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

A Fully Stochastic Primal-Dual Algorithm

Adil Salim · Pascal Bianchi · Walid Hachem

Received: date / Accepted: date

Abstract A new stochastic primal-dual algorithm for solving a composite optimization problem is proposed. It is assumed that all the functions/matrix used to define the optimization problem are given as statistical expectations. These expectations are unknown but revealed across time through i.i.d realizations. This covers the case of convex optimization under stochastic linear constraints. The proposed algorithm is proven to converge to a saddle point of the Lagrangian function. In the framework of the monotone operator theory, the convergence proof relies on recent results on the stochastic Forward Backward algorithm involving random monotone operators.

1 Introduction

Many applications in machine learning, statistics or signal processing require the solution of the following optimization problem [1,2,3]. Given two Euclidean spaces $\mathcal{X}$ and $\mathcal{V}$, solve

$$\min_{x\in\mathcal{X}} \mathsf{F}(x) + \mathsf{G}(x) + \mathsf{H}(\mathsf{L}x) \tag{1}$$

where F, G and H are lower semicontinuous convex functions such that $\mathsf{F}(x) < \infty$ for every x and L belongs to the set $\mathcal{L}(\mathcal{X}, \mathcal{V})$ of $\mathcal{X} \to \mathcal{V}$ linear operators. Consider

This work was partially supported by the Labex Digiteo-DigiCosme (OPALE project), Université Paris-Saclay.

Adil Salim, Pascal Bianchi
Télécom ParisTech
46 rue Barrault, 75013 Paris
E-mail: asalim.math@gmail.com, pascal.bianchi@telecom-paristech.fr

Walid Hachem
CNRS, Université Paris-Est Marne-la-Vallée
5 blvd. Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2
E-mail: walid.hachem@u-pem.fr

$\mathsf{c} \in \mathcal{V}$ and $\mathsf{H} = \iota_{\{\mathsf{c}\}}$, where $\iota_{\mathcal{C}}$ is the indicator function of the set $\mathcal{C}$, *i.e*, the function equal to 0 on $\mathcal{C}$ and $+\infty$ elsewhere. In this particular case, Problem (1) boils down to the linearly constrained problem

$$\min_{x \in \mathcal{X}} \mathsf{F}(x) + \mathsf{G}(x) \quad \text{s.t.} \quad \mathsf{L}x = \mathsf{c}. \tag{2}$$

In order to solve Problem (1), primal-dual methods generate a sequence of primal estimates $(x_n)_{n\in\mathbb{N}}$ and a sequence of dual estimates $(\lambda_n)_{n\in\mathbb{N}}$ jointly converging to a saddle point of the Lagrangian function. As is well known, the qualification condition

$$\mathsf{c} \in \mathrm{ri}\,(\mathsf{L}\,\mathrm{dom}\,\mathsf{G} - \mathrm{dom}\,\mathsf{H})$$

where ri is the relative interior of a set, ensures the existence of such a point [4]. There is a rich literature on such algorithms which cannot be exhaustively listed [1,2,3].

In this paper, it is assumed that the quantities that enter the minimization problem are likely to be unavailable or difficult to compute numerically. More precisely, it is assumed that the functions F and G are defined as expectations of random functions. Given a probability space $(\Xi, \mathscr{G}, \mu)$, consider two convex normal integrands (see below) $f : \Xi \times \mathcal{X} \to \mathbb{R}$ and $g : \Xi \times \mathcal{V} \to (-\infty, +\infty]$. Then, we consider that $\mathsf{F}(x) = \mathbb{E}_\mu(f(\cdot, x))$ and $\mathsf{G}(x) = \mathbb{E}_\mu(g(\cdot, x))$. In addition, let L be a measurable function from $(\Xi, \mathscr{G}, \mu)$ to $\mathcal{L}(\mathcal{X}, \mathcal{V})$ (*i.e* a random matrix), then it is assumed that $\mathsf{L} = \mathbb{E}_\mu L(\cdot)$. Finally, the Fenchel conjugate $\mathsf{H}^\star$ of H takes the form $\mathsf{H}^\star(\lambda) = \mathbb{E}_\mu(p(\cdot, \lambda))$, where p is a normal convex integrand.

In the particular case of Problem (2), let us assume that $\mathsf{c} = \mathbb{E}_\mu(c(\cdot))$ where $c(\cdot) : \Xi \to \mathcal{V}$ is a random vector. Then, since $\mathsf{H}^\star(\lambda) = \langle \lambda, \mathsf{c} \rangle$, we simply put $p(\cdot, \lambda) = \langle \lambda, c(\cdot) \rangle$.

In order to solve Problem (1), the observer is given the functions f, g, p, and L, along with a sequence of independent and identically distributed (i.i.d.) random variables (ξ_n) with the probability distribution μ. In this paper, a new stochastic primal dual algorithm based on this data is proposed to solve this problem.

