Stochastic Forward–Backward Splitting for Monotone Inclusions

Citation	Rosasco, Lorenzo, Silvia Villa, and Bang Công Vũ. "Stochastic Forward–Backward Splitting for Monotone Inclusions." Journal of Optimization Theory and Applications 169, no. 2 (February 18, 2016): 388–406.
As Published	http://dx.doi.org/10.1007/s10957-016-0893-2
Publisher	Springer US
Version	Author's final manuscript
Citable link	http://hdl.handle.net/1721.1/103419
Terms of Use	Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.

JOTA manuscript No.
(will be inserted by the editor)

Stochastic Forward-Backward Splitting for Monotone Inclusions

Lorenzo Rosasco · Silvia Villa · Bang Công Vũ

Received: date / Accepted: date

Communicated by Marc Teboulle

Abstract We propose and analyze the convergence of a novel stochastic algorithm for monotone inclusions that are sum of a maximal monotone operator and a single-valued cocoercive operator. The algorithm we propose is a natural stochastic extension of the classical forward-backward method. We provide a non-asymptotic error analysis in expectation for the strongly monotone case, as

L. Rosasco
DIBRIS, Università di Genova
Genova, Italy
lrosasco@mit.edu

L. Rosasco, S. Villa (corresponding author) and B. C. Vũ
Laboratory for Computational and Statistical Learning
Istituto Italiano di Tecnologia and Massachusetts Institute of Technology,
Cambridge, USA
silvia.villa@iit.it, cong.bang@iit.it

well as almost sure convergence under weaker assumptions. For minimization problems, we recover rates matching those obtained by stochastic extensions of so called accelerated methods. Stochastic quasi Fejér's sequences are a key technical tool to prove almost sure convergence.

Keywords Stochastic first order methods · Forward-backward splitting algorithm · Monotone inclusions · Stochastic Fejér sequences

Mathematics Subject Classification (2010) 47H05 · 90C15 · 65K10 · 90C25

1 Introduction

Maximal monotone operators have been studied extensively since [1], because of their wide applicability in pure and applied sciences [2,3]. The corresponding framework allows for a unified treatment of equilibrium problems, variational inequalities, and convex optimization, see e.g. [4,2,5]. A key problem in this context is to find a solution of an inclusion defined by a maximal monotone set-valued operator [4] and, in this paper, we assume the operator defining the inclusion to be the sum of a maximal monotone operator and a single-valued cocoercive operator. Such structured inclusions encompass fixed point problems, variational inequalities, and composite minimization problems [6, 7]. The literature on algorithmic schemes for solving structured inclusions is vast. In particular, approaches are known that separate the contribution of the two summands, notably forward-backward splitting algorithms [4,8]. Since the seminal works [9,10], forward-backward splitting methods have been considerably developed to be more flexible, faster and robust to errors, see [4, 11–15].

In this paper, we assume the single valued operator to be known only through stochastic estimates. This setting is practically relevant to consider measurements with non vanishing random noise, or cases where the computation of stochastic estimates is cheaper than the evaluation of the operator itself. While there is a rich literature on stochastic proximal gradient splitting algorithms for convex minimization problems [16,17], and various results for variational inequalities are available [18–20], we are not aware of previous studies of stochastic splitting algorithms for solving monotone inclusions, except for the concurrent papers [25,26]. In this paper, we propose a natural stochastic forward-backward splitting method and prove: 1) a non-asymptotic error analysis in expectation, and 2) strong almost sure convergence of the iterates. More specifically, under strong monotonicity assumptions, we provide non asymptotic bounds for convergence in norm and in expectation, leveraging on a non asymptotic version of Chung's lemma [21, Chapter 2, Lemma 5]. Almost sure convergence is obtained under the weaker assumption of uniform monotonicity of B using the concept of stochastic quasi-Fejèr sequences [22,23]. For variational inequalities, we obtain additional convergence results without stronger monotonicity assumptions.

A few features of our approach are worth mentioning. First, our assumptions on the stochastic estimates are weaker than those usually required in the literature, see e.g. [24]. In particular, our assumptions are different from those in [25,26], assuming an error summability condition. Second, our approach allows to avoid averaging the iterates, an aspect crucial in situations where structure is meant to induce sparse solutions and averaging can be detrimental, see e.g. [27].

The paper is organized as follows. Section 2 collects some basic definitions. In Section 3 we establish the main results of the paper. Section 4 focuses on variational inequalities and minimization problems.

2 Preliminaries and Notation

Before discussing our main contributions, we set the notation and recall basic concepts and results we need in the following.

Throughout, $(\boldsymbol{\Omega}, \boldsymbol{\mathcal{A}}, \mathsf{P})$ is a probability space, $\mathbb{N}^* = \mathbb{N}\backslash\{0\}$, $\mathcal{H}$ is a real separable Hilbert space. We denote by $\langle \cdot \mid \cdot \rangle$ and $\| \cdot \|$ the scalar product and the associated norm of $\mathcal{H}$. An operator $A\colon \mathcal{H} \to 2^{\mathcal{H}}$ is denoted by $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$. The class of all lower semicontinuous convex functions $G\colon \mathcal{H} \to \left]-\infty, +\infty\right]$ such that $\operatorname{dom} G := \big\{x \in \mathcal{H} : G(x) < +\infty\big\} \neq \varnothing$ is denoted by $\Gamma_0(\mathcal{H})$. We denote by $\sigma(X)$ the σ-field generated by a random variable $X\colon \boldsymbol{\Omega} \to \mathcal{H}$, where $\mathcal{H}$ is endowed with the Borel σ-algebra. A sequence $(\mathcal{F}_n)_{n\in\mathbb{N}}$ of sub sigma algebras of $\boldsymbol{\mathcal{A}}$ such that, for every $n \in \mathbb{N}$, $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$ is called a filtration. Let, for every $n \in \mathbb{N}$, $X_n\colon \boldsymbol{\Omega} \to \mathcal{H}$ be an integrable random variable with $\mathsf{E}[\|X_n\|] < +\infty$. The sequence $(X_n)_{n\in\mathbb{N}}$ is called a random process.

Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be a set-valued operator. The domain and the graph of A are denoted by $\operatorname{dom} A$ and $\operatorname{gra} A$ (see [4]). The set of zeros of A is denoted by

$\operatorname{zer} A := \big\{w \in \mathcal{H} : 0 \in Aw\big\}$. Moreover, A is monotone iff

$$\big(\forall\big((w,u),(y,v)\big) \in (\operatorname{gra} A)^2\big) \quad \langle w-y \mid u-v\rangle \geq 0, \tag{1}$$

and maximally monotone iff it is monotone and there exists no monotone operator $B\colon \mathcal{H} \rightrightarrows \mathcal{H}$ such that $\operatorname{gra} B$ properly contains $\operatorname{gra} A$.

Suppose that A is monotone and let $y \in \operatorname{dom} A$. We say that A is uniformly monotone at y iff there exists an increasing function $\phi\colon \left[0,+\infty\right[\to [0,+\infty]$ vanishing only at 0 such that

$$\big(\forall (w,u) \in \operatorname{gra} A\big)\big(\forall v \in Ay\big) \quad \langle w-y \mid u-v\rangle \geq \phi(\|w-y\|). \tag{2}$$

In the case when $\phi = \mu|\cdot|^2$, for some $\mu \in \left]0,+\infty\right[$, we say that A is μ-strongly monotone at y. If $A - \mu\,\mathrm{I}$ is monotone, for some $\mu \in \left]0,+\infty\right[$, we say that A is μ-strongly monotone. We say that A is strictly monotone at $y \in \operatorname{dom} A$ iff, for every $(w,u) \in \operatorname{gra} A$ and for every $v \in Ay$, $w \neq y \Rightarrow \langle w-y \mid u-v\rangle > 0$. Let $\beta \in \left]0,+\infty\right[$. A single-valued operator $B\colon \mathcal{H} \to \mathcal{H}$ is β-cocoercive iff

$$(\forall (w,y) \in \mathcal{H}^2) \quad \langle w-y \mid Bw-By\rangle \geq \beta\|Bw-By\|^2.$$

The resolvent of any maximally monotone operator A is $J_A := (\mathrm{I}+A)^{-1}$. We recall that J_A is well defined and single valued [1], and can therefore be identified with an operator $J_A\colon \mathcal{H} \to \mathcal{H}$. When $A = \partial G$ for some $G \in \Gamma_0(\mathcal{H})$, then J_A coincides with the proximity operator of G [28], which is defined as

$$\operatorname{prox}_G\colon \mathcal{H} \to \mathcal{H}\colon w \mapsto \underset{v\in\mathcal{H}}{\operatorname{argmin}}\, G(v) + \frac{1}{2}\|w-v\|^2. \tag{3}$$

We next recall the concept of stochastic quasi Fejér sequence, which was introduced and studied in the papers [29, 22, 23]. This concept provides a unified approach to prove convergence of several algorithms in convex optimization; see [4] and references therein.

