A Fully Stochastic Primal-Dual Algorithm

Pascal Bianchi, Walid Hachem, Adil Salim

A Fully Stochastic Primal-Dual Algorithm

Pascal Bianchi[1], Walid Hachem[2], and Adil Salim[3]

[1]LTCI, Télécom Paris, IP Paris, 75013, Paris, France.
[2]LIGM, CNRS, Univ. Gustave Eiffel, F-77454 Marne-la-Vallée, France
[3]Visual Computing Center, KAUST, Saudi Arabia.

27 January 2020

Abstract

A new stochastic primal-dual algorithm for solving a composite optimization problem is proposed. It is assumed that all the functions / operators that enter the optimization problem are given as statistical expectations. These expectations are unknown but revealed across time through i.i.d realizations. 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. An example of convex optimization under stochastic linear constraints is considered.

1 Introduction

Many applications in machine learning, statistics or signal processing require the solution of the following optimization problem. 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.

Assuming the truth of the qualification condition $0 \in \mathrm{ri}(\mathrm{dom}\,\mathsf{H} - \mathsf{L}\,\mathrm{dom}\,\mathsf{G})$, where dom is the domain of a function and ri is the relative interior of a set, 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 $(x, \lambda) \mapsto \mathsf{F}(x) + \mathsf{G}(x) - \mathsf{H}^\star(\lambda) + \langle \mathsf{L}x, \lambda\rangle$, where $\mathsf{H}^\star$ is the Fenchel conjugate of H. There is a rich literature on such algorithms which cannot be exhaustively listed [10, 22, 14].

In this paper, it is assumed that the quantities that enter the minimization problem are unavailable or difficult to compute numerically, and have to be replaced with random quantities. Specifically, let $(\Xi, \mathscr{G}, \mu)$ be a probability space, and let $f : \Xi \times \mathcal{X} \to \mathbb{R}$ and $g : \Xi \times \mathcal{V} \to (-\infty, +\infty]$ be two convex normal integrands (see below). Assume 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), and assume that $\mathsf{L} = \mathbb{E}_\mu L(\cdot)$. Finally, assume that $\mathsf{H}^\star$ takes the form $\mathsf{H}^\star(\lambda) = \mathbb{E}_\mu(p(\cdot, \lambda))$, where p is a normal convex integrand. In order to solve Problem (1), no one of the objects F, G, H and L is available. Instead, 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 (FB) algorithm involving random monotone operators that was introduced in [6]. 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 our knowledge, the proposed algorithm is the first method that allows to solve Problem (1) in a fully stochastic setting with weak assumptions on the noise. Existing methods typically allow to handle subproblems of Problem (1) in which some quantities used in this problem are assumed to be available or set to zero [16, 20, 21, 23]. In particular, the new algorithm generalizes the stochastic gradient algorithm, the stochastic proximal point algorithm [17, 21, 5], and the stochastic proximal gradient algorithm [1, 8]. A close paper to ours is [11], which deals with a FB algorithm with deterministic monotone operators and random additive errors. In this reference, the convergence of the iterates is established under stringent summability conditions on these errors. Random block coordinate iterations combined with the FB algorithm were also considered in [13, 7, 12].

The next section is devoted to rigorously stating the problem and the main result. An application example is also considered. Section 3 is devoted to the proof of our main theorem.

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 vector 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 $\Gamma_0(\mathcal{X})$. The set of real-valued k–summable sequences is ℓ^k.

2 Problem description and main result

We start by recalling some mathematical definitions. Let $(\Xi, \mathscr{G}, \mu)$ be a probability space where the σ-field $\mathscr{G}$ is μ-complete, and let $\mathcal{X}$ be an Euclidean space. A function $h : \Xi \times \mathcal{X} \to (-\infty, \infty]$ is said a convex normal integrand [19] if $h(s, \cdot)$ is convex, and if the set-valued mapping $s \mapsto \operatorname{epi} h(s, \cdot)$ is closed-valued and measurable in the sense of [19, Chap. 14], where epi is the epigraph of a function. We shall always assume that $h(s, \cdot) \in \Gamma_0(\mathcal{X})$ for μ–almost all $s \in \Xi$. Given $x \in \mathcal{X}$, denote as $\partial h(s, x)$ the subdifferential of $h(s, \cdot)$ at x. For $r \in [1, \infty)$, let $\mathscr{L}^r(\mu)$ be the space of the $\mathscr{G}$-measurable functions $\varphi : \Xi \to \mathcal{X}$ such that $\int \|\varphi\|^r d\mu < \infty$. If $\mu(\{s \in \Xi \,:\, \partial h(s, x) \neq \emptyset\}) < 1$, set $\mathfrak{S}^r_{\partial h(\cdot,x)} := \emptyset$, otherwise,

