Convergence de l'algorithme ADAM du point de vue des systèmes dynamiques

Anas BARAKAT, Pascal BIANCHI

LTCI, Télécom Paris, Institut polytechnique de Paris
46, rue Barrault, 75634 Paris Cedex 13, France
anas.barakat@telecom-paristech.fr, pascal.bianchi@telecom-paristech.fr

Résumé – ADAM est une variante populaire de la descente de gradient stochastique qui a pour objectif de trouver un minimiseur local d'une fonction. La fonction objective est inconnue mais un estimateur aléatoire du gradient est observé à chaque itération de l'algorithme. En dépit de la popularité de cet algorithme, ses propriétés de convergence sont peu étudiées dans la littérature. Cet article étudie le comportement dynamique d'ADAM lorsque la fonction objective est non convexe et différentiable. Nous introduisons une version à temps continu d'ADAM sous la forme d'une équation différentielle ordinaire non-autonome (EDO). L'existence et l'unicité de la solution ainsi que la convergence de cette solution vers les points stationnaires de la fonction objective sont établis. Nous montrons aussi que le système à temps continu est une approximation pertinente des itérées d'ADAM dans le sens où le processus interpolé à partir d'ADAM converge faiblement vers la solution de l'EDO lorsque le pas tend vers zéro.

Abstract – ADAM is a popular variant of the stochastic gradient descent which aims at finding a local minimizer of a function. The objective function is unknown but a random estimate of the current gradient vector is observed at each round of the algorithm. Despite the popularity of this algorithm, only few information about its convergence properties is available in the literature. This paper investigates the dynamical behavior of ADAM when the objective function is non-convex and differentiable. We introduce a continuous-time version of ADAM, under the form of a non-autonomous ordinary differential equation (ODE). The existence and the uniqueness of the solution are established, as well as the convergence of the solution towards the stationary points of the objective function. It is also proved that the continuous-time system is a relevant approximation of the ADAM iterates, in the sense that the interpolated ADAM process converges weakly to the solution to the ODE as the step size parameter approaches zero.

1 Introduction

Considérons le problème de la recherche d'un minimiseur local d'une fonction de coût s'écrivant sous la forme d'une espérance $F(x) := \mathbb{E}(f(x,\xi))$ où $x \in \mathbb{R}^d$ et $f(\,.\,,\xi)$ est une fonction éventuellement non convexe qui dépend d'une variable aléatoire ξ. La distribution de ξ est supposée inconnue, mais révélée au cours du temps à travers l'observation de copies iid $(\xi_n : n \geq 1)$ de la variable aléatoire ξ. La descente de gradient stochastique (SGD) est l'algorithme le plus classique pour chercher un tel minimiseur. Des variantes de SGD introduisant un terme de momentum sont également devenues très populaires. Dans ces méthodes, le pas est généralement supposé constant ou décroissant. Ces algorithmes ont au moins deux limites. Tout d'abord, le choix du pas est en général difficile : un grand pas engendre de grandes fluctuations de l'estimateur alors qu'un petit pas entraîne une convergence lente. De plus, un pas commun est utilisé pour toutes les coordonnées en dépit des différences possibles entre les coordonnées du gradient.

Dans l'algorithme ADAM [1], le pas est ajusté coordonnée par coordonnée en utilisant les valeurs précédentes des carrés des coordonnées du gradient. L'algorithme combine les atouts des méthodes de momentum avec ceux de la sélection adaptative du pas par coordonnée. Enfin, l'algorithme comprend une étape de débiaisage agissant sur l'estimation courante du gradient et particulièrement utile pour les premières itérations. Malgré la popularité de cet algorithme, seuls quelques travaux s'intéressent à son comportement du point de vue théorique. Cet article étudie la convergence d'ADAM du point de vue des systèmes dynamiques.

