings. Note that Assumption 2.2(ii) necessitates the almost sure convergence of the gradient to zero. This condition is verified in particular by stochastic schemes with a conditional strong growth condition, see for instance [50, 32].

2.2 Preliminary results

We will now give some preliminary results, with the aim to link the different sets related to convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ and $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$.

Proposition 2.3 *Under theorem 2.1 and theorem 2.2, we have*

(i) $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *is a.s. bounded.*

(ii) χ^∞ *is a.s. non empty and compact.*

(iii) $\chi^\infty \subset \operatorname{zer}\nabla F$ *a.s.*

(iv) $(\operatorname{dist}(\mathbf{x}_k, \chi^\infty))_{k\in\mathbb{N}}$ *converges a.s. to* 0.

Proof.

According to theorem 2.2, there exists a set $\Lambda \subset \Omega$ of probability one where, for all $\omega \in \Lambda$,

$$\lim_{k\to+\infty} F(\mathbf{x}_k(\omega)) < +\infty, \tag{2.1a}$$

$$\lim_{k\to+\infty} \|\nabla F(\mathbf{x}_k(\omega))\| = 0. \tag{2.1b}$$

Inequality (2.1a) implies that $(F(\mathbf{x}_k(\omega)))_{k\in\mathbb{N}}$ is a bounded sequence. According to Assumption 2.1, F being coercive, $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ is also bounded. It follows that the set of cluster points $\chi^\infty(\omega)$ is non empty, bounded and closed in $\mathcal{H}$, hence compact.

Let $\omega \in \Lambda$. We will now show that $\chi^\infty(\omega) \subset \operatorname{zer}\nabla F$. Let $\boldsymbol{x}_\infty \in \chi^\infty(\omega)$. There exists a subsequence $\left(\mathbf{x}_{\psi(k)}(\omega)\right)_{k\in\mathbb{N}}$ converging to $\boldsymbol{x}_\infty$ as $k \to +\infty$. Moreover, from (2.1b), $(\nabla F(\mathbf{x}_{\psi(k)}(\omega)))_{k\in\mathbb{N}}$ converges to $\mathbf{0}_N$. Since, according to theorem 2.1, the gradient of F is continuous, we thus obtain $\nabla F(\boldsymbol{x}_\infty) = \mathbf{0}_N$, and hence $\boldsymbol{x}_\infty \in \operatorname{zer}\nabla F$.

We now show that (iv) holds. By contradiction, if $(\operatorname{dist}(\mathbf{x}_k(\omega), \chi^\infty(\omega))_{k\in\mathbb{N}}$ does not converge to 0, then there exist $\varepsilon > 0$ and a subsequence $(\mathbf{x}_{\psi_1(k)}(\omega))_{k\in\mathbb{N}}$ such that

$$(\forall k \in \mathbb{N}) \quad \operatorname{dist}\left(\mathbf{x}_{\psi_1(k)}(\omega), \chi^\infty(\omega)\right) > \epsilon \tag{2.2}$$

Since $(\mathbf{x}_k(\omega))_{k\in\mathbb{N}}$ is bounded, $(\mathbf{x}_{\psi_1(k)}(\omega))_{k\in\mathbb{N}}$ is also bounded. So the set of cluster points of $(\mathbf{x}_{\psi_1(k)}(\omega))_{k\in\mathbb{N}}$ is non-empty and included in $\chi^\infty(\omega)$. Thus, there exists another subsequence $(\mathbf{x}_{(\psi_1\circ\psi_2)(k)}(\omega))_{k\in\mathbb{N}}$ and $\boldsymbol{x}'_\infty \in \chi^\infty(\omega)$ such that $\left\|\mathbf{x}_{(\psi_1\circ\psi_2)(k)}(\omega) - \boldsymbol{x}'_\infty\right\| \underset{k\to+\infty}{\longrightarrow} 0$. Hence, $\operatorname{dist}\left(\mathbf{x}_{(\psi_1\circ\psi_2)(k)}(\omega), \chi^\infty(\omega)\right) \underset{k\to+\infty}{\longrightarrow} 0$, which contradicts (2.2) and thus concludes the proof. ∎

Proposition 2.4 *Under theorem 2.1 and theorem 2.2,* $\left\{\omega \in \Omega \;\middle|\; \mathrm{F}_\infty(\omega) \in F(\operatorname{zer}\nabla F)\right\}$ *is a probability-one set.*

Proof. According to theorem 2.2 and Proposition 2.3, there exists a set $\Lambda' \subset \Omega$ of probability one where, for every $\omega \in \Lambda'$, $\lim_{k\to+\infty} F(\mathbf{x}_k(\omega)) = \mathrm{F}_\infty(\omega)$, $\chi^\infty(\omega) \neq \emptyset$, and $\chi^\infty(\omega) \subset \operatorname{zer}\nabla F$.

Then, for every $\omega \in \Lambda'$, there exist a vector $\boldsymbol{x}_\infty \in \chi^\infty(\omega)$ and a subsequence $\left(\mathbf{x}_{\psi(k)}(\omega)\right)_{k\in\mathbb{N}}$ such that $\mathbf{x}_{\psi(k)}(\omega) \underset{k\to+\infty}{\longrightarrow} \boldsymbol{x}_\infty$, and $F(\mathbf{x}_{\psi(k)}(\omega)) \underset{k\to+\infty}{\longrightarrow} \mathrm{F}_\infty(\omega)$. In addition, since F is continuous, we deduce that $F(\mathbf{x}_{\psi(k)}(\omega)) \underset{k\to+\infty}{\longrightarrow} F(\boldsymbol{x}_\infty)$.

Hence $\mathrm{F}_\infty(\omega) = F(\boldsymbol{x}_\infty) \in F(\chi^\infty)$. Using the fact that $\chi^\infty(\omega) \subset \operatorname{zer}\nabla F$, we then obtain $\mathrm{F}_\infty(\omega) \in F(\operatorname{zer}\nabla F)$, which concludes the proof. □

The next theorem is a first convergence result occurring in the particular case when $\operatorname{zer}\nabla F$ is at most countable, i.e., if it is either countable or finite.

Theorem 2.5 *Assume that theorem 2.1 and theorem 2.2 are verified and that* $(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|)_{k\in\mathbb{N}}$ *converges a.s. to* 0. *If* $\operatorname{zer}\nabla F$ *is at most countable, then sequence* $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *almost surely converges to a point belonging to this set.*

Proof. Since χ^∞ is compact a.s. and $(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|)_{k\in\mathbb{N}}$ converges a.s. to 0, then, according to the Ostrowski's Lemma [41, 26.1], we can deduce that χ^∞ is connex a.s.. Moreover, χ^∞ is a.s. non empty, and at most countable as contained in $\operatorname{zer}\nabla F$. We thus deduce that χ^∞ is a.s. reduced to a singleton $\{\mathbf{x}^*\}$, where $\mathbf{x}^* \in \operatorname{zer}\nabla F$. Since $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is a.s. bounded, it converges a.s. to $\mathbf{x}^*$, hence concluding the proof. □

3 KL theory as a baseline of improvement

When $\operatorname{zer}\nabla F$ is neither countable nor finite, there is a lack of information on the curvature of F to ensure the convergence of the general stochastic scheme to a critical point. The use of non-convex functions has encouraged the development of alternative theoretical tools, in particular in a deterministic context. One of the most famous is the class of KL inequalities [30, 6] enabling interesting gradient properties.

To this aim, we first need to introduce some notation.

For every $\zeta \in (0, +\infty]$, we denote by Φ_ζ the set of concave functions $\varphi\colon [0,\zeta) \mapsto \mathbb{R}_+$ such that $\varphi(0) = 0$, φ is continuous in 0, $\varphi \in C^1((0,\zeta))$, and, for every $s \in (0,\zeta)$, $\varphi'(s) > 0$. The link between Φ_ζ (for $\zeta > 0$) and $\Phi_{+\infty}$ is described through the following proposition.

Proposition 3.1 *Let* $\zeta \in (0, +\infty)$. *Any function* φ *of* Φ_ζ *admits a bounded extension* $\widetilde{\varphi}$ *belonging to* $\Phi_{+\infty}$.

Proof. Let $\varphi \in \Phi_\zeta$. Since φ' is positive, it follows that φ is an increasing function. Then $l_1 = \lim_{s\to\zeta^-} \varphi(s)$ exists and lies in $[0, +\infty]$. Moreover the concavity and differentiability of φ with $\varphi(0) = 0$ ensure that, for every $s \in (0,\eta)$, $\varphi(s) \leqslant s\varphi'(0)$. Passing to the limit thus gives $l_1 \leqslant \eta\varphi'(0)$ and then $l_1 < +\infty$.

Moreover, since φ' is decreasing on $(0,\zeta)$ (due to the concavity of φ) and positive, we conclude that $l_2 = \lim_{s\to\zeta^-} \varphi'(s)$ exists and lies in $[0,+\infty)$.

Finally, we deduce that there exists a function $\widetilde{\varphi}\colon [0,+\infty) \to \mathbb{R}_+$ defined as

$$(\forall s \in [0,+\infty)) \quad \widetilde{\varphi} = \begin{cases} \varphi(s) & \text{if } s \in [0,\zeta), \\ l_1 + l_2\zeta\left(1 - \frac{\zeta}{s}\right) & \text{otherwise.} \end{cases} \tag{3.1}$$

$\widetilde{\varphi}$ belongs to $\Phi_{+\infty}$ and is bounded. □

3.1 Extension of the uniformized KL theorem

In this section we extend classical KL results for differentiable functions. To this aim, we first recall the generic definition of differentiable functions satisfying KL, as introduced in, e.g., [1, 7].

Definition 3.2 *[KL inequality]* A differentiable function $f : \mathbb{R}^N \to \mathbb{R}$ satisfies the *Kurdyka-Lojasiewicz* 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 $\|\nabla f(\boldsymbol{x})\|\varphi'(f(\boldsymbol{x}) - f(\widetilde{\boldsymbol{x}})) \geqslant 1$, for every $\boldsymbol{x} \in V$ satisfying $0 < f(\boldsymbol{x}) - f(\widetilde{\boldsymbol{x}}) < \zeta$.

One major result following the KL inequality is the uniformized KL property introduced in [7, Lemma 6]. This extended KL condition has been used to prove convergence of deterministic algorithms in a non-convex setting, especially when considering block alternating approaches [7, 18].

Theorem 3.3 [Uniformized KL property] *Let C be a compact set in $\mathcal{H}$ and $f\colon \mathcal{H} \to \mathbb{R}$ be a continuous 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*

$$(\forall \overline{\boldsymbol{x}} \in C)(\forall \boldsymbol{x} \in \mathcal{H}) \quad \|\nabla f(\boldsymbol{x})\|\, \varphi'(f(\boldsymbol{x}) - f(\overline{\boldsymbol{x}})) \geqslant 1, \tag{3.2}$$

when $\operatorname{dist}(\boldsymbol{x}, C) < \varepsilon$ *and* $0 < f(\boldsymbol{x}) - f(\overline{\boldsymbol{x}}) < \zeta$.

We propose an extension of theorem 3.3 to apply KL on a finite union of compact sets, for functions that are piecewise constant on this union. This result will be instrumental to prove a.s. convergence of stochastic processes. To this aim, we introduce an additional notation. Let C be a non-empty subset of $\mathcal{H}$ and $f\colon \mathcal{H} \to \mathbb{R}$ be a function that is constant on C. Then we denote by f_C the value taken by f on C, i.e., for every $x \in C$, $f(x) = f_C$.

Theorem 3.4 [Extended uniformized KL property] *Let $C = \bigcup_{i=1}^{I} C_i$ be a union of $I \in \mathbb{N}_*$ non-empty disjoint and compact subsets $(C_i)_{1\leqslant i\leqslant I}$ of $\mathcal{H}$, and $f\colon \mathcal{H} \to \mathbb{R}$ be a differentiable function satisfying the KL property on C. We also suppose that f is constant on every C_i, for $i \in \{1,\dots,I\}$, with respective values $f_{C_1},\dots,f_{C_I}$. Then, there exist $(\varepsilon,\zeta) \in (0,+\infty)^2$ and $\varphi \in \Phi_\zeta$ such that*

$$(\forall i \in \{1,\dots,I\})(\forall \boldsymbol{x} \in \mathcal{H}) \quad \|\nabla f(\boldsymbol{x})\|\varphi'(f(\boldsymbol{x}) - f_{C_i}) \geqslant 1 \tag{3.3}$$

with $\operatorname{dist}(\boldsymbol{x}, C) < \varepsilon$ *and* $0 < f(\boldsymbol{x}) - f_{C_i} < \zeta$.

Proof. Without loss of generality we consider that $f_{C_1} \neq \cdots \neq f_{C_I}$.

