ANNALES DE L'I. H. P., SECTION B

ODILE BRANDIÈRE
MARIE DUFLO

Les algorithmes stochastiques contournent-ils les pièges ?

Annales de l'I. H. P., section B, tome 32, n° 3 (1996), p. 395-427
<http://www.numdam.org/item?id=AIHPB_1996__32_3_395_0>

Ann. Inst. Henri Poincaré,

Vol. 32, n° 3, 1996, p. 395-427 *Probabilités et Statistiques*

Les algorithmes stochastiques contournent-ils les pièges ?

par

Odile BRANDIÈRE et Marie DUFLO

Université de Marne-la-Vallée, Équipe d'analyse et de mathématiques appliquées,
2, rue de la Butte Verte, 93166 Noisy-le-grand Cedex.

RÉSUMÉ. – On étudie ici la validité d'affirmations communément admises : « un algorithme du gradient converge vers l'un des minima locaux » ou « un algorithme stochastique ne peut converger que vers un point asymptotiquement stable de l'équation différentielle associée ». On prouve qu'un algorithme contourne un piège régulier dès que le bruit est excitant dans une direction répulsive.

ABSTRACT. – Do stochastic algorithms fall into traps? We are considering the validity of commonly held claims: "a stochastic gradient algorithm only converges towards one of the local minima" or "a stochastic algorithm only converges towards an asymptotically stable solution of the associated differential equation". We prove that a stochastic algorithm does not fall into a regular trap if the noise is exciting in a repulsive direction.

INTRODUCTION

Cibles et pièges

L'approximation stochastique est essentiellement l'étude du comportement asymptotique d'une suite de vecteurs aléatoires à valeurs dans G,

A.M.S. Classification : 62 L 20, 60 F 99

ouvert convexe de $\mathbb{R}^d$, (Z_n), satisfaisant à la relation récursive perturbée :

$$Z_{n+1} = Z_n + \gamma_n h(Z_n) + \eta_{n+1}, \qquad \text{[ER]}$$

où h est une fonction continue de G dans $\mathbb{R}^d$, (η_n) une suite de « petites » perturbations; (γ_n) sera toujours une suite déterministe, positive, satisfaisant à :

$$\sum \gamma_n = \infty \qquad \text{et} \qquad \sum \gamma_n^2 < \infty.$$

Depuis l'article historique de Robbins-Monro [30] et jusqu'à des travaux plus récents, en automatique et en théorie des neurones notamment, on connaît l'intérêt d'algorithmes pour lesquels on n'a pas un choix libre de la perturbation, la variable $h(Z_n)$ n'étant observée (ou facilement calculable) qu'à un « bruit » aléatoire ξ_{n+1} près : alors $\eta_{n+1} = \gamma_n \xi_{n+1}$.

Si $\Delta = \{z;\ h(z) = 0\}$, on souhaite en général obtenir la convergence de l'algorithme vers une cible $z^* \in \Delta^*$, $\Delta^* \subseteq \Delta$.

Il existe de nombreuses méthodes visant à s'assurer que la distance de (Z_n) à l'une des composantes connexes de Δ tend, p.s., vers 0 (*cf.* [5], [8], [22], [25], ...).

Dans cet article, on se place dans l'un de ces cadres en supposant les points de $\Delta \backslash \Delta^*$ isolés; prouver que l'algorithme contourne les pièges revient alors à prouver que, pour tout $z^* \in \Delta \backslash \Delta^*$, l'événement $\Gamma(z^*) = \{(Z_n) \text{ tend vers } z^*\}$ est négligeable.

Minima d'un potentiel

Par exemple, lorsque U est un potentiel, fonction de classe C^1 de G dans $\mathbb{R}_+$ de gradient ∇U, on recherche les minima ($\Delta^* = \text{Argmin } U$) – ou les minima locaux ($\Delta^* = \text{Argminloc } U$) – de U par un algorithme du gradient :

$$Z_{n+1} = Z_n - \gamma_n \nabla U(Z_n) + \eta_{n+1}. \qquad \text{[GR]}$$

Les pièges de [GR] sont les zéros de ∇U qui ne sont pas dans Δ^*.

Afin d'atteindre les minima de U, plusieurs auteurs ont exploré les méthodes du recuit simulé, notamment Kushner [16], Pflug [22], Gelfand-Mitter ([9] et [10]), Hwang-Sheu [13]. Ainsi, Gelfand et Mitter [9] ajoutent à la perturbation $\gamma_n \xi_{n+1}$ décrite ci-dessus une excitation auxiliaire indépendante, obtenue par simulation :

$$\eta_{n+1} = \gamma_n \xi_{n+1} + s_n W_{n+1},$$

où (W_n) est une suite de vecteurs aléatoires gaussiens indépendants et de même loi, de covariance inversible, et (s_n) une suite décroissant lentement

vers 0; sous des conditions assez générales, (Z_n) converge en probabilité vers Argmin U dès que l'on choisit

$$\gamma_n = 1/n, \quad s_n = c\,(n \operatorname{Log} \operatorname{Log} n)^{-1/2},$$

c positif assez grand.

Si possible, on préfère toutefois éviter d'alourdir l'algorithme par recuit simulé. On sait alors que l'algorithme [GR] ne peut pas éviter les minima locaux (la décroissance trop rapide de (γ_n) empêche de sortir des puits entourant les minima locaux); il s'agit ici d'étudier s'il évite les autres pièges, points selles et maxima locaux.

La méthode de l'équation différentielle

Une idée féconde introduite notamment par Derevitskii-Fradkov [6], Ljung [21] et Kushner-Clark [17] est de comparer, sur l'ensemble des trajectoires telles que $\sum \eta_{n+1}$ converge, [ER] à l'équation différentielle

$$\frac{dz\,(t)}{dt} = h\,(z\,(t)), \qquad \text{[ED]}$$

z étant une fonction de classe C^1 de $\mathbb{R}_+$ dans $\mathbb{R}^d$.

Les cibles de l'algorithme sont en général les zéros de h asymptotiquement stables pour [ED] sur lesquels nous reviendrons en I.1.

Selon le théorème de Kushner-Clark [17], si $z^* \in \Delta^*$, alors la suite (Z_n) tend vers z^* dès qu'elle est bornée et revient infiniment souvent dans un compact d'un bassin d'attraction de z^*. Ce résultat a été souvent utilisé en automatique dans des exemples où $\mathbb{R}^d$ tout entier est bassin d'attraction d'un unique point de Δ^* ([2], [21], [23], ...). Mais, dans le cas général, il n'est pas forcément plus facile de vérifier la récurrence dans un compact d'un domaine d'attraction de $z^* \in \Delta^*$ (baptisée parfois « convergence au sens de Kushner-Clark ») que la convergence vers la cible z^*. Il s'agit de lever cette difficulté.

Une formulation précise du cadre que nous étudions sera donnée en I.1. et I.2. Le résultat principal est le théorème 1 énoncé en I.2. Le même problème a été étudié par Nevel'son-Has'minskii [25] (chapitre 5, p. 113-121) pour les pièges que nous appellerons « répulsifs », et récemment par Lazarev [19] pour les pièges généraux. Notre apport essentiel par rapport à ces travaux tient à la structure plus générale de la perturbation : ce point sera précisé en I.5 et confirmé par des exemples en II.

Remerciements

Les auteurs sont très reconnaisants au rapporteur et à Pierre Priouret dont les commentaires éclairés ont favorisé une amélioration significative de ce travail.

I. COMMENT LE BRUIT PERMET-IL DE CONTOURNER LES PIÈGES ?

On étudie un algorithme stochastique défini sur un espace de probabilité $(\Omega, \mathcal{A}, P)$ à valeurs dans $\mathbb{R}^d$ de la forme

$$Z_{n+1} = Z_n + \gamma_n h(Z_n) + \eta_{n+1} \qquad \text{[ER]}$$

sur l'ensemble des trajectoires $\Gamma(z^*) = \{(Z_n) \to z^*\}$ où z^* est un « piège régulier », notion dont le sens précis est donné en I.1. Pour un vecteur v de $\mathbb{R}^d$, on désigne aussi par v la matrice colonne associée; $\| \cdot \|$ est la norme euclidienne. Si A est une matrice, ${}^T A$ est sa transposée; I (ou I_d pour préciser la dimension) est la matrice identité.

I.1. Présentation des pièges réguliers

Afin de préciser la nature des « pièges », considérons d'abord l'équation différentielle [ED] définie ci-dessus. Le cas le plus simple est le cas linéaire où $h(z) = H(z - z^*)$, H matrice $d \times d$:

$$\frac{dz(t)}{dt} = H[z(t) - z^*]. \qquad (1)$$

RAPPELS RELATIFS À L'ÉQUATION DIFFÉRENTIELLE LINÉAIRE (1)

La solution de (1) est :

$$z(t) - z^* = \exp(tH)[z(0) - z^*].$$

Ses propriétés sont simples et éclairent le problème posé.

a) Si $\underline{\lambda}(H)$ et $\bar{\lambda}(H)$ sont respectivement la plus petite et la plus grande partie réelle des valeurs propres de H, les deux cas suivants sont les plus simples.

□ $\bar{\lambda}(H) < 0$ (z^* est attractif). Alors toute solution de (1) tend vers z^* si $t \to \infty$ et z^* est un point asymptotiquement stable pour l'équation différentielle (1).

□ $\underline{\lambda}(H) > 0$ (z^* est répulsif). Alors, toute solution de (1) telle que $z(0) \neq z^*$ satisfait à $\| z(t) - z^* \| \to \infty$ si $t \to \infty$.

b) Supposons $\bar{\lambda}(H) > 0 \geq \underline{\lambda}(H)$. Le polynôme minimal de H est le produit de deux polynômes M_+ et M_- dont les zéros ont respectivement des parties réelles > 0 et ≤ 0; $\mathbb{R}^d$ est la somme directe de $K_+(H) = \mathrm{Ker}[M_+(H)]$ et de $K_-(H) = \mathrm{Ker}[M_-(H)]$ où $K_+(H)$ est

l'ensemble des directions répulsives, $K_-(H)$ est l'ensemble des directions non répulsives.

Soit P^{-1} une matrice de changement de base dont les premiers vecteurs colonnes forment une base de $K_+(H)$ et les suivants une base de $K_-(H)$:

$$PH\,P^{-1} = \begin{bmatrix} J_+ & 0 \\ 0 & J_- \end{bmatrix} \quad \text{où} \quad \underline{\lambda}(J_+) > 0,\ \bar{\lambda}(J_-) \leq 0;$$

pour

$$y(t) = P\,z\ (t) = \begin{bmatrix} y_+(t) \\ y_-(t) \end{bmatrix} \qquad \text{et} \qquad y^* = P\,z^*,$$

$$\frac{d}{dt}\,y_\mp(t) = J_\mp\,(y_\mp(t) - y^*_\mp). \tag{2}$$

D'après *a*), une solution bornée de (1) satisfait à $y_+(t) = y^*_+$ pour tout t; son orbite se trouve dans le sous-espace affine

$$y^* + \mathrm{Ker}\,[M_-(H)].$$

HYPOTHÈSES DE RÉGULARITÉ

Afin de comparer, au voisinage de z^*, [ER] ou [ED] à (1), on fait les hypothèses suivantes.

[H1] Régularité de la fonction h au voisinage de z^*.

La fonction h n'est étudiée qu'au voisinage d'un « zéro régulier de h » z^*, c'est-à-dire d'un point $z^* \in \mathbb{R}^d$ tel que :

- $h(z^*) = 0$;
- *il existe un voisinage ouvert $V(z^*)$ de z^* dans lequel h est différentiable avec une différentielle lipschitzienne.*

On note : $H = Dh(z^*)$. On peut choisir $V(z^*)$ convexe; alors, sous [H1], si $z \in V(z^*)$,

$$h(z) = \int_0^1 Dh\,(tz + (1-t)\,z^*)\,dt\,[z - z^*] = [Dh(z^*) + R(z)]\,[z - z^*]$$

avec

$$R(z) = \int_0^1 Dh\,(tz + (1-t)\,z^*)\,dt - Dh(z^*),$$

$$\| R(z) \| \leq \mathrm{cte}\,\| z - z^* \|. \tag{3}$$

DÉFINITION. – *Sous l'hypothèse* [H1], *le point* z^* *est un* piège *lorsque* $\bar{\lambda}(H) > 0$. *C'est un piège* répulsif *si* $\underline{\lambda}(H) > 0$.

