The stochastic Auxiliary Problem Principle in Banach spaces: measurability and convergence

Thomas Bittar, Pierre Carpentier, Jean-Philippe Chancelier, Jérome Lonchampt

The stochastic Auxiliary Problem Principle in Banach spaces: measurability and convergence

Thomas Bittar, Pierre Carpentier, Jean-Philippe Chancelier, Jérôme Lonchampt

May 20, 2022

Abstract

The stochastic Auxiliary Problem Principle (APP) algorithm is a general Stochastic Approximation (SA) scheme that turns the resolution of an original convex optimization problem into the iterative resolution of a sequence of auxiliary problems. This framework has been introduced to design decomposition-coordination schemes but also encompasses many well-known SA algorithms such as stochastic gradient descent or stochastic mirror descent. We study the stochastic APP in the case where the iterates lie in a Banach space and we consider an additive error on the computation of the subgradient of the objective. In order to derive convergence results or efficiency estimates for a SA scheme, the iterates must be random variables. This is why we prove the measurability of the iterates of the stochastic APP algorithm. Then, we extend convergence results from the Hilbert space case to the reflexive separable Banach space case. Finally, we derive efficiency estimates for the function values taken at the averaged sequence of iterates or at the last iterate, the latter being obtained by adapting the concept of modified Fejér monotonicity to our framework.

1 Introduction

Let $\mathbb{U}$ be a reflexive separable Banach space whose norm is denoted by $\|\cdot\|$, $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and $(\mathbb{W}, \mathcal{B}(\mathbb{W}))$ be a measurable topological vector space with $\mathcal{B}(\mathbb{W})$ being the Borel σ-field on $\mathbb{W}$. We refer to [3, 6] for the definitions of basic concepts in analysis and probability theory. We consider a stochastic optimization problem of the form:

$$\min_{u \in U^{\mathrm{ad}}} \left\{ J(u) := J^C(u) + J^\Sigma(u) \right\} \text{ where } \begin{cases} J^C(u) = \mathbb{E}\left(j^C(u, \boldsymbol{W}) \right) , \\ J^\Sigma(u) = \mathbb{E}\left(j^\Sigma(u, \boldsymbol{W}) \right) . \end{cases} \tag{1}$$

where $U^{\mathrm{ad}} \subset \mathbb{U}$ is a non-empty closed convex set, $\boldsymbol{W} : \Omega \to \mathbb{W}$ is a random variable, $j^C : \mathbb{U} \times \mathbb{W} \to \mathbb{R}$ and $j^\Sigma : \mathbb{U} \times \mathbb{W} \to \mathbb{R}$ are such that $j^C(\cdot, w)$ and $j^\Sigma(\cdot, w)$ are proper, convex, and lower-semicontinuous (l.s.c.) real-valued functions for all $w \in \mathbb{W}$.

Stochastic Approximation (SA) algorithms are the workhorse for solving Problem (1). The SA technique has been originally introduced in [20, 30] as an iterative method to find

the root of a monotone function which is known only through noisy estimates. SA algorithms have been the subject of many theoretical studies [2, 19, 25, 29] and have applications in various disciplines such as machine learning, signal processing or stochastic optimal control [4, 22]. Back in 1990, with decomposition applications in mind, Culioli and Cohen [13] proposed a general SA scheme in an infinite dimensional Hilbert space based on the so-called Auxiliary Problem Principle (APP), called the stochastic APP algorithm. This algorithm also encompasses several well-known algorithms such as stochastic gradient descent, the stochastic proximal gradient algorithm or stochastic mirror descent. Recently, [15, 24] apply SA methods to solve PDE-constrained optimization problems. In this paper, we extend the stochastic APP algorithm to the Banach case.

A SA algorithm is defined by a recursive stochastic update rule. For $k \in \mathbb{N}$, the k-th iterate of a SA algorithm is a mapping $\boldsymbol{U}_k : \Omega \to \mathbb{U}$, where the range of $\boldsymbol{U}_k$ is included in U^{ad}. We denote by $\langle \cdot\,, \cdot \rangle$ the duality pairing between $\mathbb{U}$ and its topological dual space $\mathbb{U}^\star$. In the case where j^C is differentiable with respect to u, the k-th iteration of the stochastic APP algorithm computes a minimizer u_{k+1} such that:

$$u_{k+1} \in \operatorname*{arg\,min}_{u \in U^{\mathrm{ad}}} K(u) + \left\langle \varepsilon_k \nabla_u j^C(u_k, w_{k+1}) - \nabla K(u_k)\,, u \right\rangle + \varepsilon_k j^\Sigma(u, w_{k+1})\,, \tag{2}$$

where $\varepsilon_k > 0$ is a positive real, w_{k+1} is a realization of the random variable $\boldsymbol{W}$ and K is a user-defined Gateaux-differentiable[1] convex function. The role of the function K is made clear in Section 2. In the context of the APP, Problem (2) is called the *auxiliary problem* and the function K is called the *auxiliary function*. Let us now briefly expose how this scheme reduces to well-known algorithms for particular values of K and j^Σ.

The most basic SA scheme is stochastic gradient descent. Assume that $\mathbb{U}$ is a Hilbert space, $U^{\mathrm{ad}} = \mathbb{U}$ and $j^\Sigma = 0$. The k-th iteration is given by:

$$u_{k+1} = u_k - \varepsilon_k \nabla_u j^C(u_k, w_{k+1})\,. \tag{3}$$

This is exactly the stochastic APP algorithm (2) with $j^\Sigma = 0$ and $K = \frac{1}{2}\|\cdot\|^2$ where $\|\cdot\|$ is the norm induced by the inner product in $\mathbb{U}$.

In the case where j^C is differentiable and j^Σ is non-smooth but with a proximal operator that is easy to compute, proximal methods [1, 28] are particularly efficient, even in a high-dimensional Hilbert space $\mathbb{U}$. An iteration of the stochastic proximal gradient algorithm is:

$$u_{k+1} \in \operatorname*{arg\,min}_{u \in \mathbb{U}} \frac{1}{2\varepsilon_k}\|u_k - u\|^2 + \left\langle \nabla_u j^C(u_k, w_{k+1})\,, u - u_k \right\rangle + j^\Sigma(u, w_{k+1})\,. \tag{4}$$

This is again the stochastic APP algorithm with $K = \frac{1}{2}\|\cdot\|^2$ but with a non zero function j^Σ. The proximal term $\frac{1}{2\varepsilon_k}\|u_k - u\|^2$ forces the next iterate u_{k+1} to be close to u_k with respect to the norm $\|\cdot\|$. When j^Σ is the indicator of a convex set, the stochastic proximal gradient

[1] We use [3, Definition 2.43] for the Gateaux-differentiability, which requires the linearity and boundedness of the directional derivative.

method reduces to stochastic projected gradient descent and when $j^{\Sigma} = 0$, this is just the regular stochastic gradient descent (3). Proximal methods are well-suited for regularized regression problems in machine learning for example.

When $\mathbb{U}$ is only a Banach space and not a Hilbert space, Equation (3) does not make sense as $u_k \in \mathbb{U}$ while $\nabla_u j^C(u_k, w_{k+1}) \in \mathbb{U}^\star$, thus the minus operation is not defined. This difficulty is addressed with the mirror descent algorithm [26]. The original insight of the method is to map the iterate u_k to $\nabla K(u_k) \in \mathbb{U}^\star$, where K is a Gateaux-differentiable user-defined function. Then, we do a gradient step in $\mathbb{U}^\star$ and we map back the resulting point to the primal space $\mathbb{U}$. The function K is called the *mirror map* in this setting [9]. There is also a proximal interpretation of mirror descent: instead of defining proximity with the norm $\|\cdot\|$, the mirror descent algorithm and its stochastic counterpart [25] use a Bregman divergence [7] that captures the geometric properties of the problem:

$$u_{k+1} \in \underset{u \in U^{\mathrm{ad}}}{\arg\min} \frac{1}{\varepsilon_k} D_K(u, u_k) + \left\langle \nabla_u j^C(u_k, w_{k+1}) \, , u - u_k \right\rangle \ , \tag{5}$$

where D_K is the Bregman divergence associated with K:

$$D_K(u, u') = K(u) - K(u') - \left\langle \nabla K(u') \, , u - u' \right\rangle , \quad u, u' \in \mathbb{U} \ .$$

The function K is sometimes called the *distance-generating function* as it defines the proximity between u and u'. With $K = \frac{1}{2} \|\cdot\|^2$, we get back to the setting of stochastic gradient descent. The mirror descent algorithm is particularly suited to the case where $\nabla_u j^C$ has a Lipschitz constant which is large with respect to the norm $\|\cdot\|$ but small with respect to some other norm that is better suited to the geometry of the problem [25]. For example, in the finite-dimensional case, the performance of stochastic gradient descent depends on the Lipschitz constant of $\nabla_u j^C$ in the Euclidean geometry. Hence, if the problem exhibits a non-Euclidean geometric structure, stochastic mirror descent may be more efficient. Stochastic mirror descent corresponds to the stochastic APP with a general function K and $j^{\Sigma} = 0$.

The stochastic APP algorithm combines the ideas of mirror descent and of the proximal gradient method. The iteration defined by (2) can be equivalently written as:

$$u_{k+1} \in \underset{u \in U^{\mathrm{ad}}}{\arg\min} \frac{1}{\varepsilon_k} D_K(u, u_k) + \left\langle \nabla_u j^C(u_k, w_{k+1}) \, , u - u_k \right\rangle + j^{\Sigma}(u, w_{k+1}) \ .$$

In the sequel, we stick to the formulation (2) and we consider a more general version as j^C is only assumed to be subdifferentiable and we allow for an additive error on the subgradient $\partial_u j^C(u_k, w_{k+1})$. Figure 1 summarizes the relationship between the four stochastic approximation algorithms that we have introduced.

The paper is organized as follows. In Section 2, we describe the setting of the stochastic APP algorithm considered in this paper along with some examples of application. In Section 3, we prove the measurability of the iterates of the stochastic APP algorithm in a reflexive separable Banach space. The issue of measurability is not often addressed in the literature, yet it is essential from a theoretical point of view. When convergence results or

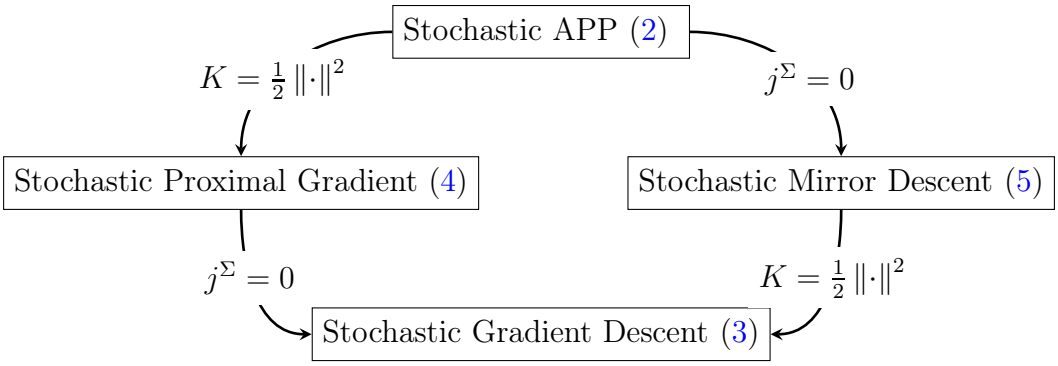

Figure 1: Links between the different stochastic approximation algorithms.

efficiency estimates are derived for SA algorithms, the iterates must be random variables so that the probabilities or the expectations that appear in the computation are well-defined. For that purpose, we carry out a precise study based on [10, 17] and we adapt some results of [32] to the infinite-dimensional case. Section 4 deals with convergence results and efficiency estimates. In §4.1, convergence results for the iterates and for the function values of the stochastic APP algorithm are extended to reflexive separable Banach spaces. These results already appear in [13] for the Hilbert case. They are also given, again in the Hilbert case, for stochastic projected gradient in [15] and stochastic mirror descent in the finite-dimensional setting [25]. In §4.2, we derive efficiency estimates for the expected function value taken either for the averaged sequence of iterates or for the last iterate. These efficiency estimates take into account the additive error on the subgradient, using the technique from [16]. To obtain convergence rates for the expected function value of the last iterate, we adapt the concept of modified Fejér monotonicity [23] to the framework of the stochastic APP algorithm. The paper ends by some concluding remarks in Section 5.

2 Description of the algorithm and examples

We describe the version of the stochastic APP algorithm that is studied in this paper and we give some examples of problems that fit in the general framework of Problem (1).

2.1 Setting of the stochastic APP algorithm

The original idea of the APP, first introduced in [11] and extended to the stochastic case in [13], is to solve a sequence of auxiliary problems whose solutions converge to the optimal solution of Problem (1). Assume that j^C is subdifferentiable with respect to u. At iteration k of the algorithm, a realization w_{k+1} of a random variable $\boldsymbol{W}_{k+1}$ is drawn. The random variables $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_{k+1}$ are independent and identically distributed as $\boldsymbol{W}$. Then, the following auxiliary problem is solved:

$$\min_{u \in U^{\mathrm{ad}}} K(u) + \langle \varepsilon_k(g_k + r_k) - \nabla K(u_k)\,, u \rangle + \varepsilon_k j^\Sigma(u, w_{k+1})\,, \tag{6}$$

where $g_k \in \partial_u j^C(u_k, w_{k+1})$ and we allow for an additive error r_k on the gradient. The term r_k represents a numerical error or a bias due to an approximation of the gradient e.g. with a finite difference scheme. The auxiliary problem is characterized by the choice of the auxiliary function K. In the introduction, we have given particular choices for K that lead to well-known algorithms. Depending on the context, the function K allows for an adaptation of the algorithm to the geometric structure of the data or it can provide decomposition properties to the algorithm, see Example 2.2. The stochastic APP algorithm is given in Algorithm 1.

Algorithm 1 Stochastic APP algorithm

1: Choose an initial point $u_0 \in U^{\mathrm{ad}}$, and a positive sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$.
2: At iteration k, draw a realization w_{k+1} of the random variable $\boldsymbol{W}_{k+1}$.
3: Solve Problem (6), denote by u_{k+1} the solution.
4: $k \leftarrow k+1$ and go back to 2.

No explicit stopping rule is provided in Algorithm 1. It is indeed difficult to know when to stop a stochastic algorithm as its properties are of statistical nature. Nevertheless, stopping rules have been developed in [36, 37] for the Robbins-Monro algorithm. In practice, the stopping criterion may be a maximal number of evaluations imposed by a budget limitation.

2.2 Some cases of interest for the stochastic APP

The structure of Problem (1) is very general and covers a wide class of problems that arise in machine learning or stochastic optimal control. We give some cases of interest that can be cast in this framework.

Example 2.1 Regularized risk minimization in machine learning.

Let $(\mathbb{X}, \mathfrak{X})$ and $(\mathbb{Y}, \mathfrak{Y})$ be two measurable spaces, where $\mathfrak{X}$ and $\mathfrak{Y}$ denote respectively the σ-fields on $\mathbb{X}$ and $\mathbb{Y}$. Let $X \subset \mathbb{X}$ and $Y \subset \mathbb{Y}$ and assume there is a probability distribution ν on $X \times Y$. Let $\{(x_i, y_i)\}_{1\le i\le N} \in (X \times Y)^N$ be a training set which consists of independent and identically distributed samples of a random vector $(\boldsymbol{X}, \boldsymbol{Y})$ following the distribution ν. Consider a convex loss function $\ell : Y \times Y \to \mathbb{R}_+$ and let $\mathbb{U}$ be a space of functions from X to Y. The goal of regularized expected loss minimization is to find a regression function $u^\sharp \in U^{\mathrm{ad}}$, where $U^{\mathrm{ad}} \subset \mathbb{U}$, such that:

$$u^\sharp \in \operatorname*{arg\,min}_{u\in U^{\mathrm{ad}}} \int_{X\times Y} \ell\big(y, u(x)\big)\nu(\mathrm{d}x, \mathrm{d}y) + R(u) \;, \tag{7}$$

where R is a regularization term. In practice, as the distribution ν is unknown, we solve an approximate problem, called the regularized empirical risk minimization problem:

$$u^\sharp \in \operatorname*{arg\,min}_{u\in U^{\mathrm{ad}}} \frac{1}{N}\sum_{i=1}^{N} \ell\big(y_i, u(x_i)\big) + R(u) \;. \tag{8}$$

Problem (8) is exactly of the form of Problem (7) if the distribution ν is taken to be the empirical measure $\nu = 1/N \sum_{i=1}^{N} \delta_{(x_i,y_i)}$, where $\delta_{(x_i,y_i)}$ denotes the measure of mass one at (x_i, y_i) and zero elsewhere. The regularized expected loss minimization Problem (7) is of the form of Problem (1) with the smooth term $J^C(u) = \int_{X\times Y} \ell\big(y, u(x)\big)\nu(\mathrm{d}x, \mathrm{d}y)$ and the possibly non-smooth term $J^\Sigma(u) = R(u)$.

Example 2.2 Decomposition aspects of the stochastic APP algorithm.

