A Perturbed Gradient Algorithm in Hilbert spaces

Kengy Barty*, Jean-Sébastien Roy†, Cyrille Strugarek‡

May 19, 2005

Abstract

We propose a perturbed gradient algorithm with stochastic noises to solve a general class of optimization problems. We 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 gradient stepsizes.

We then 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 developped in [Barty et al., 2005] is a special case of the following perturbed gradient algorithm with stochastic noises.

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

Keywords: Stochastic Quasi-Gradient, Perturbed Gradient, Infinite Dimensional Problems, Saddle Point Problems

1. Introduction

The field of stochastic approximation theory actually began with the seminal paper of Robbins and Monro ([Robbins and Monro, 1951]). 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., [Lai, 2003] for an historical survey of stochastic approximation, or [Duflo, 1997] for the many branches of this field, or [Nevel'son and Has'minskii, 1973] 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 developped, with for example [Goldstein, 1988]. An important progress in this area is the paper [Yin and Zhu, 1990], showing the convergence and giving asymptotic properties of an Hilbert-valued Robbins-Monro algorithm under a few assumptions. The important work presented in [Hiriart-Urruty, 1975] studies the role of stochastic approximation to solve general Hilbert-valued variational equations, using both probabilistic and variational arguments.

Our work aims to bring some other assumptions to ensure the convergence of stochastic approximation procedure in the general framework of infinite dimensional Hilbert spaces. It has been motivated both by the paper [Bertsekas and Tsitsiklis, 2000],

* **École Nationale des Ponts et Chaussées** (ENPC),
kengy.barty@cermics.enpc.fr

† **EDF R&D**
1, avenue du Général de Gaulle
F-92141 Clamart Cedex
jean-sebastien.roy@edf.fr

‡ **EDF R&D**
cyrille.strugarek@edf.fr
also with the École Nationale Supérieure de Techniques Avancées (ENSTA) and the École Nationale des Ponts et Chaussées (ENPC)

and by the practical need to propose efficient ways to solve infinite dimensional stochastic optimization problems.
Convergence proofs of stochastic approximation algorithms exist in various point of views. Historically, almost surely convergence proofs were given through the so called Robbins-Siegmund Lemma (see [Robbins and Siegmund, 1971]), and have been then developped by e.g. [Benvéniste et al., 1990], or [Polyak and Tsypkin, 1973]. Other approaches have been developped successfully: in the well known monograph [Clark and Kushner, 1978], stability analysis is developped and a new method based on the analysis of the underlying ordinary differential equation is provided. This method has e.g. been used in [Yin and Zhu, 1990] to derive their infinite dimensional convergence results. Following the same direction, thanks to the general results on Hilbert-valued mixingales (see [Chen and White, 1998]), the recent paper [Chen and White, 2002] provides a good understanding of infinite dimensional Robbins-Monro type procedures. They use modified stochastic approximation with boundedness properties to derive almost surely convergence results and asymptotic normality. Starting from the same ideas, we can also mention [Horn and Kulkarni, 1996] or [Delyon, 1996], founded on deterministic arguments, but limited to the finite dimensional case. Among those approaches, we will follow in this paper an approach more based on probabilistic martingale or quasimartingale arguments (see [Métivier, 1982]).
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 [Hiriart-Urruty, 1975] is the nearest of ours, by the techniques used in the proofs and the problems it adresses, but the results are a bit 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 only asymptotically. The assumptions made in [Hiriart-Urruty, 1975], 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.
The results of [Chen and White, 2002] differ from ours in that: they are not robust to any projection, except the projection on a 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. Our paper is built as follows: Section 2 adresses minimization problems. We provide in subsection 2.2 two convergence proofs with general assumptions. In subsection 2.4, we place our result in the context of stochastic approximation, and we especially compare it with the result of [Bertsekas and Tsitsiklis, 2000]. In subsection 2.5, we show how our results can be used to prove the convergence of a new algorithm introduced in a forthcoming paper ([Barty et al., 2005]), to solve infinite dimensional stochastic optimization problems. In section 3, we propose a perturbed gradient algorithm to solve general saddle point problems, and provide a convergence proof. In section 4, we provide two technical lemmas used in the convergence proofs.

2. Minimization Problems

2.1. **Algorithm.** We focus on the problem:

$$\begin{aligned} &\min_x f(x) \\ &\text{s.t. } x \in X^f. \end{aligned} \tag{1}$$

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 either a closed subspace of X, or a closed convex subset of X, and Π_{X^f} denotes the projection onto X^f.

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

Algorithm 2.1. Step $t \in \mathbb{N}$:

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

where s_t is the descent direction, w_t is the random noise (the perturbation), and γ_t is a nonnegative decreasing stepsize.

Remark 2.2 (Random noise)**.** More precisely, w_t can be seen as a random variable on some measurable space $(\Omega, \mathcal{A}, \mathbb{P})$ for all $t \in \mathbb{N}$. When we will talk in the following of almost sure convergence, it is in the sense of the product probability space $(\times_{t \in \mathbb{N}} \Omega, \otimes_{t \in \mathbb{N}} \mathbb{P})$, i.e., the sampling space.

2.2. **Convergence Proofs.**

Definition 2.3 (Coercivity)**.** A mapping $g : X \to \mathbb{R}$ is said to be *coercive* if and only if

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

We provide two convergence proofs corresponding to the two possible cases of feasible set. Theorem 2.4 focuses on the case where X^f is a closed convex subset of X, while Theorem 2.5 provides a result when X^f is a closed subspace. The calculations in the two proofs are quite similar. In Theorem 2.4, we follow the classical Robbins-Siegmund's pattern for stochastic algorithms (see [Robbins and Siegmund, 1971]), and in Theorem 2.5, we provide a more constructive proof based on the work of [Cohen and Culioli, 1990].

