Statistique et analyse des données

B. Fichet

Sur les résultats de Blum en approximation et optimisation stochastiques multidimensionnelles

Statistique et analyse des données, tome 3, n° 3 (1978), p. 57-68

<http://www.numdam.org/item?id=SAD_1978__3_3_57_0>

Statistique et Analyse des Données
3 - 1978 pp. 57, 68

SUR LES RESULTATS DE BLUM EN APPROXIMATION ET OPTIMISATION STOCHASTIQUES MULTIDIMENSIONNELLES

B.FICHET

Laboratoire de Physique. Faculté de Médecine.
Université de Marseille II.

L'extension au cas multidimensionnel des processus d'approximation stochastique de Robbins-Monro et d'optimisation stochastique de Kiefer-Wolfowitz, fut l'oeuvre de Blum. On lui doit trois importants théorèmes, deux pour l'approximation et un pour l'optimisation. Depuis, de nombreux auteurs ont établi des théorèmes de convergence tant pour ces processus que pour des processus généralisés.
Nous montrons par cette note que les résultats de Blum restent d'une grande actualité. Pour différentes régressions les convergences des processus de Robbins-Monro ou de Kiefer-Wolfowitz peuvent être déduites tant des théorèmes de Blum que de théorèmes postérieurs. Ainsi, à l'aide des théorèmes de Blum, nous démontrons la convergence pour l'approximation d'une régression linéaire donnée par une matrice définie négative, sous l'hypothèse de la variance uniformément bornée ; et même sans cette hypothèse à l'aide d'un exemple. Et nous établissons de même la convergence pour l'optimisation d'une régression de type paraboloïde.

1 - INTRODUCTION

Rappelons les processus d'approximation stochastique de Robbins-Monro et d'optimisation stochastique de Kiefer-Wolfowitz dans $\mathbb{R}^k$.
Pour l'approximation, on considère une famille $\{Y_x\}$ de vecteurs aléatoires à valeurs dans $\mathbb{R}^k$, indicée par le paramètre $x \in \mathbb{R}^k$. On suppose que : $\forall x \in \mathbb{R}^k$, $E[Y_x] = M(x)$ existe, où M (inconnue) est Borel-mesurable. On se propose alors, non pas d'approcher la fonction inconnue M, mais d'estimer la solution Θ de l'équation $M(x)=\alpha$ ($\alpha \in \mathbb{R}^k$, donné) ; dans ce but, on construit le processus de Robbins-Monro défini par :

(1.1) $X_{n+1} - X_n = a_n(Y_n - \alpha)$ $(n \in \mathbb{N}^*)$ où :

- $X_1 \in \mathbb{R}^k$ est donné.
- Y_n est un vecteur aléatoire ayant même loi de probabilité (conditionnelle) que Y_{x_n} quand x_n est la réalisation de X_n.
- $\{a_n\}$ est une suite à termes positifs vérifiant : $\sum_{n=1}^{\infty} a_n = +\infty$; $\sum_{n=1}^{\infty} a_n^2 < +\infty$

Dans le processus de Markov ainsi construit, notons que $M(X_n)$ est alors une version de $E[Y_n | X_1,\ldots,X_n] = E[Y_n | X_n]$, pour tout n.

Pour l'optimisation, on considère une famille $\{Y_x\}$ de variables aléatoires, indicée par le paramètre $x \in \mathbb{R}^k$. On suppose que : $\forall\, x \in \mathbb{R}^k$, $E[Y_x] = M(x)$ existe, où M (inconnue) est Borel-mesurable. $\{u_1,\ldots,u_k\}$ étant une base orthonormée de $\mathbb{R}^k$, pour tout réel $c > 0$, soient $Y^*_{x-cu_1}$, $Y^*_{x+cu_1}$,..., $Y^*_{x-cu_k}$, $Y^*_{x+cu_k}$, 2k variables aléatoires indépendantes de même loi que, respectivement, Y_{x-cu_1}, Y_{x+cu_1},..., Y_{x-cu_k}, Y_{x+cu_k}, et définissons $Y_{x,c}$ par : $Y_{x,c} = [(Y^*_{x+cu_1} - Y^*_{x-cu_1}),\ldots,(Y^*_{x+cu_k} - Y^*_{x-cu_k})]$

Alors, pour estimer $\Theta \in \mathbb{R}^k$ qui maximise M(x), on construit le processus de Kiefer-Wolfowitz défini par :

