STOCHASTIC FORWARD-BACKWARD AND PRIMAL-DUAL APPROXIMATION ALGORITHMS WITH APPLICATION TO ONLINE IMAGE RESTORATION

Patrick L. Combettes$^1$ and Jean-Christophe Pesquet$^2$

$^1$Sorbonne Universités – UPMC Univ. Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, Paris, France
plc@ljll.math.upmc.fr

$^2$Université Paris-Est, Laboratoire d’Informatique Gaspard Monge – CNRS UMR 8049, Champs-sur-Marne, France
jean-christophe.pesquet@univ-paris-est.fr

ABSTRACT

Stochastic approximation techniques have been used in various contexts in data science. We propose a stochastic version of the forward-backward algorithm for minimizing the sum of two convex functions, one of which is not necessarily smooth. Our framework can handle stochastic approximations of the gradient of the smooth function and allows for stochastic errors in the evaluation of the proximity operator of the nonsmooth function. The almost sure convergence of the iterates generated by the algorithm to a minimizer is established under relatively mild assumptions. We also propose a stochastic version of a popular primal-dual proximal splitting algorithm, establish its convergence, and apply it to an online image restoration problem.

Index Terms— convex optimization, nonsmooth optimization, primal-dual algorithm, stochastic algorithm, parallel algorithm, proximity operator, recovery, image restoration.

1. INTRODUCTION

A large array of optimization problems arising in signal processing involve functions belonging to $\Gamma_0(H)$, the class of proper lower semicontinuous convex function from $H$ to $[-\infty, +\infty]$, where $H$ is a finite-dimensional real Hilbert space with norm $\|\cdot\|$. In particular, the following formulation has proven quite flexible and far reaching [18].

Problem 1.1 Let $f \in \Gamma_0(H)$, let $\vartheta \in [0, +\infty[$, and let $h : H \rightarrow \mathbb{R}$ be a differentiable convex function such that $\nabla h$ is $\vartheta^{-1}$-Lipschitz continuous on $H$. The goal is to minimize

$$
\min_{x \in H} f(x) + h(x),
$$

under the assumption that the set $F$ of minimizers of $f + h$ is nonempty.

A standard method to solve Problem 1.1 is the forward-backward algorithm [6, 9, 10, 16, 18], which constructs a sequence $(x_n)_{n \in \mathbb{N}}$ in $H$ via the recursion

$$
(\forall n \in \mathbb{N}) \quad x_{n+1} = \prox_{\gamma_n f}(x_n - \gamma_n \nabla h(x_n)),
$$

where $\gamma_n \in [0, 2\vartheta]$ and $\prox_{\gamma_n f}$ is the proximity operator of function $\gamma_n f$, i.e., [3]

$$
\prox_{\gamma_n f} : x \mapsto \arg\min_{y \in H} \left( f(y) + \frac{1}{2\gamma_n} \|x - y\|^2 \right).
$$

In practice, it may happen that, at each iteration $n$, $\nabla h(x_n)$ is not known exactly and is available only through some stochastic approximation $v_n$, where only a deterministic approximation $f_n$ to $f$ is known; see, e.g., [29]. To solve (1) in such uncertain environments, we propose to investigate the following stochastic version of (2). In this algorithm, at iteration $n$, $a_n$ stands for a stochastic error term modeling inexact implementations of the proximity operator of $f_n$, $(\Omega, F, P)$ is the underlying probability space, and $L^2(\Omega, F, P; H)$ denotes the space of $H$-valued random variable $x$ such that $E\|x\|^2 < +\infty$. Our algorithmic model is the following.

Algorithm 1.2 Let $x_0$, $(u_n)_{n \in \mathbb{N}}$, and $(a_n)_{n \in \mathbb{N}}$ be random variables in $L^2(\Omega, F, P; H)$, let $(\lambda_n)_{n \in \mathbb{N}}$ be a sequence in $[0, 1]$, and let $(\gamma_n)_{n \in \mathbb{N}}$ be a sequence in $[0, 2\vartheta]$, and let $(f_n)_{n \in \mathbb{N}}$ be a sequence of functions in $\Gamma_0(H)$. For every $n \in \mathbb{N}$, set

$$
x_{n+1} = x_n + \lambda_n \left( \prox_{\gamma_n f_n}(x_n - \gamma_n u_n) + a_n - x_n \right).
$$

The first instances of the stochastic iteration (4) can be traced back to [31] in the context of the gradient descent method, i.e., when $f_n \equiv f = 0$. Stochastic approximations in the gradient method were then investigated in the Russian literature of the late 1960s and early 1970s [21, 23, 36]. Stochastic gradient methods have also been used extensively in adaptive signal processing, in control, and in machine learning.
that the following are satisfied:

\[ (\tau_n)_{n \in \mathbb{N}} \]