The convergence proof for this algorithm relies on the monotone operator theory. The algorithm is built around an instantiation of the stochastic Forward-Backward algorithm involving random monotone operators that was introduced in [5]. It is proven that the weighted means of the iterates of the algorithm, where the weights are given by the step sizes of the algorithm, converges almost surely to a saddle point of the Lagrangian function. To the authors knowledge, the proposed algorithm is the first method that allows to solve Problem (1) in a fully stochastic setting. Existing methods typically allow to handle subproblems of Problem (1) in which some quantities used to define (1) are assumed to be available or set equal to zero [6,7,8,9,10]. In particular, the new algorithm generalizes the stochastic gradient algorithm (in the case where only F is non zero), the stochastic proximal point algorithm [11,10,12] (only G is non zero), and the stochastic proximal gradient algorithm [13,14] (only $\mathsf{F} + \mathsf{G}$ is non zero).

To our knowledge, the proposed algorithm is also one of the first methods that allows to tackle stochastic linear constraints. The paper [8] studies stochastic inequality constraints for optimization over a compact set and provide regret bounds. Handling stochastic constraints online is suitable in various fields of machine learning like Neyman-Pearson classification or online portfolio optimization. For example, the Markowitz portfolio optimization problem is an instance of Problem (2) where ξ is a random variable with values in $\mathcal{X}$, $\mathsf{F}(x) = \mathbb{E}_\xi(\langle x, \xi\rangle^2)$, $\mathsf{G}(x) = \iota_\Delta(x)$ where Δ is the probability simplex, $\mathsf{L} = \mathbb{E}_\xi(\xi^T)$ and c is some real positive number. In this case, authors usually assume that $\mathsf{L} = \mathbb{E}_\xi(\xi^T)$ is fully known or estimated. The paper is organized as follows. The next section is devoted to rigorously state the main problem and the main algorithm. In section 3 the convergence proof of the algorithm is given.

*Some notations.*The notation $\mathscr{B}(\mathcal{X})$ will refer to the Borel σ-field of $\mathcal{X}$. Both the operator norm and the Euclidean norm will be denoted as $\|\cdot\|$. The distance of a point x to a set S is denoted as $\mathrm{dist}(x, S)$. As mentioned above, we denote as $\mathcal{L}(\mathcal{X}, \mathcal{V})$ the set of linear operators, identified with matrices, from $\mathcal{X}$ to $\mathcal{V}$. The set of proper, lower semicontinuous convex functions on $\mathcal{X}$ is $\Gamma_0(\mathcal{X})$.

2 Problem description

Before entering our subject, we recall some definitions regarding set-valued functions and integrals. Let $(\Xi, \mathscr{G}, \mu)$ be a probability space where the σ-field $\mathscr{G}$ is μ-complete. Given a Euclidean space $\mathcal{X}$, let $h : \Xi \rightrightarrows \mathcal{X}$ be a set valued function such that $h(s)$ is a closed set for each $s \in \Xi$. The function h is said measurable if $\{s \,:\, h(s) \cap S \neq \emptyset\} \in \mathscr{G}$ for any set $S \in \mathscr{B}(\mathcal{X})$. An equivalent definition for the mesurability of h requires that the domain $\mathrm{dom}(h) := \{s \in \Xi \,:\, h(s) \neq \emptyset\}$ of h belongs to $\mathscr{G}$, and that there exists a sequence of measurable functions $\varphi_n : \mathrm{dom}(h) \to \mathcal{X}$ such that $h(s) = \mathrm{cl}\,\{\varphi_n(s)\}_n$ for all $s \in \mathrm{dom}(h)$, where cl is the closure of a set. Such functions are called measurable selections of h. Assume now that h is measurable and that $\mu(\mathrm{dom}(h)) = 1$. Given $1 \leq p < \infty$, let $\mathscr{L}^p(\mu)$ be the space of the $\mathscr{G}$-measurable functions $\varphi : \Xi \to \mathcal{X}$ such that $\int \|\varphi\|^p d\mu < \infty$, and let

$$\mathfrak{S}_h^p := \{\varphi \in \mathscr{L}^p(\mu) \,:\, \varphi(s) \in h(s) \ \mu - \text{almost everywhere (a.e.)}\}\,.$$

If $\mathfrak{S}_h^1 \neq \emptyset$, the function h is said integrable. The selection integral of h is the set

$$\int h d\mu := \mathrm{cl}\left\{\int_\Xi \varphi d\mu \;:\; \varphi \in \mathfrak{S}_h^1\right\}. \tag{3}$$

In all the remainder, given a single-valued or a set-valued function h, the notation $\mathbb{E}_\mu h$ will refer to the integral of h with respect to μ. The meaning of this integral will be clear from the context.