(1.2) $X_{n+1} - X_n = (a_n/c_n)Y_n$ $(n \in \mathbb{N}^*)$ où :

- $X_1 \in \mathbb{R}^k$ est donné
- Y_n est un vecteur aléatoire ayant même loi de probabilité (conditionnelle) que Y_{x_n,c_n} quand x_n est la réalisation de X_n.
- $\{a_n\}$ et $\{c_n\}$ sont deux suites à termes positifs vérifiant :

$$\lim_{n\to+\infty} c_n = 0 \; ; \; \sum_{n=1}^{\infty} a_n = +\infty \; ; \; \sum_{n=1}^{\infty} a_n c_n < +\infty \; ; \; \sum_{n=1}^{\infty} (a_n/c_n)^2 < +\infty \; .$$

Similairement, Blum définit un processus par :

(1.3) $X_{n+1} - X_n = (a_n/c_n)Z_n$ $(n \in \mathbb{N}^*)$ où :

Z_n est un vecteur aléatoire ayant même loi de probabilité (conditionnelle) que Z_{x_n,c_n} et où :

$Z_{x,c} = [(Y^*_{x+cu_1} - Y^*_x),\ldots,(Y^*_{x+cu_k} - Y^*_x)]$ est défini comme $Y_{x,c}$

Les premiers théorèmes de convergence, pour les processus multidimensionnels cités, sont ceux de Blum [2]; nous les présentons au paragraphe suivant. Depuis, ces résultats ont été repris et améliorés - voir par exemple Macchi [13] chapitre I, et Daubeze[5] - ; des théorèmes de convergence, tant pour ces processus que pour des processus généralisés, établis sous des hypothèses faibles, peuvent être trouvés dans Dvoresky [9], Derman et Sacks [7], Fabian [10], Venter [16], et dans les thèses de Hiriart-Urruty [12] chapitres I, II, III, et Bertran [3] chapitre I ; pour une bibliographie sur le sujet, on peut consulter Schmetterer [15], Dupac et Ivanov [8] et Wasan [17] .
Nous nous proposons de montrer par cette note, que l'approximation et l'optimisation de certaines régressions, qui se déduisent des travaux plus récents, peuvent tout aussi bien être obtenues à l'aide des théorèmes originels de Blum.

2 - LES THEOREMES DE BLUM

Pour l'approximation nous utilisons les notations suivantes :

- $x^+ = \frac{1}{2}\left[\,|x| + x\right]$ pour toute variable réelle x.
- D^2 est l'espace des fonctions de $\mathbb{R}^k$ dans $\mathbb{R}$, admettant des dérivées partielles premières et secondes continues.
- $\forall f \in D^2$, ∇f et A_f sont respectivement le gradient et la matrice des dérivées secondes de f.

On a alors : $\forall f \in D^2$, $f(x+h) = f(x) + \langle \nabla f(x),\, h \rangle + \frac{1}{2} \langle h,\, A_f(x+t\,h)\, h \rangle$
avec : $0 \leqslant t \leqslant 1$, $t \equiv t\,(x,h)$.

Nous notons encore :

- $U_f(x) = \langle \nabla f(x),\, M(x) - \alpha \rangle$
- $\forall a \in \mathbb{R}$, $V_f(a,x) = E\left\{ \langle Y_x - \alpha, A_f\left[x + t_1\, a(Y_x - \alpha)\right] (Y_x - \alpha)\right\}$

avec $t_1 \equiv t\left[x, a(Y_x - \alpha)\right]$.

Alors, lorsque $M(\Theta) = \alpha$, pour Θ et α quelconques, les théorèmes d'approximation de Blum s'énoncent :

Théorème 1

S'il existe $f \in D^2$, satisfaisant à :

(2.1) $\forall\, x \in \mathbb{R}^k,\ f(x) \geqslant 0$

(2.2) $\forall\, \varepsilon > 0,\quad \inf_{\varepsilon \leqslant ||x-\Theta||} |f(x) - f(\Theta)| > 0$

(2.3) $\forall\, \varepsilon > 0,\quad \sup_{\varepsilon \leqslant ||x-\Theta||} U_f(x) < 0$

(2.4) $\exists\ V < +\infty : \forall\, x \in \mathbb{R}^k,\ \forall\, a \in \mathbb{R},\ V_f(a,x) \leqslant V$

Alors la suite $\{X_n\}$ *définie par (1.1) converge presque sûrement vers* Θ.