We start by applying the uniform KL property to $C_1, \ldots, C_I$. For every $i \in \{1, \ldots, I\}$, there exist $(\varepsilon_i, \zeta_i) \in (0, +\infty)^2$ and $\varphi_i \in \Phi_{\zeta_i}$ such that for every $\boldsymbol{x} \in \mathcal{H}$ verifying $\operatorname{dist}(\boldsymbol{x}, C_i) < \varepsilon_i$ and $0 < f(\boldsymbol{x}) - f_{C_i} < \zeta_i$ we have

$$\|\nabla f(\boldsymbol{x})\| \varphi_i'(f(\boldsymbol{x}) - f_{C_i}) \geqslant 1. \tag{3.4}$$

Let $\widetilde{\zeta} = \upsilon \min_{j \neq i} |f_{C_i} - f_{C_j}|$ with $\upsilon \in]0, 1/2[$. We have $\widetilde{\zeta} > 0$. In addition, using the continuity of f, for every $i \in \{1, \ldots, I\}$, there exists $\widetilde{\varepsilon}_i \in]0, \varepsilon_i[$ such that, for every $\boldsymbol{x}$ satisfying $\operatorname{dist}(\boldsymbol{x}, C_i) < \widetilde{\varepsilon}_i$, we have

$$|f(\boldsymbol{x}) - f_{C_i}| < \widetilde{\zeta}. \tag{3.5}$$

Let $\delta \in]0, 1[$ and $\varepsilon = \delta \min_{1 \leqslant i \leqslant I} \widetilde{\varepsilon}_i$. Then

$$\{\boldsymbol{x} \in \mathcal{H} \mid \operatorname{dist}(\boldsymbol{x}, C) < \varepsilon\} \subset \bigcup_{i=1}^{I} \{\boldsymbol{x} \in \mathcal{H} \mid \operatorname{dist}(\boldsymbol{x}, C_i) < \widetilde{\varepsilon}_i\}. \tag{3.6}$$

We now show that (3.3) is satisfied for ε as defined above, $\zeta = \min(\zeta_1, \ldots \zeta_I, \widetilde{\zeta})$ and $\varphi = \sum_{i=1}^{I} \varphi_i \in \Phi_\zeta$. Let $\boldsymbol{x} \in \mathcal{H}$ and $i \in \{1, \ldots, I\}$ be such that $\operatorname{dist}(\boldsymbol{x}, C) < \varepsilon$ and $0 < f(\boldsymbol{x}) - f_{C_i} < \zeta$. For every $j \in \{1, \ldots, I\} \setminus \{i\}$, since $\zeta \leqslant \widetilde{\zeta}$, using the definition of $\widetilde{\zeta}$, we have $0 < |f(\boldsymbol{x}) - f_{C_i}| \leqslant \upsilon |f_{C_i} - f_{C_j}|$. Since that $\upsilon \in]0, 1/2[$, we obtain

$$\begin{aligned}
|f(\boldsymbol{x}) - f_{C_j}| &= |(f(\boldsymbol{x}) - f_{C_i}) + (f_{C_i} - f_{C_j})| \\
&\geqslant \big| \, |f_{C_i} - f_{C_j}| - |f(\boldsymbol{x}) - f_{C_i}| \, \big| = |f_{C_i} - f_{C_j}| - |f(\boldsymbol{x}) - f_{C_i}| \\
&\geqslant (1 - \upsilon)|f_{C_i} - f_{C_j}| > \upsilon |f_{C_i} - f_{C_j}| \\
&\geqslant \widetilde{\zeta}.
\end{aligned} \tag{3.7}$$

Then $\operatorname{dist}(\boldsymbol{x}, C_j) \geqslant \widetilde{\varepsilon}_j$ (otherwise this contradicts the continuity property of f in (3.5)). So, as $\operatorname{dist}(\boldsymbol{x}, C) < \varepsilon$, (3.6) necessarily implies that $\boldsymbol{x}$ belongs to the union of sets $\bigcup_{j=1}^{I} \{\boldsymbol{x} \in \mathcal{H} \mid \operatorname{dist}(\boldsymbol{x}, C_j) < \widetilde{\varepsilon}_j\}$. This leads to $\operatorname{dist}(\boldsymbol{x}, C_i) < \widetilde{\varepsilon}_i$ and we finally deduce that $\operatorname{dist}(\boldsymbol{x}, C_i) < \varepsilon_i$ (since $\widetilde{\varepsilon}_i \leqslant \varepsilon_i$). In addition, since $\zeta \leqslant \zeta_i$, we have $0 < f(\boldsymbol{x}) - f_{C_i} < \zeta_i$, and we can apply once more the uniform KL property at C_i (i.e., (3.4) is verified). Since $\varphi_1', \ldots, \varphi_I'$ are all positives, we thus deduce that

$$\begin{aligned}
\|\nabla f(\boldsymbol{x})\| \varphi'(f(\boldsymbol{x}) - f_{C_i}) &= \|\nabla f(\boldsymbol{x})\| \sum_{j=1}^{I} \varphi_j'(f(\boldsymbol{x}) - f_{C_i}) \\
&\geqslant \|\nabla f(\boldsymbol{x})\| \varphi_i'(f(\boldsymbol{x}) - f_{C_i}) \geqslant 1.
\end{aligned}$$

This completes the proof. □

3.2 A stochastic KL property

Motivated by the results presented in the previous subsection, we make the following assumption on function F.

Assumption 3.5

(i) F satisfies the KL property on $\operatorname{zer}\nabla F$, the set of critical points of F.

(ii) There exist $I \in \mathbb{N}_*$ non-empty disjoint compact subsets $C_1, \ldots, C_I$ of $\mathcal{H}$, such that $\operatorname{zer}\nabla F = \bigcup_{i=1}^{I} C_i$.

Note that assuming that F satisfies the KL property on $\operatorname{zer}\nabla F$ (i.e., theorem 3.5(i)) is common in non-convex optimization. As emphasized, e.g., in [6], the KL inequality is satisfied for a wide class of functions, and in particular by real analytic, semi-algebraic[1] and log-exp functions. In addition, theorem 3.5(ii) is a condition that will be used for the a.s. convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$, through the existence of a uniformized KL function with respect to all the trajectories of the process.

Let

$$\Pi = \liminf_{k\to+\infty} \{\omega \in \Omega \mid F(\mathbf{x}_k(\omega)) > \mathrm{F}_\infty(\omega)\}. \tag{3.8}$$

We then deduce the following properties.

Proposition 3.6 *Assume that theorem 2.1, theorem 2.2 and theorem 3.5 hold. Then, over the set* Π*, there exist a bounded* $\varphi \in \Phi_{+\infty}$ *and an a.s. finite positive discrete random variable* K *such that, for every* $k > \mathrm{K}$*,* $\|\nabla F(\mathbf{x}_k)\|\varphi'(F(\mathbf{x}_k) - \mathrm{F}_\infty) \geqslant 1$ *a.s.*

Proof.

Let us consider

$$\mathcal{C} = \bigcup_{\omega\in\Theta\cap\Pi} \chi^\infty(\omega) \tag{3.9}$$

where

$$\Theta = \Big\{F(\mathbf{x}_k) \underset{k\to+\infty}{\to} \mathrm{F}_\infty\Big\} \bigcap \Big\{\chi^\infty \subset \operatorname{zer}\nabla F\Big\} \bigcap \Big\{\operatorname{dist}(\mathbf{x}_k, \chi^\infty) \underset{k\to+\infty}{\to} 0\Big\} \bigcap \Big\{\mathrm{F}_\infty(\Omega) \subset F(\operatorname{zer}\nabla F)\Big\} \tag{3.10}$$

is a probability-one set, owing to theorem 2.2, theorem 2.3 and theorem 2.4. The continuity of ∇F (theorem 2.1) also ensures that $\operatorname{zer}\nabla F$ is closed and then $\overline{\mathcal{C}} \subset \operatorname{zer}\nabla F$, where $\overline{\mathcal{C}}$ denotes the closure of set $\mathcal{C}$. Considering $\mathbb{J} = \{i \in \{1, \ldots, I\} \mid \overline{\mathcal{C}} \cap C_i \neq \emptyset\}$, it follows that $\overline{\mathcal{C}} = \overline{\mathcal{C}} \cap \operatorname{zer}\nabla F = \bigcup_{i\in\mathbb{J}} \overline{\mathcal{C}} \cap C_i$, where, for every $i \in \mathbb{J}$, the set $\overline{\mathcal{C}} \cap C_i$ is non-empty, bounded and closed. Moreover, for every $i \in \mathbb{J}$, we have $F(\overline{\mathcal{C}} \cap C_i) = \{F_{C_i}\}$.

According to theorem 3.5, since F satisfies the KL property on $\operatorname{zer}\nabla F$, we can apply theorem 3.4. As a consequence there exist $\varepsilon_{\mathcal{C}} > 0$, $\zeta_{\mathcal{C}} > 0$ and $\varphi_{\mathcal{C}} \in \Phi_{\zeta_{\mathcal{C}}}$, such that, for every $\boldsymbol{x} \in \mathcal{H}$ and $i \in \mathbb{J}$ satisfying $\operatorname{dist}(\boldsymbol{x}, \overline{\mathcal{C}}) < \varepsilon_{\mathcal{C}}$ and $0 < F(\boldsymbol{x}) - F_{C_i} < \zeta_{\mathcal{C}}$,

$$\|\nabla F(\boldsymbol{x})\| \; \varphi'_{\mathcal{C}}(F(\boldsymbol{x}) - F_{C_i}) \geqslant 1. \tag{3.11}$$

[1] A function is semi-algebraic if its graph is a finite union of sets defined by a finite number of polynomial inequalities.

According to theorem 3.1, $\varphi_{\mathcal{C}}$ has a bounded extension $\widetilde{\varphi}_{\mathcal{C}}$ belonging to $\Phi_{+\infty}$. Then, $\widetilde{\varphi}_{\mathcal{C}}$ also satisfies, for any $\boldsymbol{x} \in \mathcal{H}$ and $i \in \mathbb{J}$ such that $\operatorname{dist}(\boldsymbol{x}, \overline{\mathcal{C}}) < \varepsilon_{\mathcal{C}}$ and $0 < F(\boldsymbol{x}) - F_{C_i} < \zeta_{\mathcal{C}}$,

$$\|\nabla F(\boldsymbol{x})\| \, \widetilde{\varphi}'_{\mathcal{C}}(F(\boldsymbol{x}) - F_{C_i}) \geqslant 1. \tag{3.12}$$

Considering this, we now define the positive discrete random variable

$$\mathrm{K} = \min \left\{ l \in \mathbb{N}_* \mid (\forall p \geqslant l) \ \operatorname{dist}(\mathbf{x}_p, \overline{\mathcal{C}}) \leqslant \varepsilon_{\mathcal{C}} \text{ and } 0 < F(\mathbf{x}_p) - \mathrm{F}_\infty < \zeta_{\mathcal{C}} \right\}. \tag{3.13}$$

The latter is finite over $\Theta \cap \Pi$. In addition, according to Assumption 3.5, for every $\omega \in \Theta \cap \Pi$, there exists $i \in \mathbb{J}$ such that $\mathrm{F}_\infty(\omega) = F_{C_i}$, and (3.12) finally leads to

$$(\forall \omega \in \Theta \cap \Pi) \quad (\forall k > \mathrm{K}(w)) \quad \|\nabla F(\mathbf{x}_k(\omega))\| \, \widetilde{\varphi}'_{\mathcal{C}} \left(F(\mathbf{x}_k(\omega)) - \mathrm{F}_\infty(\omega)\right) \geqslant 1. \tag{3.14}$$

The fact that $\mathbb{P}(\Theta \cap \Pi) = P(\Pi)$ concludes the proof. □

In [35, Assumption 3.9, C.2], the authors have highlighted that the construction of a stochastic KL inequality was based on the set $\mathcal{C}$ as introduced in (3.9) without however going further in the reasoning. In addition, in [35, 21], the authors assumed the existence of a uniformized function with respect to all trajectories to enable their convergence results to hold.

4 An almost sure convergence result based on KL theory

In this section, we give conditions that a stochastic process must satisfy to ensure its convergence to a critical point of F. To this aim, we introduce additional notations that will be used in the remainder of this work. We associate the initial probabilistic space $(\Omega, \mathcal{F}, \mathbb{P})$ with a filtration $(\mathcal{F}_k)_{k \in \mathbb{N}}$. Moreover, for a given $k \in \mathbb{N}$ and subject to existence, we denote by $\mathbb{E}(.|\mathcal{F}_k)$ the conditional expectation operator associated with the sub sigma-algebra $\mathcal{F}_k$. We also denote by F^* the minimal value of F.

4.1 Main assumption and summability criterion

The following assumption is the backbone of our convergence theorem, given in section 4.2.

Assumption 4.1 Let $(\mathbf{x}_k)_{k \in \mathbb{N}}$ be a stochastic process.

(i) $(F(\mathbf{x}_k))_{k \in \mathbb{N}}$ and $\left(\|\nabla F(\mathbf{x}_k)\|^2\right)_{k \in \mathbb{N}}$ are two $\mathcal{F}_k$-measurable and integrable processes.