Definition 2.1 [23] Let S be a non-empty subset of $\mathcal{H}$. A random process $(w_n)_{n\in\mathbb{N}^*}$ in $\mathcal{H}$ is stochastic quasi-Fejér monotone with respect to the set S if $\mathsf{E}[\|w_1\|^2] < +\infty$ and there exists $(\varepsilon_n)_{n\in\mathbb{N}^*} \in \ell^1_+(\mathbb{N}^*)$ such that

$$(\forall w \in S)(\forall n \in \mathbb{N}^*) \quad \mathsf{E}[\|w_{n+1} - w\|^2 | \sigma(w_1, \ldots, w_n)] \leq \|w_n - w\|^2 + \varepsilon_n \quad \text{a.s.}$$

3 Main Results

The following is the main problem studied in the paper.

Problem 3.1 Let $A\colon \mathcal{H} \rightrightarrows \mathcal{H}$ be maximally monotone, let $\beta \in]0, +\infty[$ and let $B\colon \mathcal{H} \to \mathcal{H}$ be β-cocoercive. Assume that $\operatorname{zer}(A + B) \neq \varnothing$. The goal is to find $\overline{w} \in \mathcal{H}$ such that

$$0 \in A\overline{w} + B\overline{w}. \tag{4}$$

3.1 Algorithm

We propose the following stochastic forward-backward splitting algorithm for solving Problem 3.1. The key difference with respect to the classical setting is that we assume to have access only to a stochastic estimate of B.

Algorithm 3.1 *Let $(\gamma_n)_{n\in\mathbb{N}^*}$ be a sequence in $]0, +\infty[$, $(\lambda_n)_{n\in\mathbb{N}^*}$ be a sequence in $[0, 1]$, and $(\mathfrak{B}_n)_{n\in\mathbb{N}^*}$ be a $\mathcal{H}$-valued random process such that, for*

every $n \in \mathbb{N}^*$, $\mathsf{E}[\|\mathfrak{B}_n\|^2] < +\infty$. *Let* $w_1 \colon \Omega \to \mathcal{H}$ *be a random variable such that* $\mathsf{E}[\|w_1\|^2] < +\infty$ *and set*

$$(\forall n \in \mathbb{N}^*) \quad \left\lfloor \begin{array}{l} z_n = w_n - \gamma_n \mathfrak{B}_n \\ y_n = J_{\gamma_n A} z_n \\ w_{n+1} = (1-\lambda_n) w_n + \lambda_n y_n. \end{array} \right. \tag{5}$$

We will consider the following conditions for the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}^*}$, where $\mathcal{F}_n = \sigma(w_1, \ldots, w_n)$, for every $n \in \mathbb{N}^*$.

(A1) For every $n \in \mathbb{N}^*$, $\mathsf{E}[\mathfrak{B}_n|\mathcal{F}_n] = Bw_n$.

(A2) There exist $(\alpha_n)_{n\in\mathbb{N}^*}$ in $]0,+\infty[$ and $\delta \in \;]0,+\infty[$ such that, for every $n \in \mathbb{N}^*$, $\mathsf{E}[\|\mathfrak{B}_n - Bw_n\|^2|\mathcal{F}_n] \leq \delta^2(1 + \alpha_n\|Bw_n\|^2)$.

(A3) There exists $\varepsilon \in]0,+\infty[$ such that $(\forall n \in \mathbb{N}^*)$ $\gamma_n \leq (2-\epsilon)\beta/(1+2\delta^2\alpha_n)$.

(A4) Let $\overline{w}$ be a solution of Problem 3.1 and let, $\chi_n^2 = \lambda_n\gamma_n^2\big(1 + 2\alpha_n\|B\overline{w}\|^2\big)$, for every $n \in \mathbb{N}^*$. Then the following hold:

$$\sum_{n\in\mathbb{N}^*} \lambda_n\gamma_n = +\infty \quad \text{and} \quad \sum_{n\in\mathbb{N}^*} \chi_n^2 < +\infty. \tag{6}$$

Remark 3.1

(i) If, for every $n \in \mathbb{N}^*$, $\mathfrak{B}_n = Bw_n$, Algorithm 3.1 reduces to the well known forward–backward splitting in [30, Section 6]. However, under Assumptions (A1)-(A2)-(A3)-(A4), weak convergence of $(w_n)_{n\in\mathbb{N}^*}$ is not guaranteed since (A4) implies $\inf \gamma_n = 0$, while to apply the classic theory we need $\inf \gamma_n > 0$. Under our assumptions, only ergodic convergence of $(w_n)_{n\in\mathbb{N}^*}$ has been proved in the deterministic case; see [10,31].

(ii) A stochastic forward-backward splitting algorithm for monotone inclusions has been recently analyzed in [25,26], under rather different assumptions. Indeed, they consider a fixed stepsize and a summability condition on $\mathsf{E}[\|\mathfrak{B}_n - Bw_n\|^2|\mathcal{F}_n]$. In the case $A = \partial G$ and $B = \nabla F$, for some G and $F \in \Gamma_0(\mathcal{H})$ such that F is differentiable with β^{-1}-Lipschitz continuous gradient, Algorithm 3.1 is a variant of the algorithm in [16], also studied in [17].

(iii) Condition (A2) is a more general than the condition usually assumed in the context of stochastic optimization, where $\alpha_n = 0$.

(iv) If $A = 0$, $(A4)$ becomes $\sum_{n\in\mathbb{N}^*} \lambda_n\gamma_n = +\infty$ and $\sum_{n\in\mathbb{N}^*} \lambda_n\gamma_n^2 < +\infty$. The latter are the usual conditions required for stochastic gradient descent algorithms; see e.g. [32].

Example 3.1 Let $(\mathcal{G}, \mathcal{B}, P)$ be a probability space, let $b\colon \mathcal{H}\times\mathcal{G} \to \mathcal{H}$ be a measurable function such that $\int_{\mathcal{G}} \|b(w,y)\|P(dy) < +\infty$, and suppose that B satisfies

$$(\forall w \in \mathcal{H}) \quad Bw = \int_{\mathcal{G}} b(w,y)P(dy). \tag{7}$$

If an independent and identically distributed sequence $(y_n)_{n\in\mathbb{N}^*}$ of realizations of the random vector y is available, then one can take $\mathfrak{B}_n = b(w_n, y_n)$. If in addition B is a gradient operator and $\mathcal{G}$ is finite dimensional, we are in the classical setting of stochastic optimization [24].

3.2 Almost Sure Convergence

In this section we describe our main results about almost sure convergence of the iterates of Algorithm 3.1. All the proofs are postponed to Section 3.4.

Proposition 3.1 *Suppose that* (A1), (A2), (A3), *and* (A4) *are satisfied. Let* $(w_n)_{n\in\mathbb{N}^*}$ *be the sequence generated by Algorithm 3.1 and let* $\overline{w}$ *be a solution of Problem 3.1. Then the following hold:*

(i) *The sequence* $(\mathsf{E}[\|w_n-\overline{w}\|^2])_{n\in\mathbb{N}^*}$ *converges to a finite value.*

(ii) $\sum_{n\in\mathbb{N}^*}\lambda_n\gamma_n\mathsf{E}[\langle w_n-\overline{w}\mid Bw_n-B\overline{w}\rangle]<+\infty$. *Consequently,*

$$\varliminf_{n\to\infty}\mathsf{E}[\langle w_n-\overline{w}\mid Bw_n-B\overline{w}\rangle]=0 \quad \textit{and} \quad \varliminf_{n\to\infty}\mathsf{E}[\|Bw_n-B\overline{w}\|^2]=0.$$

(iii) $\displaystyle\sum_{n\in\mathbb{N}^*}\lambda_n\mathsf{E}[\|w_n-y_n-\gamma_n(\mathfrak{B}_n-B\overline{w})\|^2]<+\infty$ *and* $\displaystyle\sum_{n\in\mathbb{N}^*}\lambda_n\mathsf{E}[\|w_n-y_n\|^2]<+\infty$.

Proposition 3.1 states similar properties to those stated for the forward-backward splitting algorithm in [13]. These properties are key to prove almost sure convergence, which is stated in the next theorem.

Theorem 3.2 *Suppose that conditions* (A1), (A2), (A3), *and* (A4) *are satisfied. Let* $(w_n)_{n\in\mathbb{N}^*}$ *be the sequence generated by Algorithm 3.1 and let* $\overline{w}$ *be a solution of Problem 3.1. Then the following hold:*

(i) $(w_n)_{n\in\mathbb{N}^*}$ *is stochastic quasi-Fejèr monotone with respect to* $\operatorname{zer}(A+B)$.

(ii) *There exists an integrable random variable* $\zeta_{\overline{w}}$ *such that* $\|w_n-\overline{w}\|^2\to\zeta_{\overline{w}}$ *a.s.*

(iii) *If* B *is uniformly monotone at* $\overline{w}$, *then* $w_n\to\overline{w}$ *a.s.*

(iv) *If B is strictly monotone at $\overline{w}$ and weakly continuous, then there exists $\Omega_1 \in \boldsymbol{\mathcal{A}}$ such that $\mathsf{P}(\Omega_1) = 1$, and, for every $\omega \in \Omega_1$, there exists a subsequence $(w_{t_n}(\omega))_{n\in\mathbb{N}^*}$ such that $w_{t_n}(\omega) \rightharpoonup \overline{w}$.*

Almost sure convergence is the one traditionally studied in the stochastic optimization literature. However, most papers focus on the finite dimensional setting, and require boundedness of the variance of the stochastic estimate of the gradients or subgradients (namely, $\alpha_n = 0$ in assumption (A2)). Weak almost sure convergence of the iterates generated by the stochastic forward-backward splitting algorithm can be derived from the results in [25,26], without additional monotonicity assumptions on A or B, under more restrictive assumptions on the stochastic error, with a nonvanishing stepsize.