Théorème 2

S'il existe $f \in D^2$ *satisfaisant à (2.1), (2.2) et à :*

(2.5) $\forall \delta > 0\ \exists\ \lambda(\delta) > 0 : \forall\, a \in \mathbb{R},\quad \sup_{\delta \leqslant ||x-\Theta||} \left[U_f(x) + \lambda(\delta)\, V^+_f(a,x)\right] < 0$

(2.6) $\exists\ \varepsilon > 0,\ V < +\infty : \forall\, a \in \mathbb{R},\ \sup_{0 \leqslant ||x-\Theta|| < \varepsilon} V_f(a,x) \leqslant V$

Alors la suite $\{X_n\}$ *définie par (1.1) converge presque sûrement vers* Θ.

Notons que la démonstration tient encore si (2.5) n'est satisfaite que pour $\delta = \varepsilon$ définie par (2.6), pourvu que (2.3) soit aussi satisfaite. Signalons enfin que notre énoncé du théorème 2 diffère légèrement de celui de Blum.

Lorsque Θ est quelconque, le théorème d'optimisation de Blum s'énonce :

Théorème 3

Si :

(2.7) *M admet des dérivées partielles premières et secondes continues*

(2.8) $\forall\, x \in \mathbb{R}^k,\ \sigma_x^2 = var\,(Y_x) \leqslant \sigma^2 < +\infty$

(2.9) $\forall\, \varepsilon > 0,\ \exists\ \rho(\varepsilon) > 0 : ||x-\Theta|| \geqslant \varepsilon \Longrightarrow M(x) - M(\Theta) \leqslant -\rho(\varepsilon)$ *et* $||\nabla M(x)|| \geqslant \rho(\varepsilon)$

(2.10) $\forall\, x \in \mathbb{R}^k,\ \forall i,j = 1,\ldots,k \mid \partial^2 M(x)/\partial x_i\, \partial x_j \leqslant C$

Alors la suite $\{X_n\}$ *définie par (1.3) converge presque sûrement vers* Θ.

Notons que la convergence est également assurée pour le processus défini par (1.2) ; à quelques détails prés, la démonstration est semblable à celle de Blum.

3 - APPLICATION A L'APPROXIMATION DE CERTAINES REGRESSIONS LINEAIRES.

Nous envisageons le cas d'une régression linéaire caractérisée par une matrice définie négative. Ce cas est ébauché par Blum, dans son exemple 2 - déduit du théorème 2 - et que nous présentons pour Θ et α quelconques sous la forme du :

Corollaire 1

Si :

(3.1) $M(x) = Bx$ *ou* B *est une matrice définie négative.*

(3.2) $\exists\ \varepsilon > 0,\ \exists\ C > 0 : ||x-\Theta|| \leqslant \varepsilon \Longrightarrow E\{||Y_x - \alpha||^2\} \leqslant C$

(3.3) $\exists\ \rho(\Theta) > 0 : ||x-\Theta|| > \varepsilon \Longrightarrow \langle x-\Theta,\ B(x-\Theta)\rangle + \rho(0)\, E\{||Y_x - \alpha||^2\} \leqslant 0$

Alors la suite $\{X_n\}$ *définie par (1.1) converge presque sûrement vers* Θ.

Les conditions données par Blum dépendent encore du second membre α de l'équation et de la solution Θ. Nous pouvons maintenant démontrer le :

Corollaire 2

Si (3.1) est vérifiée, et, notant Y_x^i *la ième composante de* Y_x, *si :*

(3.4) $\exists\ \sigma > 0 : \forall\ x \in \mathbb{R}^k,\ \forall i = 1,\ldots,k\quad (\sigma_x^i)^2 = var(Y_x^i) \leqslant \sigma^2$

Alors la suite $\{X_n\}$ *définie par (1.1) converge presque sûrement vers* Θ.

Démonstration :

Appliquons le corollaire 1. On a :

$$E\{\| Y_x - \alpha\|^2\} = E\{\| Y_x\|^2\} - 2 < Bx, \alpha > + \|\alpha\|^2$$

$$\leqslant k\sigma^2 + \|Bx\|^2 + 2\,\|Bx\| .\|\alpha\| + \|\alpha\|^2$$

$$\leqslant k\sigma^2 + |||B|||^2 .\ \|x\|^2 + 2|||B|||.\|\alpha\| .\|x\| + \|\alpha\|^2$$

où $|||.|||$ est la norme classique d'une application linéaire.