(ii) There exist $(u_k)_{k \in \mathbb{N}}, (v_k)_{k \in \mathbb{N}}, (r_k)_{k \in \mathbb{N}}, (s_k)_{k \in \mathbb{N}}, (t_k)_{k \in \mathbb{N}}$ five non-negative deterministic sequences and $(\mathrm{w}_k)_{k \in \mathbb{N}}$ a non-negative $\mathcal{F}_k$-measurable integrable process, such that

$$\sum_{k=0}^{+\infty} u_k < +\infty, \ \inf_{k \in \mathbb{N}} v_k > 0, \ \sum_{k=0}^{+\infty} r_k < +\infty, \ \sum_{k=0}^{+\infty} t_k < +\infty, \ \sum_{k=0}^{+\infty} \mathrm{w}_k < \ +\infty \ \text{ a.s,} \tag{4.1}$$

and such that, for every $k \in \mathbb{N}$, we have

$$\mathbb{E}[F(\mathbf{x}_{k+1}) - F^* | \mathcal{F}_k] \leqslant (1 + u_k)(F(\mathbf{x}_k) - F^*) - v_k \|\nabla F(\mathbf{x}_k)\|^2 + \mathrm{w}_k \text{ a.s.}, \tag{4.2}$$

and

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\| \mid \mathcal{F}_k] \leqslant r_k \sqrt{F(\mathbf{x}_k) - F^*} + s_k \|\nabla F(\mathbf{x}_k)\| + t_k \text{ a.s.}. \tag{4.3}$$

theorem 4.1(i) is a common assumption, satisfied in particular when the randomness appearing at iteration $k \in \mathbb{N}$ is independent from what happened during the previous iterations. In theorem 4.1(ii), condition (4.2) ensures that $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ is a quasi-supermartingale [44]. Additionally, (4.3) ensures that the norm difference between consecutive iterates is upper bounded, relatively to the norm of the gradient of F. In practice, these two inequalities often imply an underlying assumption on the nature of the stochastic phenomenon at stake. Note that theorem 4.1 is slightly stronger than theorem 2.2, as emphasized by the next proposition.

Proposition 4.2 *If theorem 2.1 and theorem 4.1 are satisfied, then theorem 2.2 holds.*

Proof. Condition (4.2) and the fact that F is bounded from below (since it is coercive) allows us to invoke the Robbins-Siegmund Lemma [44] which directly ensures that theorem 2.2 is verified with $\sum_{k=0}^{+\infty} v_k \|\nabla F(\mathbf{x}_k)\|^2 < +\infty$ a.s.. The fact that $\inf_{k\in\mathbb{N}} v_k > 0$ then directly gives $\lim_{k\to+\infty} \|\nabla F(\mathbf{x}_k)\| = 0$ a.s.. ∎

Before going further, we need to introduce additional notation. Let $(p_k)_{k\in\mathbb{N}}$ be the positive sequence defined by

$$(\forall k \in \mathbb{N}) \quad p_k = \prod_{i=0}^{k} (1 + u_i), \tag{4.4}$$

where $(u_i)_{i\in\mathbb{N}}$ is defined in theorem 4.1(ii). Let, for very $k \in \mathbb{N}_*$,

$$\begin{array}{llll} L_k\colon & \mathcal{H} \times \mathbb{R}^k & \to & \mathbb{R} \\ & (\boldsymbol{x}, w_0, \ldots, w_{k-1}) & \mapsto & (F(\boldsymbol{x}) - F^*) p_{k-1}^{-1} - \sum_{i=0}^{k-1} w_i p_i^{-1} \end{array} \tag{4.5}$$

Defining such a function enables to handle a supermartingale where the conditional decrease is only relative to the gradient, instead of the quasi-supermartingale appearing in condition (4.2). In this context, for every $k \in \mathbb{N}_*$ and $(w_0, \ldots, w_{k-1}) \in \mathbb{R}^k$, $L_k(\cdot, w_0, \ldots, w_{k-1})$ plays the role of a Lyapunov function which can be use to get an equivalent formulation of (4.2).

Lemma 4.3 *Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a stochastic process, and let $(\mathrm{w}_k)_{k\in\mathbb{N}}$ be an integrable $\mathcal{F}_k$-measurable non-negative process satisfying (4.1). Let also $(v_k)_{k\in\mathbb{N}}$ be a non-negative deterministic sequence satisfying (4.1), and let $(p_k)_{k\in\mathbb{N}}$ be the sequence defined in (4.4). Condition (4.2) is equivalent to*

$$\begin{aligned} (\forall k \in \mathbb{N}_*) \quad & \mathbb{E}[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) | \mathcal{F}_k] \\ & \leqslant L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \frac{v_k}{p_k} \|\nabla F(\mathbf{x}_k)\|^2 \text{ a.s.}. \end{aligned} \tag{4.6}$$

Proof. For every $k \geqslant 1$, $\boldsymbol{x} \in \mathcal{H}$ and $(w_0, \ldots, w_{k-1}) \in \mathbb{R}^k$, we have

$$F(\boldsymbol{x}) - F^* = p_{k-1}\Big(L_k(\boldsymbol{x}, w_0, \ldots, w_{k-1}) + \sum_{i=0}^{k-1} w_i p_i^{-1}\Big).$$

Using this equality, for every $k \geqslant 1$, (4.2) is then equivalent to

$$p_k\Big(\mathbb{E}[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k] + \sum_{i=0}^{k} \mathrm{w}_i p_i^{-1}\Big)$$
$$\leqslant (1+u_k)p_{k-1}\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) + \sum_{i=0}^{k-1} \mathrm{w}_i p_i^{-1}\Big) - v_k\|\nabla F(\mathbf{x}_k)\|^2 + \mathrm{w}_k \text{ a.s.}$$

Dividing this inequality by $p_k = p_{k-1}(1+u_k) > 0$, we obtain

$$\mathbb{E}[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k]$$
$$\leqslant L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \Big(\sum_{i=0}^{k} \mathrm{w}_i p_i^{-1} - \sum_{i=0}^{k-1} \mathrm{w}_i p_i^{-1}\Big) - \frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k)\|^2 + \mathrm{w}_k p_k^{-1} \text{ a.s.} \quad (4.7)$$

The equivalence between (4.7) and (4.6) then follows from the fact that $\sum_{i=0}^{k} \mathrm{w}_i p_i^{-1} - \sum_{i=0}^{k-1} \mathrm{w}_i p_i^{-1} = \mathrm{w}_k p_k^{-1}$.

□

Remark 4.4 Let $k \in \mathbb{N}_*$ and $(\mathrm{w}_i)_{0 \leqslant i \leqslant k-1}$ be defined as in theorem 4.1(ii). If the a.s. finite limit F_∞ of $(F(\mathbf{x}_l))_{l \in \mathbb{N}}$ exists, then the a.s. finite limit $\mathrm{L}_{k,\infty}$ of process $(L_k(\mathbf{x}_l, \mathrm{w}_1, \ldots \mathrm{w}_{k-1}))_{l \in \mathbb{N}}$ exists as well, and is given by

$$\mathrm{L}_{k,\infty} = (\mathrm{F}_\infty - F^*)p_k^{-1} - \sum_{i=0}^{k-1} \mathrm{w}_i p_i^{-1}. \quad (4.8)$$

In the case when F_∞ exists, we define, for every $\gamma > 0$, the following set of events:

$$\Xi_\gamma = \Big\{\omega \in \Omega \;\Big|\; (\forall k \geqslant 1)\;\; \mathrm{F}_\infty(\omega) < F(\mathbf{x}_k(\omega))$$
$$\text{and } \big|\mathrm{L}_{k,\infty} - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\big| \leqslant \gamma\frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k(\omega))\|^2\Big\}. \quad (4.9)$$

In the remainder, to obtain summability and convergence results, we will assume that there exists $\gamma \in (0,1)$ such that $\mathbb{P}(\Xi_\gamma) = 1$. Intuitively, this assumption means that F_∞ must a.s. be a lower bound of process $(F(\mathbf{x}_k))_{k \in \mathbb{N}}$. The second inequality in (4.9) indicates that, at iteration $k \in \mathbb{N}$, the error on the identification of $\mathrm{L}_{k,\infty}$ (equivalently, of F_∞) is upper bounded by the squared norm of ∇F.

Proposition 4.5 *Suppose that theorem 2.1, theorem 3.5 and theorem 4.1 hold. Moreover, assume that* $\left(s_k p_k v_k^{-1}\right)_{k\in\mathbb{N}}$ *is a non-increasing sequence and that there exists* $\gamma \in (0,1)$ *such that* $\mathbb{P}(\Xi_\gamma) = 1$. *Let* $\varphi \in \Phi_{+\infty}$ *be a bounded function, and*

$$\Gamma_\varphi := \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k} \Big(\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) - \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\Big). \quad (4.10)$$

Then $\Gamma_\varphi < +\infty$ *a.s..*

Proof.

According to theorem 4.2, theorem 2.2 holds, so, for every $k \in \mathbb{N}_*$, $\mathrm{L}_{k,\infty}$, the a.s limit of process $(L_k(\mathbf{x}_l, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}))_{l\in\mathbb{N}_*}$ exists and is given by (4.8). According to theorem 3.5 and theorem 2.4, F_∞ belongs to a finite set. In addition, according to theorem 4.1, $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ and $(\mathrm{w}_k)_{k\in\mathbb{N}}$ are integrable. Thus, for every $k \in \mathbb{N}_*$, $L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)$ and $\mathrm{L}_{k+1,\infty}$ are integrable, and hence $\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right]$ and $\mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]$ are well-defined.

On the one hand, there exists $\gamma \in (0,1)$ such that $\mathbb{P}(\Xi_\gamma) = 1$. Using the definition of Ξ_γ, we therefore have $\mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k] - \mathrm{L}_{k,\infty} \geqslant -\gamma \frac{v_k}{p_k} \|\nabla F(\mathbf{x}_k)\|^2$ a.s.. Combining this inequality with (4.6), we obtain, for every $k \geqslant 1$,

$$L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty} - \Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big) \geqslant (1-\gamma)\frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k)\|^2 \text{ a.s..} \quad (4.11)$$

Since $1 - \gamma > 0$, we deduce that

$$(\forall k \in \mathbb{N}) \quad L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty} \geqslant \mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k] \text{ a.s..} \quad (4.12)$$

On the other hand, following the definitions (4.5) and (4.8) with $\mathbb{P}(\Xi_\gamma) = 1$, for every $k \in \mathbb{N}_*$, the two random variables in (4.12), $(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty})$ and $(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k])$ are a.s. non-negative. Let $\varphi \in \Phi_{+\infty}$ be bounded. Then, the function φ can be applied to the a.s. positive random variables $(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty})$ and $(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k])$.

Furthermore, φ being an increasing function, (4.12) leads to

$$(\forall k \in \mathbb{N}) \quad \varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) \geqslant \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big) \text{ a.s.,} \quad (4.13)$$

and consequently Γ_φ is a.s. well-defined, non-negative as an infinite sum of a.s. non-negative terms. We can subsequently take the expectation of these quantities, and use the monotone convergence theorem [31] to obtain

$$\mathbb{E}\left[\Gamma_\varphi\right] = \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k} \mathbb{E}\Big[\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) - \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\Big]. \tag{4.14}$$

Since φ is bounded, we can use the linearity of the expectation, leading to

$$\mathbb{E}\left[\Gamma_\varphi\right] = \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k} \mathbb{E}\left[\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big)\right] - \frac{s_k p_k}{v_k} \mathbb{E}\left[\varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\right]. \tag{4.15}$$

Using the inverse conditional Jensen's inequality [31] applied to φ gives

$$\begin{aligned}
&\varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big) \\
&\quad = \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}|\mathcal{F}_k\right]\Big) \\
&\quad \geqslant \mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)|\mathcal{F}_k\right].
\end{aligned} \tag{4.16}$$

Taking the expectation $\mathbb{E}$ of these quantities leads to

$$\begin{aligned}
&\mathbb{E}\left[\varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\right] \\
&\quad \geqslant \mathbb{E}\left[\mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)|\mathcal{F}_k\right]\right] \\
&\quad = \mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)\right].
\end{aligned} \tag{4.17}$$

Combining (4.15) with (4.17), and since $\left(\frac{s_k p_k}{v_k}\right)_{k\in\mathbb{N}}$ is non-increasing, we obtain

$$\begin{aligned}
\mathbb{E}\left[\Gamma_\varphi\right] &\leqslant \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k} \mathbb{E}\left[\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big)\right] \\
&\quad - \frac{s_k p_k}{v_k} \mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)\right] \\
&\leqslant \sum_{k=1}^{+\infty} \frac{s_k p_k}{v_k} \mathbb{E}\left[\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big)\right] \\
&\quad - \frac{s_{k+1} p_{k+1}}{v_{k+1}} \mathbb{E}\left[\varphi\Big(L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) - \mathrm{L}_{k+1,\infty}\Big)\right].
\end{aligned} \tag{4.18}$$