I.2. Hypothèses relatives à l'algorithme

Les hypothèses suivantes relatives aux pas et aux perturbations aléatoires sont classiques.

[H2] Propriétés des perturbations

On suppose que la perturbation (η_n) *est de la forme :*

$$\eta_{n+1} = c_n [r_{n+1} + \varepsilon_{n+1}].$$

[H2*a*] *Les suites* (ε_n) *et* (r_n) *sont des vecteurs aléatoires de dimension* d *définis sur l'espace de probabilité* $(\Omega, \mathcal{A}, P)$ *adaptés à une filtration* $\mathbb{F} = (\mathcal{F}_n)$; Z_0 est $\mathcal{F}_0$*-mesurable et l'on a les propriétés suivantes,* p.s. *sur* $\Gamma(z^*) = \{\omega;\ (Z_n(\omega)) \to z^*\}$:

- $\sum \| r_n \|^2 < \infty$;
- $\limsup_n E(\| \varepsilon_{n+1} \|^2 \,|\, \mathcal{F}_n) < \infty$ *et* $E(\varepsilon_{n+1} \,|\, \mathcal{F}_n) = 0$;

(*cette propriété s'exprime par* « *sur* $\Gamma(z^*)$, (ε_n) *est un* bruit adapté à $\mathbb{F}$ *ayant un moment conditionnel d'ordre* 2 *fini* »).

[H2*b*] Excitation dans une direction répulsive de $H = Dh(z^*)$.

Notant $\varepsilon_n^{(r)}$ *la projection de* ε_n *sur* $K_+(H)$ *parallèlement à* $K_-(H)$,

$$\liminf E(\| \varepsilon_{n+1}^{(r)} \| \,|\, \mathcal{F}_n) > 0, \text{ p.s. } sur\ \Gamma(z^*).$$

[H3] Propriétés des pas

Les suites (γ_n) *et* (c_n) *sont des suites déterministes et positives, choisies de telle sorte que,*

$$\gamma_n = 0(c_n), \sum \gamma_n = \infty, \sum c_n^2 < \infty.$$

Dans les paragraphes I.3 et I.4 nous prouverons le théorème suivant qui est le point principal de cet article.

THÉORÈME 1. – *Sous les hypothèses* [H1, H2, H3], *si le point* z^* *est un piège, l'événement* $\Gamma(z^*)$ *est négligeable.*

HEURISTIQUE

Le contenu de ce théorème est intuitif. L'excitation du bruit dans une direction répulsive interdit à l'algorithme ce qui était possible pour une solution de [ED] : converger vers un piège répulsif en restant dans un sous-espace attractif.

REMARQUES RELATIVES [H2] ET [H3]

1) *Version simplifiée de la propriété d'excitation*

Lorsque le bruit a, sur $\Gamma(z^*)$, un moment conditionnel d'ordre $a > 2$, c'est-à-dire :

$$\limsup E(\| \varepsilon_{n+1} \|^a \mid \mathcal{F}_n) < \infty, \text{ p.s. sur } \Gamma(z^*),$$

on a, par l'inégalité de Hölder,

$$E(\| \varepsilon_{n+1}^{(r)} \|^2 \mid \mathcal{F}_n) \leq (E(\| \varepsilon_{n+1}^{(r)} \| \mid \mathcal{F}_n))^{[a-2]/[a-1]} (E(\| \varepsilon_{n+1}^{(r)} \|^a \mid \mathcal{F}_n))^{1/[a-1]}.$$

On peut donc remplacer la condition [H2*b*] par la condition suivante, souvent plus maniable : p.s. sur $\Gamma(z^*)$,

$$\liminf E(\| \varepsilon_{n+1}^{(r)} \|^2 \mid \mathcal{F}_n) > 0;$$

c'est en particulier le cas si :

$$\liminf \lambda_{\min} E(\varepsilon_{n+1}{}^T \varepsilon_{n+1} \mid \mathcal{F}_n) > 0, \text{ p.s. sur } \Gamma(z^*).$$

2) *Un outil destiné à simplifier les démonstrations*

Dans la suite, il s'agit de prouver $P(\Gamma(z^*)) = 0$.

Il suffira de faire les démonstrations en supposant que, pour trois constantes $K < \infty$, $A > 0$ et $B < \infty$, on a :

$$\sum \| r_{n+1} \|^2 \leq K, \tag{4}$$

et, quel que soit n, p.s. :

$$0 < A \leq E(\| \varepsilon_{n+1}^{(r)} \| \mid \mathcal{F}_n) \qquad \text{et} \qquad E(\| \varepsilon_{n+1} \|^2 \mid \mathcal{F}_n) \leq B. \tag{5}$$

Pour obtenir (4), il suffit de remarquer que $\Gamma(z^*)$ est la réunion de ses intersections avec $B_K = \{\sum_n \| r_{n+1} \|^2 \leq K\}$ pour K entier, puis de remplacer r_n par $\tilde{r}_n = r_n$ si $\sum_{j=0} \| r_j \|^2 \leq K$, $\tilde{r}_n = 0$ sinon. Un théorème prouvé avec la suite $(\tilde{r}_n)$ le sera, sur B_K, avec (r_n).

Pour obtenir la réduction (5), on utilise l'astuce suivante due à Lai et Wei [18]. Soit, sur un espace de probabilité $(\Omega', \mathcal{A}', P')$ une suite (δ_n) de vecteurs aléatoires de dimension d indépendants et de même loi, bornés et centrés. On définit sur $(\Omega \times \Omega', \mathcal{A} \otimes \mathcal{A}', P \otimes P')$ la filtration $\tilde{\mathbb{F}} = (\tilde{\mathcal{F}}_n)$ avec $\tilde{\mathcal{F}}_n = \mathcal{F}_n \otimes \sigma(\delta_j;\ j \leq n)$. Soit :

$$\Gamma_n = \{A \leq E(\| \varepsilon_{n+1}^{(r)} \| \mid \mathcal{F}_n\},\ E(\| \varepsilon_{n+1} \|^2 \mid \mathcal{F}_n\} \leq B\}.$$

Pour tout $\alpha > 0$, on peut trouver A, B et un entier p tels que :

$$P(\Gamma(z^*) \setminus \bigcap_{n \geq p} \Gamma_n) \leq \alpha.$$

Posons :

$$\tilde{\varepsilon}_{n+1}(\omega, \omega') = \varepsilon_{n+1}(\omega) 1_{\Gamma_n}(\omega) + \delta_{n+1}(\omega') 1_{\Gamma_n^c}(\omega).$$

La suite $(\tilde{\varepsilon}_{n+p})$ est adaptée à $(\tilde{\mathcal{F}}_{n+p})$ et satisfait à (5). L'algorithme [ER] dans lequel on remplace ε_{n+p} par $\tilde{\varepsilon}_{n+p}$ pour $n \geq 0$ coïncide avec l'algorithme étudié sur $\bigcap_{n \geq p} \Gamma_n$. Si l'on prouve que les pièges sont, p.s., évités sous la condition (5) pour tous A, B, p, alors $P(\Gamma(z^*)) \leq \alpha$ pour tout α et $P(\Gamma(z^*)) = 0$.

Les relations (4) et (5) seront donc utilisées dans les démonstrations sans restreindre le cadre des théorèmes.

3) Sous [H2-3], $\sum c_n^2 < \infty$ et les séries $\sum c_n^2 \| \varepsilon_{n+1} \|^2$ et $\sum c_n \varepsilon_{n+1}$ convergent p.s. sur $\Gamma(z^*)$. Il suffit de voir, si (5) est satisfaite, que l'espérance de la première est finie et que la seconde est une martingale de carré intégrable convergente p.s.

4) L'hypothèse [H3] est la condition courante des pas décroissants pour des algorithmes stochastiques.

I.3. Piège répulsif

LEMME 2. – *Soit H une matrice complexe* répulsive *dont la plus petite des parties réelles des valeurs propres $\underline{\lambda}(H)$ est > 0.*

Si $(\varphi_n)_{n \geq 0}$ est une suite à valeurs dans $\mathbb{C}^d$, bornée et satisfaisant à une relation récursive :

$$\varphi_{n+1} = (I + \gamma_n H) \varphi_n + \xi_{n+1},$$

alors, pour p assez grand et $B_n = (I + \gamma_n H) \dots (I + \gamma_p H)$, la série $\sum_{j=p}^{\infty} B_j^{-1} \xi_{j+1}$ converge vers $[-\varphi_p]$.

Démonstration – Soit L un réel tel que $0 < L < \underline{\lambda}(H)$. Il existe une norme de $\mathbb{C}^d$, $| \cdot |$, telle que, si γ est un réel positif assez petit ($\gamma \leq \bar{\gamma}$), pour tout vecteur v de $\mathbb{C}^d$:

$$|(I + \gamma H) v| \geq (1 + \gamma L) |v|.$$

On le voit facilement par un changement de base transformant H en une matrice dont les termes de la diagonale sont les valeurs propres, ceux situés

immédiatement sous la diagonale égaux à 0 ou à $t > 0$ arbitrairement petit, les autres termes étant nuls.

Puisque la suite (γ_n) tend vers 0, on peut choisir l'entier p tel que, pour $n \geq p$, $\gamma_n \leq \bar{\gamma}$ et :

$$|B_n v| \geq (1 + L\gamma_n) \dots (1 + L\gamma_p) |v|;$$

autrement dit $|B_n^{-1}| \leq [(1 + L\gamma_n) \dots (1 + L\gamma_p)]^{-1}$.

Comme $\sum \gamma_n = \infty$, $(1 + L\gamma_n) \dots (1 + L\gamma_p) \to \infty$.

Le lemme résulte alors de l'égalité :

$$\varphi_n = B_{n-1} \left(\varphi_p + \sum_{j=p}^{n-1} B_j^{-1} \xi_{j+1} \right). \quad \blacksquare$$

Avant d'examiner le cas multidimensionnel, voici une démonstration plus simple valable dans le cas unidimensionnel.

LEMME 3. – *Piège répulsif unidimensionnel. Avec les hypothèses* [H1-2-3], *on suppose que* $d = 1$ *et que* $h'(z^*) = \lambda > 0$. *Alors* $P(\Gamma(z^*)) = 0$.

Démonstration. – *a*) Il existe, sur un voisinage W de 0, une fonction φ dérivable avec une dérivée lipschitzienne telle que

$$\varphi'(z)\, h(z + z^*) = \lambda \varphi(z). \tag{6}$$

La fonction φ doit satisfaire à $\varphi(0) = 0$, $\varphi'(0) = 1$.

Posons $\varphi(z) = z \exp[G(z)]$ avec $G(0) = 0$; l'équation (6) devient

$$[h(z + z^*)][1 + z\, G'(z)] = \lambda z.$$

En tenant compte de (3), pour $z \neq 0$,

$$G'(z) = g(z) = -R(z + z^*)/(z[\lambda + R(z + z^*)]).$$

La fonction g est continue et bornée au voisinage de 0 sauf éventuellement en 0; la fonction $G(z) = \int_0^z g(u)\, du$ convient.

b) Soit $V_1(z^*)$ un voisinage de z^* supposé ouvert, convexe et contenu dans $V(z^*) \cap \{z;\ z - z^* \in W\}$. Considérons pour p entier :

$$\Gamma_p = \{\omega;\ \omega \in \Gamma(z^*) \quad \text{et} \quad Z_n(\omega) \in V_1(z^*) \text{ pour tout } n \geq p\}.$$

Il suffit de prouver que Γ_p est négligeable quel que soit p.