We now state our problem. A function $h : \Xi \times \mathcal{X} \to (-\infty, \infty]$ is said a convex normal integrand if $h(s, \cdot)$ is convex, and if the set-valued mapping

$s \mapsto \operatorname{epi} h(s,\cdot)$ is closed-valued and measurable, where epi is the epigraph of a function. Let $f : \Xi \times \mathcal{X} \to (-\infty, \infty]$ be a convex normal integrand, and assume that $\int |f(s,x)|\, \mu(ds) < \infty$ for all $x \in \mathcal{X}$. Consider the convex function $\mathsf{F}(x)$ defined on $\mathcal{X}$ as the Lebesgue integral $\mathsf{F}(x) = \mathbb{E}_\mu f(\cdot, x)$. Denoting as $\partial f(s,x)$ the subdifferential of $f(s,\cdot)$ with respect to x, it is known that the set-valued function $\partial f(\cdot, x)$ is measurable, $\mathfrak{S}^1_{\partial f(\cdot,x)} \neq \emptyset$, and $\partial \mathsf{F}(x) = \mathbb{E}_\mu \partial f(\cdot, x)$ for each $x \in \mathcal{X}$, where the integral is the selection integral defined above [15,16].

Let $g : \Xi \times \mathcal{X} \to (-\infty, \infty]$ be another convex normal integrand, and let $\mathsf{G}(x) = \mathbb{E}_\mu g(\cdot, x)$, where the integral is defined as the sum

$$\int_{\{s\,:\, g(s,x) \in [0,\infty)\}} g(s,x)\, \mu(ds) + \int_{\{s\,:\, g(s,x) \in]-\infty, 0[\}} g(s,x)\, \mu(ds) + I(x)\,,$$

and

$$I(x) = \begin{cases} +\infty, \text{ if } \mu(\{s : g(s,x) = \infty\}) > 0, \\ 0, \quad \text{otherwise}\,, \end{cases}$$

and where the convention $(+\infty) + (-\infty) = +\infty$ is used. The function G is a lower semi continuous convex function if $\mathsf{G}(x) > -\infty$ for all x, which we assume. We shall also assume that G is proper. Note that this implies that $g(s,\cdot) \in \Gamma_0(\mathcal{X})$ for μ-almost all s. It is also known that $\partial g(\cdot, x)$ is measurable for each x [15]. We assume that $\partial \mathsf{G}(x) = \mathbb{E}_\mu \partial g(\cdot, x)$, where the right hand member is set to $\emptyset$ for the values of x for which $\mathfrak{S}^1_{\partial g(\cdot,x)} = \emptyset$. Before proceeding in the problem statement, it is useful to provide sufficient conditions under which this interchange of the expectation and the subdifferentiation is possible. By [16], this will be the case if the following conditions hold: *i)* the set-valued mapping $s \mapsto \operatorname{cl} \operatorname{dom} g(s,\cdot)$ is constant μ-a.e., where $\operatorname{dom} g(s,\cdot)$ is the domain of $g(s,\cdot)$, *ii)* $\mathsf{G}(x) < \infty$ whenever $x \in \operatorname{dom} g(s,\cdot)$ μ-a.e., *iii)* there exists $x_0 \in \mathcal{X}$ at which G is finite and continuous. Another case where this interchange is permitted is the following. Let m be a positive integer, and let $\mathcal{C}_1, \ldots \mathcal{C}_m$ be a collection of closed and convex subsets of $\mathcal{X}$. Let $\mathcal{C} = \cap_{i=k}^m \mathcal{C}_k \neq \emptyset$, and assume that the normal cone $N_{\mathcal{C}}(x)$ of $\mathcal{C}$ at x satisfies the identity $N_{\mathcal{C}}(x) = \sum_{k=1}^m N_{\mathcal{C}_k}(x)$ for each $x \in \mathcal{X}$, where the summation is the usual set summation. As is well known, this identity holds true under a qualification condition of the type $\cap_{k=1}^m \operatorname{ri} \mathcal{C}_k \neq \emptyset$ (see also [17] for other conditions). Now, assume that $\Xi = \{1, \ldots, m\}$ and that μ is an arbitrary probability measure putting a positive weight on each $\{k\} \subset \Xi$. Let $g(s,x)$ be the indicator function

$$g(s,x) = \iota_{\mathcal{C}_s}(x) \text{ for } (s,x) \in \Xi \times \mathcal{X}. \tag{4}$$

Then it is obvious that g is a convex normal integrand, $\mathsf{G} = \iota_{\mathcal{C}}$, and $\partial \mathsf{G}(x) = \mathbb{E}_\mu \partial g(\cdot, x)$. We can also combine these two types of conditions: let $(\Sigma, \mathscr{T}, \nu)$ be a probability space, where $\mathscr{T}$ is ν-complete, and let $h : \Sigma \times \mathcal{X} \to (-\infty, \infty]$ be a convex normal integrand satisfying the conditions *i)–iii)* above. Consider the closed and convex sets $\mathcal{C}_1, \ldots, \mathcal{C}_m$ introduced above, and let α be a probability measure on the set $\{0, \ldots, m\}$ such that $\alpha(\{k\}) > 0$ for each $k \in \{0, \ldots, m\}$. Now, set $\Xi = \Sigma \times \{0, \ldots, m\}$, $\mu = \nu \otimes \alpha$, and define $g : \Xi \times \mathcal{X} \to (-\infty, \infty]$ as