From (4.18), since φ is an increasing and bounded function on $\mathbb{R}_+$, we deduce that

$$0 \leqslant \mathbb{E}\left[\Gamma_\varphi\right] \leqslant \frac{s_1 p_1}{v_1} \mathbb{E}\left[\varphi\Big(L_1(\mathbf{x}_1, \mathrm{w}_0) - \mathrm{L}_{1,\infty}\Big)\right] < +\infty. \tag{4.19}$$

Hence Γ_φ admits a finite expectation and consequently is a.s. finite. □

4.2 Main result

Theorem 4.6 *Under theorem 2.1, theorem 3.5 and theorem 4.1, if* $\left(s_k p_k v_k^{-1}\right)_{k\in\mathbb{N}}$ *is a non-increasing sequence and if there exists* $\gamma \in (0,1)$ *such that* $\mathbb{P}(\Xi_\gamma) = 1$*, then*

(i) $\sum_{k=1}^{+\infty} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| < +\infty$ *a.s.,*

(ii) *the sequence* $(\mathbf{x}_k)_{k\in\mathbb{N}}$ *converges a.s. to a critical point of* F*.*

Proof. According to theorem 4.2, theorem 2.2 holds. In addition, since $\Xi_\gamma \subset \Pi$, where Π is defined in (3.8), and $\mathbb{P}(\Xi_\gamma) = 1$, we have $\mathbb{P}(\Pi) = 1$. Thus theorem 3.6 can be applied, and there exist $\varphi \in \Phi_{+\infty}$ and an a.s. finite non-negative discrete random variable K such that

$$(\forall k > \mathrm{K}) \quad \|\nabla F(\mathbf{x}_k)\|\varphi'(F(\mathbf{x}_k) - \mathrm{F}_\infty) \geqslant 1 \text{ a.s..} \tag{4.20}$$

According to the definitions of L_k and $\mathrm{L}_{k,\infty}$ in (4.5) and (4.8), respectively, we have

$$(\forall k \in \mathbb{N}_*) \quad F(\mathbf{x}_k) - \mathrm{F}_\infty = p_{k-1}\big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\big), \tag{4.21}$$

where $(\mathrm{w}_0, \ldots, \mathrm{w}_{k-1})$ are defined in theorem 4.1(ii). Hence (4.20) rewrites

$$(\forall k > \mathrm{K}) \quad \|\nabla F(\mathbf{x}_k)\|\varphi'\Big(p_{k-1}\big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\big)\Big) \geqslant 1 \text{ a.s..} \tag{4.22}$$

According to (4.4), $(p_k)_{k\in\mathbb{N}}$ is lower-bounded by 1. Then, since φ is concave, hence φ' is decreasing, and we have

$$(\forall k > \mathrm{K}) \quad \|\nabla F(\mathbf{x}_k)\|\varphi'\big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\big) \geqslant 1 \text{ a.s..} \tag{4.23}$$

In addition, owing to the concavity of φ, for every $(a,b) \in \mathbb{R}^2$, we have $\varphi(a) - \varphi(b) \geqslant \varphi'(a)(a-b)$. So, for every $k > \mathrm{K}$, taking $a = L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}$ and $b = \mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]$ in (4.23), we obtain

$$\begin{aligned}
&\varphi\left(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\right) \\
&\qquad - \varphi\left(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\right) \\
&\geqslant \varphi'(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}) \\
&\qquad \times \Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty} \\
&\qquad - \Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big)\Big) \text{ a.s.}
\end{aligned} \tag{4.24}$$

Since Proposition 4.5, holds, we can use inequality (4.11). Hence, injecting successively (4.11) and (4.23) in (4.24) leads, for every $k > \mathrm{K}$, to

$$\begin{aligned}
&\varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) \\
&\qquad - \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k)|\mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty}|\mathcal{F}_k]\Big) \\
&\geqslant (1-\gamma)\frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k)\|^2\varphi'\left(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\right) \\
&\geqslant (1-\gamma)\frac{v_k}{p_k}\|\nabla F(\mathbf{x}_k)\| \text{ a.s.}
\end{aligned} \tag{4.25}$$

Combining (4.3) with (4.25) almost surely gives, for every $k > \mathrm{K}$,

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\| | \mathcal{F}_k] \leqslant (1-\gamma)^{-1} v_k^{-1} s_k p_k \Big(& \varphi\Big(L_k(\mathbf{x}_k, \mathrm{w}_0, \ldots, \mathrm{w}_{k-1}) - \mathrm{L}_{k,\infty}\Big) \\ & - \varphi\Big(\mathbb{E}\left[L_{k+1}(\mathbf{x}_{k+1}, \mathrm{w}_0, \ldots, \mathrm{w}_k) | \mathcal{F}_k\right] - \mathbb{E}[\mathrm{L}_{k+1,\infty} | \mathcal{F}_k]\Big)\Big) \\ & + r_k \sqrt{F(\mathbf{x}_k) - F^*} + t_k. \end{aligned} \tag{4.26}$$

Using Proposition 4.5, the summability of $(r_k)_{k\in\mathbb{N}}$ and $(t_k)_{k\in\mathbb{N}}$, and the fact that $\left(\sqrt{F(\mathbf{x}_k) - F^*}\right)_{k\in\mathbb{N}}$ a.s. converges to $\sqrt{\mathrm{F}_\infty - F^*}$, the right hand side term in (4.26) is a.s. summable. Then, since K is also a.s. finite, we can deduce that $\sum_{k=1}^{+\infty} \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\| | \mathcal{F}_k] < +\infty$ a.s.. Finally, since $(\|\mathbf{x}_{k+1} - \mathbf{x}_k\|)_{k\in\mathbb{N}}$ is a positive sequence, we can apply Levy's sharpening of Borel-Cantelli Lemma [37, Ch.1, Th.21], leading to $\sum_{k=1}^{+\infty} \|\mathbf{x}_{k+1} - \mathbf{x}_k\| < +\infty$ a.s.. It then follows that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ is a.s. a Cauchy sequence. Moreover, according to Proposition 2.3, $(\mathbf{x}_k)_{k\in\mathbb{N}}$ a.s. has a critical point of F as an accumulation point. Hence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ a.s. converges to this critical point. ☐

5 Stochastic gradient schemes

In the previous sections we have presented a general framework to show convergence of the iterates of stochastic schemes for solving (1.1) in a non-convex context, when F satisfies theorem 2.1 and theorem 3.5. In this section we give examples of stochastic gradient algorithms satisfying theorem 4.1, hence with convergence guaranteed by theorem 4.6.

A wide class of stochastic gradient schemes for solving (1.1) are of the form of

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{U}_k \mathbf{f}_k, \tag{5.1}$$

where, for every $k \in \mathbb{N}$, $\mathbf{f}_k \in \mathcal{H}$ usually denotes a stochastic approximation of the gradient of F at $\mathbf{x}_k$, $\alpha_k \in (0, +\infty)$ is the step-size (also called learning rate), and $\mathbf{U}_k \colon \mathcal{H} \to \mathcal{H}$ is a stochastic self-adjoint linear operator usually used as preconditioner (e.g., stochastic Newton or quasi-Newton schemes). Algorithms of the form of (5.1) include variants of the popular SGD [50, 46, 19, 29] as well as schemes incorporating second-order information like stochastic Hessian approximations [52, 36], or more generally preconditioned SGD algorithms [33].

5.1 Conditions on the approximated gradient

Let $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be the canonical filtration, i.e., for every $k \in \mathbb{N}$ we have $\mathcal{F}_k = \sigma(\mathbf{x}_0, \ldots \mathbf{x}_k)$.

For every $k \in \mathbb{N}$, $\mathbf{f}_k$ in (5.1) aims at approximating the true gradient $\nabla F(\mathbf{x}_k)$. In general, it is assumed to be integrable, and that there exist three non-negative deterministic sequences $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$, and $(c_k)_{k\in\mathbb{N}}$ such that

$$\mathbb{E}[\|\mathbf{f}_k\|^2 \mid \mathcal{F}_k] \leqslant a_k(F(\mathbf{x}_k) - F^*) + b_k \|\nabla F(\mathbf{x}_k)\|^2 + c_k \quad \text{a.s..} \tag{5.2}$$

This condition is very generic, and is satisfied by multiple schemes from the literature. For instance, in [51, 28], condition (5.2) is satisfied when $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$, and $(c_k)_{k\in\mathbb{N}}$ are constant, equal to $a \geqslant 0$, $b \geqslant 0$ and $c \geqslant 0$, respectively. When, in particular $a = 0$, and $c = 0$, condition (5.2) is equivalent to the conditional strong growth condition [50]:

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}[\|\mathbf{f}_k\|^2 \mid \mathcal{F}_k] \leqslant b\|\nabla F(\mathbf{x}_k)\|^2 \quad \text{a.s.}. \tag{5.3}$$

A relaxed case, only assuming $a = 0$, has also been investigated in [5, 10]. Finally, condition (5.2) also extents those of [17], in the differentiable case, where, for every $k \in \mathbb{N}$, $a_k = 0$, but without assuming that $(b_k)_{k\in\mathbb{N}}$ and $(c_k)_{k\in\mathbb{N}}$ are constant.

5.2 Conditions on the step-size

In (5.1), $(\alpha_k)_{k\in\mathbb{N}}$ is a positive sequence playing the role of step-sizes. When they are chosen to be constant, it has been shown in [50] that theorem 4.1 is verified if the unknown gradients $(\mathbf{f}_k)_{k\in\mathbb{N}}$ satisfy the strong growth condition (5.3). If this condition is not satisfied, in particular for non-constant step-sizes, but a more general condition holds (e.g., (5.2)), $(\alpha_k)_{k\in\mathbb{N}}$ must be decreasing to ensure the robustness of the perturbation on the induced variance [10].

5.3 Conditions on the preconditioning operator

In (5.1), $(\mathbf{U}_k)_{k\in\mathbb{N}}$ is a sequence of stochastic preconditioning self-adjoint linear operators, with an a.s. uniformly bounded spectrum, i.e., there exist two positive sequences $(\mu_k)_{k\in\mathbb{N}}$ and $(\nu_k)_{k\in\mathbb{N}}$ such that

$$(\forall k \in \mathbb{N})(\forall \boldsymbol{x} \in \mathcal{H}) \quad \mu_k\|\boldsymbol{x}\|^2 \leqslant \langle \boldsymbol{x} \mid \mathbf{U}_k\boldsymbol{x}\rangle \leqslant \nu_k\|\boldsymbol{x}\|^2 \quad \text{a.s.} \tag{5.4}$$

The most famous class of preconditioning matrices for gradient descent schemes is the Quasi-Newton one. In this context, for every $k \in \mathbb{N}$, $\mathbf{U}_k$ is chosen to be an approximation of the Hessian of F evaluated at the current iterate $\mathbf{x}_k$ (assuming that F is twice-differentiable). The celebrated quasi-Newton BFGS method is one of them. The convergence behavior of $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$ has been studied in [52] in a non-convex and stochastic setting. However, to the best of our knowledge, the a.s. convergence of the sequence of iterates $(\mathbf{x}_k)_{k\in\mathbb{N}}$ has not been studied yet in the literature.

5.4 A general framework for convergence of SGD algorithms in a non-convex context

In this section, we link the noisy gradient condition (5.2) to the descent conditions of theorem 4.1. In particular, we show that under mild conditions on the parameters involved in (5.1), (5.2) and (5.4), theorem 4.6 holds.

Proposition 5.1 *Let F be a β-Lipschitz differentiable function, for some $\beta > 0$, and let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a process verifying (5.1), (5.2), and (5.4) with respect to $(\mathcal{F}_k)_{k\in\mathbb{N}}$. Assume that*

$$\inf_{k\in\mathbb{N}} \mu_k > 0, \quad \sum_{k=0}^{+\infty} \alpha_k \nu_k a_k^{1/2} < +\infty, \quad \sum_{k=0}^{+\infty} \alpha_k \nu_k c_k^{1/2} < +\infty, \tag{5.5}$$

and that one of the two following statement holds:

(i) *The sequence $(\mathbf{f}_k)_{k\in\mathbb{N}}$ is conditionally unbiased, i.e., for every $k \in \mathbb{N}$, $\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] = \nabla F(\mathbf{x}_k)$ a.s.; and*

$$\inf_{k\in\mathbb{N}} \alpha_k \left(\mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\right) > 0. \tag{5.6}$$

(ii) *The sequence $(\mathbf{f}_k)_{k\in\mathbb{N}}$ verifies*

$$\sum_{k=0}^{+\infty} \frac{\alpha_k \nu_k^2}{\mu_k} \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|^2 < +\infty \quad \text{a.s.} \tag{5.7}$$

and there exists $\varrho \in (0,1)$ such that

$$\inf_{k\in\mathbb{N}} \alpha_k \left(\varrho\mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\right) > 0. \tag{5.8}$$

If the sequence $(\mathbf{U}_k)_{k\in\mathbb{N}}$ is chosen to be adapted to $(\mathcal{F}_k)_{k\in\mathbb{N}}$ (i.e., for every $k \in \mathbb{N}$, $\mathbf{U}_k$ is $\mathcal{F}_k$-measurable), then $(\mathbf{x}_k)_{k\in\mathbb{N}}$ satisfies theorem 4.1.

