Convergence d'un algorithme du gradient proximal stochastique à pas constant et généralisation aux opérateurs monotones aléatoires

Adil Salim, Pascal Bianchi, Walid Hachem

Convergence d'un algorithme du gradient proximal stochastique à pas constant et généralisation aux opérateurs monotones aléatoires

Adil SALIM[1], Pascal BIANCHI[1], Walid HACHEM[2]

[1]LTCI, Télécom ParisTech, Université Paris-Saclay
46, rue Barrault, 75634 Paris Cedex 13, France

[2]CNRS / LIGM (UMR 8049), Université Paris-Est Marne-la-Vallée
5 Boulevard Descartes, Cité Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2, France.
adil.salim,pascal.bianchi@telecom-paristech.fr, walid.hachem@u-pem.fr

Résumé – L'algorithme du gradient proximal permet de trouver les minimiseurs d'une somme $F + G$ de deux fonctions convexes propres et fermées, l'une étant supposée dérivable. Cet article introduit une version stochastique de cet algorithme. Les itérations font intervenir une suite iid de deux fonctions aléatoires, dont les espérances coïncident respectivement avec F et G, ainsi que des projections aléatoires sur des ensembles convexes fermés. L'objectif est de fournir une analyse de convergence, dans un contexte adaptatif où le pas de l'algorithme est supposé constant. On montre que, en moyenne de Césaro, la probabilité pour que les itérées soient hors d'un voisinage des minimiseurs souhaités est arbitrairement faible lorsque le nombre d'itérations tend vers l'infini, et dans la limite de pas faibles. Le comportement ergodique des itérées est également étudié. Enfin, l'algorithme est étendu au contexte plus général des opérateurs maximaux monotones aléatoires.

Abstract – The proximal gradient algorithm allows to find the minimizers of a sum $F + G$ of two proper closed convex functions, one of them being differentiable. This paper introduces a stochastic version of the proximal gradient algorithm. The iterations involve an iid sequence of two random functions, whose expectations coincide with F and G respectively, as well as random projections onto closed convex sets. The aim is to provide a convergence analysis in an adaptive context where the step size of the algorithm is constant. We prove that, in Césaro mean, the probability that the iterates are away from the sought minimizers is small when the number of iterations tends to infinity and in the limit of small step sizes. The ergodic behavior is studied as well. Finally, the algorithm is extended to the context of random maximal monotone operators.

1 Introduction

L'algorithme du gradient stochastique permet de minimiser de manière approchée une fonction de coût s'écrivant sous la forme d'une espérance $x \mapsto \mathbb{E}(f(\xi, x))$ où ξ est une variable aléatoire (v.a.), $f(\xi, .) : \mathbb{R}^N \to \mathbb{R}$ est une fonction de classe C^1 sur $\mathbb{R}^N$, que nous supposerons en outre convexe, et N est un entier. Son utilisation est pertinente dans le cas où, par exemple, l'espérance n'est pas calculable, mais révélée au cours du temps, au travers de l'observation d'une suite de copies indépendantes et identiquement distribuées (ξ_n) de ξ. Les itérations sont de la forme $x_{n+1} = x_n - \gamma_n \nabla f(\xi_{n+1}, x_n)$ où (γ_n) est une suite positive qui représente les pas de l'algorithme. Dans le contexte du traitement adaptatif du signal, il est fréquent de supposer que le pas est constant, c'est à dire $\gamma_n \equiv \gamma$. Dans ce cas de figure, la suite (x_n) ne converge généralement pas au sens presque sûr lorsque $n \to \infty$, mais "fluctue" dans un voisinage de l'ensemble des minimiseurs de la fonction de coût (cet ensemble étant supposé non vide).

L'objectif de cet article est d'analyser une généralisation de l'algorithme du gradient stochastique : l'algorithme du gradient *proximal* stochastique. Dans sa version déterministe standard, l'algorithme du gradient proximal est un algorithme couramment utilisé en optimisation convexe et en apprentissage statistique. Soit $F : \mathbb{R}^N \to \mathbb{R}$ une fonction convexe de classe C^1 sur $\mathbb{R}^N$ et soit $G : \mathbb{R}^N \to (-\infty, \infty]$ une fonction propre, convexe et semi continue inférieurement (notation : $G \in \Gamma_0$). Supposons que $F + G$ possède un minimiseur, en d'autres termes, que l'ensemble des zéros $Z(\nabla F + \partial G)$ de $\nabla F + \partial G$ soit non vide. On se donne pour objectif de trouver un minimiseur de $F + G$ à l'aide d'un algorithme itératif. L'algorithme du gradient proximal s'écrit