$$g(s,x) = \begin{cases} \alpha(0)^{-1} h(u,x) \text{ if } k = 0, \\ \iota_{\mathcal{C}_k}(x) \qquad\quad \text{otherwise}, \end{cases}$$

where $s = (u, k) \in \Sigma \times \{0, \dots, m\}$. Then it is clear that

$$\mathsf{G}(x) = \frac{1}{\alpha(0)} \int_{\Sigma} h(u, x)\nu(du) + \iota_{\mathcal{C}}(x)\,,$$

and

$$\partial \mathsf{G}(x) = \mathbb{E}_{\mu} \partial g(\cdot, x) = \frac{1}{\alpha(0)} \mathbb{E}_{\nu} \partial h(\cdot, x) + \sum_{k=1}^{m} N_{\mathcal{C}_k}(x)\,.$$

To proceed with our problem statement, we introduce another convex normal integrand $p : \Xi \times \mathcal{Z} \to (-\infty, \infty]$ and assume that the function p has verbatim the same properties as g, after replacing the space $\mathcal{X}$ with $\mathcal{V}$. We also denote H the Fenchel conjugate of $\mathsf{P}(\lambda) = \mathbb{E}_{\mu} p(\cdot, \lambda)$, so that $\mathsf{H}^{\star}(\lambda) = \mathbb{E}_{\mu} p(\cdot, \lambda)$.

Finally, let $L : \Xi \to \mathcal{L}(\mathcal{X}, \mathcal{V})$ be an operator-valued measurable function. Let us assume that $\|L\|$ is μ-integrable, and let us introduce the Lebesgue integral $\mathsf{L} = \mathbb{E}_{\mu} L$.

Having introduced these functions, our purpose is to find a solution $x \in \mathcal{X}$ of Problem (1), where the set of such points is assumed non empty. To solve this problem, the observer is given the functions f, g, p, L, and a sequence of i.i.d random variables $(\xi_n)_{n \in \mathbb{N}}$ from a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to $(\Xi, \mathscr{G})$ with the probability distribution μ.

Denote as $\mathrm{prox}_h(x) = \arg\min_{y \in \mathcal{X}} h(y) + \|y - x\|^2/2$ the Moreau's proximity operator of a function $h \in \Gamma_0(\mathcal{X})$. We also denote as $\partial_0 h(x)$ the least norm element of the set $\partial h(x)$, which is known to exist and to be unique [4]. Similarly, $\partial_0 f(s, x)$ will refer to the least norm element of $\partial f(s, x)$ which was introduced above. We shall also denote as $\widetilde{\nabla} f(s, x)$ a measurable subgradient of $f(s, \cdot)$ at x. More precisely, $\widetilde{\nabla} f : (\Xi \times \mathcal{X}, \mathscr{G} \otimes \mathscr{B}(\mathcal{X})) \to (\mathcal{X}, \mathscr{B}(\mathcal{X}))$ is a measurable function such that for each $x \in \mathcal{X}$, $\widetilde{\nabla} f(\cdot, x) \in \mathfrak{S}^1_{\partial f(\cdot, x)}$ (recall that this set is non empty). A possible choice for $\widetilde{\nabla} f(s, x)$ is $\partial_0 f(s, x)$ (see [5, §2.3 and §3.1] for the measurability issues). Turning back to Problem (1), our purpose will be to find a saddle point of the Lagrangian $(x, \lambda) \mapsto \mathsf{F}(x) + \mathsf{G}(x) - \mathsf{H}^{\star}(\lambda) + \langle \mathsf{L}x, \lambda \rangle$. Denoting as $\mathcal{S} \subset \mathcal{X} \times \mathcal{V}$ the set of these saddle points, an element (x, λ) of $\mathcal{S}$ is characterized by the inclusions

$$\begin{cases} 0 \in \partial \mathsf{F}(x) + \partial \mathsf{G}(x) + \mathsf{L}^T \lambda, \\ 0 = \quad -\mathsf{L}x \quad\quad + \partial \mathsf{H}^{\star}(\lambda)\,. \end{cases} \tag{5}$$

Consider a sequence of positive weights $(\gamma_n)_{n \in \mathbb{N}}$. The algorithm proposed here consists in the following iterations applied to the random vector $(x_n, \lambda_n) \in \mathcal{X} \times \mathcal{V}$.

We also give the instance of the main algorithm that allows to solve Problem (2) (which is a instance of Problem (1)).

The convergence of Algorithm 1 is stated by the following theorem.

Theorem 1 *Consider the Problem* (1)*, and let the following assumptions hold true.*

1. *The step size sequence satisfies* $(\gamma_n) \in \ell^2 \setminus \ell^1$*, and* $\gamma_{n+1}/\gamma_n \to 1$ *as* $n \to \infty$.