$$\mathfrak{S}^r_{\partial h(\cdot,x)} := \{\varphi \in \mathscr{L}^r(\mu) \,:\, \varphi(s) \in \partial h(s, x) \ \mu - \text{almost everywhere (a.e.)}\}$$

is the set of the so-called r–integrable selections of the measurable set-valued function $s \mapsto \partial h(s, x)$. Denoting as cl the closure of a set, the so-called selection integral of $\partial h(\cdot, x)$ is the set

$$\mathbb{E}_\mu \partial h(\cdot, x) := \operatorname{cl} \left\{ \int_\Xi \varphi d\mu \ : \ \varphi \in \mathfrak{S}^1_{\partial h(\cdot,x)} \right\} \tag{2}$$

that might be empty. Note that we use the same notation $\mathbb{E}_\mu$ for these set-valued expectations and for the classical single-valued expectations.

We now state our problem. Let $f : \Xi \times \mathcal{X} \to (-\infty, \infty]$ be a convex normal integrand, assume that $\mathbb{E}_\mu |f(\cdot, x)| < \infty$ for all $x \in \mathcal{X}$, and consider the convex function $\mathsf{F}(x) := \mathbb{E}_\mu f(\cdot, x)$ which domain is $\mathcal{X}$. Let $g : \Xi \times \mathcal{X} \to (-\infty, \infty]$ be a convex normal integrand, and let $\mathsf{G}(x) := \mathbb{E}_\mu g(\cdot, x)$, where the integral $\mathbb{E}_\mu$ 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

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

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 assume that G is proper. In a similar manner, let $p : \Xi \times \mathcal{V} \to (-\infty, \infty]$ be a convex normal integrand, assume that $\mathsf{P} : \lambda \mapsto \mathbb{E}_\mu p(\cdot, \lambda)$ belongs to $\Gamma_0(\mathcal{V})$, and let H be its Fenchel conjugate (thus, $\mathsf{H}^\star = \mathsf{P}$). Finally, let $L : \Xi \to \mathcal{L}(\mathcal{X}, \mathcal{V})$ be an operator-valued measurable function, assume that $\|L\|$ is μ-integrable, and let $\mathsf{L} := \mathbb{E}_\mu L$.

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. Specifically, $\widetilde{\nabla} f : (\Xi\times\mathcal{X}, \mathscr{G}\otimes\mathscr{B}(\mathcal{X})) \to (\mathcal{X}, \mathscr{B}(\mathcal{X}))$ is a measurable function such that for each $x \in \mathcal{X}$, $\widetilde{\nabla} f(\cdot,x) \in \mathfrak{S}^1_{\partial f(\cdot,x)}$, which is known to be non empty thanks to the integrability assumption $\mathbb{E}_\mu|f(\cdot,x)| < \infty$ [18]. A possible choice for $\widetilde{\nabla} f(s,x)$ is $\partial_0 f(s,x)$ [6, §2.3 and §3.1]. 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

$$\left\{\begin{array}{rcccl} 0 & \in & \partial\mathsf{F}(x) + \partial\mathsf{G}(x) & & +\mathsf{L}^T\lambda, \\ 0 & \in & & -\mathsf{L}x & +\partial\mathsf{H}^\star(\lambda)\,. \end{array}\right. \tag{3}$$

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 $(x_n, \lambda_n) \in \mathcal{X}\times\mathcal{V}$.

$$\begin{aligned} x_{n+1} &= \mathrm{prox}_{\gamma_{n+1} g(\xi_{n+1},\cdot)}\left(x_n - \gamma_{n+1}(\widetilde{\nabla} f(\xi_{n+1}, x_n) + L(\xi_{n+1})^T\lambda_n)\right), \\ \lambda_{n+1} &= \mathrm{prox}_{\gamma_{n+1} p(\xi_{n+1},\cdot)}\left(\lambda_n + \gamma_{n+1} L(\xi_{n+1}) x_n\right)\,. \end{aligned} \tag{4}$$

The convergence of Algorithm (4) is stated by the next theorem in terms of weighted averaged estimates