Let $n > 0$ be a given positive integer. Suppose that $\mathbb{U} = \mathbb{U}_1 \times \ldots \times \mathbb{U}_n$ and $U^{\mathrm{ad}} = U_1^{\mathrm{ad}} \times \ldots \times U_n^{\mathrm{ad}}$ with $U_i^{\mathrm{ad}} \subset \mathbb{U}_i$ for all $i \in \{1, \cdots, n\}$. Moreover, assume that j^Σ is an additive function, that is, $j^\Sigma(u, \boldsymbol{W}) = \sum_{i=1}^{n} j_i^\Sigma(u^i, \boldsymbol{W})$ with $u^i \in \mathbb{U}_i$, whereas j^C induces a non-additive coupling. In this case, Problem (1) is:

$$\min_{u\in U^{\mathrm{ad}}} J^C(u) + \sum_{i=1}^{n} J_i^\Sigma(u^i) \ ,$$

where $J_i^\Sigma(u^i) = \mathbb{E}\left(j_i^\Sigma(u^i, \boldsymbol{W})\right)$. We apply the stochastic APP algorithm with an additive auxiliary function $K(u) = \sum_{i=1}^{n} K_i(u^i)$. Let $\bar{u} \in \mathbb{U}$ be given, a canonical choice for K_i is:

$$K_i(u^i) = J^C(\bar{u}^{1:i-1}, u^i, \bar{u}^{i+1:n}), \quad i \in \{1, \ldots, n\} \ ,$$

where $\bar{u}^{i:j} = (\bar{u}^i, \ldots, \bar{u}^j)$ for $1 \le i \le j \le n$ and $\bar{u}^{1:0}$ denotes the empty vector by convention. Another classical choice is $K = \frac{1}{2}\|\cdot\|^2$. With an additive function K, the auxiliary problem (6) can be split into n independent subproblems that can be solved in parallel. At iteration k of the stochastic APP algorithm, the i-th subproblem is:

$$\min_{u^i\in U_i^{\mathrm{ad}}} K_i(u^i) + \left\langle \varepsilon_k(g_k^i + r_k^i) - \nabla K_i(u_k^i) \ , u^i \right\rangle + \varepsilon_k j_i^\Sigma(u^i, w_{k+1}) \ , \tag{9}$$

where $g_k^i \in \partial_{u^i} j^C(u_k, w_{k+1})$ and r_k^i is an additive error on $\partial_{u^i} j^C(u_k, w_{k+1})$. This example shows that the stochastic APP encompasses decomposition techniques.

3 Measurability of the iterates of the stochastic APP algorithm

Convergence results for SA algorithms often consist in proving the almost sure convergence of the sequence of iterates $\left\{\boldsymbol{U}_k\right\}_{k\in\mathbb{N}}$ to the optimal value $u^\sharp$. Other results provide non-asymptotic bounds for the expectation of function values $\mathbb{E}\big(J(\boldsymbol{U}_k) - J(u^\sharp)\big)$ or the quadratic mean $\mathbb{E}\left(\left\|\boldsymbol{U}_k - u^\sharp\right\|^2\right)$ for example. In order for these expectations and probabilities to be well-defined, $\boldsymbol{U}_k$ must be a measurable mapping from Ω to $\mathbb{U}$. However, as far as we know, the current literature does not provide constructive conditions under which the measurability of $\boldsymbol{U}_k$ is ensured. We aim at filling this theoretical gap by proving the measurability of the iterates of the stochastic APP algorithm.

3.1 A general measurability result

This section is devoted to the proof of a general measurability result in Theorem 3.23. We obtain the measurability of the iterates of the stochastic APP algorithm as a consequence in Theorem 3.27.

Recall that $(\Omega, \mathcal{A}, \mathbb{P})$ is a probability space and that $(\mathbb{W}, \mathcal{B}(\mathbb{W}))$ is a measurable topological vector space. The Banach space $\mathbb{U}$ is equipped with the Borel σ-field $\mathcal{B}(\mathbb{U})$. The topological dual of $\mathbb{U}$, equipped with the topology of the norm induced by the primal norm, is denoted by $\mathbb{U}^\star$, and its Borel σ-field is $\mathcal{B}(\mathbb{U}^\star)$. We consider the following problem:

$$\min_{u \in U^{\mathrm{ad}}} \left\{ \Phi(\omega, u) := K(u) + \left\langle \boldsymbol{\varphi}(\omega)\, , u \right\rangle + \varepsilon j^{\Sigma}\big(u, \boldsymbol{W}(\omega)\big) \right\} \ , \tag{10}$$

where $\varepsilon > 0$ is a given positive real number and $\boldsymbol{\varphi} : \Omega \to \mathbb{U}^\star$ is a given measurable function. The goal is to show the existence of a measurable mapping $\widetilde{\boldsymbol{U}}$ such that for all $\omega \in \Omega$, $\widetilde{\boldsymbol{U}}(\omega) \in \arg\min_{u \in U^{\mathrm{ad}}} \Phi(\omega, u)$. The mapping $\omega \mapsto \arg\min_{u \in U^{\mathrm{ad}}} \Phi(\omega, u)$ is a set-valued mapping.

3.1.1 Some tools from the theory of set-valued mappings

We recall some results from the theory of set-valued mappings that are used to state and prove the measurability result of Theorem 3.23. The definitions and propositions are mostly taken from [10, 17]. Theorem 3.23 requires $\mathbb{U}$ to be a reflexive separable Banach space. However, all results from §3.1.1 are more generally valid for $\mathbb{U}$ being a Polish space. For two sets X, Y, we denote by $\Gamma : X \rightrightarrows Y$ a set-valued mapping Γ from X to Y. This means that for $x \in X$, $\Gamma(x) \subset Y$ or in other words that $\Gamma(x) \in \mathcal{P}(Y)$ where $\mathcal{P}(Y)$ is the power set of Y.

Definition 3.1 (Measure completion) *Let $(\Omega, \mathcal{A})$ be a measurable space.*

- *Let μ be a measure on $(\Omega, \mathcal{A})$. The* μ-completion *of $\mathcal{A}$ is the σ-field $\mathcal{A}_\mu$ generated by $\mathcal{A} \cup \{A' \in \mathcal{P}(\Omega) \,|\, A' \subset A, A \in \mathcal{A} \text{ and } \mu(A) = 0\}$, that is, the union of $\mathcal{A}$ and the μ-negligible sets. The σ-field $\mathcal{A}$ is said to be* complete *for the measure μ if $\mathcal{A} = \mathcal{A}_\mu$.*
- *The σ-field $\hat{\mathcal{A}}$ of universally measurable sets is defined by $\hat{\mathcal{A}} = \bigcap_\mu \mathcal{A}_\mu$ where μ ranges over the set of positive σ-finite measures on the measurable space $(\Omega, \mathcal{A})$.*

Definition 3.2 (Measurable selection) *Let $(\Omega, \mathcal{A})$ be a measurable space and $\mathbb{U}$ be a separable Banach space. Let $\Gamma : \Omega \rightrightarrows \mathbb{U}$ be a set-valued mapping. A function $\gamma : \Omega \to \mathbb{U}$ is a measurable selection of Γ if $\gamma(\omega) \in \Gamma(\omega)$ for all $\omega \in \Omega$ and γ is measurable.*

Definition 3.3 (Measurable mapping) *Let $(\Omega, \mathcal{A})$ be a measurable space and $\mathbb{U}$ be a separable Banach space. A set-valued mapping $\Gamma : \Omega \rightrightarrows \mathbb{U}$ is Effros-measurable if, for every open set $O \subset \mathbb{U}$, we have:*

$$\Gamma^-(O) = \left\{ \omega \in \Omega,\ \Gamma(\omega) \cap O \neq \emptyset \right\} \in \mathcal{A} \ .$$

Remark 3.4 *The Effros-measurability of a set-valued mapping $\Gamma : \Omega \rightrightarrows \mathbb{U}$ is equivalent to the measurability of Γ viewed as a function from Ω to $\mathcal{P}(\mathbb{U})$.*

Proposition 3.5 *[10, Theorem III.9] Let $(\Omega, \mathcal{A})$ be a measurable space and $\mathbb{U}$ be a separable Banach space. Let $\Gamma : \Omega \rightrightarrows \mathbb{U}$ be a non-empty-valued and closed-valued mapping. Then the following statements are equivalent:*

(i) Γ *is Effros-measurable.*

(ii) Γ *admits a Castaing representation: there exists a sequence of measurable functions $\{\gamma_n\}_{n\in\mathbb{N}}$ such that for all $\omega \in \Omega$, $\Gamma(\omega) = \mathrm{cl}\left\{\gamma_n(\omega),\ n \in \mathbb{N}\right\}$ where* cl *denotes the closure of a set.*

Remark 3.6 *An important consequence of Proposition 3.5 is that any Effros-measurable mapping admits a measurable selection. This result is usually known as the Kuratowski–Ryll-Nardzewski selection theorem [21].*

Proposition 3.7 *[10, Proposition III.23: Sainte-Beuve's projection theorem] Let $(\Omega, \mathcal{A})$ be a measurable space and $(\mathbb{U}, \mathcal{B}(\mathbb{U}))$ be a separable Banach space equipped with its Borel σ-field. Let $G \in \mathcal{A} \otimes \mathcal{B}(\mathbb{U})$. Denote by $\mathrm{proj}_\Omega(G)$ the projection of G on Ω. Then, $\mathrm{proj}_\Omega(G) \in \hat{\mathcal{A}}$, where we recall that $\hat{\mathcal{A}}$ is the σ-field of universally measurable sets.*

Proposition 3.8 *[10, Proposition III.30] Let $(\Omega, \mathcal{A}, \mu)$ be a measure space where $\mathcal{A}$ is a complete σ-field, that is, $\mathcal{A} = \mathcal{A}_\mu$ and let $\mathbb{U}$ be a separable Banach space. Let $\Gamma : \Omega \rightrightarrows \mathbb{U}$ be a non-empty valued and closed-valued mapping. The following statements are equivalent:*

(i) Γ *is Effros-measurable.*

(ii) For every closed set $C \subset \mathbb{U}$, we have:

$$\Gamma^-(C) = \{\omega \in \Omega,\ \Gamma(\omega) \cap C \neq \emptyset\} \in \mathcal{A}\ .$$

Remark 3.9 *When $\mathbb{U}$ is finite-dimensional, Proposition 3.8 is true in any measurable space $(\Omega, \mathcal{A})$; that is, the completeness assumption of the σ-field $\mathcal{A}$ is not needed [32, Theorem 14.3]. In the infinite-dimensional setting, (ii) implies (i) remains true in any measurable space $(\Omega, \mathcal{A})$ [10, Proposition III.11]. The completeness assumption is only required to prove (i) implies (ii) when $\mathbb{U}$ is infinite-dimensional. Essentially, in the finite-dimensional case, the proof of (i) implies (ii) relies on the fact that $\mathbb{U}$ is locally compact. In the infinite-dimensional case, $\mathbb{U}$ is not locally compact and the proof uses the Sainte-Beuve's projection theorem.*

Definition 3.10 (Graph and epigraph) *Let $(\mathbb{X}, \mathcal{X})$ be a measurable space and $\mathbb{U}$ be a Banach space. Let $h : \mathbb{X} \to \mathbb{R} \cup \{+\infty\}$ be a function and $\Gamma : \mathbb{X} \rightrightarrows \mathbb{U}$ be a set-valued mapping.*

- *The graph and the epigraph of h are respectively defined by:*

$$\mathrm{gph}\, h = \left\{(x, \alpha) \in \mathbb{X} \times \mathbb{R},\ h(x) = \alpha\right\}\ ,$$
$$\mathrm{epi}\, h = \left\{(x, \alpha) \in \mathbb{X} \times \mathbb{R},\ h(x) \leq \alpha\right\}\ .$$

- *The graph of* Γ *is defined by:*

$$\operatorname{gph}\Gamma = \left\{(x,u) \in \mathbb{X} \times \mathbb{U},\ u \in \Gamma(x)\right\}.$$

Definition 3.11 (Normal integrand) *Let* $(\Omega, \mathcal{A})$ *be a measurable space and* $\mathbb{U}$ *be a Banach space. A function* $f : \Omega \times \mathbb{U} \to \mathbb{R} \cup \{+\infty\}$ *is a normal integrand if it satisfies the following conditions:*

(i) For all $\omega \in \Omega$, $f(\omega, \cdot)$ *is l.s.c.,*

(ii) The epigraphical mapping $S_f : \Omega \rightrightarrows \mathbb{U} \times \mathbb{R}$ *defined by* $S_f(\omega) = \operatorname{epi} f(\omega, \cdot)$ *is Effros-measurable.*

Remark 3.12 *The point (i) of Definition 3.11 is equivalent to* S_f *being closed-valued. In this paper, we consider the definition of the normal integrand used by Hess [17]. It differs from the definition of Castaing [10] where the point (ii) is replaced by the* $\mathcal{A} \otimes \mathcal{B}(\mathbb{U})$*-measurability of* f*. We shall see in Proposition 3.17 that the Effros-measurability of the epigraphical mapping* S_f *implies the* $\mathcal{A} \otimes \mathcal{B}(\mathbb{U})$*-measurability of* f*. Note also that if* $\mathcal{A}$ *is complete for a positive* σ*-finite measure* μ*, these two definitions are equivalent, see [10, Proposition III.30].*

Definition 3.13 (Carathéodory integrand) *Let* $(\Omega, \mathcal{A})$ *be a measurable space and* $\mathbb{U}$ *be a separable Banach space. A function* $f : \Omega \times \mathbb{U} \to \mathbb{R}$ *(finite-valued) is a Carathéodory integrand if it satisfies the following conditions:*

(i) For all $u \in \mathbb{U}$, $f(\cdot, u)$ *is measurable.*

(ii) For all $\omega \in \Omega$, $f(\omega, \cdot)$ *is continuous.*

Proposition 3.14 *[17, Proposition 2.5] If* f *is a Carathéodory integrand, then it is a normal integrand.*

Proposition 3.15 *[10, Proposition III.13] Let* $(\Omega, \mathcal{A})$ *be a measurable space and* $(\mathbb{U}, \mathcal{B}(\mathbb{U}))$ *be a separable Banach space equipped with its Borel* σ*-field. If* $\Gamma : \Omega \rightrightarrows \mathbb{U}$ *is an Effros-measurable, closed-valued mapping, then* $\operatorname{gph}\Gamma \in \mathcal{A} \otimes \mathcal{B}(\mathbb{U})$*.*

We now recall a technical result on the Borel σ-field of a product space that is used in the proof of subsequent propositions.

Proposition 3.16 *[5, Proposition 7.13] Let* $\left\{\left(\mathbb{X}_i, \mathcal{B}(\mathbb{X}_i)\right)\right\}_{i\in\mathbb{N}}$ *be a sequence of measurable separable topological spaces equipped with their Borel* σ*-fields. For* $n \in \mathbb{N}$*, let* $\mathbb{Y}_n = \prod_{i=1}^n \mathbb{X}_i$ *and let* $\mathbb{Y} = \prod_{i\in\mathbb{N}} \mathbb{X}_i$*. Then, the Borel* σ*-field of the product space* $\mathbb{Y}_n$ *(resp.* $\mathbb{Y}$*) coincides with the product of the Borel* σ*-fields of* $\{\mathbb{X}_i\}_{i=1}^n$ *(resp.* $\{\mathbb{X}_i\}_{i\in\mathbb{N}}$*), that is:*

$$\mathcal{B}\left(\mathbb{Y}_n\right) = \bigotimes_{i=1}^{n} \mathcal{B}(\mathbb{X}_i) \quad \textit{and} \quad \mathcal{B}\left(\mathbb{Y}\right) = \bigotimes_{i\in\mathbb{N}} \mathcal{B}(\mathbb{X}_i).$$

The following proposition shows that a normal integrand $f : \Omega \times \mathbb{U} \to \mathbb{R} \cup \{+\infty\}$, as defined in [17], is jointly $\mathcal{A} \otimes \mathcal{B}(\mathbb{U})$-measurable. This result is given in [32, Corollary 14.34] when $\mathbb{U} = \mathbb{R}^n$ but is extended here in the Banach case.

Proposition 3.17 *Let $(\Omega, \mathcal{A})$ be a measurable space and $(\mathbb{U}, \mathcal{B}(\mathbb{U}))$ be a separable Banach space equipped with its Borel σ-field. If $f : \Omega \times \mathbb{U} \to \mathbb{R} \cup \{+\infty\}$ is a normal integrand, then f is $\mathcal{A} \otimes \mathcal{B}(\mathbb{U})$-measurable.*

Proof. The function f is a normal integrand so its epigraphical mapping S_f is Effros-measurable and closed-valued. Moreover $\mathbb{U}$ is separable, so by Proposition 3.15, we get that:

$$\operatorname{gph} S_f = \big\{(\omega, u, \alpha) \in \Omega \times \mathbb{U} \times \mathbb{R},\ f(\omega, u) \leq \alpha\big\} \in \mathcal{A} \otimes \mathcal{B}(\mathbb{U} \times \mathbb{R}) \,.$$

Using that $\mathbb{U}$ and $\mathbb{R}$ are separable, we have $\mathcal{B}(\mathbb{U} \times \mathbb{R}) = \mathcal{B}(\mathbb{U}) \otimes \mathcal{B}(\mathbb{R})$ by Proposition 3.16. Then, for each $\alpha \in \mathbb{R}$, we get:

$$f^{-1}\big(]-\infty, \alpha]\big) = \big\{(\omega, u) \in \Omega \times \mathbb{U},\ f(\omega, u) \leq \alpha\big\} \in \mathcal{A} \otimes \mathcal{B}(\mathbb{U}) \,.$$

This shows that f is $\mathcal{A} \otimes \mathcal{B}(\mathbb{U})$-measurable. □

The following proposition is an adaptation of [32, Proposition 14.45(c)] on the composition operations on normal integrands to the Banach case. The separability of $\mathbb{U}$ is a crucial assumption that is used explicitly in the proof of Proposition 3.18 and that appears in most of the results of this part. Essentially, as only a countable union of measurable sets is measurable, countable dense subsets of a separable space are often used in proofs of measurability. Moreover, in the infinite-dimensional setting, we must assume the completeness of the σ-field $\mathcal{A}$ because we appeal to Proposition 3.8 in the proof.