Dans le paragraphe 2, nous rappelons l'algorithme ADAM en temps discret et nous exposons les principales hypothèses. Dans le paragraphe 3, nous introduisons une version en temps continu de l'algorithme ADAM sous la forme d'une équation différentielle ordinaire non-autonome (EDO). Nous présentons nos principaux résultats dans le paragraphe 4. Nous établissons l'existence et l'unicité de la solution de l'EDO, un résultat non trivial étant donné l'irrégularité du champ moyen. Nous montrons ensuite la convergence de cette solution vers l'ensemble des points stationnaires de la fonction objective F. Enfin, nous montrons que les itérées d'ADAM suivent le comportement de l'EDO non-autonome dans le régime asymptotique où le pas γ d'ADAM est petit. Plus précisément, nous considérons le processus interpolé $\mathsf{z}^\gamma(t)$ associé à la version discrète d'ADAM, qui consiste en une interpolation linéaire par morceaux des ité-

Algorithme 1 ADAM$(\gamma, \alpha, \beta, \varepsilon)$**.**

Entrées : données x_i, nombre d'itérations n_{iter}.
Paramètres : $\gamma > 0, \varepsilon > 0$, $(\alpha, \beta) \in [0,1)^2$.
Initialisation : $x_0 \in \mathbb{R}^d$, $m_0 = 0$, $v_0 = 0$.
for $n = 1$ **to** n_{iter} **do**
$m_n = \alpha m_{n-1} + (1-\alpha)\nabla f(x_{n-1}, \xi_n)$
$v_n = \beta v_{n-1} + (1-\beta)\nabla f(x_{n-1}, \xi_n)^2$
$\hat{m}_n = m_n/(1-\alpha^n)$ {étape de débiaisage}
$\hat{v}_n = v_n/(1-\beta^n)$ {étape de débiaisage}
$x_n = x_{n-1} - \gamma \hat{m}_n/(\varepsilon + \sqrt{\hat{v}_n})$.
end for

rées. Nous montrons que lorsque γ tend vers zéro, le processus interpolé z^γ converge faiblement [1] vers la solution de l'EDO non-autonome.

2 L'algorithme ADAM

Notations. Si x, y sont deux vecteurs de $\mathbb{R}^d$, on note xy, x/y, x^α, $|x|$ les vecteurs de $\mathbb{R}^d$ dont la k-ème coordonnée est donnée par $x_k y_k$, x_k/y_k, x_k^α, $|x_k|$. Les inégalités du type $x \leq y$ sont à lire coordonnée par coordonnée. Pour tout $v \in (0,+\infty)^d$, notons $\|x\|_v^2 = \sum_k v_k x_k^2$. Si (E, d) est un espace métrique, $z \in E$ et A est un ensemble non vide de E, on utilise la notation $\mathsf{d}(z, A) := \inf\{\mathsf{d}(z, z') : z' \in A\}$.

2.1 Les itérées

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité et soit $(\Xi, \mathfrak{S})$ un autre espace mesurable. Soit $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ une fonction mesurable, où d est un entier naturel. Pour une valeur fixée de ξ, l'application $x \mapsto f(x, \xi)$ est supposée différentiable et son gradient par rapport à x est noté $\nabla f(x, \xi)$. Nous définissons $\mathcal{Z} := \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ et $\mathcal{Z}_+ := \mathbb{R}^d \times \mathbb{R}^d \times [0, +\infty)^d$. ADAM génère une suite (x_n, m_n, v_n) sur $\mathcal{Z}_+$ (voir Algorithme 1).

Soit $z_n := (x_n, m_n, v_n)$ pour tout n. Cette suite satisfait : $z_n = T_{\gamma,\alpha,\beta}(n, z_{n-1}, \xi_n)$, pour tout $n \geq 1$, où pour tout $z = (x, m, v)$ dans $\mathcal{Z}_+$, $\xi \in \Xi$, nous définissons :
$T_{\gamma,\alpha,\beta}(n, z, \xi) :=$

$$\begin{pmatrix} x - \frac{\gamma(1-\alpha^n)^{-1}(\alpha m + (1-\alpha)\nabla f(x,\xi))}{\varepsilon + (1-\beta^n)^{-1/2}(\beta v + (1-\beta)\nabla f(x,\xi)^2)^{1/2}} \\ \alpha m + (1-\alpha)\nabla f(x,\xi) \\ \beta v + (1-\beta)\nabla f(x,\xi)^2 \end{pmatrix} . \quad (1)$$