Proof. Following a similar proof as for [14, Lemma 2], it can be shown that the conditional expectations of all manipulated random variable are well-defined and satisfy theorem 4.1(i). To show that theorem 4.1(ii) is satisfied, we need to show that both conditions (4.2) and (4.3) hold. To this aim, we will first derive generic inequalities. F being β-Lipschitz differentiable, we can apply the descent lemma [4], and we obtain

$$(\forall k \in \mathbb{N}) \quad F(\mathbf{x}_{k+1}) \leqslant F(\mathbf{x}_k) - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbf{f}_k \rangle + \frac{\alpha_k^2 \beta}{2} \|\mathbf{U}_k \mathbf{f}_k\|^2. \tag{5.9}$$

Subtracting F^* in (5.9), and passing to the conditional expectation lead to

$$\begin{aligned}(\forall k \in \mathbb{N}) \quad & \mathbb{E}[F(\mathbf{x}_{k+1}) - F^* \,|\mathcal{F}_k] \\ & \leqslant \mathbb{E}\Big[F(\mathbf{x}_k) - F^* - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbf{f}_k \rangle + \frac{\alpha_k^2 \beta}{2} \|\mathbf{U}_k \mathbf{f}_k\|^2 \;\Big|\; \mathcal{F}_k\Big] \quad \text{a.s.} \\ & = F(\mathbf{x}_k) - F^* - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] \rangle + \frac{\alpha_k^2 \beta}{2} \mathbb{E}\big[\|\mathbf{U}_k \mathbf{f}_k\|^2 | \mathcal{F}_k\big] \quad \text{a.s.},\end{aligned} \tag{5.10}$$

where the equality is obtained using the linearity of the conditional expectation, and the $\mathcal{F}_k$-measurability of $F(\mathbf{x}_k)$, $\nabla F(\mathbf{x}_k)$ and $\mathbf{U}_k$, for all $k \in \mathbb{N}$. Moreover, from (5.2) and (5.4), it follows that, for every $k \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{U}_k \mathbf{f}_k\|^2 | \mathcal{F}_k\big] & \leqslant \nu_k^2 \mathbb{E}\big[\|\mathbf{f}_k\|^2 | \mathcal{F}_k\big] \\ & \leqslant \nu_k^2 a_k \big(F(\mathbf{x}_k) - F^*\big) + \nu_k^2 b_k \|\nabla F(\mathbf{x}_k)\|^2 + \nu_k^2 c_k \quad \text{a.s.}.\end{aligned} \tag{5.11}$$

Injecting (5.11) in (5.10) gives, for every $k \in \mathbb{N}$,

$$\mathbb{E}[F(\mathbf{x}_{k+1}) - F^* | \mathcal{F}_k] \leqslant \left(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\right) (F(\mathbf{x}_k) - F^*) \\ - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbb{E}[\mathbf{f}_k | \mathcal{F}_k] \rangle \frac{\alpha_k^2 \beta \nu_k^2 b_k}{2} \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2}. \quad (5.12)$$

We will now first show that condition (4.3) is verified in both cases (i) and (ii). We will then show that (4.2) is also satisfied, using separately case (i) or case (ii).

According to the scheme in definition (5.1), we have

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{F}_k] = \alpha_k^2 \mathbb{E}[\|\mathbf{U}_k \mathbf{f}_k\|^2 | \mathcal{F}_k]. \quad (5.13)$$

Hence, injecting (5.13) in (5.11), we obtain for all $k \in \mathbb{N}$

$$\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{F}_k] \leqslant \alpha_k^2 \nu_k^2 a_k (F(\mathbf{x}_k) - F^*) + \alpha_k^2 \nu_k^2 b_k \|\nabla F(\mathbf{x}_k)\|^2 + \alpha_k^2 \nu_k^2 c_k \text{ a.s.}. \quad (5.14)$$

Using the conditional version of Jensen's inequality [31] leads to

$$\begin{aligned} (\forall k \in \mathbb{N}) \quad & \mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\| | \mathcal{F}_k] \\ & \leqslant \sqrt{\mathbb{E}[\|\mathbf{x}_{k+1} - \mathbf{x}_k\|^2 | \mathcal{F}_k]} \\ & \leqslant \sqrt{\alpha_k^2 \nu_k^2 a_k (F(\mathbf{x}_k) - F^*) + \alpha_k^2 \nu_k^2 b_k \|\nabla F(\mathbf{x}_k)\|^2 + \alpha_k^2 \nu_k^2 c_k} \\ & \leqslant \alpha_k \nu_k a_k^{1/2} \sqrt{F(\mathbf{x}_k) - F^*} + \alpha_k \nu_k b_k^{1/2} \|\nabla F(\mathbf{x}_k)\| + \alpha_k \nu_k c_k^{1/2} \text{ a.s.} \end{aligned} \quad (5.15)$$

Let, for every $k \in \mathbb{N}$, $r_k := \alpha_k \nu_k a_k^{1/2}$, $s_k = \alpha_k \nu_k b_k^{1/2}$, and $t_k = \alpha_k \nu_k c_k^{1/2}$. Sequences $(r_k)_{k \in \mathbb{N}}, (s_k)_{k \in \mathbb{N}}, (t_k)_{k \in \mathbb{N}}$ are non-negative, and the summabilities of $(r_k)_{k \in \mathbb{N}}$ and $(t_k)_{k \in \mathbb{N}}$ are due to (5.5). Hence (4.3) is satisfied.

We now distinguish case (i) from case (ii) to show that condition (4.3) is also satisfied.

Case (i): Since $(\mathbf{f}_k)_{k \in \mathbb{N}}$ is conditionally unbiased in (5.12), we obtain, for every $k \in \mathbb{N}$,

$$\begin{aligned} & \mathbb{E}[F(\mathbf{x}_{k+1}) - F^* | \mathcal{F}_k] \\ & \leqslant \left(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\right) (F(\mathbf{x}_k) - F^*) \\ & \quad - \alpha_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \nabla F(\mathbf{x}_k) \rangle + \frac{\alpha_k^2 \beta \nu_k^2 b_k}{2} \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2} \\ & \leqslant \left(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\right) (F(\mathbf{x}_k) - F^*) \\ & \quad - \alpha_k \mu_k \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 b_k}{2} \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2} \\ & = \left(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\right) (F(\mathbf{x}_k) - F^*) \\ & \quad - \alpha_k \left(\mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\right) \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2} \text{ a.s.} \end{aligned} \quad (5.16)$$

where majoration of $\langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \nabla F(\mathbf{x}_k)\rangle$ directly comes from (5.4). Majoration in (5.16) has the same structure as in (4.2), where, for every $k \in \mathbb{N}$,

$$u_k := \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}, \quad v_k := \alpha_k \left(\mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\right), \quad \mathrm{w}_k := \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2}. \tag{5.17}$$

Moreover, thanks to (5.5) and (5.6), we deduce that $(u_k)_{k\in\mathbb{N}}, (v_k)_{k\in\mathbb{N}}, (\mathrm{w}_k)_{k\in\mathbb{N}}$ are non-negative sequences, and that $\sum_k u_k < +\infty$, $\inf_k v_k > 0$, and $\sum_k \mathrm{w}_k < +\infty$. Hence condition (4.2) is satisfied under condition (i).

Case (ii): Using (5.4) and the Cauchy-Schwarz inequality, we have, for every $k \in \mathbb{N}$,

$$\begin{aligned}
&\langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbb{E}[\mathbf{f}_k|\mathcal{F}_k]\rangle \\
&= \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \nabla F(\mathbf{x}_k)\rangle + \Big\langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \Big(\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\Big)\Big\rangle \\
&\geqslant \mu_k \|\nabla F(\mathbf{x}_k)\|^2 - \big\|\nabla F(\mathbf{x}_k)\big\| \, \Big\|\mathbf{U}_k \Big(\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\Big)\Big\| \\
&\geqslant \mu_k \|\nabla F(\mathbf{x}_k)\|^2 - \nu_k \|\nabla F(\mathbf{x}_k)\| \, \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|.
\end{aligned} \tag{5.18}$$

We consider inequality $ab \leqslant a^2 c + b^2/(4c)$, that holds for any $a \geqslant 0$, $b \geqslant 0$ and $c > 0$. Let $k \in \mathbb{N}$, and take $a = \|\nabla F(\mathbf{x}_k)\|$, $b = \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|$, $c = (1-\varrho)\mu_k \nu_k^{-1}$. Then, for every $k \in \mathbb{N}$, we have

$$\begin{aligned}
&\|\nabla F(\mathbf{x}_k)\| \, \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\| \\
&\qquad \leqslant \frac{(1-\varrho)\mu_k}{\nu_k} \|\nabla F(\mathbf{x}_k)\|^2 + \frac{\nu_k}{4(1-\varrho)\mu_k} \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|^2.
\end{aligned} \tag{5.19}$$

Combining this inequality with (5.18) leads to

$$\begin{aligned}
&(\forall k \in \mathbb{N}) \quad \langle \nabla F(\mathbf{x}_k) \mid \mathbf{U}_k \mathbb{E}[\mathbf{f}_k|\mathcal{F}_k]\rangle \\
&\qquad \geqslant \varrho \mu_k \|\nabla F(\mathbf{x}_k)\|^2 - \frac{\nu_k^2}{4(1-\varrho)\mu_k} \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|^2.
\end{aligned} \tag{5.20}$$

Injecting (5.20) in (5.12), we obtain

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{x}_{k+1}) - F^*|\mathcal{F}_k] \\
&\leqslant \left(1 + \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}\right)(F(\mathbf{x}_k) - F^*) - \alpha_k \left(\varrho \mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\right) \|\nabla F(\mathbf{x}_k)\|^2 \\
&\qquad + \frac{\alpha_k \nu_k^2}{4(1-\varrho)\mu_k} \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\|^2 + \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2} \quad \text{a.s..}
\end{aligned} \tag{5.21}$$

Let, for every $k \in \mathbb{N}$, $u_k = \frac{\alpha_k^2 \beta \nu_k^2 a_k}{2}$, $v_k = \alpha_k \left(\varrho \mu_k - \frac{\alpha_k \beta \nu_k^2 b_k}{2}\right)$, and $\mathrm{w}_k = \frac{\alpha_k^2 \beta \nu_k^2 c_k}{2} + \frac{\alpha_k \nu_k^2}{4\mu_k(1-\varrho)} \|\mathbb{E}(\mathbf{f}_k|\mathcal{F}_k) - \nabla F(\mathbf{x}_k)\|^2$. According to (5.5) (5.7) and (5.8), $(u_k)_{k\in\mathbb{N}}$, $(v_k)_{k\in\mathbb{N}}$, $(\mathrm{w}_k)_{k\in\mathbb{N}}$ are non-negative sequences, such that $\sum_k u_k < +\infty$, $\inf_k v_k > 0$, and $\sum_k \mathrm{w}_k < +\infty$ a.s. Hence (4.2) is satisfied under condition (ii).

□

5.5 Application to some state-of-the-art algorithms

In this section we review a few state-of-the-art SGD algorithms of the form of (5.1), those convergence in a non-convex setting is ensured by Proposition (5.1).

5.5.1 SGD with constant stepsize from [50]

The algorithm proposed in [50] consists of an SGD scheme of the form of (5.1), with constant stepsize. Precisely, for every $k \in \mathbb{N}$, $\alpha_k = \alpha$, with $\alpha > 0$, and without preconditioning matrix, i.e.,

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \alpha\, \mathbf{f}_k. \tag{5.22}$$

The a.s. convergence of the iterates (5.22) can be shown using Proposition 5.1. First, the authors in [50] show that scheme (5.22) verifies a descent condition in a non-necessary convex setting assuming that (5.2) holds with, for every $k \in \mathbb{N}$, $a_k = c_k = 0$ and $b_k = b$, for some $b > 0$. This ensures the convergence of $(F(\mathbf{x}_k))_{k\in\mathbb{N}}$. Second, condition (5.4) is directly satisfied with, for every $k \in \mathbb{N}$, $\mu_k = \nu_k = 1$. Finally, to the extent that the gradient approximations $(\mathbf{f}_k)_{k\in\mathbb{N}}$ are unbiased, condition (i) of Proposition 5.1 holds for $\alpha \in \big(0, 2/(Lb)\big)$. Hence, according to Proposition 5.1, Assumption 4.1 is satisfied and the a.s. convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a critical point is ensured by Theorem 4.6.