Proposition 3.18 *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space where $\mathcal{A}$ is a complete σ-field, that is, $\mathcal{A} = \mathcal{A}_\mu$. Let $(\mathbb{W}, \mathcal{B}(\mathbb{W}))$ be a topological measurable space and $(\mathbb{U}, \mathcal{B}(\mathbb{U}))$ be a separable Banach space equipped with its Borel σ-field. Let $h : \mathbb{U} \times \mathbb{W} \to \mathbb{R} \cup \{+\infty\}$ be l.s.c. and $\boldsymbol{W} : \Omega \to \mathbb{W}$ be a measurable mapping. Then:*

$$f : (\omega, u) \in \Omega \times \mathbb{U} \mapsto h\big(u, \boldsymbol{W}(\omega)\big) \in \mathbb{R} \cup \{+\infty\}$$

is a normal integrand.

Proof. We have that h is l.s.c. so $f(\omega, \cdot) = h(\cdot, \boldsymbol{W}(\omega))$ is l.s.c. for all $\omega \in \Omega$. It remains to prove that the epigraphical mapping S_f is Effros-measurable. As h is l.s.c., the set $\operatorname{epi} h$ is closed. Define:

$$G : (\omega, u, \alpha) \in \Omega \times \mathbb{U} \times \mathbb{R} \mapsto (u, \boldsymbol{W}(\omega), \alpha) \in \mathbb{U} \times \mathbb{W} \times \mathbb{R} \,.$$

Then, let:

$$\begin{aligned} Q(\omega) &= \big[(\mathbb{U} \times \mathbb{R}) \times \operatorname{epi} h\big] \cap \operatorname{gph} G(\omega, \cdot, \cdot) \,, \\ &= \big\{ \big((u, \alpha), (u, \boldsymbol{W}(\omega), \alpha)\big) \text{ such that } h(u, \boldsymbol{W}(\omega)) \leq \alpha,\ (u, \alpha) \in \mathbb{U} \times \mathbb{R}\big\} \,, \\ &= \big\{ \big((u, \alpha), (u, \boldsymbol{W}(\omega), \alpha)\big) \text{ such that } f(\omega, u) \leq \alpha,\ (u, \alpha) \in \mathbb{U} \times \mathbb{R}\big\} \,. \end{aligned}$$

Now, define the projection operator P as:

$$P : (\mathbb{U} \times \mathbb{R}) \times (\mathbb{U} \times \mathbb{W} \times \mathbb{R}) \to (\mathbb{U} \times \mathbb{R}) \,,$$
$$((u, \alpha), (v, w, \beta)) \mapsto (u, \alpha)$$

so that we have:

$$S_f(\omega) = \big\{(u, \alpha) \in \mathbb{U} \times \mathbb{R},\ f(\omega, u) \leq \alpha\big\} = P\big(Q(\omega)\big) \,.$$

- Let Γ be the set valued mapping defined by $\Gamma : \omega \in \Omega \mapsto \operatorname{gph} G(\omega, \cdot, \cdot) \in (\mathbb{U} \times \mathbb{R}) \times (\mathbb{U} \times \mathbb{W} \times \mathbb{R})$. We show that Γ is Effros-measurable. As $\mathbb{U}$ is separable, there exists a countable dense subset $\{(b_n, r_n),\ n \in \mathbb{N}\}$ of $\mathbb{U} \times \mathbb{R}$. For $n \in \mathbb{N}$, let $\gamma_n(\omega) = \big((b_n, r_n), G(\omega, b_n, r_n)\big)$. As $G(\omega, b_n, r_n) = (b_n, \boldsymbol{W}(\omega), r_n)$ and $\boldsymbol{W}$ is measurable, we get that γ_n is measurable. Then, we have $\Gamma(\omega) = \operatorname{cl}\{\gamma_n(\omega),\ n \in \mathbb{N}\}$. Hence, $\{\gamma_n\}_{n \in \mathbb{N}}$ is a Castaing representation of Γ. Moreover, Γ is closed-valued and non-empty valued so by Proposition 3.5, we deduce that Γ is Effros-measurable.

- Let $C \subset (\mathbb{U} \times \mathbb{R}) \times (\mathbb{U} \times \mathbb{W} \times \mathbb{R})$ be a closed set. We have:

$$Q^-(C) = \big\{\omega \in \Omega, \big[(\mathbb{U} \times \mathbb{R}) \times \operatorname{epi} h\big] \cap \Gamma(\omega) \cap C \neq \emptyset\big\} \,,$$
$$= \Gamma^-\left(C \cap \big[(\mathbb{U} \times \mathbb{R}) \times \operatorname{epi} h\big]\right) \,.$$

  As $\operatorname{epi} h$ is closed, the set $C \cap \big[(\mathbb{U} \times \mathbb{R}) \times \operatorname{epi} h\big]$ is closed. By assumption, the σ-field $\mathcal{A}$ is complete and we have shown that Γ is Effros-measurable, therefore by Proposition 3.8, we get that $\Gamma^-\left(C \cap \big[(\mathbb{U} \times \mathbb{R}) \times \operatorname{epi} h\big]\right) = Q^-(C) \in \mathcal{A}$. Hence, Q is Effros-measurable.

- Finally, for every open set $V \subset \mathbb{U} \times \mathbb{R}$, as $S_f(\omega) = P\big(Q(\omega)\big)$, we have:

$$S_f^-(V) = \big\{\omega \in \Omega,\ Q(\omega) \cap P^{-1}(V) \neq \emptyset\big\} \,.$$

  The projection P is continuous so $P^{-1}(V)$ is open. As Q is Effros-measurable, we get that $S_f^-(V) \in \mathcal{A}$, that is, S_f is Effros-measurable.

This completes the proof. □

We now give the main results that are used to prove the measurability of the iterates of the stochastic APP. The following proposition is a slight extension of [18, Proposition 4.2(c)].

Proposition 3.19 *Let $(\Omega, \mathcal{A})$ be a measurable space and $(\mathbb{U}, \mathcal{B}(\mathbb{U}))$ be a separable Banach space equipped with its Borel σ-field. Let U^{ad} be a closed subset of $\mathbb{U}$. Let $f : \Omega \times \mathbb{U} \to \mathbb{R} \cup \{+\infty\}$ be an $\mathcal{A} \otimes \mathcal{B}(\mathbb{U})$-measurable function. Let $M : \Omega \rightrightarrows \mathbb{U}$ be the argmin set-valued mapping:*

$$M(\omega) = \operatorname*{arg\,min}_{u \in U^{\mathrm{ad}}} f(\omega, u) \,.$$

Assume that the argmin mapping M is non-empty valued, then M admits an $\hat{\mathcal{A}}$-measurable selection.

Proof. Let $\alpha \in \mathbb{R}$ and $m(\omega) = \min_{u \in U^{\mathrm{ad}}} f(\omega, u)$. The function m is well-defined as M is non-empty valued. Let:

$$H = (\Omega \times U^{\mathrm{ad}}) \cap \{(\omega, u) \in \Omega \times \mathbb{U},\ f(\omega, u) < \alpha\} \ .$$

We have:

$$\{\omega \in \Omega,\ m(\omega) < \alpha\} = \mathrm{proj}_{\Omega}(H) \ ,$$

where $\mathrm{proj}_{\Omega}(H)$ is the projection of H on Ω. As f is $\mathcal{A} \otimes \mathcal{B}(\mathbb{U})$-measurable and U^{ad} is closed hence measurable, we get that $H \in \mathcal{A} \otimes \mathcal{B}(\mathbb{U})$. From Proposition 3.7, we deduce that $m^{-1}(]-\infty, \alpha[)$ is $\hat{\mathcal{A}}$-measurable so that m is $\hat{\mathcal{A}}$-measurable. As $\mathcal{A} \subset \hat{\mathcal{A}}$, we have that $\mathcal{A} \otimes \mathcal{B}(\mathbb{U}) \subset \hat{\mathcal{A}} \otimes \mathcal{B}(\mathbb{U})$, therefore the function f is $\hat{\mathcal{A}} \otimes \mathcal{B}(\mathbb{U})$-measurable. We can write:

$$M(\omega) = \left\{u \in U^{\mathrm{ad}},\ f(\omega, u) = m(\omega)\right\} \ ,$$

so, $\mathrm{gph}\, M = \left\{(\omega, u) \in \Omega \times U^{\mathrm{ad}}, f(\omega, u) = m(\omega)\right\}$. Therefore, $\mathrm{gph}\, M$ is $\hat{\mathcal{A}} \otimes \mathcal{B}(\mathbb{U})$-measurable as the inverse image of $\{0\}$ under the $\hat{\mathcal{A}} \otimes \mathcal{B}(\mathbb{U})$-measurable mapping $(\omega, u) \mapsto f(\omega, u) - m(\omega)$. Let O be an open subset of $\mathbb{U}$. We have:

$$M^{-}(O) = \mathrm{proj}_{\Omega}\left((\Omega \times O) \cap \mathrm{gph}\, M\right) \ .$$

As $(\Omega \times O) \cap \mathrm{gph}\, M \in \hat{\mathcal{A}} \otimes \mathcal{B}(\mathbb{U})$, by Proposition 3.7, we get that $M^{-}(O) \in \hat{\hat{\mathcal{A}}} = \hat{\mathcal{A}}$. Hence, M is Effros-measurable for the σ-field $\hat{\mathcal{A}}$ and is non-empty-valued by assumption, so by Proposition 3.5, M admits an $\hat{\mathcal{A}}$-measurable selection. □

Corollary 3.20 *Let $(\Omega, \mathcal{A}, \mu)$ be a complete probability space, i.e. $\mathcal{A} = \mathcal{A}_{\mu}$. Let $(\mathbb{U}, \mathcal{B}(\mathbb{U}))$ be a separable Banach space equipped with its Borel σ-field. Let $f : \Omega \times \mathbb{U} \to \mathbb{R} \cup \{+\infty\}$ be an $\mathcal{A} \otimes \mathcal{B}(\mathbb{U})$-measurable function. Suppose that the argmin mapping M is non-empty valued. Then, M admits an $\mathcal{A}$-measurable selection.*

Proof. As μ is a positive σ-finite measure, we have $\hat{\mathcal{A}} = \bigcap_{\mu} \mathcal{A}_{\mu} \subset \mathcal{A}_{\mu} = \mathcal{A}$. By Proposition 3.19, M admits an $\hat{\mathcal{A}}$-measurable selection, which is also an $\mathcal{A}$-measurable selection. □

Proposition 3.21 *[17, Theorem 4.6] Let $(\Omega, \mathcal{A})$ be a measurable space and $\mathbb{U}$ be a separable Banach space with separable topological dual $\mathbb{U}^{\star}$. Let $f : \Omega \times \mathbb{U} \to \mathbb{R} \cup \{+\infty\}$ be a proper normal integrand and $\boldsymbol{U} : \Omega \to \mathbb{U}$ be a measurable mapping. Then, the set-valued mapping $D_{\boldsymbol{U}} : \Omega \rightrightarrows \mathbb{U}^{\star}$ defined by*

$$\begin{aligned} D_{\boldsymbol{U}}(\omega) &= \partial_u f\big(\omega, \boldsymbol{U}(\omega)\big) \\ &= \left\{v \in \mathbb{U}^{\star},\ f(\omega, u) \geq f(\omega, \boldsymbol{U}(\omega)) + \left\langle v\, , u - \boldsymbol{U}(\omega)\right\rangle,\ \forall u \in \mathbb{U}\right\} , \end{aligned}$$

is Effros-measurable.

3.1.2 Existence of a measurable selection for the argmin mapping of Φ

In this section, we make use of the tools introduced in §3.1.1 to prove our main measurability result. We introduce the argmin set-valued mapping $M : \Omega \rightrightarrows \mathbb{U}$ for Problem (10):

$$M(\omega) = \underset{u \in U^{\mathrm{ad}}}{\arg\min} \left\{ \Phi(\omega, u) := K(u) + \left\langle \boldsymbol{\varphi}(\omega)\,, u \right\rangle + \varepsilon j^{\Sigma}\big(u, \boldsymbol{W}(\omega)\big) \right\} . \tag{11}$$

We consider the following assumptions:

(A1) The space $\mathbb{U}$ is a reflexive, separable Banach space. This implies in particular that $\mathbb{U}^\star$ is separable.

(A2) U^{ad} is a non-empty closed convex subset of $\mathbb{U}$.

(A3) $j^{\Sigma} : \mathbb{U} \times \mathbb{W} \to \mathbb{R}$ is jointly l.s.c. and for all $w \in \mathbb{W}$, $j^{\Sigma}(\cdot, w)$ is proper and convex.

(A4) The function $K : \mathbb{U} \to \mathbb{R}$ is proper, convex, l.s.c. and Gateaux-differentiable on an open set containing U^{ad}.

(A5) For all $\omega \in \Omega$, the function $u \mapsto \Phi(\omega, u)$ is coercive on U^{ad} meaning that when $\|u\| \to +\infty$ with $u \in U^{\mathrm{ad}}$, we have $\Phi(\omega, u) \to +\infty$. This assumption is automatically satisfied if U^{ad} is bounded.

(A6) The σ-field $\mathcal{A}$ is complete for the measure $\mathbb{P}$, that is, $\mathcal{A} = \mathcal{A}_{\mathbb{P}}$.

(A7) The function $\boldsymbol{W} : \Omega \to \mathbb{W}$ is measurable.

(A8) The function $\boldsymbol{\varphi} : \Omega \to \mathbb{U}^\star$ is measurable.

The objective of this part is to prove that M defined in Equation (11) admits a measurable selection. We start by a classical theorem from optimization theory giving conditions for the existence and uniqueness of a minimizer $\Phi(\omega, \cdot)$.

Theorem 3.22 *Let $\omega \in \Omega$. Under Assumptions (A1)-(A5), $M(\omega)$ is non-empty, closed and convex. Moreover, if K is strongly convex, then $M(\omega)$ is a singleton, meaning that $\Phi(\omega, \cdot)$, defined in (11), has a unique minimizer.*

Proof. The objective function $\Phi(\omega, \cdot)$ is the sum of three convex, l.s.c. functions, it is then convex and l.s.c. By (A5), $\Phi(\omega, \cdot)$ is also coercive. As $\mathbb{U}$ is a reflexive Banach space (A1) and U^{ad} is non-empty, closed and convex (A2), the set of minimizers $M(\omega)$ is non-empty [8, Corollary III.20]. The convexity of $\Phi(\omega, \cdot)$ ensures that $M(\omega)$ is convex and the lower-semicontinuity of $\Phi(\omega, \cdot)$ ensures that $M(\omega)$ is closed.

If K is strongly convex, then $\Phi(\omega, \cdot)$ is strongly convex, hence the minimizer of $\Phi(\omega, \cdot)$ is unique so $M(\omega)$ is a singleton.[2] □

[2] In the case where K is strongly convex, the coercivity assumption is not needed as it is implied by the strong convexity of $\Phi(\omega, \cdot)$.

Theorem 3.23 *Under Assumptions (A1)-(A8), the mapping M defined in Equation (11) admits a measurable selection.*

Proof. We start by proving that $\Phi(\omega, u) = K(u) + \langle \boldsymbol{\varphi}(\omega), u \rangle + \varepsilon j^{\Sigma}\big(u, \boldsymbol{W}(\omega)\big)$ is a normal integrand:

- As the function K is l.s.c. (A4), $(\omega, u) \mapsto K(u)$ is a normal integrand. Indeed, its epigraphical mapping $\omega \mapsto \{(u, \alpha) \in \mathbb{U} \times \mathbb{R},\ K(u) \leq \alpha\}$ is a constant function of ω and is then measurable.
- The Banach space $\mathbb{U}$ is separable (A1) and $\mathcal{A}$ is complete (A6). The space $\mathbb{U}^\star$ equipped with its Borel σ-field $\mathcal{B}(\mathbb{U}^\star)$ is a measurable space. The function $\boldsymbol{\varphi}$ is measurable (A8) and the function $(u, v) \in \mathbb{U} \times \mathbb{U}^\star \mapsto \langle v, u \rangle \in \mathbb{R}$ is continuous hence, in particular, l.s.c.. Then, Proposition 3.18 applies, showing that the function $(\omega, u) \mapsto \langle \boldsymbol{\varphi}(\omega), u \rangle$ is a normal integrand.
- With the same reasoning, using that $\mathbb{U}$ is separable (A1), $\boldsymbol{W}$ is measurable (A7), $\mathcal{A}$ is complete (A6) and j^{Σ} is l.s.c. (A3), we use Proposition 3.18 with $h = j^{\Sigma}$ to deduce that $(\omega, u) \mapsto j^{\Sigma}\big(u, \boldsymbol{W}(\omega)\big)$ is a normal integrand.