2.2 Hypothèses

Hypothèse 2.1. L'application $f : \mathbb{R}^d \times \Xi \to \mathbb{R}$ satisfait les hypothèses suivantes :

i) Pour tout $x \in \mathbb{R}^d$, $f(x, \,.\,)$ est $\mathfrak{S}$-mesurable.

ii) Pour presque tout ξ, la fonction $f(\,.\,, \xi)$ est de classe C^1.

iii) Il existe $x_* \in \mathbb{R}^d$ tel que $\mathbb{E}(|f(x_*, \xi)|) < \infty$ et $\mathbb{E}(\|\nabla f(x_*, \xi)\|^2) < \infty$.

iv) Pour tout compact $K \subset \mathbb{R}^d$, il existe $L_K > 0$ tel que pour tout $(x, y) \in K^2$, $\mathbb{E}(\|\nabla f(x, \xi) - \nabla f(y, \xi)\|^2) \leq L_K^2 \|x - y\|^2$.

Sous l'hypothèse 2.1, il est facile de montrer que les fonctions $F : \mathbb{R}^d \to \mathbb{R}$ et $S : \mathbb{R}^d \to \mathbb{R}^d$, données par :

$$F(x) := \mathbb{E}(f(x, \xi)) \quad (2)$$

$$S(x) := \mathbb{E}(\nabla f(x, \xi)^2) \quad (3)$$

sont bien définies. De plus, F est de classe C^1 et par le théorème de convergence dominée, $\nabla F(x) = \mathbb{E}(\nabla f(x, \xi))$ pour tout x. Enfin, l'hypothèse 2.1 implique que ∇F et S sont des fonctions localement lipschitziennes. Concernant l'hypothèse 2.1-ii), notons que le cas d'une fonction $f(\,.\,, \xi)$ non différentiable (pour presque tout ξ) est aussi intéressant en pratique, mais l'étude est plus difficile et pourra faire l'objet d'un travail futur.

Hypothèse 2.2. F est coercive.

Hypothèse 2.3. Pour tout $x \in \mathbb{R}^d$, $S(x) > 0$.

Notons

$$\mathcal{S} := \nabla F^{-1}(\{0\})$$

l'ensemble des points critiques de F. Comme F est coercive et de classe C^1, l'ensemble $\mathcal{S}$ est non vide. L'hypothèse 2.3 signifie qu'il n'existe *pas* de point $x \in \mathbb{R}^d$ satisfaisant $\nabla f(x, \xi) = 0$ avec probabilité un. En pratique, cette hypothèse est légère.

2.3 Régime asymptotique

Notre objectif est d'étudier la suite $z_n = (x_n, m_n, v_n)$. Dans cet article, nous nous intéressons au régime du pas constant car le pas γ d'ADAM est fixé tout au long des itérations en pratique. Par conséquent, la suite $z_n^\gamma := z_n$ dépend du choix de γ (la valeur recommandée par défaut est $\gamma = 0.001$), et *ne* converge en général *pas* presque sûrement lorsque n tend vers l'infini. Afin d'établir un résultat de convergence sous l'hypothèse du pas constant, nous examinons le comportement asymptotique de la *famille* de processus $(n \mapsto z_n^\gamma)_{\gamma>0}$ indexée par γ, dans le régime où $\gamma \to 0$. Nous adoptons à cette fin la méthode de l'EDO, bien connue en approximation [2, 3]. Nous définissons le processus interpolé z^γ,comme la fonction linéaire par morceaux définie sur $[0, +\infty) \to \mathcal{Z}_+$ pour tout $t \in [n\gamma, (n+1)\gamma)$ par

$$\mathsf{z}^\gamma(t) := z_n^\gamma + (z_{n+1}^\gamma - z_n^\gamma)\left(\frac{t - n\gamma}{\gamma}\right) . \quad (4)$$

La méthode de l'EDO a pour but d'établir la convergence faible de la famille de processus aléatoires $(\mathsf{z}^\gamma)_{\gamma>0}$ lorsque γ tend vers zéro, vers un système déterministe à temps continu défini par une EDO. Introduite ci-dessous (Eq. (EDO)), elle représentera la version à temps continu d'ADAM .