$$\mathsf{x}_{n+1} = \mathrm{prox}_{\gamma G}(\mathsf{x}_n - \gamma \nabla F(\mathsf{x}_n)) \tag{1}$$

où

$$\mathrm{prox}_{\gamma G}(x) := \arg\min_{w \in \mathbb{R}^N} G(w) + \frac{1}{2\gamma}\|w - x\|^2$$

est l'*opérateur proximal* de Moreau et où $\gamma > 0$ est un pas d'adaptation. Si ∇F est Lipschitz et si γ est suffisamment petit, cet algorithme converge vers un point de $Z(\nabla F + \partial G)$.

Dans le paragraphe 2, nous exposons le problème d'optimisation et introduisons une version stochastique de l'algorithme du gradient proximal. Nous présentons une analyse du comportement des itérées dans le régime asympotique où premièrement $n \to \infty$ et où, deuxièmement, le pas de l'algorithme γ tend vers zéro. Dans le paragraphe 3, nous généralisons l'algorithme au contexte des opérateurs monotones. Les opérateurs

monotones sont des fonctions multivaluées qui généralisent la notion de sous-différentielle d'une fonction convexe propre et semi-continue inférieurement. Nous établissons un résultat de convergence et fournissons quelques éléments de preuve.

2 Le gradient proximal stochastique à pas constant

Soit $(\Sigma, \mathscr{F}, \rho)$ un espace de probabilité et soit $f : \Sigma \times \mathbb{R}^N \to \mathbb{R}$ une fonction telle que $f(\cdot, x)$ soit mesurable pour tout $x \in \mathbb{R}^N$ et telle que $f(s, \cdot)$ soit convexe et de classe C^1 pour tout $s \in \Sigma$. Nous supposons $\int |f(s,x)|\, \rho(ds) < \infty$ pour tout $x \in \mathbb{R}^N$ et nous définissons sur $\mathbb{R}^N$ la fonction $F(x) := \int f(s,x)\, \rho(ds)$. Cette fonction est dérivable et satisfait $\nabla F(x) = \int \nabla f(s,x)\, \rho(ds)$ grâce à la convexité de $f(s,\cdot)$. Soit $h : \Sigma \times \mathbb{R}^N \to \mathbb{R}$ un intégrande normal convexe (*cf.* [6], en particulier, $h(s,\cdot)$ est convexe). Pour simplifier, nous supposons ici que h soit fini partout, mais cette hypothèse peut être relâchée. Nous supposons également que $\int |h(s,x)|\rho(ds) < \infty$ pour tout $x \in \mathbb{R}^N$, et nous considérons la fonction convexe $H(x) := \int h(s,x)\, \rho(ds)$ définie sur $\mathbb{R}^N$. Enfin, soit $m \in \mathbb{N}^*$ et soit $\{\mathcal{C}_1, \dots, \mathcal{C}_m\}$ une famille de fermés convexes de $\mathbb{R}^N$. Nous supposons que l'intersection $\bigcap_{i=1}^m \mathrm{ri}(\mathcal{C}_i)$ des intérieurs relatifs des $\mathcal{C}_i$ soit non vide.

Notre but est de résoudre d'une manière approchée le problème d'optimisation

$$\min_{x \in \mathcal{C}} F(x) + H(x), \quad \mathcal{C} := \bigcap_{i=1}^m \mathcal{C}_i \tag{2}$$

dont nous supposons le minimum atteint.

Pour résoudre ce problème, nous mettons en œuvre une version stochastique de l'algorithme du gradient proximal. Soit (u_n) une suite indépendante et identiquement distribuée (iid) sur Σ de loi ρ et soit (I_n) une suite iid de loi α sur l'ensemble $\{0, 1, \dots, m\}$. Nous supposons que $\alpha(k) = \mathbb{P}(I_1 = k) > 0$ pour tout k et que (I_n) et (u_n) soient indépendantes. Afin de résoudre le problème (2), nous considérons les itérées

$$x_{n+1} = \begin{cases} \mathrm{prox}_{\alpha(0)^{-1}\gamma h(u_{n+1},\cdot)}(x_n - \gamma \nabla f(u_{n+1}, x_n)) & \\ & \text{si } I_{n+1} = 0, \\ \Pi_{\mathcal{C}_{I_{n+1}}}(x_n - \gamma \nabla f(u_{n+1}, x_n)) & \text{sinon,} \end{cases} \tag{3}$$

où $\gamma > 0$ est un pas et où $\Pi_{\mathcal{C}_i}$ est le projecteur sur $\mathcal{C}_i$.