Dès lors, $\forall\varepsilon>0$, (3.2) est vérifiée, puisque : $\|x-\Theta\| \leqslant \varepsilon \Longrightarrow \|x\| \leqslant \varepsilon + \|\Theta\|$.

Seule donc, reste à vérifier (3.3) pour un certain $\varepsilon > 0$.

B étant définie négative : $\exists\ b > 0 : \forall\ x \in \mathbb{R}^k\ <x, Bx> \leqslant -b\,\|x\|^2$

d'où : $< x-\Theta, B(x-\Theta)> \leqslant -b\,\|x-\Theta\|^2 \leqslant -b\,(\|x\| - \|\Theta\|)^2$

Alors $\rho(\Theta)$ étant à déterminer :

$$< x-\Theta, B(x-\Theta) > + \rho(\Theta)\, E\{\| Y_x - \alpha\|^2\}$$

$$\leqslant -b(\|x\| - \|\Theta\|)^2 + \rho(\Theta)\,[k\sigma^2 + |||B|||^2.\|x\|^2 + 2\,|||B|||.\,\|\alpha\| .\|x\| + \|\alpha\|^2]$$

$$\leqslant [-b+\rho(\Theta)|||B|||^2]\ \|x\|^2 + 2\,[b\,\|\Theta\| + \rho(\Theta)|||B|||.\,\|\alpha\|]\,\|x\|$$

$$+ [-b\,\|\Theta\|^2 + \rho(\Theta)k\sigma^2 + \rho(\Theta)\|\alpha\|^2] \equiv K(x) \qquad \text{(posé)}$$

Si nous choisissons $0 < \rho(\Theta) < b/\ |||B|||^2$

$K(x) \to -\infty$ quand $\|x\| \to \infty$

Donc il existe $\varepsilon' > 0$ tel que :

$\|x\| \geqslant \varepsilon' \Longrightarrow < x - \Theta, B(x-\Theta)> + \rho(\Theta)\, E\{\|Y_x - \alpha\|^2\} \leqslant K(x) \leqslant 0$

Dès lors, la démonstration est complète avec $\varepsilon=\varepsilon' + \|\Theta\|$ et $\rho(\Theta)$ précédemment déterminé. □

Ainsi, pour une régression linéaire caractérisée par une matrice définie négative, la convergence est assurée quel que soit α (ou Θ), sous la seule hypothèse de la variance uniformément bornée. Wolfowitz [18] a montré que celle-ci ne pouvait être omise dans l'étude originelle de Robbins et Monro. Friedman [11] fut le premier à

démontrer la convergence sans cette hypothèse, mais en transformant le processus.
L'exemple suivant montre que les résultats de Blum, bien antérieurs à ceux de Friedman, ne l'exigeaient déjà pas.

Exemple

$\forall\, p \in \mathbb{R}$, soit f_p la fonction de $\mathbb{R}$ dans $\mathbb{R}$ définie par :

$$\forall\, t \in \mathbb{R}, f_p(t) = \begin{cases} (1-m)^2\, m\, e^{-m(1-m)t} & \text{si } t \geqslant 0 \\ m^2\,(1-m) e^{m(1-m)t} & \text{si } t < 0 \end{cases}$$

$$\text{avec } m = \begin{cases} \left[p+2 - \sqrt{4+p^2}\right]/2p & \text{si } p \neq 0 \\ 1/2 & \text{si } p = 0 \end{cases}$$

Il est aisé de voir que lorsque p croît de $-\infty$ à $+\infty$, m décroît de 1 à 0, et que f_p est une densité de probabilité.
Si X est une variable aléatoire dont la loi de probabilité admet une densité f_p, de simples calculs montrent que :

$$E[X] = p \; ; \quad E\left[X^2\right] = \frac{2}{m^2(1-m)^2} \; ; \; \text{Var}\,(X) = \frac{-4m^2 + 4m + 1}{m^2(1-m)^2}$$