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

Theorem 2.1 *Consider Problem* (1)*, and let the following assumptions hold.*

1. *The step size sequence satisfies* $(\gamma_n) \in \ell^2 \setminus \ell^1$*, and* $\gamma_{n+1}/\gamma_n \to 1$ *as* $n \to \infty$*.*
2. *The function* G *satisfies* $\partial\mathsf{G}(x) = \mathbb{E}_\mu \partial g(\cdot, x)$ *for each* $x \in \mathcal{X}$*.*
3. *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* $\varphi_p \in \mathfrak{S}^{2m}_{\partial p(\cdot,\lambda_\star)}$ *such that*

$$\mathbb{E}_\mu\varphi_f + \mathbb{E}_\mu\varphi_g + \mathsf{L}^T\lambda_\star = 0, \text{ and } -\mathsf{L}x_\star + \mathbb{E}_\mu\varphi_p = 0. \tag{5}$$

Moreover, for every point $(x_\star, \lambda_\star) \in \mathcal{S}$*, there exist three functions* $\varphi_f \in \mathfrak{S}^2_{\partial f(\cdot,x_\star)}$*,* $\varphi_g \in \mathfrak{S}^2_{\partial g(\cdot,x_\star)}$*, and* $\varphi_p \in \mathfrak{S}^2_{\partial p(\cdot,\lambda_\star)}$ *such that* (5) *holds.*

4. *For any compact set* $K \subset \operatorname{dom} \partial \mathsf{G}$*, there exist* $\varepsilon \in (0,1]$ *and* $x_0 \in \operatorname{dom} \partial \mathsf{G}$ *such that*

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

5. *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 3, 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* $\mathbb{E}_\mu \|\widetilde{\nabla} f(\cdot, x)\|^4 \leq C(1 + \|x\|^{2m})$*.*

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

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

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

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

where $\Pi_g(s, \cdot)$ *is the projection operator onto* $\operatorname{cl}(\operatorname{dom} \partial g(s, \cdot))$*, and where* m *is the integer provided by Assumption 3.*

8. *Assumptions 2, 4, 6 and 7 hold true when the function* g *is replaced with* p *and the space* $\mathcal{X}$ *is replaced with* $\mathcal{V}$*.*

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 requires that the interchange of the expectation $\mathbb{E}_\mu g(\cdot, x)$ and the subdifferentiation be possible. Let us provide some sufficient conditions for this to be true. By [18], 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 of practical importance 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 $\mathcal{C} = \cap_{i=k}^{m} \mathcal{C}_k$ be non empty, 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 [3] 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

$$g(s, x) = \iota_{\mathcal{C}_s}(x) \text{ for } (s, x) \in \Xi \times \mathcal{X}. \tag{6}$$

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

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

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

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

and

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

Assumption 3 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 (5) and the two inclusions in (3). Focusing on the first inclusion and using Assumption 2, there exist $a \in \partial \mathsf{F}(x_\star) = \mathbb{E}_{\mu} \partial f(\cdot, x_\star)$ and $b \in \partial \mathsf{G}(x_\star) = \mathbb{E}_{\mu} \partial g(\cdot, x_\star)$ such that $0 = a + b + \mathsf{L}^T \lambda_\star$. Then, Assumption 3 states that a and b can be taken in such a way 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 $b = \mathbb{E}_{\mu} \varphi_g$. A sufficient condition for the existence of the selections satisfying Assumption 3 is the following [8]: there exists an open neighborhood $\mathcal{N}_x$ of $x_\star$ and an open neighborhood $\mathcal{N}_\lambda$ of $\lambda_\star$ such that $\forall x \in \mathcal{N}_x$, $\int f(s, x)^{2m} \mu(ds) < \infty$ and $\int g(s, x)^{2m} \mu(ds) < \infty$, and $\forall \lambda \in \mathcal{N}_\lambda$, $\int p(s, x)^{2m} \mu(ds) < \infty$. Note also that the larger is m, and the weaker is Assumption 7.

Assumption 4 is relatively weak and easy to check. It is interesting to compare it with Assuption 5. It is indeed much weaker than the latter,

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 (4) 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 5 is satisfied under the more restrictive assumption that $\nabla f(s,\cdot)$ is L-Lipschitz continuous without any bounded gradient assumption.

Assumption 6 is quite weak, and is studied *e.g* in [15]. This assumption is easy to illustrate in the case where $g(s,x) = \iota_{\mathcal{C}_s}(x)$ as in (6). Following [3], we say that the subsets $(\mathcal{C}_1,\dots,\mathcal{C}_m)$ are linearly regular if there exists $C>0$ such that for every x,

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

Sufficient conditions for a collection of sets to satisfy the above condition can be found in [3] 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 7. As $\gamma \to 0$, it is known that $\operatorname{prox}_{\gamma g(s,\cdot)}(x)$ converges to $\Pi_g(s,x)$ for every (s,x). Assumption 7 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 [5].