Theorem 2.4 (Projection on a closed convex subset)**.** *(i) Assume that $x \mapsto f(x)$ is strongly convex with modulus $B > 0$, and lower semicontinuous. Hence, it is Gâteaux-differentiable. Assume moreover that X^f is a closed convex subset of X. Then,* (1) *has a unique solution denoted by x^*.*
(ii) Let $(\mathcal{F}_t)$ be a filtration, and assume that for all $t \in \mathbb{N}$, x_t and s_t are $\mathcal{F}_t$-measurable.
(iii) Assume that $\nabla f(\cdot)$ is Lipschitz continuous with modulus L.
(iv) Assume that there exist $c, \kappa > 0$, such that for all $t \in \mathbb{N}$,

$$\langle s_t, x_t - x^* \rangle \leq \kappa \left(f(x^*) - f(x_t) \right), \tag{2a}$$

$$\|s_t\| \leq c \left(1 + \|\nabla f(x_t)\| \right). \tag{2b}$$

(v) Assume that there are $b \geq 0$, $A > 0$ and a nonnegative sequence (ϵ_t), such that for all $t \in \mathbb{N}$,

$$\|\mathbb{E}\left(w_t | \mathcal{F}_t \right)\| \leq b \epsilon_t \left(1 + \|\nabla f(x_t)\| \right), \tag{3a}$$

$$\mathbb{E}\left(\|w_t\|^2 | \mathcal{F}_t \right) \leq A \left(1 + \frac{1}{\epsilon_t} \|\nabla f(x_t)\|^2 \right). \tag{3b}$$

(vi) Assume that the sequences (γ_t) and (ϵ_t) are such that:

$$\forall t \in \mathbb{N},\ \gamma_t, \epsilon_t > 0, \quad \sum_{t \in \mathbb{N}} \gamma_t = +\infty, \quad \sum_{t \in \mathbb{N}} \frac{(\gamma_t)^2}{\epsilon_t} < +\infty, \quad \sum_{t \in \mathbb{N}} b \gamma_t \epsilon_t < +\infty. \tag{4}$$

Then $f(x_t) \to f(x^*)$ a.s., as t goes to infinity, and (x_t) strongly converges to the unique solution x^* of (1).

Proof : We follow Robbins-Siegmund's scheme (see [Robbins and Siegmund, 1971]). Let $t \in \mathbb{N}$. It holds:

$$\begin{aligned}\|x_{t+1} - x^*\|^2 =& \|\Pi_{X^f}\left(x_t + \gamma_t(s_t + w_t)\right) - x^*\|^2\\ \leq& \|x_t - x^* + \gamma_t(s_t + w_t)\|^2, \text{ by the nonexpansiveness of } \Pi\\ \leq& \|x_t - x^*\|^2 + (\gamma_t)^2\|s_t + w_t\|^2 + 2\gamma_t\langle s_t + w_t, x_t - x^*\rangle\\ \leq& \|x_t - x^*\|^2 + 2(\gamma_t)^2\|s_t\|^2 + 2(\gamma_t)^2\|w_t\|^2\\ &+ 2\gamma_t\langle w_t, x_t - x^*\rangle + 2\gamma_t\kappa\left(f(x^*) - f(x_t)\right)\end{aligned} \tag{5}$$

by assumption (2a), and using the classical inequality $(x+y)^2 \leq 2x^2 + 2y^2$. Using once again this inequality, together with assumption (2b) yields:

$$\begin{aligned}\|x_{t+1} - x^*\|^2 \leq& \|x_t - x^*\|^2 + 4c^2(\gamma_t)^2\left(1 + \|\nabla f(x_t)\|^2\right)\\ &+ 2(\gamma_t)^2\|w_t\|^2 + 2\gamma_t\langle w_t, x_t - x^*\rangle + 2\gamma_t\kappa\left(f(x^*) - f(x_t)\right)\end{aligned} \tag{6}$$

The strong convexity of f reads:

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

By optimality, one has $\langle\nabla f(x^*), x^* - x_t\rangle \leq 0$. Thus,

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

Let us now take the conditional expectation with respect to $\mathcal{F}_t$ in (6), and use the last equation about strong convexity. One obtains:

$$\begin{aligned}\mathbb{E}\left(\|x_{t+1} - x^*\|^2|\mathcal{F}_t\right) \leq& \|x_t - x^*\|^2 + 4c^2(\gamma_t)^2\left(1 + \|\nabla f(x_t)\|^2\right)\\ &+ 2(\gamma_t)^2\mathbb{E}\left(\|w_t\|^2|\mathcal{F}_t\right) + 2\gamma_t\|\mathbb{E}\left(w_t|\mathcal{F}_t\right)\|\|x_t - x^*\|\\ &- B\kappa\gamma_t\|x_t - x^*\|^2,\\ \leq& \|x_t - x^*\|^2 + 4c^2(\gamma_t)^2 + 8c^2(\gamma_t)^2\|\nabla f(x^*)\|^2\\ &+ 8c^2(\gamma_t)^2\|\nabla f(x_t) - \nabla f(x^*)\|^2 + 2A(\gamma_t)^2\left(1 + \frac{1}{\epsilon_t}\|\nabla f(x_t)\|^2\right)\\ &+ 2b\gamma_t\epsilon_t\left(1 + \|\nabla f(x_t)\|\right)\|x_t - x^*\| - B\kappa\gamma_t\|x_t - x^*\|^2\end{aligned} \tag{7}$$

by assumptions (3a)–(3b). We will now use the inequality $x \leq 1 + x^2$ and the Lipschitz property of $\nabla f(\cdot)$. Thus,