Soit B une matrice définie négative d'ordre (k,k). Notons $B^i x$ la ième composante de Bx (B^i est une forme linéaire).
Supposons que :
$\forall\, x \in \mathbb{R}^k$, $\forall\, i = 1,\ldots,k$, la loi de Y_x^i admet une densité f_{p_i}, avec $p_i \equiv B^i x$.
Alors les relations précédentes montrent que :

$$M(x) = E\left[Y_x\right] = Bx \; ; \quad \sup_{x \in R^k} \text{Var}\,(Y_x^i) = +\infty$$

En appliquant le corollaire 1, montrons la convergence du processus (1.1) pour tout Θ.
La variation de m en fonction de p, montre que :

$$\forall\, A > 0, \; \exists\, r(A) : |p| \leqslant A \Longrightarrow r(A) \leqslant m \leqslant 1 - r(A)$$

avec $0 < r(A) < 1/2$

Notant $\forall i = 1,\ldots,k \; M_2^i = E\left[(Y_x^i)^2\right]$;

$$E\left\{\| Y_x - \alpha \|^2\right\} \leqslant \sum_{i=1}^{k} M_2^i + 2\, \|\alpha\| \,.\, |||B||| \,.\, \|x\| + \|\alpha\|^2$$

$$\text{Or : } \|x\| \leqslant \varepsilon \Longrightarrow |p_i| \leqslant |||B^i|||.\varepsilon \Longrightarrow M_2^i \leqslant 2/\left[r(|||B^i|||\;\varepsilon)\right]^4.$$

Dès lors, $\forall\, \varepsilon > 0$ (3.2) est vérifiée.

Seule reste à vérifier (3.3) pour un certain $\varepsilon > 0$.

$\rho(\Theta)$ étant à déterminer :

$$< x-\Theta,\ B(x-\Theta) > + \rho(\Theta)\ E\{\| Y_x - \alpha\|^2\}$$

$$\leqslant -b(\| x\| - \| \Theta\|)^2 + \rho(\Theta)\ \Big[\sum_{i=1}^{k} M_2^i + 2\ \| \alpha\| .\ |||B|||\ .\| x\| + \| \alpha\|^2\ \Big]$$

$$\leqslant -b\ \| x\|^2 + \rho(\Theta) \sum_{i=1}^{k} M_2^i + 2\ \big[b\ \| \Theta\| + \rho(\Theta)\ \| \alpha\| .|||B|||\ \big]\ \| x\| + \big[-b\ \| \Theta\|^2 + \rho(\Theta)\ \| \alpha\|^2\big]$$

Cherchons une borne supérieure pour M_2^i.

Un calcul simple montre que :

$M_2^i \sim 2p_i^2$ quand $p_i \to +\infty$; $M_2^i \sim 2p_i^2$ quand $p_i \to -\infty$.

Soit $\delta > 0$ donné ; alors, il existe $A > 0$:

$$|p_i| > A \Longrightarrow M_2^i \leqslant 2\ (1+\delta)\ |||B^i|||^2\ \ \| x\|^2$$

$$|p_i| \leqslant A \Longrightarrow M_2^i \leqslant 2/[r(A)]^4$$

Soit $\varepsilon' > 0$ tel que : $2/\ [r(A)]^4 \leqslant 2(1+\delta)\ K^2\ \varepsilon'^2$ où $K = \max\limits_i\ |||B^i|||$.

Alors $\| x\| \geqslant \varepsilon' \Longrightarrow M_2^i \leqslant 2\ (1+\delta)\ K^2\ \| x\|^2$ $i=1,\ldots,k$.

Dès lors, la démonstration suit celle du corollaire 2. avec $\rho(\Theta) < b/2\ k(1+\delta)\ K^2$. □

Remarque

Comme exemple d'application du théorème 1, Blum envisage le cas d'une régression linéaire dans une boule, et constante à l'extérieur, soit :

$$M(x) = \begin{cases} Bx & \text{si } \| x\| \leqslant \rho \\ \big[\rho/\| x\|\ \big]Bx & \text{si } \| x\| > \rho \end{cases}$$

où B est une matrice (k,k) définie négative.

Avec $f(x) = \| x\|^2$, Blum démontre, sous l'hypothèse (3.4) de la variance uniformément bornée, la convergence du processus (1.1).

Mais cet exemple reste d'une portée limitée si on ne peut établir la convergence pour tout Θ vérifiant $\| \Theta\| < \rho$. Si, par une naturelle généralisation, on prend $f(x) = \| x-\Theta\|^2$, il est clair que (2.1) et (2.2) sont vérifiées, et que, comme dans [2], (2.4) l'est aussi. Montrons par un contre-exemple, que (2.3) peut ne pas être satisfaite si $k>1$.