Remark 3.2 Under the same assumptions as in Theorem 3.2, suppose in addition that B is strictly monotone at $\overline{w}$. The assumptions of Theorem 3.2(iv) are satisfied when either $\mathcal{H}$ is finite dimensional or B is bounded and linear.

3.3 Nonasymptotic Bounds

In this section we focus on convergence in expectation. We provide results for the case when either A or B is strongly monotone. We derive a nonasymptotic bound for $\mathsf{E}[\|w_n - \overline{w}\|^2]$ similarly to what has been done for the stochastic gradient algorithm for the case of minimization of a smooth function in the finite dimensional case [33, Theorem 1]. In the next theorem we will consider the following assumption.

Assumption 3.3 *Let $\overline{w}$ be a solution of Problem 3.1. Furthermore, suppose that A is ν-strongly monotone and B is μ-strongly monotone at $\overline{w}$, for some $(\nu, \mu) \in [0, +\infty[^2$ such that $\nu + \mu > 0$.*

To state the results more concisely, for every $c \in \mathbb{R}$, we define the function

$$\varphi_c \colon \,]0, +\infty[\, \to \mathbb{R} \colon t \mapsto \begin{cases} (t^c - 1)/c & \text{if } c \neq 0; \\ \\ \log t & \text{if } c = 0. \end{cases} \tag{8}$$

Theorem 3.4 *Let $(\underline{\lambda}, \overline{\alpha}) \in \,]0, +\infty[^2$ and let $(w_n)_{n\in\mathbb{N}^*}$ be the sequence generated by Algorithm 3.1. Assume that conditions* (A1), (A2), (A3), *and Assumption 3.3 are satisfied and suppose that $\inf_{n\in\mathbb{N}^*} \lambda_n \geq \underline{\lambda}$, $\sup_{n\in\mathbb{N}^*} \alpha_n \leq \bar{\alpha}$, and that $\gamma_n = c_1 n^{-\theta}$ for some $\theta \in \,]0, 1]$ and for some $c_1 \in]0, +\infty[$. Set*

$$t = 1 - 2^{\theta-1} \geq 0, \quad c = \frac{c_1\underline{\lambda}(2\nu + \mu\varepsilon)}{(1+\nu)^2}, \quad \text{and} \quad \tau = \frac{2\delta^2 c_1^2(1 + \overline{\alpha}\|B\overline{w}\|)}{c^2}. \tag{9}$$

Let n_0 be the smallest integer such that for every integer $n \geq n_0 > 1$, it holds $\max\{c, c_1\}n^{-\theta} \leq 1$. Define, $(\forall n \in \mathbb{N}^*)\, s_n = \mathsf{E}[\|w_n - \overline{w}\|^2]$. Then, for every $n \geq 2n_0$, the following hold:

(i) *Suppose that $\theta \in \,]0, 1[$. Then*

$$s_{n+1} \leq \Big(\tau c^2 \varphi_{1-2\theta}(n) + s_{n_0} \exp\Big(\frac{c n_0^{1-\theta}}{1-\theta}\Big)\Big) \exp\Big(\frac{-ct(n+1)^{1-\theta}}{1-\theta}\Big) + \frac{\tau 2^\theta c}{(n-2)^\theta} \tag{10}$$

(ii) *Suppose that $\theta = 1$. Then*

$$s_{n+1} \leq s_{n_0}\Big(\frac{n_0}{n+1}\Big)^c + \frac{\tau c^2}{(n+1)^c}\Big(1 + \frac{1}{n_0}\Big)^c \varphi_{c-1}(n)\,. \tag{11}$$

(iii) *The sequence $(s_n)_{n\in\mathbb{N}^*}$ satisfies*

$$s_n = \begin{cases} O(n^{-\theta}), & \text{if } \theta \in \left]0,1\right[, \\ O(n^{-c}) + O(n^{-1}), & \text{if } \theta = 1,\ c \neq 1, \\ O(n^{-c}) + O(n^{-1}\log n), & \text{if } \theta = 1,\ c = 1. \end{cases} \tag{12}$$

Theorem 3.4 implies that, even without assuming $(A4)$, in the strongly monotone case there is convergence in quadratic mean for every $\theta \in \left]0,1\right]$. The constants in (10) and (11) depend on the monotonicity constant of $A+B$. By (12) it follows that the best rate is obtained with $\theta = 1$, for a choice of c_1 ensuring $c > 1$.

3.4 Proofs of the Main Results

We start with a result characterizing the asymptotic behavior of stochastic quasi-Fejér monotone sequences. The following statement is given in [34, Lemma 2.3] without a proof. A version of Proposition 3.2 in the finite dimensional setting can also be found in [23]. The concept of stochastic Fejér sequences has been revisited and extended in a Hilbert space setting in [25].

Proposition 3.2 *Let S be a non-empty closed subset of $\mathcal{H}$, and let $(w_n)_{n\in\mathbb{N}^*}$ be stochastic quasi-Fejér monotone with respect to S. Then the following hold.*

(i) *Let $w \in S$. Then, there exist $\zeta_w \in \mathbb{R}$ and an integrable random variable $\xi_w \in \mathcal{H}$ such that $\mathsf{E}[\|w_n - w\|^2] \to \zeta_w$ and $\|w_n - w\|^2 \to \xi_w$ almost surely.*

(ii) *$(w_n)_{n\in\mathbb{N}^*}$ is bounded a.s.*

(iii) *The set of weak subsequential limits of $(w_n)_{n\in\mathbb{N}^*}$ is non-empty a.s.*

We next prove Proposition 3.1.

Proof (of Proposition 3.1) Let $n \in \mathbb{N}^*$. Since $\overline{w}$ is a solution of Problem 3.1 we have

$$\overline{w} = J_{\gamma_n A}(\overline{w} - \gamma_n B\overline{w})\,. \tag{13}$$

It follows from (5) and the convexity of $\|\cdot\|^2$ that

$$\|w_{n+1} - \overline{w}\|^2 \leq (1-\lambda_n)\|w_n - \overline{w}\|^2 + \lambda_n\|y_n - \overline{w}\|^2. \tag{14}$$

Since $J_{\gamma_n A}$ is firmly non-expansive by [4, Proposition 23.7], setting

$$u_n = w_n - y_n - \gamma_n(\mathfrak{B}_n - B\overline{w}). \tag{15}$$

we have

$$\begin{aligned} \|y_n - \overline{w}\|^2 &\leq \|(w_n - \overline{w}) - \gamma_n(\mathfrak{B}_n - B\overline{w})\|^2 - \|u_n\|^2 \qquad (16)\\ &= \|w_n - \overline{w}\|^2 - 2\gamma_n \left\langle w_n - \overline{w} \mid \mathfrak{B}_n - B\overline{w}\right\rangle + \gamma_n^2\|\mathfrak{B}_n - B\overline{w}\|^2 - \|u_n\|^2. \end{aligned}$$

Since $\mathsf{E}[\|\mathfrak{B}_n\|^2] < +\infty$ by assumption, we derive that $\mathsf{E}[\|\mathfrak{B}_n - B\overline{w}\|^2] < +\infty$. On the other hand, by induction we get that $\mathsf{E}[\|w_n - \overline{w}\|^2] < +\infty$ and hence $\mathsf{E}[\|w_n - \overline{w}\|] < +\infty$ and therefore $\mathsf{E}[|\langle w_n - \overline{w} \mid \mathfrak{B}_n - B\overline{w}\rangle|] < +\infty$, so that $\mathsf{E}[\langle w_n - \overline{w} \mid \mathfrak{B}_n - B\overline{w}\rangle\,|\mathcal{F}_n]$ is well-defined. Assumption (A1) yields

$$\begin{aligned} \mathsf{E}[\langle w_n - \overline{w} \mid \mathfrak{B}_n - B\overline{w}\rangle] &= \mathsf{E}[\mathsf{E}[\langle w_n - \overline{w} \mid \mathfrak{B}_n - B\overline{w}\rangle\,|\mathcal{F}_n]\\ &= \mathsf{E}[\langle w_n - \overline{w} \mid \mathsf{E}[\mathfrak{B}_n - B\overline{w}|\mathcal{F}_n]\rangle]\\ &= \mathsf{E}[\langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle]. \qquad (17) \end{aligned}$$