The first objective of the present work is to provide a thorough convergence analysis of the stochastic forward-backward algorithm described in Algorithm 1.2. In particular, our results do not require that the proximal parameter sequence \((\gamma_n)_{n \in \mathbb{N}}\) be vanishing. A second goal of our paper is to show that the extension of Algorithm 1.2 for solving monotone inclusion problems allows us to derive a stochastic version of a recent primal-dual algorithm [39] (see also [17, 19]). Note that our algorithm is different from the random block-coordinate approaches developed in [4, 30], and that it is more in the spirit of the adaptive method of [28].

The organization of the paper is as follows. Section 2 contains our main result on the convergence of the iterates of a sequence \((X_n)_{n \in \mathbb{N}}\) of random variables such that \(\mathcal{X} = (\mathcal{X}_n)_{n \in \mathbb{N}}\) is a sequence of random variables such that

\[
(\forall n \in \mathbb{N}) \quad X_n \in \mathcal{F} \quad \text{and} \quad (\sigma(X_0, \ldots, X_n)) \subseteq \mathcal{X}_n \subseteq \mathcal{X}_{n+1}.
\] (5)

Furthermore, \(\ell_+^2(\mathcal{X})\) designates the set of sequences of \([0, +\infty]-valued\) random variables \((\xi_n)_{n \in \mathbb{N}}\) such that, for every \(n \in \mathbb{N}\), \(\xi_n\) is \(\mathcal{X}_n\)-measurable, and we define

\[
\ell_+^{1/2}(\mathcal{X}) = \left\{ (\xi_n)_{n \in \mathbb{N}} \in \ell_+^1(\mathcal{X}) \mid \sum_{n \in \mathbb{N}} \langle \xi_n \rangle < +\infty \right\}.
\] (6)

and

\[
\ell_+^\infty(\mathcal{X}) = \left\{ (\xi_n)_{n \in \mathbb{N}} \in \ell_+^1(\mathcal{X}) \mid \sup_{n \in \mathbb{N}} \xi_n < +\infty \right\}.
\] (7)

We now state our main convergence result.

**Theorem 2.1** Consider the setting of Problem 1.1, let \((\tau_n)_{n \in \mathbb{N}}\) be a sequence in \([0, +\infty]\), let \((x_n)_{n \in \mathbb{N}}\) be a sequence generated by Algorithm 1.2, and let \(\mathcal{X} = (\mathcal{X}_n)_{n \in \mathbb{N}}\) be a sequence of sub-sigma-algebras satisfying (5). Suppose that the following are satisfied:

(a) \(\sum_{n \in \mathbb{N}} \lambda_n \sqrt{\mathbb{E}(|u_n|^2 | \mathcal{X}_n)} < +\infty\).

(b) \(\sum_{n \in \mathbb{N}} \sqrt{\lambda_n} \mathbb{E}(u_n | \mathcal{X}_n) - \nabla h(x_n) < +\infty\).

(c) For every \(z \in F\), there exists \((\zeta_n(z))_{n \in \mathbb{N}} \in \ell_+^{1/2}(\mathcal{F})\) such that \(\lambda_n \zeta_n(z) \in \ell_+^1(\mathcal{X})\) and

\[
(\forall n \in \mathbb{N}) \quad \mathbb{E}(\|u_n - \mathbb{E}(u_n | \mathcal{X}_n)\|^2 | \mathcal{X}_n) \\
\leq \tau_n \|\nabla h(x_n) - \nabla h(z)\|^2 + \zeta_n(z).
\] (8)

(d) There exist sequences \((\alpha_n)_{n \in \mathbb{N}}\) and \((\beta_n)_{n \in \mathbb{N}}\) in \([0, +\infty]\) such that \(\sum_{n \in \mathbb{N}} \sqrt{\alpha_n} < +\infty\), \(\sum_{n \in \mathbb{N}} \beta_n < +\infty\), and

\[
(\forall n \in \mathbb{N}) \quad \mathbb{E}(|\text{prox}_{\alpha_n} x - \text{prox}_{\beta_n} x|) \leq \alpha_n ||x|| + \beta_n.
\] (9)

(e) \(\inf_{n \in \mathbb{N}} \gamma_n > 0\), \(\sup_{n \in \mathbb{N}} \tau_n < +\infty\), and

\[
\sup_{n \in \mathbb{N}} (1 + \tau_n) \gamma_n < 2\theta.
\]

(f) Either \(\inf_{n \in \mathbb{N}} \lambda_n > 0\) or \(\gamma_n \equiv \gamma; \sum_{n \in \mathbb{N}} \tau_n < +\infty\), and

\[
\sum_{n \in \mathbb{N}} \lambda_n = +\infty
\].