We now consider an application example of Theorem 2.1.

Example 1 *Let $\mathsf{c} \in \mathcal{V}$. Setting $\mathsf{H} = \iota_{\{\mathsf{c}\}}$, where $\iota_{\mathcal{C}}$ is the indicator function of the set $\mathcal{C}$, Problem* (1) *boils down to the linearly constrained problem*

$$\min_{x\in\mathcal{X}} \mathsf{F}(x) + \mathsf{G}(x) \quad s.t. \quad \mathsf{L}x = \mathsf{c}. \tag{7}$$

If we assume that $\mathsf{c} = \mathbb{E}_\mu(c(\cdot))$ where $c(\cdot): \Xi \to \mathcal{V}$ is a random vector, then our problem amounts to randomizing the constraints and to handling these stochastic constraints online. Such a context is encountered in various fields of machine learning, as the Neyman-Pearson classification, or in online so-called Markowicz portfolio optimization.

Since $\mathsf{H}^\star(\lambda) = \langle \lambda, \mathsf{c}\rangle$, we simply need to put $p(\cdot,\lambda) = \langle \lambda, c(\cdot)\rangle$, and Algorithm (4) *becomes:*

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

To go further, let us particularize Problem (7) *to the case of the Markowicz portfolio optimization, and check the assumptions of Theorem 2.1 to complete*

the picture. In this case, ξ is a $\mathcal{X}$–valued random variable with a second moment, $\mathsf{F}(x) = \mathbb{E}_\mu\langle x, \xi\rangle^2$, $\mathsf{G}(x) = \iota_\Delta(x)$ where Δ is the probability simplex, $\mathsf{L} = \mathbb{E}_\mu(\xi^T)$, and c is some real positive number. Note that it is usually assumed that $\mathsf{L} = \mathbb{E}_\mu(\xi^T)$ is fully known or estimated, which we don't do here. We of course assume that the qualification condition $\mathsf{c} \in \mathrm{ri}\,\mathsf{L}\Delta$ holds true.

Assumptions 2 and 4 of the statement of Theorem 2.1 are immediate for both g and p. One can check that Assumption 3 is satisfied for $m = 2$ if we assume that $\mathbb{E}_\mu\|\xi\|^4 < \infty$, which also ensures the truth of Assumption 5. Assumptions 6 and 7 are trivially satisfied for g and p, since $\mathrm{prox}_{\gamma g(s,\cdot)} = \Pi_g(s,\cdot)$, and since $p(s,\cdot)$ has a full domain.

3 Proof of Theorem 2.1

The proof of Theorem 2.1 makes use of the monotone operator theory. We begin by recalling some basic facts on monotone operators. All the results below can be found in [9, 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 ∂h of a function $h \in \Gamma_0(\mathcal{X})$. In this case, the resolvent $(I + \gamma\partial h)^{-1}$ for $\gamma > 0$ coincides with the proximity operator $\mathrm{prox}_{\gamma h}$. 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 FB 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 [5, 6, 8]. A random element of $\mathscr{M}(\mathcal{X})$ is a measurable function $M : \Xi \to \mathscr{M}(\mathcal{X})$ in the sense of [2], where $(\Xi, \mathscr{G}, \mu)$ is the probability

space introduced at the beginning of Section 2. In particular, when $h : \Xi \times \mathcal{X} \to (-\infty, \infty]$ is a convex normal integrand such as $h(s, \cdot)$ is proper μ-a.e., $M(s) = \partial h(s, \cdot)$ is a random element of $\mathscr{M}(\mathcal{X})$. Moreover, when $M(s)$ 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 $\mathscr{M}(\mathcal{X})$. If we fix $x \in \mathcal{X}$ and we denote as $M(s, x)$ its image by $M(s)$, then the set-valued function $s \mapsto M(s, x)$ is measurable, and its (set-valued) expectation $\mathsf{M}(x) = \mathbb{E}_\mu M(\cdot, x)$ is defined similarly to Equation (2) [2, 5, 6]. Note that M is monotone but not necessarily maximal.

We now enter the proof of Theorem 2.1. 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 the operator that takes (x, λ) to

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

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, $A(s)$ is a random element of $\mathscr{M}(\mathcal{Y})$. 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

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

($B_2(s)$ 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 random element of $\mathscr{M}(\mathcal{Y})$.