Or, sur Γ_p, pour $n \geq p$:

$$\varphi(Z_{n+1} - z^*) = \varphi(Z_n - z^*) + \varphi'(Z_n - z^*)[Z_{n+1} - Z_n] + \rho_{n+1},$$

où $\rho_{n+1} = O([Z_{n+1} - Z_n]^2)$. Posons $\varphi_n = \varphi(Z_n - z^*)$:

$$\varphi_{n+1} = (1 + \lambda\gamma_n)\,\varphi_n + \xi_{n+1},$$

$$\xi_{n+1} = \rho_{n+1} + c_n\,\varphi'(Z_n - z^*)[r_{n+1} + \varepsilon_{n+1}].$$

Or ρ_{n+1} est $\mathcal{F}_{n+1}$-mesurable et, sur Γ_p, $\rho_{n+1} = O(\gamma_n^2 + c_n^2 \,\|\, \varepsilon_{n+1} \,\|^2)$;

$$\xi_{n+1} = c_n\,\varphi'(Z_n - z^*)\varepsilon_{n+1} + c_n\, r_{n+1}^1,$$

où, p.s. sur $\Gamma(z^*)$, $\sum (r_{n+1}^1)^2 < \infty$ et $(\varphi'(Z_n - z^*)\,\varepsilon_{n+1}) = (e_{n+1})$ satisfait aux mêmes hypothèses que (ε_{n+1}).

D'après le lemme 2 appliqué à la suite $(\varphi_{n+p})_{n\geq 0}$ et le théorème A de l'appendice $P(\Gamma_p) = 0$. ■

PROPOSITION 4. – *Piège répulsif multidimensionnel. On donne une matrice $d \times d$ à coefficients réels et répulsive, H.*

On considère un algorithme à valeurs dans $\mathbb{R}^d$:

$$Z_{n+1} = (I + \gamma_n H_n)\, Z_n + c_n\,[\varepsilon_{n+1} + r_{n+1}].$$

On fait les hypothèses [H3] *sur les pas.*

On considère un ensemble de trajectoires noté Λ sur lequel les propriétés énoncées en [H2*a*] *et* [H2*b*] *sont vérifiées.*

On suppose enfin que (H_n) est une suite de matrices aléatoires $d \times d$ adaptée à $\mathbb{F}$. Alors, $\Gamma = \Lambda \cap \{(Z_n)$ tend vers 0 et (H_n) tend vers $H\}$ est négligeable.

Remarque. – Il est facile de voir que, selon cette proposition, un piège répulsif régulier est contourné.

En effet, si z^* est un tel piège, et si $H = Dh(z^*)$, on choisit le voisinage $V(z^*)$ de l'hypothèse [H1] ouvert, convexe, et tel que, pour $z \in V(z^*)$,

$$h(z) = [H + R(z)]\,(z - z^*),$$

avec, selon (3), $\| R(z) \| \leq \text{Cte} \,\| z - z^* \|$.

Sur $\Gamma_p = \{\omega;\ \omega \in \Gamma(z^*),\ Z_n(\omega) \in V(z^*)$ pour $n \geq p\}$, on a, pour $n \geq p$,

$$(Z_{n+1} - z^*) = (I + \gamma_n H_n)\,(Z_n - z^*) + c_n\,[\varepsilon_{n+1} + r_{n+1}],$$

$$\| H_n - H \| = \| R(Z_n) \| \leq \text{Cte} \,\| Z_n - z^* \|.$$

La proposition 4 prouvera ainsi que Γ_p est négligeable, donc que $\Gamma(z^*) \cup \Gamma_p$ est négligeable.

Démonstration de la proposition 4. – La remarque 2) de I.2 s'applique et l'on suppose effectuées les réductions (4) et (5).

a) *Préliminaire.* – La matrice H étant répulsive, il existe, selon un lemme de Lyapounov, une matrice Q symétrique et définie positive telle que $QH + {}^T HQ = 2I$. Pour y et z dans $\mathbb{R}^d$,

$$ {}^T(y+z)\,Q\,y \leq ({}^T(y+z)\,Q\,(y+z)\,{}^T y\,Q\,y)^{1/2},$$

$$ ({}^T(y+z)\,Q\,(y+z))^{1/2} - ({}^T y\,Q\,y)^{1/2} \geq ({}^T y\,Q\,z)\,({}^T y\,Q\,y)^{-1/2}.$$

Posant $U_n = ({}^T Z_n\,QZ_n)^{1/2}$ et $\Delta_n = H_n - H$:

$$ Z_{n+1} = (I + \gamma_n H)\,Z_n + \gamma_n\,\Delta_n\,Z_n + c_n\,(\varepsilon_{n+1} + r_{n+1}),$$

$$ U_{n+1} - U_n \geq \gamma_n\,[\| Z_n \|^2/U_n] + \gamma_n\,{}^T Z_n\,Q\,\Delta_n\,Z_n/U_n + c_n\,[{}^T Z_n\,Q\,(\varepsilon_{n+1} + r_{n+1})/U_n];$$

$$ [\| Z_n \|^2 + {}^T Z_n\,Q\,\Delta_n\,Z_n]/U_n \geq \| Z_n \|\,(1 - \| Q\,\Delta_n \|)/[\lambda_{\max} Q]^{1/2}.$$

L'ensemble Γ est la réunion de ses intersections Γ_p avec

$$ \{\sup_{n\geq p} \| Q\,\Delta_n \| \leq 1/2 \quad \text{et} \quad \sup_{n\geq p} \| Z_n \| \leq 1\}.$$

Sur Γ_p, pour $M \geq n \geq p$ et $K = 2\,[\lambda_{\max} Q]^{1/2}$:

$$ \sum_{j=n}^{M} \gamma_j \| Z_j \| \leq KU_{M+1} - K \sum_{j=n}^{M} c_j\,[{}^T Z_j\,Q\,(\varepsilon_{j+1} + r_{j+1})/U_j].$$

Posons $\alpha_n^2 = \sum_{j=n}^{\infty} c_j^2$.

Avec les réductions (4) et (5), par l'inégalité de Kolmogorov,

$$ E\left(\sup_{M\geq n} \left\| \sum_{j=n}^{M} c_j\,[{}^T Z_j\,Q\,(\varepsilon_{j+1} + r_{j+1})/U_j] \right\|^2\right) \leq \text{Cte}\,(\alpha_n^2).$$

Comme $E\,(1_{\Gamma_p} \| U_{M+1} \|^2)$ tend vers 0 si $M \to \infty$, on obtient :

$$ E\left(\left(1_{\Gamma_p} \sum_{j=n}^{\infty} \gamma_j \| Z_j \|\right)^2\right) \leq \text{Cte}\,\alpha_n^2.$$

b) Soit $0 < L < \underline{\lambda}(H)$ et une norme $|\cdot|$ de $\mathbb{R}^d$ telle que, si $0 < \gamma < \bar{\gamma}$,

$$ |(I + \gamma H)^{-1}| \leq (1 + L\gamma)^{-1};$$

on prend p entier, tel que, pour $n \geq p$, $\gamma_n < \bar{\gamma}$.

Pour $n \geq p$, on a, sur Γ_p,

$$Z_{n+1} = (I + \gamma_n H) Z_n + \gamma_n R_n + c_n (\varepsilon_{n+1} + r_{n+1}),$$

avec $|R_n| \leq |\Delta_n| |Z_n|$.

Soit $B_n = (1 + \gamma_n H) \dots (1 + \gamma_p H)$ pour $n \geq p$, $B_{p-1} = I$.

On écrit alors sur Γ_p, pour $n \geq p$, selon le lemme 2 :

$$-Z_p = \sum_{n=p}^{\infty} B_n^{-1} (\gamma_n R_n + c_n [\varepsilon_{n+1} + r_{n+1}]).$$

Sur Γ_p, posons $S_N = \sum_{n=N}^{\infty} (\gamma_n R_n + c_n [\varepsilon_{n+1} + r_{n+1}])$;

$$-Z_p = \sum_{n=p}^{\infty} B_n^{-1} (S_n - S_{n+1}) = \sum_{n=p}^{\infty} [B_n^{-1} - B_{n-1}^{-1}] S_n + S_p;$$

$$\sum_{n=p}^{\infty} (R_n^1 + c_n [\varepsilon_{n+1} + r_{n+1}]) = -Z_p,$$

avec

$$R_n^1 = \gamma_n R_n - [B_{n-1}^{-1} - B_n^{-1}] S_n.$$

Or :

$$[B_{n-1}^{-1} - B_n^{-1}] = B_n^{-1} \gamma_n H;$$

$$|B_{n-1}^{-1} - B_n^{-1}| \leq [(1 + \gamma_n L) \dots (1 + \gamma_p L)]^{-1} \gamma_n |H|$$
$$= O(\gamma_n \exp[-L(\gamma_p + \dots + \gamma_n)]);$$

$$E\left(1_{\Gamma_p} \sum_{n=N}^{\infty} |R_n^1|\right) \leq E\left(1_{\Gamma_p} \sum_{n=N}^{\infty} \gamma_n |R_n|\right)$$
$$+ \text{Cte}\, E\left(1_{\Gamma_p} \sum_{n=N}^{\infty} |B_{n-1}^{-1} - B_n^{-1}| |S_n|\right)$$
$$\leq E\left(1_{\Gamma_p} \sup_{n \geq N} |\Delta_n| \sum_{n=N}^{\infty} \gamma_n |Z_n|\right)$$
$$+ \sum_{n=N}^{\infty} |B_{n-1}^{-1} - B_n^{-1}| E(1_{\Gamma_p} |S_n|).$$

Pour étudier cette majoration, on fait quatre remarques :

□ d'après *a*),

$$E\left(1_{\Gamma_p} \sup_{n\geq N} |\Delta_n| \sum_{n=N}^{\infty} \gamma_n |Z_n|\right) \leq \text{Cte}\, \alpha_n \left(E\left(1_{\Gamma_p} \sup_{n\geq N} |\Delta_n|^2\right)\right)^{1/2};$$

□ $(\sup_{n\geq N} |\Delta_n|)$ est, sur Γ_p, une suite majorée par une constante qui tend vers 0 si $N \to \infty$;

□ $\sum_{n=p}^{\infty} |B_{n-1}^{-1} - B_n^{-1}| < \infty$;

□ $E(1_{\Gamma_p} |S_n|) \leq \text{Cte}\, \alpha_n$.

D'où : $E\left(1_{\Gamma_p} \sum_{n=N}^{\infty} |R_n^1|\right) = o(\alpha_N)$.

On est dans le cadre d'application du théorème A de l'appendice et $P(\Gamma_p) = 0$; $\Gamma = \cup\Gamma_p$ est négligeable. ■

I.4. Piège régulier général

Il s'agit enfin de prouver que les pièges réguliers quelconques sont contournés.

REDRESSEMENT

Pour l'étude du cas général, on a recours dans la théorie des équations différentielles à des propriétés de « redressement » de h au voisinage de z^*. Reprenant la transformation linéaire P définie en I.1 dans le cas général, on se ramène à :

$$\begin{aligned} Y_{n+1} &= P(Z_{n+1} - z^*) = Y_n + \gamma_n P\, h(P^{-1}[Y_n + P z^*]) + P\, \eta_{n+1} \\ &= Y_n + \gamma_n\, g(Y_n) + c_n P(r_{n+1} + \varepsilon_{n+1}). \end{aligned}$$

Soit δ la dimension de Ker $M_+(H)$; on associe à $y \in \mathbb{R}^d$, $y^+ \in \mathbb{R}^\delta$ et $y^- \in \mathbb{R}^{d-\delta}$ tels que ${}^T y = ({}^T y^+, {}^T y^-)$. Au voisinage de 0,

$$g(y) = \begin{bmatrix} g_+(y) \\ g_-(y) \end{bmatrix} = PH\, P^{-1} y + q(y) = \begin{bmatrix} J_+ & 0 \\ 0 & J_- \end{bmatrix} \begin{bmatrix} y^+ \\ y^- \end{bmatrix} + \begin{bmatrix} q_+(y) \\ q_-(y) \end{bmatrix},$$

q de classe C^1, $q(0) = 0$, $Dq(0) = 0$, Dq lipschitzienne.