Avant de décrire l'EDO, nous précisons le régime asymptotique dans lequel nous conduisons notre étude. En effet, contrairement à SGD, ADAM dépend de deux paramètres α et β en plus du pas γ. L'article [1] recommande d'utiliser des constantes α et β proches de un ($\alpha = 0.9$ et $\beta = 0.999$). Il est donc légitime de supposer que α et β tendent vers un, lorsque γ tend

1. dans l'espace des fonctions continues sur $[0, +\infty)$ muni de la topologie de la convergence uniforme sur les compacts.

vers zéro. Cela revient à considérer $\alpha := \bar{\alpha}(\gamma)$ et $\beta := \bar{\beta}(\gamma)$, où $\bar{\alpha}$ et $\bar{\beta}$ sont des fonctions sur $\mathbb{R}_+$ à valeurs dans $[0,1)$ telles que $\bar{\alpha}(\gamma)$ et $\bar{\beta}(\gamma)$ convergent vers un lorsque $\gamma \to 0$.

Nous supposons en plus l'hypothèse suivante.

Hypothèse 2.4. Les fonctions $\bar{\alpha} : \mathbb{R}_+ \to [0,1)$ et $\bar{\beta} : \mathbb{R}_+ \to [0,1)$ sont telles que les limites suivantes existent :

$$a := \lim_{\gamma \downarrow 0} \frac{1-\bar{\alpha}(\gamma)}{\gamma}, \quad b := \lim_{\gamma \downarrow 0} \frac{1-\bar{\beta}(\gamma)}{\gamma}\,. \tag{5}$$

De plus, $a > 0$ et $b > 0$, et la condition $b \leq 4a$ est satisfaite.

Cette dernière condition $b \leq 4a$ est compatible avec les paramètres par défaut recommandés par [1]. Dans notre modèle, nous remplaçons dorénavant l'application $T_{\gamma,\alpha,\beta}$ par $T_{\gamma,\bar{\alpha}(\gamma),\bar{\beta}(\gamma)}$.

Soit x_0 un élément fixé de $\mathbb{R}^d$. Pour tout $\gamma > 0$ fixé, nous définissons la suite (z_n^γ) générée par ADAM avec un pas fixé $\gamma > 0$:

$$z_n^\gamma := T_{\gamma,\bar{\alpha}(\gamma),\bar{\beta}(\gamma)}(n, z_{n-1}^\gamma, \xi_n)\,, \tag{6}$$

avec $z_0^\gamma = (x_0, 0, 0)$.

3 Le système à temps continu

Afin d'étudier le comportement de la suite (z_n^γ) définie par (6), nous réécrivons les itérations d'ADAM sous la forme équivalente suivante, pour tout $n \geq 1$:

$$z_n^\gamma = z_{n-1}^\gamma + \gamma h_\gamma(n, z_{n-1}^\gamma) + \gamma \Delta_n^\gamma\,, \tag{7}$$

où on définit pour tout $\gamma > 0$, $z \in \mathcal{Z}_+$,

$$h_\gamma(n,z) := \gamma^{-1}\mathbb{E}(T_{\gamma,\bar{\alpha}(\gamma),\bar{\beta}(\gamma)}(n,z,\xi) - z)\,, \tag{8}$$

et où $\Delta_n^\gamma := \gamma^{-1}(z_n^\gamma - z_{n-1}^\gamma) - h_\gamma(n, z_{n-1}^\gamma)$. Notons que (Δ_n^γ) est une suite d'incréments de martingale i.e. $\mathbb{E}(\Delta_n^\gamma | \mathcal{F}_{n-1}) = 0$ pour tout $n \geq 1$, où $\mathcal{F}_n$ désigne la tribu engendrée par les variables aléatoires $\xi_1, \ldots, \xi_n$. Définissons $h : (0,+\infty) \times \mathcal{Z}_+ \to \mathcal{Z}$ pour tout $t > 0$ et tout $z = (x,m,v)$ de $\mathcal{Z}_+$ par :

$$h(t,z) = \begin{pmatrix} -\frac{(1-e^{-at})^{-1} m}{\varepsilon + \sqrt{(1-e^{-bt})^{-1} v}} \\ a(\nabla F(x) - m) \\ b(S(x) - v) \end{pmatrix}, \tag{9}$$