Cet algorithme peut présenter de l'intérêt dans les situations où les fonctions $F(x) = \mathbb{E} f(u_1, x)$ et $H(x) = \mathbb{E} h(u_1, x)$ ne sont pas accessibles ou sont difficiles à calculer, mais où des suites « bruitées » $(f(u_n,\cdot))_n$ et $(h(u_n,\cdot))_n$ sont observables. De plus, si le nombre m des ensembles-contraintes $\mathcal{C}_i$ est élevé, l'opération de projection sur $\mathcal{C}$ peut s'avérer complexe, alors que les projections sur les $\mathcal{C}_i$ sont le plus souvent simples (typiquement, les $\mathcal{C}_i$ sont des demi espaces). Dans cette situation, il peut être utile de remplacer la projection sur $\mathcal{C}$ par des projections aléatoires sur les $\mathcal{C}_i$.

D'autres applications de l'algorithme (3) peuvent être envisagées. Dans [7], des variantes de cet algorithme relatives à des problèmes d'optimisation sur des graphes – citons les problèmes de « trend filtering » ou de « graph inpainting » – sont également étudiées.

Le problème (2) équivaut au problème de la recherche d'un élément de $Z(\nabla F + \partial G)$ où $G(x) := \sum_{k=1}^m \iota_{\mathcal{C}_k}(x) + H(x)$ et où ι_S est l'indicatrice d'un ensemble S au sens de la théorie de l'optimisation. Comme l'algorithme (3) est un algorithme à pas constant, il ne converge pas en général vers $Z(\partial G + \nabla F)$. Néanmoins, si γ est petit, les itérées x_n restent proches de cet ensemble pour les grandes valeurs de n. Il en sera de même des moyennes empiriques

$$\bar{x}_n := \frac{1}{n+1} \sum_{k=0}^n x_k \,.$$

Ces résultats sont décrits d'une manière plus précise dans le theorème suivant.

Nous rappelons qu'une fonction réelle q est dite coercive si $\lim_{\|x\|\to\infty} q(x) = \infty$; cette fonction est dite super coercive si $\lim_{\|x\|\to\infty} q(x)/\|x\| = \infty$. Nous rappelons également que l'*enveloppe de Moreau* de coefficient $\gamma > 0$ d'une fonction $q \in \Gamma_0$ est

$$q_\gamma(x) := \min_w q(w) + \frac{1}{2\gamma}\|w - x\|^2.$$

Enfin, nous notons $d(x, S)$ la distance d'un point x à un ensemble S.

Théorème 1. Supposons vraies les hypothèses suivantes :

- H1 Il existe $x_\star \in Z(\nabla F + \partial G)$ qui satisfait les conditions suivantes : $\nabla f(\cdot, x_\star) \in \mathcal{L}^2(\rho)$, il existe $\varphi \in \mathcal{L}^2(\rho)$ telle que $\varphi(s) \in \partial h(s, x_\star)$ ρ-presque partout, et $-\int(\varphi + \nabla f(\cdot, x_\star))d\rho \in \boldsymbol{N}_{\mathcal{C}}$ où $\boldsymbol{N}_{\mathcal{C}}$ est le cône normal à $\mathcal{C}$ en $x_\star$,
- H2 Il existe $c > 0$ tel que pour tout $x \in \mathbb{R}^N$,

$$\int \langle \nabla f(s,x) - \nabla f(s, x_\star), x - x_\star \rangle\, \rho(ds) \geq c \int \|f(s,x) - f(s,x_\star)\|^2\, \rho(ds),$$