$$\begin{aligned}\mathbb{E}\left(\|x_{t+1} - x^*\|^2|\mathcal{F}_t\right) \leq& \left(1 + 8c^2(\gamma_t)^2L^2 + 4A\frac{(\gamma_t)^2}{\epsilon_t}L^2 + 2b\gamma_t\epsilon_t(1 + L + \|\nabla f(x^*)\|\right)\|x_t - x^*\|^2\\ &+ 2(\gamma_t)^2A + 4c^2(\gamma_t)^2\left(1 + \|\nabla f(x^*)\|^2\right) + 4A\frac{(\gamma_t)^2}{\epsilon_t}\|\nabla f(x^*)\|^2\\ &+ 2b\gamma_t\epsilon_t(1 + \|\nabla f(x^*)\|) - B\kappa\gamma_t\|x_t - x^*\|^2\\ =& (1 + \alpha_t)\|x_t - x^*\|^2 + \beta_t - \delta_t,\end{aligned} \tag{8}$$

with α_t and β_t the terms of summable sequences by assumption (4), and

$$\delta_t = B\kappa\gamma_t\|x_t - x^*\|^2 \geq 0.$$

We can hence apply Robbins-Siegmund's Lemma ([Robbins and Siegmund, 1971]), which yields:

$$\begin{aligned}&\|x_t - x^*\|^2 \text{ converges a.s. when } t \to \infty, \text{ and,}\\ &\sum_{t\in\mathbb{N}}\gamma_t\|x_t - x^*\|^2 < +\infty.\end{aligned}$$

Hence, $(\|x_t - x^*\|^2)$ converges to 0 when t goes to infinity, and it completes the proof. □

Theorem 2.5 (Projection on a subspace)**.** *(i) Assume that $x \mapsto f(x)$ is convex and lower semicontinuous. Hence, it is Gâteaux-differentiable. If moreover f is coercive on X, and if X^f is a closed subspace of X, then (1) has a solution set denoted by X^*.*
(ii) Let $(\mathcal{F}_t)$ be a filtration, and assume that for all $t \in \mathbb{N}$, x_t and s_t are $\mathcal{F}_t$-measurable.
(iii) Assume that $\nabla f(\cdot)$ is Lipschitz continuous with modulus L.
(iv) Assume that there exist $\kappa, c > 0$ such that for all $x^* \in X^*$, for all $t \in \mathbb{N}$,

$$\langle s_t, x_t - x^* \rangle \leq \kappa \left(f(x^*) - f(x_t) \right), \tag{9a}$$

$$\|s_t\| \leq c \left(1 + \|\nabla f(x_t)\| \right). \tag{9b}$$

(v) Assume that there are $b \geq 0$, $A > 0$ and a nonnegative sequence (ϵ_t), such that for all $t \in \mathbb{N}$,

$$\|\mathbb{E}\left(w_t | \mathcal{F}_t \right)\| \leq b\epsilon_t \left(1 + \|\nabla f(x_t)\| \right), \tag{10a}$$

$$\mathbb{E}\left(\|w_t\|^2 | \mathcal{F}_t \right) \leq A \left(1 + \frac{1}{\epsilon_t} \|\nabla f(x_t)\|^2 \right). \tag{10b}$$

(vi) Assume that the sequences (γ_t) and (ϵ_t) are such that:

$$\forall t \in \mathbb{N},\ \gamma_t, \epsilon_t > 0, \quad \sum_{t \in \mathbb{N}} \gamma_t = +\infty, \quad \sum_{t \in \mathbb{N}} \frac{(\gamma_t)^2}{\epsilon_t} < +\infty, \quad \sum_{t \in \mathbb{N}} b\gamma_t \epsilon_t < +\infty. \tag{11}$$

Then $f(x_t) \to f(x^*)$ a.s., as t goes to infinity, and every cluster point of (x_t) belongs to X^*.
(vii) Moreover, if f is strongly convex, (x_t) strongly converges to the unique solution x^* of (1).

Proof : We use the scheme introduced by [Cohen and Culioli, 1990], using a Lyapunov function. 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 $t \in \mathbb{N}$.

$$\begin{aligned}
\Lambda_{t+1} - \Lambda_t =& \frac{1}{2}\|x_{t+1} - x_t\|^2 + \langle x_{t+1} - x_t, x_t - x^* \rangle \\
\leq& \frac{(\gamma_t)^2}{2} \|s_t + w_t\|^2 + \gamma_t \langle s_t + w_t, x_t - x^* \rangle
\end{aligned} \tag{12}$$

These inequalities are obtained using the nonexpansiveness and the linearity of the projection over a subspace. We take now the conditional expectation with respect to $\mathcal{F}_t$ in (12)

$$\begin{aligned}
\mathbb{E}\left(\Lambda_{t+1} | \mathcal{F}_t \right) - \Lambda_t \leq& \frac{(\gamma_t)^2}{2} \mathbb{E}\left(\|s_t + w_t\|^2 | \mathcal{F}_t \right) + \gamma_t \langle s_t, x_t - x^* \rangle \\
&+ \gamma_t \langle \mathbb{E}\left(w_t | \mathcal{F}_t \right), x_t - x^* \rangle, \\
\leq& \frac{(\gamma_t)^2}{2} \mathbb{E}\left(\|s_t + w_t\|^2 | \mathcal{F}_t \right) + \gamma_t \kappa \left(f(x^*) - f(x_t) \right) \\
&+ b\epsilon_t \gamma_t \|x_t - x^*\| + b\epsilon_t \gamma_t \|\nabla f(x_t)\| \|x_t - x^*\|, \\
\leq& \frac{(\gamma_t)^2}{2} \mathbb{E}\left(\|s_t + w_t\|^2 | \mathcal{F}_t \right) + \gamma_t \kappa \left(f(x^*) - f(x_t) \right) \\
&+ b\epsilon_t \gamma_t \|x_t - x^*\| + b\epsilon_t \gamma_t \|\nabla f(x_t) - \nabla f(x^*)\| \|x_t - x^*\| \\
&+ b\epsilon_t \gamma_t \|\nabla f(x^*)\| \|x_t - x^*\|, \\
\leq& \frac{(\gamma_t)^2}{2} \mathbb{E}\left(\|s_t + w_t\|^2 | \mathcal{F}_t \right) + \gamma_t \kappa \left(f(x^*) - f(x_t) \right) \\
&+ b\epsilon_t \gamma_t (1 + L + \|\nabla f(x^*)\|) \|x_t - x^*\|^2 + b\epsilon_t \gamma_t \left(1 + \|\nabla f(x^*)\| \right).
\end{aligned} \tag{13}$$