Considérons le système différentiel :

$$\begin{cases} dy^+/dt = g_+(y) \\ dy^-/dt = g_-(y). \end{cases}$$

D'après un lemme de Poincaré datant de 1886 ([29]; *cf.* Hartman [11] p. 228-242, lemme 5.1, exercice 5.1 et corollaire 5.2), il existe une fonction G définie sur un voisinage de 0 dans $\mathbb{R}^{d-\delta}$ à valeurs dans $\mathbb{R}^{\delta}$, différentiable et telle que :

□ G est différentiable avec une matrice jacobienne DG lipschitzienne, $G(0) = 0$, $DG(0) = 0$;

□ si $y(0)$ est dans un voisinage W de 0 et si $y^+(0) = G(y^-(0))$,

$$y^+(t) = G(y^-(t)) \quad \text{pour} \quad 0 \leq t \leq t_0,\ t_0 > 0.$$

La seconde propriété signifie que, dans un voisinage W de 0

□ $y^+ = G(y^-)$ implique $g_+(y) = DG(y^-)\, g_-(y)$.

On choisit W ouvert et convexe.

Pour $y \in W$, notons $u^+ = y^+ - G(y^-)$, $u^- = y^-$, $u = (u^+, u^-)$;

$$\begin{aligned} F(u^+, u^-) &= g_+(y) - DG(y^-)\, g_-(y) \\ &= g_+(u^+ + G(u^-), u^-) - DG(u^-)\, g_-(u^+ + G(u^-), u^-); \end{aligned}$$

$F(0, u^-) = 0$.

Notant D_+ la différentielle partielle par rapport à u^+, on a :

$$\begin{aligned} F(u^+, u^-) &= F(u^+, u^-) - F(0, u^-) \\ &= \left(\int_0^1 D_+ F(tu^+, u^-)\, dt \right) \begin{bmatrix} u^+ \\ 0 \end{bmatrix} = (J_+ + \Delta(u))\, u^+, \end{aligned}$$

où $\Delta(u)$ est une matrice telle que $\| \Delta(u) \| \leq \text{Cte} \| u \|$.

TRANSFORMATION DE L'ALGORITHME

Les réductions (4) et (5) de I.2, étant effectuées, soit $V^1(z^*)$ un voisinage de z^* ouvert et convexe tel que :

$$V^1(z^*) \subseteq V(z^*) \cap \{z;\ P(z - z^*) \in W\},$$

$$\Gamma_p = \{\omega;\ \omega \in \Gamma(z^*),\ Z_n(\omega) \in V^1(z^*) \text{ pour } n \geq p\};$$

$$\Gamma(z^*) = \cup \Gamma_p.$$

Posons $U_n^+ = Y_n^+ - G(Y_n^-)$, $U_n^- = Y_n^-$, $U_n = (U_n^+, U_n^-)$.

On a, sur Γ_p, pour $n \geq p$,

$$G(Y_{n+1}^-) = G(Y_n^-) + DG(Y_n^-)\, [Y_{n+1}^- - Y_n^-] + O(\| Y_{n+1} - Y_n \|^2);$$

$$U_{n+1}^+ - U_n^+ = \gamma_n (g_+(Y_n) - DG(Y_n^-)\, g_-(Y_n)) + c_n (\rho_{n+1} + e_{n+1})$$

où $e_{n+1} = [P\,\varepsilon_{n+1}]^+ - DG(Y_n^-)\,[P\,\varepsilon_{n+1}]^-$ et (ρ_n) est une suite adaptée à $\mathbb{F}$ telle que $\sum \| \rho_{n+1} \|^2 < \infty$, p.s. sur $\Gamma(z^*)$;

$$U_{n+1}^+ = U_n^+ + \gamma_n [J_+ + \Delta_n] U_n^+ + c_n [\rho_{n+1} + e_{n+1}],$$

où $\Delta_n = \Delta(U_n)$.

Les hypothèses de la proposition 4 s'appliquent à (U_n^+). Il ne reste en effet qu'à examiner les propriétés du bruit et :

$$E(\| DG([Y_n]^-)\,[P\,\varepsilon_{n+1}]^- \|^2 \mid \mathcal{F}_n) \leq \| DG([Y_n]^-) \|^2 E(\| [P\,\varepsilon_{n+1}]^- \|^2 \mid \mathcal{F}_n).$$

Avec les hypothèses [H2a], sur $\Gamma(z^*)$, $DG([Y_n]^-)$ tend vers 0 et ce dernier majorant tend, p.s. vers 0; d'où, p.s. sur $\Gamma(z^*)$,

$$\limsup E(\| e_{n+1} \|^2 \mid \mathcal{F}_n) < \infty;$$

grâce à l'hypothèse [H2b],

$$\liminf E(\| e_{n+1} \|^2 \mid \mathcal{F}_n) = \liminf E(\| [P\,\varepsilon_{n+1}]^+ \| \mid \mathcal{F}_n) > 0.$$

Le théorème est démontré. ■

I.5. Commentaires

EXCITATION D'ORDRE P DU BRUIT

Soit p un entier > 1; [ER] implique :

$$Z_{(n+1)p} = T_{n+1} = T_n + a_n\,h(T_n) + \zeta_{n+1},$$

avec $a_n = \gamma_{np} + \ldots + \gamma_{(n+1)p-1}$; $\eta_{n+1} = c_n(\varepsilon_{n+1} + r_{n+1})$ se traduit par :

$$\begin{aligned}\zeta_{n+1} &= c_{np}[\varepsilon_{np+1} + r_{np+1}] + \ldots + c_{(n+1)p-1}[\varepsilon_{(n+1)p} + r_{(n+1)p}] \\ &\quad + \gamma_{np+1}[h(Z_{np+1}) - h(T_n)] \\ &\quad + \ldots + \gamma_{(n+1)p-1}[h(Z_{(n+1)p-1}) - h(T_n)] \\ &= c_{np}\,\varepsilon_{np+1} + \ldots + c_{(n+1)p-1}\,\varepsilon_{(n+1)p} + c_{np}\,\rho_{n+1};\end{aligned}$$

on suppose que c_{n+1}/c_n et γ_{n+1}/γ_n tendent vers 1 si $n \to \infty$; sur $\Gamma(z^*)$, $\sum \| \rho_{n+1} \|^2 < \infty$. Le théorème 1 s'applique encore en remplaçant la propriété d'excitation du bruit [H2b] par :

$$\liminf E(\| \varepsilon_{n+1}^{(r)} + \ldots + \varepsilon_{n+p}^{(r)} \| \mid \mathcal{F}_n) > 0.$$

Comme

$$E\left(\| \varepsilon_{n+1}^{(r)} + \ldots + \varepsilon_{n+p}^{(r)} \|^2 \mid \mathcal{F}_n\right) = E\left(\sum_{j=1}^{p} \| \varepsilon_{n+j}^{(r)} \|^2 \mid \mathcal{F}_n\right),$$

on obtient, en tenant compte de la remarque 1 de I.2, le corollaire aisément maniable suivant.

COROLLAIRE 5. – *On suppose que le bruit a, sur* $\Gamma(z^*)$, *un moment conditionnel d'ordre* > 2 *fini et que* c_{n+1}/c_n *et* γ_{n+1}/γ_n *tendent vers* 1 *si* n *tend vers l'infini.*

Le théorème 1 *reste valable en remplaçant* [H2b] *par la condition d'excitation suivante : il existe un entier* p *tel que,* p.s. *sur* $\Gamma(z^*)$, $\liminf E\left(\sum_{j=1}^{p} \| \varepsilon_{n+j}^{(r)} \|^2 \mid \mathcal{F}_n\right) > 0.$

UN AUTRE REDRESSEMENT

Il est possible de généraliser la méthode utilisée en I.3. pour prouver le lemme 3 de la manière suivante. Soit λ une valeur propre de H de partie réelle > 0 et v un vecteur propre de ${}^T H$ associé. Supposons qu'il existe une fonction φ définie sur un voisinage de 0, à valeurs dans $\mathbb{C}$, φ supposée différentiable avec un gradient $\nabla \varphi$ lipschitzien et solution de l'équation de Poincaré

$$[P_\lambda] \qquad {}^T\nabla \varphi(z)\, h(z - z^*) = \lambda \varphi(z).$$

Alors la démonstration du lemme 3 est inchangée, en utilisant le corollaire B de l'appendice, et prouve le théorème 1.

L'étude de l'équation de Poincaré résulte de la linéarisation de l'équation différentielle [ED] lorsqu'elle est possible (*cf.* Hartman [11], p. 256-271).

Cette méthode – que nous avions utilisée dans un premier temps – est plus simple ... mais a l'inconvénient d'introduire des hypothèses plus fortes de différentiabilité de h au voisinage de z^* ainsi que l'hypothèse de non résonance de H qui s'avère difficile à vérifier. Le redressement utilisé en I.4 est celui qu'utilise Lazarev [19].

RÉSULTATS ANTÉRIEURS

La méthode utilisée ci-dessus nous semble plus facile à comprendre que celle suivie dans [25] et [19] qui repose notamment, dans le cas répulsif, sur la construction compliquée d'une fonction de Lyapounov.

Avec nos notations, les hypothèses de Nevel'son-Has'minskii [25] dans le cas répulsif et celles de Lazarev [19] dans le cas général diffèrent surtout sur les perturbations. Dans [19], les conditions de régularité sur h sont

les mêmes que les nôtres et l'on impose aux pas la condition plus forte $\sum \gamma_n^2 \left(\sum_{j=n}^{\infty} c_j^2 \right)^{-1/2} < \infty$, ce qui n'est pas très important puisque l'on a le choix des pas. Par contre les conditions imposées au bruit (ε_n) et aux termes résiduels (r_n) sont plus restrictives que les nôtres. Dans [19], on a :

$$r_n = q(n, Z_n) \text{ avec, pour } q(n) = \sup \{q(n, z);\ z \in V(z^*)\},$$

$$\sum \gamma_n q(n) \left(\sum_{j=n}^{\infty} c_j^2 \right)^{-1/2} < \infty,$$

$$E(\varepsilon_{n+1}{}^T \varepsilon_{n+1} \mid \mathcal{F}_n) = \Gamma(n, Z_n)$$

avec $\lambda_{\min} \Gamma(n, z) \geq a > 0$ pour tout entier n et $z \in V(z^*)$.

Nous verrons en II divers cas où l'élargissement de ces hypothèses relatives à la perturbation n'est pas superflu, tant sur la structure du bruit que sur le terme résiduel.

II. COMMENT APPLIQUER CE QUI PRÉCÉDE ?

II.1 Cadre général

Dans tout ce paragraphe II, on se place sur un espace de probabilité $(\Omega, \mathcal{A}, P)$ muni d'une filtration $\mathbb{F} = (\mathcal{F}_n)$; on considère une suite adaptée à $\mathbb{F}$, (H_n) de vecteurs aléatoires de dimension d. On étudie un algorithme à valeurs dans un ouvert G de $\mathbb{R}^d$:

$$Z_{n+1} = Z_n + \gamma_n H_{n+1} + \eta_{n+1}^1, \tag{7}$$

sur l'ensemble de trajectoires $\Gamma(z^*) = \{(Z_n) \to z^*\}$, $z^* \in G$; (γ_n) est une suite déterministe positive, $\sum \gamma_n = \infty$, $\sum \gamma_n^2 < \infty$.

La perturbation (η_n^1) est adaptée à $\mathbb{F}$; elle est souvent, mais pas toujours, prise nulle; Z_0 est mesurable par rapport à $\mathcal{F}_0$. On suppose qu'un prédicteur raisonnable de H_{n+1} à l'instant n est $h(Z_n)$, h fonction continue de G dans $\mathbb{R}^d$.