The function Φ is then a normal integrand as the sum of three normal integrands. As $\mathbb{U}$ is separable (A1), we use Proposition 3.17 to get that Φ is $\mathcal{A} \otimes \mathcal{B}(\mathbb{U})$-measurable. In addition, using (A2)-(A5) to apply Theorem 3.22 ensures that M is non-empty valued. Moreover, the σ-field $\mathcal{A}$ is complete for $\mathbb{P}$ (A6). Hence, by Corollary 3.20, we conclude that $M : \omega \mapsto \arg\min_{u \in U^{\mathrm{ad}}} \Phi(\omega, u)$ admits a measurable selection. □

Corollary 3.24 *Under Assumptions (A1)-(A8) and if we additionally assume that K is strongly convex, then for all $\omega \in \Omega$, $\Phi(\omega, \cdot)$, defined in (11), has a unique minimizer and the mapping:*

$$\widetilde{\boldsymbol{U}}(\omega) = \operatorname*{arg\,min}_{u \in U^{\mathrm{ad}}} \Phi(\omega, u) \in \mathbb{U}$$

is measurable, that is, $\widetilde{\boldsymbol{U}}$ is a random variable.

3.2 Application to the stochastic APP algorithm

We aim at studying the iterations of the stochastic APP in terms of random variables so we consider the argmin set-valued mapping $M : \Omega \rightrightarrows \mathbb{U}$ defined by:

$$M(\omega) = \operatorname*{arg\,min}_{u \in U^{\mathrm{ad}}} K(u) + \big\langle \varepsilon(\boldsymbol{G}(\omega) + \boldsymbol{R}(\omega)) - \nabla K(\boldsymbol{U}(\omega)), u \big\rangle + \varepsilon j^{\Sigma}(u, \boldsymbol{W}(\omega)), \tag{12}$$

with $\varepsilon > 0$, $\boldsymbol{U}(\omega) \in U^{\mathrm{ad}}$, $\boldsymbol{W}(\omega) \in \mathbb{W}$, $\boldsymbol{G}(\omega) \in \partial_u j^C(\boldsymbol{U}(\omega), \boldsymbol{W}(\omega))$ and $\boldsymbol{R}(\omega) \in \mathbb{U}^\star$. An iteration of the stochastic APP algorithm consists in solving Problem (12), which is exactly of the form of Problem (11) with:

$$\boldsymbol{\varphi}(\omega) = \varepsilon\big(\boldsymbol{G}(\omega) + \boldsymbol{R}(\omega)\big) - \nabla K\big(\boldsymbol{U}(\omega)\big). \tag{13}$$

In addition to (A1)-(A7), we assume now:

(A9) The function $j^C : \mathbb{U} \times \mathbb{W} \to \mathbb{R}$ that appears in Problem (1) is jointly l.s.c. and for all $w \in \mathbb{W}$, $j^C(\cdot, w)$ is proper, convex and subdifferentiable on an open set containing U^{ad}.

(A10) The mappings $\boldsymbol{U} : \Omega \to U^{\mathrm{ad}}$ and $\boldsymbol{R} : \Omega \to \mathbb{U}^\star$ are measurable.

In (A10), we assume that the mappings $\boldsymbol{U}$ and $\boldsymbol{R}$ are random variables. We cannot do the same for the mapping $\boldsymbol{G}$ as it must satisfy $\boldsymbol{G}(\omega) \in \partial_u j^C(\boldsymbol{U}(\omega), \boldsymbol{W}(\omega))$ for all $\omega \in \Omega$. In the following proposition, we ensure that there exists a measurable mapping satisfying this relationship.

Proposition 3.25 *Under Assumptions (A1), (A6), (A7), (A9), (A10), the subgradient mapping $\Gamma : \omega \mapsto \partial_u j^C(\boldsymbol{U}(\omega), \boldsymbol{W}(\omega)) \subset \mathbb{U}^\star$ admits a measurable selection $\boldsymbol{G} : \Omega \to \mathbb{U}^\star$.*

Proof. Let $f(\omega, u) = j^C\big(u, \boldsymbol{W}(\omega)\big)$ for $\omega \in \Omega,\ u \in \mathbb{U}$.

- Using that $\mathbb{U}$ is separable (A1), $\boldsymbol{W}$ is measurable (A7), $\mathcal{A}$ is complete (A6) and j^C is l.s.c. (A9), Proposition 3.18 with $h = j^C$ shows that f is a normal integrand.
- We have that for all $\omega \in \Omega$, $\Gamma(\omega) = \partial_u f\big(\omega, \boldsymbol{U}(\omega)\big)$. With (A9), we get that $f(\omega, \cdot)$ is proper for all $\omega \in \Omega$. We have that $\mathbb{U}$ and $\mathbb{U}^\star$ are separable (A1), $\boldsymbol{U}$ is measurable (A10) and f is a normal integrand, so by Proposition 3.21, Γ is Effros-measurable.

Assumption (A9) ensures that Γ is non-empty valued. In addition, Γ is Effros-measurable and closed-valued in $\mathbb{U}^\star$ which is separable. By Proposition 3.5, Γ admits a measurable selection. This means that there exists a measurable function $\boldsymbol{G} : \Omega \to \mathbb{U}^\star$ such that for all $\omega \in \Omega$, $\boldsymbol{G}(\omega) \in \Gamma(\omega) = \partial_u j^C(\boldsymbol{U}(\omega), \boldsymbol{W}(\omega))$. □

In the sequel, $\boldsymbol{G}$ denotes a measurable selection of Γ. In order to apply Theorem 3.23 to prove that the iterates of the stochastic APP algorithm are measurable, we must ensure that Assumption (A8) is satisfied, that is, we must show that the mapping $\boldsymbol{\varphi}$ defined in (13) is measurable. We prove in Proposition 3.26 that Assumption (A8) can be deduced from the other assumptions.

Proposition 3.26 *Under Assumptions (A1), (A4), (A7), (A9), (A10), the function $\boldsymbol{\varphi}$ is measurable.*

Proof. We have already seen in the proof of Theorem 3.23 that $\Lambda : (\omega, u) \mapsto K(u)$ is a normal integrand. Assumption (A4) ensures that $\Lambda(\omega, \cdot)$ is proper for all $\omega \in \Omega$. We have that $\mathbb{U}$ and $\mathbb{U}^\star$ are separable (A1), $\boldsymbol{U}$ is measurable (A10), so $\omega \mapsto \nabla_u \Lambda\big(\omega, \boldsymbol{U}(\omega)\big) = \nabla K\big(\boldsymbol{U}(\omega)\big)$ is measurable by Proposition 3.21. Finally, $\boldsymbol{R}$ is also measurable (A10), so $\boldsymbol{\varphi}$ is measurable as a sum of measurable functions. □

From Theorem 3.23 and Proposition 3.26, we have obtained that under Assumptions (A1)-(A7), (A9),(A10), the mapping M defined in (12) admits a measurable selection. Now, we

give the measurability result for the iterates of the stochastic APP algorithm, which is defined by the following recursion for $\omega \in \Omega$ and $k \in \mathbb{N}$:

$$\begin{aligned} M_0(\omega) &= \{u_0\} \subset U^{\mathrm{ad}} \,, \\ M_{k+1}(\omega) &= \underset{u \in U^{\mathrm{ad}}}{\arg\min}\, K(u) + \left\langle \varepsilon_k \big(\boldsymbol{G}_k(\omega) + \boldsymbol{R}_k(\omega)\big) - \nabla K\big(\boldsymbol{U}_k(\omega)\big)\,, u \right\rangle \\ &\qquad + \varepsilon_k j^{\Sigma}\big(u, \boldsymbol{W}_{k+1}(\omega)\big) \,, \end{aligned} \tag{14}$$

Theorem 3.27 *Under Assumptions (A1)-(A7), (A9), (A10), for all $k \in \mathbb{N}$, the mapping M_k that defines the k-th iteration of the stochastic APP algorithm (14) admits a measurable selection.*

Proof. The mapping M_0 admits a measurable selection defined by $\boldsymbol{U}_0(\omega) = u_0$. Then, by iteratively using the fact that (12) admits a measurable selection, we deduce that for all $k \in \mathbb{N}$, M_k admits a measurable selection. □

Corollary 3.28 *Assume that (A1)-(A7), (A9), (A10) are satisfied and that the auxiliary mapping K is strongly convex. Then, for all $k \in \mathbb{N}$, the unique mapping $\boldsymbol{U}_k$ that defines the k-th iterate of the stochastic APP algorithm is measurable.*

Proof. If K is strongly convex, from Corollary 3.24, we get that M_k is single-valued, so the iterate $\boldsymbol{U}_k$ is uniquely defined. The measurability of $\boldsymbol{U}_k$ follows from Theorem 3.27. □

Remark 3.29 *In [32, Chapter 14], Rockafellar exposes a whole set of measurability results in the case where $\mathbb{U}$ is finite-dimensional. The finite-dimensional framework allows to avoid some technicalities of the infinite-dimensional case. In particular, the completeness assumption (A6) is not needed as shown by [32, Proposition 14.37] which is the analogous of Proposition 3.19 in the finite-dimensional case.*

Remark 3.30 *In Problem (1), when $\mathbb{U}$ is a Hilbert space, $U^{\mathrm{ad}} = \mathbb{U}$, $j^{\Sigma} = 0$ and j^C is assumed to be differentiable with respect to u, we can use stochastic gradient descent. Then, we have the explicit formula:*

$$\boldsymbol{U}_{k+1} = \boldsymbol{U}_k - \varepsilon_k \nabla_u j^C(\boldsymbol{U}_k, \boldsymbol{W}_{k+1}) \,. \tag{15}$$

Under Assumptions (A1), (A7), (A9), the measurability of the iterates is directly obtained by induction using the explicit formula (15).

4 Convergence results and efficiency estimates

In this section, we prove the convergence of the stochastic APP algorithm for solving Problem (1). In addition, we give efficiency estimates for the convergence of function values. Some technical results for the proofs of this section are given in the appendix.

4.1 Convergence of the stochastic APP algorithm

Let $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$ be a filtration with $\mathcal{F}_k = \sigma\left(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_k\right)$, where $\left(\boldsymbol{W}_1, \ldots, \boldsymbol{W}_k\right)$ are the random variables that appear in the successive iterations of the stochastic APP algorithm (14) defined on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Recall that, in (14), $\boldsymbol{G}_k \in \partial_u j^C(\boldsymbol{U}_k, \boldsymbol{W}_{k+1})$ [3] is an unbiased stochastic gradient, whereas $\boldsymbol{R}_k$ represents a bias on the gradient.

Convergence results for the stochastic APP algorithm are already proved in [12, 13] when $\mathbb{U}$ is a Hilbert space (possibly infinite-dimensional) and when there is no bias $\boldsymbol{R}_k$. In [15], convergence of the projected stochastic gradient descent is proved in a Hilbert space and with a bias $\boldsymbol{R}_k$. For stochastic mirror descent in the finite-dimensional setting, convergence results and efficiency estimates can be found in [25], but no bias is considered. Here, we present convergence results for the stochastic APP algorithm in a reflexive separable Banach space and we allow for a bias $\boldsymbol{R}_k$, hence generalizing previous results.

In addition to (A1)-(A7), (A9), (A10), we make the following assumptions:

(A11) The functions $j^C(\cdot, w) : \mathbb{U} \to \mathbb{R}$ and $j^\Sigma(\cdot, w) : \mathbb{U} \to \mathbb{R}$ have linearly bounded subgradient in u, uniformly in $w \in \mathbb{W}$:

$$\begin{cases} \exists c_1, c_2 > 0\ , \ \ \forall (u, w) \in U^{\mathrm{ad}} \times \mathbb{W}\ , \ \ \forall r \in \partial_u j^C(u, w)\ , \ \ \|r\| \leq c_1 \|u\| + c_2\ . \\ \exists d_1, d_2 > 0\ , \ \ \forall (u, w) \in U^{\mathrm{ad}} \times \mathbb{W}\ , \ \ \forall s \in \partial_u j^\Sigma(u, w)\ , \ \ \|s\| \leq d_1 \|u\| + d_2\ . \end{cases}$$

(A12) The objective function J is coercive on U^{ad}.

(A13) The function K is b-strongly convex for $b > 0$, meaning that for all $u, v \in \mathbb{U}$:

$$K(v) \geq K(u) + \left\langle \nabla K(u)\, , v - u \right\rangle + \frac{b}{2} \|u - v\|^2\ ,$$

and ∇K is L_K-Lipschitz continuous with $L_K > 0$, that is, for all $u, v \in \mathbb{U}$:

$$\|\nabla K(v) - \nabla K(u)\|_\star \leq L_K \|v - u\|\ ,$$

where $\|\cdot\|_\star$ is the dual norm on $\mathbb{U}^\star$.

(A14) The sequence of step sizes $\{\varepsilon_k\}_{k\in\mathbb{N}}$, with $\varepsilon_k > 0$ for all k, satisfies $\sum_k \varepsilon_k = +\infty$ and $\sum_k \varepsilon_k^2 < +\infty$.

(A15) Each $\boldsymbol{R}_k$ is measurable with respect to $\mathcal{F}_{k+1}$, the sequence of random variables $\left\{\boldsymbol{R}_k\right\}_{k\in\mathbb{N}}$ is $\mathbb{P}$-almost surely ($\mathbb{P}$-a.s.) bounded,[4] and we have:

$$\sum_{k\in\mathbb{N}} \varepsilon_k \mathbb{E}\big(\left\|\boldsymbol{R}_k\right\| \ \big|\ \mathcal{F}_k \big) < +\infty \quad \mathbb{P}\text{-a.s.}$$

[3] In this expression, the $\in$ relationship is to be understood ω *by* ω.

[4] The set $\left\{\omega \in \Omega,\ \{\boldsymbol{R}_k(\omega)\}_{k\in\mathbb{N}} \text{ is unbounded}\right\}$ is negligible.

(A16) For all integers $k \geq 1$, the integrand $\tau_k : \mathbb{U} \times \Omega \to \mathbb{R}$ defined by $\tau_k(u, \omega) = (j^C + j^\Sigma)(u, \boldsymbol{W}_k(\omega))$ for all $(u, \omega) \in \mathbb{U} \times \Omega$ is $\mathcal{A}$-quasi-integrable [35]. That is, for each $k \geq 1$, there exists an integrable mapping $\psi_k : \Omega \to \mathbb{R}$ such that $\psi_k \leq \tau_k(u, \cdot)$ for all $u \in \mathbb{U}$.

We make some comments on Assumptions (A11)-(A16):

- Assumption (A11) is a relaxation of the standard assumption of bounded gradients, used in [25] for example.
- Assumption (A12) is used to ensure the existence of solutions to Problem (1) when U^{ad} is an unbounded domain. When U^{ad} is bounded, Assumption (A12) is automatically satisfied.
- Assumption (A13) is related to the user-defined function K and not to the intrinsic characteristics of the minimization problem, thus it is not restrictive.
- Assumptions (A14) and (A15) are standard to ensure the convergence of SA schemes. In particular, (A15) ensures that the noise on the gradient vanishes sufficiently fast.
- Assumption (A16) is a technical assumption used to ensure that the conditional expectation of the integrand τ_k is defined. As a sufficient condition, assuming that $j^C + j^\Sigma$ is nonnegative would ensure Assumption (A16).

Assumptions (A1)-(A3), (A9) and (A12) ensure that J is well-defined, convex, l.s.c., coercive and attains its minimum on U^{ad}. Hence, Problem (1) has a non-empty set of solutions $U^\sharp$. From now on, K is supposed to be b-strongly convex, so by Corollary 3.28, the problem solved at each iteration k of the stochastic APP algorithm admits a unique solution $\boldsymbol{U}_{k+1}$, which is measurable.

We start by a technical lemma which gives a key inequality that will be used for the proof of convergence of the stochastic APP algorithm in Theorem 4.2 and to derive efficiency estimates in Theorems 4.5 and 4.7.

Lemma 4.1 *Let $v \in U^{\mathrm{ad}}$ and consider the Lyapunov function:*

$$\ell_v(u) = K(v) - K(u) - \langle \nabla K(u), v - u \rangle, \quad u \in U^{\mathrm{ad}}. \tag{16}$$

Let $\{u_k\}_{k\in\mathbb{N}}$ be the sequence of iterates of Algorithm 1 corresponding to the realization $\{w_k\}_{k\in\mathbb{N}}$ of the stochastic process $\{\boldsymbol{W}_k\}_{k\in\mathbb{N}}$. Then, under Assumptions (A9), (A11) and (A13), there exist constants $\alpha, \beta, \gamma, \delta > 0$ such that, for all $k \in \mathbb{N}$:

$$\begin{aligned}\ell_v(u_{k+1}) \leq \Big(1 + \alpha\varepsilon_k^2 + \frac{2}{b}\varepsilon_k\|r_k\|\Big)\ell_v(u_k) + \beta\varepsilon_k^2\ell_v(u_{k+1}) \\ + \Big(\gamma\varepsilon_k^2 + \varepsilon_k\|r_k\| + \delta(\varepsilon_k\|r_k\|)^2\Big) \\ + \varepsilon_k\big((j^C + j^\Sigma)(v, w_{k+1}) - (j^C + j^\Sigma)(u_k, w_{k+1})\big),\end{aligned} \tag{17}$$

where we recall that $b > 0$ is the strong convexity constant of K, ε_k is the step size and r_k is the additive error on the stochastic gradient at iteration k of the algorithm.

