# Hilbert-valued Perturbed Subgradient Algorithms

Kengy Barty
EDF R&D – 1, avenue du Général de Gaulle – F-92141 Clamart Cedex, France
email: kengy.barty@edf.fr

Jean-Sébastien Roy
EDF R&D – 1, avenue du Général de Gaulle – F-92141 Clamart Cedex, France
email: jean-sebastien.roy@edf.fr

Cyrille Strugarek
École Nationale des Ponts et Chaussées – 5-7, avenue Blaise Pascal – Cité Descartes, Champs-sur-Marne
F-77455 Marne-la-Vallée Cedex 2, France
email: strugare@cermics.enpc.fr

We propose a Hilbert-valued perturbed subgradient algorithm with stochastic noises, and provide a convergence proof for this algorithm, under classical assumptions on the descent direction, and new assumptions on the stochastic noises. Instead of requiring the stochastic noises to correspond to martingale increments, we only require these noises to be asymptotically so. Furthermore, the variance of these noises is allowed to grow infinitely under the control of a decreasing sequence linked with the subgradient stepsizes.

This algorithm can be used to solve stochastic closed loop control problems without any a priori discretization of the uncertainty such as linear decision rules or tree representations. It can also be used as a way to perform stochastic dynamic programming without state-space discretization or a priori functional bases (i.e., approximate dynamic programming). Both problems arise frequently for example in power systems scheduling or option pricing. This article focuses on the theorical foundations of the algorithm. The reader is directed to articles [BRS05a] and [BRS05b] for detailed practical experimentations.

In the second part of the paper, we compare this new approach and assumptions with classical ones in the stochastic approximation literature.

As an application of this general setting, we show how the algorithm to solve infinite dimensional stochastic optimization problems developed in [BRS05a] is a special case of our perturbed subgradient algorithm with stochastic noises.

In a last part, we provide a general perturbed subgradient algorithm to solve saddle point problems, and provide a convergence proof under mild assumptions, in the same spirit as the previous theorem.

*Key words*: Stochastic Quasi-Gradient ; Perturbed Subgradient ; Infinite Dimensional Problems ; Nonsmooth optimization

*MSC2000 Subject Classification*: Primary: 62L20; Secondary: 93E20, 93E35

*OR/MS subject classification*: Primary: Programming (Stochastic); Secondary: Programming (Algorithms)

**1. Introduction** Infinite dimensional optimization problems typically appear in the field of stochastic programming or stochastic dynamic programming. In these research fields, the variable of interest is functional, since it is either an optimal control variable (feedback) or Bellman functions. Power systems scheduling as well as option pricing involve this type of difficulties. There is hence a big challenge in proposing efficient methods for solving infinite dimensional problems. Essentially, solutions of such problems can only be estimated, and a natural way to solve it is to use stochastic approximation.

The field of stochastic approximation theory actually began with the seminal paper of Robbins and Monro ([RM51]). Thanks to the various and numerous applications, stochastic approximation has been studied very thoroughly, and the results, either general or more applied, are today well known, especially in the case of finite dimensional stochastic approximation (see, e.g., [Lai03] for an historical survey of stochastic approximation, or [Duf97] for the many branches of this field, or [NH73] for their important monograph).

A lot of various assumptions on stochastic approximation algorithms already exist, and our goal is not to make this field more complicated, but to propose some new general assumptions particularly adapted to stochastic optimal control and infinite dimensional problems. The study of Hilbert-valued stochastic approximations has also been developed, with for example [Ré73a, Ré73b], [Sa80], and further [Gol88]. An important progress in this area is the paper [YZ90], showing the convergence and giving asymptotic properties of an Hilbert-valued Robbins-Monro algorithm under assumptions mimicking usual finite-dimensional assumptions. However, in all these cases, it is not possible to take into account a projection onto a convex subset during the iterations of the algorithms. The important work presented in

[HU75] studies the role of stochastic approximation to solve general Hilbert-valued variational equations, using both probabilistic and variational arguments. In those infinite-dimensional papers, the noise assumptions are practically impossible to verify.
Our work aims at bringing some other assumptions to ensure the convergence of stochastic approximation procedure in the general framework of infinite dimensional Hilbert spaces. It has been motivated by the practical need to propose efficient ways to solve infinite dimensional stochastic optimization problems.
Indeed, most of the existing results cannot be practically applied in infinite dimensional optimization problems such as stochastic programming or stochastic dynamic programming. The existing results are either only available in a finite dimensional setting, or their assumptions are not practically implementable for infinite dimensional problems.

Convergence proofs of stochastic approximation algorithms exist from various point of views. Historically, convergence proofs were given through the so called Robbins-Siegmund Lemma (see [RS71]), and have been then developed by, e.g., [BMP90], or [PT73]. Other approaches have been developed successfully: in the well known monograph [CK78], a stability analysis is developed and the method based on the analysis of the underlying ordinary differential equation, introduced by [DF74], is thoroughly studied. This method has, e.g., been used in [YZ90] to derive their infinite dimensional convergence results. Following the same direction, thanks to general results on Hilbert-valued mixingales (see [CW98]), the recent paper [CW02] provides a comprehensive framework for infinite dimensional Robbins-Monro type procedures. They use modified stochastic approximation with boundedness properties to derive almost sure convergence results and asymptotic normality. Starting from the same ideas, we can also mention [HK96] or [Del96], founded on deterministic arguments, but limited to the finite dimensional case. Another original approach valid for the finite dimensional setting is proposed in [BT00]. Among those approaches, we will follow in this paper an approach more based on probabilistic martingale or quasimartingale arguments (see [Mé82]).

In this paper, we focus on the theoretical and general setting of the stochastic approximation procedure we suggest, centered on the solution of stochastic optimization problems. The paper [HU75] is the nearest to ours, by the techniques used in the proofs and the problems it adresses, but the results are significantly different from ours. The biggest difference is the explicit introduction in our paper of stochastic noises which are not from the beginning martingale increments, but are only asymptotically so. The assumptions made in [HU75] (Theorem 5.1 and Theorem 5.2) involve the whole sequence of the noises, and can hence be difficult to verify. Starting from the same ideas, we propose other assumptions which lead to the same result, but only involve instantaneous perturbations, and are more verifiable practically.
The results of [CW02] differ from ours in that: they are not robust to any projection, except the projection on a particular finite dimensional subspace of the original Hilbert space; they focus on modified stochastic approximation procedures with boundedness properties; they provide more restrictive assumptions on the perturbation sequences, and need the differentiability of the cost function whereas we only need subgradients.

Our paper is organized as follows: Section 3 adresses nonsmooth minimization problems. We provide in subsection 3.1 a convergence proof with general assumptions. In subsection 3.3, we place our result in the context of stochastic approximation and projected subgradient algorithms, and we especially compare it with the result of [AIS98]. In subsection 3.4, we show how our results can be used to prove the convergence of a new algorithm introduced in a forthcoming paper ([BRS05a]), to solve infinite dimensional stochastic optimization problems in practice. In section 4, we propose a perturbed gradient algorithm to solve general saddle point problems, and provide a convergence proof.

## 2. Auxiliary lemmas

We here provide two technical lemmas we will use in the following convergence proofs of Theorems 3.1 and 4.1. These two lemmas were introduced in [Coh84].