où a, b sont les constantes définies dans l'hypothèse 2.4. Nous montrons que, pour tout (t,z) fixé, $h(t,z)$ coincide avec la limite de $h_\gamma(\lfloor t/\gamma \rfloor, z)$ quand $\gamma \downarrow 0$. Cette remarque ainsi que l'Eq. (7) suggèrent que le processus interpolé z^γ suit l'équation différentielle non-autonome quand $\gamma \downarrow 0$:

$$\dot{z}(t) = h(t, z(t))\,. \tag{EDO}$$

Formellement, nous démontrons que la famille des $(\mathsf{z}^\gamma : \gamma \in (0, \gamma_0])$ (où $\gamma_0 > 0$ est une constante fixée), interprétée comme une famille de variables aléatoires sur $C([0,+\infty), \mathcal{Z}_+)$ muni de la topologie de la convergence uniforme sur les compacts, converge faiblement vers la solution de l'(EDO) quand $\gamma \to 0$, sous des hypothèses techniques. Ainsi, l' (EDO) est une approximation légitime du comportement de z^γ quand γ est petit.

L'existence, l'unicité et la bornitude de la solution de l'(EDO) ne découlent pas trivialement de théorèmes connus. Ceci est dû à au moins deux difficultés : $h(\,.\,, z)$ n'est pas continue en zéro pour $z \in \mathcal{Z}_+$ fixé, et $h(t,\,.\,)$ n'est pas localement lipschitzienne, pour $t > 0$ fixé.

4 Principaux résultats

4.1 Temps continu : étude de l'ODE

Dans ce paragraphe, nous étudions l'équation différentielle non-autonome (EDO).

Définition 4.1. Soit $x_0 \in \mathbb{R}^d$. Une appplication continue $z : [0,+\infty) \to \mathcal{Z}_+$ est une solution globale de l'(EDO) avec condition initiale $(x_0,0,0)$ si z est de classe C^1 sur $(0,+\infty)$, si Eq. (EDO) est satisfaite pour tout $t > 0$, et si $z(0) = (x_0,0,0)$.

Théorème 4.1 (Existence et unicité)**.** Supposons que les hypothèses 2.1, 2.2, 2.3 et 2.4 sont satisfaites. Soit $x_0 \in \mathbb{R}^d$. L' (EDO) avec condition initiale $(x_0,0,0)$ admet une unique solution globale $z : [0,+\infty) \to \mathcal{Z}_+$. De plus, $z([0,+\infty))$ est un sous-ensemble borné de $\mathcal{Z}_+$.

Théorème 4.2 (Convergence)**.** Supposons que les hypothèses 2.1, 2.2, 2.3 et 2.4 sont satisfaites. Supposons de plus que $F(\mathcal{S})$ est d'intérieur vide. Soit $x_0 \in \mathbb{R}^d$ et soit $z : t \mapsto (x(t), m(t), v(t))$ la solution globale de l' (EDO) avec condition initiale $(x_0,0,0)$. Alors, l'ensemble $\mathcal{S}$ est non vide et

$$\lim_{t\to\infty} \mathsf{d}(x(t), \mathcal{S}) = 0\,.$$

De plus, $\lim_{t\to\infty} m(t) = 0$ et $\lim_{t\to\infty} S(x(t)) - v(t) = 0$.

Nous formulons à présent quelques commentaires. Dans la suite, nous noterons $z(t) = (x(t), m(t), v(t))$ l'unique solution globale de l' (EDO) issue de $(x_0,0,0)$.

Fonction de Lyapunov. La preuve du Th. 4.1 (voir [4, Paragraphe 5] pour plus de détails) repose sur l'existence d'une fonction de Lyapunov pour l'équation différentielle non- autonome (EDO). Nous rappelons qu'une fonction de Lyapunov est une fonction continue $V : (0,+\infty) \times \mathcal{Z}_+ \to \mathbb{R}$ telle que $t \mapsto V(t, z(t))$ est décroissante sur $(0,+\infty)$. Une telle fonction V est donnée par :

$$V(t,z) := F(x) + \frac{1}{2}\left\|m\right\|^2_{U(t,v)^{-1}}\,, \tag{10}$$

pour tout $t > 0$ et tout $z = (x,m,v)$ dans $\mathcal{Z}_+$, où $U : (0,+\infty) \times [0,+\infty)^d \to \mathbb{R}^d$ est une application donnée par :

$$U(t,v) := a(1-e^{-at})\left(\varepsilon + \sqrt{\frac{v}{1-e^{-bt}}}\right). \tag{11}$$