Contre-exemple :

Définissons B comme une matrice diagonale, d'éléments diagonaux $\lambda_1,\ldots,\lambda_k$ négatifs.

Si nous trouvons x tel que $< x-\Theta, M(x)-\alpha > > 0$, (2.3) ne sera pas satisfaite.

Notons $x = (x^1,\ldots,x^k)$, $\Theta = (\Theta^1,\ldots,\Theta^k)$,
et choisissons x et Θ tels que :

$$\|x\| > \rho,\ 0 < \rho\, x^1/\|x\| < \Theta^1 < x^1$$

(ex : k=2, $\rho=x^1 = 1$, $x^2 = \sqrt{3}$, $\Theta^1 = 3/4$, $\Theta^2 = 0$)

Ceci entraîne :

$$\lambda_1(x^1-\Theta^1)(\rho x^1/\|x\| - \Theta^1) > 0$$

(ex : $-\lambda_1/16 > 0$)

et puisque $\|\Theta\| < \rho$, il existe un indice i supérieur à 1, tel que :

$$|\Theta^i| < \rho\ |x^i|\ /\|x\| < |x^i|.$$

Fixons $\lambda_j, j=2,\ldots,k$ (ex : $\lambda_2 = -1$)

Puisque B est diagonale et que $\|x\| > \rho$,

$$\frac{1}{2}U_f(x) = \lambda_1(x^1 - \Theta^1)(\rho x^1/\|x\| - \Theta^1) + \sum_{j=2}^{k} \lambda_j(x^j - \Theta^j)(\rho x^j/\|x\| - \Theta^j).$$

Si le second terme est positif ou nul (impossible si k=2), $U_f(x) > 0$;

et, s'il est négatif, il suffit de choisir :

$$\lambda_1 < -\frac{\sum_{j=2}^{k} \lambda_j(x^j-\Theta^j)(\rho x^j/\|x\| - \Theta^j)}{(x^1-\Theta^1)(\rho x^1/\|x\| - \Theta^1)} < 0 \qquad (\text{ex : } \lambda_1 < -24)$$

pour avoir encore $U_f(x) > 0$.

□

Si k=1, $|\Theta|<\rho$, il est aisé de prouver la convergence ; mais c'est un résultat antérieur à l'article de Blum - voir [1]- .

4 - APPLICATION A L'OPTIMISATION D'UNE REGRESSION DE TYPE PARABOLOIDE

Derman [6] établit le premier la convergence du processus de Kiefer-Wolfowitz vers le point où une regression parabolique M(x) atteint son maximum. Outre que le résultat n'est qu'unidimensionnel des hypothèses restrictives sont exigées et la convergence n'a lieu qu'en probabilité. Par application du théorème 3 de Blum, antérieur aux travaux de Derman, on obtient un résultat infiniment plus fort ; c'est notre :

Corollaire 3

Si (2.8) est satisfaite et si :

(4.1) $M(x) \equiv K - K' \| A(x-\Theta) \|^2$

où $K \in \mathbb{R}$, $K' > 0$, *A est une matrice d'ordre (k,k) régulière.*

Alors les suites $\{X_n\}$ *définies par (1.2) et (1.3) convergent presque sûrement vers* Θ.

Pour la démonstration, nous avons besoin du :

Lemme.
Si A est une matrice régulière d'ordre (k,k) :
$\forall \varepsilon > 0 \ \exists \ q(A,\varepsilon) > 0 : \| x \| \geqslant \varepsilon \Longrightarrow \| Ax \| \geqslant q(A,\varepsilon)$.

Démonstration :
S'il n'en était pas ainsi, il existerait $\varepsilon > 0$ tel que pour tout $q > 0$ il existerait x_q vérifiant : $\| x_q \| \geqslant \varepsilon$ et $\| Ax_q \| < q$.
L'application $x \longmapsto \| Ax \|$ est continue. Sur le compact défini par $\| x \| = \varepsilon$ elle atteint sa borne inférieure. Alors, puisque A est régulière, il existe R tel que :

$$\| x \| = \varepsilon \Longrightarrow \| Ax \| \geqslant R > 0.$$

Choisissons $q < R$.