Lemma 2.1 *Let $(\mu_k)$ be a sequence of nonnegative real numbers. Let $(\alpha_k)$ and $(\beta_k)$ be sequences of nonnegative real numbers such that $\sum_{k\in\mathbb{N}} \alpha_k < +\infty$ and $\sum_{k\in\mathbb{N}} \beta_k < +\infty$. If we have:*

$$\forall k \in \mathbb{N},\ \mu_{k+1} - \mu_k \leq \alpha_k \mu_k + \beta_k,$$

*then the sequence* $(\mu_k)_{k\in\mathbb{N}}$ *is bounded.*

Proof. The proof is classical and can be found, e.g., in [Coh84]. □

Lemma 2.2 *Let* $(\Omega, \mathcal{F}, \mathbb{P})$ *be some probability space, equipped with a filtration* $(\mathcal{F}_k)$*. Let* $J$ *be a mapping from an Hilbert space* $H$ *to the real line* $\mathbb{R}$*. Let* $(u_k)$ *be a sequence of random variables with values in* $H$*, such that for all* $k \in \mathbb{N}$*,* $u_k$ *is* $\mathcal{F}_k$*-measurable, and* $(\gamma_k)$ *a sequence of nonnegative real numbers such that:*

*(i)* $\sum_{k\in\mathbb{N}} \gamma_k = +\infty$,

*(ii)* $\exists \mu \in \mathbb{R}$, $\sum_{k\in\mathbb{N}} \gamma_k (J(u_k) - \mu) < +\infty$, *and* $\forall k \in \mathbb{N}$, $J(u_k) - \mu \geq 0$, *a.s.*

*(iii)* $\exists \delta > 0$, $\forall k \in \mathbb{N}$, $J(u_k) - \mathbb{E}(J(u_{k+1})|\mathcal{F}_k) \leq \delta\gamma_k$, *a.s.*

*Then* $(J(u_k))$ *a.s. converges to* $\mu$*.*

Proof. For all $\alpha \in \mathbb{R}$, let us define the subset $N_\alpha$ of $\mathbb{N}$ such that:

$$N_\alpha := \{k \in \mathbb{N} \;:\; J(u_k) - \mu \leq \alpha, \text{ a.s.}\}.$$

We will also denote by $N_\alpha^c$ the complementary set of $N_\alpha$ in $\mathbb{N}$. Assumptions $(i - ii)$ imply that $N_\alpha$ is not finite.
Following $(ii)$, we have:

$$+\infty > \sum_{k\in\mathbb{N}} \gamma_k (J(u_k) - \mu) \geq \sum_{k\in N_\alpha^c} \gamma_k (J(u_k) - \mu) \geq \alpha \sum_{k\in N_\alpha^c} \gamma_k.$$

It proves that for all $\beta > 0$, there is some $n_\beta \in \mathbb{N}$ such that $\sum_{k\in N_\alpha^c,\, k\geq n_\beta} \gamma_l \leq \beta$.
Let $\epsilon > 0$. Take $\alpha = \epsilon/2$ and $\beta = \epsilon/(2\delta)$. For all $k \geq n_\beta$, we have two possibilities:

- If $k \in N_\alpha$, then $J(u_k) - \mu \leq \alpha < \epsilon$.
- If $k \in N_\alpha^c$, let $m$ be the smallest element of $N_\alpha$ such that $m \geq k$ (we know that it exists since $N_\alpha$ is not finite). We can hence write:

$$\begin{aligned} J(u_k) - \mu =& J(u_k) - \mathbb{E}(J(u_m)|\mathcal{F}_k) + \mathbb{E}(J(u_m)|\mathcal{F}_k) - \mu \\ =& \mathbb{E}\left( \sum_{l=k}^{m-1} J(u_l) - \mathbb{E}(J(u_{l+1})|\mathcal{F}_l) \middle| \mathcal{F}_k \right) + \mathbb{E}(J(u_m)|\mathcal{F}_k) - \mu, \\ \leq& \delta \left( \sum_{l=k}^{m-1} \gamma_l \right) + \alpha \leq \delta \left( \sum_{l\in N_\alpha^c,\, l\geq n_\beta} \gamma_l \right) + \alpha \leq \epsilon, \end{aligned}$$

and it concludes the proof. □

We end by a lemma on quasi-Féjer sequences introduced in the finite dimensional setting in [Erm66]. It can be seen as a probabilistic version of a result proposed by [AIS98].

Lemma 2.3 *Let* $H$ *be an Hilbert space, and* $V$ *a nonempty subset of* $H$*. Let* $(x_k)$ *be a sequence of random variable with values in* $H$*. Define for all* $k \in \mathbb{N}$*,* $\mathcal{F}_k = \sigma(x_l,\ l \leq k)$ *the sigma fields generated by the sequence. Assume that*

$$\forall x^* \in V,\ \exists (\delta_k) \subset \mathbb{R}^+,\ \sum_{k\in\mathbb{N}} \delta_k < +\infty,\ \exists \tilde{k} \in \mathbb{N},\ \forall k \geq \tilde{k},\ \mathbb{E}\left(\|x_{k+1} - x^*\|^2|\mathcal{F}_k\right) \leq \|x_k - x^*\|^2 + \delta_k,\ \textit{almost surely.}$$

*Then, it holds that*

*(i)* $(x_k)$ *is a.s. bounded,*

*(ii)* $(\|x_k - x^*\|^2)$ *converges a.s. for all* $x^* \in V$*,*

*(iii) if all weak accumulation points of* $(x_k)$ *belong a.s. to* $V$*, then* $(x_k)$ *is a.s. weakly convergent, i.e., it has a.s. a unique accumulation point.*

Proof. The presence of conditional expectations does not modify the proof of [AIS98], Definition 1 and Proposition 1. It just forces the inequalities to be valid only almost surely. □

## 3. Minimization Problems

**3.1 Algorithm** We focus on the problem:

$$\min_{x\in X} f(x) \qquad (1)$$
$$\text{s.t. } x \in X^f.$$

where:

- $X$ is some Hilbert space with inner product and norm respectively denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$,
- $f : X \to \mathbb{R}$ is a convex mapping,
- $X^f$ is a closed convex subset of $X$ (we will distinguish the case where $X^f$ is a closed subspace), and $\Pi_{X^f}$ denotes the projection onto $X^f$.

We write the following perturbed subgradient algorithm for problem (1), i.e.,

Algorithm 3.1 *Step 0: let $x_0 \in X^f$.*
*Step $t+1 \in \mathbb{N}$:*

$$x_{t+1} = \Pi_{X^f}\left(x_t + \gamma_t(s_t + w_t)\right),$$

where $-s_t$ typically belongs to the convex subdifferential of $f$ at point $x_t$ (what will be called in the following a descent direction), $w_t$ is a random noise (the perturbation), and $\gamma_t$ is a nonnegative deterministic decreasing stepsize. More precisely, $(w_t)$ is a sequence of random variables on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $X$, such that $(x_t)$ becomes itself a sequence of random variables with values in $X$. Hence, the convergence of Algorithm 3.1 can only be stated in a probabilistic sense, and it will be given here in terms of almost sure convergence. Associated with that algorithm, we can define a filtration $(\mathcal{F}_t)$ on $(\Omega, \mathcal{A}, \mathbb{P})$ by letting:

$$\forall t \in \mathbb{N},\ \mathcal{F}_t := \sigma(x_0, \ldots, x_t).$$

**3.2 Convergence Proof**

Definition 3.1 (Coercivity) *A mapping $h : X \to \mathbb{R}$ is said to be* coercive *if and only if*

$$\lim_{\|x\|\to\infty} h(x) = +\infty.$$

We provide a convergence proof for Algorithm 3.1 in two main cases corresponding to two different constraints, namely $X^f$ being a closed vector subspace of $X$ and $X^f$ being a closed convex subset of $X$ (and not a subspace).

Theorem 3.1 *(i) Assume that $f$ is convex and coercive. Then $\partial f(x) \neq \emptyset$ for all $x \in X^f$ which is either a closed convex subset or a closed vector subspace of $X$.*
*(ii) Assume that for all $t \in \mathbb{N}$, $s_t$ is $\mathcal{F}_t$-measurable.*
*(iii) Assume that $f$ has linearly bounded subgradients, i.e.,*

$$\exists a_1, a_2 \geq 0,\ \forall x \in X^f,\ \forall v \in \partial f(x),\ \|v\| \leq a_1\|x\| + a_2. \qquad (2)$$

*(iv) Assume that there exists $\kappa > 0$, such that for all $t \in \mathbb{N}$,*

$$-\frac{1}{\kappa}s_t \in \partial f(x_t) \qquad (3)$$

*(v) Assume that there are $b \geq 0$, $A > 0$ and two deterministic nonnegative sequences $(\epsilon_t)$ and $(\eta_t)$ such that for all $t \in \mathbb{N}$ there exists $v_t \in \partial f(x_t)$ such that,*

$$\|\mathbb{E}(w_t|\mathcal{F}_t)\| \leq b\eta_t\left(1 + \|v_t\|\right), \qquad (4a)$$

$$\mathbb{E}\left(\|w_t\|^2|\mathcal{F}_t\right) \leq A\left(1 + \frac{1}{\epsilon_t}\|v_t\|^2\right). \qquad (4b)$$