- H3 La fonction $F+G$ satisfait l'une des propriétés suivantes :
  - (a) $F + G$ est coercive,
  - (b) $F + G$ est super coercive,
- H4 Pour tout compact $\mathcal{K} \subset \mathbb{R}^N$, il existe $\varepsilon > 0$ tel que

$$\sup_{x \in \mathcal{K} \cap \mathcal{C}} \int \|\partial h_0(s,x)\|^{1+\varepsilon}\, \rho(ds) < \infty,$$

  où $\partial h_0(s,\cdot)$ est l'élément de norme euclidienne minimale de $\partial h(s,\cdot)$,
- H5 Il existe une boule fermée dans $\mathbb{R}^N$ telle que $\|\nabla f(s,x)\| \leq M(s)$ pour tout x dans cette boule, où $M(s)$ est une fonction ρ-intégrable. De plus, pour tout compact $\mathcal{K} \subset \mathbb{R}^N$, il existe $\varepsilon > 0$ tel que

$$\sup_{x \in \mathcal{K}} \int \|\nabla f(s,x)\|^{1+\varepsilon}\, \rho(ds) < \infty \,,$$

H6 Les ensembles $\mathcal{C}_1, \ldots, \mathcal{C}_m$ sont *linéairement réguliers*, en d'autres termes, il existe $\kappa > 0$ tel que

$$\forall x \in \mathbb{R}^N, \quad \max_{i\in[m]} d(x, \mathcal{C}_i) \geq \kappa d(x, \mathcal{C}),$$

H7 Il existe $\gamma_0 > 0$ tel que pour tout $\gamma \in]0, \gamma_0]$ et pour tout $x \in \mathbb{R}^N$,

$$\int \left(\|\nabla h_\gamma(s,x)\| + \|\nabla f(s,x)\|\right) \rho(ds) \leq C(1 + |F(x) + H_\gamma(x)|)\,,$$

où $h_\gamma(s,\cdot)$ est l'enveloppe de Moreau de $h(s,\cdot)$.

Alors, pour toute variable aléatoire x_0 telle que $\mathbb{E}[x_0^2] < \infty$,

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}\left[d(x_k, Z(\nabla F + \partial G)) > \varepsilon\right] \xrightarrow[\gamma\to 0]{} 0\,.$$

Par ailleurs, si l'hypothèse H3–(b) est satisfaite,

$$\limsup_{n\to\infty} \mathbb{P}\left[d\left(\bar{x}_n, Z(\nabla F + \partial G)\right) \geq \varepsilon\right] \xrightarrow[\gamma\to 0]{} 0, \text{ et}$$
$$\limsup_{n\to\infty} d\left(\mathbb{E}[\bar{x}_n], Z(\nabla F + \partial G)\right) \xrightarrow[\gamma\to 0]{} 0\,.$$

L'hypothèse H2 nécessite un commentaire. Supposons qu'il existe une fonction mesurable $\beta(s) \geq 0$ telle que $\|\nabla f(s,x) - \nabla f(s,x')\| \leq \beta(s)\|x - x'\|$ pour tout x, x', s. Alors, par le théorème de Baillon-Haddad, $\langle \nabla f(s,x) - \nabla f(s,x'), x - x'\rangle \geq \frac{1}{\beta(s)}\|\nabla f(s,x) - \nabla f(s,x')\|^2$. Ainsi, l'hypothèse H2 est évidemment satisfaite dans le cas où $\beta(s)$ est bornée.

3 Généralisation aux opérateurs maximaux monotones aléatoires

3.1 Le opérateurs maximaux monotones et l'algorithme « Forward Backward »

Les opérateurs maximaux monotones [3], présentés succinctement ici, généralisent les sous-différentiels des fonctions dans Γ_0. Soit $\mathsf{A} : \mathbb{R}^N \to 2^{\mathbb{R}^N}$ un opérateur multivoque. Le domaine et le graphe de A sont les sous-ensembles respectifs de $\mathbb{R}^N$ et de $\mathbb{R}^N \times \mathbb{R}^N$ définis comme $\mathrm{dom}(\mathsf{A}) := \{x \in \mathbb{R}^N : \mathsf{A}(x) \neq \emptyset\}$ et $\mathrm{gr}(\mathsf{A}) := \{(x,y) \in \mathbb{R}^N \times \mathbb{R}^N : y \in \mathsf{A}(x)\}$. L'opérateur A est dit *monotone* si $\forall x, x' \in \mathrm{dom}(\mathsf{A}), \forall y \in \mathsf{A}(x), \forall y' \in \mathsf{A}(x')$, $\langle y - y', x - x'\rangle \geq 0$. Un opérateur monotone A de domaine non vide est dit *maximal* si $\mathrm{gr}(\mathsf{A})$ est un élément maximal pour l'inclusion dans l'ensemble des graphes des opérateurs monotones.

