Un usage de l'approximation stochastique pour l'estimation récursive

Kengy Barty [a], Jean-Sébastien Roy [b], Cyrille Strugarek [b,1]

[a] *CERMICS, École nationale des ponts et chaussées, 6–8, avenue Blaise-Pascal, cité Descartes, 77455 Marne la Vallée cedex 2, France*

[b] *EDF R&D, 1, avenue du Général de Gaulle, 92141 Clamart cedex, France*

Reçu le 6 octobre 2005 ; accepté après révision le 3 novembre 2006

Présenté par Paul Deheuvels

Résumé

Cette Note a pour objet de présenter une technique générique de démonstration de la convergence d'une large classe d'estimateurs récursifs, par l'usage de techniques d'approximation stochastique. À titre d'illustration de cette méthode, on donne une preuve alternative de convergence de l'estimateur de densité introduit par Wagner et Wolverton. La preuve proposée part de résultats généraux sur les algorithmes stochastiques à valeurs dans un espace de Hilbert. Puis, on montre comment cette nouvelle preuve peut être facilement réutilisée pour montrer la convergence d'estimateurs récursifs de la densité ou de fonctions de régression y compris en présence de contraintes.

Abstract

On the use of stochastic approximation in recursive estimation. We propose in this Note a way to study the convergence of statistical recursive estimates. As an illustration, we provide a convergence proof for the well-known recursive density estimate introduced by Wagner and Wolverton. This alternative proof is based on results on Hilbert-valued stochastic approximation schemes, and provides an insight on recursive density estimation seen as minimization problems. Indeed, recursive statistical estimates such as Wagner-Wolverton's estimate can be seen as stochastic algorithms with values in a functional space which turns out to be here an Hilbert space. Hence, we can embed the field of recursive statistical estimation into the field of Hilbert-valued stochastic approximation and propose a systematic approach to prove the convergence of statistical recursive estimates, like density estimates or regression function estimates. Moreover, this systematic approach enables to take into account various constraints on the estimates like bounds or measurability.

Abridged English version

In 1969, Wagner and Wolverton (cf. [15]) introduce for the first time a recursive density estimator. Contrarily to the related Parzen-Rosenblatt estimator (cf. [13,9]), whose sample size must be set in advance,

Email addresses: kengy.barty@cermics.enpc.fr (Kengy Barty), jean-sebastien.roy@edf.fr (Jean-Sébastien Roy), cyrille.strugarek@edf.fr (Cyrille Strugarek).

[1] Ce travail a également été soutenu par l'École Nationale Supérieure de Techniques Avancées (ENSTA) et l'École Nationale des Ponts et Chaussées (ENPC).

it has the advantage to converge along the iterations as additional samples are available. In this Note, we follow the works of Révész (cf. [10,11,12]), and study the convergence of such statistical recursive estimates using new results on Hilbert-valued stochastic approximation schemes.

Indeed, recursive statistical estimates such as Wagner-Wolverton estimate can be seen as stochastic algorithms with values in a functional space which turns out to be here an Hilbert space. Hence, we can embed the field of recursive statistical estimation into the field of Hilbert-valued stochastic approximation and propose a systematic approach to prove the convergence of statistical recursive estimates, like density estimates or regression function estimates. Moreover, this systematic approach enables to take into account various constraints on the estimates like bounds or measurability. We briefly describe below our general convergence result on Hilbert-valued stochastic approximation and provide an application to Wagner-Wolverton's estimator.

We consider a general minimization problem with constraints

$$\min_{x\in X} J(x)$$
$$\text{s.t. } x \in X^f,$$

where X is supposed to be an Hilbert space, the feasible set X^f is assumed to be non-empty, closed and convex, and $J : X \to \mathbb{R}$ is a convex, lower semi-continuous and coercive mapping.

We introduce the following stochastic perturbed gradient algorithm in order to approximate the solution of the problem. Let $x_0 \in X^f$, then

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

where Π_{X^f} is the projection operator into X^f, (s_t) and (w_t) are sequences of realizations of random variables on X, s_t is supposed to be a descent direction, w_t is a perturbation of the descent direction, and γ_t is a nonincreasing sequence of nonnegative real stepsizes.