*If $X^f$ is a closed convex set but not a subspace, then assume also that there exist a bounded mapping $g : \mathbb{R} \to \mathbb{R}$, and for all $t \in \mathbb{N}$ some $v_t \in \partial f(x_t)$ such that,*

$$\mathbb{E}\left(\|w_t\| | \mathcal{F}_t\right) \leq g(\|v_t\|). \tag{4c}$$

*(vi) Assume that the sequences $(\gamma_t)$, $(\epsilon_t)$ and $(\eta_t)$ are such that:*

$$\forall t \in \mathbb{N},\ \gamma_t, \epsilon_t > 0, \quad \sum_{t\in\mathbb{N}} \gamma_t = +\infty, \quad \sum_{t\in\mathbb{N}} (\gamma_t)^2 < +\infty, \quad \sum_{t\in\mathbb{N}} \frac{(\gamma_t)^2}{\epsilon_t} < +\infty, \quad \sum_{t\in\mathbb{N}} b\gamma_t\eta_t < +\infty. \tag{5}$$

*Then the problem* (1) *has solutions, and denoting its solution set by $S$ and its optimal value by $f_S$, $f(x_t) \to f_S$ almost surely, as $t$ goes to infinity, and $(x_t)$ almost surely weakly converges to a point of $S$.*
*(vii) If moreover $f$ is strongly convex (with modulus $B > 0$), then $S = \{x^*\}$ and $(x_t)$ strongly converges almost surely to $x^*$.*

Proof. We use the scheme introduced by [CC90], using a Lyapunov function. Let $x^* \in S$, and let, for all $x \in X$, $\Lambda(x) := \frac{1}{2}\|x - x^*\|^2$ be our Lyapunov function. We will study its evolution over the iterations. For all $t \in \mathbb{N}$, we will denote $\Lambda_t = \Lambda(x_t)$. Let $t \in \mathbb{N}$.

$$\Lambda_{t+1} - \Lambda_t = \frac{1}{2}\|x_{t+1} - x_t\|^2 + \langle x_{t+1} - x_t, x_t - x^* \rangle. \tag{6}$$

By definition of $x_{t+1}$ (see Algorithm 3.1) and nonexpansiveness of the projection, it comes

$$\Lambda_{t+1} = \frac{1}{2}\|\Pi_{X^f}\left(x_t + \gamma_t(s_t + w_t)\right) - \Pi_{X^f}\left(x^*\right)\|^2 \leq \frac{1}{2}\|x_t + \gamma_t(s_t + w_t) - x^*\|^2. \tag{7}$$

Using Pythagore's inequality, one gets therefore

$$\Lambda_{t+1} - \Lambda_t \leq \frac{(\gamma_t)^2}{2}\|s_t + w_t\|^2 + \gamma_t\langle s_t + w_t, x_t - x^* \rangle. \tag{8}$$

Note that assumption (3) and convexity of $f$ imply that

$$\langle s_t, x_t - x^* \rangle \leq \kappa\left(f(x^*) - f(x_t)\right).$$

We take now the conditional expectation with respect to $\mathcal{F}_t$ in (8)

$$\begin{aligned}
\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right) - \Lambda_t \leq & \frac{(\gamma_t)^2}{2}\mathbb{E}\left(\|s_t + w_t\|^2|\mathcal{F}_t\right) + \gamma_t\langle s_t, x_t - x^* \rangle \\
& + \gamma_t\langle \mathbb{E}\left(w_t|\mathcal{F}_t\right), x_t - x^* \rangle, \\
\leq & \frac{(\gamma_t)^2}{2}\mathbb{E}\left(\|s_t + w_t\|^2|\mathcal{F}_t\right) + \gamma_t\kappa\left(f(x^*) - f(x_t)\right) \\
& + b\eta_t\gamma_t\|x_t - x^*\|\left(1 + \|v_t\|\right) \text{ , by assumptions (3),(4a)} \\
\leq & \frac{(\gamma_t)^2}{2}\mathbb{E}\left(\|s_t + w_t\|^2|\mathcal{F}_t\right) + \gamma_t\kappa\left(f(x^*) - f(x_t)\right) \\
& + b\eta_t\gamma_t\|x_t - x^*\|\left(1 + a_1\|x_t - x^*\| + a_1\|x^*\| + a_2\right), \text{ by assumption (2)}
\end{aligned} \tag{9}$$

We now use on the norms the classical scalar inequality for $a \in \mathbb{R}$, $a \leq 1 + a^2$, and we get from (9)

$$\begin{aligned}
\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right) - \Lambda_t \leq & \frac{(\gamma_t)^2}{2}\mathbb{E}\left(\|s_t + w_t\|^2|\mathcal{F}_t\right) + \gamma_t\kappa\left(f(x^*) - f(x_t)\right) \\
& + b\eta_t\gamma_t\left(1 + a_1 + a_1\|x^*\| + a_2\right)\|x_t - x^*\|^2 + b\eta_t\gamma_t\left(1 + a_1\|x^*\| + a_2\right).
\end{aligned} \tag{10}$$

We now focus on the first term of the right hand side. By the classical inequality for $a, b \in \mathbb{R}$, $(a + b)^2 \leq 2(a^2 + b^2)$:

$$\begin{aligned}
\frac{(\gamma_t)^2}{2}\mathbb{E}\left(\|s_t + w_t\|^2|\mathcal{F}_t\right) \leq & (\gamma_t)^2\left(\|s_t\|^2 + \mathbb{E}\left(\|w_t\|^2|\mathcal{F}_t\right)\right), \\
\leq & (\gamma_t)^2\kappa\left(2((a_2 + a_1\|x^*\|)^2 + (a_1)^2\|x_t - x^*\|^2)\right) \\
& + (\gamma_t)^2 A\left(1 + \frac{1}{\epsilon_t}\|v_t\|^2\right), \text{ by assumptions (3),(2),(4b)} \\
\leq & (\gamma_t)^2\kappa\left(2((a_2 + a_1\|x^*\|)^2 + (a_1)^2\|x_t - x^*\|^2)\right) \\
& + (\gamma_t)^2 A\left(1 + \frac{2}{\epsilon_t}((a_2 + a_1\|x^*\|)^2 + (a_1)^2\|x_t - x^*\|^2)\right) \\
\leq & \left(C_1(\gamma_t)^2 + C_2\frac{(\gamma_t)^2}{\epsilon_t}\right)\|x_t - x^*\|^2 + \left(C_3(\gamma_t)^2 + C_4\frac{(\gamma_t)^2}{\epsilon_t}\right)
\end{aligned} \tag{11}$$

with $C_1, C_2, C_3, C_4$ nonnegative deterministic scalars. We now go back to equation (10), and we obtain:

$$\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right) - \Lambda_t \leq \alpha_t\Lambda_t + \beta_t + \gamma_t\kappa\left(\underbrace{f(x^*) - f(x_t)}_{\leq 0,\text{ by optimality}}\right) \leq \alpha_t\Lambda_t + \beta_t, \tag{12}$$

with:

$$\alpha_t = 2b\eta_t\gamma_t\left(1 + a_1 + a_1\|x^*\| + a_2\right) + 2C_1(\gamma_t)^2 + 2C_2\frac{(\gamma_t)^2}{\epsilon_t},$$
$$\beta_t = b\eta_t\gamma_t\left(1 + a_1\|x^*\| + a_2\right) + C_3(\gamma_t)^2 + C_4\frac{(\gamma_t)^2}{\epsilon_t}.$$

Thus, $(\alpha_t)$ and $(\beta_t)$ form two summable sequences (see assumption (5)). Let us take the expectation in (12), and denote $\lambda_t = \mathbb{E}\left(\Lambda_t\right)$. It yields:

$$\lambda_{t+1} - \lambda_t \leq \alpha_t\lambda_t + \beta_t + \gamma_t\kappa\mathbb{E}\left(\underbrace{f(x^*) - f(x_t)}_{\leq 0,\text{ by optimality}}\right). \tag{13}$$

Using Lemma 2.1 (see section 2), it implies that $\lambda_t$ is bounded by, say, some $M > 0$. We now prove that $\Lambda_t$ is a convergent quasimartingale. Indeed:

- By definition, $\Lambda_t$ is $\mathcal{F}_t$ measurable for all $t \in \mathbb{N}$.
- By definition, for all $t \in \mathbb{N}$, $\Lambda_t \geq 0$, and therefore $\inf_{t\in\mathbb{N}}\mathbb{E}\left(\Lambda_t\right) \geq 0$.
- Let for all $t \in \mathbb{N}$, $D_t := \{\mathbb{E}\left(\Lambda_{t+1} - \Lambda_t|\mathcal{F}_t\right) > 0\}$. Define $1_{D_t} : \Omega \to \{0,1\}$ by $1_{D_t}(\omega) = 1$ if $\omega \in D_t$ and $1_{D_t}(\omega) = 0$ if $\omega \notin D_t$. $1_{D_t}$ is $\mathcal{F}_t$-measurable. Hence, with (12), we have:
$$\begin{aligned}\sum_{t\in\mathbb{N}}\mathbb{E}\left(1_{D_t}\cdot\left(\Lambda_{t+1} - \Lambda_t\right)\right) =& \sum_{t\in\mathbb{N}}\mathbb{E}\left(1_{D_t}\cdot\mathbb{E}\left(\Lambda_{t+1} - \Lambda_t|\mathcal{F}_t\right)\right),\\ \leq& \sum_{t\in\mathbb{N}}\mathbb{E}\left(1_{D_t}\left(\alpha_t\Lambda_t + \beta_t\right)\right),\\ \leq& \sum_{t\in\mathbb{N}}\left(\alpha_t M + \beta_t\right) < +\infty.\end{aligned}$$
- Since $\Lambda_t \geq 0$, it is clear that $\sup_{t\in\mathbb{N}}\mathbb{E}\left(\min(\Lambda_t, 0)\right) < +\infty$. Consequently, using the result of [Mé82] (pp. 49-51), the sequence $(\Lambda_t)$ is a quasimartingale and converges a.s. to some integrable random variable. Hence, it is a.s. bounded, and hence, by definition, and using assumption (2), the sequences $(x_t)$ and $(s_t)$ are a.s. bounded in $X$.

We now prove that $(f(x_t))$ a.s. converges to $f(x^*)$. Coming back to (13), we obtain:

$$\kappa\gamma_t\mathbb{E}\left(f(x_t) - f(x^*)\right) \leq \alpha_t\lambda_t + \beta_t + \lambda_t - \lambda_{t+1}.$$

We sum this inequality for $t = 0, \ldots, n$:

$$\begin{aligned}\kappa\sum_{t=0}^{n}\gamma_t\mathbb{E}\left(f(x_t) - f(x^*)\right) \leq& \lambda_0 - \lambda_{n+1} + \sum_{t=0}^{n}(\alpha_t M + \beta_t),\\ \leq& M + M\sum_{t=0}^{n}\alpha_t + \sum_{t=0}^{n}\beta_t.\end{aligned} \tag{14}$$

We make $n \to \infty$:

$$\sum_{t\in\mathbb{N}}\gamma_t\mathbb{E}\left(f(x_t) - f(x^*)\right) < +\infty.$$

By optimality, all the terms under the expectation are a.s. nonnegative. Thus, almost surely:

$$\sum_{t\in\mathbb{N}}\gamma_t\left(f(x_t) - f(x^*)\right) < +\infty. \tag{15}$$

We now want to use Lemma 2.2. Let $t \in \mathbb{N}$. By convexity of $f$, since $-\frac{1}{\kappa} s_t \in \partial f(x_t)$,

$$\begin{aligned} f(x_t) - f(x_{t+1}) \leq & \frac{-1}{\kappa} \langle s_t, x_t - x_{t+1} \rangle, \\ = & \frac{-1}{\kappa} \langle s_t, x_t - \Pi_{X_f} \left( x_t + \gamma_t (s_t + w_t) \right) \rangle. \end{aligned} \tag{16}$$

Again, we distinguish between two cases:

- If $X^f$ is a closed vector subspace of $X$, the projection mapping is self-adjoint and linear and hence, (16) reads:
$$f(x_t) - f(x_{t+1}) \leq \frac{\gamma_t}{\kappa} \langle \Pi_{X^f} \left( s_t \right), s_t + w_t \rangle.$$
By taking the conditional expectation with respect to $\mathcal{F}_t$, one gets
$$f(x_t) - \mathbb{E} \left( f(x_{t+1}) | \mathcal{F}_t \right) \leq \frac{\gamma_t}{\kappa} \langle \Pi_{X^f} \left( s_t \right), s_t + \mathbb{E} \left( w_t | \mathcal{F}_t \right) \rangle,$$
since the other random variables are all $\mathcal{F}_t$-measurable. Since $(s_t)$ and $(x_t)$ are a.s. bounded on $X$, one obtains with assumptions (2),(4a) that there is some $\delta > 0$ such that:
$$f(x_t) - \mathbb{E} \left( f(x_{t+1}) | \mathcal{F}_t \right) \leq \gamma_t \delta. \tag{17}$$

- If $X^f$ is a closed convex subset of $X$, (16) reads
$$f(x_t) - f(x_{t+1}) \leq \frac{\gamma_t}{\kappa} \|s_t\| \times \|s_t + w_t\|,$$
by using the nonexpansiveness of the projection and Cauchy-Schwartz inequality. By taking now the conditional expectation with respect to $\mathcal{F}_t$, and using assumption (4c), since $(s_t)$ and $(x_t)$ are a.s. bounded, there exists some deterministic constant $\delta > 0$ such that
$$f(x_t) - \mathbb{E} \left( f(x_{t+1}) | \mathcal{F}_t \right) \leq \gamma_t \delta. \tag{18}$$

Hence, we can in any case apply Lemma 2.2, with (15) and (17) or (18), which yields

$$\lim_{t \to \infty} f(x_t) = f(x^*) \quad \text{almost surely.} \tag{19}$$

Let $\bar{x}$ be a cluster point of $(x_t)$. Hence there is some subsequence $(x_{\phi(t)})$ which weakly converges to $\bar{x}$. Since $X^f$ is convex and closed, $\bar{x} \in X^f$, and by lower semicontinuity of $f$, it holds:

$$f(\bar{x}) \leq \liminf_{t \to \infty} f(x_{\phi(t)}) = f(x^*),$$

hence, $\bar{x} \in S$, i.e., every weak accumulation point of $(x_t)$ belongs to $S$.
Moreover, by boundedness of $(x_t)$ and inequality (12), it holds that

$$\forall x^* \in S,\ \mathbb{E} \left( \|x_{t+1} - x^*\|^2 | \mathcal{F}_t \right) \leq \|x_t - x^*\|^2 + \underbrace{\beta_t + \alpha_t \max_{s \in \mathbb{N}} \|x_t - x^*\|^2}_{\delta_t} \qquad \text{a.s.} \tag{20}$$

with $(\delta_t)$ a summable sequence. Lemma 2.3 and the result on accumulation points of $(x_t)$ provides that almost surely, $(x_t)$ weakly converges.
Suppose now that $f$ is strongly convex with modulus $B > 0$. In this case, $S$ reduces to a singleton $\{x^*\}$. By definition, one has

$$f(x_t) - f(x^*) \geq \langle v^*, x_t - x^* \rangle + \frac{B}{2} \|x^* - x_t\|^2, \tag{21}$$

for all $v^* \in \partial f(x^*)$. The unique solution $\{x^*\}$ is moreover characterized by the optimality condition :

$$\exists v^* \in \partial f(x^*),\ \forall x \in X^f,\ \langle v^*, x - x^* \rangle \geq 0.$$

Applying (21) to the subgradient corresponding to the previous variational inequality gives therefore

$$f(x_t) - f(x^*) \geq \frac{B}{2} \|x^* - x_t\|^2, \tag{22}$$

which shows with (19) the strong convergence of $(x_t)$ to $x^*$ almost surely, and completes the proof. □

Remark 3.1 (Strong convexity) *Following the work of [BL72] we can weaken point (vii) of Theorem 3.1 related to the strong convexity assumption. Indeed, if the function $f$ is only required to be strictly convex, the strong convergence of $(x_t)$ towards the unique solution $x^*$ of problem* (1) *can also be proved. For the sake of simplicity and clarity of the proof, we have here preferred to make the strong convexity assumption.*