Then the following hold for every \(z \in F\) and for some \(F\)-valued random variable \(x\):

(i) \(\sum_{n \in \mathbb{N}} \lambda_n \|\nabla h(x_n) - \nabla h(z)\|^2 < +\infty\) \(\mathbb{P}\-a.s.\)

(ii) \(\sum_{n \in \mathbb{N}} \lambda_n \|x_n - \gamma_n \nabla h(x_n) - \text{prox}_{\gamma_n}(x_n - \gamma_n \nabla h(x_n))\|^2 < +\infty\) \(\mathbb{P}\-a.s.\)

(iii) \((x_n)_{n \in \mathbb{N}}\) converges almost surely to \(x\).

In the deterministic case, Theorem 2.1(iii) can be found in [7, Corollary 6.5]. The proof the above stochastic version is based on the theoretical tools of [12] (see [13] for technical details and extensions to infinite-dimensional Hilbert spaces).

It should be noted that the existing works which are the most closely related to ours do not allow any approximation of the function \(f\) and make some additional restrictive assumptions. For example, in [1, Corollary 8] and [33], \((\gamma_n)_{n \in \mathbb{N}}\) is a decreasing sequence. In [1, Corollary 8], [33], and [34], no error term is allowed in the numerical evaluations of the proximity operators \((a_n \equiv 0)\). In addition, in the former work, it is assumed that \((x_n)_{n \in \mathbb{N}}\) is bounded, whereas the two latter ones assume that the approximation of the gradient of \(h\) is unbiased, that is

\[
(\forall n \in \mathbb{N}) \quad \mathbb{E}(u_n | \mathcal{X}_n) = \nabla h(x_n).
\] (10)

### 3. STOCHASTIC PRIMAL-DUAL SPLITTING

The subdifferential

\[
\partial f : x \mapsto \{ u \in H \mid \langle u, y - x \rangle \leq f(y) \}
\]

of a function \(f \in \Gamma_0(H)\) is an example of a maximally monotone operator [3]. Forward-backward splitting has been developed in the more general framework of solving monotone
inclinations [3, 7]. This powerful framework makes it possible to design efficient primal-dual strategies for optimization problems; see for instance [17, 25] and the references therein. More precisely, we are interested in the following optimization problem [11, Section 4].

**Problem 3.1** Let $f \in \Gamma_0(H)$, let $\vartheta \in [0, +\infty]$, let $h : H \to \mathbb{R}$ be convex and differentiable with a $\vartheta^{-1}$-Lipschitz-continuous gradient, and let $q$ be a strictly positive integer. For every $k \in \{1, \ldots, q\}$, let $G_k$ be a finite-dimensional Hilbert space, let $g_k \in \Gamma_0(G_k)$, and let $L_k : H \to G_k$ be linear. Let $G = G_1 \oplus \cdots \oplus G_q$ be the direct Hilbert sum of $G_1, \ldots, G_q$, and suppose that there exists $\mathbf{x} \in H$ such that

$$0 \in \partial f(\mathbf{x}) + \sum_{k=1}^{q} L_k \partial g_k(L_k \mathbf{x}) + \nabla h(\mathbf{x}). \quad (12)$$

Let $F$ be the set of solutions to the problem

$$\min_{\mathbf{x} \in H} f(\mathbf{x}) + \sum_{k=1}^{q} g_k(L_k \mathbf{x}) + h(\mathbf{x}) \quad (13)$$

and let $F^*$ be the set of solutions to the dual problem

$$\min_{\mathbf{v} \in G} (f^* \square h^*)(\mathbf{v}) = -\sum_{k=1}^{q} L_k^\star \mathbf{v}_k + \sum_{k=1}^{q} g_k^\star(\mathbf{v}_k), \quad (14)$$

where $\square$ denotes the infimal convolution operation, $\varphi^*$ is the Legendre conjugate of a function $\varphi$, and $\mathbf{v} = (v_1, \ldots, v_q)$ designates a generic point in $G$. The objective is to find a point in $F \times F^*$.

We are interested in the case when only stochastic approximations of the gradients of $h$ and approximations of the function $f$ are available to solve Problem 3.1. The following algorithm, which can be viewed as a stochastic extension of those of [5,8,17,19,22,24,39], will be the focus of our investigation.