Inequality (13) is obtained by using assumptions (10a)–(9a), the classical inequality $x \leq 1 + x^2$, the Lipschitz property of ∇f, and Cauchy-Schwarz's inequality. We now focus on the first term of the right hand side:

$$\begin{aligned}\frac{(\gamma_t)^2}{2}\mathbb{E}\left(\|s_t+w_t\|^2|\mathcal{F}_t\right) \leq&(\gamma_t)^2\left(\|s_t\|^2+\mathbb{E}\left(\|w_t\|^2|\mathcal{F}_t\right)\right),\\ \leq&(\gamma_t)^2\left(2c^2(1+\|\nabla f(x_t)\|^2)+A\left(1+\frac{1}{\epsilon_t}\|\nabla f(x_t)\|^2\right)\right),\\ \leq&(\gamma_t)^2C+2(\gamma_t)^2\left(2c^2+\frac{A}{\epsilon_t}\right)\left(\|\nabla f(x_t)-\nabla f(x^*)\|^2+\|\nabla f(x^*)\|^2\right),\\ \leq&(\gamma_t)^2\left(C'+\frac{C''}{\epsilon_t}\right)+2(\gamma_t)^2L^2\left(2c^2+\frac{A}{\epsilon_t}\right)\|x_t-x^*\|^2, \qquad (14)\end{aligned}$$

with C' and C'' two positive deterministic scalars. (14) is obtained by using the classical inequality $(x+y)^2 \leq 2x^2+2y^2$, assumptions (9b)–(10b), and the Lipschitz property of ∇f. We now go back to equation (13), and we obtain:

$$\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right)-\Lambda_t \leq \alpha_t\Lambda_t+\beta_t+\gamma_t\kappa\left(f(x^*)-f(x_t)\right) \qquad (15)$$

with:

$$\begin{aligned}\alpha_t =&4(\gamma_t)^2L^2\left(2c^2+\frac{A}{\epsilon_t}\right)+2b\epsilon_t\gamma_t\left(1+L+\|\nabla f(x^*)\|\right),\\ \beta_t =&(\gamma_t)^2\left(C'+\frac{C''}{\epsilon_t}\right)+b\epsilon_t\gamma_t\left(1+\|\nabla f(x^*)\|\right).\end{aligned}$$

Thus, (α_t) and (β_t) form two summable sequences. Let us take the expectation in (15), and denote $\lambda_t=\mathbb{E}\left(\Lambda_t\right)$. It yields:

$$\lambda_{t+1}-\lambda_t \leq \alpha_t\lambda_t+\beta_t+\gamma_t\kappa\underbrace{\mathbb{E}\left(f(x^*)-f(x_t)\right)}_{\leq 0,\ \text{by optimality}}. \qquad (16)$$

Using Lemma 4.1 (see the following subsection), it shows that (λ_t) is bounded, by, say, some $M>0$. We now prove that (Λ_t) is a convergent quasimartingale. Indeed:

- By definition, (Λ_t) is $(\mathcal{F}_t)$ adapted.
- By definition, for all $t\in\mathbb{N}$, $\Lambda_t\geq 0$, i.e., $\inf_{t\in\mathbb{N}}\mathbb{E}\left(\Lambda_t\right)\geq 0$.
- Let for all $t\in\mathbb{N}$, $C_t:=\{\mathbb{E}\left(\Lambda_{t+1}-\Lambda_t|\mathcal{F}_t\right)>0\}$. Clearly, 1_{C_t} is $\mathcal{F}_t$-measurable. Hence, with (15), we have:

$$\begin{aligned}\sum_{t\in\mathbb{N}}\mathbb{E}\left(1_{C_t}\cdot(\Lambda_{t+1}-\Lambda_t)\right)\leq&\sum_{t\in\mathbb{N}}\mathbb{E}\left(1_{C_t}\cdot\mathbb{E}\left(\Lambda_{t+1}-\Lambda_t|\mathcal{F}_t\right)\right),\\ \leq&\sum_{t\in\mathbb{N}}\mathbb{E}\left(1_{C_t}\left(\alpha_t\Lambda_t+\beta_t\right)\right),\\ \leq&\sum_{t\in\mathbb{N}}\left(\alpha_tM+\beta_t\right)<+\infty.\end{aligned}$$

- It is also clear that $\sup_{t\in\mathbb{N}}\mathbb{E}\left(\min(\Lambda_t,0)\right)<+\infty$. Consequently, using a result of [Métivier, 1982] (pp. 49-51), (Λ_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 Lipschitz property of ∇f, the sequences (x_t) and $(\nabla f(x_t))$ are a.s. bounded in X. By assumption (9b), the sequence (s_t) is also a.s. bounded in X.