Remark 3.2 (Random stepsizes) *The stepsizes $(\rho_t)$ and $(\epsilon_t)$ introduced in Theorem 3.1 can be taken as random sequences with nonnegative values, such that for all $t \in \mathbb{N}$, $\rho_t$ and $\epsilon_t$ are $\mathcal{F}_t$-measurable. Indeed, the main result we use in the proof, namely Métivier's Proposition on quasimartingales, is available with $(\mathcal{F}_t)$ adapted sequences for the stepsizes. This remark allows possible online definition for these stepsizes, depending on the past $\sigma$-fields.*

Remark 3.3 (Boundedness of the noise) *Assumption* (4c) *is necessary only in the case of a closed convex constraint set. However, it is possible to relax this assumption when $f$ is strongly convex, by using classical arguments directly on the Lyapunov function $\Lambda$ introduced in the proof and invoking Robbins-Siegmund Lemma (see [RS71]).*

Remark 3.4 (Descent direction) *Assumption* (3) *may be replaced by the following weaker ones:*

$$\begin{aligned} \forall t \in \mathbb{N},\ \forall x^* \in S,\ \langle s_t, x_t - x^*\rangle \leq & \kappa\left(f(x^*) - f(x_t)\right), \\ \forall t \in \mathbb{N},\ \exists v_t \in \partial f(x_t),\ \|s_t\| \leq & c\left(1 + \|v_t\|\right). \end{aligned}$$

*However, for the simplicity of the statement, we preferred to directly assume that $-s_t/\kappa \in \partial f(x_t)$, which, together with* (2)*, implies those equations.*

Remark 3.5 (Kiefer-Wolfowitz) *Another stochastic approximation algorithm suitable for differentiable and finite-dimensional problems and referred to as Kiefer-Wolfowitz algorithm (see [KW52]) computes an approximation of the true gradient, on the basis of finite differences. There are in this algorithm two stepsizes, the one corresponding to the descent step $\gamma_t$, and the other corresponding to the finite difference approximation. These two steps are required to satisfy joint decreasing assumptions, which are exactly the same as* (5)*, if you consider the finite difference stepsize to correspond to $(\eta_t)^2$, when $m = 1$.*

Remark 3.6 (Convex subset and linear subspaces) *We distinguish in the assumptions the cases where $X^f$ is a general convex subset of the Hilbert space $X$, and the cases where it has moreover a linear subspace structure. Since the linear subspaces are convex, this distinction may seem unnecessary. However, the assumptions on the perturbations $(w_t)$ may be weakened in the subspace case, due to the special properties of the projection mapping on a linear subspace. It is the reason why the two cases are separated in the convergence theorem.*

**3.3 Comparison with existing results** Among most of literature concerning stochastic approximation algorithms, beginning with [RM51], it is hard to find a result really appropriate to a comparison with our result. The contributions of Ermoliev in this field (e.g., [Erm66, Erm66], or [Erm76]) are the closest from ours in the spirit: they aim at solving convex constrained optimization problems by variational techniques, but remain in a finite dimensional setting. The results of [Ré73a, Ré73b] and [CW02] are close to ours, but the algorithms do not present the same abilities, especially concerning general projections. A comparison with those results would not be sensible.

On the contrary, the results in the infinite dimensional setting for deterministic projected subgradients algorithms are easier to compare with our result. We can especially cite the work [AIS98]. It provides a convergence theorem for projected $\varepsilon$-subgradient algorithms. This theorem relies on convexity assumption and local boundedness assumptions for the subdifferential. Our result may be seen as a perturbed or stochastic version of this result, where our assumption (2) plays the role (in a more restricted way) of the local boundedness of the subdifferential. Assumptions on the decreasing stepsize sequences are essentially the same. Moreover, if we replace in the theorem 3.1 the subgradients where they appear in the assumptions by $\nu_t$-subgradients, with $(\nu_t)$ another decreasing sequence, we could prove the same convergence result.

**3.4 Application to closed loop problems** We here assume that $X = L^2(\mathbb{R}^m, \mathbb{R}^p, \mathbb{P})$, and that there is some random variable denoted by $\boldsymbol{\xi}$ and a convex, lower semicontinuous and differentiable in its first component mapping $j : \mathbb{R}^p \times \mathbb{R}^m \to \mathbb{R}$ such that:

$$\forall x \in L^2(\mathbb{R}^m, \mathbb{R}^p, \mathbb{P}),\ f(x) = \mathbb{E}\left(j(x(\boldsymbol{\xi}), \boldsymbol{\xi})\right).$$

Let $X^f$ be a closed convex subset of $X$. We hence focus on the problem :

$$\min_{x \in X^f} \mathbb{E}\left(j(x(\boldsymbol{\xi}), \boldsymbol{\xi})\right). \tag{23}$$

Notice that since $j$ is convex, then so is $f$, and hence, $f$ and $j$ are differentiable, and it holds:

$$\forall x \in X,\ \nabla f(x)(\cdot) = \nabla_x j(x(\cdot), \cdot).$$

Such problems are often referred to as *closed loop stochastic optimization problems.* A recent work [BRS05a] focused on this problem, and proposed a stochastic gradient type algorithm to solve this problem, based on the use of kernels, i.e., mappings $K_t : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$. Their algorithm is the following:

Algorithm 3.2 *Step t:*

- *Draw $\boldsymbol{\xi}_{t+1}$ identically, independently from the past drawings,*
- *Update:*

$$x_{t+1}(\cdot) = \Pi_{X^f}\left(x_t(\cdot) - \rho_t \nabla_x j(x_t(\boldsymbol{\xi}_{t+1}), \boldsymbol{\xi}_{t+1}) K_t(\boldsymbol{\xi}_{t+1}, \cdot)\right),$$

They provide a convergence proof for this algorithm. We claim here that this algorithm (whose abilities and applications are developed in [BRS05a]) is a special case of Algorithm 3.1. Indeed, let us define:

- $\mathcal{F}_t := \sigma(x_0, \ldots, x_t) = \sigma(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_t)$
- $s_t := -\nabla_x j(x_t(\cdot), \cdot)$,
- $w_t := \nabla_x j(x_t(\cdot), \cdot) - \nabla_x j(x_t(\boldsymbol{\xi}_{t+1}), \boldsymbol{\xi}_{t+1}) \frac{1}{\epsilon_t} K_t(\boldsymbol{\xi}_{t+1}, \cdot)$.

Then, Algorithm 3.2 can be rewritten as:

$$x_{t+1} = \Pi_{X^f}\left(x_t + \rho_t \epsilon_t (s_t + w_t)\right),$$

which corresponds exactly to Algorithm 3.1 with $\gamma_t = \rho_t \epsilon_t$, and $\eta_t = (\epsilon_t)^{1/m}$ to satisfy the noise assumptions.

Clearly, assumptions on convexity of $f$ and (3) are satisfied with our choice of $s_t$. We have also to assume that $j(\cdot, \xi)$ has uniformly (in $\xi$) linearly bounded gradients. We now focus on assumptions (4a)–(4b)–(4c) and (5).
In [BRS05a], the kernel functions are assumed to be such that:

$$\forall t \in \mathbb{N},\ \|s_t - \mathbb{E}\left(s_t(\boldsymbol{\xi}) \frac{1}{\epsilon_t} K_t(\boldsymbol{\xi}, \cdot)\right)\| \leq b_1 \eta_t \left(1 + \|s_t\|\right),$$

$$\forall x \in \mathbb{R}^m,\ \mathbb{E}\left((K_t(x, \boldsymbol{\xi}))^2\right) \leq b_2 \epsilon_t, \tag{24}$$

with two deterministic positive scalars $b_1$ and $b_2$. The stepsizes are assumed to decrease to 0 and to satisfy:

$$\epsilon_t, \rho_t > 0, \quad \sum_{t \in \mathbb{N}} \epsilon_t \rho_t = +\infty, \quad \sum_{t \in \mathbb{N}} \rho_t \epsilon_t \eta_t < +\infty, \quad \sum_{t \in \mathbb{N}} (\rho_t)^2 \epsilon_t < +\infty. \tag{25}$$

Clearly, (24) and (25) ensure that assumptions (4a)–(4b)–(4c) and (5) of Theorem 3.1 are satisfied. Hence, Algorithm 3.2 converges.