Assuming s_t is indeed a descent direction, if the perturbation w_t is asymptotically unbiased and its variance does not increase too fast, and assuming the stepsize γ_t, the bias and the variance of w_t follow some relations usual for stochastic algorithms, we are able to prove the weak convergence of x_t to its optimum value, and the strong convergence when J is strongly convex (cf. [1]).

This theorem appears generic enough to provide a systematic way to prove the convergence of, e.g., Wagner-Wolverton's estimator or Deheuvels's estimator (cf. [3,4]). For example, Wagner-Wolverton's estimator of a density f is given by

$$\forall x \in \mathbb{R}^m,\ f_{t+1}(x) = \frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{h_i^m}K\left(\frac{x-X_i}{h_i}\right) = f_t(x) - \frac{1}{t+1}\left(f_t(x) - \frac{1}{h_{t+1}^m}K\left(\frac{x-X_{t+1}}{h_{t+1}}\right)\right),$$

where (X_t) is a sequence of i.i.d random variables of density f, (h_t) is a sequence of positive reals, and K a probability density on $\mathbb{R}^m$.

To apply the previous theorem, we consider $J(g) = \frac{1}{2}\|g - f\|^2$, with $g \in L^2(\mathbb{R}^m, \mathbb{R}, f)$, and the algorithm $g_{t+1} = g_t + \gamma_t\left(s_t + w_t\right)$, with $s_t = f - g_t$, $w_t(\cdot) = K_{t+1}(\cdot, X_{t+1}) - f(\cdot)$, $K_t(x,y) = \frac{1}{h_t^m}K\left(\frac{x-y}{h_t}\right)$, and $\gamma_t = 1/(t+1)$. In this case, g_t and f_t as defined by Wagner-Wolverton's estimator are equal. Moreover, the hypotheses of the previous theorem are easily verified which lead to the convergence of the estimator.

Similarly, this theorem can also be used to prove the convergence of the recursive regression estimators introduced by Révész, and extend their convergence to the case of the constrained regression.

1. Introduction

1.1. *Contexte*

En 1969, Wagner et Wolverton (cf. [15]) introduisent pour la première fois un estimateur de densité récursif. Cet estimateur comporte quelques similitudes avec l'estimateur antérieur de Parzen-Rosenblatt (cf. [13,9]), notamment de par l'utilisation dans les deux cas de noyaux. Cependant, là où l'estimateur de Parzen-Rosenblatt est un estimateur *à nombre de tirages* (ou taille d'échantillon) *fixé*, l'estimateur de Wagner-Wolverton permet quant à lui de tenir compte des tirages au fur et à mesure de leur arrivée, sans en avoir fixé le nombre à l'avance, et d'obtenir ainsi une propriété de convergence d'un estimateur enrichi autant qu'on le souhaite.

De nombreux travaux ont exploré de manière assez exhaustive l'estimation de densité non-paramétrique du type Parzen-Rosenblatt, affinant les propriétés de convergence de cet estimateur en fonction de la régularité de la densité à estimer, et donnant de nombreux résultats tant asymptotiques (grandes déviations, théorèmes centraux limites, lois du logarithme itéré) que de convergence (en norme L^p, presque partout, etc.). On peut consulter au sujet de l'estimation non-paramétrique par exemple les ouvrages de Devroye [7] ou Tsybakov [14].

En revanche, les travaux autour de la convergence de la version non-paramétrique récursive de l'estimation de densité évoquée plus haut sont plus rares. Outre le travail initial de Wagner et Wolverton, on peut principalement citer Révész (cf. [10,11,12]), Deheuvels (cf. [2,5]), et Devroye (cf. [6]). De plus, chacune de ces contributions propose une analyse de la convergence et un raffinement des hypothèses dédiées à l'estimateur particulier considéré. Comme l'a souligné Révész ([10,11,12]) l'estimation récursive, tant pour les fonctions de régression que pour les densités, peut être plongée dans le cadre de l'approximation stochastique.