Soit I l'opérateur identité et soit A^{-1} l'inverse de l'opérateur A, défini par le fait que $(x,y) \in \mathrm{gr}(\mathsf{A}^{-1}) \Leftrightarrow (y,x) \in \mathrm{gr}(\mathsf{A})$. Il est bien connu que A appartient à l'ensemble $\mathscr{M}$ des opérateurs maximaux monotones sur $\mathbb{R}^N$ si et seulement si pour tout $\gamma > 0$, la *résolvante* $(I + \gamma\mathsf{A})^{-1}$ est une contraction définie sur tout l'espace $\mathbb{R}^N$. En particulier, cette résolvante est univoque. Dans le cas où $\mathsf{A} = \partial q$ où $q \in \Gamma_0$, l'opérateur $(I + \gamma\mathsf{A})^{-1}$ n'est autre que l'opérateur proximal $\mathrm{prox}_{\gamma q}$.

L'ensemble des zéros d'un opérateur $\mathsf{A} \in \mathscr{M}$ est l'ensemble $Z(\mathsf{A}) := \{x \in \mathbb{R}^N : 0 \in \mathsf{A}(x)\}$. Etant donné deux opérateurs $\mathsf{A}, \mathsf{B} \in \mathscr{M}$, un algorithme itératif de « splitting » est un algorithme de recherche d'un élément de $Z(\mathsf{A} + \mathsf{B})$ au sein duquel les opérations sur A et les opérations sur B sont réalisées séparément. L'algorithme dit du Forward Backward (FB) est un algorithme de splitting conçu pour les situations où B est univoque. Il s'écrit

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

où γ est un pas positif. La convergence de cet algorithme est assurée si B satisfait une condition dite de co coercivité et si γ est suffisamment petit.

Dans le cas où $\mathsf{A} = \partial G$ et $\mathsf{B} = \nabla F$ où F et G sont les fonctions décrites dans l'introduction de cet article, l'algorithme FB coïncide avec l'algorithme du gradient proximal (1). Au delà de ce cas, d'autres d'algorithmes d'optimisation comme l'algorithme primal-dual, ADMM, etc. s'avèrent être des instances de l'algorithme FB. Dans la suite de cet article, nous considérons une version aléatoire de l'algorithme FB. Cet algorithme généralise l'algorithme (3) et ouvrira également la voie à des versions aléatoires des algorithmes primaux-duaux, d'algorithmes d'optimisation sur les graphes ou d'autres. Ces applications sont en cours d'étude.

3.2 L'algorithme FB aléatoire

Nous remplaçons maintenant les opérateurs A et B par des opérateurs monotones maximaux aléatoires. Etant donné un espace de probabilité $(\Omega, \mathscr{G}, \mu)$, l'opérateur B est remplacé par une fonction mesurable $B : \Omega \times \mathbb{R}^N \to \mathbb{R}^N$ telle que $B(\omega,\cdot)$ soit continue et monotone (en tant qu'opérateur univoque) sur $\mathbb{R}^N$. En supposant que pour tout $x \in \mathbb{R}^N$, $B(\cdot, x) \in \mathcal{L}^1(\mu)$, nous définissons aussi *l'opérateur moyen*

$$\mathcal{B}(x) := \int B(\omega, x)\, \mu(d\omega)\,.$$

Nous supposons que $\mathcal{B}$ est continu. En particulier, il est maximal [3]. L'opérateur A est remplacé par une fonction aléatoire $A : \Omega \to \mathscr{M}$ (les questions liées à la mesurabilité de cette fonction multivoque sont traitées dans [1, 2]). Dans la suite, nous noterons $A(\omega, x)$ l'image de x par l'opérateur $A(\omega)$. Nous considérons la situation où les ensembles $D(\omega) := \mathrm{dom}\, A(\omega)$ peuvent être différents. L'*intersection essentielle* des $D(\omega)$, notée $\mathcal{D}$, est alors définie par le fait que $x \in \mathcal{D} \Leftrightarrow \mu(\{\omega : x \in D(\omega)\}) = 1$. Pour tout $x \in \mathcal{D}$, nous supposons que l'ensemble

$$\mathcal{S}_x := \{\varphi \in \mathcal{L}^1(\mu) : \varphi(\omega) \in A(\omega, x)\ \mu - \text{p.p.}\}$$

des *sélections intégrables* de $A(\cdot, x)$ soit non vide. Sur $\mathcal{D}$, l'opérateur moyen $\mathcal{A}$ sera alors

$$\mathcal{A}(x) := \overline{\left\{\int \varphi d\mu : \varphi \in \mathcal{S}_x\right\}}$$

(une telle intégrale s'appelle une *intégrale de sélection*). Il est facile de constater que l'opérateur $\mathcal{A}$ ainsi défini est monotone.