An interesting application of Algorithm 3.2 appears when $X^f$ is the intersection of a closed convex set $X_c$ and a linear subspace $X_v$ stable under projection on $X_c$. Indeed, thanks to the Proposition 3.1, one can rewrite the algorithm as follows:

$$x_{t+1}(\cdot) = \Pi_{X_c}\left\{x_t(\cdot) - \rho_t \Pi_{X_v}\left(\nabla_x j(x_t(\boldsymbol{\xi}_{t+1}), \boldsymbol{\xi}_{t+1}) K_t(\boldsymbol{\xi}_{t+1}, \cdot)\right)\right\}.$$

To sum up, if the projection on the convex set is easy to compute, a preprocessing of the kernels $K_t$ may help to compute the projection on the linear subspace. The following proposition on projections is known, and its proof simply relies on the very definition of a projection on a convex set.

Proposition 3.1 (Projection on an intersection) *Let $X^f = X_v \cap X_c$, with $X_v$ a closed vector subspace of $X$ and $X_c$ a closed convex subset of $X$. Assume that $X^f$ is not empty, and that $\Pi_{X_c}(X_v) \subset X_v$. Then it holds:*

$$\Pi_{X_v \cap X_c} = \Pi_{X_c} \circ \Pi_{X_v}.$$

Proof. One uses the variational inequality characterizing the projection, namely:

$$\forall x \in X,\ \forall y \in X_v \cap X_c,\ \langle x - \Pi_{X_v \cap X_c}(x), y - \Pi_{X_v \cap X_c}(x)\rangle \leq 0.$$

Let $x \in X$, and $y \in X_v \cap X_c$. Then, one has

$$\begin{aligned}\langle x - \Pi_{X_c}\left(\Pi_{X_v}(x)\right), y - \Pi_{X_c}\left(\Pi_{X_v}(x)\right)\rangle =&\langle \Pi_{X_v}(x) - \Pi_{X_c}\left(\Pi_{X_v}(x)\right), y - \Pi_{X_c}\left(\Pi_{X_v}(x)\right)\rangle \\ &+ \langle x - \Pi_{X_v}(x), y - \Pi_{X_c}\left(\Pi_{X_v}(x)\right)\rangle\end{aligned}$$

The first term of the right hand-side is negative by characterization of the projection on $X_c$ of $\Pi_{X_v}(x)$. On the other hand, one has by assumption that $\Pi_{X_c}\left(\Pi_{X_v}(x)\right) \in X_v$, and hence

$$\langle x - \Pi_{X_v}(x), y - \Pi_{X_c}\left(\Pi_{X_v}(x)\right)\rangle = \langle \Pi_{X_v}\left(x - \Pi_{X_v}(x)\right), y - \Pi_{X_c}\left(\Pi_{X_v}(x)\right)\rangle = 0,$$

since $\Pi_{X_v}$ is linear and self-adjoint. It concludes the proof. □

## 4. Saddle Point Problems

**4.1 Algorithm** We focus here on the problem:

$$\min_{x \in X} \max_{p \in P} L(x,p), \qquad (26)$$
$$\text{s.t. } x \in X^f,\ p \in P^f,$$

where

- $X$ and $P$ are two Hilbert spaces with respective inner product and norm denoted by $\langle \cdot, \cdot \rangle_X$, $\langle \cdot, \cdot \rangle_P$ and $\| \cdot \|_X$, $\| \cdot \|_P$,
- $L : X \times P \to \mathbb{R}$ is a convex-concave mapping,
- $X^f, P^f$ are either closed convex subsets or closed subspaces of $X$ and $P$ respectively, and $\Pi_{\cdot}(\cdot)$ will denote the projection.

We write the following perturbed subgradient algorithm for problem (26):

Algorithm 4.1 *Step $t \in \mathbb{N}$:*

$$\begin{aligned}x_{t+1} =&\Pi_{X^f}\left(x_t + \gamma_t^x(s_t + w_t)\right), \\ p_{t+1} =&\Pi_{P^f}\left(p_t + \gamma_t^p(r_t + v_t)\right).\end{aligned}$$

$s_t$ is hence as before a descent direction, while $r_t$ is an ascent direction, and $w_t, v_t$ are the perturbations. The nonnegative stepsizes $\gamma_t^x, \gamma_t^p$ will be in the following the same.

**4.2 Convergence Proof** We have the following theorem:

Theorem 4.1 (Saddle Point Problems) *(i) Assume that $L(\cdot, p) : X \to \mathbb{R}$ is convex for all $p \in P$, and that $L(x, \cdot) : P \to \mathbb{R}$ is concave for all $x \in X$. Assume moreover that $X^f$ and $P^f$ are closed convex subsets of $X$ and $P$, and that there exists a saddle point $(x^*, p^*)$ to $L$ over $X^f \times P^f$.*
*(ii) Let $(\mathcal{F}_t)$ be a filtration, and assume that for all $t \in \mathbb{N}$, $x_t, s_t, p_t$ and $r_t$ are $\mathcal{F}_t$-measurable.*
*(iii) Assume that for all $(x,p) \in X^f \times P^f$, $\partial_x L(x,p)$ and $\partial_p L(x,p)$ are not empty, and that there exist $a_1, a_2 > 0$ such that*

$$\forall (x,p) \in X^f \times P^f,\ \forall u_x \in \partial_x L(x,p),\ \|u_x\|_X \leq a_1 \|x\|_X + a_2, \qquad (27a)$$
$$\forall (x,p) \in X^f \times P^f,\ \forall u_p \in \partial_p L(x,p),\ \|u_p\|_P \leq a_1 \|p\|_P + a_2, \qquad (27b)$$

*(iv) Assume that there exist $c, \kappa > 0$ such that for all $t \in \mathbb{N}$,*

$$\begin{aligned} \langle s_t, x_t - x^* \rangle_X \leq & \kappa \left( L(x^*, p_t) - L(x_t, p_t) \right), & \text{(28a)} \\ \langle r_t, p_t - p^* \rangle_P \leq & \kappa \left( L(x_t, p_t) - L(x_t, p^*) \right), & \text{(28b)} \\ \exists u_t^x \in \partial_x L(x_t, p_t),\ \|s_t\|_X \leq & c \left( 1 + \|u_t^x\| \right), & \text{(28c)} \\ \exists u_t^p \in \partial_p L(x_t, p_t),\ \|r_t\|_P \leq & c \left( 1 + \|u_t^p\| \right). & \text{(28d)} \end{aligned}$$

*(v) Assume that there are $b_x, b_p \geq 0$, $A > 0$ and nonnegative sequences $(\epsilon_t^x, \eta_t^x)$ and $(\epsilon_t^p, \eta_t^p)$ such that for all $t \in \mathbb{N}$, there exist $(u_t^x, u_t^p) \in \partial_x L(x_t, p_t) \times \partial_p L(x_t, p_t)$ and it holds*

$$\begin{aligned} \|\mathbb{E}\left( w_t | \mathcal{F}_t \right)\|_X \leq & b_x \eta_t^x \left( 1 + \|u_t^x\|_X \right), & \text{(29a)} \\ \|\mathbb{E}\left( v_t | \mathcal{F}_t \right)\|_P \leq & b_p \eta_t^p \left( 1 + \|u_t^p\|_P \right), & \text{(29b)} \\ \mathbb{E}\left( \|w_t\|_X^2 | \mathcal{F}_t \right) \leq & A \left( 1 + \frac{1}{\epsilon_t^x} \|u_t^x\|_X^2 \right), & \text{(29c)} \\ \mathbb{E}\left( \|v_t\|_P^2 | \mathcal{F}_t \right) \leq & A \left( 1 + \frac{1}{\epsilon_t^p} \|u_t^p\|_P^2 \right). & \text{(29d)} \end{aligned}$$

*If $X^f$ (resp. $P^f$) is a closed convex subset and not a subspace, assume also that there exist a bounded mapping $g_x : \mathbb{R} \to \mathbb{R}$ (resp. $g_t$), and for all $t \in \mathbb{N}$ some $u_t^x \in \partial_x L(x_t, p_t)$ (resp. $u_t^p \in \partial_p L(x_t, p_t)$) such that,*

$$\mathbb{E}\left( \|w_t\|_X | \mathcal{F}_t \right) \leq g_x(\|u_t^x\|_X),\ (\textit{resp. } \mathbb{E}\left( \|v_t\|_P | \mathcal{F}_t \right) \leq g_p(\|u_t^p\|_P)). \tag{29e}$$