Nous proposons dans cette Note un prolongement des travaux de Révész, grâce à un résultat général d'approximation stochastique hilbertienne donné en Section 2, et donnons à titre d'illustration en Section 3 une preuve alternative de convergence de l'estimateur de Wagner et Wolverton. Cette technique de preuve donne une manière systématique de prouver la convergence d'estimateurs récursifs, comme en témoigne l'application en Section 4 à un autre estimateur introduit par Deheuvels. Au delà de ces cas particuliers, le théorème général que nous utilisons permet d'étudier la convergence de tout estimateur fonctionnel récursif, avec ou sans contraintes, ce qui est intéressant par exemple dans le cas d'estimateurs d'espérances conditionnelles ou de fonctions de régression contraintes.

1.2. *Estimation récursive de la densité*

Le but est d'estimer la densité f sur $\mathbb{R}^m$ d'une variable aléatoire, sur la base de tirages successifs i.i.d. de cette variable aléatoire. L'estimateur proposé par Wagner et Wolverton s'écrit :

$$\forall x \in \mathbb{R}^m,\ f_{t+1}(x) = \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{h_i^m} K\left(\frac{x - X_i}{h_i}\right), \tag{1}$$

avec $(X_i)_{1 \leq i \leq t+1}$ une suite de variables aléatoires i.i.d. de densité f, $(h_i)_{1 \leq i \leq t+1}$ une suite de réels positifs et K une densité de probabilité sur $\mathbb{R}^m$. La formule (1) peut se réinterpréter comme l'algorithme suivant :

$$\begin{aligned} &\text{Tirer } X_{t+1} \text{ indépendamment de } (X_1, \ldots, X_t) \text{ de densité } f, \\ &\forall x \in \mathbb{R}^m,\ f_{t+1}(x) = f_t(x) - \frac{1}{t+1}\left(f_t(x) - \frac{1}{h_{t+1}^m} K\left(\frac{x - X_{t+1}}{h_{t+1}}\right)\right). \end{aligned} \tag{2}$$

Wagner et Wolverton ont prouvé le théorème de convergence suivant :
Théorème 1.1 (Wagner et Wolverton, 1969) *Supposons que :*
(i) La suite de réels positifs (h_t) est telle que :

$$1 \geq h_1 \geq \cdots \geq 0,\ \sum_{t \in \mathbb{N}} \frac{h_t}{t} < +\infty,\ \sum_{t \in \mathbb{N}} \frac{1}{t^2 h_t^m} < +\infty. \tag{3}$$

(ii) L'application $K : \mathbb{R}^m \to \mathbb{R}$ est une densité de probabilité satisfaisant

$$\forall x \in \mathbb{R}^m,\ K(x) \leq A < +\infty, \qquad \int \|x\|_m K(x) dx = B < +\infty. \tag{4a}$$

(iii) La densité de probabilité f est lipschitzienne de constante L.
(iv) Alors,

$$\int_{\mathbb{R}^m} (f_t(x) - f(x))^2 f(x) dx \to 0, \text{ p.s. } \textit{quand } t \to +\infty.$$

2. Gradient perturbé dans un espace de Hilbert

Avant de passer à la preuve proprement dite du Théorème 1.1, nous introduisons dans cette section un résultat général de convergence d'algorithmes d'approximation stochastique dans le contexte hilbertien. Ce résultat, prouvé dans [1], a fait l'objet de récents travaux des auteurs sur l'approximation stochastique, dont les sources historiques sont principalement les articles [8,10,16].

Nous allons ici nous placer dans le cadre général suivant. Le problème est de résoudre :

$$\begin{aligned} &\min_{x \in X} J(x) \\ &\text{s.c. } x \in X^f. \end{aligned} \tag{5}$$

avec X un espace de Hilbert muni d'un produit scalaire et d'une norme notés $\langle \cdot, \cdot \rangle$ et $\| \cdot \|$, $J : X \to \mathbb{R}$ une application convexe, X^f un convexe fermé de X, et Π_{X^f} la projection sur X^f. On se donne également $(\Omega, \mathcal{A}, \mathbb{P})$ un espace de probabilité, et $(\mathcal{F}_t)$ une filtration sur cet espace. On propose alors l'algorithme de gradient perturbé suivant pour résoudre le problème 5. On se donne $x_0 \in X$.