We now prove that $(f(x_t))$ a.s. converges to $f(x^*)$. Coming back to (16), we obtain:

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

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

$$\begin{aligned}\kappa\sum_{t=0}^{n}\gamma_t\mathbb{E}\left(f(x_t)-f(x^*)\right)\leq&\lambda_0-\lambda_{n+1}+\sum_{t=0}^{n}(\alpha_tM+\beta_t),\\ \leq&M+M\sum_{t=0}^{n}\alpha_t+\sum_{t=0}^{n}\beta_t. \qquad (17)\end{aligned}$$

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:

$$\sum_{k\in\mathbb{N}} \gamma_t \left(f(x_t) - f(x^*)\right) < \infty. \tag{18}$$

We now want to use Lemma 4.2. Let $l \in \mathbb{N}$. By convexity of f,

$$\begin{aligned} f(x_l) - f(x_{l+1}) \leq& \langle \nabla f(x_l), x_l - x_{l+1}\rangle, \\ =& -\gamma_l \langle \nabla f(x_l), \Pi_{X^f}\left(s_l + w_l\right)\rangle. \end{aligned} \tag{19}$$

We take now the conditional expectation with respect to $\mathcal{F}_l$:

$$\begin{aligned} f(x_l) - \mathbb{E}\left(f(x_{l+1})|\mathcal{F}_l\right) \leq& -\gamma_l \langle \nabla f(x_l), \Pi_{X^f}\left(s_l + \mathbb{E}\left(w_l|\mathcal{F}_l\right)\right)\rangle, \\ \leq& \gamma_l \|\nabla f(x_l)\| \left(\|s_l\| + \|\mathbb{E}\left(w_l|\mathcal{F}_l\right)\|\right), \\ \leq& \gamma_l \|\nabla f(x_l)\| \left(\|s_l\| + b\epsilon_l(1 + \|\nabla f(x_l)\|)\right) \\ \leq& \gamma_l \delta, \end{aligned} \tag{20}$$

with $\delta > 0$, since we already know that $(\|s_t\|)$ and $(\|\nabla f(x_t)\|)$ are bounded. Hence, we can apply Lemma 4.2, with (18) and (20), which yields

$$\lim_{t\to\infty} f(x_t) = f(x^*) \tag{21}$$

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

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

hence, $\bar{x} \in X^*$.
Suppose now that f is strongly convex with modulus $B > 0$. In this case, X^* reduces to a singleton $\{x^*\}$. By definition,

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

By optimality, $\langle \nabla f(x^*), x_t - x^*\rangle \geq 0$. (22) gives therefore the strong convergence of (x_t) to x^*, and it completes the proof. □

Remark 2.6 (Strong convexity)**.** Following the work of [Berliocchi and Lasry, 1972] we can weaken in point (vii) of Theorem 2.5 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 cle却ity of the proof, we here preferred to make the strong convexity assumption. This assumption can be weakened in a strict convexity assumption.

Remark 2.7 (Random stepsizes)**.** The stepsizes (ρ_t) and (ϵ_t) introduced in Theorems 2.4 and 2.5 can be taken as random sequences with nonnegative values, such that for all $t \in \mathbb{N}$, ρ_t and ϵ_t are $\mathcal{F}_t$-measurable. Indeed, the main results we use in the proofs, such as Robbins-Siegmund's Lemma or Métivier's Proposition on quasimartingales, are available with $(\mathcal{F}_t)$ adapted sequences for the stepsizes. This remark allows possible online definition for these stepsizes, depending on the past σ-fields.

2.3. **A typical theorem.**

Among the lot of literature concerning stochastic approximation algorithms, beginning with [Robbins and Monro, 1951], we focus on the work of [Bertsekas and Tsitsiklis, 2000]. As we mention in the introduction, the results of [Chen and White, 2002] are near from ours, but the algorithms do not present the same abilities, and a comparison between the convergence results would not be sensible. Bertsekas and Tsitsiklis show the following theorem:

Theorem 2.8 (Bertsekas-Tsitsiklis)**.** *(i) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable mapping, such that ∇f is Lipschitz continous.*
(ii) Define (x_t) to be the sequence generated by:

$$x_{t+1} = x_t + \gamma_t(s_t + w_t),$$

where γ_t is a deterministic stepsize, s_t is a descent direction, and w_t is a random noise term. Let $\mathcal{F}_t$ be an increasing sequence of σ-fields.
(iii) Assume that for all $t \in \mathbb{N}$, x_t and s_t are $\mathcal{F}_t$-measurable,
(iv) Assume that there exist two positive scalars c_1 and c_2 such that for all $t \in \mathbb{N}$:

$$c_1 \|\nabla f(x_t)\|^2 \leq - \langle \nabla f(x_t), s_t \rangle, \tag{23a}$$

$$\|s_t\| \leq c_2 \left(1 + \|\nabla f(x_t)\|\right) \tag{23b}$$

(v) Assume that for all $t \in \mathbb{N}$,

$$\mathbb{E}\left(w_t | \mathcal{F}_t\right) = 0, \tag{24a}$$

$$\mathbb{E}\left(\|w_t\|^2 | \mathcal{F}_t\right) \leq A\left(1 + \|\nabla f(x_t)\|^2\right) \tag{24b}$$

where $A > 0$ is a positive deterministic constant.
(vi) Assume that the stepsizes are such that:

$$\forall t \in \mathbb{N},\ \gamma_t > 0,\ \sum_{t \in \mathbb{N}} \gamma_t = +\infty,\ \sum_{t \in \mathbb{N}} (\gamma_t)^2 < +\infty. \tag{25}$$

(vii) Then, either $f(x_t) \to -\infty$ or $f(x_t)$ converges to a finite value, and $\lim_{t\to\infty} \nabla f(x_t) = 0$. Furthermore, every limit point of (x_t) is a stationary point of f.

See [Bertsekas and Tsitsiklis, 2000] for the proof. The main interest of this result is that there is no assumption on the boundedness of f from below. It renders the problem difficult, since any classical result on supermartingales is available for this reason. We are but especially interested in a comparison of the assumptions of this theorem. In the case of f convex, the conclusions of Theorem 2.8 can be adapted to fit with conclusions of Theorem 2.5.