5.5.2 SGD from [17]

In [17], the authors provide a stochastic version of the forward-backward algorithm for minimizing the sum of a Lipschitz-differentiable function, with constant $\beta > 0$, and a non-necessarily smooth function, both being convex. Assuming that the non-smooth function is zero, then this algorithm boils down to an SGD algorithm of the form of (5.1), without preconditioning matrices, i.e.,

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \lambda_k \gamma_k \mathbf{f}_k, \tag{5.23}$$

where, for every $k \in \mathbb{N}$, $\lambda_k \in]0, 1]$ and $\gamma_k \in (0, 2/\beta)$. By contrast with the usual SGD algorithms, in [17] the sequence of approximated gradients $(\mathbf{f}_k)_{k\in\mathbb{N}}$ is not supposed to be unbiased, but to verify the following relaxed condition :

$$\sum_{k=0}^{+\infty} \sqrt{\lambda_k} \|\mathbb{E}[\mathbf{f}_k|\mathcal{F}_k] - \nabla F(\mathbf{x}_k)\| < +\infty \quad \text{a.s..} \tag{5.24}$$

The a.s. convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a minimizer of F is ensured by [17, Thm. 2.1], under few technical assumptions (including F convex).

Under the same assumptions, using Proposition 5.1 and Theorem 4.6, we can show that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ a.s. converges to a critical point of F, without needing F to be convex. Indeed, first, [17, Thm. 2.1 - cond. (c)] ensures that (5.2) is satisfied, where $a_k \equiv 0$, $(b_k)_{k\in\mathbb{N}}$ is a bounded sequence, and

$(c_k)_{k\in\mathbb{N}}$ satisfies $\sum_{k=0}^{+\infty}\lambda_k c_k < +\infty$. Second, condition (5.4) is satisfied with, for every $k \in \mathbb{N}$, $\mu_k = \nu_k = 1$. Finally, if $\inf_{k\in\mathbb{N}}\lambda_k > 0$, as in [17, Thm. 2.1 - cond. (f)], then (5.24) implies condition (5.7), hence condition (ii) in Proposition 5.1 holds. Consequently, according to Proposition 5.1, Assumption 4.1 is satisfied and the a.s. convergence of $(\mathbf{x}_k)_{k\in\mathbb{N}}$ to a critical point is ensured by Theorem 4.6. This result is novel, up to our knowledge. Actually, almost-sure convergence based on sequence $(\mathbf{x}_k)_{k\in\mathbb{N}}$ for stochastic gradient schemes remains scarcely studied in the non-convex case. Recent results in [23, 14] proposed an alternative proof technique without KL inequality by constraining the topology of the stationary points of F.

5.5.3 Stochastic quasi-Newton with constant stepsize from [36]

In the case where F is twice differentiable, the author of [36] proposed an alternative version of the usual stochastic Quasi-Newton algorithm [52] of the form

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \eta \mathbf{U}_k \mathbf{f}_k, \tag{5.25}$$

where $\eta > 0$ is a constant stepsize. In [36], the authors show that any process generated by scheme (5.25) satisfies Assumption 2.1. Under the same assumptions, we can show that $(\mathbf{x}_k)_{k\in\mathbb{N}}$ a.s. converges to a critical point of F, using Proposition 5.1 and Theorem 4.6. First, the authors of [52] assume that the gradients $(\mathbf{f}_k)_{k\in\mathbb{N}}$ are unbiased and satisfy (5.2), with, for every $k \in \mathbb{N}$, $a_k = c_k = 0$ and $b_k = b$, for some $b > 0$. Second, condition (5.4) is satisfied with, for every $k \in \mathbb{N}$, $\mu_k = \mu > 0$ and $\nu_k = \nu > 0$, since the preconditioning matrix $\mathbf{U}_k$ is built to approximate the Hessian of F at $\mathbf{x}_k$ and to have a uniformly bounded spectrum. Finally, (i) in Proposition 5.1 is verified when $\alpha \in \big(0, 2\mu/(\beta b \nu^2)\big)$.

6 Stochastic proximal-gradient scheme examples

In this section, we will give examples of stochastic proximal algorithms from the literature satisfying theorem 4.1, hence benefiting from the convergence guarantees given in theorem 4.6. Similarly to section 5, we will present a general framework for investigating convergence of stochastic proximal schemes in a non-convex setting.

We consider the case where problem (1.1) can be written as

$$\underset{\boldsymbol{x}\in\mathcal{H}}{\text{minimize}}\, F(\boldsymbol{x}) = G(\boldsymbol{x}) + H(\boldsymbol{x}), \tag{6.1}$$

where $F\colon \mathcal{H} \to \mathbb{R}$ verifies theorem 2.1 and theorem 3.5, and is expressed as the sum of two functions $G\colon \mathcal{H} \to \mathbb{R}$ and $H\colon \mathcal{H} \to]-\infty, +\infty]$, where G is assumed to be differentiable, and H is assumed to be convex, proper and lower semi-continuous. In this context, a suitable scheme to solve (6.1) is to consider a proximal-gradient algorithm (also known as forward-backward, or ISTA). It alternates, at each iteration, between a gradient-descent step on the differentiable function and a proximal step on the non-smooth function. The proximity operator of a proper,

lower semi-continuous, convex function H at a point $\boldsymbol{x} \in \mathcal{H}$ is denoted by $\operatorname{prox}_H(\boldsymbol{x})$, and is defined as the unique minimizer of $H + \frac{1}{2\gamma_k}\|\boldsymbol{x} - \cdot\|^2$ on $\mathcal{H}$ [39, 3].

A stochastic proximal-gradient algorithm can be expressed as

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \left(\mathbf{P}_k(\mathbf{x}_k - \gamma_k \mathbf{g}_k) - \mathbf{x}_k\right), \tag{6.2}$$

where for every $k \in \mathbb{N}$, $\mathbf{g}_k \in \mathcal{H}$ is a stochastic approximation of the gradient of G at $\mathbf{x}_k$, γ_k and λ_k are two positive stepsizes, and $\mathbf{P}_k \colon \mathcal{H} \to \mathcal{H}$ is a stochastic approximation of the proximity operator of H computed at the output of the gradient descent step $(\mathbf{x}_k - \gamma_k \mathbf{g}_k)$.

The stochastic approximations of the proximity operators can be useful in the context when the computation of the proximity operator itself is too demanding, e.g., due to the structure of H (for instance where H is a sum of a very large number of components). A typical example of operators $(\mathbf{P}_k)_{k\in\mathbb{N}}$ is the one encountered in federated algorithms [48]. In this context, H reads as a convex combination $H = \sum_{i=1}^{I} \omega_i H_i$, where $I > 1$, $(\omega_i)_{1\leqslant i\leqslant I} \in (0,1)^I$ with $\sum_{i=1}^{I} \omega_i = 1$, and, for every $i \in \{1,\ldots,I\}$, $H_i \colon \mathcal{H} \to]-\infty, +\infty]$. Then, the most common choice consists in adopting, at every iteration $k \in \mathbb{N}$, $\mathbf{P}_k = \sum_{i\in\mathbb{I}_k} \omega_i \operatorname{prox}_{\gamma_k H_i}$, where $\mathbb{I}_k$ is a random subset of $\{1,\ldots,I\}$. Another popular choice of approximation relies on the (deterministic) notion of ϵ-subdifferentiability [49].

Scheme (6.2) can be seen as a stochastic gradient scheme by rewriting it as

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \lambda_k \mathbf{f}_k, \tag{6.3}$$

where, at each iteration $k \in \mathbb{N}$, $\mathbf{f}_k = \gamma_k^{-1}\left(\mathbf{x}_k - \mathbf{P}_k(\mathbf{x}_k - \gamma_k \mathbf{g}_k)\right)$. Thus, $\mathbf{f}_k$ can be interpreted as an estimation of the gradient of the whole function $\nabla F(\mathbf{x}_k)$. In the literature, this viewpoint is adopted to prove the convergence of some versions of the standard stochastic proximal algorithm (i.e., when, for every $k \in \mathbb{N}$, $\lambda_k = 1$) [38].

6.1 Conditions on the approximations

Let $(\mathcal{F}_k)_{k\in\mathbb{N}}$ be the canonical filtration as defined in section 5.1. Let $\mathbf{y}$ be an integrable random variable from Ω to $\mathcal{H}$. We assume that approximation sequences $(\mathbf{g}_k)_{k\in\mathbb{N}}$ and $\big(\mathbf{P}_k(\mathbf{y})\big)_{k\in\mathbb{N}}$ are conditionally unbiased, i.e.,

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}[\mathbf{g}_k|\mathcal{F}_k] = \nabla G(\mathbf{x}_k) \quad \text{and} \quad \mathbb{E}[\mathbf{P}_k(\mathbf{y}) - \operatorname{prox}_{\gamma_k H}(\mathbf{y})|\mathcal{F}_k] = 0, \tag{6.4}$$

and that there exist two deterministic sequences $(d_k)_{k\in\mathbb{N}}$ and $(e_k)_{k\in\mathbb{N}}$ such that

$$(\forall k \in \mathbb{N}) \quad \mathbb{E}\big[\|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|^2|\mathcal{F}_k\big] \leqslant d_k\|\nabla F(\mathbf{x}_k)\|^2 + e_k, \tag{6.5}$$

$$\mathbb{E}\big[\|\mathbf{P}_k(\mathbf{y}) - \operatorname{prox}_{\gamma_k H}(\mathbf{y})\|^2|\mathcal{F}_k\big] \leqslant d_k\|\nabla F(\mathbf{x}_k)\|^2 + e_k. \tag{6.6}$$

6.2 A general framework for convergence of differentiable proximal-gradient algorithms in a non-convex context

In this section we introduce a result to enable the convergence study of proximal-gradient based methods, for non-convex minimization problems. In order to use the results developed in section 3 and section 4, we will consider differentiable functions G and H.

Proposition 6.1 *Let F be of the form of (6.1), where G and H are Lipschitz-differentiable functions, with respective constants $\beta_G > 0$ and $\beta_H > 0$, and H is convex. Let $(\mathbf{x}_k)_{k\in\mathbb{N}}$ be a process generated by (6.2), satisfying conditions (6.4), (6.5), and (6.6), with respect to $(\mathcal{F}_k)_{k\in\mathbb{N}}$. Assume that*

$$\inf_{k\in\mathbb{N}} \lambda_k\gamma_k > 0, \quad \sup_{k\in\mathbb{N}} \lambda_k \leqslant 1, \quad \sup_{k\in\mathbb{N}} \gamma_k < 1/\beta_H, \quad \sum_{k=0}^{+\infty} e_k^{1/2} < +\infty. \tag{6.7}$$

Let $\beta = \beta_G + \beta_H$, and $(\rho_k)_{k\in\mathbb{N}}$, $(\sigma_k)_{k\in\mathbb{N}}$ be two sequences defined as

$$(\forall k \in \mathbb{N}) \quad \rho_k = (1-\beta_H\gamma_k)^{-1} \quad \text{and} \quad \sigma_k = \frac{\gamma_k\rho_k}{2}\left((\sqrt{2}+1)\beta_H + 4\beta\lambda_k\rho_k\right), \tag{6.8}$$

with

$$\sup_{k\in\mathbb{N}} \ \sigma_k + d_k\left(\sigma_k + \lambda_k\gamma_k^{-1}\beta\right) < 1. \tag{6.9}$$

Then $(\mathbf{x}_k)_{k\in\mathbb{N}}$ verifies theorem 4.1.

Proof. To prove that theorem 4.1 holds, we will use the equivalent form (6.3) of the proximal-gradient scheme (6.2). Hence the proof will be very similar to the one of theorem 5.1. Precisely, we consider scheme (5.1), where, for every $k \in \mathbb{N}$, $\alpha_k = \gamma_k\lambda_k$ and $\mathbf{U}_k = \boldsymbol{I}$, and the sequence $(\mathbf{f}_k)_{k\in\mathbb{N}}$ is an approximation of $(\nabla F(\mathbf{x}_k))_{k\in\mathbb{N}}$ given by

$$(\forall k \in \mathbb{N}) \quad \mathbf{f}_k = \mathbf{d}_k + \Delta_k, \tag{6.10}$$

$$\mathbf{d}_k = \gamma_k^{-1}\left(\mathbf{x}_k - \operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k\mathbf{g}_k)\right), \tag{6.11}$$

$$\Delta_k = \gamma_k^{-1}\left(\operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k\mathbf{g}_k) - \mathbf{P}_k(\mathbf{x}_k - \gamma_k\mathbf{g}_k)\right). \tag{6.12}$$

The first step of the proof consists in giving a majoration of the approximated gradient sequence $(\mathbf{f}_k)_{k\in\mathbb{N}}$ as a function of the true gradient and the error terms involved in the assumptions.