Décroissance de la fonction de coût à l'origine. Comme F n'est pas une fonction de Lyapunov pour l' (EDO), il n' y a pas de garantie de décroissance de $F(x(t))$ par rapport à t. Néanmoins, ce résultat est vrai à l'origine. En effet, on peut montrer que $\lim_{t\downarrow 0} V(t, z(t)) = F(x_0)$ (voir [4, Prop. 5.4]). Par conséquent,

$$\forall t \geq 0,\ F(x(t)) \leq F(x_0)\,. \tag{12}$$

La version en temps continu d'ADAM *ne peut* donc *qu'améliorer* le point initial x_0. Cette propriété est une conséquence de l'étape de débiaisage dans ADAM (voir Algorithme 1). Les

premières itérations de l'algorithme peuvent dégrader l'estimation initiale x_0 si le débiaisage n'est pas effectué.

Dérivées à l'origine. La preuve du Th. 4.1 montre aussi que $x(t)$ est de classe C^1 sur $[0, +\infty)$. La dérivée initiale est donnée par $\dot{x}(0) = -\nabla F(x_0)/(\varepsilon + \sqrt{S(x_0)})$ (voir [4, Lemme 5.1]). Il s'agit ici aussi d'une caractéristique remarquable d'ADAM. En l'absence de débiaisage, la dérivée initiale $\dot{x}(0)$ serait une fonction des paramètres initiaux m_0, v_0, des hyperparamètres qu'il faudrait alors bien choisir. Le débiaisage permet de s'affranchir de ces hyperparamètres, la derivée initiale étant naturellement fixée. Quand ε est petit et que la variance de $\nabla f(x_0, \xi)$ est faible (*i.e.*, $S(x_0) \simeq \nabla F(x_0)^2$), la dérivée initiale est environ égale à $-\nabla F(x_0)/|\nabla F(x_0)|$. Cela suggère que les premières itérations d'ADAM sont comparables à la variante de la descente de gradient avec le *signe*.

ADAM et Heavy Ball with Friction (HBF). Nous montrons à travers notre preuve que $x(t)$ est deux fois dérivable et satisfait pour tout $t > 0$,

$$c_1(t)\,\ddot{x}(t) + c_2(t)\dot{x}(t) + \nabla F(x(t)) = 0\,, \qquad (13)$$

où $c_1(t) := a^{-2}U(t, v(t))$ et $c_2(t)$, qui peut être explicité (l'expression est omise), satisfait $c_2(t) > \frac{\dot{U}(t,v(t)))}{2a^2}$ pour tout $t > 0$. Ainsi, l'Eq. (13) suggère que $x(t)$ peut être interprétée comme la solution d'un problème HBF généralisé où la masse de la particule et la viscosité dépendent du temps [5, 6, 7].

Hypothèses. Bien que l'on suppose vraies les hypothèses 2.1, 2.2, 2.3 et 2.4, les résultats de ce paragraphe ne sont pas spécifiques à la définition de F et S dans les Eq. (2)-(3). Les résultats sont vrais pour des applications arbitraires $F : \mathbb{R}^d \to \mathbb{R}^d$, $S : \mathbb{R}^d \to \mathbb{R}^d$, en supposant que F est de classe C^1 et S, ∇F sont localement lipschitziennes, au lieu de l'hypothèse 2.1.

4.2 Temps discret : convergence faible d'ADAM

Hypothèse 4.1. Pour tout compact $K \subset \mathbb{R}^d$, il existe $r_K > 0$ tel que

$$\sup_{x \in K} \mathbb{E}(\|\nabla f(x, \xi)\|^{2+r_K}) < \infty\,.$$

Hypothèse 4.2. La suite $(\xi_n : n \geq 1)$ est une suite de copies iid de ξ.