*(vi) Assume that the sequences $(\gamma_t)$, $(\epsilon_x^t)$, $(\epsilon_t^p)$, $(\eta_t^x)$ and $(\eta_t^p)$ are all strictly nonnegative and verify:*

$$\sum_{t \in \mathbb{N}} \gamma_t = +\infty,\ \sum_{t \in \mathbb{N}} (\gamma_t)^2 < +\infty,\ \sum_{t \in \mathbb{N}} b_x \gamma_t \eta_t^x < +\infty,\ \sum_{t \in \mathbb{N}} b_p \gamma_t \eta_t^p < +\infty,\ \sum_{t \in \mathbb{N}} \frac{(\gamma_t)^2}{\epsilon_t^x} < +\infty,\ \sum_{t \in \mathbb{N}} \frac{(\gamma_t)^2}{\epsilon_t^p} < +\infty. \tag{30a}$$

*Then, $(x_t)$ and $(p_t)$ are a.s. bounded, and almost surely, $L(x_t, p^*) \to L(x^*, p^*)$, and $L(x^*, p_t) \to L(x^*, p^*)$ as $t$ goes to infinity. Moreover, if $L(\cdot, p^*)$ is strongly convex, $(x_t)$ strongly converges almost surely to $x^*$.*

Proof. We follow the same scheme as in the proof of Theorem 3.1. The proof may therefore seem routine, but it is necessary to write it because of the interaction between the iterates $x_t$ and $p_t$, given by assumptions (28a)–(28b).

Let us define for all $t \in \mathbb{N}$, $\Lambda_t$ our Lyapunov function to be:

$$\Lambda_t = \|x_t - x^*\|_X^2 + \|p_t - p^*\|_P^2.$$

Using the same calculations as those leading to (8), we obtain in any case:

$$\begin{aligned} \Lambda_{t+1} \leq & \Lambda_t + (\gamma_t)^2 \left( \|s_t + w_t\|_X^2 + \|r_t + v_t\|_P^2 \right) \\ & + 2\gamma_t \left( \langle s_t + w_t, x_t - x^* \rangle_X + \langle r_t + v_t, p_t - p^* \rangle_P \right). \end{aligned} \tag{31}$$

With the classical scalar inequality $(a+b)^2 \leq 2(a^2 + b^2)$, and by assumptions (28a)–(28b), we get from (31):

$$\begin{aligned} \Lambda_{t+1} \leq & \Lambda_t + 2(\gamma_t)^2 \left( \|s_t\|_X^2 + \|w_t\|_X^2 \right) \\ & + 2(\gamma_t)^2 \left( \|r_t\|_P^2 + \|v_t\|_P^2 \right) \\ & + 2\gamma_t \kappa \left( L(x^*, p_t) - L(x_t, p_t) + L(x_t, p_t) - L(x_t, p^*) \right) \\ & + 2\gamma_t \left( \langle w_t, x_t - x^* \rangle_X + \langle v_t, p_t - p^* \rangle_P \right). \end{aligned} \tag{32}$$

Moreover, by assumptions (28c)–(28d), we get:

$$\begin{aligned} \|s_t\|_X^2 \leq & 2c^2 \left( 1 + \|u_t^x\|_X^2 \right), \\ \|r_t\|_P^2 \leq & 2c^2 \left( 1 + \|u_t^p\|_P^2 \right). \end{aligned}$$

Using assumption (27) one hence obtains:

$$\begin{aligned} \|s_t\|_X^2 \leq & 2c^2 \left( 1 + 2(a_1)^2 \|x_t - x^*\|_X^2 + 2(a_2 + a_1 \|x^*\|_X)^2 \right), \\ \|s_t\|_X^2 \leq & 2c^2 \left( 1 + 2(a_1)^2 \|p_t - p^*\|_P^2 + 2(a_2 + a_1 \|p^*\|_P)^2 \right). \end{aligned}$$

Finally, define $a_3 = 4c^2(a_1)^2$ and $a_4^x = 2c^2\left(1+2(a_2+a_1\|x^*\|_X)^2\right)$ and analogously for $a_4^p$, and we obtain

$$\|s_t\|_X^2 \leq a_3\|x_t - x^*\|_X^2 + a_4^x, \tag{33a}$$
$$\|s_t\|_X^2 \leq a_3\|p_t - p^*\|_P^2 + a_4^p. \tag{33b}$$

Similarly, assumptions (27) and (29c)–(29d) read

$$\mathbb{E}\left(\|w_t\|_X^2|\mathcal{F}_t\right) \leq A\left(1+\frac{2}{\epsilon_t^x}\left((a_1)^2\|x_t-x^*\|_X^2+(a_2+a_1\|x^*\|_X)^2\right)\right) \tag{34a}$$
$$\mathbb{E}\left(\|v_t\|_P^2|\mathcal{F}_t\right) \leq A\left(1+\frac{2}{\epsilon_t^p}\left((a_1)^2\|p_t-p^*\|_P^2+(a_2+a_1\|p^*\|_P)^2\right)\right) \tag{34b}$$

We now take the conditional expectation with respect to $\mathcal{F}_t$ in (32), and apply inequalities (33)–(34). It yields,

$$\begin{aligned}\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right) \leq& \Lambda_t + 2(\gamma_t)^2\left(a_3\|x_t-x^*\|_X^2 + a_4^x + a_3\|p_t-p^*\|_P^2 + a_4^p\right)\\ &+2(\gamma_t)^2\left(A\left(1+\frac{2}{\epsilon_t^x}\left((a_1)^2\|x_t-x^*\|_X^2+(a_2+a_1\|x^*\|_X)^2\right)\right)\right)\\ &+2(\gamma_t)^2\left(A\left(1+\frac{2}{\epsilon_t^p}\left((a_1)^2\|p_t-p^*\|_P^2+(a_2+a_1\|p^*\|_P)^2\right)\right)\right)\\ &+2\gamma_t\kappa\left(L(x^*,p_t)-L(x_t,p_t)+L(x_t,p_t)-L(x_t,p^*)\right)\\ &+2\gamma_t\left(\|\mathbb{E}\left(w_t|\mathcal{F}_t\right)\|\times\|x_t-x^*\|_X+\|\mathbb{E}\left(v_t|\mathcal{F}_t\right)\|\times\|p_t-p^*\|_P\right)\end{aligned} \tag{35}$$

Assumptions (29a)–(29b) provide bounds for the last summands of (35), and we finally obtain

$$\begin{aligned}\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right) \leq& \Lambda_t + 2(\gamma_t)^2\left(a_3\|x_t-x^*\|_X^2 + a_4^x + a_3\|p_t-p^*\|_P^2 + a_4^p\right)\\ &+2(\gamma_t)^2\left(A\left(1+\frac{2}{\epsilon_t^x}\left((a_1)^2\|x_t-x^*\|_X^2+(a_2+a_1\|x^*\|_X)^2\right)\right)\right)\\ &+2(\gamma_t)^2\left(A\left(1+\frac{2}{\epsilon_t^p}\left((a_1)^2\|p_t-p^*\|_P^2+(a_2+a_1\|p^*\|_P)^2\right)\right)\right)\\ &+2\gamma_t\kappa\left(L(x^*,p_t)-L(x_t,p_t)+L(x_t,p_t)-L(x_t,p^*)\right)\\ &+2b_x\eta_t^x\gamma_t\left(a_1\|x_t-x^*\|_X+a_2+a_1\|x^*\|_X\right)\|x_t-x^*\|_X\\ &+2b_p\eta_t^p\gamma_t\left(a_1\|p_t-p^*\|_P+a_2+a_1\|p^*\|_P\right)\|p_t-p^*\|_P\end{aligned} \tag{36}$$

Moreover, the following classical scalar inequality holds: $ab \leq \frac{a^2+b^2}{2}$. Hence, (36) reads:

$$\begin{aligned}\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right) \leq& \Lambda_t + \beta_t + \alpha_t\left(\|x_t-x^*\|_X^2+\|p_t-p^*\|_P^2\right) + 2\gamma_t\kappa\left(L(x^*,p_t)-L(x_t,p^*)\right),\\ \leq& \Lambda_t\left(1+\alpha_t\right)+\beta_t+2\gamma_t\kappa\left(L(x^*,p_t)-L(x_t,p^*)\right),\end{aligned} \tag{37}$$

with $(\alpha_t)$ and $(\beta_t)$ two summable sequences defined in the same way as in the Proof of Theorem 3.1. Using the saddle point assumption in $(x^*,p^*)$, one get with (37):

$$\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right) \leq \Lambda_t\left(1+\alpha_t\right)+\beta_t+2\gamma_t\kappa\left(L(x^*,p^*)-L(x_t,p^*)\right) \text{ and,} \tag{38a}$$
$$\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right) \leq \Lambda_t\left(1+\alpha_t\right)+\beta_t+2\gamma_t\kappa\left(L(x^*,p_t)-L(x^*,p^*)\right). \tag{38b}$$

Moreover, it is also clear by the saddle point assumption that:

$$L(x^*,p_t)-L(x^*,p^*) \leq 0, \text{ and, } L(x^*,p^*)-L(x_t,p^*) \leq 0.$$

At this point, using the same quasimartingale arguments as before, we get that the sequence $(\Lambda_t)$ is a quasimartingale and converges a.s. to some integrable random variable. Hence, it is a.s. bounded and hence, $(x_t)$ and $(p_t)$ are a.s. bounded in $X$ and $P$ respectively. Using assumptions (27),(28), $(s_t)$ and $(r_t)$ are also a.s. bounded.