Proof. By (A13), K is b-strongly convex implying that:

$$\frac{b}{2}\|u-v\|^2 \leq \ell_v(u)\ . \tag{18}$$

This shows that ℓ_v is lower bounded and coercive. Let $k \in \mathbb{N}$, as u_{k+1} is solution of (6), it solves the following variational inequality, characterizing the minimum of the sum of a Gateaux-differentiable and a non-differentiable function [14, Chapter II, Proposition 2.2]: for all $u \in U^{\mathrm{ad}}$,

$$\begin{aligned}\big\langle \nabla K(u_{k+1}) - \nabla K(u_k) + \varepsilon_k(g_k + r_k)\,, u - u_{k+1}\big\rangle \\ + \varepsilon_k(j^{\Sigma}(u, w_{k+1}) - j^{\Sigma}(u_{k+1}, w_{k+1})) \geq 0\ .\end{aligned} \tag{19}$$

We have:

$$\begin{aligned}\ell_v(u_{k+1}) - \ell_v(u_k) = \underbrace{K(u_k) - K(u_{k+1}) - \big\langle \nabla K(u_k)\,, u_k - u_{k+1}\big\rangle}_{T_1} \\ + \underbrace{\big\langle \nabla K(u_k) - \nabla K(u_{k+1})\,, v - u_{k+1}\big\rangle}_{T_2}\ .\end{aligned} \tag{20}$$

As K is convex (A13), we get $T_1 \leq 0$. The optimality condition (19) at $u = v$ implies:

$$\begin{aligned}T_2 &\leq \varepsilon_k\big\langle g_k + r_k\,, v - u_{k+1}\big\rangle + \varepsilon_k\big(j^{\Sigma}(v, w_{k+1}) - j^{\Sigma}(u_{k+1}, w_{k+1})\big) \\ &\leq \varepsilon_k\Big(\underbrace{\big\langle g_k\,, v - u_k\big\rangle + j^{\Sigma}(v, w_{k+1}) - j^{\Sigma}(u_k, w_{k+1})}_{T_3} + \underbrace{\big\langle r_k\,, v - u_k\big\rangle}_{T_4} \\ &\qquad + \underbrace{\big\langle g_k + r_k\,, u_k - u_{k+1}\big\rangle + j^{\Sigma}(u_k, w_{k+1}) - j^{\Sigma}(u_{k+1}, w_{k+1})}_{T_5}\Big)\ .\end{aligned}$$

- As $j^C(\cdot, w_{k+1})$ is convex (A9), we get:

$$T_3 \leq \big(j^C + j^{\Sigma}\big)(v, w_{k+1}) - \big(j^C + j^{\Sigma}\big)(u_k, w_{k+1})\ .$$

- By Cauchy-Schwarz inequality, using $a \leq a^2 + 1$ for $a \geq 0$ and (18), we get:

$$T_4 \leq \|r_k\|\|v - u_k\| \leq \|r_k\|\big(\|v - u_k\|^2 + 1\big) \leq \|r_k\| + \frac{2}{b}\ell_v(u_k)\|r_k\|\ .$$

- The optimality condition (19) at $u = u_k$ and the strong monotonicity of ∇K, that arises from (A13), imply:

$$\begin{aligned}b\big\|u_{k+1} - u_k\big\|^2 \leq \varepsilon_k\big(\langle g_k + r_k\,, u_k - u_{k+1}\rangle \\ + j^{\Sigma}(u_k, w_{k+1}) - j^{\Sigma}(u_{k+1}, w_{k+1})\big)\ ,\end{aligned} \tag{21}$$

  where we recognize $\varepsilon_k T_5$ as the right-hand side. Using the linearly bounded subgradient property of j^{Σ} (A11) with the result of Proposition A.4, we get:

$$\begin{aligned}\big|j^{\Sigma}(u_k, w_{k+1}) - j^{\Sigma}(u_{k+1}, w_{k+1})\big| \\ &\leq \Big(d_1 \max\big\{\|u_k\|, \|u_{k+1}\|\big\} + d_2\Big)\|u_k - u_{k+1}\|\ , \\ &\leq \Big(d_1\big(\|u_k\| + \|u_{k+1}\|\big) + d_2\Big)\|u_k - u_{k+1}\|\ .\end{aligned}$$

With Cauchy-Schwarz inequality on the first term of T_5, we have:

$$T_5 \leq \|g_k + r_k\| \|u_k - u_{k+1}\| + \big(d_1\|u_k\| + d_1\|u_{k+1}\| + d_2\big)\|u_k - u_{k+1}\| \,.$$

By the triangular inequality and Assumption (A11) for j^C, we deduce that there exist positive constants e_1, e_2 and e_3 such that:

$$T_5 \leq \big(e_1\|u_k\| + e_2\|u_{k+1}\| + e_3 + \|r_k\|\big)\|u_{k+1} - u_k\| \,.$$

By Inequality (21), we then get:

$$\big\|u_{k+1} - u_k\big\| \leq \frac{\varepsilon_k}{b}\big(e_1\|u_k\| + e_2\|u_{k+1}\| + e_3 + \|r_k\|\big) \,, \tag{22}$$

and therefore by a repeated use of $(a+b)^2 \leq 2(a^2+b^2)$,

$$T_5 \leq \frac{4\varepsilon_k}{b}\big(e_1^2\,\|u_k\|^2 + e_2^2\,\|u_{k+1}\|^2 + e_3^2 + \|r_k\|^2\big) \,.$$

We bound $\|u_k\|$ (resp. $\|u_{k+1}\|$) by $\|u_k - v\| + \|v\|$ (resp. $\|u_{k+1} - v\| + \|v\|$) and use (18) to infer that there exist positive constants α, β, γ, δ such that:

$$T_5 \leq \varepsilon_k\big(\alpha\ell_v(u_k) + \beta\ell_v(u_{k+1}) + \gamma + \delta\|r_k\|^2\big) \,.$$

Collecting the bounds for T_1, T_3, T_4 and T_5, we get the desired result. □

When no bias is present, $r_k = 0$, we retrieve the same inequality as in [12, §2.5.1]. In the proofs of the subsequent theorems, Inequality (17) is fundamental to derive boundedness properties or convergence results for the Lyapunov function ℓ_v.

We give convergence results for the stochastic APP algorithm, in terms of function values as well as for the iterates. The proof is similar to that in [12, 13] (case of a Hilbert space, no bias considered). The assumption that the Banach $\mathbb{U}$ is reflexive (A1) allows for a similar treatment as in the Hilbert case. The additional contribution of the bias is already taken care of by Inequality (17). We denote by $J^\sharp$ the value of J on the non-empty set of solutions $U^\sharp$ of Problem (1).

Theorem 4.2 *Under Assumptions (A1)-(A7), (A9)-(A16), we have the following statements:*

- *The sequence of random variables* $\big\{J(\boldsymbol{U}_k)\big\}_{k\in\mathbb{N}}$ *converges to* $J^\sharp$ *almost surely.*
- *The sequence of iterates* $\big\{\boldsymbol{U}_k\big\}_{k\in\mathbb{N}}$ *of the stochastic APP algorithm is almost surely bounded and every weak cluster point of a bounded realization of this sequence belongs to the optimal set* $U^\sharp$.

Proof. Let $u^\sharp \in U^\sharp$ be a solution of Problem (1).

1. **Upper bound on the variation of the Lyapunov function.** We write the inequality of Lemma 4.1 at $v = u^\sharp$ in terms of random variables and reorganize the terms as follows:

$$\underbrace{(1-\beta\varepsilon_k^2)\ell_{u^\sharp}(\boldsymbol{U}_{k+1})}_{\boldsymbol{A}_{k+1}} \le \underbrace{\Big(1+\alpha\varepsilon_k^2+\frac{2}{b}\varepsilon_k\|\boldsymbol{R}_k\|\Big)\ell_{u^\sharp}(\boldsymbol{U}_k)}_{\boldsymbol{B}_k} + \underbrace{\Big(\gamma\varepsilon_k^2+\varepsilon_k\|\boldsymbol{R}_k\|+\delta(\varepsilon_k\|\boldsymbol{R}_k\|)^2\Big)}_{\boldsymbol{C}_k} - \underbrace{\varepsilon_k\big((j^C+j^\Sigma)(\boldsymbol{U}_k,\boldsymbol{W}_{k+1})-(j^C+j^\Sigma)(u^\sharp,\boldsymbol{W}_{k+1})\big)}_{\boldsymbol{D}_k}\,, \quad (23)$$

this last inequality being valid $\mathbb{P}$-a.s.. We assume without loss of generality that $(1-\beta\varepsilon_k^2)>0$ as it is true for k large enough. As $\boldsymbol{D}_k$ takes only finite values since j^C+j^Σ takes finite values, we obtain from Equation (23) that almost surely $\boldsymbol{A}_{k+1}+\boldsymbol{D}_k \le \boldsymbol{B}_k+\boldsymbol{C}_k$. It is classical to define extended conditional expectation for nonnegative random variables or more generally for $\mathcal{A}$-quasi-integrable random variables [35, p. 339]. Thus the (extended) conditional expectation with respect to $\mathcal{F}_k$ is well defined for each of the three terms $\boldsymbol{A}_{k+1}$, $\boldsymbol{B}_k$, $\boldsymbol{C}_k$. Moreover, as $\boldsymbol{B}_k$ and $\boldsymbol{C}_k$ are both nonnegative, we have that $\mathbb{E}\big(\boldsymbol{B}_k+\boldsymbol{C}_k \,\big|\, \mathcal{F}_k\big) = \mathbb{E}\big(\boldsymbol{B}_k \,\big|\, \mathcal{F}_k\big)+\mathbb{E}\big(\boldsymbol{C}_k \,\big|\, \mathcal{F}_k\big)$. We now prove that the conditional expectation of $\boldsymbol{D}_k$ with respect to $\mathcal{F}_k$ exists and satisfies $\mathbb{E}\big(\boldsymbol{D}_k \,\big|\, \mathcal{F}_k\big) = \varepsilon_k(J(\boldsymbol{U}_k)-J(u^\sharp))$. For that purpose, we consider the integrand $\tau_k : \mathbb{U}\times\Omega \to \mathbb{R}$ defined by $\tau_k(u,\omega) = (j^C+j^\Sigma)(u,\boldsymbol{W}_{k+1}(\omega))$ for all $(u,\omega)\in\mathbb{U}\times\Omega$ and recall some results from [35]. The integrand τ_k is $\mathcal{A}$-quasi-integrable (by Assumption (A16)) and l.s.c. (using the assumptions on j^C and j^Σ). Therefore, there exists an $\mathcal{A}$-quasi-integrable integrand $\tau_k^{\mathcal{F}_k}$ which gives the conditional expectation with respect to $\mathcal{F}_k$ of the integrand τ_k, that is $\tau_k^{\mathcal{F}_k}(u,\omega) = \mathbb{E}\big(\tau_k(u,\cdot) \,\big|\, \mathcal{F}_k\big)(\omega)$ for all $(u,\omega)\in\mathbb{U}\times\Omega$ [35, Proposition 12]. We have

$$\tau_k^{\mathcal{F}_k}(u,\omega) = \mathbb{E}\big(\tau_k(u,\cdot) \,\big|\, \mathcal{F}_k\big)(\omega) = \mathbb{E}\big((j^C+j^\Sigma)(u,\boldsymbol{W}_{k+1}) \,\big|\, \mathcal{F}_k\big)(\omega)\,,$$

where the conditional expectation is an expectation since $\boldsymbol{W}_{k+1}$ is independent of the σ-field $\mathcal{F}_k$, so that

$$\tau_k^{\mathcal{F}_k}(u,\omega) = \mathbb{E}\big((j^C+j^\Sigma)(u,\boldsymbol{W}_{k+1})\big) = J(u)\,,$$

that is, $\tau_k^{\mathcal{F}_k}$ does not depend on ω. Moreover, given any $\mathcal{F}_k$-measurable random variable $\boldsymbol{Y}$ we have [35, Proposition 13] that

$$\tau_k^{\mathcal{F}_k}(\boldsymbol{Y}(\omega),\omega) = \mathbb{E}\big(\tau_k(\boldsymbol{Y}(\cdot),\cdot) \,\big|\, \mathcal{F}_k\big)(\omega)\,,$$

which, using the fact that $\boldsymbol{U}_k$ is $\mathcal{F}_k$-measurable, gives for all $\omega\in\Omega$ that

$$\begin{aligned}
\varepsilon_k\big(J(\boldsymbol{U}_k(\omega))-J(u^\sharp)\big) &= \varepsilon_k\big(\tau_k^{\mathcal{F}_k}(\boldsymbol{U}_k(\omega),\omega)-\tau_k^{\mathcal{F}_k}(u^\sharp,\omega)\big)\\
&= \varepsilon_k\big(\mathbb{E}\big(\tau_k(\boldsymbol{U}_k(\cdot),\cdot) \,\big|\, \mathcal{F}_k\big)(\omega)-\mathbb{E}\big(\tau_k(u^\sharp,\cdot) \,\big|\, \mathcal{F}_k\big)(\omega)\big)\\
&= \varepsilon_k\mathbb{E}\big(\tau_k(\boldsymbol{U}_k(\cdot),\cdot)-\tau_k(u^\sharp,\cdot) \,\big|\, \mathcal{F}_k\big)(\omega)\\
&\qquad\qquad (\text{as } \mathbb{E}\big(\tau_k(u^\sharp,\cdot) \,\big|\, \mathcal{F}_k\big) = J(u^\sharp) \text{ is finite})\\
&= \varepsilon_k\mathbb{E}\big((j^C+j^\Sigma)(\boldsymbol{U}_k,\boldsymbol{W}_{k+1})-(j^C+j^\Sigma)(u^\sharp,\boldsymbol{W}_{k+1}) \,\big|\, \mathcal{F}_k\big)(\omega)\\
&= \mathbb{E}\big(\boldsymbol{D}_k \,\big|\, \mathcal{F}_k\big)(\omega)\,.
\end{aligned}$$

Thus, the conditional expectations with respect to $\mathcal{F}_k$ of $\boldsymbol{A}_{k+1}$, $\boldsymbol{B}_k$, $\boldsymbol{C}_k$ and $\boldsymbol{D}_k$ are properly defined. To conclude it remains to show that $\mathbb{E}\big(\boldsymbol{A}_{k+1} + \boldsymbol{D}_k \;\big|\; \mathcal{F}_k\big) = \mathbb{E}\big(\boldsymbol{A}_{k+1} \;\big|\; \mathcal{F}_k\big) + \mathbb{E}\big(\boldsymbol{D}_k \;\big|\; \mathcal{F}_k\big)$ which is the case as $\mathbb{E}\big(\boldsymbol{D}_k \;\big|\; \mathcal{F}_k\big) = \varepsilon_k\big(J(\boldsymbol{U}_k) - J(u^\sharp)\big)$ is almost surely finite and nonnegative. Then we can take the conditional expectation on both sides of $\boldsymbol{A}_{k+1} + \boldsymbol{D}_k \leq \boldsymbol{B}_k + \boldsymbol{C}_k$ with respect to the σ-field $\mathcal{F}_k$ and use the just proved additivity properties to obtain

$$(1 - \boldsymbol{\beta}_k)\mathbb{E}\big(\ell_{u^\sharp}\big(\boldsymbol{U}_{k+1}\big) \;\big|\; \mathcal{F}_k\big) + \varepsilon_k\big(J(\boldsymbol{U}_k) - J(u^\sharp)\big) \leq (1 + \boldsymbol{\alpha}_k)\ell_{u^\sharp}\big(\boldsymbol{U}_k\big) + \boldsymbol{\gamma}_k \;, \tag{24}$$

where we have:

$$\begin{aligned}
\boldsymbol{\alpha}_k &= \alpha\varepsilon_k^2 + \frac{2}{b}\varepsilon_k\mathbb{E}\big(\|\boldsymbol{R}_k\| \;\big|\; \mathcal{F}_k\big) \;, \\
\boldsymbol{\beta}_k &= \beta\varepsilon_k^2 \;, \\
\boldsymbol{\gamma}_k &= \gamma\varepsilon_k^2 + \varepsilon_k\mathbb{E}\big(\|\boldsymbol{R}_k\| \;\big|\; \mathcal{F}_k\big) + \delta\left(\varepsilon_k\mathbb{E}\big(\|\boldsymbol{R}_k\| \;\big|\; \mathcal{F}_k\big)\right)^2 \;,
\end{aligned}$$

and where we have used that $\boldsymbol{U}_k$ is $\mathcal{F}_k$-measurable, to obtain that $\mathbb{E}\big(\ell_{u^\sharp}\big(\boldsymbol{U}_k\big) \;\big|\; \mathcal{F}_k\big) = \ell_{u^\sharp}\big(\boldsymbol{U}_k\big)$. By Assumptions (A14) and (A15), $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ are the terms of convergent series. Recall that $J(\boldsymbol{U}_k) - J(u^\sharp)$ is almost surely nonnegative as $u^\sharp$ is solution of (1). The right hand side of Equation (24) is almost surely finite and since the left hand side is the sum of two positive terms, each of them is almost surely finite. Thus we also have almost surely that

$$\begin{aligned}
\mathbb{E}\big(\ell_{u^\sharp}\big(\boldsymbol{U}_{k+1}\big) \;\big|\; \mathcal{F}_k\big) \leq (1 + \boldsymbol{\alpha}_k)\ell_{u^\sharp}\big(\boldsymbol{U}_k\big) + \boldsymbol{\beta}_k\mathbb{E}\big(\ell_{u^\sharp}\big(\boldsymbol{U}_{k+1}\big) \;\big|\; \mathcal{F}_k\big) \\
+ \boldsymbol{\gamma}_k - \varepsilon_k\big(J(\boldsymbol{U}_k) - J(u^\sharp)\big) \;.
\end{aligned} \tag{25}$$