**Algorithm 3.2** Let $\rho \in [0, +\infty[$, let $(f_n)_{n \in \mathbb{N}}$ be a sequence of functions in $\Gamma_0(H)$, let $(\lambda_n)_{n \in \mathbb{N}}$ be a sequence in $[0, 1]$ such that $\sum_{n \in \mathbb{N}} \lambda_n = +\infty$, and, for every $k \in \{1, \ldots, q\}$, let $\sigma_k \in [0, +\infty[$. Let $x_0, (u_n)_{n \in \mathbb{N}}$, and $(b_n)_{n \in \mathbb{N}}$ random variables in $L^2(\Omega, \mathcal{F}, \mathbb{P}; H)$, and let $v_0$ and $(e_n)_{n \in \mathbb{N}}$ be random variables in $L^2(\Omega, \mathcal{F}, \mathbb{P}; G)$. Iterate

$$
\begin{align*}
    y_n & = \text{prox}_{\rho \varphi} \left( x_n - \rho \left( \sum_{k=1}^{q} L_k^\star v_{k,n} + u_n \right) \right) + b_n \\
    x_{n+1} & = x_n + \lambda_n (y_n - x_n) \\
    \text{for } k = 1, \ldots, q \\
    w_{k,n} & = \text{prox}_{\rho \varphi} \left( v_{k,n} + \sigma_k L_k (2y_n - x_n) \right) + e_{k,n} \\
    v_{k,n+1} & = v_{k,n} + \lambda_n (w_{k,n} - v_{k,n}).
\end{align*}
$$

(15)

One of main benefits of the proposed algorithm is that it allows us to solve jointly the primal problem (13) and the dual one (14) in a fully decomposed fashion, where each function and linear operator is activated individually. In particular, it does not require any inversion of some linear operator related to the operators $(L_k)_{1 \leq k \leq q}$ arising in the original problem. The convergence of the algorithm is guaranteed by the following result which follows from [13, Proposition 5.3].

**Proposition 3.3** Consider the setting of Problem 3.1, let $\mathcal{E} = (\mathcal{X}_n)_{n \in \mathbb{N}}$ be a sequence of sub-sigma-algebras of $\mathcal{F}$, and let $(x_n)_{n \in \mathbb{N}}$ and $(v_n)_{n \in \mathbb{N}}$ be sequences generated by Algorithm 3.2. Suppose that the following are satisfied:

(a) $(\forall n \in \mathbb{N})$ $\sigma(x_n, v_n)_{0 \leq n' \leq n} \subset \mathcal{X}_n \subset \mathcal{X}_{n+1}$.

(b) $\sum_{n \in \mathbb{N}} \lambda_n \sqrt{\mathbb{E}([b_n]^{2})} \lambda_n \sigma([c_n]^{2}) + \nabla h(x_n) < +\infty$ and $\sum_{n \in \mathbb{N}} \lambda_n \sqrt{\mathbb{E}([c_n]^{2})} + \nabla h(x_n) < +\infty$.

(c) $\sum_{n \in \mathbb{N}} \sqrt{\mathbb{E}([u_n]^{2})} \sqrt{\mathbb{E}([v_n]^{2})} \lambda_n \sigma([\lambda_n s_n]^{2}) < +\infty$.

(d) There exists a summable sequence $(\tau_n)_{n \in \mathbb{N}}$ in $[0, +\infty]$ such that, for every $x \in F$, there exists $(\zeta_n(x))_{n \in \mathbb{N}} \in L^{1_{\frac{3}{2}}}([X])$ such that $(\lambda_n \zeta_n(x))_{n \in \mathbb{N}} \in L^{1_{\frac{3}{2}}}([X])$ and

$$(\forall n \in \mathbb{N}) \quad \mathbb{E}((u_n - E(u_n | X_n))^2 | X_n) = \tau_n \| \nabla h(x_n) - \nabla h(x) \|^2 + \lambda_n \zeta_n(x). \quad (16)$$

(e) There exist sequences $(\alpha_n)_{n \in \mathbb{N}}$ and $(\beta_n)_{n \in \mathbb{N}}$ in $[0, +\infty]$ such that $\sum_{n \in \mathbb{N}} \lambda_n \alpha_n < +\infty$, $\sum_{n \in \mathbb{N}} \lambda_n \beta_n < +\infty$, and

$$(\forall n \in \mathbb{N}) (\forall x \in H) \quad \| \text{prox}_{\rho \varphi} x - \text{prox}_{\rho \varphi} x \| \leq \alpha_n \| x \| + \beta_n. \quad (17)$$

Then, for some $F$-valued random variable $x$ and some $F^*$-valued random variable $v$, $(x_n)_{n \in \mathbb{N}}$ converges almost surely to $x$ and $(v_n)_{n \in \mathbb{N}}$ converges almost surely to $v$.