Moreover, by making the same calculations as those leading to (15), we obtain:

$$\sum_{t\in\mathbb{N}}\gamma_t\left(L(x_t,p^*)-L(x^*,p^*)\right) < +\infty, \tag{39a}$$
$$\sum_{t\in\mathbb{N}}\gamma_t\left(L(x^*,p^*)-L(x^*,p_t)\right) < +\infty. \tag{39b}$$

By convexity of $L(\cdot, p^*)$ and concavity of $L(x^*, \cdot)$, we make the same calculations as in (16)–(17), which are still valid by the boundedness of the sequences and the assumptions of the theorem, and finally get by Lemma 2.2:

$$\lim_{t\to\infty} L(x_t, p^*) = L(x^*, p^*) \quad \text{almost surely, and} \tag{40a}$$

$$\lim_{t\to\infty} L(x^*, p_t) = L(x^*, p^*) \quad \text{almost surely.} \tag{40b}$$

Lower semicontinuity of $L(\cdot, p^*)$ and upper semicontinuity of $L(x^*, \cdot)$ yield the weak convergence of $(x_t, p_t)$ to $(x^*, p^*)$ in the closed convex case.
Finally, if $L(\cdot, p^*)$ is strongly convex, by the same equation as (22), we obtain that $(x_t)$ strongly converges to $x^*$. □

## 5. Conclusions

We proposed here a general framework for the convergence analysis of Hilbert-valued perturbed subgradient algorithms. We proved the convergence of such schemes under convexity and subdifferentiability assumptions on the cost function. The perturbations of the subgradients were only required to be asymptotically martingale increments instead of being so all along the iterations. Furthermore, we allowed projections at each iteration on the feasible set which could be either a closed convex subset or a closed vector subspace of the Hilbert space.
We then extended this framework to the solution of saddle-point problems, and proved the convergence of perturbed Arrow-Hurwicz type subgradient algorithms.
Solving stochastic optimization problems with measurability constraints is a natural use of our infinite dimensional stochastic approximation scheme. In this case, the projection on the measurable functions may be done directly in the subgradient perturbation (under, e.g., a mollifying kernel applied to the subgradient mapping), making the algorithm easily implementable.

**Acknowledgments.** We thank P. Carpentier, G. Cohen, M. Minoux and T. Pennanen for useful discussions and comments on an earlier version of this paper. We also thank three anonymous referees for their useful comments on this paper, including references to related earlier works.

## References

[AIS98] Y. Alber, A. Iusem, and M. Solodov. On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. *Mathematical Programming*, 81:23–35, 1998.

[BRS05a] K. Barty, J.-S. Roy, and C. Strugarek. A stochastic gradient type algorithm for closed loop problems. *submitted*, 2005. http://hera.rz.hu-berlin.de/speps/artikel/ClosedLoopSGV2.pdf.

[BRS05b] K. Barty, J.-S. Roy, and C. Strugarek. Temporal difference learning with kernels for pricing american-style options. *submitted*, 2005. http://www.optimization-online.org/DB_HTML/2005/05/1133.html.

[BMP90] A. Benvéniste, M. Métivier, and P. Priouret. *Adaptive Algorithms and stochastic approximation.* Springer Verlag, New York, 1990.

[BL72] H. Berliocchi and J.-M. Lasry. Nouvelles applications des mesures paramétrées. *C. R. Acad. Sci., Paris*, 274:1623–1626, 1972.

[BT00] D.P. Bertsekas and J.N. Tsitsiklis. Gradient convergence in gradient methods. *SIAM J. Optim.*, 10(3):627–642, 2000.

[CC90] G. Cohen and J.-C. Culioli. Decomposition Coordination Algorithms for Stochastic Optimization. *SIAM J. Control Optimization*, 28(6):1372–1403, 1990.

[CK78] D.S. Clark and H.J. Kushner. *Stochastic Approximation for constrained and unconstrained systems.* Springer Verlag, New York, 1978.

[Coh84] G. Cohen. *Décomposition et Coordination en optimisation déterministe différentiable et non-différentiable.* Thèse de doctorat d'État, Université de Paris IX Dauphine, 1984.

[CW98] X. Chen and H. White. Laws of large numbers for Hilbert space-valued mixingales with applications. *Econometric Theory*, 12:284–304, 1998.

[CW02] X. Chen and H. White. Asymptotic properties of some projection-based Robbins-Monro procedures in a Hilbert space. *Stud. Nonlinear Dyn. Econom.*, 6:1–53, 2002.

[Del96] B. Delyon. General results on the convergence of stochastic algorithms. *IEEE Trans. Autom. Control*, 41(9):1245–1255, 1996.

[DF74] D. Derevitskii and A. Fradkov. Two models for analyzing the dynamics of adaptation algorithms. *Autom. Remote Control*, 35:59–67, 1974.

[Duf97] M. Duflo. *Random Iterative Models.* Springer Verlag, Berlin, 1997.

[Erm66] Y. Ermoliev. Methods of solution of nonlinear extremal problems. *Cybernetics*, 2(4):1–17, 1966.

[Erm66] Y. Ermoliev. On the method of generalized stochastic gradients and quasi-Féjer sequences. *Cybernetics*, 5(2):73–84, 1969.

[Erm76] Y. Ermoliev. *Methods of Stochastic Programming.* (In Russian) Nauka, Moscow, 1976.

[Gol88] L. Goldstein. Minimizing noisy functionals in Hilbert spaces : an extension of the Kiefer-Wolfowitz procedure. *J. Theor. Probab.*, 1:189–204, 1988.

[HK96] C. Horn and S.R. Kulkarni. An alternative proof for convergence of stochastic approximation algorithms. *IEEE Trans. Autom. Control*, 41(3):419–424, 1996.

[HU75] J.-B. Hiriart-Urruty. Algorithmes de résolution d'équations et d'inéquations variationnelles. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 33:167–186, 1975.

[KW52] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.

[Lai03] T.L. Lai. Stochastic Approximation. *Ann. Stat.*, 31(2):391–406, 2003.

[Mé82] M. Métivier. *Semimartingales.* De Gruyter, Berlin, 1982.

[NH73] M.B. Nevel'son and R.Z. Has'minskii. *Stochastic Approximation and recursive estimation.* American Mathematical Society, Providence, RI, 1973.

[PT73] B.T. Polyak and Y.Z. Tsypkin. Pseudogradient adaptation and training algorithms. *Autom. Remote Control*, 12:83–94, 1973.

[Ré73a] P. Révész, Robbins-Monro procedure in a Hilbert space and its application in the theory of learning processes, I. *Studia Sci. Math.Hungar.* 8, 391–398, 1973.

[Ré73b] P. Révész, Robbins-Monro procedure in a Hilbert space, II. *Studia Sci. Math. Hungar.* 8, 469–472, 1973.

[RM51] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[RS71] H. Robbins and D. Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In J.S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, New York, 1971.

[Sa80] G. Salov. On a stochastic approximation theorem in a Hilbert space and its applications. *Theory Probab. Appl.*, 24:413–419,1980.

[YZ90] G. Yin and Y.M. Zhu. On H-valued Robbins-Monro processes. *J. Multivariate Anal.*, 34:116–140, 1990.