Théorème 4.3. Supposons les hypothèses 2.1, 2.2, 2.3, 2.4, 4.1, 4.2 vraies. Soit $x_0 \in \mathbb{R}^d$. Pour tout $\gamma > 0$, soit $(z_n^\gamma : n \in \mathbb{N})$ la suite aléatoire définie par les itérations d'ADAM (6) et $z_0^\gamma = (x_0, 0, 0)$. Soit z^γ le processus interpolé correspondant, défini par l'Eq. (4). Enfin, notons z l'unique solution globale de l'(EDO) issue de $(x_0, 0, 0)$. Alors,

$$\forall T > 0,\ \forall \delta > 0,\ \lim_{\gamma \downarrow 0} \mathbb{P}\left(\sup_{t \in [0,T]} \|\mathsf{z}^\gamma(t) - z(t)\| > \delta \right) = 0\,.$$

Le théorème 4.3 signifie que la famille de processus aléatoires $(\mathsf{z}^\gamma :\ \gamma > 0)$ converge en probabilité vers l'unique solution de l' (EDO) issue de $(x_0, 0, 0)$, quand $\gamma \downarrow 0$. La convergence en probabilité est établie dans l'espace $C([0, +\infty), \mathcal{Z}_+)$ des fonctions continues sur $[0, +\infty)$ muni de la topologie de la convergence uniforme sur les compacts. Ainsi, le système non-autonome (EDO) est une approximation pertinente du comportement des itérées $(z_n^\gamma : n \in \mathbb{N})$ pour un petit pas γ.

Remarque 1. Lorsque le pas γ est constant, la suite (x_n^γ) ne converge pas au sens presque sûr quand $n \to \infty$. La convergence ne peut avoir lieu que dans le double régime asymptotique où $n \to \infty$ puis $\gamma \to 0$. Pour les systèmes autonomes, un tel régime a été étudié dans les travaux [8, 9]. En particulier, [9] suggère que l'on peut espérer le comportement en temps long suivant :

$$\forall \delta > 0,\ \lim_{\gamma \downarrow 0} \limsup_{n \to \infty} \frac{1}{n+1} \sum_{k=0}^{n} \mathbb{P}(\mathsf{d}(x_n^\gamma, \mathcal{S}) > \delta) = 0\,. \quad (14)$$

Cependant, pour montrer (14), l'étude de la convergence en temps long menée dans [9] doit être revisitée pour le cas des chaînes de Markov non homogènes. Cette généralisation est possible mais dépasse le cadre de cet article. Nous laissons la preuve de (14) pour de futurs travaux.

Références

[1] D. P. Kingma and J. Ba. Adam : A method for stochastic optimization. International Conference on Learning Representations (ICLR), 2015

[2] Kushner, H. J. and Yin, G. G. *Stochastic approximation and recursive algorithms and applications*. Applications of Mathematics (New York), vol. 35, Springer-Verlag, New York, 2003.

[3] M. Benaïm and S. J. Schreiber. Ergodic properties of weak asymptotic pseudotrajectories for semiflows. Journal of Dynamics and Differential Equations, vol. 12, n.3, p. 579–598, 2000

[4] A. Barakat and P. Bianchi. Convergence of the ADAM algorithm from a Dynamical System Viewpoint. arXiv preprint arXiv :1810.02263, 2018.

[5] H. Attouch, X. Goudou and P. Redont. The heavy ball with friction method, I. The continuous dynamical system : global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. Communications in Contemporary Mathematics, vol. 2, n. 01, p. 1–34, 2000.

[6] A. Cabot, H. Engler and S. Gadat. On the long time behavior of second order differential equations with asymptotically small dissipation. Transactions of the American Mathematical Society, vol. 361, n. 11, p. 5983–6017, 2009.

[7] S. Gadat, F. Panloup and S. Saadane. Stochastic heavy ball. Electronic Journal of Statistics, vol. 12, n. 1, p. 461–529, 2018

[8] J.-C. Fort and G. Pagès. Asymptotic behavior of a Markovian stochastic algorithm with constant step. SIAM Journal on Control and Optimization, vol. 37, n. 5, p. 1456–1482, 1999.

[9] P. Bianchi, W. Hachem and A. Salim. Constant Step Stochastic Approximations Involving Differential Inclusions : Stability, Long-Run Convergence and Applications. arXiv preprint arXiv :1612.03831, 2016