4. APPLICATION TO ONLINE SIGNAL RECOVERY

We consider the recovery of a signal $\mathbf{x} \in H = \mathbb{R}^N$ from the observation model

$$((\forall n \in \mathbb{N}) \quad z_n = K_n \mathbf{x} + e_n, \quad (18)$$

where $K_n$ is a $\mathbb{R}^{M \times N}$-valued random matrix and $e_n$ is a $\mathbb{R}^M$-valued random noise vector. The objective is to recover $\mathbf{x}$ from $(K_n, z_n)_{n \in \mathbb{N}}$, which is assumed to be an identically distributed sequence. Such recovery problems have been addressed in [14]. In this context, we propose to solve the primal problem (13) with $q = 1$ and

$$(\forall x \in \mathbb{R}^N) \quad h(x) = \frac{1}{2} \mathbb{E}([K_0 x - z_0]^2), \quad (19)$$

where $K_0$ is a $\mathbb{R}^{M \times N}$-valued random matrix and $z_0$ is a $\mathbb{R}^M$-valued random noise vector.
while functions $f$ and $g_1 \circ L_1$ are used to promote prior formation on the target solution. Since the statistics of sequence $(K_n, z_n)_{n \in \mathbb{N}}$ are not assumed to be known a priori and have to be learnt online, at iteration $n \in \mathbb{N}$, we empirically estimate

$$u_n = \frac{1}{m_{n+1}} \sum_{n'=0}^{m_{n+1}-1} K^T_n (K_{n'} x_n - z_{n'})$$

of $\nabla h(x_n)$. The following statement, which can be deduced from [13, Section 5.2], illustrates the applicability of the result of Section 3.

**Proposition 4.1** Consider the setting of Problem 3.1 and Algorithm 3.2, where $f_n \equiv f$, $b_n \equiv 0$, and $e_n \equiv 0$. $(m_n)_{n \in \mathbb{N}}$ be a strictly increasing sequence in $\mathbb{N}$ such that $m_n = O(n^{1+\delta})$ with $\delta \in [0, +\infty]$, and let

$$(\forall n \in \mathbb{N}) \quad X_n = \sigma (x_0, w_0, (K_{n'}, e_{n'})_{0 \leq n' < m_n}).$$

Suppose that the following are satisfied:

(a) The domain of $f$ is bounded.

(b) $(K_n, e_n)_{n \in \mathbb{N}}$, is an independent and identically distributed (i.i.d.) sequence such that $E \|K_0\|^4 < +\infty$ and $E \|e_0\|^4 < +\infty$.

(c) $\lambda_n = O(n^{-\kappa})$, where $\kappa \in [1-\delta, 1] \cap [0, 1]$. Then Assumptions (a)-(c) in Proposition 3.3 hold.

Based on this result, we apply Algorithm 3.2 to a practical scenario in which a grayscale image of size $256 \times 256$ with pixel values in $[0, 255]$ is degraded by a stochastic blur. The stochastic operator corresponds to a uniform i.i.d. subsampling of a uniform $5 \times 5$ blur, performed in the discrete Fourier domain. More precisely, the value of the frequency response at each frequency bin is kept with probability 0.3 or it is set to zero. In addition, the image is corrupted by an additive zero-mean white Gaussian noise with standard deviation equal to 5. The average signal-to-noise ratio (SNR) is initially equal to 3.4 dB.

In our restoration approach, the function $f$ is the indicator function of the set $[0, 255]^N$, while $g_1 \circ L_1$ is a classical isotropic total variation regularizer, where $L_1$ is the concatenation of the horizontal and vertical discrete gradient operators. Fig. 1 displays the original image, the restored image, as well as two realizations of the degraded images. The SNR for the restored image is equal to 28.1 dB. Fig. 2 shows the convergence behavior of the algorithm. In these experiments, in accordance with Proposition 4.1, we have chosen

$$\begin{cases} m_n = n^{1.1} \\ \lambda_n = (1 + (n/500)^{0.95})^{-1}. \end{cases}$$

\[ \text{(22)} \]

5. CONCLUSION

We have proposed two stochastic proximal splitting algorithms for solving nonsmooth convex optimization problems. These methods require only approximations of the functions used in the formulation of the optimization problem, which is of the utmost importance for solving online signal processing problems. The almost sure convergence of these algorithms has been established. The stochastic version of the primal-dual algorithm that we have investigated has been evaluated in an online image restoration problem in which the data are blurred by a stochastic point spread function and corrupted with noise.