Moreover, using assumption (A2) and the cocoercivity of B, we have

$$
\begin{aligned}
&\mathsf{E}[\|\mathfrak{B}_n - B\overline{w}\|^2] = \\
&= \mathsf{E}[\|Bw_n - B\overline{w}\|^2] + \mathsf{E}[\|\mathfrak{B}_n - Bw_n\|^2] + 2\mathsf{E}[\langle Bw_n - B\overline{w}, \mathfrak{B}_n - Bw_n\rangle] \\
&\leq \mathsf{E}[\|Bw_n - B\overline{w}\|^2] + \delta^2(1 + \alpha_n \mathsf{E}[\|Bw_n\|^2]) \\
&\qquad\qquad\qquad\qquad + 2\mathsf{E}\left[\langle Bw_n - B\overline{w}, \mathsf{E}[\mathfrak{B}_n - Bw_n|\mathcal{F}_n\rangle]\right] \\
&\leq (1 + 2\delta^2\alpha_n)\mathsf{E}[\|Bw_n - B\overline{w}\|^2] + \delta^2(1 + 2\alpha_n\|B\overline{w}\|^2) \\
&\leq \frac{1 + 2\delta^2\alpha_n}{\beta}\mathsf{E}[\langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle] + 2\delta^2(1 + 2\alpha_n\|B\overline{w}\|^2).
\end{aligned}
\tag{18}
$$

Recalling the definition of ε, from (14), (16), (17), and (18) we get that

$$
\begin{aligned}
\mathsf{E}[\|w_{n+1} - \overline{w}\|^2] &\leq (1 - \lambda_n)\mathsf{E}[\|w_n - \overline{w}\|^2] + \lambda_n \mathsf{E}[\|y_n - \overline{w}\|^2] \\
&\leq \mathsf{E}[\|w_n - \overline{w}\|^2] - \gamma_n\lambda_n\left(2 - \frac{\gamma_n(1 + 2\delta^2\alpha_n)}{\beta}\right)\cdot \\
&\quad \cdot \mathsf{E}[\langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle] + 2\delta^2\chi_n^2 - \lambda_n \mathsf{E}[\|u_n\|^2] \\
&\leq \mathsf{E}[\|w_n - \overline{w}\|^2] - \varepsilon\gamma_n\lambda_n \mathsf{E}[\langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle] + 2\delta^2\chi_n^2 \\
&\quad - \lambda_n \mathsf{E}[\|u_n\|^2].
\end{aligned}
\tag{19}
$$

(i): Since the sequence $(\chi_n^2)_{n\in\mathbb{N}^*}$ is summable by assumption $(A4)$, we derive from (19) that $(\mathsf{E}[\|w_{n+1} - \overline{w}\|^2])_{n\in\mathbb{N}^*}$ converges to a finite value.

(ii): It follows from (19) that

$$
\sum_{n\in\mathbb{N}^*} \gamma_n\lambda_n \mathsf{E}[\langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle] < +\infty. \tag{20}
$$

Since $\sum_{n\in\mathbb{N}^*} \lambda_n\gamma_n = +\infty$ by $(A4)$, we get $\underline{\lim}\, \mathsf{E}[\langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle] = 0$, which implies, by cocoercivity, $\underline{\lim}\, \mathsf{E}[\|Bw_n - B\overline{w}\|^2] = 0$.

Since B is cocoercive, it is Lipschitzian. Therefore, by (i), there exists $M \in \,]0,+\infty[$ such that

$$(\forall n \in \mathbb{N}^*) \quad \mathsf{E}[\langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle] \leq \beta^{-1}\mathsf{E}[\|w_n - \overline{w}\|^2] \leq M. \tag{21}$$

Hence, we derive from (A4) and (18) that

$$\sum_{n\in\mathbb{N}^*} \lambda_n \gamma_n^2 \mathsf{E}[\|\mathfrak{B}_n - B\overline{w}\|^2] < +\infty. \tag{22}$$

(iii) It follows from (19) that $\sum_{n\in\mathbb{N}^*} \gamma_n\lambda_n \mathsf{E}[\|u_n\|^2] < +\infty$. Finally, by (22) we obtain

$$\sum_{n\in\mathbb{N}^*} \lambda_n \mathsf{E}[\|w_n - y_n\|^2] \leq 2\sum_{n\in\mathbb{N}} \lambda_n \mathsf{E}[\|u_n\|^2] + 2\sum_{n\in\mathbb{N}^*} \lambda_n\gamma_n^2 \mathsf{E}[\|\mathfrak{B}_n - B\overline{w}\|^2] < +\infty.$$

Therefore, (iii) is proved.

Next we prove Theorem 3.2, which is based on Propositions 3.2 and 3.1.

Proof (Theorem 3.2) (i) Let $n \in \mathbb{N}^*$. Reasoning as in the proof of Proposition 3.1, we have

$$\begin{aligned}&\|y_n - \overline{w}\|^2 \\ &\qquad \leq \|w_n - \overline{w}\|^2 - 2\gamma_n \langle w_n - \overline{w} \mid \mathfrak{B}_n - B\overline{w}\rangle + \gamma_n^2\|\mathfrak{B}_n - B\overline{w}\|^2 - \|u_n\|^2,\end{aligned} \tag{23}$$

where $u_n = w_n - y_n - \gamma_n(\mathfrak{B}_n - B\overline{w})$ is defined as in (15).

We next estimate the conditional expectation with respect to $\mathcal{F}_n$ of each term in the right hand side of (23). Since w_n is $\mathcal{F}_n$-measurable, using condition (A1),

$$\begin{aligned}\mathsf{E}[\langle w_n - \overline{w} \mid \mathfrak{B}_n - B\overline{w}\rangle \,|\mathcal{F}_n] &= \langle w_n - \overline{w} \mid \mathsf{E}[\mathfrak{B}_n - B\overline{w}|\mathcal{F}_n\rangle \\ &= \langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle\,.\end{aligned} \tag{24}$$

Noting that Bw_n is $\mathcal{F}_n$-measurable since w_n is $\mathcal{F}_n$-measurable and B is continuous, and using condition (A2) and cocoercivity of B, we derive

$$\begin{aligned}
&\mathsf{E}[\|\mathfrak{B}_n - B\overline{w}\|^2|\mathcal{F}_n] \\
&\quad = \mathsf{E}[\|\mathfrak{B}_n - Bw_n\|^2|\mathcal{F}_n] + \mathsf{E}[\|Bw_n - B\overline{w}\|^2|\mathcal{F}_n] \\
&\qquad\qquad\qquad\qquad + \mathsf{E}[\langle Bw_n - B\overline{w}, \mathfrak{B}_n - Bw_n\rangle|\mathcal{F}_n] \\
&\quad \leq \delta^2(1+\alpha_n\|Bw_n\|^2) + \|Bw_n - B\overline{w}\|^2 \\
&\quad \leq \|Bw_n - B\overline{w}\|^2 + \delta^2(1+2\alpha_n\|Bw_n - B\overline{w}\|^2 + 2\alpha_n\|B\overline{w}\|^2) \\
&\quad \leq \frac{(1+2\delta^2\alpha_n)}{\beta}\langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle + \delta^2(1+2\alpha_n\|B\overline{w}\|^2), \qquad (25)
\end{aligned}$$

Now, note that by convexity we have

$$\|w_{n+1} - \overline{w}\|^2 \leq (1-\lambda_n)\|w_n - \overline{w}\|^2 + \lambda_n\|y_n - \overline{w}\|^2. \qquad (26)$$

Taking the conditional expectation and invoking (23), (24), (25), we obtain,

$$\begin{aligned}
&\mathsf{E}[\|w_{n+1} - \overline{w}\|^2|\mathcal{F}_n] \leq (1-\lambda_n)\|w_n - \overline{w}\|^2 + \lambda_n\mathsf{E}[\|y_n - \overline{w}\|^2|\mathcal{F}_n] \\
&\quad \leq \|w_n - \overline{w}\|^2 - \gamma_n\lambda_n\left(2 - \frac{\gamma_n(1+2\delta^2\alpha_n)}{\beta}\right)\langle Bw_n - B\overline{w} \mid w_n - \overline{w}\rangle \\
&\qquad\qquad\qquad\qquad + 2\delta^2\chi_n^2 - \lambda_n\mathsf{E}[\|u_n\|^2|\mathcal{F}_n] \\
&\quad \leq \|w_n - \overline{w}\|^2 - \varepsilon\gamma_n\lambda_n\langle Bw_n - B\overline{w} \mid w_n - \overline{w}\rangle + 2\delta^2\chi_n^2 - \lambda_n\mathsf{E}[\|u_n\|^2|\mathcal{F}_n].
\end{aligned}$$

Hence $(w_n)_{n\in\mathbb{N}^*}$ is stochastic quasi-Fejér monotone with respect to the set $\operatorname{zer}(A+B)$, which is nonempty, closed, and convex.

(ii): It follows from Proposition 3.2(i) that $(\|w_n - \overline{w}\|^2)_{n\in\mathbb{N}^*}$ converges a.s to some integrable random variable $\zeta_{\overline{w}}$.