2. **Convergence analysis.** Applying Corollary A.3 of Robbins-Siegmund theorem, we get that the sequence of random variables $\big\{\ell_{u^\sharp}\big(\boldsymbol{U}_k\big)\big\}_{k\in\mathbb{N}}$ converges $\mathbb{P}$-a.s. to a finite random variable $\boldsymbol{\ell}_{u^\sharp}^\infty$ and we have:

$$\sum_{k=0}^{+\infty} \varepsilon_k\big(J(\boldsymbol{U}_k) - J(u^\sharp)\big) < +\infty \;\; \mathbb{P}\text{-a.s.} \;. \tag{26}$$

3. **Limits of sequences.** The sequence $\big\{\ell_{u^\sharp}(\boldsymbol{U}_k)\big\}_{k\in\mathbb{N}}$ is $\mathbb{P}$-a.s. bounded, so by (18), we get that the sequence $\big\{\boldsymbol{U}_k\big\}_{k\in\mathbb{N}}$ is also $\mathbb{P}$-a.s. bounded. Assumption (A11) then implies that the sequence $\big\{\boldsymbol{G}_k\big\}_{k\in\mathbb{N}}$ is also $\mathbb{P}$-a.s. bounded. Finally, as the sequence $\big\{\boldsymbol{R}_k\big\}_{k\in\mathbb{N}}$ is assumed to be $\mathbb{P}$-a.s. bounded (A15), we deduce from (22) that the sequence $\big\{\|\boldsymbol{U}_{k+1} - \boldsymbol{U}_k\|/\varepsilon_k\big\}_{k\in\mathbb{N}}$ is also $\mathbb{P}$-a.s. bounded. This last property ensures that Assumption (c) of Proposition A.5, is satisfied. Assumption (b) of Proposition A.5 is exactly (26) and Assumption (a) is satisfied as we have (A14). On a bounded set containing the sequence $\big\{\boldsymbol{U}_k\big\}_{k\in\mathbb{N}}$, for instance the convex hull of this sequence, the function J is Lipschitz continuous by Proposition A.4. This ensures the continuity assumption required to apply Proposition A.5. We conclude that $\big\{J(\boldsymbol{U}_k)\big\}_{k\in\mathbb{N}}$ converges almost surely to $J(u^\sharp) = J^\sharp$, the optimal value of Problem (1).

Let Ω_0 be the negligible subset of Ω on which the sequence $\big\{\ell_{u^\sharp}\big(\boldsymbol{U}_k\big)\big\}_{k\in\mathbb{N}}$ is unbounded and Ω_1 be the negligible subset of Ω on which the relation (26) is not satisfied. We have $\mathbb{P}\big(\Omega_0 \cup \Omega_1\big) = 0$. Let $\omega \notin \Omega_0 \cup \Omega_1$. The sequence $\{u_k\}_{k\in\mathbb{N}}$ associated to this element ω is bounded and each u_k is in U^{ad}, a closed subset of $\mathbb{U}$. As $\mathbb{U}$ is reflexive (A1), there exists a weakly converging subsequence

$\{u_{\xi(k)}\}_{k\in\mathbb{N}}$. Note that $\{\xi(k)\}_{k\in\mathbb{N}}$ depends on ω. Let $\overline{u}$ be the weak limit of the sequence $\{u_{\xi(k)}\}_{k\in\mathbb{N}}$. The function J is l.s.c. and convex, it is then weakly l.s.c. by [14, Corollary 2.2]. Thus we have:

$$J(\overline{u}) \leq \liminf_{k\to+\infty} J(u_{\xi(k)}) = J(u^\sharp) \ .$$

We conclude that $\overline{u} \in U^\sharp$. □

When the differential of K is weakly continuous, we can prove stronger convergence results for the sequence of iterates of the stochastic APP algorithm. These results already appear in [12] and remain valid for our more general version of the algorithm.

Theorem 4.3 *Consider again (A1)-(A7), (A9)-(A15) and suppose that the differential of K is weakly continuous. Then, the sequence of iterates $\{\boldsymbol{U}_k\}$ converges weakly $\mathbb{P}$-a.s. to a single element of $U^\sharp$. If moreover, the function J^C is strongly convex, then, the sequence of iterates $\{\boldsymbol{U}_k\}$ converges strongly $\mathbb{P}$-a.s. to the unique solution $u^\sharp$ of Problem (1).*

Proof. Consider the case where the differential of K is weakly continuous. Let $\{u_k\}_{k\in\mathbb{N}}$ be a sequence generated by the algorithm. Suppose that there exist two subsequences $\{u_{\xi(k)}\}_{k\in\mathbb{N}}$ and $\{u_{\psi(k)}\}_{k\in\mathbb{N}}$ converging weakly respectively to two solutions $\overline{u}_\xi$ and $\overline{u}_\psi$ of the problem, with $\overline{u}_\xi \neq \overline{u}_\psi$. Then we have:

$$\begin{aligned} K(\overline{u}_\psi) - K(u_{\xi(k)}) - \left\langle \nabla K(u_{\xi(k)}) \, , \overline{u}_\psi - u_{\xi(k)} \right\rangle = K(\overline{u}_\psi) - K(\overline{u}_\xi) - \left\langle \nabla K(u_{\xi(k)}) \, , \overline{u}_\psi - \overline{u}_\xi \right\rangle \\ + \left(K(\overline{u}_\xi) - K(u_{\xi(k)}) - \left\langle \nabla K(u_{\xi(k)}) \, , \overline{u}_\xi - u_{\xi(k)} \right\rangle \right) . \quad (27) \end{aligned}$$

By the point 2 of the proof of Theorem 4.2,

$$\begin{aligned} \lim_{k\to+\infty} K(\overline{u}_\psi) - K(u_{\xi(k)}) - \left\langle \nabla K(u_{\xi(k)}) \, , \overline{u}_\psi - u_{\xi(k)} \right\rangle = \lim_{k\to+\infty} \ell_{\overline{u}_\psi}(u_k) = \ell_{\overline{u}_\psi} \ , \\ \lim_{k\to+\infty} K(\overline{u}_\xi) - K(u_{\xi(k)}) - \left\langle \nabla K(u_{\xi(k)}) \, , \overline{u}_\xi - u_{\xi(k)} \right\rangle = \lim_{k\to+\infty} \ell_{\overline{u}_\xi}(u_k) = \ell_{\overline{u}_\xi} \ , \end{aligned}$$

therefore by weak continuity of ∇K and strong convexity of K, we get:

$$\ell_{\overline{u}_\psi} - \ell_{\overline{u}_\xi} = K(\overline{u}_\psi) - K(\overline{u}_\xi) - \left\langle \nabla K(\overline{u}_\xi) \, , \overline{u}_\psi - \overline{u}_\xi \right\rangle \geq \frac{b}{2} \|\overline{u}_\xi - \overline{u}_\psi\|^2 \ .$$

Inverting the roles of $\overline{u}_\psi$ and $\overline{u}_\xi$, by a similar calculation as previously we get:

$$\ell_{\overline{u}_\xi} - \ell_{\overline{u}_\psi} \geq \frac{b}{2} \|\overline{u}_\xi - \overline{u}_\psi\|^2 \ ,$$

We then deduce that $\overline{u}_\xi = \overline{u}_\psi$, which contradicts the initial assumption. We conclude that all weakly converging subsequences of the sequence $\{u_k\}$ converge to the same limit, hence we have the weak convergence of the whole sequence $\{u_k\}$ to a single element of $U^\sharp$.

Now consider the case where J^C is strongly convex, with constant a. Then, Problem (1) admits a unique solution $u^\sharp$ which is characterized by the following variational inequality:

$$\exists r^\sharp \in \partial J^C(u^\sharp) \ , \ \ \forall u \in U^{\mathrm{ad}} \ , \ \ \left\langle r^\sharp \, , u - u^\sharp \right\rangle + J^\Sigma(u) - J^\Sigma(u^\sharp) \geq 0 \ .$$

The strong convexity assumption on J^C yields:

$$J(\boldsymbol{U}_k) - J(u^\sharp) \geq \langle r^\sharp \, , \boldsymbol{U}_k - u^\sharp \rangle + \frac{a}{2}\|\boldsymbol{U}_k - u^\sharp\|^2 + J^\Sigma(\boldsymbol{U}_k) - J^\Sigma(u^\sharp) \geq \frac{a}{2}\|\boldsymbol{U}_k - u^\sharp\|^2 \, .$$

As $\left\{J(\boldsymbol{U}_k)\right\}_{k\in\mathbb{N}}$ converges almost surely to $J(u^\sharp)$, we get that $\|\boldsymbol{U}_k - u^\sharp\|$ converges to zero. Thus, we have the strong convergence of the sequence $\left\{\boldsymbol{U}_k\right\}_{k\in\mathbb{N}}$ to the unique solution $u^\sharp$ of the problem. □

4.2 Efficiency estimates

In this section, we derive efficiency estimates for the convergence of the expectation of function values. In Theorem 4.5, we consider the expected function value taken for the averaged iterates following the technique of Polyak-Ruppert [29, 34]. We take a step size ε_k of the order $\mathcal{O}\left(k^{-\theta}\right)$ with $1/2 < \theta < 1$, ensuring the convergence of the algorithm, and leading to a better convergence rate than with a small step size $\varepsilon_k = \mathcal{O}\left(k^{-1}\right)$. The efficiency estimate is obtained using a similar technique as in [25] but without requiring the boundedness of U^{ad}. Moreover, we are able to take into account the bias on the gradient with the following assumption, inspired from [16]:

(A17) For $k \in \mathbb{N}$, let $Q_k = \operatorname{ess\,sup}_{\omega\in\Omega} \left\|\boldsymbol{R}_k(\omega)\right\|$ be the essential supremum of $\left\|\boldsymbol{R}_k\right\|$ and assume that:

$$\sum_{k\in\mathbb{N}} Q_k \varepsilon_k < \infty \, .$$

We start by a lemma that proves the boundedness of the expectation of the Lyapunov function. This result will be used multiple times in this part.

Lemma 4.4 *Under Assumptions (A1)-(A7), (A9)-(A17), the sequence of expectations of the Lyapunov function* $\left\{\mathbb{E}\left(\ell_{u^\sharp}\left(\boldsymbol{U}_k\right)\right)\right\}_{k\in\mathbb{N}}$ *is bounded and the sequence* $\left\{\mathbb{E}\left(J(\boldsymbol{U}_k)\right)\right\}_{k\in\mathbb{N}}$ *takes finite values.*

Proof. From Inequality (25), using the fact that $\left\|\boldsymbol{R}_k\right\| \leq Q_k$ almost surely and using the fact that $J(\boldsymbol{U}_k) - J(u^\sharp)$ is nonnegative, we obtain

$$\mathbb{E}\left(\ell_{u^\sharp}\left(\boldsymbol{U}_{k+1}\right) \mid \mathcal{F}_k\right) \leq (1+\alpha_k)\ell_{u^\sharp}\left(\boldsymbol{U}_k\right) + \beta_k \mathbb{E}\left(\ell_{u^\sharp}\left(\boldsymbol{U}_{k+1}\right) \mid \mathcal{F}_k\right) + \gamma_k \, , \tag{28}$$

where

$$\alpha_k = \alpha\varepsilon_k^2 + \frac{2}{b}\varepsilon_k Q_k \, , \quad \beta_k = \beta\varepsilon_k^2 \, , \quad \gamma_k = (\gamma + \delta Q_k^2)\varepsilon_k^2 + Q_k\varepsilon_k \, . \tag{29}$$

Then, taking the extended expectation (all random variables are nonnegative) on both sides of Inequality (28) yields:

$$\mathbb{E}\left(\ell_{u^\sharp}\left(\boldsymbol{U}_{k+1}\right)\right) \leq (1+\alpha_k)\mathbb{E}\left(\ell_{u^\sharp}\left(\boldsymbol{U}_k\right)\right) + \beta_k \mathbb{E}\left(\ell_{u^\sharp}\left(\boldsymbol{U}_{k+1}\right)\right) + \gamma_k \, .$$

From (A14) and (A17), α_k, β_k and γ_k are the terms of convergent series. Using a deterministic version of Corollary A.3, we get that the sequence $\left\{\mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_k)\right)\right\}_{k\in\mathbb{N}}$ converges and is bounded.

Now, we can use again Inequality (25) and the fact that, from the previous point, $\mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_{k+1})\right)$ is finite to obtain that

$$\varepsilon_k \mathbb{E}\left(J(\boldsymbol{U}_k) - J(u^\sharp)\right) \leq (1+\alpha_k)\mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_k)\right) + (\beta_k - 1)\mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_{k+1})\right) + \gamma_k \;, \tag{30}$$

from which we obtain in particular that $\mathbb{E}\left(J(\boldsymbol{U}_k) - J(u^\sharp)\right)$ is finite. □

Theorem 4.5 *Suppose that Assumptions (A1)-(A7), (A9)-(A17) are satisfied. Let $n \in \mathbb{N}$ and let $\left\{\boldsymbol{U}_k\right\}_{k\in\mathbb{N}}$ be the sequence of iterates of the stochastic APP algorithm. Define the averaged iterate as:*

$$\widetilde{\boldsymbol{U}}_i^n = \sum_{k=i}^n \eta_k^i \boldsymbol{U}_k \quad \textit{with} \quad \eta_k^i = \frac{\varepsilon_k}{\sum_{l=i}^n \varepsilon_l} \;.$$

Suppose that for all $k \in \mathbb{N}$, $\varepsilon_k = ck^{-\theta}$ with $1/2 < \theta < 1$ and a constant $c > 0$. Then for any minimizer $u^\sharp$ of J, we have:

$$\mathbb{E}\left(J\left(\widetilde{\boldsymbol{U}}_1^n\right) - J\left(u^\sharp\right)\right) = \mathcal{O}\left(n^{\theta-1}\right) \;.$$

In particular, the rate of convergence can be arbitrarily close to the order $n^{-1/2}$ if θ is chosen to be arbitrarily close to $1/2$.

Proof. From Lemma 4.4, we get that Inequality (30) is satisfied and the sequence $\left\{\mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_k)\right)\right\}_{k\in\mathbb{N}}$ is bounded. Then, there exists a constant $M \geq 0$ such that $\mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_k)\right) \leq M$ for all $k \in \mathbb{N}$. Summing (30) over $i \leq k \leq n$ and using $\mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_k)\right) \leq M$, we get:

$$\sum_{k=i}^n \varepsilon_k \mathbb{E}\left(J\left(\boldsymbol{U}_k\right) - J\left(u^\sharp\right)\right) \leq M + \sum_{k=i}^n \left(M(\alpha+\beta) + \gamma + \delta Q_k^2\right)\varepsilon_k^2 + \left(\frac{2}{b}M + 1\right) Q_k \varepsilon_k \;.$$

In the sequel, let $R = M(\alpha+\beta)+\gamma$ and $S = \frac{2}{b}M + 1$. By convexity of J, we get:

$$\mathbb{E}\left(J\left(\widetilde{\boldsymbol{U}}_i^n\right) - J(u^\sharp)\right) \leq \frac{M + \sum_{k=i}^n \left(R + \delta Q_k^2\right)\varepsilon_k^2 + S Q_k \varepsilon_k}{\sum_{k=i}^n \varepsilon_k} \;.$$

We have $\varepsilon_k = ck^{-\theta}$ with $1/2 < \theta < 1$ and:

$$\sum_{k=1}^n k^{-\theta} \geq \frac{(n+1)^{1-\theta} - 1}{1-\theta} \geq \widetilde{C}_\theta n^{1-\theta} \;,$$

for some $\widetilde{C}_\theta > 0$. Moreover, from (A14) and (A17), ε_k^2, $Q_k\varepsilon_k$ and $Q_k^2\varepsilon_k^2$ are the terms of convergent series. Thus, there exists a constant $C_\theta > 0$ such that:

$$\mathbb{E}\left(J\left(\widetilde{\boldsymbol{U}}_1^n\right) - J(u^\sharp)\right) \leq \frac{C_\theta}{n^{1-\theta}} \;,$$

which gives the desired rate of convergence. □

Theorem 4.5 proves a convergence rate of order $\mathcal{O}\left(n^{\theta-1}\right)$ for the stochastic APP algorithm without assuming strong convexity of the objective. This rate appears for stochastic gradient descent in [2] where it is stated that the combination of large step sizes of order $\mathcal{O}\left(n^{-\theta}\right)$ with $1/2 < \theta < 1$, together with averaging lead to the best convergence behavior. A similar rate is also given for stochastic proximal gradient in [33].

In the following theorem, we show that this rate also holds when we consider the expected function value taken at the last iterate $\boldsymbol{U}_n$ instead of the averaged iterate $\widetilde{\boldsymbol{U}}_1^n$. Using the concept of modified Fejér monotone sequences, Lin and al. [23] have given convergence rates of the expected function value of the last iterate for many algorithms, such as the projected subgradient method or the proximal gradient algorithm. The idea of modified Fejér sequence is adapted to the stochastic case in [33, Theorem 3.1]. We further adapt this concept for the stochastic APP algorithm. Before stating the result, we need a technical lemma.