L'erreur de prédiction est $\pi_{n+1} = H_{n+1} - h(Z_n)$ et :

$$Z_{n+1} = Z_n + \gamma_n h(Z_n) + \eta_{n+1},$$

$$\eta_{n+1} = \gamma_n \pi_{n+1} + \eta_{n+1}^1;$$

c'est un algorithme [ER] du type de celui qui est étudié en I.

ALGORITHME DE ROBBINS-MONRO [30]

Les cas usuels correspondent à des algorithmes de recherche de certains des zéros de h (*algorithme de Robbins-Monro).* Lorsque U est un potentiel, fonction de classe C^1 de $V(z^*)$ dans $\mathbb{R}$, l'algorithme est un *algorithme du gradient* lorsque $h = -\nabla U$.

ALGORITHME DE KIEFER-WOLFOWITZ [15]

Soit U un potentiel de classe C^2 de différentielle seconde lipschitzienne au voisinage de z^*; il arrive que l'on sache réaliser des expériences d'effet moyen $U(z)$ mais pas des expériences d'effet moyen $-\nabla U(z)$. On peut alors procéder comme suit.

Soient (b_n) une suite qui décroît vers 0 et $(f^j)_{1\leq j\leq d}$ la base canonique de $\mathbb{R}^d$. À l'instant n, Z_n étant choisi, on réalise 2 d expériences H_{n+1}^{j+} et H_{n+1}^{j-}, $1 \leq j \leq d$, à valeurs réelles indépendantes entre elles conditionnellement au passé; on suppose que l'expérience $H_{n+1}^{j\mp}$ a été réalisée avec le contrôle $Z_n \mp b_n f^j$ et qu'un prédicteur convenable de son résultat est $U(Z_n \mp b_n f^j)$.

Un algorithme de Kiefer-Wolfowitz relatif à $Z_n = (Z_n^j)_{1\leq j\leq d}$ s'écrit alors :

$$Z_{n+1}^j = Z_n^j - [\gamma_n/b_n](H_{n+1}^{j+} + H_{n+1}^{j-}) + \eta_{n+1}^{1,j}.$$

Sur $\Gamma(z^*)$, pour n assez grand :

$$Z_{n+1}^j = Z_n^j - [\gamma_n/b_n](U(Z_n + b_n f^j) - U(Z_n - b_n f^j) + \pi_{n+1}^j) + \eta_{n+1}^{1,j},$$

où $\pi_{n+1} = (\pi_{n+1}^j)$ est une erreur de prédiction;

$$Z_{n+1} = Z_n - 2\gamma_n \nabla U(Z_n) + \eta_{n+1},$$
$$\eta_{n+1} = c_n[\pi_{n+1} + \gamma_n b_n^3 q(n, Z_n)] + \eta_{n+1}^1,$$

$c_n = [\gamma_n/b_n]$, $\| q(n, z) \| \leq$ Cte si $z \in V(z^*)$ et si n est assez grand. On supposera $\sum b_n^6 < \infty$. Si U est seulement supposé de classe C^1, ∇U étant lipschitzienne, on remplace b_n^3 par b_n^2 avec $\sum b_n^4 < \infty$.

Les hypothèses [H3] sont satisfaites avec $\gamma_n = \gamma/n$ et $b_n = n^{-b}$, $1/6 < b < 1/2$ dans le premier cas, $1/4 < b < 1/2$ dans le second.

ALGORITHME NORME [28]

On modifie un algorithme de Robbins-Monro pour lequel $\eta^1 = 0$, en lui imposant d'être normé :

$$Z_{n+1} = [Z_n + \gamma_n H_{n+1}]/\| Z_n + \gamma_n H_{n+1} \|.$$

Lorsque (H_n) est bornée sur $\Gamma(z^*)$, on obtient pour n assez grand :

$$Z_{n+1} = Z_n + \gamma_n(- < Z_n,\ H_{n+1} > Z_n + H_{n+1}) + \gamma_n r^1_{n+1},$$

r^1_{n+1} étant $\mathcal{F}_{n+1}$-mesurable et $r^1_{n+1} = O(\gamma_n)$ sur $\Gamma(z^*)$. Autrement dit :

$$Z_{n+1} = Z_n + \gamma_n g(Z_n) + \eta_{n+1},$$

avec $g(z) = -\langle z,\ h(z)\rangle z + h(z)$,

$$\eta_{n+1} = \gamma_n(-\langle Z_n,\ \pi_{n+1}\rangle Z_n + \pi_{n+1} + r^1_{n+1}).$$

On s'est ramené à un algorithme [ER] pour lequel h a été remplacée par g.

Dans ces diverses situations, z^* étant respectivement un piège régulier de h, de $-\nabla U$ ou de g, le théorème 1 est le corollaire 5 s'appliquent, à condition de vérifier les hypothèses relatives à la perturbation.

II.2. Modèle de régression

Considérons un modèle de régression général adapté à $\mathbb{F}$, pour lequel Z_n est une variable explicative à l'instant n :

$$H_{n+1} = h(Z_n) + \varepsilon_{n+1},$$

la suite (ε_n) est, sur $\Gamma(z^*)$, un bruit adapté à $\mathbb{F}$ ayant un moment conditionnel d'ordre 2 fini, selon l'expression définie en I.2 (hypothèse [H2]); la fonction h est inconnue.

Par exemple, supposons que :

$$H_{n+1} = H(X_{n+1},\ Z_n)$$

pour une suite (X_n) adaptée à $\mathbb{F}$ de fonctions mesurables à valeurs dans un espace $(K,\ \mathcal{K})$ de loi μ, X_{n+1} étant indépendante de $\mathcal{F}_n$. On est dans le cadre précédent, si, tout $z \in V(z^*)$, $[H(\cdot,\ z)]^2$ est μ-intégrable;

$$\bullet\ h(z) = \int H(x,\ z)\, d\mu(x).$$

Dans ce cas particulier, on est dans le cadre étudié par Lazarev pour les algorithmes de Robbins-Monro et de Kiefer-Wolfowitz, mais pas pour l'algorithme normé à cause du terme résiduel.

Dans le cadre de II.1, $\pi_{n+1} = \varepsilon_{n+1}$, et, pour les trois algorithmes décrits ci-dessus, avec $\eta^1 = 0$, le piège z^* est, p.s., évité dès que (ε_n) est un bruit excitant dans une direction répulsive.

EXEMPLE : ANALYSE EN COMPOSANTES PRINCIPALES

Soit $(a(j))_{1\leq j\leq N}$ un nuage de N points de $\mathbb{R}^d$ de C la matrice :

$$C = \frac{1}{N}\sum_{j=1}^{N} a(j)^T a(j).$$

On suppose que C a des valeurs propres toutes distinctes; si le nuage est centré, l'analyse en composantes principales est la recherche des vecteurs propres normés associés aux plus grandes valeurs propres. Voici deux algorithmes pour la recherche de l'axe principal, inspirés par les réseaux de neurones.

Un algorithme du gradient

On cherche quels vecteurs $z \in \mathbb{R}^d$ minimisent :

$$V(z) = \frac{1}{N}\sum_{j=1}^{N} \| a(j) - z^T z\, a(j) \|^2 = \text{Trace } (I - z^T z)\, C\, (I - z^T z).$$

Soit

$$v(z) = \| a - z^T z\, a \|^2 = \| a \|^2 + \langle a,\, z \rangle^2 (\| z \|^2 - 2);$$

$$-\frac{1}{2} Dv(z) = (a^T a\,(2 - \| z \|^2) - \langle a,\, z \rangle^2)\, z,$$

$$-\frac{1}{2} Dv(z) = (C\,(2 - \| z \|^2) - {}^T z\, C\, z)\, z.$$

Notons $\lambda_1 > \ldots > \lambda_d > 0$ les valeurs propres de C et $v^1, \ldots, v^d$ des vecteurs propres normés et deux à deux orthogonaux associés.

L'ensemble des zéros de DV contient 0 et les vecteurs $\mp v^j$, $1 \leq j \leq d$. Dans la base $(v^1, \ldots, v^d)$, soient $z_1, \ldots, z_d$ les composantes de z; la i-ième composante de $DV/2$ est

$$z \to -2\lambda_i z_i + z_i \sum_{j=1}^{d} [\lambda_j + \lambda_i]\, z_j^2 = \frac{1}{2}\, [\partial V/\partial z_i]\, V(z);$$

$$\frac{1}{2}\, [\partial^2 V/\partial^2 z_i]\, V(z) = -2\lambda_i + \sum_{j=1,\, j\neq i}^{d} (\lambda_i + \lambda_j)\, z_j^2 + 6\lambda_i z_i^2;$$

et, pour $j \neq i$,

$$\frac{1}{2}\, [\partial^2 V/\partial z_i\, \partial z_j]\, V(z) = 2 z_i z_j (\lambda_i + \lambda_j).$$

$D^2 V(0) = -4C$, 0 est un maximum de V; et, dans cette base, $D^2 V(\mp v^k)$ est une matrice diagonale dont les termes valent $2(\lambda_k - \lambda_i)$ pour $i \neq k$ et $4\lambda_k$ pour $i = k$.

On choisit au hasard une suite de points du nuage c'est-à-dire $(a(X_n))$, pour une suite (X_n) de variables aléatoires uniformes sur $\{1, \ldots, N\}$ et indépendantes. Soit (γ_n) une suite positive décroissante satisfaisant à [H3], par exemple $\gamma_n = \gamma/n$.

Pour $Z_0 \in \mathbb{R}^d$ non nul et $C_n = a(X_n)^T a(X_n)$, considérons l'algorithme

$$\begin{aligned} Z_{n+1} &= Z_n + \gamma_n (1 + \| Z_n \|^2)^{-1} (C_{n+1} (2 - \| Z_n \|^2) - {}^t Z_n C_{n+1} Z_n) Z_n \\ &= Z_n - \frac{1}{2} \gamma_n (1 + \| Z_n \|^2)^{-1} (DV(Z_n) + \varepsilon_{n+1}); \end{aligned}$$

on montre alors de manière classique que (Z_n) converge, p.s. vers 0 ou vers l'un des vecteurs propres unitaires de C (*cf.* [3]).

Les pièges de l'algorithme sont 0 et les vecteurs propres $\mp v^k$, $k > 1$, points selles de V. Dès que chacun des pièges est évité, on est assuré de la convergence, p.s., de l'algorithme vers un vecteur unitaire de l'axe principal.

Soit un piège $\mp v^k$; pour $i < k$, v^i est vecteur propre de $-D^2 V(\mp v^k)$ associé à $2(\lambda_i - \lambda_k) > 0$. On a :

$$\begin{aligned} \langle v^i, \varepsilon_{n+1} \rangle &= {}^T v^i ((C_{n+1} - C)(2 - \| Z_n \|^2) - {}^T Z_n (C_{n+1} - C) Z_n) Z_n \\ &= (\langle v^i, a(X_{n+1}) \rangle \langle Z_n, a(X_{n+1}) \rangle \\ &\quad - \lambda_i \langle v^i, Z_n \rangle)(2 - \| Z_n \|^2) \\ &\quad + (-\langle Z_n, a(X_{n+1}) \rangle^2 + {}^T Z_n C Z_n) \langle Z_n, v^i \rangle; \end{aligned}$$

d'où, sur $\{(Z_n) \to \mp v^k\}$:

$$\lim E(\langle v^i, \varepsilon_{n+1} \rangle^2 \mid X_1, \ldots, X_n) > 0.$$

Comme le bruit est borné, l'hypothèse [H2b] est satisfaite et le piège $\mp v^k$ est contourné.

Pour le piège 0, la condition d'excitation du bruit n'est plus satisfaite : une démonstration directe est nécessaire. On dispose d'une majoration grossière K des normes des vecteurs du nuage.