(iii) Since B is uniformly monotone at $\overline{w}$, there exists an increasing function $\phi\colon\ [0,+\infty[\ \to [0,+\infty[$ vanishing only at 0 such that

$$\langle Bw_n - B\overline{w} \mid w_n - \overline{w}\rangle \geq \phi(\|w_n - \overline{w}\|). \tag{27}$$

and thus $\overline{w}$ is the unique solution of Problem 3.1. We derive from Proposition 3.1 (ii) and (27) that

$$\sum_{n\in\mathbb{N}^*} \lambda_n\gamma_n \mathsf{E}[\phi(\|w_n - \overline{w}\|)] < \infty, \tag{28}$$

and hence

$$\sum_{n\in\mathbb{N}^*} \lambda_n\gamma_n \phi(\|w_n - \overline{w}\|) < \infty \quad \text{a.s.} \tag{29}$$

Since $(\lambda_n\gamma_n)_{n\in\mathbb{N}^*}$ is not summable by (A4), we have $\underline{\lim}\,\phi(\|w_n - \overline{w}\|) = 0$ a.s. Consequently, taking into account (ii), there exist $\Omega_1 \subset \Omega$ and an integrable random variable $\zeta_{\overline{w}}$ in $\mathcal{H}$ such that $P(\Omega_1) = 1$, and, for every $\omega \in \Omega_1$, $\underline{\lim}\,\phi(\|w_n(\omega) - \overline{w}\|) = 0$ and $\|w_n(\omega) - \overline{w}\|^2 \to \zeta_{\overline{w}}$. Let $\omega \in \Omega_1$. Then, there exists a subsequence $(k_n)_{n\in\mathbb{N}^*}$ such that $\phi(\|w_{k_n}(\omega) - \overline{w}\|) \to 0$, which implies that $\|w_{k_n}(\omega) - \overline{w}\| \to 0$, and therefore $w_n(\omega) \to \overline{w}$. Since ω is arbitrary in Ω_1, the statement follows.

(iv): By Proposition 3.1(i), $\underline{\lim}\,\mathsf{E}[\|Bw_n - B\overline{w}\|^2] = 0$, and hence there exists a subsequence $(k_n)_{n\in\mathbb{N}^*}$ such that $\lim_{n\to\infty} \mathsf{E}[\|Bw_{k_n} - B\overline{w}\|^2] = 0$. Therefore, there exists a subsequence $(p_n)_{n\in\mathbb{N}^*}$ of $(k_n)_{n\in\mathbb{N}^*}$ such that $\|Bw_{p_n} - B\overline{w}\|^2 \to 0$ almost surely. Thus, it follows from (ii) and Proposition 3.2(iii) that there exists $\Omega_1 \in \boldsymbol{\mathcal{A}}$ such that $\mathsf{P}(\Omega_1) = 1$ and, for every $\omega \in \Omega_1$, $(w_n(\omega))_{n\in\mathbb{N}^*}$ has weak cluster points and $\|Bw_{p_n}(\omega) - B\overline{w}\|^2 \to 0$. Fix $\omega \in \Omega_1$ and let $\overline{z}(\omega)$ be a weak cluster point of $(w_{p_n}(\omega))_{n\in\mathbb{N}^*}$, then there exists a subsequence $(w_{q_{p_n}}(\omega))_{n\in\mathbb{N}^*}$

such that $w_{q_{p_n}}(\omega) \rightharpoonup \overline{z}(\omega)$. Since B is weakly continuous, $Bw_{q_{p_n}}(\omega) \rightharpoonup B\overline{z}(\omega)$. Therefore, $B\overline{w} = B\overline{z}(\omega)$, and hence $\langle B\overline{z}(\omega) - B\overline{w} \mid \overline{z}(\omega) - \overline{w}\rangle = 0$. Since B is strictly monotone at $\overline{w}$, we obtain, $\overline{w} = \overline{z}(\omega)$. This shows that $w_{q_{p_n}}(\omega) \rightharpoonup \overline{w}$. Defining $(t_n)_{n\in\mathbb{N}^*}$ by setting, for every $n \in \mathbb{N}^*$, $t_n = q_{p_n}$ the statement follows.

The following lemma establishes a non asymptotic bound for numerical sequences satisfying a given recursive inequality. This is a non asymptotic version of Chung's lemma [21, Chapter 2, Lemma 5] (see also [33]).

Lemma 3.1 *Let α be in $]0,1]$, and let c and τ be in $]0,+\infty[$, let $(\eta_n)_{n\in\mathbb{N}^*}$ be the sequence defined by $(\forall n \in \mathbb{N}^*)$ $\eta_n = cn^{-\alpha}$. Let $(s_n)_{n\in\mathbb{N}^*}$ be such that*

$$(\forall n \in \mathbb{N}^*) \quad 0 \leq s_{n+1} \leq (1-\eta_n)s_n + \tau\eta_n^2. \tag{30}$$

Let n_0 be the smallest integer such that, for every $n \geq n_0 > 1$, it holds $\eta_n \leq 1$, set $t = 1 - 2^{\alpha-1} \geq 0$, and define $\varphi_{1-2\alpha}$ and φ_{c-1} as in (8). Then, for every $n \geq 2n_0$, if $\alpha \in\,]0,1[$,

$$s_{n+1} \leq \left(\tau c^2 \varphi_{1-2\alpha}(n) + s_{n_0} \exp\left(\frac{cn_0^{1-\alpha}}{1-\alpha}\right)\right) \exp\left(\frac{-ct(n+1)^{1-\alpha}}{1-\alpha}\right) + \frac{\tau 2^\alpha c}{(n-2)^\alpha} \tag{31}$$

and, if $\alpha = 1$,

$$s_{n+1} \leq s_{n_0}\left(\frac{n_0}{n+1}\right)^c + \frac{\tau c^2}{(n+1)^c}\left(1+\frac{1}{n_0}\right)^c \varphi_{c-1}(n). \tag{32}$$

Proof Note that, for every $n \in \mathbb{N}^*$ and for every integer $m \leq n$:

$$\sum_{k=m}^{n} k^{-\alpha} \geq \varphi_{1-\alpha}(n+1) - \varphi_{1-\alpha}(m), \tag{33}$$

where $\varphi_{1-\alpha}$ is defined by (8). Since all terms in (30) are positive for $n \geq n_0$, by applying the recursion $n - n_0$ times we have

$$s_{n+1} \leq s_{n_0} \prod_{k=n_0}^{n} (1-\eta_k) + \tau \sum_{k=n_0}^{n} \prod_{i=k+1}^{n} (1-\eta_i)\eta_k^2. \tag{34}$$

Let us estimate the first term in the r.h.s. of (34). Since $1 - x \leq \exp(-x)$ for every $x \in \mathbb{R}$, from (33), we derive

$$\begin{aligned} s_{n_0} \prod_{k=n_0}^{n} (1-\eta_k) &= s_{n_0} \prod_{k=n_0}^{n} \Big(1 - \frac{c}{k^\alpha}\Big) \leq s_{n_0} \exp\left(-c \sum_{k=n_0}^{n} k^{-\alpha}\right) \\ &\leq \begin{cases} s_{n_0} \Big(\dfrac{n_0}{n+1}\Big)^c & \text{if } \alpha = 1, \\ \\ s_{n_0} \exp\Big(\dfrac{c}{1-\alpha}(n_0^{1-\alpha} - (n+1)^{1-\alpha})\Big) & \text{if } 0 < \alpha < 1. \end{cases} \end{aligned} \tag{35}$$

To estimate the second term on the right hand side of (34), let us first consider the case $\alpha < 1$, and let $m \in \mathbb{N}^*$ such that $n_0 \leq n/2 \leq m+1 \leq (n+1)/2$. We have

$$\begin{aligned} \sum_{k=n_0}^{n} \prod_{i=k+1}^{n} (1-\eta_i)\eta_k^2 &= \sum_{k=n_0}^{m} \prod_{i=k+1}^{n} (1-\eta_i)\eta_k^2 + \sum_{k=m+1}^{n} \prod_{i=k+1}^{n} (1-\eta_i)\eta_k^2 \\ &\leq \exp\big(-\sum_{i=m+1}^{n} \eta_i\big) \sum_{k=n_0}^{m} \eta_k^2 + \eta_m \sum_{k=m+1}^{n} \left(\prod_{i=k+1}^{n} (1-\eta_i) - \prod_{i=k}^{n} (1-\eta_i)\right) \\ &\leq \exp\big(-\sum_{i=m+1}^{n} \eta_i\big) \sum_{k=n_0}^{m} \eta_k^2 + \eta_m \\ &\leq c^2 \exp\Big(\frac{c}{1-\alpha}((m+1)^{1-\alpha} - (n+1)^{1-\alpha})\Big)\varphi_{1-2\alpha}(n) + \eta_m & (36) \\ &\leq c^2 \exp\Big(\frac{-ct(n+1)^{1-\alpha}}{1-\alpha}\Big)\varphi_{1-2\alpha}(n) + \frac{2^\alpha c}{\mu(n-2)^\alpha}. & (37) \end{aligned}$$

Hence, combining (35) and (37), for $\alpha \in]0,1[$ we get

$$s_{n+1} \leq \Big(\tau c^2 \varphi_{1-2\alpha}(n) + s_{n_0} \exp\Big(\frac{cn_0^{1-\alpha}}{1-\alpha}\Big)\Big) \exp\Big(\frac{-ct(n+1)^{1-\alpha}}{1-\alpha}\Big) + \frac{\tau 2^\alpha c}{(n-2)^\alpha} \tag{38}$$

We next estimate the second term in the right hand side of (34) in the case $\alpha = 1$. We have

$$\sum_{k=n_0}^{n} \prod_{i=k+1}^{n} (1-\eta_i)\eta_k^2 \leq \frac{c^2}{(n+1)^c}\Big(1+\frac{1}{n_0}\Big)^c \varphi_{c-1}(n).$$

Therefore, for $\alpha = 1$, we obtain,

$$s_{n+1} \leq s_{n_0}\Big(\frac{n_0}{n+1}\Big)^c + \frac{\tau c^2}{(n+1)^c}\Big(1+\frac{1}{n_0}\Big)^c \varphi_{c-1}(n), \tag{39}$$

which completes the proof.