2.4. **Comparison.** We compare here Theorems 2.8 and 2.5. Assumptions in these theorems are of three different types: assumptions on the descent direction, assumptions on the stochastic noises, and assumptions on the stepsizes.

(1) Descent direction: the difference consists in the two equations

$$\langle s_t, x_t - x^* \rangle \leq \kappa \left(f(x^*) - f(x_t)\right), \text{ and, } c_1 \|\nabla f(x_t)\|^2 \leq -\langle \nabla f(x_t), s_t \rangle.$$

Since the mapping f is assumed to be convex in Theorem 2.5, these two assumptions are not so different. In our infinite dimensional setting, the most essential assumption is of course the convexity of f which in return gives the Gâteaux-differentiability of f.

(2) Stochastic noises: clearly, assumptions (10a)–(10b) are weaker than assumptions (24a)–(24b). Indeed, Theorem 2.8 requires the stochastic noises to be martingale increments (cf. (24a)), what Theorem 2.5 only asymptotically requires, and in a more relaxed form (cf. (10a)). Analogously, the variance of the stochastic noises are allowed in Theorem 2.5 to grow a priori infinitely, and in any case to be much bigger than in Theorem 2.8.

(3) Stepsizes: assumptions on stepsizes are quite the same. The descent stepsize γ_t must decrease to zero, not too fast, but not too slow. The difference between the two theorems is hence very slight in that point of view. The last assumption involving the constant $b \geq 0$ in Theorem 2.5 is interesting. Indeed, if one is in the case of (24a), one has in (10a) the constant $b = 0$, and hence, by taking $\epsilon_t = 1$ for all $t \in \mathbb{N}$, (10a) becomes exactly (24a),

which once again shows that Theorem 2.5 can be seen as a generalization of Theorem 2.8.

As a conclusion, the main new feature of Theorem 2.5 is to relax classical assumptions on the stochastic noises, and to provide a convergence proof even in the infinite dimensional case with projections, which was not made in [Bertsekas and Tsitsiklis, 2000].

2.5. **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 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 vector subspace of X. We hence focus on the problem :

$$\min_{x \in X^f} \mathbb{E}\left(j(x(\boldsymbol{\xi}), \boldsymbol{\xi})\right). \tag{26}$$

Notice that since j is convex, then so is f, and hence, f and j are Gâteaux-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 [Barty et al., 2005] 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 2.9. Step t:

- Draw $\boldsymbol{\xi}_{t+1}$ identically, independently from the past draws,
- 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 developped in [Barty et al., 2005]) is a special case of Algorithm 2.1. Indeed, let us define:

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

Then, Algorithm 2.9 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 2.1 with $\gamma_t = \rho_t \epsilon_t$. Clearly, assumptions on convexity of f and (9a)–(9b) are satisfied with our choice of s_t. We now focus son assumptions (10a)–(10b)–(11).

In [Barty et al., 2005], the kernel functions are assumed to be such that:

$$\begin{aligned} \forall t \in \mathbb{N},\ \|s_t - \mathbb{E}\left(s_t(\boldsymbol{\xi}) \frac{1}{\epsilon_t} K_t(\boldsymbol{\xi}, \cdot)\right)\| &\leq b_1 \epsilon_t \left(1 + \|s_t\|\right), \\ \forall x \in \mathbb{R}^m,\ \mathbb{E}\left((K_t(x, \boldsymbol{\xi}))^2\right) &\leq b_2 \epsilon_t, \end{aligned} \tag{27}$$

with two deterministic positive scalars b_1 and b_2. The stepsizes are assumed to be such that:

$$\epsilon_t, \rho_t > 0, \quad \sum_{t \in \mathbb{N}} \epsilon_t \rho_t = +\infty, \quad \sum_{t \in \mathbb{N}} \rho_t (\epsilon_t)^2 < +\infty, \quad \sum_{t \in \mathbb{N}} (\rho_t)^2 \epsilon_t < +\infty. \tag{28}$$

Clearly, (27) and (28) ensure that assumptions (10a)–(10b)–(11) of Theorem 2.5 are satisfied. Hence, Algorithm 2.9 converges, which was already proved in [Barty et al., 2005].

Remark 2.10 (Kiefer-Wolfowitz)**.** Another stochastic approximation algorithm, referred to as Kiefer-Wolfowitz algorithm (see [Kiefer and Wolfowitz, 1952]), computes an approximation of the true gradient, on the basis of finite differences. There are hence in this algorithm two stepsizes, the one corresponding to the descent step γ_t, and the other corresponding to the finite difference approximation. These two steps are required to satisfy joint decreasing assumptions, which are exactly the sames as (28), if you consider the finite difference stepsize to correspond to $(\epsilon_t)^2$, i.e. to the kernel approximation stepsize in Algorithm 2.9.

3. Saddle Point Problems

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

$$\min_{x \in X} \max_{p \in P} L(x,p), \qquad \text{s.t. } x \in X^f,\ p \in P^f, \tag{29}$$

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 gradient algorithm for problem (29):

Algorithm 3.1. 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 γ_t^x, γ_t^p will be in the following the same.