Notant $a_n = a(X_n)$,

$$\begin{aligned} Z_{n+1} &= Z_n + \gamma_n (1 + \| Z_n \|^2)^{-1} (\langle a_{n+1}, Z_n \rangle (2 - \| Z_n \|^2) a_{n+1} \\ &\quad - \langle a_{n+1}, Z_n \rangle^2 Z_n) \\ \| Z_{n+1} - Z_n \| &\leq 2 \gamma_n K^2 (1 + \| Z_n \|^2)^{-1} \| Z_n \| (1 + \| Z_n \|^2) \\ &\leq 2 \gamma_n K^2 \| Z_n \|, \\ \| Z_{n+1} \| &\geq \| Z_n \| (1 - 2 \gamma_n K^2); \end{aligned}$$

pour $\| Z_n \|^2 \leq 1/2$,

$$\begin{aligned}\| Z_{n+1} \|^2 &= \| Z_n \|^2 + 4\gamma_n (1 + \| Z_n \|^2)^{-1} (\langle a_{n+1}, Z_n \rangle^2 (1 - \| Z_n \|^2)) \\ &\quad + \gamma_n^2 (1 + \| Z_n \|^2)^{-2} (\langle a_{n+1}, Z_n \rangle^2 \| a_{n+1} \|^2 (2 - \| Z_n \|^2)^2 \\ &\quad + \langle a_{n+1}, Z_n \rangle^4 (-4 + 3 \| Z_n \|^2)) \\ &\geq \| Z_n \|^2 + 2\gamma_n (\langle a_{n+1}, Z_n \rangle^2 (1 - \| Z_n \|^2)) \\ &\quad + \frac{1}{4} \gamma_n^2 \langle a_{n+1}, Z_n \rangle^2 \| a_{n+1} \|^2 (4 - 8 \| Z_n \|^2) \geq \| Z_n \|^2.\end{aligned}$$

Si l'on choisit γ_0 assez petit ($\gamma_0 < 1/2K^2$), si $\| Z_n \|^2 > 1/2$, $Z_{n+1} \neq 0$ et si $\| Z_n \|^2 \leq 1/2$, $\| Z_{n+1} \| \geq \| Z_n \|$. Ainsi, la suite (Z_n) ne s'annule pas et ne peut pas tendre vers 0.

Pour γ_0 choisi assez petit, le piège 0 est évité.

Un algorithme normé

Considérons l'algorithme normé suivant dû à Oja [28] :

$$Z_{n+1} = (Z_n + \gamma_n a(X_{n+1})^T a(X_{n+1}) Z_n) / \| Z_n + \gamma_n a(X_{n+1})^T a(X_{n+1}) Z_n \|.$$

D'après II.1, il s'écrit :

$$Z_{n+1} = Z_n + \gamma_n g(Z_n) + \gamma_n [r_{n+1} + \varepsilon_{n+1}],$$

avec $g(z) = (C - {}^T z C z) z$. Les zéros de g sont les mêmes que ceux de h introduits ci-dessus; tous sont des pièges réguliers sauf $\mp v^1$; Oja et Karhunen montrent la convergence, p.s., vers l'un de ces zéro de g ([28]; *voir* aussi [5]). Les pièges sont évités si γ_0 est assez petit : la preuve donnée ci-dessus est presque inchangée.

PROPOSITION 6. – *Les deux algorithmes précédents convergent, p.s., vers l'un des deux vecteurs unitaires de l'axe principal.*

D'autres algorithmes analogues ont été proposés pour la recherche successive de vecteurs unitaires des axes principaux : *voir* par exemple Oja [27] ou Hornik-Kuan [12]; à l'exception d'une preuve partielle donnée dans [28], ces auteurs se contentent de vérifier que les cibles (p vecteurs unitaires des p premiers axes principaux) correspondent aux zéros asymptotiquement stables de l'équation différentielle associée. L'analyse des pièges de ces divers algorithmes peut être entreprise à l'aide de ce qui précède. Brandière [3] mène à bien cette étude dans le cas d'un algorithme du gradient.

II.3. Petites perturbations markoviennes

Le cadre qui suit est à peu près celui du livre de Benveniste-Métivier-Priouret [2].

On reprend les schémas de II.1, mais ici, dans la formule (7),

$$H_{n+1} = H(X_{n+1}, Z_n),$$

où (X_n) est une suite, adaptée à $\mathbb{F}$, d'observations aléatoires à valeurs dans un espace mesuré $(K, \mathcal{K})$; on suppose que la loi de X_{n+1} conditionnelle à $\mathcal{F}_n$ est $p(X_n, Z_n; \cdot)$, pour une probabilité de transition p de $K \times G$ muni du produit de $\mathcal{K}$ et de la tribu borélienne dans $(K, \mathcal{K})$.

Notons $p_z(x; \cdot) = p(x, z; \cdot)$. On connaît divers critères de « récurrence rapide » ou de « stabilité de modèles itératifs » assurant la validité de l'hypothèse suivante (*cf.* [24], [26] pour la récurrence; [2], [7] pour les modèles lipschitziens).

[HM] Propriétés de stationnarité

Il existe un voisinage $V(z^*)$ de z tel que, pour tout $z \in V(z^*)$ les propriétés suivantes soient satisfaites :

a) Il existe une probabilité μ_z sur $(K, \mathcal{K})$ invariante par p_z, unique.

b) $H(\cdot, z)$ est de carré intégrable pour μ; on note

$$h(z) = \int H(x, z)\, d\mu_z(x).$$

c) Il existe une fonction G définie sur $K \times V(z^*)$ à valeurs réelles, borélienne si $K \times V(z^*)$ est muni du produit de $\mathcal{K}$ et de la tribu borélienne, telle que $G(\cdot, z)$ soit solution de l'équation de Poisson associée à $H(\cdot, z)$:

$$H(x, z) - h(z) = G(x, z) - \int G(y, z)\, p_z(x; dy);$$

de plus :

$$\sup \left\{ \int \| G(x, z) \|^2 p_z(x; dy);\ z \in V(z^*),\ x \in K \right\} < \infty.$$

Pour la chaîne de Markov de transition p_z, μ_z est une loi stationnaire. Lorsque $Z_n \in V(z^*)$, sous les hypothèses [HM], il est donc naturel de considérer $h(Z_n)$ comme un prédicteur de $H(X_{n+1}, Z_n)$. L'erreur de prédiction s'écrit :

$$\pi_{n+1} = H(X_{n+1}, Z_n) - h(Z_n) = G(X_{n+1}, Z_n) - \int p(X_{n+1}, Z_n; dy)\, G(y, Z_n).$$

Pour n assez grand, sur $\Gamma(z^*)$,

$$Z_{n+1} = Z_n + \gamma_n H(X_{n+1}, Z_n) + \eta^1_{n+1} = Z_n + \gamma_n h(Z_n) + \gamma_n \pi_{n+1} + \eta^1_{n+1}.$$

La « petite perturbation markovienne » $\pi = (\pi_n)$ peut s'étudier de la manière suivante.

Soit $\varepsilon_{n+1} = G(X_{n+1}, Z_n) - \int p(X_n, Z_n; dy)\, G(y, Z_n)$; (ε_n) est un bruit adapté à $\mathbb{F}$ ayant un moment conditionnel d'ordre 2. Posons :

$$\begin{aligned} Y_{n+1} &= Z_{n+1} + \gamma_n \int p(X_{n+1}, Z_n; dy)\, G(y, Z_n); \\ Y_{n+1} &= Y_n + \gamma_n h(Y_n) + \gamma_n (\varepsilon_{n+1} + r_{n+1}) + \eta^1_{n+1}, \end{aligned} \tag{8}$$

avec

$$\begin{aligned} r_{n+1} = {} & (h(Z_n) - h(Y_n)) + \int p(X_n, Z_n; dy)\, G(y, Z_n) \\ & - \int p(X_n, Z_{n-1}; dy)\, G(y, Z_{n-1}) \\ & + [1 - \gamma_{n-1}/\gamma_n] \int p(X_n, Z_{n-1}; dy)\, G(y, Z_{n-1}). \end{aligned}$$

Sous diverses hypothèses assez fortes de régularité de ce modèle, on parvient à prouver que :

□ $(y_n - Z_n)$ tend vers 0;

□ $z \to h(z)$ et, pour tout x, $z \to \int p(x, z; dy)\, H(x, z)$ sont lipschitziennes;

□ la suite (r_{n+1}) qui est $\mathcal{F}_{n+1}$-mesurable satisfait à

$$\sum \gamma_n \| r_{n+1} \| \text{ et } \sum \| r_{n+1} \|^2 < \infty, \text{ p.s. sur } \Gamma(z^*),$$

lorsque l'on choisit $\gamma_n = \gamma n^{-b}$, $1/2 < b \leq 1$, $\gamma > 0$.

Voir en [2] une axiomatique adaptée et des exemples; le cas où p_z ne dépend pas de z est évidemment le plus simple.

Alors, sur $\Gamma(z^*)$, $E(\| \varepsilon_{n+1} - H(X_{n+1}, z^*) \|^2 \mid \mathcal{F}_n)$ tend, p.s., vers 0.

Si la suite $(H(X_{n+1}, z^*))$ est excitante dans une direction répulsive et si la perturbation complémentaire η^1 est nulle, le piège est contourné.

EXEMPLE : ESTIMATION D'UN MODÈLE DE GIBBS

Description du modèle. – On se place dans le cadre étudié par Younes ([31], [32]).

L'observation est un champ de Gibbs $X = (X^{(s)})_{s \in S}$, $K = F^S$, S étant un ensemble fini de sites et F un ensemble fini de niveaux de gris; K est muni d'une loi de Gibbs, définie pour $x \in K$ par

$$\mu_\theta(x) = \exp(\langle \theta, U(x) \rangle + \operatorname{Log} Z(\theta)),$$

où U est une fonction de K dans $\mathbb{R}^d$ connue et θ un paramètre inconnu à estimer.

On note $\bar{x}(s) = \{x^{(u)}; u \neq s\}$. On sait simuler une chaîne de Markov de loi stationnaire μ_θ, par exemple par un *échantillonneur de Gibbs aléatoire,* dont le site transformé est tiré au hasard. La transition de cet échantillonneur s'écrit :

$$p(x, \theta; y) = [1/\operatorname{card} S] \exp \langle \theta, U(y) \rangle \Big(\sum_{\bar{z}(s) = \bar{x}(s)} \exp \langle \theta, U(z) \rangle \Big)^{-1}$$

si $\bar{y}(s) = \bar{x}(s)$ pour un $s \in S$, $p(x, \theta; y) = 0$ sinon.

On suppose qu'une partie $S(2)$ des sites est cachée; on note

$$S(1) = S \backslash S(2),\ X^1 = (X^{(s)})_{s \in S(1)},\ X^2 = (X^{(s)})_{s \in S(2)}.$$

On peut aussi simuler une chaîne de Markov à valeurs dans $K_2 = F^{S(2)}$ dont la loi stationnaire est la loi de X conditionnelle à $X^1 = x^1$ par un *échantillonneur de Gibbs aléatoire conditionnel.* En posant

$$\bar{x}^2(s) = \{y;\ y^{(u)} = x^{(u)} \text{ pour } u \in S(2),\ u \neq s\},$$

la transition de ce nouvel échantillonneur s'écrit :

$$q(x^2, \theta; y^2) = [1/\operatorname{card} S(2)] \exp \langle \theta, U(x^1, y^2) \rangle \Big(\sum_{\bar{z}^2(s) = \bar{x}^2(s)} \exp \langle \theta, U(x^1, z^2) \rangle \Big)^{-1}$$

si $\bar{y}^2(s) = \bar{x}^2(s)$ pour un $s \in S(2)$, $q(x^2, \theta; y^2) = 0$ sinon.

Ces chaînes de Markov sont irréductibles : la probabilité de transition d'un point à un autre en un nombre de pas égal au cardinal de S est toujours > 0.

On prouve ([31], [32], [13]) que, pour toute fonction H de K dans $\mathbb{R}^d$, on a, pour chacune de ces chaînes des solutions de l'équation de Poisson $x \to a(\theta, x)$ et $x^2 \to b(\theta, x^2)$, satisfaisant à :

$$\| a(\theta, \cdot) \| + \| [\partial/\partial\theta] a(\theta, \cdot) \| + \| b(\theta, \cdot) \| + \| [\partial/\partial\theta] b(\theta, \cdot) \| \leq \operatorname{Cte} e^{c \| \theta \|},$$

pour une constante $c > 0$.