We are now ready to prove Theorem 3.4.

Proof (Theorem 3.4) Since $\mu + \nu > 0$, then $A + B$ is strongly monotone at $\overline{w}$. Hence, $\operatorname{zer}(A+B) = \{\overline{w}\}$. Let $n \in \mathbb{N}^*$. Since $\gamma_n A$ is $\gamma_n \nu$-strongly monotone, by [4, Proposition 23.11] $J_{\gamma_n A}$ is $(1+\gamma_n\nu)$-cocoercive, and then

$$\begin{aligned} \|y_n - \overline{w}\|^2 &= \|J_{\gamma_n A}(w_n - \gamma_n \mathfrak{B}_n) - J_{\gamma_n A}(\overline{w} - \gamma_n B\overline{w})\|^2 \\ &\leq \frac{1}{(1+\gamma_n\nu)^2}\|(w_n - \overline{w}) - \gamma_n(\mathfrak{B}_n - B\overline{w})\|^2. \end{aligned}$$

Next, proceeding as in the proof of Proposition 3.1 and recalling (17)-(18), we obtain

$$\begin{aligned} \mathsf{E}[\|y_n - \overline{w}\|^2] \leq &\frac{1}{(1+\gamma_n\nu)^2}\left(\mathsf{E}[\|w_n - \overline{w}\|^2] - \gamma_n\left(2 - \gamma_n \frac{1+2\delta^2\alpha_n}{\beta}\right)\cdot\right. \\ &\left.\cdot\, \mathsf{E}[\langle w_n - \overline{w} \mid Bw_n - B\overline{w}\rangle] + 2\gamma_n^2\delta^2(1+\alpha_n\|B\overline{w}\|^2)\right). \end{aligned} \tag{40}$$

Since B is strongly monotone of parameter μ at $\overline{w}$,

$$\langle Bw_n - B\overline{w} \mid w_n - \overline{w}\rangle \geq \mu\|w_n - \overline{w}\|^2 \,. \tag{41}$$

Therefore, from (40), recalling the definition of ε in $(A3)$, we get

$$\lambda_n \mathsf{E}[\|y_n - \overline{w}\|^2] \leq \frac{\lambda_n}{(1+\gamma_n\nu)^2}\Big((1-\gamma_n\mu\epsilon)\mathsf{E}[\|w_n - \overline{w}\|^2] + 2\delta^2\chi_n^2\Big). \tag{42}$$

Hence, by definition of w_{n+1},

$$\mathsf{E}[\|w_{n+1} - \overline{w}\|^2] \leq \left(1 - \frac{\lambda_n\gamma_n(2\nu + \gamma_n\nu^2 + \mu\epsilon)}{(1+\gamma_n\nu)^2}\right)\mathsf{E}[\|w_n - \overline{w}\|^2] + \frac{2\delta^2\chi_n^2}{(1+\gamma_n\nu)^2}. \tag{43}$$

Now, suppose that $n \geq n_0$. Since $\gamma_n \leq \gamma_{n_0} = c_1 n_0^{-\theta} \leq 1$, we have

$$\frac{\lambda_n\gamma_n(2\nu + \gamma_n\nu^2 + 2\mu\epsilon)}{(1+\gamma_n\nu)^2} \geq \frac{\underline{\lambda}(2\nu + \mu\varepsilon)}{(1+\nu)^2}\gamma_n = cn^{-\theta}. \tag{44}$$

On the other hand,

$$\frac{2\delta^2\chi_n^2}{(1+\gamma_n\nu)^2} \leq 2\delta^2(1+\overline{\alpha}\|B\overline{w}\|^2)c_1^2 n^{-2\theta} \,. \tag{45}$$

Then, putting together (43), (44), and (45), we get

$$\mathsf{E}[\|w_{n+1} - \overline{w}\|^2] \leq (1-\eta_n)\mathsf{E}[\|w_n - \overline{w}\|^2] + \tau\eta_n^2, \tag{46}$$

with $\tau = 2\delta^2 c_1^2(1+\overline{\alpha}\|B\overline{w}\|^2)/c^2$ and $\eta_n = cn^{-\theta}$.

(i)&(ii): Inequalities (10) and (11) follow from (46) by applying Lemma 3.1.

(iii) For $\theta \in \,]0,1[$, the statement follows from (10). For $\theta = 1$, the statement follows from (11) and (8).

4 Special Cases

In this section, we study two special instances of Problem 3.1, namely variational inequalities and minimization problems. Moreover, for variational inequalities, we prove an additional result showing that a suitably defined merit function [35] goes to zero when evaluated on the iterates of the stochastic forward-backward algorithm. This merit function has been used to quantify the inaccuracy of an approximation of the solution in [18].

4.1 Variational Inequalities

In this section we focus on a special case of Problem 3.1, assuming that A is the subdifferential of $G \in \Gamma_0(\mathcal{H})$.

Problem 4.1 Let $B\colon \mathcal{H} \to \mathcal{H}$ be β-cocoercive, for some $\beta \in \left]0,+\infty\right[$, let G be a function in $\Gamma_0(\mathcal{H})$. The problem is to solve the following variational inequality [36,5,4]

$$\text{find } \overline{w} \in \mathcal{H} \text{ such that} \quad (\forall w \in \mathcal{H}) \quad \langle \overline{w} - w \mid B\overline{w} \rangle + G(\overline{w}) \leq G(w), \tag{47}$$

under the assumption that (47) has at least one solution.

Several stochastic algorithms for variational inequalities have been studied on finite dimensional spaces: the sample average approximation [37,38] (see also references therein), the mirror proximal stochastic approximation algorithm [18], and stochastic proximal methods [19,20].

Problem 4.1 reduces to a particular case of Problem 3.1 with $A = \partial G$ and Algorithm 3.1 can be specialized according to the fact that, $(\forall n \in \mathbb{N}^*)$, $J_{\gamma_n A} = \operatorname{prox}_{\gamma_n G}$.

When G is the indicator function of a non-empty, closed, convex subset C of $\mathcal{H}$, Problem 4.1 reduces to the problem of solving a classic variational inequality [36,39], namely to find $\overline{w}$ such that

$$(\forall w \in C) \quad \langle B\overline{w} \mid \overline{w} - w \rangle \leq 0. \tag{48}$$

Proximal algorithms are often used to solve this problem; see [4, Chapter 25] and references therein. Note that, by [40, Lemma 1], since cocoercivity of B implies Lipschitz continuity, $\overline{w}$ is a solution of (48) if and only if

$$(\forall w \in C) \quad \langle Bw \mid \overline{w} - w \rangle \leq 0\,. \tag{49}$$

As it has been done in [18], it is therefore natural to quantify the inaccuracy of a candidate solution $u \in \mathcal{H}$ by the merit function

$$V(u) = \sup_{w \in C} \langle Bw \mid u - w \rangle\,. \tag{50}$$

In particular, note that $(\forall u \in \mathcal{H})$ $V(u) \geq 0$ and $V(u) = 0$ if and only u is a solution of (49). We will consider convergence properties of the following iteration, which differs from the one in Algorithm 3.1 only by the averaging step.

Algorithm 4.1 *Let C be a nonempty bounded closed convex subset of $\mathcal{H}$. Let $(\gamma_t)_{t\in\mathbb{N}^*}$ be a sequence in $]0,+\infty[$. Let $(\lambda_t)_{t\in\mathbb{N}^*}$ be a sequence in $[0,1]$, and let*

$(\mathfrak{B}_t)_{t\in\mathbb{N}^*}$ *be a* $\mathcal{H}$*-valued random process such that* $(\forall n \in \mathbb{N}^*)$ $\mathsf{E}[\|\mathfrak{B}_n\|^2] < +\infty$*. Let* $w_1\colon \Omega \to \mathcal{H}$ *be a random variable such that* $\mathsf{E}[\|w_1\|^2] < +\infty$ *and set*

$$(\forall n \in \mathbb{N}^*) \quad \left\lfloor \begin{array}{l} \text{For } t = 1,\ldots,n \\ \left\lfloor \begin{array}{l} z_t = w_t - \gamma_t \mathfrak{B}_t \\ y_t = P_C z_t \\ w_{t+1} = (1-\lambda_t) w_t + \lambda_t y_t \end{array} \right. \\ \overline{w}_n = \big(\sum_{t=1}^n \gamma_t \lambda_t w_t\big) / \sum_{t=1}^n (\gamma_t \lambda_t). \end{array} \right. \tag{51}$$

The next theorem gives an estimate of the function V when evaluated on the expectation of $\overline{w}_n$. Note that we do not impose any additional monotonicity property on B.