Dans la suite, nous supposerons que cet opérateur soit également maximal. Des conditions qui garantissent la maximalité de $\mathcal{A}$ sont présentées dans [3, Chap. II.6] ou dans [1, Prop. 3.1].

Notre objectif est d'étudier l'algorithme suivant. Etant donné une suite (ξ_n) de variables aléatoires iid à valeurs dans Ω et de loi μ, nous étudions le comportement des itérées

$$x_{n+1} = (I + \gamma A(\xi_{n+1}, \cdot))^{-1}(x_n - \gamma B(\xi_{n+1}, x_n)) \qquad (4)$$

où $\gamma > 0$. Nous montrerons que si γ est petit, les itérées x_n sont proches de $Z(\mathcal{A}+\mathcal{B})$ (supposé non vide) pour les grandes valeurs de n.

Avant d'étudier la convergence de cet algorithme, revenons au problème du paragraphe 2 et posons $\Omega = \Sigma \times \{0, \dots, m\}$ et $\mu = \rho \otimes \alpha$. En écrivant $\omega = (u, i)$, nous posons $B(\omega, x) = \nabla_x f(u, x)$. En définissant la fonction $g : \Omega \times \mathbb{R}^N \to]-\infty, \infty]$ de la manière suivante :

$$g(\omega, x) := \begin{cases} \alpha(0)^{-1} h(u, x) & \text{si } i = 0, \\ \iota_{\mathcal{C}_i} & \text{sinon,} \end{cases}$$

nous constatons que $g(\omega, \cdot) \in \Gamma_0$ pour tout ω et nous posons $A(\omega, \cdot) = \partial_x g(\omega, \cdot)$. Il est important de remarquer que l'opérateur moyen $\mathcal{A}(\cdot) = \mathbb{E}_\mu \partial_x g(\omega, \cdot)$ n'est autre que $\partial G(\cdot) = \partial \mathbb{E}_\mu g(\omega, \cdot)$, les hypothèses sur g rendant l'échange entre $\mathbb{E}_\mu$ et ∂ licite (*cf.* [8]). Remarquons aussi que cet échange garantit la maximalité de $\mathcal{A}$, la fonction G étant un élément de Γ_0.

En remarquant que $\Pi_{\mathcal{C}_i} = \mathrm{prox}_{\iota_{\mathcal{C}_i}}$, la construction de $A(\omega, \cdot)$ et de $B(\omega, \cdot)$ que nous venons de faire nous montre que l'algorithme (3) est une instance de l'algorithme (4).

3.3 Approche et résultat général

Notre approche pour étudier le comportement dynamique de l'algorithme (4) s'inspire de la méthode de l'équation différentielle ordinaire, méthode bien connue dans le domaine de l'approximation stochastique. Comme l'opérateur $\mathcal{A}$ est maximal et $\mathcal{B}$ est continu sur $\mathbb{R}^N$, l'opérateur $\mathcal{A}+\mathcal{B}$ est maximal [3]. Un résultat classique de la théorie des opérateurs monotones nous dit alors que pour tout $x_0 \in \mathcal{D}$, *l'inclusion différentielle* (ID)

$$\begin{cases} \dot{\mathsf{x}}(t) & \in & -(\mathcal{A}+\mathcal{B})(\mathsf{x}(t)) \\ \mathsf{x}(0) & = & x_0 \end{cases}$$

admet une unique solution absolument continue $\mathsf{x} : \mathbb{R}_+ \to \mathcal{D}$ [3]. De plus, la fonction $\Phi : \mathcal{D} \times \mathbb{R}_+ \to \mathcal{D}, (x_0, t) \mapsto \mathsf{x}(t)$ où x est la solution de l'ID de valeur initiale x_0 peut être étendue à $\overline{\mathcal{D}} \times \mathbb{R}_+$ et constitue un *semiflot* sur cet ensemble.