Description de l'algorithme SEM. – Après une observation $X^1 = x^1$, on cherche un estimateur du maximum de vraisemblance, Arg min v, avec

$$v(\theta) = -\mathrm{Log}\, P_\theta (X^1 = x^1) = -\mathrm{Log} \sum_{x^2 \in K_2} \mu_\theta (x^1, x^2);$$

$$\nabla v(\theta) = E_\theta (U(X)) - E_\theta (U(x^1, X^2) \mid X^1 = x^1);$$

et, si Γ_θ désigne la covariance, la différentielle seconde est :

$$D^2 v(\theta) = -\Gamma_\theta (U(X)) + \Gamma_\theta (U(x^1, X^2) \mid X^1 = x^1).$$

On suppose la matrice $\sum_{x \in K} U(x)^T U(x)$ inversible; alors $\Gamma_\theta (U(X))$ est définie positive et, pour le champs de Gibbs observable sans partie cachée, ∇v ne s'annule qu'en un point qui est l'estimateur du maximum de vraisemblance; ce cas, étudié en [31], ne comporte pas de piège.

L'algorithme SEM du champs de Gibbs partiellement observé étudié par Younes ([32], dans un cadre plus général) est un algorithme du gradient stochastique :

$$Z_{n+1} = Z_n - [\gamma/n] (U(X_{n+1}) - U(x^1, \xi_{n+1})),$$

où, à l'instant n, X_{n+1} et ξ_{n+1} sont générés de manière indépendante conditionnellement au passé, X_{n+1} de loi $p(X_n, Z_n; \cdot)$ et ξ_{n+1} de loi $q(\xi_n, Z_n; \cdot)$.

Comme U est bornée, $\| U(\cdot) \| \leq \| U \|$, $\| Z_n \| \leq \mathrm{Cte} + 2\gamma \| U \| \mathrm{Log} n$ et

$$\exp ([\| U \| + c] \| Z_n \|) = O(n^b),$$

où γ peut être choisi assez petit pour que $b < 1/2$.

L'algorithme peut alors s'écrire, selon la formule (8) précédente,

$$Z_{n+1} + u_{n+1} = Y_{n+1} = Y_n - [\gamma/n] (\nabla v(Y_n) + \varepsilon_{n+1} + r_{n+1}),$$

où, p.s., $(u_n) \to 0$, $\sum \gamma_n \varepsilon_{n+1}$ converge,

$$\sum (\| r_{n+1} \|^2 + \gamma_n \| r_{n+1} \|) < \infty.$$

Lorsque $\| \theta \| \to \infty$,

$$\langle \nabla v(\theta), \theta \rangle \simeq \alpha (\theta / \| \theta \|) \| \theta \|,$$

$$\alpha(\theta) = \sup_x \langle \theta, U(x) \rangle - \sup_{x^2} \langle \theta, U(x^1, x^2) \rangle.$$

[HG] *Hypothèses sur la vraisemblance*

□ $\inf\{\alpha(\theta) : \|\theta\| = 1\} > 0$;

□ *les zéros de* ∇v *sont isolés.*

La première de ces hypothèses est une propriété un peu plus forte que la propriété déjà introduite : « $\sum_{x \in K} U(x)^T U(x)$ est inversible ». C'est une propriété de contraction à l'infini grâce à laquelle la suite (Z_n) est bornée; la seconde des hypothèses [HG] implique alors la convergence presque sûre de (Z_n) vers l'un des zéros de ∇v (*cf.* [5] et [8]).

De plus, pour $j = \text{card } S$,

$$E([U(X_{n+j}) - U(x^1, \xi_{n+j})]^T [U(X_{n+j}) - U(x^1, \xi_{n+j})] \mid \mathcal{F}_n) \geq \int p_j(X_n, Z_n; dx)\, U(x)^T U(x);$$

si $\nabla v(z^*) = 0$, sur $\Gamma(z^*)$, la limite inférieure de ce minorant est > 0, et le corollaire 5 s'applique.

PROPOSITION 7. – *Sous les hypothèses* [HG], *l'algorithme SEM du champs de Gibbs partiellement observé converge, p.s., vers un maximum local de la vraisemblance.*

II.4. Vers le recuit simulé

Dans les divers exemples précédents, on s'est ramené à un algorithme

$$Z_{n+1} = Z_n + \gamma_n h(Z_n) + c_n [\varepsilon_{n+1} + r_{n+1}] + \eta^1_{n+1},$$

pour lequel (ε_n) est un bruit sur $\Gamma(z^*)$ et (r_n) une suite adaptée à $\mathbb{F}$ telle que $\sum r_n^2 < \infty$, p.s. sur $\Gamma(z^*)$. Si l'on n'est pas assuré de l'excitation de (ε_n) dans une direction répulsive, on peut prendre $\eta^1_{n+1} = c_n e_{n+1}$, où (e_n) est une suite de vecteurs aléatoires indépendants, ayant un moment d'ordre > 2 fini et de même loi supposée centrée et de covariance Γ inversible. La suite (e_n) peut être engendrée par simulation, e_{n+1} indépendante de $\mathcal{F}_n$. Soit $\mathcal{G}_n$ la tribu engendrée par $\mathcal{F}_n$ et $(e_k)_{k \leq n}$; p.s. sur $\Gamma(z^*)$,

$$\liminf \lambda_{\min} E([\varepsilon_{n+1} + e_{n+1}]^T [\varepsilon_{n+1} + e_{n+1}] \mid \mathcal{G}_n) \geq \lambda_{\min} \Gamma > 0.$$

Alors, si (ε_n) a, sur $\Gamma(z^*)$, un moment conditionnel d'ordre > 2 fini, le bruit $(\varepsilon_n + e_n)$ a la propriété d'excitation requise dans [H2b]; le piège z^* est, p.s., contourné.

Cette méthode est analogue à celle du recuit simulé.

Comme on l'a vu en introduction, le recuit simulé reste en général indispensable pour obliger un algorithme du gradient à contourner les minima locaux.

Ainsi, on prouve dans [13] que l'algorithme SEM du champ de Gibbs partiellement observé perturbé par $\beta (n \operatorname{Log} \operatorname{Log} n)^{-1/2} e_{n+1}$ converge en probabilité vers un maximum de la vraisemblance (sous les hypothèses [HG] et avec β assez grand).

Il est toutefois utile d'éviter le recuit si l'on sait (comme dans le cas de l'analyse en composantes principales ou dans la recherche des minima d'une fonction qui n'a pas de minima locaux) qu'il suffit de contourner les pièges pour converger, p.s., vers l'une des cibles recherchées. On sait en effet étudier la vitesse de convergence vers les points attractifs pour l'algorithme et cette vitesse aurait été ralentie par un éventuel recuit.

APPENDICE

Loi d'une série régressive perturbée

Il est souvent utile d'étudier la loi d'une somme d'une série régressive; on peut dans divers cadres prouver qu'elle ne charge aucun point.

Ce type de résultat est obtenu par Lévy [20] et Jessen-Winter [14] lorsque (ε_n) est une suite de variables aléatoires indépendantes de même loi, centrée et de carré intégrable. Des versions relatives aux martingales sont données par Barlow [1], Burkholder [4], Lai-Wei [18] et Wei [33]. En voici la version utilisée dans notre article, proche de celle prouvée sans le terme résiduel Φ par Lai et Wei.

La démonstration qui suit, due à Bernard Delyon dans le cas unidimensionnel et à Pierre Priouret dans le cas multidimensionnel, est meilleure que la nôtre; qu'ils soient remerciés de nous autoriser à la reproduire.

THÉORÈME A. – *Soit* $\mathbb{F} = (\mathcal{F}_n)$ *une filtration et* $\varepsilon = (\varepsilon_n)$ *un « bruit » à valeurs dans* $\mathbb{R}^d$ *adapté à* $\mathbb{F}$: *pour tout* n,

$$E(\| \varepsilon_{n+1} \|^2 \mid \mathcal{F}_n) < \infty,\ E(\varepsilon_{n+1} \mid \mathcal{F}_n) = 0.$$

Soit $\Phi = (\Phi_n)$ *une autre suite de vecteurs aléatoires de dimension* d *adaptée à* $\mathbb{F}$ *et* (c_n) *une suite réelle déterministe avec une infinité de termes non nuls et telle que* $\sum | c_n |^2 < \infty$.

Désignons par H un ensemble de trajectoires pour lesquelles on a :

□ *les conditions de « Marcinkiewick-Zygmund » sur le bruit :*

$$\limsup_n E(\| \varepsilon_{n+1} \|^2 \mid \mathcal{F}_n) < \infty, \quad \liminf E(\| \varepsilon_{n+1} \| \mid \mathcal{F}_n) > 0;$$

□ *Φ est la somme de deux suites adaptées à $\mathbb{F}$, (r_n) et (R_n) avec*

$$\sum \| r_n \|^2 < \infty \text{ et } E\left(1_H \sum_{n=N}^{\infty} \| c_n R_n \|\right) = o\left(\sum_{n=N}^{\infty} | c_n |^2\right)^{1/2}.$$

Sur H, la série $\sum_{n=1}^{\infty} c_n (\Phi_n + \varepsilon_n)$ converge, p.s., vers une variable aléaoire finie L et, pour toute variable aléatoire Y mesurable par rapport à $\mathcal{F}_p$, p entier quelconque, on a :

$$P(H \cap (Y = L)) = 0.$$

Démonstration. – *a*) Simplifications du problème

□ On se ramène à $p = 0$, en considérant la filtration $(\mathcal{F}_{n+p})_{n \geq 0}$ et la relation :

$$\sum_{j=1}^{\infty} c_{j+P} (\Phi_{j+p} + \varepsilon_{j+p}) = \left(Y - \sum_{k=0}^{p} c_k (\Phi_k + \varepsilon_k)\right).$$

□ On peut prendre $Y = 0$; si $q = \inf \{r;\ r \geq p,\ | c_r | > 0\}$, on remplace (Φ_n) par $(\tilde{\Phi}_n)$ avec $\Phi_q - Y/c_q = \tilde{\Phi}_q$, $\Phi_n = \tilde{\Phi}_n$ pour $n \neq q$.

□ Enfin, selon la remarque 2) de I.2, on peut supposer que :

$$\sum \| r_n \|^2 \leq \text{Cte},$$

et, pour deux constantes A et B, p.s. pour tout n :

$$0 < A \leq E(\| \varepsilon_{n+1} \| \mid \mathcal{F}_n) \qquad \text{et} \qquad E(\| \varepsilon_{n+1} \|^2 \mid \mathcal{F}_n) \leq B.$$

□ Dans le cadre précédent, la série régressive $\sum c_n \varepsilon_{n+1}$ converge p.s. Soit $L = \sum_{k=1} c_k (\Phi_k + \varepsilon_k)$.

b) Soit

$$\rho_n = \sum_{k=n}^{\infty} c_k \varepsilon_k \text{ et } \alpha_n^2 = \sum_{k=n}^{\infty} c_k^2.$$

On a :

$$E(\| \rho_n \|^2 \mid \mathcal{F}_{n-1}) \leq B\, \alpha_n^2.$$

Montrons que : $E(\| \rho_n \|^{3/2} \mid \mathcal{F}_{n-1}) \geq C_2\, \alpha_n^{3/2}$, C_2 constante > 0. Remarquons d'abord que, pour $0 < \alpha < 1$ et $a_i \geq 0$, $1 \leq i \leq d$,

$$d^{\alpha-1}(a_1^\alpha + \ldots + a_d^\alpha) \leq (a_1 + \ldots + a_d)^\alpha \leq a_1^\alpha + \ldots + a_d^\alpha;$$

on utilise l'inégalité de Hölder pour la première inégalité et la relation $t^\alpha + (1-t)^\alpha \geq 1$ si $0 \leq t \leq 1$ pour la seconde.