Theorem 4.2 (Ergodic convergence) *In the setting of problem* (47)*, assume that* $G = \iota_C$ *for some nonempty bounded closed convex set* C *in* $\mathcal{H}$*. Let* $(\overline{w}_n)_{n\in\mathbb{N}^*}$ *be the sequence generated by Algorithm 4.1 and suppose that conditions* (A1), (A2), *and* (A3) *hold. Set*

$$\theta_0 = \sup_{u\in C} \frac{1}{2}\mathsf{E}[\|w_1 - u\|^2] \text{ and } \theta_{1,n} = \frac{1}{2}\sum_{t=1}^n \big(\lambda_t \gamma_t^2 (1+\delta^2\alpha_t)\mathsf{E}[\|Bw_t\|^2] + \delta^2 \lambda_t \gamma_t^2\big), \tag{52}$$

then

$$V(\mathsf{E}[\overline{w}_n]) \leq (\theta_0 + \theta_{1,n})\bigg(\sum_{t=1}^n \lambda_t \gamma_t\bigg)^{-1}. \tag{53}$$

Moreover, suppose that the condition (*A*4) *is also satisfied. Then,*

$$\lim_{n\to+\infty} V(\mathsf{E}[\overline{w}_n]) = 0. \tag{54}$$

In particular, if $(\forall t \in \mathbb{N}^*)$ $\lambda_t = 1$ *and* $\gamma_t = t^{-\theta}$ *for some* $\theta \in\,]1/2, 1[$*, we get*

$$V(\mathsf{E}[\overline{w}_n]) = O(n^{\theta-1}). \tag{55}$$

Proof Since C is a non-empty closed convex set, P_C is non-expansive. Hence, from the convexity of $\| \cdot \|^2$, for every $t \in \mathbb{N}^*$ and every $u \in C$,

$$\begin{aligned}\|w_{t+1} - u\|^2 &\leq (1-\lambda_t)\|w_t - u\|^2 + \lambda_t \|P_C(w_t - \gamma_t \mathfrak{B}_t) - P_C u\|^2 \\ &\leq (1-\lambda_t)\|w_t - u\|^2 + \lambda_t \|w_t - u - \gamma_t \mathfrak{B}_t\|^2 \\ &\leq \|w_t - u\|^2 - 2\lambda_t\gamma_t \langle w_t - u \mid \mathfrak{B}_t\rangle + \lambda_t\gamma_t^2 \|\mathfrak{B}_t\|^2.\end{aligned}$$

We derive from conditions $(A1)$ that $\mathsf{E}[\langle w_t - u \mid \mathfrak{B}_t\rangle\, |\mathcal{F}_t] = \langle w_t - u \mid Bw_t\rangle$ and from (A3) that

$$\begin{aligned}&\mathsf{E}[\|\mathfrak{B}_t\|^2|\mathcal{F}_t] \leq \\ &\qquad \leq \mathsf{E}[\|\mathfrak{B}_t - Bw_t\|^2|\mathcal{F}_t] + \mathsf{E}[\|Bw_t\|^2|\mathcal{F}_t] + 2\mathsf{E}[\langle \mathfrak{B}_t - Bw_t \mid Bw_t\rangle\, |\mathcal{F}_t] \\ &\qquad \leq \|Bw_t\|^2 + \delta^2(1 + \alpha_t\|Bw_t\|^2). \qquad (56)\end{aligned}$$

Therefore, (56) and the monotonicity of B yield

$$\begin{aligned}2\lambda_t\gamma_t \langle w_t - u \mid Bu\rangle &\leq \|w_t - u\|^2 - \mathsf{E}[\|w_{t+1} - u\|^2|\mathcal{F}_t] \\ &\qquad + \lambda_t\gamma_t^2(1 + \delta^2\alpha_t)\|Bw_t\|^2) + \delta^2\lambda_t\gamma_t^2, \qquad (57)\end{aligned}$$

which implies that

$$\begin{aligned}&2\mathsf{E}[\langle \overline{w}_n - u \mid Bu\rangle] \\ &\leq \Big(\sum_{t=1}^{n} \lambda_t\gamma_t\Big)^{-1}\Big(\mathsf{E}[\|w_1 - u\|^2 + \sum_{t=1}^{n}\Big(\lambda_t\gamma_t^2(1 + \sigma^2\lambda_t\alpha_t)\mathsf{E}[\|Bw_t\|^2] + \sigma^2\lambda_t\gamma_t^2)\Big)\Big).\end{aligned}$$

Thus, $\sup_{u\in C} \mathsf{E}[\langle \overline{w}_n - u \mid Bu\rangle] \leq (\theta_0 + \theta_{1,n})\big(\sum_{t=1}^{n} \lambda_t\gamma_t\big)^{-1}$, which proves (53). Finally, since C is bounded, $\theta_0 < +\infty$. Now, additionally assume that $(A4)$ is satisfied. Then $\sum_{t=1}^{+\infty} \lambda_t\gamma_t = +\infty$, hence, to get (54), it is enough to prove that $(\theta_0 + \theta_{1,n})_{n\in\mathbb{N}^*}$ is bounded. Since $(A4)$ implies that $\sum_{t=1}^{+\infty} \lambda_t\gamma_t^2 < +\infty$ and

$\sum_{t=1}^{+\infty} \lambda_t \gamma_t^2 \alpha_t < +\infty$, we are left to prove that $(\mathsf{E}[\|Bw_t\|^2])_{t\in\mathbb{N}^*}$ is bounded. This directly follows from Proposition 3.1(i). The last assertion of the statement follows from (53) with, for every $t \in \mathbb{N}^*$, $\gamma_t = t^{-\theta}$ and $\lambda_t = 1$.

Remark 4.1 Under slightly different assumptions, an alternative method to solve Problem 4.1 is the mirror-prox algorithm in [18]. With respect to forward-backward, the mirror-prox algorithm requires two projections per iteration, rather than one. With such procedure, $\mathsf{E}[V(\overline{w}_n)] \to 0$; see [18]. In general, $V(\mathsf{E}[\overline{w}_n]) \leq \mathsf{E}[V(\overline{w}_n)]$.

4.2 Minimization Problems

In this section, we specialize the results in Section 3 to minimization problems. In the special case of composite minimization, stochastic implementations of first order methods received much attention [16,17,27,41–43] for the ease of implement and the low memory requirement of each iteration. In particular, [42] derives an optimal rate of convergence for the objective function values. Similar accelerated algorithms have been also studied in the machine learning community [44–49].

Problem 4.2 Let $\beta \in \,]0, +\infty[$, let $G \in \Gamma_0(\mathcal{H})$, and let $F\colon \mathcal{H} \to \mathbb{R}$ be a convex differentiable function, with a β^{-1}-Lipschitz continuous gradient. The problem is to

$$\underset{w\in\mathcal{H}}{\text{minimize}}\; F(w) + G(w), \tag{58}$$

under the assumption that the set of solution to (58) is non-empty.

Problem 4.2 is a specific instance of Problem 3.1, with $A = \partial G$ and $B = \nabla F$. Indeed, ∇F is cocoercive thanks to the Baillon-Haddad Theorem [4, Corollary 18.16]. Algorithm 3.1 can therefore be specialized to the minimization setting, with, for every $n \in \mathbb{N}^*$, $J_{\gamma_n A} = \operatorname{prox}_{\gamma_n G}$. When G is an indicator function, and $(\forall n \in \mathbb{N}^*)$ $\lambda_n = 1$, related results have been obtained in [50]. Theorem 3.4 applied to Problem 4.2 is the extension to the nonsmooth case of [33, Theorem 1]. Algorithm 3.1 for minimization is closely related to the FOBOS algorithm studied in [16] (see also [17]). The main difference with these papers is that our convergence results consider convergence of the iterates with no averaging, without boundedness assumptions. The asymptotic rate $O(n^{-1})$ for the iterates improves the $O((\log n)/n)$ rate derived from [16, Corollary 10] for the average of the iterates and it coincides with the one that can be derived by applying optimal methods [42], and the methods in [51–54]. In stochastic optimization, the study of almost sure convergence has a long history; see e.g. [55–58] and references therein. Recent results on convergence of projected stochastic gradient algorithm can be found in [59,60,34].

5 Conclusions

We studied a stochastic version of the forward-backward splitting algorithm, providing various convergence results in the strongly and uniformly monotone case. The monotone inclusions framework is key to derive convergence of primal-dual algorithms in the deterministic setting, and we believe that the extension to the stochastic case is an interesting research direction.

Acknowledgements This material is based upon work supported by CBMM, funded by NSF STC award CCF-1231216. L. Rosasco acknowledges the financial support of the Italian Ministry of Education, University and Research FIRB project RBFR12M3AC and B. C. Vu the Vietnam National Foundation for Science and Technology Development (NAFOSTED), Grant No. 102.01-2014.02. S. Villa is member of the Gruppo Nazionale per l'Analisi Matematica, la Probabilità e le loro Applicazioni (GNAMPA) of the Istituto Nazionale di Alta Matematica (INdAM).