3.2. **Convergence Proof.** Once again, it would be possible to do two proofs as in section 2.2, but we prefer here to restrain ourselves to the case where X^f and P^f are closed subspaces of X and P. We have the following theorem:

Theorem 3.2 (Saddle Point Problems)**.** *(i) Assume that $L(\cdot, p) : X \to \mathbb{R}$ is convex and lower semicontinuous for all $p \in P$, and that $L(x, \cdot) : P \to \mathbb{R}$ is concave and upper semicontinuous for all $x \in X$. Hence, they are Gâteaux-differentiable. Assume moreover that X^f and P^f are closed subspaces of X and P, and that there exist 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 $\nabla_x L(\cdot, \cdot)$ is Lipschitz continuous with modulus l_x and that $\nabla_p L(\cdot, \cdot)$ is Lipschitz continuous with modulus l_p.
(iv) Assume that there exist $c_x, c_p, \kappa > 0$ such that for all $t \in \mathbb{N}$,

$$\langle s_t, x_t - x^* \rangle_X \leq \kappa \left(L(x^*, p_t) - L(x_t, p_t) \right), \tag{30a}$$

$$\langle r_t, p_t - p^* \rangle_P \leq \kappa \left(L(x_t, p_t) - L(x_t, p^*) \right), \tag{30b}$$

$$\| s_t \|_X \leq c_x \left(1 + \| \nabla_x L(x_t, p_t) \| \right), \tag{30c}$$

$$\| r_t \|_P \leq c_p \left(1 + \| \nabla_p L(x_t, p_t) \| \right). \tag{30d}$$

(v) Assume that there are $b_x, b_p \geq 0$, $A_x, A_p > 0$ and nonnegative sequences (ϵ_t^x) and (ϵ_t^p) such that for all $t \in \mathbb{N}$,

$$\|\mathbb{E}\left(w_t|\mathcal{F}_t\right)\|_X \leq b_x \epsilon_t^x \left(1 + \|\nabla_x L(x_t, p_t)\|_X\right), \tag{31a}$$

$$\|\mathbb{E}\left(v_t|\mathcal{F}_t\right)\|_P \leq b_p \epsilon_t^p \left(1 + \|\nabla_p L(x_t, p_t)\|_P\right), \tag{31b}$$

$$\mathbb{E}\left(\|w_t\|_X^2|\mathcal{F}_t\right) \leq A_x \left(1 + \frac{1}{\epsilon_t^x}\|\nabla_x L(x_t, p_t)\|_X^2\right), \tag{31c}$$

$$\mathbb{E}\left(\|v_t\|_P^2|\mathcal{F}_t\right) \leq A_p \left(1 + \frac{1}{\epsilon_t^p}\|\nabla_p L(x_t, p_t)\|_P^2\right). \tag{31d}$$

(vi) Assume that the sequences (γ_t^x), (γ_t^p), (ϵ_x^t), (ϵ_t^p) are all strictly nonnegative and such that:

$$\forall t \in \mathbb{N},\ \gamma_t^x = \gamma_t^p = \gamma_t, \tag{32a}$$

and that

$$\sum_{t\in\mathbb{N}} \gamma_t = +\infty,\ \sum_{t\in\mathbb{N}} b_x \gamma_t \epsilon_t^x < +\infty,\ \sum_{t\in\mathbb{N}} b_p \gamma_t \epsilon_t^p < +\infty\ \sum_{t\in\mathbb{N}} \frac{(\gamma_t)^2}{\epsilon_t^x} < +\infty,\ \sum_{t\in\mathbb{N}} \frac{(\gamma_t)^2}{\epsilon_t^p} < +\infty. \tag{32b}$$

Then, (x_t) and (p_t) are a.s. bounded, and $L(x_t, p^*) \to L(x^*, p^*)$, 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 to x^*.

Proof : We follow the same scheme as in the proof of Theorem 2.5. Let us define for all $t \in \mathbb{N}$, Λ_t our Lyapunov function to be:

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

Using the linearity and nonexpansiveness of the projection over a closed subspace, and Pythagore's inequality leads:

$$\begin{aligned}
\Lambda_{t+1} \leq & \|x_t - x^*\|_X^2 + \|p_t - p^*\|_P^2 \\
& + (\gamma_t)^2 \|s_t + w_t\|_x^2 + (\gamma_t)^2 \|r_t + v_t\|_P^2 \\
& + 2\gamma_t \langle s_t + w_t, x_t - x^*\rangle_X + 2\gamma_t \langle r_t + v_t, p_t - p^*\rangle_P, \qquad (33)\\
\leq & \Lambda_t + (\gamma_t)^2 \left(\|s_t + w_t\|_X^2 + \|r_t + v_t\|_P^2\right) \\
& + 2\gamma_t \left(\langle s_t + w_t, x_t - x^*\rangle_X + \langle r_t + v_t, p_t - p^*\rangle_P\right) \qquad (34)
\end{aligned}$$

With the classical inequality $(a+b)^2 \leq 2(a^2 + b^2)$, and by assumptions (30a)–(30b), we get from (34):

$$\begin{aligned}
\Lambda_{t+1} \leq & \Lambda_t + 2(\gamma_t)^2 \left(\|s_t\|_X^2 + \|w_t\|_X^2\right) \\
& + 2(\gamma_t)^2 \left(\|r_t\|_P^2 + \|v_t\|_P^2\right) \\
& + 2\gamma_t \left(L(x^*, p_t) - L(x_t, p_t) + L(x_t, p_t) - L(x_t, p^*)\right) \\
& + 2\gamma_t \left(\langle w_t, x_t - x^*\rangle_X + \langle v_t, p_t - p^*\rangle_P\right) \qquad (35)
\end{aligned}$$

Moreover, by assumptions (30c)–(30d), we get:

$$\begin{aligned}
\|s_t\|_X^2 \leq & 2c_x^2 \left(1 + \|\nabla_x L(x_t, p_t)\|_X^2\right), \\
\|r_t\|_P^2 \leq & 2c_p^2 \left(1 + \|\nabla_p L(x_t, p_t)\|_P^2\right).
\end{aligned}$$