Algorithm 1 The Main Algorithm : Solving Problem (1)

$$x_{n+1} = \mathrm{prox}_{\gamma_{n+1} g(\xi_{n+1},\cdot)} \left(x_n - \gamma_{n+1} (\widetilde{\nabla} f(\xi_{n+1}, x_n) + L(\xi_{n+1})^T \lambda_n) \right),$$
$$\lambda_{n+1} = \mathrm{prox}_{\gamma_{n+1} p(\xi_{n+1},\cdot)} \left(\lambda_n + \gamma_{n+1} L(\xi_{n+1}) x_n \right) .$$

Algorithm 2 Stochastic Linear Constraints : Solving Problem (2)

$$x_{n+1} = \mathrm{prox}_{\gamma_{n+1} g(\xi_{n+1},\cdot)} \left(x_n - \gamma_{n+1} (\widetilde{\nabla} f(\xi_{n+1}, x_n) + L(\xi_{n+1})^T \lambda_n) \right),$$
$$\lambda_{n+1} = \lambda_n + \gamma_{n+1} \left(L(\xi_{n+1}) x_n - c(\xi_{n+1}) \right) .$$

2. *There exists an integer $m \geq 2$ that satisfies the following conditions:*
   - *The function L is in $\mathscr{L}^{2m}(\mu)$.*
   - *There exists a point $(x_\star, \lambda_\star) \in \mathcal{S}$, and three functions $\varphi_f \in \mathfrak{S}^{2m}_{\partial f(\cdot, x_\star)}$, $\varphi_g \in \mathfrak{S}^{2m}_{\partial g(\cdot, x_\star)}$, and $\varphi_p \in \mathfrak{S}^{2m}_{\partial p(\cdot, \lambda_\star)}$ which*

$$\mathbb{E}_\mu \varphi_f + \mathbb{E}_\mu \varphi_g + \mathsf{L}^T \lambda_\star = 0, \ \textit{and} \ \ - \mathsf{L} x_\star + \mathbb{E}_\mu \varphi_p = 0. \tag{6}$$

   The last assumption is verified for $m = 1$ and for each point $(x_\star, \lambda_\star) \in \mathcal{S}$.
3. *For any compact set K of $\mathrm{dom}\, \partial \mathsf{G}$, there exist $\varepsilon \in (0, 1]$ and $x_0 \in \mathrm{dom}\, \partial \mathsf{G}$ such that*

$$\sup_{x \in K} \mathbb{E} \|\partial_0 g(\cdot, x)\|^{1+\varepsilon} < +\infty, \ \textit{and} \ \ \mathbb{E} \|\partial_0 g(\cdot, x_0)\|^{1+1/\varepsilon} < +\infty.$$

4. *Writing $D_{\partial g}(s) = \mathrm{dom}\, \partial g(s, \cdot)$, there exists $C > 0$ such that for all $x \in \mathcal{X}$,*

$$\mathbb{E}_\mu \, \mathrm{dist}(x, D_{\partial g}(\cdot))^2 \geq C \, \mathrm{dist}(x, \mathrm{dom}\, \partial \mathsf{G})^2.$$

5. *There exists $C > 0$ such that for any $x \in \mathcal{X}$ and any $\gamma > 0$,*

$$\int \| \mathrm{prox}_{\gamma g(s,\cdot)}(x) - \mathit{\Pi}_g(s, x)\|^4 \mu(ds) \leq C \gamma^4 (1 + \|x\|^{2m}),$$

   where $\mathit{\Pi}_g(s, \cdot)$ is the projection operator onto $\mathrm{cl}(\mathrm{dom}\, \partial g(s, \cdot))$, and where m is the integer provided by Assumption 2.

   Assumptions similar to 3–5 are made on the function p and P.

6. *There exists a measurable $\Xi \to \mathbb{R}_+$ function β such that β^{2m} is μ-integrable, where m is the integer provided by Assumption 2, and such that for all $x \in \mathcal{X}$,*

$$\|\widetilde{\nabla} f(s, x)\| \leq \beta(s)(1 + \|x\|).$$

   Moreover, there exists a constant $C > 0$ such that $\mathbb{E}_\mu \|\widetilde{\nabla} f(\cdot, x)\|^4 \leq C(1 + \|x\|^{2m})$.

Consider the sequence of iterates (x_n, λ_n) *produced by the algorithm* (1), *and define the averaged estimates*

$$\bar{x}_n = \frac{\sum_{k=1}^n \gamma_k x_k}{\sum_{k=1}^n \gamma_k}, \text{ and } \bar{\lambda}_n = \frac{\sum_{k=1}^n \gamma_k \lambda_k}{\sum_{k=1}^n \gamma_k}.$$

Then, the sequence (x_n, λ_n) *is bounded in* $\mathscr{L}^{2m}(\Omega)$ *and the sequence* $(\bar{x}_n, \bar{\lambda}_n)$ *converges almost surely (a.s.) to a random variable* (X, Λ) *supported by* $\mathcal{S}$.