Alors : $q > \| Ax_q \| = \left[\| x_q \| / \varepsilon \right] \| A \left[\varepsilon \, x_q / \| x_q \| \right] \| \geqslant \left[\| x_q \| / \varepsilon \right] R \geqslant R$

D'où la contradiction.

Démonstration du corollaire :
Appliquant le théorème 3., seules (2.9) et (2.10) sont à vérifier.
(2.10) est triviale.
Utilisant le lemme :
$\forall \varepsilon > 0$, $\| x-\Theta \| \geqslant \varepsilon \Longrightarrow M(x) - M(\Theta) \leqslant - K' \, q^2 (A,\varepsilon)$; la première partie de (2.9) est vérifiée.
Un calcul simple montre que :
$\nabla M(x) = - 2 K'A'A(x-\Theta)$ où A' est la transposée de A.

Alors, appliquant le lemme :
$\forall \varepsilon > 0$, $\| x-\Theta \| \geqslant \varepsilon \Longrightarrow \| Ax-A\Theta \| \geqslant q(A,\varepsilon) \Longrightarrow \| \nabla M(x) \| \geqslant 2K' \, q[A', q(A,\varepsilon)]$.

La seconde partie de (2.9) est prouvée et la démonstration est complète. □

5 - UN EXEMPLE NUMERIQUE

Nous proposons un exemple numérique, destiné à illustrer la convergence du processus de Kiefer-Wolfowitz. Les résultats ont été obtenus par simulation. Pour des exemples relevant de cas concrets, on pourra consulter par exemple [14].

Nous nous plaçons dans $\mathbb{R}^{10}$, de point courant $x = (x^1,\ldots,x^{10})$. La régression est de la forme $M(x) = K - \sum_{i=1}^{10} (a_i\, x^i + b_i)^2$ avec :

$K \in \mathbb{R}$, $\forall i=1,\ldots,10$ $a_i \in \mathbb{R}_+^*$, $b_i \in \mathbb{R}$.

Cette régression admet donc un maximum unique pour $\Theta = (\Theta^1,\ldots,\Theta^{10})$ tel que :

$\Theta^i = - b_i/a_i$, $i=1,\ldots,10$.

On suppose que $\forall x \in \mathbb{R}^{10}$, Y_x suit une loi uniforme sur $]\, M(x)-a,\ M(x)+a[$ $(a>o)$; pour x donné, une observation de Y_x est obtenue par génération d'un "nombre au hasard" sur $]0,1[$ et transformation linéaire affine de ce nombre.

On vérifie aisément que les hypothèses du théorème 3 de Blum, sont satisfaites.

Les suites $\{a_n\}$ et $\{c_n\}$ intervenant dans les processus, sont les suites classiques, de type $1/n$ et $1/n^{1/3}$; toutefois, nous les "initialisons" à des valeurs correspondant à n égal à un nombre entier positif donné ; de sorte que nous prenons :

$a_n = \frac{1}{\nu+n}$; $c_n = \frac{1}{(\nu+n)^{1/3}}$ $(\nu \in \mathbb{N}$ donné).

Pour la régression, les valeurs numériques choisies sont :

K=5, a=10

a_i, i=1,...,10	2	1	5	8	3	7	4	9	6	5
b_i, i=1,...,10	35	0	-4	-17	11	-83	-25	50	10	2
Θ^i, i=1,..., 10	-17,5	0,00	0,80	2,12	-3,66	11,85	6,25	-5,55	-1,66	-0,40

Nous donnons deux exemples, pour deux valeurs de ν particulières. Par chacun d'eux, le vecteur X_1 de début est l'origine ; les processus définis en (1.2) et (1.3) sont respectivement notés P2 et P3.

On obtient alors :

Cas $\nu = 0$

n	Processus	x_n^i, i=1,...,10									
1	P3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
9	P3	3×10^2	-9	-90	-1×10^7	-3×10^3	-3×10^7	2×10^4	1×10^8	-1×10^5	-1×10^4
3000	P3	3×10^2	-9	-90	-1×10^7	-3×10^3	-3×10^7	2×10^4	1×10^8	-1×10^5	-1×10^4