We will plug these inequalities in (35), and take the conditional expectation with respect to $\mathcal{F}_t$. It yields, by assumptions (31c)–(31d) and (31a)–(31b):

$$\begin{aligned}\mathbb{E}\left(\Lambda_{t+1}|\mathcal{F}_t\right) \leq &\Lambda_t + 2(\gamma_t)^2\left(2c_x^2 + A_x + \|\nabla_x L(x_t,p_t)\|_X^2\left(2c_x^2 + \frac{A_x}{\epsilon_t^x}\right)\right)\\ &+ 2(\gamma_t)^2\left(2c_p^2 + A_p + \|\nabla_p L(x_t,p_t)\|_P^2\left(2c_p^2 + \frac{A_p}{\epsilon_t^p}\right)\right)\\ &+ 2\gamma_t\left(L(x^*,p_t) - L(x_t,p^*)\right)\\ &+ 2\gamma_t\left(b_x\epsilon_t^x(1+\|\nabla_x L(x_t,p_t)\|_X)\|x_t - x^*\|_X\right)\\ &+ 2\gamma_t\left(b_p\epsilon_t^p(1+\|\nabla_p L(x_t,p_t)\|_P)\|p_t - p^*\|_P\right) \qquad (36)\end{aligned}$$

We will now use the Lipschitz property of $\nabla_x L$ and $\nabla_p L$. Indeed, one has:

$$\begin{aligned}\|\nabla_x L(x_t,p_t)\|_X \leq &\|\nabla_x L(x_t,p_t) - \nabla_x L(x^*,p^*)\|_X + \|\nabla_x L(x^*,p^*)\|_X\\ \leq &l_x\|x_t - x^*\|_X + l_x\|p_t - p^*\|_P + \|\nabla_x L(x^*,p^*)\|_X,\end{aligned}$$

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

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

with (α_t) and (β_t) two summable sequences defined in the same way as in the Proof of Theorem 2.5:

$$\begin{aligned}\alpha_t =&8(l_x\gamma_t)^2\left(2c_x^2 + \frac{A_x}{\epsilon_t^x}\right) + 8(l_p\gamma_t)^2\left(2c_p^2 + \frac{A_p}{\epsilon_t^p}\right)\\ &+ b_x\gamma_t\epsilon_t^x(4+3l_x) + b_p\gamma_t\epsilon_t^p(4+3l_p)\\ \beta_t =&4\|\nabla_x L(x^*,p^*)\|_X^2(\gamma_t)^2\left(2c_x^2 + \frac{A_x}{\epsilon_t^x}\right) + 4\|\nabla_p L(x^*,p^*)\|_P^2(\gamma_t)^2\left(2c_p^2 + \frac{A_p}{\epsilon_t^p}\right)\\ &+ 2b_x\gamma_t\epsilon_t^x(1+\|\nabla_x L(x^*,p^*)\|_X)\\ &+ 2b_p\gamma_t\epsilon_t^p(1+\|\nabla_p L(x^*,p^*)\|_P) + 2(\gamma_t)^2(2c_x^2 + A_x) + 2(\gamma_t)^2(2c_p^2 + A_p)\end{aligned}$$

Using the saddle point assumption in (x^*,p^*), one get with (37):

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

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

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 (Λ_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 the Lipschitz property of $\nabla_x L$ and $\nabla_p L$, we get also that $(\nabla_x L(x_t,p^*))$, $(\nabla_p L(x^*,p_t))$, (s_t) and (r_t) are a.s. bounded.
Moreover, by making the same calculations as those leading to (18), we obtain:

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

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

By convexity of $L(\cdot,p^*)$ and concavity of $L(x^*,\cdot)$, we make the same calculations as in (19)–(20), which are still valid by the boundedness of the gradient sequences and the assumptions of the theorem, and finally get by Lemma 4.2:

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

$$\lim_{t\to\infty} L(x^*,p_t) = L(x^*,p^*). \qquad (40b)$$

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^*).
Finally, if $L(\cdot, p^*)$ is strongly convex, by the same equation as (22), we obtain that (x_t) strongly converges to x^*. □

4. Technical Lemmas

We here provide two technical lemmas we use in the preceding convergence proof of theorem 2.5.

Lemma 4.1. *Let (x_k) be a sequence of nonnegative real numbers. Let (α_k) and (β_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:*

$$\forall k \in \mathbb{N},\ x_{k+1} - x_k \le \alpha_k x_k + \beta_k,$$

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

The proof can be found in [Cohen, 1984].

Lemma 4.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be some probability space, equipped with a filtration $(\mathcal{F}_k)$. Let J be a real valued mapping from an Hilbert space H. 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 (γ_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* $\forall k \in \mathbb{N}$, $J(u_k) - \mu \ge 0$, *a.s.*
(iii) $\exists \delta > 0$, $\forall k \in \mathbb{N}$, $J(u_k) - \mathbb{E}(J(u_{k+1})|\mathcal{F}_k) \le \delta\gamma_k$, *a.s.*

Then $(J(u_k))$ a.s. converges to μ.

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

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

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

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

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

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

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

□

5. Conclusion

As a conclusion, the general framework we proposed here for perturbed gradient algorithms with stochastic noises allows to consider many practical cases, like infinite dimensional stochastic optimization problems. In this framework, we provide a constructive convergence proof for our algorithm.
Moreover, the assumptions we require on the stochastic noises are weaker than

those usually required, since we only require asymptotically assumptions that are usually required all along the iterations. Furthermore, our setting is compatible with the projection either on a closed vector subspace or on a closed convex subset of the initial Hilbert space in which we optimize. This is an interesting feature for measurability constraints like nonanticipativity constraints of multi-stage stochastic programs, which can be taken into account in our framework.

References

[Barty et al., 2005] Barty, K., Roy, J.-S., and Strugarek, C. (2005). A stochastic gradient type algorithm for closed loop problems. *submitted to SPEPS.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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