Revenons au processus (x_n) produit par l'algorithme (4). Soit $\mathsf{x}_\gamma(t)$ le processus continu obtenu par l'interpolation linéaire par morceaux des itérées x_n avec un pas d'interpolation égal à γ. La première étape de l'approche consiste à démontrer que x_γ est proche de la solution de l'ID pour les petites valeurs de γ. Plus précisément, nous montrons que le processus aléatoire x_γ converge étroitement vers une solution de l'ID pour la métrique de la convergence (uniforme) sur les compacts de $\mathbb{R}_+$. La convergence étroite ne suffit pas pour contrôler le comportement à long terme des itérées x_n. Un résultat supplémentaire de *stabilité* est nécessaire. Afin d'obtenir ce résultat, nous regardons (x_n) comme une chaîne de Markov fellerienne dont le noyau de transition est paramétré par γ. Nous nous donnons pour objectif de démontrer que l'ensemble des mesures invariantes de la chaîne est non vide, que l'ensemble de ces mesures invariantes obtenues quand γ parcourt un certain intervalle $]0, \gamma_0]$ est *tendu* et que ces mesures chargent principalement un petit voisinage de l'ensemble $\mathcal{D}$. Dans le cadre du paragraphe 2, les hypothèses H2, H3, H6 et H7 servent à obtenir ces résultats. La convergence étroite de x_γ d'une part et la tension des mesures invariantes d'autre part nous permettent de montrer que les points d'accumulation de ces mesures invariantes quand $\gamma \downarrow 0$ sont également *invariantes pour le semi flot* Φ (voir [5] pour des contextes similaires). Ce résultat nous permet de contrôler le comportement à long terme des mesures empiriques des itérées ainsi que leurs espérances. Nous aboutissons ainsi aux résultats de convergence énoncés par le théorème 2 ci-dessous, énoncé de manière informelle par manque de place (voir [2] pour un énoncé rigoureux).

Théorème 2. Avec des hypothèses qui généralisent les hypothèses H1–H7 du théorème 1, pour tout $\varepsilon > 0$,

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}(d(x_k, \mathcal{U}) > \varepsilon) \xrightarrow[\gamma\to 0]{} 0$$

où $\mathcal{U}$ est l'union des supports des mesures invariantes pour le semiflot Φ. Si l'opérateur $\mathcal{A}+\mathcal{B}$ est *demipositif* [4],

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}\left(d(x_k, Z(\mathcal{A}+\mathcal{B})) > \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0\,.$$

De plus, que $\mathcal{A}+\mathcal{B}$ soit demipositif ou non,

$$\limsup_{n\to\infty} \mathbb{P}\left(d\left(\bar{x}_n, Z(\mathcal{A}+\mathcal{B})\right) \geq \varepsilon\right) \xrightarrow[\gamma\to 0]{} 0\,,$$
$$\limsup_{n\to\infty} d\left(\mathbb{E}[\bar{x}_n], Z(\mathcal{A}+\mathcal{B})\right) \xrightarrow[\gamma\to 0]{} 0\,.$$

Références

[1] P. Bianchi and W. Hachem. Dynamical behavior of a stochastic Forward-Backward algorithm using random monotone operators. *J. Optim. Theory Appl.*, 171(1) :90–120, 2016.

[2] P. Bianchi, W. Hachem, and A. Salim. A constant step Forward-Backward algorithm involving random maximal monotone operators. *arXiv preprint arXiv :1702.04144*, 2017.

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

[4] R. E. Bruck, Jr. Asymptotic convergence of nonlinear contraction semigroups in Hilbert space. *J. Funct. Anal.*, 18 :15–26, 1975.

[5] J.-C. Fort and G. Pagès. Asymptotic behavior of a Markovian stochastic algorithm with constant step. *SIAM J. Control Optim.*, 37(5) :1456–1482 (electronic), 1999.

[6] R. T. Rockafellar and R. J.-B. Wets. On the interchange of subdifferentiation and conditional expectations for convex functionals. *Stochastics*, 7(3) :173–182, 1982.

[7] A. Salim, P. Bianchi, and W. Hachem. Snake : a stochastic proximal gradient algorithm for regularized problems over large graphs. in preparation, 2017.

[8] D. W. Walkup and R. J.-B. Wets. Stochastic programs with recourse. II : On the continuity of the objective. *SIAM J. Appl. Math.*, 17 :98–103, 1969.