Lemma 4.6 *Under Assumptions (A1)-(A7), (A9)-(A17), there exists a constant $M' \geq 0$ such that $\mathbb{E}\left(\ell_{\boldsymbol{U}_j}\left(\boldsymbol{U}_k\right)\right) \leq M'$ for all $j, k \in \mathbb{N}$.*

Proof. As ∇K is L_K-Lipschitz continuous by (A13), we have the following inequality (see for example [27, Lemma 1.2.3])

$$K(v) \leq K(u) + \left\langle \nabla K(u), v - u \right\rangle + \frac{L_K}{2} \|u - v\|^2 ,$$

and hence, for all $u, v \in U^{\mathrm{ad}}$,

$$\ell_v(u) \leq \frac{L_K}{2}\|u - v\|^2 \leq L_K\left(\|u - u^\sharp\|^2 + \|v - u^\sharp\|^2\right) ,$$

the last inequality arising from the standard norm inequality $\|a+b\|^2 \leq 2(\|a\|^2 + \|b\|^2)$. Then using the b-strong convexity of K and Equation (18), we obtain

$$\ell_v(u) \leq \frac{2L_K}{b}\left(\ell_{u^\sharp}(u) + \ell_{u^\sharp}(v)\right) . \tag{31}$$

Writing Inequality (31) in terms of random variables, v and u being respectively replaced by $\boldsymbol{U}_j$ and $\boldsymbol{U}_k$, and taking the extended expectation (all quantities are positive) leads to:

$$\begin{aligned}\mathbb{E}\left(\ell_{\boldsymbol{U}_j}(\boldsymbol{U}_k)\right) &\leq \frac{2L_K}{b}\left(\mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_k)\right) + \mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_j)\right)\right) , \\ &\leq \frac{4L_K}{b} M ,\end{aligned}$$

since we have by Lemma 4.4 that the quantities $\mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_k)\right)$ are bounded by a constant M. □

Theorem 4.7 *Suppose that Assumptions (A1)-(A7), (A9)-(A17) are satisfied. Let $n \in \mathbb{N}$ and let $\left\{\boldsymbol{U}_k\right\}_{k\in\mathbb{N}}$ be the sequence of iterates of the stochastic APP algorithm. Suppose that*

for all $k \in \mathbb{N}$, $\varepsilon_k = ck^{-\theta}$ with $1/2 < \theta < 1$ and a constant $c > 0$. Assume also that $Q_k \leq qk^{-\nu}$ for $\nu > 1 - \theta$ and a constant $q > 0$. Then, for any minimizer $u^\sharp$ of J we have:

$$\mathbb{E}\left(J\left(\boldsymbol{U}_n\right) - J\left(u^\sharp\right)\right) = \mathcal{O}\left(n^{\theta-1}\right) \ .$$

In particular, the rate of convergence can be arbitrarily close to the order $n^{-1/2}$ if θ is chosen to be arbitrarily close to $1/2$.

Proof. For $k \in \mathbb{N}$, let $a_k = \mathbb{E}\left(J\left(\boldsymbol{U}_k\right) - J\left(u^\sharp\right)\right)$, which is finite by Lemma 4.4. Fix $n \in \mathbb{N}$, from Lemma A.1 applied to the sequence $\{\varepsilon_k a_k\}_{k\in\mathbb{N}}$ (see Appendix), we have that:

$$\varepsilon_n a_n = \frac{1}{n}\Big(\sum_{k=1}^{n} \varepsilon_k a_k\Big) + \sum_{i=1}^{n-1} \frac{1}{i(i+1)}\Bigg(\Big(\sum_{k=n-i+1}^{n} \varepsilon_k a_k\Big) - i\varepsilon_{n-i}a_{n-i}\Bigg) \ .$$

Moreover, we have that:

$$\begin{aligned}
\Big(\sum_{k=n-i+1}^{n} \varepsilon_k a_k\Big) - i\varepsilon_{n-i}a_{n-i} &= \sum_{k=n-i+1}^{n} \left(\varepsilon_k a_k - \varepsilon_{n-i}a_{n-i}\right) \\
&= \sum_{k=n-i+1}^{n} \left(\varepsilon_k(a_k - a_{n-i}) + (\varepsilon_k - \varepsilon_{n-i})a_{n-i}\right) \\
&\leq \sum_{k=n-i+1}^{n} \varepsilon_k(a_k - a_{n-i}) \\
&= \sum_{k=n-i+1}^{n} \varepsilon_k \mathbb{E}\left(J\left(\boldsymbol{U}_k\right) - J\left(\boldsymbol{U}_{n-i}\right)\right) \ ,
\end{aligned}$$

where the last inequality follows from the fact that the sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ is decreasing and that by optimality of $u^\sharp$, we have $a_{n-i} \geq 0$. We therefore obtain that:

$$\varepsilon_n a_n \leq \frac{1}{n}\Big(\sum_{k=1}^{n} \varepsilon_k a_k\Big) + \sum_{i=1}^{n-1} \frac{1}{i(i+1)} \sum_{k=n-i+1}^{n} \varepsilon_k \mathbb{E}\left(J\left(\boldsymbol{U}_k\right) - J\left(\boldsymbol{U}_{n-i}\right)\right) \ . \tag{32}$$

We bound the terms of (32):

1. From Lemma 4.4, Inequality (30) is satisfied and there exists a constant $M \geq 0$ such that $\mathbb{E}\left(\ell_{u^\sharp}\left(\boldsymbol{U}_k\right)\right) \leq M$ for all $k \in \mathbb{N}$. Summing (30) from 1 to n and using that $\beta_n - 1 \leq 0$ for n large enough, we get:

$$\begin{aligned}
\sum_{k=1}^{n} \varepsilon_k a_k &\leq M_1 + \sum_{k=2}^{n} \Big((\alpha_k + \beta_{k-1})\mathbb{E}\left(\ell_{u^\sharp}\left(\boldsymbol{U}_k\right)\right) + \gamma_k\Big) \\
&\qquad + (\beta_n - 1)\mathbb{E}\left(\ell_{u^\sharp}\left(\boldsymbol{U}_{n+1}\right)\right) \\
&\leq M_1 + \sum_{k=2}^{n} M(\alpha_k + \beta_{k-1}) + \gamma_k \ .
\end{aligned} \tag{33}$$

   with $M_1 = (1 + \alpha_1)\ell_{u^\sharp}\left(u_1\right) + \gamma_1$.

2. In order to bound the second term in the right hand side of (32), we start from Inequality (17) in Lemma 4.1 that we write in terms of random variables, v being replaced by $\boldsymbol{U}_{n-i}$. This leads to the following inequality

$$\underbrace{(1-\beta\varepsilon_k^2)\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_{k+1})}_{\boldsymbol{A}'_{k+1}} \leq \underbrace{\Big(1+\alpha\varepsilon_k^2+\frac{2}{b}\varepsilon_k\|\boldsymbol{R}_k\|\Big)\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_k)}_{\boldsymbol{B}'_k} + \underbrace{\Big(\gamma\varepsilon_k^2+\varepsilon_k\|\boldsymbol{R}_k\|+\delta(\varepsilon_k\|\boldsymbol{R}_k\|)^2\Big)}_{\boldsymbol{C}'_k} - \underbrace{\varepsilon_k\big((j^C+j^\Sigma)(\boldsymbol{U}_k,\boldsymbol{W}_{k+1})-(j^C+j^\Sigma)(\boldsymbol{U}_{n-i},\boldsymbol{W}_{k+1})\big)}_{\boldsymbol{D}'_k} . \quad (34)$$

This inequality is similar to Inequality (23), but the term $\boldsymbol{D}'_k$ cannot be assumed to be nonnegative and we need to adapt the steps used in the proof of Theorem 4.2. Using Lemma 4.6 we obtain that the terms $\boldsymbol{A}'_{k+1}$ and $\boldsymbol{B}'_k$ are integrable so that their conditional expectation is well defined. Moreover, since $k \geq n-i$, the random variable $\boldsymbol{U}_{n-i}$ is $\mathcal{F}_k$-measurable and thus $\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_k)$ is also $\mathcal{F}_k$-measurable. For the term $\boldsymbol{D}'_k$, we consider separately $\boldsymbol{D}'^{,1}_k = \varepsilon_k(j^C+j^\Sigma)(\boldsymbol{U}_k,\boldsymbol{W}_{k+1})$ and $\boldsymbol{D}'^{,2}_k = \varepsilon_k(j^C+j^\Sigma)(\boldsymbol{U}_{n-i},\boldsymbol{W}_{k+1})$. Following the steps of the proof of Theorem 4.2, we obtain that both terms $\boldsymbol{D}'^{,1}_k$ and $\boldsymbol{D}'^{,2}_k$ admit explicit (extended) conditional expectations with respect to the σ-field $\mathcal{F}_k$ which are given by $\mathbb{E}\big(\boldsymbol{D}'^{,1}_k \mid \mathcal{F}_k\big)(\omega) = \varepsilon_k J(\boldsymbol{U}_k(\omega))$ and $\mathbb{E}\big(\boldsymbol{D}'^{,2}_k \mid \mathcal{F}_k\big)(\omega) = \varepsilon_k J(\boldsymbol{U}_{n-i}(\omega))$. Now, using Lemma 4.4 we have that $J(\boldsymbol{U}_k)$ and $J(\boldsymbol{U}_{n-i})$ are both integrable and therefore the conditional expectation of $\boldsymbol{D}'_k$ is well defined as the difference of two integrable functions. We can therefore take the conditional expectation w.r.t. $\mathcal{F}_k$ on both sides of Inequality (34) and we obtain

$$\mathbb{E}\big(\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_{k+1}) \mid \mathcal{F}_k\big) \leq (1+\boldsymbol{\alpha}_k)\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_k) + \boldsymbol{\beta}_k\mathbb{E}\big(\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_{k+1}) \mid \mathcal{F}_k\big) + \boldsymbol{\gamma}_k - \varepsilon_k\big(J(\boldsymbol{U}_k)-J(\boldsymbol{U}_{n-i})\big) .$$

In this last inequality, replacing the random variables $(\boldsymbol{\alpha}_k,\boldsymbol{\beta}_k,\boldsymbol{\gamma}_k)$ by their deterministic upper bounds $(\alpha_k,\beta_k,\gamma_k)$ given by Equation (29), and taking the expectation of the left and right terms, we obtain:

$$\mathbb{E}\left(\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_{k+1})\right) \leq (1+\alpha_k)\mathbb{E}\left(\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_k)\right) + \beta_k\mathbb{E}\left(\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_{k+1})\right) + \gamma_k - \varepsilon_k\mathbb{E}\left(J\left(\boldsymbol{U}_k\right)-J\left(\boldsymbol{U}_{n-i}\right)\right) , \quad (35)$$

We have already seen by Lemma 4.6 that $\mathbb{E}\left(\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_{k+1})\right)$ is finite. We have also seen by Lemma 4.4 that $a_k = \mathbb{E}\left(J\left(\boldsymbol{U}_k\right)-J\left(u^\sharp\right)\right)$ is finite, therefore so is $\mathbb{E}\left(J\left(\boldsymbol{U}_k\right)-J\left(\boldsymbol{U}_{n-i}\right)\right) = a_k - a_{n-i}$. All quantities in (35) are finite and we can write:

$$\varepsilon_k\mathbb{E}\left(J\left(\boldsymbol{U}_k\right)-J\left(\boldsymbol{U}_{n-i}\right)\right) \leq (1+\alpha_k)\mathbb{E}\left(\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_k)\right) + (\beta_k-1)\mathbb{E}\left(\ell_{\boldsymbol{U}_{n-i}}(\boldsymbol{U}_{k+1})\right) + \gamma_k , \quad (36)$$

From Lemma 4.6, there exists a constant $M' \geq 0$ such that $\mathbb{E}\left(\ell_{\boldsymbol{U}_j}(\boldsymbol{U}_k)\right) \leq M'$ for all $j, k \in \mathbb{N}$. Summing (36) from $n-i$ to n and using that $\beta_n - 1 \leq 0$ for n large enough, we get:

$$\sum_{k=n-i+1}^{n} \varepsilon_k \mathbb{E}\left(J\left(\boldsymbol{U}_k\right) - J\left(\boldsymbol{U}_{n-i}\right)\right) \leq \gamma_{n-i} + \sum_{k=n-i+1}^{n} M'\left(\alpha_k + \beta_{k-1}\right) + \gamma_k \,. \tag{37}$$

Define:

$$\bar{\alpha}_k = \alpha\varepsilon_k^2 + \frac{2}{b}\varepsilon_k q k^{-\nu}, \quad \bar{\gamma}_k = \left(\gamma + \delta q^2 k^{-2\nu}\right)\varepsilon_k^2 + \varepsilon_k q k^{-\nu} \,.$$

As $Q_k \leq q k^{-\nu}$, we have $\alpha_k \leq \bar{\alpha}_k$ and $\gamma_k \leq \bar{\gamma}_k$. Let $\xi_k = (M + M')(\bar{\alpha}_k + \beta_{k-1}) + \bar{\gamma}_k$. Moreover, note that:

$$\sum_{i=1}^{n-1} \frac{1}{i(i+1)} \sum_{k=n-i+1}^{n} \xi_k = \sum_{k=2}^{n} \sum_{i=n-k+1}^{n-1} \left(\frac{1}{i} - \frac{1}{i+1}\right)\xi_k = \sum_{k=2}^{n} \frac{\xi_k}{n-k+1} - \frac{1}{n}\sum_{k=2}^{n} \xi_k \,. \tag{38}$$

We plug (33) and (37) into (32) and use that $M(\alpha_k + \beta_k) + \gamma_k \leq \xi_k$ and $M'(\alpha_k + \beta_k) + \gamma_k \leq \xi_k$ along with (38). This yields:

$$\begin{aligned}
\varepsilon_n a_n &\leq \frac{M_1}{n} + \frac{1}{n}\sum_{k=2}^{n} \xi_k + \sum_{k=2}^{n} \frac{1}{n-k+1}\xi_k - \frac{1}{n}\sum_{k=2}^{n} \xi_k + \sum_{i=1}^{n-1} \frac{\gamma_{n-i}}{i(i+1)} \\
&= \frac{M_1}{n} + \sum_{k=2}^{n} \frac{1}{n-k+1}\xi_k + \sum_{i=1}^{n-1} \frac{\gamma_{n-i}}{i(i+1)} \,.
\end{aligned} \tag{39}$$

From the assumptions on ε_k, $\{\xi_k\}_{k\in\mathbb{N}}$ is non-increasing. Thus,

$$\begin{aligned}
\sum_{k=2}^{n} \frac{1}{n-k+1}\xi_k &\leq \xi_{\lfloor \frac{n}{2}+1 \rfloor} \sum_{n/2+1 \leq k \leq n} \frac{1}{n-k+1} + \frac{2}{n} \sum_{2 \leq k < n/2+1} \xi_k \,, \\
&\leq \xi_{\lfloor \frac{n}{2}+1 \rfloor} \left(\log\left(\frac{n}{2}\right) + 1\right) + \frac{2}{n}\sum_{k=2}^{n} \xi_k \,.
\end{aligned} \tag{40}$$

By assumption, $\varepsilon_n = c n^{-\theta}$ and we have:

$$\begin{aligned}
\xi_k &\leq \left((M + M')(\alpha + \beta) + \gamma + \delta q^2 k^{-2\nu}\right) c^2 (k-1)^{-2\theta} + \left(\frac{2}{b}(M + M') + 1\right) c q k^{-(\nu+\theta)} \\
&\leq \xi (k-1)^{-\mu} \,,
\end{aligned}$$

for $\mu = \min\{2\theta, \nu + \theta\}$ and some constant $\xi > 0$ so that,

$$\frac{1}{\varepsilon_n}\sum_{k=2}^{n} \frac{1}{n-k+1}\xi_k \leq 2^{\mu}\frac{\xi}{c}(n-1)^{\theta-\mu}\left(\log\left(\frac{n}{2}\right) + 1\right) + 2\frac{\xi}{c} n^{\theta-1} \sum_{k=1}^{n-1} k^{-\mu} \,. \tag{41}$$

As $\theta > 1/2$ and $\nu > 1 - \theta$, we have $\mu > 1$ so,

$$\sum_{k=1}^{n-1} k^{-\mu} \leq \frac{\mu}{\mu - 1} \, .$$

Using a similar computation as in (40) and that $\gamma_i \leq \kappa i^{-\mu}$ for some constant $\kappa > 0$, we deduce that there exist constants $\Gamma_1, \Gamma_2 > 0$ such that:

$$\frac{1}{\varepsilon_n} \sum_{i=1}^{n-1} \frac{\gamma_{n-i}}{i(i+1)} \leq \Gamma_1 n^{\theta-\mu} + \Gamma_2 n^{\theta-2} \, . \tag{42}$$

Gathering (41) and (42) into (39), we get:

$$a_n \leq M_1 n^{\theta-1} + \Xi_1 (n-1)^{\theta-\mu} \left(\log\left(\frac{n}{2}\right) + 1 \right) + \Xi_2 n^{\theta-1} + \Gamma_1 n^{\theta-\mu} + \Gamma_2 n^{\theta-2} \, , \tag{43}$$

with $\Xi_1 = 2^\mu \frac{\xi}{c}$ and $\Xi_2 = 2 \frac{\xi}{c} \frac{\mu}{\mu-1}$. Finally, as $\theta - \mu < \theta - 1$, we get that $a_n = \mathcal{O}\left(n^{\theta-1}\right)$. This concludes the proof. □

Remark 4.8 *Inequality* (30) *(which holds in fact for any* $u \in U^{\mathrm{ad}}$ *in place of* $u^\sharp$*) is the counterpart of modified Fejér monotonicity [23]. The main differences are that* (30) *involves a Bregman divergence instead of the Euclidean distance. Moreover, there are coefficients* $\alpha_k, \beta_k > 0$ *that slightly degrade the inequality compared to what we obtain with Fejér monotone sequences where* $\alpha_k = \beta_k = 0$*. The summability of* α_k *and* β_k *in addition with the boundedness of the expectation of the Bregman divergence* $\left\{ \mathbb{E}\left(\ell_{u^\sharp}(\boldsymbol{U}_k) \right) \right\}_{k \in \mathbb{N}}$ *allow us to proceed in the same way as in [23, 33] to get the convergence rate of Theorem 4.7.*

5 Conclusion

We have studied the stochastic APP algorithm in a reflexive separable Banach case. This framework generalizes many stochastic optimization algorithms for convex problems. We have proved the measurability of the iterates of the algorithm, hence filling a theoretical gap to ensure that the quantities we manipulate when deriving efficiency estimates are well-defined. We have shown the convergence of the stochastic APP algorithm in the case where a bias on the gradient is considered. Finally, efficiency estimates are derived while taking the bias into account. Assuming a sufficiently fast decay of this bias, we get a convergence rate for the expectation of the function values that is similar to that of well-known stochastic optimization algorithms when no bias is present, such as stochastic gradient descent [2], stochastic mirror descent [25] or the stochastic proximal gradient algorithm [33]. Future work will consist in an application the stochastic APP algorithm to an optimization problem in a Banach space with decomposition aspects in mind.