Since $f(\cdot, x)$ is Lebesgue-integrable for all $x \in \mathcal{X}$ by construction, it is known that $\partial \mathsf{F}(x) = \mathbb{E}_\mu \partial f(\cdot, x)$ [18]. Moreover, $\partial \mathsf{G}(x) = \mathbb{E}_\mu \partial g(\cdot, x)$ and $\partial \mathsf{H}^\star(\lambda) = \mathbb{E}_\mu \partial p(\cdot, \lambda)$ by Assumptions 2 and 8. Thus, the operators $\mathsf{A}((x, \lambda)) = \mathbb{E}_\mu A(\cdot, (x, \lambda))$ and $\mathsf{B}((x, \lambda)) = \mathbb{E}_\mu B(\cdot, (x, \lambda))$ can be written as

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

thus, these monotone operators are both maximal. By [4, Cor. 24.4], we also get that $\mathsf{A} + \mathsf{B}$ belong to $\mathscr{M}(\mathcal{Y})$. Moreover, recalling the system of inclusions (3), 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 FB 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 (4). On the other hand, this algorithm is an instance of the random FB algorithm studied in [6]. By checking the assumptions of Theorem 2.1 one by one, one sees that the assumptions of [6, Th. 3.1 and Cor. 3.1] are verified. Theorem 2.1 follows.

Remark 1 *The convergence stated by Theorem 2.1 concerns the averaged sequence* $(\bar{x}_n,\bar{\lambda}_n)$*. One can ask whether the sequence* (x_n,λ_n) *itself converges to* $\mathcal{S}$*. This would happen if the operator* $\mathsf{A}+\mathsf{B}$ *were so-called demipositive [6]. This happens when,* e.g.*,* $\mathsf{F}+\mathsf{G}$ *is strongly convex and* H *is smooth (proof omitted). Unfortunately, demipositivity of* $\mathsf{A}+\mathsf{B}$ *is not always guaranteed.*

References

[1] Y. F. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Journal of Machine Learning Research*, 18(1):310–342, 2017.

[2] H. Attouch. Familles d'opérateurs maximaux monotones et mesurabilité. *Annali di Matematica Pura ed Applicata*, 120(1):35–111, 1979.

[3] 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.

[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. Ergodic convergence of a stochastic proximal point algorithm. *SIAM Journal on Optimization*, 26(4):2235–2260, 2016.

[6] 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.

[7] P. Bianchi, W. Hachem, and F. Iutzeler. A coordinate descent primal-dual algorithm and application to distributed asynchronous optimization. *IEEE Transactions on Automatic Control*, 61(10):2947–2957, Oct 2016.

[8] P. Bianchi, W. Hachem, and A. Salim. A constant step Forward-Backward algorithm involving random maximal monotone operators. *Journal of Convex Analysis*, 26(2):397–436, 2019.

[9] H. Brézis. *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*. North-Holland mathematics studies. Elsevier Science, Burlington, MA, 1973.

[10] 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.

[11] P. L. Combettes and J.-C. Pesquet. Stochastic approximations and perturbations in forward-backward splitting for monotone operators. *Pure and Applied Functional Analysis*, 1(1):13–37, January 2016.

[12] P. L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping ii: mean-square and linear convergence. *Mathematical Programming*, 174(1):433–451, Mar 2019.

[13] P.L. Combettes and J.-C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248, 2015.

[14] 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.

[15] I. Necoara, P. Richtarik, and A. Patrascu. Randomized projection methods for convex feasibility problems: conditioning and convergence rates. *arXiv preprint arXiv:1801.04873*, 2018.

[16] 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.

[17] A. Patrascu and I. Necoara. Nonasymptotic convergence of stochastic proximal point algorithms for constrained convex optimization. *Journal of Machine Learning Research*, May 2017.

[18] 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.

[19] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998.

[20] L. Rosasco, S. Villa, and B. C. Vũ. Stochastic inertial primal-dual algorithms. *arXiv preprint arXiv:1507.00852*, 2015.

[21] P. Toulis, T. Horel, and E. M. Airoldi. Stable robbins-monro approximations through stochastic proximal updates. *arXiv preprint arXiv:1510.00967*, 2015.

[22] B. C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. *Advances in Computational Mathematics*, 38(3):667–681, 2013.

[23] H. Yu, M. Neely, and X. Wei. Online convex optimization with stochastic constraints. In *Advances in Neural Information Processing Systems*, pages 1427–1437, 2017.