Let us now discuss our assumptions. Assumption 1 is standard in the decreasing step case. Assumption 2 is a moment assumption that is generally easy to check. Note that this assumption requires the set of saddle points $\mathcal{S}$ to be non empty. Notice the relation between Equations (6) and the two inclusions in (5). Focusing on the first inclusion, there exist $a \in \partial F(x_\star) = \mathbb{E}_\mu \partial f(\cdot, x_\star)$ and $b \in \partial G(x_\star) = \mathbb{E}_\mu \partial g(\cdot, x_\star)$ such that $0 = a + b + \mathsf{L}^T \lambda_\star$. Then, Assumption 2 states that there are two measurable selections φ_f and φ_g of $\partial f(\cdot, x_\star)$ and $\partial g(\cdot, x_\star)$ respectively which are both in $\mathscr{L}^{2m}(\mu)$ and which satisfy $a = \mathbb{E}_\mu \varphi_f$ and $b = \mathbb{E}_\mu \varphi_g$. Not also that the larger is m, and the weaker is Assumption 5.

Assumption 3 is relatively weak and easy to check. This assumption on the functions g and p is much weaker than Assumption 6, which assumes that the growth of $\widetilde{\nabla} f(s, \cdot)$ is not faster than linear. This is due to the fact that g and p enter the algorithm (1) through the proximity operator while the function f is used explicitly in this algorithm (through its (sub)gradient). This use of the functions f is reminiscent of the well-known Robbins-Monro algorithm, where a linear growth is needed to ensure the algorithm stability. Note that Assumption 6 is satisfied under the more restrictive assumption that $\nabla f(s, \cdot)$ is L-Lipschitz continuous without any bounded gradient assumption.

Assumption 4 is quite weak, and is studied *e.g* in [18]. This assumption is easy to illustrate in the case where $g(s, x) = \iota_{\mathcal{C}_s}(x)$ as in (4). Following [17], we say that the subsets $(\mathcal{C}_1, \ldots, \mathcal{C}_m)$ are linearly regular if there exists $C > 0$ such that for every x,

$$\max_{i=1\ldots m} \operatorname{dist}(x, \mathcal{C}_i) \geq C \operatorname{dist}(x, \mathcal{C}).$$

Sufficient conditions for a collection of sets to satisfy the above condition can be found in [17] and the references therein. Note that this condition implies that $N_{\mathcal{C}}(x) = \sum_{i=1}^m N_{\mathcal{C}_i}(x)$. Let us finally discuss Assumption 5. As $\gamma \to 0$, it is known that $\operatorname{prox}_{\gamma g(s,\cdot)}(x)$ converges to $\mathit{\Pi}_g(s, x)$ for every (s, x). Assumption 5 provides a control on the convergence rate. This assumption holds under the sufficient condition that for μ-almost every s and for every $x \in \operatorname{dom} \partial g(s, \cdot)$,

$$\|\partial g_0(s, x)\| \leq \beta(s)(1 + \|x\|^{m/2}),$$

where β is a positive random variable with a finite fourth moment [12].

3 Proof of Theorem 1

The proof of Theorem 1 employs the monotone operator theory. We begin by recalling some basic facts on monotone operators. All the results below can be found in [19,4] without further mention.

A set-valued mapping $\mathsf{A} : \mathcal{X} \rightrightarrows \mathcal{X}$ on the Euclidean space $\mathcal{X}$ will be called herein an operator. An operator with singleton values is identified with a function. As above, the domain of A is $\mathrm{dom}(\mathsf{A}) = \{x \in \mathcal{X} \,:\, \mathsf{A}(x) \neq \emptyset\}$. The graph of A is $\mathrm{gr}(\mathsf{A}) = \{(x,y) \in \mathcal{X} \times \mathcal{X} \,:\, y \in \mathsf{A}(x)\}$. The operator A is said monotone if $\forall (x,y), (x',y') \in \mathrm{gr}(\mathsf{A})$, $\langle y - y', x - x' \rangle \geq 0$. A monotone operator with non empty domain is said maximal if $\mathrm{gr}(\mathsf{A})$ is a maximal element for the inclusion ordering in the family of the monotone operator graphs. Let I be the identity operator, and let A^{-1} be the inverse of A, which is defined by the fact that $(x,y) \in \mathrm{gr}(\mathsf{A}^{-1}) \Leftrightarrow (y,x) \in \mathrm{gr}(\mathsf{A})$. An operator A belongs to the set $\mathscr{M}(\mathcal{X})$ of the maximal monotone operators on $\mathcal{X}$ if and only if for each $\gamma > 0$, the so-called resolvent $(I + \gamma \mathsf{A})^{-1}$ is a contraction defined on the whole space $\mathcal{X}$. In particular, it is single-valued. A typical element of $\mathscr{M}(\mathcal{X})$ is the subdifferential $\partial \mathsf{G}$ of a function $\mathsf{G} \in \Gamma_0(\mathcal{X})$. In this case, the resolvent $(I + \gamma \partial \mathsf{G})^{-1}$ for $\gamma > 0$ coincides with the proximity operator $\mathrm{prox}_{\gamma \mathsf{G}}$. A skew-symmetric element of $\mathcal{L}(\mathcal{X}, \mathcal{X})$ can also be checked to be an element of $\mathscr{M}(\mathcal{X})$.