Since H is differentiable, according to the Fermat's rule [3] we have

$$(\mathbf{x}_k - \gamma_k\mathbf{g}_k) - \operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k\mathbf{g}_k) = \gamma_k\nabla H\left(\operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k\mathbf{g}_k)\right). \tag{6.13}$$

Combining (6.11) and (6.13), it follows that

$$\mathbf{d}_k = \mathbf{g}_k + \nabla H\left(\operatorname{prox}_{\gamma_k H}(\mathbf{x}_k - \gamma_k\mathbf{g}_k)\right) = \mathbf{g}_k + \nabla H(\mathbf{x}_k - \gamma_k\mathbf{d}_k). \tag{6.14}$$

Since $\nabla F = \nabla G + \nabla H$, using the triangular inequality, and the fact that H has a β_H-Lipschitzian gradient

$$\begin{aligned}\|\mathbf{d}_k\| &= \|\nabla F(\mathbf{x}_k) + (\mathbf{g}_k - \nabla G(\mathbf{x}_k)) + \nabla H(\mathbf{x}_k - \gamma_k \mathbf{d}_k) - \nabla H(\mathbf{x}_k)\| \\ &\leqslant \|\nabla F(\mathbf{x}_k)\| + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\| + \beta_H \gamma_k \|\mathbf{d}_k\|. \end{aligned} \tag{6.15}$$

According to the definition of ρ_k in (6.8), this leads to

$$\|\mathbf{d}_k\| \leqslant \rho_k \left(\|\nabla F(\mathbf{x}_k)\| + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|\right). \tag{6.16}$$

Combining (6.10) and (6.16) gives

$$\|\mathbf{f}_k\| \leqslant \rho_k \left(\|\nabla F(\mathbf{x}_k)\| + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|\right) + \|\Delta_k\|. \tag{6.17}$$

Moreover, passing to the square, we also deduce that

$$\|\mathbf{f}_k\|^2 \leqslant 4\rho_k^2 \left(\|\nabla F(\mathbf{x}_k)\|^2 + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|^2\right) + 2\|\Delta_k\|^2. \tag{6.18}$$

The second step of the proof aims at establishing stochastic descent properties for the scheme at stake, to show that conditions (4.2) and (4.3) hold.

First, we apply the descent Lemma [5] to G, using the fact that $\mathbf{x}_{k+1} - \mathbf{x}_k = -\gamma_k \mathbf{f}_k$,

$$G(\mathbf{x}_{k+1}) \leqslant G(\mathbf{x}_k) - \gamma_k \lambda_k \langle \nabla G(\mathbf{x}_k) \mid \mathbf{f}_k \rangle + \frac{\beta_G \gamma_k^2 \lambda_k^2}{2} \|\mathbf{f}_k\|^2, \tag{6.19}$$

Then, adding $H(\mathbf{x}_{k+1})$ on both sides of (6.19) gives

$$F(\mathbf{x}_{k+1}) \leqslant G(\mathbf{x}_k) - \gamma_k \lambda_k \langle \nabla G(\mathbf{x}_k) \mid \mathbf{f}_k \rangle + H(\mathbf{x}_{k+1}) + \frac{\beta_G \gamma_k^2 \lambda_k^2}{2} \|\mathbf{f}_k\|^2. \tag{6.20}$$

Since H is convex and differentiable, using again $\mathbf{x}_{k+1} - \mathbf{x}_k = -\gamma_k \mathbf{f}_k$, we also have

$$H(\mathbf{x}_{k+1}) \leqslant H(\mathbf{x}_k) - \gamma_k \lambda_k \langle \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k \rangle. \tag{6.21}$$

Combining (6.20) and (6.21) gives a descent inequality, given by, for every $k \in \mathbb{N}$,

$$F(\mathbf{x}_{k+1}) \leqslant F(\mathbf{x}_k) - \left(\gamma_k \lambda_k \langle \nabla G(\mathbf{x}_k) + \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k \rangle - \frac{\beta_G \gamma_k^2 \lambda_k^2}{2} \|\mathbf{f}_k\|^2\right). \tag{6.22}$$

Note that, for every $k \in \mathbb{N}$, we have

$$\begin{aligned}&\gamma_k \lambda_k \langle \nabla G(\mathbf{x}_k) + \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k \rangle \\ &\quad = \gamma_k \lambda_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k \rangle - \gamma_k \lambda_k \langle \nabla H(\mathbf{x}_k) - \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k \rangle \\ &\quad \geqslant \gamma_k \lambda_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k \rangle - \gamma_k \lambda_k \big\|\nabla H(\mathbf{x}_k) - \nabla H(\mathbf{x}_{k+1})\big\| \, \|\mathbf{f}_k\| \qquad (6.23) \\ &\quad \geqslant \gamma_k \lambda_k \langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k \rangle - \frac{\beta_H \gamma_k^2 \lambda_k^2}{2} \|\mathbf{f}_k\|^2, \qquad (6.24)\end{aligned}$$

where (6.23) is obtained using the Cauchy-Schwarz inequality, and (6.24) using the Lipschitz continuity of ∇H and (6.3). Hence, since $\beta = \beta_G + \beta_H$, we have

$$\gamma_k\lambda_k\langle \nabla G(\mathbf{x}_k) + \nabla H(\mathbf{x}_{k+1}) \mid \mathbf{f}_k\rangle - \frac{\beta_G\gamma_k^2\lambda_k^2}{2}\|\mathbf{f}_k\|^2 \geqslant \gamma_k\lambda_k\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle - \frac{\gamma_k^2\lambda_k^2\beta}{2}\|\mathbf{f}_k\|^2. \quad (6.25)$$

On the one hand, using (6.18), we obtain

$$-\frac{\gamma_k^2\lambda_k^2\beta}{2}\|\mathbf{f}_k\|^2 \geqslant -\gamma_k^2\lambda_k^2\beta\Big(\|\Delta_k\|^2 + 2\rho_k^2\Big(\|\nabla F(\mathbf{x}_k)\|^2 + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|^2\Big)\Big). \quad (6.26)$$

On the other hand, using the definition of $\mathbf{f}_k$ in (6.10), equation (6.14), and the fact that $\nabla F = \nabla G + \nabla H$, we have

$$\begin{aligned}
\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle &= \langle \nabla F(\mathbf{x}_k) \mid \mathbf{d}_k + \Delta_k\rangle \\
&= \langle \nabla F(\mathbf{x}_k) \mid \nabla F(\mathbf{x}_k) - \nabla G(\mathbf{x}_k) - \nabla H(\mathbf{x}_k) \\
&\qquad\qquad + \mathbf{g}_k + \nabla H\big(\mathbf{x}_k - \gamma_k\mathbf{d}_k\big) + \Delta_k\rangle \\
&= \|\nabla F(\mathbf{x}_k)\|^2 - \big\langle \nabla F(\mathbf{x}_k) \mid \nabla H(\mathbf{x}_k) - \nabla H\big(\mathbf{x}_k - \gamma_k\mathbf{d}_k\big)\big\rangle \\
&\qquad\qquad + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle. \qquad (6.27)
\end{aligned}$$

Using the Cauchy-Schwarz inequality and the Lipschitz continuity of ∇H, we obtain

$$\begin{aligned}
\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle &\geqslant \|\nabla F(\mathbf{x}_k)\|^2 - \beta_H\gamma_k\|\nabla F(\mathbf{x}_k)\|\|\mathbf{d}_k\| \\
&\qquad\qquad + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle \\
&\geqslant \|\nabla F(\mathbf{x}_k)\|^2 - \beta_H\gamma_k\rho_k\|\nabla F(\mathbf{x}_k)\|\Big(\|\nabla F(\mathbf{x}_k)\| + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|\Big) \\
&\qquad\qquad + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle \qquad (6.28) \\
&= \big(1 - \beta_H\gamma_k\rho_k\big)\|\nabla F(\mathbf{x}_k)\|^2 - \beta_H\gamma_k\rho_k\|\nabla F(\mathbf{x}_k)\|\,\|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\| \\
&\qquad\qquad + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle.
\end{aligned}$$

where (6.28) is obtained using (6.16). For any $(a, b, c) \in \mathbb{R}_+^3$, we have $ab \leqslant ca^2 + b^2/(4c)$. Taking $a = \|\nabla F(\mathbf{x}_k)\|, b = \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|, c = (\sqrt{2} - 1)/2$, we obtain

$$\begin{aligned}
\langle \nabla F(\mathbf{x}_k) \mid \mathbf{f}_k\rangle &\geqslant \Big(1 - \beta_H\gamma_k\rho_k\frac{\sqrt{2}+1}{2}\Big)\|\nabla F(\mathbf{x}_k)\|^2 \\
&\qquad - \beta_H\gamma_k\rho_k\frac{\sqrt{2}+1}{2}\|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|^2 + \langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle. \quad (6.29)
\end{aligned}$$

Combining (6.25), (6.26), and (6.29), we obtain

$$\begin{aligned}
&\gamma_k\lambda_k\big\langle \nabla G(\mathbf{x}_k) + \nabla H\big(\mathbf{x}_{k+1}\big) \mid \mathbf{f}_k\big\rangle - \frac{\beta_G\gamma_k^2\lambda_k^2}{2}\|\mathbf{f}_k\|^2 \\
&\geqslant \gamma_k\lambda_k\Big(1 - \beta_H\gamma_k\rho_k\frac{\sqrt{2}+1}{2}\Big)\|\nabla F(\mathbf{x}_k)\|^2 - \beta_H\lambda_k\gamma_k^2\rho_k\frac{\sqrt{2}+1}{2}\|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|^2 \\
&+ \gamma_k\lambda_k\langle \nabla F(\mathbf{x}_k) \mid \mathbf{g}_k - \nabla G(\mathbf{x}_k) + \Delta_k\rangle \\
&\qquad\qquad - \gamma_k^2\lambda_k^2\beta\Big(\|\Delta_k\|^2 + 2\rho_k^2\Big(\|\nabla F(\mathbf{x}_k)\|^2 + \|\mathbf{g}_k - \nabla G(\mathbf{x}_k)\|^2\Big)\Big).
\end{aligned}$$

Using the definition of $(\sigma_k)_{k\in\mathbb{N}}$ given in (6.8) we thus obtain

$$\begin{aligned}
&\gamma_k\lambda_k\langle\nabla G(\mathbf{x}_k)+\nabla H(\mathbf{x}_{k+1})\mid\mathbf{f}_k\rangle-\frac{\beta_G\gamma_k^2\lambda_k^2}{2}\|\mathbf{f}_k\|^2\\
&\geqslant\gamma_k\lambda_k\left(1-\sigma_k\right)\|\nabla F(\mathbf{x}_k)\|^2-\gamma_k\lambda_k\sigma_k\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|^2\\
&\qquad\qquad-\gamma_k^2\lambda_k^2\beta\|\Delta_k\|^2+\gamma_k\lambda_k\langle\nabla F(\mathbf{x}_k)\mid\mathbf{g}_k-\nabla G(\mathbf{x}_k)+\Delta_k\rangle. \qquad (6.30)
\end{aligned}$$

Since, for every $k\in\mathbb{N}$, $\|\nabla F(\mathbf{x}_k)\|^2$ is $\mathcal{F}_k$-measurable, passing to conditional expectation in (6.30) and using successively (6.4), (6.5) and (6.6) gives

$$\begin{aligned}
&\mathbb{E}\left[\gamma_k\lambda_k\langle\nabla G(\mathbf{x}_k)+\nabla H(\mathbf{x}_{k+1})\mid\mathbf{f}_k\rangle-\frac{\beta_G\gamma_k^2\lambda_k^2}{2}\|\mathbf{f}_k\|^2\;\middle|\mathcal{F}_k\right]\\
&\quad\geqslant\gamma_k\lambda_k\left(1-\sigma_k\right)\|\nabla F(\mathbf{x}_k)\|^2-\gamma_k\lambda_k\sigma_k\,\mathbb{E}\big[\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|^2\;\big|\mathcal{F}_k\big]\\
&\qquad\qquad-\gamma_k^2\lambda_k^2\beta\,\mathbb{E}\big[\|\Delta_k\|^2\;\big|\mathcal{F}_k\big]\\
&\quad\geqslant\gamma_k\lambda_k\left(1-\sigma_k\right)\|\nabla F(\mathbf{x}_k)\|^2-\gamma_k\lambda_k\sigma_k\Big(d_k\|\nabla F(\mathbf{x}_k)\|^2+e_k\Big)\\
&\qquad\qquad-\lambda_k^2\beta\Big(d_k\|\nabla F(\mathbf{x}_k)\|^2+e_k\Big)\\
&\quad=\gamma_k\lambda_k\Big(1-\sigma_k\big(1+d_k\big)-\lambda_k\gamma_k^{-1}\beta d_k\Big)\|\nabla F(\mathbf{x}_k)\|^2\\
&\qquad\qquad-\gamma_k\lambda_k e_k\Big(\sigma_k+\lambda_k\gamma_k^{-1}\beta\Big)\;\text{ a.s.}. \qquad (6.31)
\end{aligned}$$

Taking the conditional expectation in (6.22), combining it with (6.31), we obtain

$$\begin{aligned}
\mathbb{E}\big[F(\mathbf{x}_{k+1})\;\big|\mathcal{F}_k\big]&\leqslant F(\mathbf{x}_k)-\gamma_k\lambda_k\Big(1-\sigma_k\big(1+d_k\big)-\lambda_k\gamma_k^{-1}\beta d_k\Big)\|\nabla F(\mathbf{x}_k)\|^2\\
&\qquad+\gamma_k\lambda_k e_k\Big(\sigma_k+\lambda_k\gamma_k^{-1}\beta\Big)\;\text{ a.s.}. \qquad (6.32)
\end{aligned}$$

Let, for every $k\in\mathbb{N}$, $u_k=0$, $v_k=\gamma_k\lambda_k\Big(1-\sigma_k\big(1+d_k\big)-\lambda_k\gamma_k^{-1}\beta d_k\Big)$, and $\mathrm{w}_k=\gamma_k\lambda_k e_k\Big(\sigma_k+\lambda_k\gamma_k^{-1}\beta\Big)$. We have $\sum_k u_k<+\infty$, according to (6.7) and (6.9) we have $\inf_k v_k>0$, and $\sum_k \mathrm{w}_k<+\infty$. Then, inequality (6.32) is of the form of (4.2).