Notant ρ_n^i, $1 \leq i \leq d$, les composantes de ρ_n,

$$E(\| \rho_n \|^{3/2} \mid \mathcal{F}_{n-1}) = E(([\rho_n^1]^2 + \ldots + [\rho_n^d]^2)^{3/4} \mid \mathcal{F}_{n-1})$$
$$\geq d^{-1/4} \sum_{i=1}^{d} E([\rho_n^i]^{3/2}\, \mathcal{F}_{n-1}).$$

Par une inégalité de Burkholder, pour une constante $C_1 > 0$,

$$E([\rho_n^i]^{3/2} \mid \mathcal{F}_{n-1}) \geq C_1\, E\left(\left(\sum_{k=n}^{\infty} c_k^2\, [\varepsilon_k^i]^2\right)^{3/4} \mid \mathcal{F}_{n-1}\right).$$

Par l'inégalité de Hölder,

$$\alpha_n^{-2} \sum_{k=n}^{\infty} c_k^2\, |\varepsilon_k^i|^{3/2} \leq \left(\alpha_n^{-2} \sum_{k=n}^{\infty} c_k^2\, [\varepsilon_k^i]^2\right)^{3/4};$$

de plus : $\displaystyle\sum_{i=1}^{d} |\varepsilon_k^i|^{3/2} \geq \| \varepsilon_k \|^{3/2}$.

Pour $C_2 = C_1\, d^{-1/4}\, A^{3/2}$,

$$E(\| \rho_n \|^{3/2} \mid \mathcal{F}_{n-1}) \geq C_1\, d^{-1/4}\, \alpha_n^{-1/2} \sum_{k=n}^{\infty} c_k^2 \left(E\left(\sum_{i=1}^{d} |\varepsilon_k^i|^{3/2} \mid \mathcal{F}_{n-1}\right)\right)$$
$$\geq C_1\, d^{-1/4}\, \alpha_n^{-1/2} \sum_{k=n}^{\infty} c_k^2\, E(\| \varepsilon_k \|^{3/2} \mid \mathcal{F}_{n-1})$$
$$\geq C_2\, \alpha_n^{3/2}.$$

On en déduit que $E(\| \rho_n \| \mid \mathcal{F}_{n-1}) \geq C\, \alpha_n$, C constante > 0, par les inégalités :

$$C_2^2\, \alpha_n^3 \leq (E(\| \rho_n \|^{3/2} \mid \mathcal{F}_{n-1}))^2 \leq E(\| \rho_n \| \mid \mathcal{F}_{n-1})\, E(\| \rho_n \|^2 \mid \mathcal{F}_{n-1}).$$

c) Posons :

$$S_n = \sum_{k=1}^{n} c_k (\Phi_k + \varepsilon_k),\ \tau_n = \sum_{k=n}^{\infty} c_k \Phi_k;$$
$$T_n = L - S_{n-1} = \rho_n + \tau_n.$$

Soit $a \to U(a)$ une fonction borélienne de $\mathbb{R}^d \backslash 0$ dans l'ensemble des matrices $d \times d$ orthogonales telle que :

$$U(a)\,[a/\| a \|] = e_1,$$

e_1 étant le premier vecteur de la base canonique, ${}^T e_1 = (1, 0, \ldots, 0)$.

Le premier vecteur ligne de $U(a)$ est $a/\| a \|$.

Sur $G = \{L = 0\} \cap H$, pour tout n, $S_{n-1} + T_n = 0$ et :

$$\| T_n \|\, e_1 + U(S_{n-1})\, T_n = 0,$$
$$\| \| \rho_n \|\, e_1 + U(S_{n-1})\, \rho_n \| \leq | \| \rho_n \| - \| T_n \| | + \| \rho_n - T_n \| \leq 2 \| \tau_n \|.$$

Pour $G_n = \{P(G \mid \mathcal{F}_n) > 0\}$, $E(|\, 1_G - 1_{G_n}\,|)$ tend vers 0 si $n \to \infty$.

$$P(G_{n-1}) \leq [C\,\alpha_n]^{-1} \| E(1_{G_{n-1}} E(\| \rho_n \|\, e_1 | \mathcal{F}_{n-1})) \|$$
$$= [C\,\alpha_n]^{-1} \| E(1_{G_{n-1}} E(\| \rho_n \|\, e_1 + U(S_{n-1})\, \rho_n \mid \mathcal{F}_{n-1})) \|$$
$$\leq [C\,\alpha_n]^{-1} \| E(1_G E(\| \| \rho_n \|\, e_1 + U(S_{n-1})\, \rho_n \| \mid \mathcal{F}_{n-1}))$$
$$+ E(|\, 1_G - 1_{G_n} \mid E(2 \| \rho_n \| \mid \mathcal{F}_{n-1}))$$
$$\leq \text{Cte}\ \alpha_n^{-1} (E(\| \rho_n \|^2)\, E(|\, 1_G - 1_{G_n}\,|))^{1/2} + \text{Cte}\ \alpha_n^{-1} E(1_G \| \tau_n \|).$$

Avec la simplification effectuée en *a*), $\sum \| r_n \|^2 \leq \text{Cte}$, et

$$E(1_G \| \tau_n \|) \leq \text{Cte}\ \alpha_n \left(\left(\sum_{k=n}^{\infty} \| r_k \|^2 \right)^{1/2} \right)$$
$$+ \text{Cte}\ E\left(1_H \left(\sum_{k=n}^{\infty} \| c_k R_k \| \right) \right) = o(\alpha_n).$$

La suite $(P(G_n))$ tend vers 0 et $P(G) = 0$. ■

COROLLAIRE B. – *Le théorème A est encore valable en remplaçant les suites* (c_n), (ε_n) *et* (Φ_n) *par des suites complexes unidimensionnelles.*

Démonstration. – On se ramène à c_n réel en écrivant

$$c_n (\varepsilon_n + \Phi_n) = | c_n | (\varepsilon'_n + \Phi'_n).$$

On applique alors le théorème A au modèle bidimensionnel formé par les parties réelles et imaginaires. ■

RÉFÉRENCES

[1] W. J. BARLOW, *Coefficient properties of random variable sequences.* Ann. of probability, vol. **3**, 1975, p. 840-848.

[2] A., BENVENISTE, M. MÉTIVIER et P. PRIOURET, *Algorithmes adaptatifs et approximations stochastiques.* Masson, 1987.

[3] O. BRANDIÈRE, *Un algorithme du gradient pour l'analyse en composantes principales.* Comptes Rendus Académie des Sciences, **321**, série I, 1995, p. 233-236.

[4] D. L. BURKHOLDER, *Independent sequences with the Stein property,* Ann. of math. stat., vol. **39**, 1968, p. 1282-1288.

[5] B. DELYON, *A deterministic approach to stochastic approximation,* IRISA, prépublication n° 789, 1994.

[6] D. P. DEREVITSKII et A. L. FRADKOV, *Two models for analysing the dynamics of adaption algorithms.* Avtomatika i Telemekhanika, 1, 1974, p. 67-75; en anglais, Automation and remote control, vol. **1**, 1974, p. 59-67.

[7] M. DUFLO, *Méthodes récursives aléatoires.* Masson, 1990.

[8] J. C. FORT et G. PAGÈS, *Sur la convergence presque sûre d'algorithmes stochastiques : le théorème de Kushner-Clark revisité.* Prépublication du SAMOS 33, Université Paris 1, 1994.

[9] S. B. GELFAND et S. K. MITTER, *Recursive stochastic algorithms for global optimization in* $\mathbb{R}^d$. SIAM J. control and optimization, vol. **29**, 1991, p. 999-1018.

[10] S. B. GELFAND et S. K. MITTER, *Metropolis-type annealing algorithms for global optimization in* $\mathbb{R}^d$. SIAM J. control and optimization, vol. **31**, 1993, p. 111-131.

[11] P. HARTMAN, *Ordinary Differential Equations,* Wiley, 1964; seconde édition, 1982.

[12] K. HORNIK et C. M. KUAN, *Convergence analysis of local feature extraction algorithms.* Neural networks, vol. **5**, 1992, p. 229-240.

[13] C. R. HWANG et S. J. SHEU, *On the behaviour of a stochastic algorithm with annealing.* Rapport technique, Academia sinica, Taiwan, 1990.

[14] B. JESSEN et A. WINTNER, *Distribution functions and the Riemann zeta function.* Trans. american math. soc., vol. **38**, 1935, p. 725-734.

[15] J. KIEFER et J. WOLFOWITZ, *Stochastic estimation of the maximum of a regression fonction.* Ann. math. stat., vol. **23**, 1952, p. 462-466.

[16] H. J. KUSHNER, *Asymptotic global behavior for stochastic approximation and diffusions with slowly decreasing noise effects: Global minimization via Monte Carlo.* SIAM J. Appl. Math., **47**, 1987, p. 169-185.

[17] H. J. KUSHNER et D. S. CLARK, *Stochastic approximation for constrained and unconstrained systems.* Applied math. science series, vol. **26**, Springer, 1978.

[18] T. Z. LAI et C. Z. WEI, *A note on martingale difference sequences satisfying the local Marcinkiewicz-Zygmund condition.* Bull. of the institute of mathematics, Academia Sinica, vol. **11**, 1983, p. 1-13.

[19] V. A. LAZAREV, *Convergence of stochastic-approximation procedures in the case of a regression equation with several roots.* Problems of Information Transmission, vol. **28**, 1992, p. 66-78; en russe, Problemy Peredachi Informatsii, vol. **28**, 1992, p. 75-88.

[20] P. LÉVY, *Sur les séries dont les termes sont des variables éventuellement indépendantes.* Studia math., vol. **3**, 1931, p. 119-155.

[21] L. LJUNG, Analysis of recursive stochastic algorithms. *IEEE Trans. Automatic Control,* vol. **22**, 1977, p. 551-575.

[22] L. LJUNG, G. PFLUG et H. WALK, *Stochastic approximation of random systems.* Birkhäuser, 1992.

[23] L. LJUNG et T. SODERSTRÖM, *Theory and practice of recursive identification,* MIT Press, 1983.

[24] S. P. MEYN et R. L. TWEEDIE, *Markov chains and stochastic stability.* Springer, 1993.

[25] M. B. NEVEL'SON et R. Z. HAS'MINSKII, *Stochastic approximation and recursive estimations.* Nauka, Moscou, 1972 – Translation of math. monographs, vol. **47**, American Mathematical Society, 1973.

[26] E. NUMMELIN, *General irreductible Markov chains and nonnegative operators.* Cambridge university press, 1984.

[27] E. OJA, *Principal networks, principal components, and linear neural networks.* Neural networks, vol. **5**, 1992, p. 927-935.

[28] E. OJA et J. KARHUNEN, *On stochastic approximation of the eigenvalues of the expectation of a random matrix.* J. of math. analysis and appl., vol. **106**, 1985, p. 69-84.

[29] H. POINCARÉ, *Mémoire sur les courbes définies par une équation différentielle (IV).* J. Math. Pures Appl., vol. **4**, 1886, p. 151-217.

[30] H. ROBBINS et S. MONRO, *A stochastic approximation method.* Ann. math. stat., vol. **22**, 1951, p. 400-407.

[31] L. YOUNES, *Estimation and annealing gor Gibbsian fields.* Ann. Inst. Henri Poincaré, vol. **24**, 1988, p. 269-294.

[32] L. YOUNES, *Parametric inference of imperfectly observed Gibbsian fields.* Probability theory and related fields, vol. **82**, 1989, p. 625-645.

[33] C. Z. WEI, *Martingale transforms with non-atomic limits and stochastic approximation.* Probability theory and related fieds, vol. **95**, 1993, p. 103-114.

Référence complémentaire (1995). Après la rédaction de ce texte, nous avons pris connaissance de l'article suivant dont l'esprit est proche du nôtre (avec des hypothèses plus restrictives) :

R. PEMANTLE, *Non convergence to unstable points in urn models and stochastic approximations.* Ann. of Probability, vol. **18**, 1990, p. 698-712.