The set of zeros of an operator A on $\mathcal{X}$ is the set $Z(\mathsf{A}) = \{x \in \mathcal{X} \,:\, 0 \in \mathsf{A}(x)\}$. The sum of two operators A and B is the operator $\mathsf{A} + \mathsf{B}$ whose image at x is the set sum of $\mathsf{A}(x)$ and $\mathsf{B}(x)$. Given two operators $\mathsf{A}, \mathsf{B} \in \mathscr{M}(\mathcal{X})$, where B is single-valued with domain $\mathcal{X}$, the so-called Forward-Backward algorithm is an iterative algorithm for finding a point in $Z(\mathsf{A} + \mathsf{B})$. It reads

$$x_{n+1} = (I + \gamma \mathsf{A})^{-1}(x_n - \gamma \mathsf{B}(x_n))$$

where γ is a positive step.

In the sequel, we shall be interested by random elements of $\mathscr{M}(\mathcal{X})$ as used in [12,5,14]. Consider a function $A : \Xi \to \mathscr{M}(\mathcal{X})$, where $(\Xi, \mathscr{G}, \mu)$ is the probability space introduced at the beginning of Section 2. By the maximality of $A(s)$, the graph $\mathrm{gr}(A(s))$ is known to be a closed subset of $\mathcal{X} \times \mathcal{X}$. By saying that $A(\cdot)$ is a $\mathscr{M}(\mathcal{X})$-valued random variable, we mean that the function $s \mapsto \mathrm{gr}(A(s))$ is measurable according to the definition of Section 2. When $A(s) = \partial h(s, \cdot)$, where $h : \Xi \times \mathcal{X} \to (-\infty, \infty]$ is a convex normal integrand such as $h(s, \cdot)$ is proper μ-a.e., A is a random element of $\mathscr{M}(\mathcal{X})$. Finally, when $A(s)$ is a skew-symmetric element of $\mathcal{L}(\mathcal{X}, \mathcal{X})$ which is measurable in the usual sense (as a $\Xi \to \mathcal{L}(\mathcal{X}, \mathcal{X})$ function), then it is also a random element of $\mathscr{M}(\mathcal{X})$.

We now enter the proof of Theorem 1. Let us set $\mathcal{Y} = \mathcal{X} \times \mathcal{V}$, and endow this Euclidean space with the standard scalar product. By writing $(x, \lambda) \in \mathcal{Y}$, it will be understood that $x \in \mathcal{X}$ and $\lambda \in \mathcal{V}$.

For each $s \in \Xi$, define the set-valued operator $A(s)$ on $\mathcal{Y}$ as

$$A(s, (x, \lambda)) = \begin{bmatrix} \partial g(s, x) \\ \partial p(s, \lambda) \end{bmatrix},$$

where $A(s,(x,\lambda))$ is the image of (x,λ) by $A(s)$. Fixing $s \in \Xi$, the operator $A(s,(x,\lambda))$ coincides with the subdifferential of the convex normal integrand $g(s,x)+p(s,\lambda)$ with respect to (x,λ). Thus, the map $s \mapsto A(s)$ is a measurable $\Xi \to \mathscr{M}(\mathcal{Y})$ function. Let us also define the operator $B(s)$ as

$$B(s,(x,\lambda)) = \begin{bmatrix} \partial f(s,x) & +L(s)^T\lambda \\ -L(s)x & \end{bmatrix}.$$

We can write $B(s) = B_1(s) + B_2(s)$, where

$$B_1(s,(x,\lambda)) = \begin{bmatrix} \partial f(s,x) \\ 0 \end{bmatrix}, \qquad B_2(s) = \begin{bmatrix} 0 & L(s)^T \\ -L(s) & 0 \end{bmatrix}$$

($B_2(s)$ is a linear skew-symmetric operator written in a matrix form in $\mathcal{Y}$). For each $s \in \Xi$, both these operators belong to $\mathscr{M}(\mathcal{Y})$, and $\operatorname{dom} B_2(s) = \mathcal{Y}$. Thus, $B(s) \in \mathscr{M}(\mathcal{Y})$ by [4, Cor. 24.4]. Moreover, since both B_1 and B_2 are measurable, B is a $\mathscr{M}(\mathcal{Y})$-valued random variable.