It remains to show that (4.3) holds. According to (6.3) and (6.17), and using Jensen's inequality, for every $k\in\mathbb{N}$, we have

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_{k+1}-\mathbf{x}_k\|\;|\mathcal{F}_k]\\
&\quad\leqslant\gamma_k\lambda_k\rho_k\Big(\|\nabla F(\mathbf{x}_k)\|+\mathbb{E}\big[\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|\;\big|\mathcal{F}_k\big]+\gamma_k\lambda_k\mathbb{E}\big[\|\Delta_k\|\;\big|\mathcal{F}_k\big]\;\text{ a.s.}\\
&\quad\leqslant\gamma_k\lambda_k\rho_k\Big(\|\nabla F(\mathbf{x}_k)\|+\sqrt{\mathbb{E}\big[\|\mathbf{g}_k-\nabla G(\mathbf{x}_k)\|^2\;\big|\mathcal{F}_k\big]}\Big)+\gamma_k\lambda_k\sqrt{\mathbb{E}\big[\|\Delta_k\|^2\;\big|\mathcal{F}_k\big]}\;\text{ a.s.}
\end{aligned}$$

According to conditions (6.5) and (6.6), we then obtain

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{x}_{k+1}-\mathbf{x}_k\|\;|\mathcal{F}_k]\\
&\leqslant\gamma_k\lambda_k\rho_k\left(\|\nabla F(\mathbf{x}_k)\|+\sqrt{d_k\|\nabla F(\mathbf{x}_k)\|^2+e_k}\right)+\gamma_k\lambda_k\sqrt{d_k\|\nabla F(\mathbf{x}_k)\|^2+e_k}\\
&\leqslant\gamma_k\lambda_k\left(\rho_k+\rho_k d_k^{1/2}+d_k^{1/2}\right)\|\nabla F(\mathbf{x}_k)\|+\gamma_k\lambda_k(\rho_k+1)e_k^{1/2}\;\text{ a.s.}, \qquad (6.33)
\end{aligned}$$

where the last majoration is obtained using identity $\sqrt{a+b} \leqslant \sqrt{a}+\sqrt{b}$, that holds for any $a, b \geqslant 0$. Let, for every $k \in \mathbb{N}$, $r_k = 0$, $s_k = \gamma_k \lambda_k \left(\rho_k + \rho_k d_k^{1/2} + d_k^{1/2}\right)$, and $t_k = \gamma_k \lambda_k (\rho_k + 1) e_k^{1/2}$. We have $\sum_k r_k < +\infty$ and, according to (6.7) and (6.8), $\sum_k t_k < +\infty$. Hence, (6.33) is of the form of condition (4.3).

Hence we conclude that theorem 4.1 holds.

□

6.3 Application to some state-of-the-art algorithms

The stochastic aspect of process (6.2) enables two types of uncertainties. First, it takes into account uncertainty arising from stochastic approximations of the gradient of G. Second, it can handle uncertainty resulting from the use of an approximation of the proximity operator of H. The combination of these two types of uncertainties in a single scheme has been taken into account in some works in the literature.

6.3.1 Stochastic FB scheme

The authors of [17] show the a.s. convergence of the iterates of a stochastic FB scheme of the form of (6.2), assuming that G is convex, and considering the following error on true proximity operator, i.e.,

$$(\forall k \in \mathbb{N})(\forall \boldsymbol{u} \in \mathcal{H}) \quad \mathbf{P}_k(\boldsymbol{u}) = \operatorname{prox}_{\gamma_k H_k}(\boldsymbol{u}) + \mathbf{p}_k, \tag{6.34}$$

where $(H_k)_{k\in\mathbb{N}}$ are successive convex approximations of H and $(\mathbf{p}_k)_{k\in\mathbb{N}}$ is a conditionally summable sequence. The latter study generalises those of [45] for which no error term on the proximal operator was considered.

It is worth noticing that, up to our knowledge, the a.s. convergence of the iterates when G is non-convex has only been studied when $\lambda_k = 1$ and without any error on the proximal mapping [25, 34], i.e.,

$$(\forall k \in \mathbb{N})(\forall \boldsymbol{u} \in \mathcal{H}) \quad \mathbf{P}_k(\boldsymbol{u}) = \operatorname{prox}_{\gamma_k H}(\boldsymbol{u}). \tag{6.35}$$

Hence, the general results we presented in section 6.2 ensuring the a.s. convergence of the iterates when G is non-convex, and allowing stochastic approximations of the proximity operator of H, appear to be new.

6.3.2 Stochastic proximal scheme

Let, for every $k \in \mathbb{N}$, $\lambda_k = 1$, $\gamma_k = 1$, and let $G = 0$, and $H = \sum_{i=1}^{I} \omega_i H_i$, where $(\omega_i)_{1 \leqslant i \leqslant I} \in]0,1[^I$ with $\sum_{i=1}^{I} \omega_i = 1$ and, for every $i \in \{1, \ldots, I\}$, H_i Lipschitz-differentiable. Then scheme (6.2) boils down to a stochastic proximal algorithm of the form of

$$(\forall k \in \mathbb{N}) \quad \mathbf{x}_{k+1} = \mathbf{P}_k(\mathbf{x}_k). \tag{6.36}$$

Such algorithm is very similar to the federated-prox algorithm introduced in [48]. In particular, in [48], the authors propose to choose

$$(\forall k \in \mathbb{N})(\forall \boldsymbol{u} \in \mathcal{H}) \quad \mathbf{P}_k(\boldsymbol{u}) = \operatorname{prox}_{\sum_{i \in \mathbb{S}_k} \omega_i H_i}(\boldsymbol{u}), \tag{6.37}$$

where, for every $k \in \mathbb{N}$, $\mathbb{S}_k$ is a subset of $\{1, \ldots, I\}$. The authors show that, assuming the functions $(H_i)_{1 \leqslant i \leqslant I}$ satisfy some bounded dissimilarity condition, the expectation of the global objective function conditionally to the subsets of indices, decreases at each iteration.

The results presented in section 6.2 provide stronger convergence guaranties than in [48], at the price of slightly stronger assumptions on the choice of $(\mathbf{P}_k)_{k \in \mathbb{N}}$.

References

[1] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine 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.

[2] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

[3] Heinz H Bauschke, Patrick L Combettes, et al. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2011.

[4] Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, 2nd edition, 1999.

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

[6] Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of lojasiewicz inequalities and applications. *arXiv preprint arXiv:0802.0826*, 2008.

[7] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, Aug 2014.

[8] Silvia Bonettini, Marco Prato, and Simone Rebegoldi. Convergence of inexact forward–backward algorithms using the forward–backward envelope. *SIAM Journal on Optimization*, 30(4):3069–3097, 2020.

[9] Silvia Bonettini, Marco Prato, and Simone Rebegoldi. New convergence results for the inexact variable metric forward–backward method. *Applied Mathematics and Computation*, 392:125719, 2021.

[10] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[11] Mathieu Chalvidal, Emilie Chouzenoux, Jean-Baptiste Fest, and Claire Lefort. Block delayed majorize-minimize subspace algorithm for large scale image restoration. Technical report, 2022. https://hal.science/hal-03814231.

[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, Nov. 2016.

[13] Emilie Chouzenoux and Jean-Baptiste Fest. Convergence analysis of block majorize-minimize subspace approach. Technical report, 2022. https://hal.science/hal-03920026.

[14] Emilie Chouzenoux and Jean-Baptiste Fest. SABRINA: A stochastic subspace majorization-minimization algorithm. *Journal of Optimization Theory and Applications*, 195:919–952, 2022.

[15] Emilie Chouzenoux, Anna Jezierska, Jean-Christophe Pesquet, and Hugues Talbot. A majorize-minimize subspace approach for $\ell_2 - \ell_0$ image regularization. *SIAM Journal on Imaging Sciences*, 6(1):563–591, 2013.

[16] Emilie Chouzenoux, Jean-Christophe Pesquet, and Audrey 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.

[17] Patrick L Combettes and Jean-Christophe 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.

[18] Damek Davis. The asynchronous PALM algorithm for nonsmooth nonconvex problems. Technical report, 2016. https://arxiv.org/abs/1604.00526.

[19] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

[20] Marie Duflo. *Random Iterative Models*. Springer-Verlag, Berlin Heidelberg, 2013.

[21] Ilyas Fatkhullin, Jalal Etesami, Niao He, and Negar Kiyavash. Sharp analysis of stochastic optimization under global Kurdyka-Lojasiewicz inequality. Technical report, 2022. https://arxiv.org/abs/2210.01748.

[22] P. Frankel, G. Garrigos, and J. Peypouquet. Splitting methods with variablemetric for Kurdyka-Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.

[23] Sébastien Gadat and Ioana Gavra. Asymptotic study of stochastic adaptive algorithm in nonconvex landscape. *Journal of Machine Learning Research*, 23:1–54, 2022.

[24] Sébastien Gadat and Fabien Panloup. Optimal non-asymptotic bound of the Ruppert-Polyak averaging without strong convexity. Technical report, 2017. https://arxiv.org/abs/1709.03342.

[25] Caroline Geiersbach and Teresa Scarinci. Stochastic proximal gradient methods for nonconvex problems in Hilbert spaces. *Computational Optimization and Applications*, 78(3):705–740, 2021.

[26] Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. *Advances in Neural Information Processing Systems*, 32, 2019.

[27] EG Gladyshev. On stochastic approximation. *Theory of Probability & Its Applications*, 10(2):275–278, 1965.

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

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR 2015)*, 2015.

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

[31] Jean-François Le Gall. Intégration, probabilités et processus aléatoires. *Ecole Normale Supérieure de Paris*, 2006.

[32] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[33] Xi-Lin Li. Preconditioned stochastic gradient descent. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1454–1466, 2017.

[34] Xiao Li and Andre Milzarek. A unified convergence theorem for stochastic optimization methods. In *Proceedings of Thirty-Sixth Conference on Neural Information Processing Systems (NEURIPS 2022)*, 2022.

[35] Xiao Li, Andre Milzarek, and Junwen Qiu. Convergence of random reshuffling under the Kurdyka-Lojasiewicz inequality. Technical report, 2021. https://arxiv.org/abs/2110.04926.

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

[37] Paul-André Meyer. *Martingales and Stochastic Integrals I*, volume 284. Springer, 2006.

[38] Konstantin Mishchenko, Ahmed Khaled, and Peter Richtárik. Proximal and federated random reshuffling. In *Proceedings of the International Conference on Machine Learning (ICML 2022)*, pages 15718–15749, 2022.

[39] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[40] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 1999.

[41] Alexander M Ostrowski. *Solution of Equations in Euclidean and Banach Spaces*. Academic Press, 1973.

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

[43] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[44] Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. *Optimizing Methods in Statistics*, pages 233–257, 1971.

[45] Lorenzo Rosasco, Silvia Villa, and Bang Công Vũ. Convergence of stochastic proximal gradient algorithm. *Applied Mathematics and Optimization*, 82:891–917, 2014.

[46] Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. *Advances in Neural Information Processing Systems*, 25, 2012.

[47] David Saad. Online algorithms and stochastic approximations. *Online Learning*, 5(3):6, 1998.

[48] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. Technical report, 2018. https://arxiv.org/abs/1812.06127.

[49] Saverio Salzo and Silvia Villa. Inexact and accelerated proximal point algorithms. *Journal of Convex Analysis*, 19(4):1167–1192, 2012.

[50] Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition. Technical report, 2013. https://arxiv.org/abs/1308.6370.

[51] Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Proceedings of the Conference on Learning Theory (COLT 2021)*, pages 3935–3971, 2021.

[52] Xiao Wang, Shiqian Ma, Donald Goldfarb, and Wei Liu. Stochastic quasi-Newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017.