References

1. Minty, G.J.: Monotone (nonlinear) operators in Hilbert space. Duke Math. J. **29**, 341–346 (1962)
2. Brézis, H.: Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert. North-Holland Publishing Co., Amsterdam-London; American Elsevier Publishing Co., Inc., New York (1973)
3. Pascali, D., Sburlan, S.: Nonlinear mappings of monotone type. Martinus Nijhoff Publishers, The Hague; Sijthoff & Noordhoff International Publishers, Alphen aan den Rijn (1978)
4. Bauschke, H.H., Combettes, P.L.: Convex analysis and monotone operator theory in Hilbert spaces. Springer, New York (2011)
5. Zeidler, E.: Nonlinear functional analysis and its applications. II/B. Springer-Verlag, New York (1990)
6. Rockafellar, R.T.: On the maximal monotonicity of subdifferential mappings. Pacific J. Math. **33**, 209–216 (1970)
7. Rockafellar, R.T.: Monotone operators associated with saddle-functions and minimax problems. In: Nonlinear Functional Analysis (Proc. Sympos. Pure Math., Vol. XVIII, Part 1, Chicago, Ill., 1968), pp. 241–250. Amer. Math. Soc., Providence, R.I. (1970)
8. Combettes, P.L.: Solving monotone inclusions via compositions of nonexpansive averaged operators. Optim. **53**(5-6), 475–504 (2004)

9. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979)
10. Passty, G.B.: Ergodic convergence to a zero of the sum of monotone operators in Hilbert space. J. Math. Anal. Appl. **72**(2), 383–390 (1979)
11. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
12. Combettes, P.L., Vũ, B.C.: Variable metric quasi-Fejér monotonicity. Nonlinear Anal. **78**, 17–31 (2013)
13. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (electronic) (2005)
14. Nesterov, Y.: Gradient methods for minimizing composite objective function. CORE Discussion Paper 2007/76, Catholic University of Louvain (2007)
15. Villa, S., Salzo, S., Baldassarre, L., Verri, A.: Accelerated and inexact forward-backward algorithms. SIAM J. Optim. **23**(3), 1607–1633 (2013)
16. Duchi, J., Singer, Y.: Efficient online and batch learning using forward backward splitting. J. Mach. Learn. Res. **10**, 2899–2934 (2009)
17. Atchade, Y.F., Fort, G., Moulines, E.: On stochastic proximal gradient algorithms. arXiv:1402.2365 (2014)
18. Juditsky, A., Nemirovski, A., Tauvel, C.: Solving variational inequalities with stochastic mirror-prox algorithm. Stoch. Syst. **1**(1), 17–58 (2011)
19. Jiang, H., Xu, H.: Stochastic approximation approaches to the stochastic variational inequality problem. IEEE Trans. Automat. Control **53**(6), 1462–1475 (2008)
20. Koshal, J., Nedic, A., Shanbhag, U.V.: Regularized iterative stochastic approximation methods for stochastic variational inequality problems. IEEE Trans. Automat. Contr. **58**(3), 594–609 (2013)
21. Polyak, B.: Introduction to Optimization. Optimization Software, New York (1987)
22. Ermol'ev, Y.M.: On the method of generalized stochastic gradients and quasi-Fejér sequences. Cybernetics **5**(2), 208–220 (1969)
23. Ermol'ev, Y.M., Tuniev, A.D.: Random Fejér and quasi-Fejér sequences. Theory of Optimal Solutions– Akademiya Nauk Ukrainskoĭ SSR Kiev **2**, 76–83 (1968)

24. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM J. Optim. **19**(4), 1574–1609 (2008)
25. Combettes, P.L., Pesquet, J.C.: Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. SIAM J. Optim **25**(2), 1221–1248 (2015)
26. Combettes, P.L., Pesquet, J.C.: Stochastic approximations and perturbations in forward-backward splitting for monotone operators. Pure Appl. Funct. Anal. (to appear)
27. Lin, Q., Chen, X., Peña, J.: A sparsity preserving stochastic gradient methods for sparse regression. Comput. Optim. Appl. **to appear** (2014)
28. Moreau, J.J.: Fonctions convexes duales et points proximaux dans un espace hilbertien. C. R. Acad. Sci. Paris **255**, 2897–2899 (1962)
29. Ermol′ev, J.M.: The convergence of random quasi-Féjer sequences. Kibernetika (Kiev) (4), 70–71 (1971)
30. Combettes, P.L.: Solving monotone inclusions via compositions of nonexpansive averaged operators. Optimization **53**, 475–504 (2004)
31. Attouch, H., Czarnecki, M., Peypouquet, J.: Coupling forward-backward with penalty schemes and parallel splitting for constrained variational inequalities. SIAM J. Optim. **21**(4), 1251–1274 (2011)
32. Bertsekas, D.P., Tsitsiklis, J.N.: Gradient convergence in gradient methods with errors. SIAM J. Optim. **10**(3), 627–642 (electronic) (2000)
33. Bach, F., Moulines, E.: Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In: Advances in Neural Information Processing Systems, vol. 24 (2011)
34. Barty, K., Roy, J.S., Strugarek, C.: Hilbert-valued perturbed subgradient algorithms. Math. Oper. Res. **32**(3), 551–562 (2007)
35. Auslender, A.: Optimisation. Masson, Paris-New York-Barcelona (1976)
36. Lions, J.L., Stampacchia, G.: Variational inequalities. Comm. Pure Appl. Math. **20**, 493–519 (1967)
37. Shapiro, A.: Monte Carlo sampling methods. In: Stochastic programming, *Handbooks Oper. Res. Management Sci.*, vol. 10, pp. 353–425. Elsevier Sci. B. V., Amsterdam (2003)

38. Chen, X., Wets, R.J.B., Zhang, Y.: Stochastic variational inequalities:residual minimization smoothing sample average approximations. SIAM J. Optim. **22**(2), 649–673 (2012)

39. Martinet, B.: Regularisation d'inequations variationelles par approximations successives. Revue Francaise d?Informatique et de Recherche Opérationelle **4**, 154–159 (1970)

40. Browder, F.E.: Nonlinear monotone operators and convex sets in Banach spaces. Bull. Amer. Math. Soc. **71**, 780–785 (1965)

41. Devolder, O.: Stochastic first order methods in smooth convex optimization. Tech. rep., Center for Operations Research and econometrics (2011)

42. Lan, G.: An optimal method for stochastic composite optimization. Math. Program. **133**(1-2, Ser. A), 365–397 (2012)

43. Duchi, J., Agarwal, A., Johansson, M., Jordan, M.: Ergodic mirror descent. SIAM J. Optim. **22**(4), 1549–1578 (2012)

44. Kwok, J.T., C. Hu, Pan, W.: Accelerated gradient methods for stochastic optimization and online learning. In: Advances in Neural Information Processing Systems, vol. 22, pp. 781–789 (2009)

45. Xiao, L.: Dual averaging methods for regularized stochastic learning and online optimization. J. Mach. Learn. Res. **11**, 2543–2596 (2010)

46. Bottou, L., Le Cun, Y.: Online learning for very large data sets. Appl. Stoch. Model. Bus. **21**(2), 137–151 (2005)

47. Shalev-Shwartz, S., Singer, Y., Srebro, N.: Pegasos: Primal estimated sub-gradient solver for SVM. In: Proceedings ICML (2007)

48. Shalev-Shwartz, S., Srebro, N.: SVM optimization: inverse dependence on training set size. In: Proceedings ICML (2008)

49. Zhang, T.: Multi-stage convex relaxation for learning with sparse regularization. In: Advances in Neural Information Processing Systems, pp. 1929–1936 (2008)

50. Yousefian, F., Nedić, A., Shanbhag, U.V.: On stochastic gradient and subgradient methods with adaptive steplength sequences. Automatica J. IFAC **48**(1), 56–67 (2012)

51. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: a generic algorithmic framework. SIAM J. Optim. **22**(4), 1469–1492 (2012)

52. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization ii: shrinking procedures and optimal algorithms. SIAM J. Optim. **23**(4), 2061–2089 (2013)

53. Hazan, E., Kale, S.: Beyond the regret minimization barrier:optimal algorithms for stochastic strongly-convex optimization. J. Mach. Learn. Res. **15**, 2489–2512 (2014)

54. Juditsky, A., Nesterov, Y.E.: Deterministic and stochastic primal-dual subgradient methods for minimizing uniformly convex functions. Stochastic Systems **4**(1), 44–80 (2014)

55. Robbins, H., Siegmund, D.: A convergence theorem for non negative almost supermartingales and some applications. In: Optimizing methods in statistics, pp. 233–257. Academic Press, New York (1971)

56. Kushner, H.J., Clark, D.S.: Stochastic approximation methods for constrained and unconstrained systems, vol. 26. Springer-Verlag, New York (1978)

57. Benveniste, A., Métivier, M., Priouret, P.: Adaptive algorithms and stochastic approximations. Springer-Verlag, Berlin (1990)

58. Chen, X., White, H.: Asymptotic properties of some projection-based robbins-monro procedures in a hilbert space. Stud. Nonlinear Dyn. Econom. **6**, 1–53 (2002)

59. Bennar, A., Monnez, J.M.: Almost sure convergence of a stochastic approximation process in a convex set. Int. J. Appl. Math. **20**(5), 713–722 (2007)

60. Monnez, J.M.: Almost sure convergence of stochastic gradient processes with matrix step sizes. Statist. Probab. Lett. **76**(5), 531–536 (2006)

61. Combettes, P.L., Pesquet, J.C.: Proximal thresholding algorithm for minimization over orthonormal bases. SIAM J. Optim. **18**(4), 1351–1376 (2007)

62. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B **58**(1), 267–288 (1996)