Acknowledgements

The authors would like to thank the two anonymous reviewers for their careful reading and for their comments and questions which helped to improve and clarify the manuscript.

A Technical results used in the proofs

Lemma A.1 *Let $\{a_i\}_{i\in\mathbb{N}}$ be a sequence in $\mathbb{R}$. Let $n \in \mathbb{N}$ and for $i \in \{0, 1, \dots, n-1\}$, let $s_i = \sum_{k=n-i}^{n} a_k$. Then,*

$$a_n = \frac{s_{n-1}}{n} + \sum_{i=1}^{n-1} \frac{1}{i(i+1)}(s_{i-1} - i a_{n-i}) \, .$$

The proof of Lemma A.1 is a straightforward computation and is left to the reader.

Theorem A.2 *[31, Robbins-Siegmund] Consider four sequences of nonnegative random variables $\{\boldsymbol{\Lambda}_k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\alpha}_k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\beta}_k\}_{k\in\mathbb{N}}$ and $\{\boldsymbol{\eta}_k\}_{k\in\mathbb{N}}$, that are all adapted to a given filtration $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$. Moreover, suppose that:*

$$\mathbb{E}\big(\boldsymbol{\Lambda}_{k+1} \,\big|\, \mathcal{F}_k\big) \leq (1 + \boldsymbol{\alpha}_k)\boldsymbol{\Lambda}_k + \boldsymbol{\beta}_k - \boldsymbol{\eta}_k, \ \forall k \in \mathbb{N} \, ,$$

and that $\sum_k \boldsymbol{\alpha}_k < +\infty$ and $\sum_k \boldsymbol{\beta}_k < +\infty$ almost surely. Then, the sequence of random variables $\{\boldsymbol{\Lambda}_k\}_{k\in\mathbb{N}}$ converges almost surely to a finite random variable $\boldsymbol{\Lambda}^\infty$, and we have in addition that $\sum_k \boldsymbol{\eta}_k < +\infty$ almost surely.

An extension of Robbins-Siegmund theorem is given by the following corollary.

Corollary A.3 *Consider the following sequences of nonnegative random variables $\{\boldsymbol{\Lambda}_k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\alpha}_k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\beta}_k\}_{k\in\mathbb{N}}$, $\{\boldsymbol{\gamma}_k\}_{k\in\mathbb{N}}$, and $\{\boldsymbol{\eta}_k\}_{k\in\mathbb{N}}$, that are all adapted to a given filtration $\{\mathcal{F}_k\}_{k\in\mathbb{N}}$. Moreover suppose that:*

$$\mathbb{E}\big(\boldsymbol{\Lambda}_{k+1} \,\big|\, \mathcal{F}_k\big) \leq \big(1 + \boldsymbol{\alpha}_k\big)\boldsymbol{\Lambda}_k + \boldsymbol{\beta}_k \mathbb{E}\big(\boldsymbol{\Lambda}_{k+1} \,\big|\, \mathcal{F}_k\big) + \boldsymbol{\gamma}_k - \boldsymbol{\eta}_k \, ,$$

and that $\sum_k \boldsymbol{\alpha}_k < +\infty$, $\sum_k \boldsymbol{\beta}_k < +\infty$ and $\sum_k \boldsymbol{\gamma}_k < +\infty$ almost surely. Then, the sequence of random variables $\{\boldsymbol{\Lambda}_k\}_{k\in\mathbb{N}}$ converges almost surely to a finite random variable $\boldsymbol{\Lambda}^\infty$, and we have in addition that $\sum_k \boldsymbol{\eta}_k < +\infty$ almost surely.

Proof. Consider a realization of the different sequences satisfying the assumptions of the corollary, and define three sequences $\{\widetilde{\alpha}_k\}_{k\in\mathbb{N}}$, $\{\widetilde{\gamma}_k\}_{k\in\mathbb{N}}$ and $\{\widetilde{\eta}_k\}_{k\in\mathbb{N}}$ such that:

$$1 + \widetilde{\alpha}_k = \frac{1 + \alpha_k}{1 - \beta_k} \, , \ \widetilde{\gamma}_k = \frac{\gamma_k}{1 - \beta_k} \, , \ \widetilde{\eta}_k = \frac{\eta_k}{1 - \beta_k} \, .$$

As the sequence $\{\beta_k\}$ converges to zero, we have that $\beta_k \leq 1/2$ for k large enough. For such k, we get:

$$\frac{1}{1 - \beta_k} \leq 1 + 2\beta_k \quad \text{and} \quad 1 \leq \frac{1}{1 - \beta_k} \leq 2 \, .$$

Then, we deduce that $\widetilde{\alpha}_k \leq 2(\alpha_k + \beta_k)$, $\widetilde{\gamma}_k \leq 2\gamma_k$ and $\widetilde{\eta}_k \geq \eta_k$. The conclusions of the corollary are then obtained by applying Theorem A.2 directly. □

Proposition A.4 *Let $\mathbb{U}$ be a Banach space and consider a function $J : \mathbb{U} \to \mathbb{R}$ that is subdifferentiable on a non-empty, closed, convex subset U^{ad} of $\mathbb{U}$, with linearly bounded subgradient. Then, there exist $c_1 > 0$ and $c_2 > 0$ such that:*

$$\forall (u, v) \in U^{\mathrm{ad}} \times U^{\mathrm{ad}} \ , \ \left| J(u) - J(v) \right| \leq \Big(c_1 \max\left\{ \|u\|, \|v\| \right\} + c_2 \Big) \, \|u - v\| \ . \tag{44}$$

In particular, J is Lipschitz continuous on every bounded subset that is contained in U^{ad}.

Proof. Let $(u, v) \in U^{\mathrm{ad}} \times U^{\mathrm{ad}}$. From the definition of subdifferentiability, we get that for all $r \in \partial J(u)$ and for all $s \in \partial J(v)$:

$$\langle s \, , u - v \rangle \leq J(u) - J(v) \leq \langle r \, , u - v \rangle \ ,$$

and therefore:

$$\left| J(u) - J(v) \right| \leq \max\left\{ \langle r \, , u - v \rangle, \langle s \, , v - u \rangle \right\} .$$

Using Cauchy-Schwarz inequality and the linearly bounded subgradient assumption, we get the desired result. □

Proposition A.5 *Let $\mathbb{U}$ be a Banach space and $J : \mathbb{U} \to \mathbb{R}$ be a Lipschitz continuous function with constant $L > 0$. Let $\{u_k\}_{k \in \mathbb{N}}$ be a sequence of elements in $\mathbb{U}$ and let $\{\varepsilon_k\}_{k \in \mathbb{N}}$ be a real positive sequence such that:*

(a) $\sum_{k \in \mathbb{N}} \varepsilon_k = +\infty$,

(b) $\exists \mu \in \mathbb{R}$, $\sum_{k \in \mathbb{N}} \varepsilon_k \left| J(u_k) - \mu \right| < +\infty$,

(c) $\exists \delta > 0$, $\forall k \in \mathbb{N}$, $\|u_{k+1} - u_k\| \leq \delta \varepsilon_k$.

Then, the sequence $\left\{ J(u_k) \right\}_{k \in \mathbb{N}}$ converges to μ.

Proof. For $\alpha > 0$, define $N_\alpha = \left\{ k \in \mathbb{N}, \ |J(u_k) - \mu| \leq \alpha \right\}$ and $N_\alpha^{\complement} = \mathbb{N} \setminus N_\alpha$.

(i) From Assumption (b), we have:

$$+\infty > \sum_{k \in \mathbb{N}} \varepsilon_k \left| J(u_k) - \mu \right| \geq \sum_{k \in N_\alpha^{\complement}} \varepsilon_k \left| J(u_k) - \mu \right| \geq \alpha \sum_{k \in N_\alpha^{\complement}} \varepsilon_k \ .$$

Hence, for all $\beta > 0$, there exists $n_\beta \in \mathbb{N}$ such that $\sum_{k \geq n_\beta, k \in N_\alpha^{\complement}} \varepsilon_k \leq \beta$.

(ii) From Assumption (a), we have $\sum_{k \in \mathbb{N}} \varepsilon_k = \sum_{k \in N_\alpha} \varepsilon_k + \sum_{k \in N_\alpha^{\complement}} \varepsilon_k = +\infty$ but we have just proved that the last sum in the above equality is finite, hence the first sum of the right hand side is infinite, which implies that N_α is infinite.

Let $\epsilon > 0$, choose $\alpha = \epsilon/2$ and $\beta = \epsilon/(2L\delta)$. Let n_β be the integer defined in (i). For $k \geq n_\beta$, there are two possible cases:

- $k \in N_\alpha$: then, by definition of N_α, we have $\big|J(u_k) - \mu\big| \leq \alpha < \epsilon$.
- $k \notin N_\alpha$: let m be the smallest element of N_α such that $m \geq k$, this element exists by (ii). Using the fact that J is Lipschitz continuous, we get:

$$\big|J(u_k) - \mu\big| \leq \big|J(u_k) - J(u_m)\big| + \big|J(u_m) - \mu\big| \leq L\|u_k - u_m\| + \alpha \; .$$

  Now, with Assumption (c) and condition (i), it comes:

$$\big|J(u_k) - \mu\big| \leq L\delta\left(\sum_{l=k}^{m-1} \varepsilon_l\right) + \alpha \leq L\delta\left(\sum_{l \geq n_\beta, l \in N_\alpha^{\complement}} \varepsilon_l\right) + \alpha \leq \epsilon \; .$$

Hence, we get $\big|J(u_k) - \mu\big| \leq \epsilon$ for all $k \geq n_\beta$, giving the desired result. □

References

[1] Atchade, Y. F., Fort, G., and Moulines, E. On Perturbed Proximal Gradient Algorithms. *Journal of Machine Learning Research 18* (2017), 1–33.

[2] Bach, F., and Moulines, E. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Advances in Neural Information Processing Systems* (2011), pp. 451–459.

[3] Bauschke, H. H., and Combettes, P. L. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS books in mathematics. Springer, New York, 2011.

[4] Benveniste, A., Metivier, M., Priouret, P., and Wilson, S. S. *Adaptive algorithms and stochastic approximations*. No. 22 in Stochastic modelling and applied probability. Springer-Verl, Berlin, 2012.

[5] Bertsekas, D. P., and Shreve, S. E. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 1996.

[6] Billingsley, P. *Probability and measure*, 3rd ed. Wiley series in probability and mathematical statistics. Wiley, New York, 1995.

[7] Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics 7*, 3 (1967), 200–217.

[8] Brézis, H. *Analyse fonctionnelle: Théorie et applications*. Mathématiques appliquées pour la maîtrise. Dunod, 2005.

[9] Bubeck, S. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning 8*, 3-4 (2015), 231–357.

[10] Castaing, C., and Valadier, M. *Convex Analysis and Measurable Multifunctions*, vol. 580 of *Lecture Notes in Mathematics*. Springer Berlin Heidelberg, 1977.

[11] Cohen, G. Optimization by decomposition and coordination: A unified approach. *IEEE Transactions on Automatic Control 23*, 2 (1978), 222–232.

[12] Culioli, J.-C. *Algorithmes de décomposition/coordination en optimisation stochastique*. PhD thesis, Ecole Nationale Supérieure des Mines de Paris, 1987.

[13] Culioli, J.-C., and Cohen, G. Decomposition/Coordination Algorithms in Stochastic Optimization. *SIAM Journal on Control and Optimization 28*, 6 (1990), 1372–1403.

[14] Ekeland, I., and Temam, R. *Convex Analysis and Variational Problems*. Society for Industrial and Applied Mathematics, 1999.

[15] Geiersbach, C., and Pflug, G. C. Projected Stochastic Gradients for Convex Constrained Problems in Hilbert Spaces. *SIAM Journal on Optimization 29*, 3 (2019), 2079–2099.

[16] Geiersbach, C., and Scarinci, T. Stochastic Proximal Gradient Methods for Nonconvex Problems in Hilbert Spaces. *arXiv:2001.01329 [math]* (Jan. 2020). arXiv: 2001.01329.

[17] Hess, C. On the Measurability of the Conjugate and the Subdifferential of a Normal Integrand. *Journal of Convex Analysis 2*, 1-2 (1995), 153–165.

[18] Hess, C. Epi-convergence of sequences of normal integrands and strong consistency of the maximum likelihood estimator. *The Annals of Statistics 24*, 3 (1996), 1298–1315.

[19] Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. Non-asymptotic Analysis of Biased Stochastic Approximation Scheme. In *Proceedings of the Thirty-Second Conference on Learning Theory* (2019), vol. 99, pp. 1944–1974.

[20] Kiefer, J., and Wolfowitz, J. Stochastic Estimation of the Maximum of a Regression Function. *The Annals of Mathematical Statistics 23*, 3 (1952), 462–466.

[21] Kuratowski, K., and Ryll-Nardzewski, C. A general theorem on selectors. *Bull. Acad. Polon. Sci. Sér. Sci. Math. Astronom. Phys. 13*, 6 (1965), 397–403.

[22] Kushner, H. J., and Yin, G. *Stochastic approximation algorithms and applications*. No. 35 in Applications of Mathematics. Springer, New York, 1997.

[23] Lin, J., Rosasco, L., Villa, S., and Zhou, D.-X. Modified Fejér sequences and applications. *Computational Optimization and Applications 71*, 1 (2018), 95–113.

[24] Martin, M., Nobile, F., and Tsilifis, P. A Multilevel Stochastic Gradient method for PDE-constrained Optimal Control Problems with uncertain parameters. *arXiv:1912.11900 [math]* (2019).

[25] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust Stochastic Approximation Approach to Stochastic Programming. *SIAM Journal on Optimization 19*, 4 (2009), 1574–1609.

[26] Nemirovski, A., and Yudin, D. B. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983.

[27] Nesterov, Y. *Introductory Lectures on Convex Optimization*, vol. 87 of *Applied Optimization*. Springer US, Boston, MA, 2004.

[28] Parikh, N., and Boyd, S. P. Proximal Algorithms. *Foundations and Trends® in Optimization 1*, 3 (2014), 127–239.

[29] Polyak, B. T., and Juditsky, A. B. Acceleration of Stochastic Approximation by Averaging. *SIAM Journal on Control and Optimization 30*, 4 (1992), 838–855.

[30] Robbins, H., and Monro, S. A Stochastic Approximation Method. *The Annals of Mathematical Statistics 22*, 3 (1951), 400–407.

[31] Robbins, H., and Siegmund, D. A Convergence Theorem for Non Negative Almost Supermartingales and Some Applications. In *Optimizing Methods in Statistics*. Springer, New York, NY, 1971, pp. 233–257.

[32] Rockafellar, R. T., and Wets, R. J.-B. *Variational analysis*. No. 317 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2004.

[33] Rosasco, L., Villa, S., and Vũ, B. C. Convergence of Stochastic Proximal Gradient Algorithm. *Applied Mathematics & Optimization* (2019), 1–27.

[34] Ruppert, D. Efficient estimations from a slowly convergent Robbins-Monro process. Tech. rep., Cornell University Operations Research and Industrial Engineering, 1988.

[35] Thibault, L. Espérances conditionnelles d'intégrandes semi-continus. *Annales de l'I.H.P, section B 17*, 4 (1981), 337–350.

[36] Wada, T., and Fujisaki, Y. A stopping rule for stochastic approximation. *Automatica 60* (2015), 1–6.

[37] Yin, G. A stopping rule for the Robbins-Monro method. *Journal of Optimization Theory and Applications 67*, 1 (1990), 151–173.