Cas $\nu = 1000$

n	Processus	x_n^i, i=1,...,10									
1	P2	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
	P3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
9	P2	-2,33	0,07	0,59	1,83	-1,15	9,69	2,60	-5,13	-0,97	-0,21
	P3	-1,54	-0,40	-0,22	1,36	-0,70	6,34	1,46	-4,27	-0,82	-0,43
250	P2	-16,96	0,40	0,49	2,11	-4,09	11,70	6,17	-5,63	-1,52	-0,52
	P3	-14,12	-0,41	0,81	2,20	-3,58	11,99	6,17	-5,46	-1,55	-0,37
1000	P2	-17,79	-0,30	0,69	2,10	-3,51	11,73	6,01	-5,67	-1,81	-0,21
	P3	-16,98	-0,09	0,46	2,04	-4,08	11,76	5,84	-5,51	-1,68	-0,56
4000	P2	-17,17	-1,27	0,69	2,08	-3,62	11,94	6,11	-5,71	-1,41	-0,43
	P3	-17,44	0,50	0,60	2,02	-3,96	11,97	6,03	-5,69	-1,61	-0,51

Dans le deuxième cas la convergence apparaît clairement, alors qu'elle est pratiquement inexistante dans le premier cas. Ceci montre l'importance d'un choix convenable des suites $\{a_n\}$ et $\{c_n\}$. Il est certain qu'un choix entraînant un "éloignement prononcé" dès les premières itérations, peut rendre la convergence trop lente - voire, l'interdire en pratique, si les erreurs d'arrondis compensent l'évolution - ; notre premier cas en est une illustration. Pour un étude de la rapidité de convergence liée au choix de la suite $\{a_n\}$dans le cas de l'approximation, on pourra consulter, par exemple [4].
Si dans notre deuxième cas la convergence des processus est manifeste, nous devons toutefois convenir que cette convergence est assez lente. Notons enfin que les deux processus donnent des résultats à peu près semblables ; et n'oublions pas que le processus P3 ne nécessite que (k+1) observations à chaque itération, alors que le processus P2 en nécessite 2k.

BIBLIOGRAPHIE

1 - J.R.BLUM - *Approximation methods which converge with probability one.*
Ann. Math. Stat. Vol.25 (1954) pp.382-386.

2 - J.R.BLUM - *Multidimensional stochastic approximation methods.*
Ann. Math. Stat. Vol.25 (1954) pp.737-744.

3 - J.P.BERTRAN - Thèse d'Etat (1975) Nancy.

4 - CHUNG - *On a stochastic approximation method.*
Ann. Math. Stat. Vol.25 (1954) pp.463-483.

5 - P.DAUBEZE - Thèse de 3ème Cycle (1974) Toulouse.

6 - C.DERMAN - *An application of Chung's lemma to the Kiefer-Wolfowitz stochastic approximation procedure.*
Ann. Math. Stat. (1956) pp.532-536.

7 - C.DERMAN and J.SACKS - *On Dvoresky's stochastic approximation theorem.*
Ann. Math. Stat. Vol.30 (1959) pp.601-606.

8 - V.DUPAC et V.V.IVANOV - *Aplikace Matematiky.*
Svazek 22 (1977) pp.134-146.

9 - A.DVORESKY - *On stochastic approximation*
Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. Vol.1, pp.39-56, University of California Press (1956).

10 - V.FABIAN - *Stochastic approximation methods.*
Czechoslovak mathematical journal (1960) pp.123-159.

11 - S.FRIEDMAN - *On stochastic approximations.*
Ann. Math. Stat. (1963) pp.343-346.

12 - J.B.HIRIART-URRUTY - *Contributions à la programation mathématique : cas déterministe et stochastique.*
Thèse d'Etat (1977) Clermont II.

13 - C.MACCHI - Thèse d'Etat (1972) Paris VI.

14 - J.M.MONNEZ - *Approximation stochastique : Le processus de Robbins-Monro. Mise à jour des résultats et quelques compléments.*
Thèse de 3ème cycle (1975) Nancy I.

15 - L.SCHMETTERER - *L'approximation stochastique.*
Cours de D.E.A. 1ère édition (1968), 2ème édition (1972) Clermont.

16 - J.H.VENTER - *On Dvoresky stochastic approximation theorems.*
Ann. Math. Stat. Vol.37 (1966) pp.1534-1544.

17 - M.T.WASAN - *Stochastic approximation*
Cambridge University Press (1969).

18 - J.WOLFOWITZ - *On the stochastic approximation method of Robbins and Monro.*
Ann. Math. Stat. Vol.23 (1952) pp.457-461.