Now, from the assumptions on the functions f, g, and p, we see that the operators $\mathsf{A} = \mathbb{E}_\mu A$ and $\mathsf{B} = \mathbb{E}_\mu B$, where $\mathbb{E}_\mu$ is the selection integral (3), are written as

$$\mathsf{A}(x,\lambda) = \begin{bmatrix} \partial \mathsf{G}(x) \\ \partial \mathsf{H}^\star(\lambda) \end{bmatrix}, \text{ and } \mathsf{B}(x,\lambda) = \begin{bmatrix} \partial \mathsf{F}(x) & +\mathsf{L}^T\lambda \\ -\mathsf{L}x & \end{bmatrix}.$$

For the same reasons as for the operators $A(s)$ and $B(s)$, it holds that A, B, and $\mathsf{A}+\mathsf{B}$ belong to $\mathscr{M}(\mathcal{Y})$. Moreover, recalling the system of inclusions (5), we also obtain that $\mathcal{S} = Z(\mathsf{A}+\mathsf{B})$.

Defining the function

$$b(s,(x,\lambda)) = \begin{bmatrix} \widetilde{\nabla} f(s,x) & +L(s)^T\lambda \\ -L(s)x & \end{bmatrix}$$

(obviously, $b(s,(x,\lambda)) \in B(s,(x,\lambda))$ μ-a.e.), let us consider the following version of the Forward-Backward algorithm

$$(x_{n+1},\lambda_{n+1}) = (I + \gamma_{n+1}A(\xi_{n+1},\cdot))^{-1}\left((x_n,\lambda_n) - \gamma_{n+1}b(\xi_{n+1},(x_n,\lambda_n))\right).$$

On the one hand, one can easily check that this is exactly Algorithm (1). On the other hand, this algorithm is an instance of the random Forward-Backward algorithm studied in [5]. By checking the assumptions of Theorem 1 one by one, one sees that the assumptions of [5, Th. 3.1 and Cor. 3.1] are verified. Theorem 1 follows.

Remark 1 The convergence stated by Theorem 1 concerns the averaged sequence $(\bar{x}_n, \bar{\lambda}_n)$. One can ask whether the sequence (x_n,λ_n) itself converges to $\mathcal{S}$. A counterexample is provided by the particular case $\mathcal{X} = \mathcal{V} = \mathbb{R}$, $f = g = p = 0$, and $L = 1$ (proof omitted). A pointwise convergence would have been possible if $\mathsf{A}+\mathsf{B}$ were so-called *demipositive* [5]. Note that in the previous counterexample, $\mathsf{A}+\mathsf{B} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ is not demipositive.

Remark 2 Constant step Forward-Backward algorithms usually require the operator B to be so-called *cocoercive*. This property is not needed if a decreasing step size is used [20, 5].

References

1. A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, 2011.
2. B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.
3. L. Condat. A primal-dual splitting method for convex optimization involving Lipschitzian, proximable and linear composite terms. *Journal of Optimization Theory and Applications*, 158(2):460–479, 2013.
4. H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011.
5. P. Bianchi and W. Hachem. Dynamical behavior of a stochastic forward-backward algorithm using random monotone operators. *Journal of Optimization Theory and Applications*, 171(1):90–120, 2016.
6. H. Ouyang, N. He, L. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *International Conference on Machine Learning*, pages 80–88, 2013.
7. L. Rosasco, S. Villa, and B. C. Vũ. Stochastic inertial primal-dual algorithms. *arXiv preprint arXiv:1507.00852*, 2015.
8. H. Yu, M. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, pages 1427–1437, 2017.
9. P. L. Combettes and J. C. Pesquet. Stochastic forward-backward and primal-dual approximation algorithms with application to online image restoration. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 1813–1817. IEEE, 2016.
10. P. Toulis, T. Horel, and E. M. Airoldi. Stable robbins-monro approximations through stochastic proximal updates. *arXiv preprint arXiv:1510.00967*, 2015.
11. A. Patrascu and I. Necoara. Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization. *Journal of Machine Learning Research*, May 2017.
12. P. Bianchi. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.
13. Y. F. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(1):310–342, 2017.
14. P. Bianchi, W. Hachem, and A. Salim. A constant step Forward-Backward algorithm involving random maximal monotone operators. *To appear in Journal of Convex Analysis*, 2019.
15. H. Attouch. Familles d'opérateurs maximaux monotones et mesurabilité. *Annali di Matematica Pura ed Applicata*, 120(1):35–111, 1979.
16. R. T. Rockafellar and R. J.-B. Wets. On the interchange of subdifferentiation and conditional expectations for convex functionals. *Stochastics*, 7(3):173–182, 1982.
17. H. H. Bauschke, J. M. Borwein, and W. Li. Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization. *Mathematical Programming*, 86(1):135–160, 1999.
18. I. Necoara, P. Richtarik, and A. Patrascu. Randomized projection methods for convex feasibility problems: conditioning and convergence rates. *arXiv preprint arXiv:1801.04873*, 2018.
19. H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland mathematics studies. Elsevier Science, Burlington, MA, 1973.
20. J. Peypouquet and S. Sorin. Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time. *Journal of Convex Analysis*, 17(3-4):1113–1163, 2010.