Algorithme 2.1 *Étape $t \in \mathbb{N}$:*

$$\begin{aligned} &\textit{Calculer } s_t \textit{ et } w_t \textit{ variables aléatoires à valeurs dans } X, \\ &x_{t+1} = \Pi_{X^f} \left(x_t + \gamma_t (s_t + w_t) \right), \end{aligned}$$

avec s_t une direction de descente, w_t une perturbation sur la direction de descente, et γ_t une suite décroissante de réels positifs. On dispose du résultat de convergence suivant :

Théorème 2.2 *(i) Supposons que $x \mapsto J(x)$ soit convexe, et semi-continue inférieurement. Si de plus $J(x) \to +\infty$ quand $\|x\| \to +\infty$, et si X^f est un convexe fermé non vide de X, alors le problème (5) admet un ensemble de solutions non vide qu'on notera S.*

Supposons de plus que

(ii) $(\mathcal{F}_t)$ soit une filtration sur $(\Omega, \mathcal{A}, \mathbb{P})$, et que pour tout $t \in \mathbb{N}$, x_t et s_t soient $\mathcal{F}_t$-mesurables ;

(iii) J soit différentiable et que $\nabla J(\cdot)$ soit lipschitzienne de module L ;

(iv) il existe $\kappa, c > 0$ tels que pour tout $x^* \in S$, pour tout $t \in \mathbb{N}$,

$$\langle s_t, x_t - x^* \rangle \le \kappa \left(J(x^*) - J(x_t) \right), \quad \|s_t\| \le c \left(1 + \|\nabla J(x_t)\| \right); \tag{6a}$$

(v) il existe $b \ge 0$, $A > 0$ et deux suites de réels positifs (ϵ_t, η_t), tels que pour tout $t \in \mathbb{N}$,

$$\|\mathbb{E} \left(w_t | \mathcal{F}_t \right) \| \le b \eta_t, \quad \mathbb{E} \left(\|w_t\|^2 | \mathcal{F}_t \right) \le A \left(1 + \frac{1}{\epsilon_t} \right); \tag{7a}$$

(vi) les suites (γ_t), (ϵ_t) and (η_t) décroissent vers 0 et soient telles que :

$$\forall t \in \mathbb{N},\ \gamma_t, \epsilon_t > 0, \quad \sum_{t \in \mathbb{N}} \gamma_t = +\infty, \quad \sum_{t \in \mathbb{N}} \frac{(\gamma_t)^2}{\epsilon_t} < +\infty, \quad \sum_{t \in \mathbb{N}} b \gamma_t \eta_t < +\infty. \tag{8}$$

Alors $J(x_t) \to J(x^*)$ p.s., et (x_t) converge faiblement vers x^* p.s., quand $t \to \infty$. De plus, si J est fortement convexe, (x_t) converge fortement vers l'unique solution x^* de (5).

3. Preuve alternative pour l'estimateur de Wagner et Wolverton

Remarquons tout d'abord que comme f est une densité lipschitzienne, f est de carré et de cube intégrables. Soit $\mathcal{L}^2(\mathbb{R}^m, \mathbb{R}, f) := \left\{ g : \mathbb{R}^m \to \mathbb{R} \ : \ \int_{\mathbb{R}^m} g(x)^2 f(x) dx < +\infty \right\}$, on note alors $L^2(\mathbb{R}^m, \mathbb{R}, f)$ l'espace quotient de $\mathcal{L}^2(\mathbb{R}^m, \mathbb{R}, f)$ par la relation d'égalité $f(x)dx$-presque sûre. Par simplicité, on notera $\langle \cdot, \cdot \rangle$ et $\| \cdot \|$ le produit scalaire et la norme dans $L^2(\mathbb{R}^m, \mathbb{R}, f)$, qu'on notera dans cette preuve L^2. Regardons le problème d'optimisation fonctionnel suivant :

$$\min_{g \in L^2} J(g) := \frac{1}{2}\|g - f\|^2. \tag{9}$$

Comme f est dans L^2, il est clair que l'unique solution de (9) est f. Un algorithme de gradient perturbé sur le problème (9) peut s'écrire : $g_{t+1} = g_t + \gamma_t (s_t + w_t)$, avec $s_t = f - g_t$, $w_t(\cdot) = K_{t+1}(\cdot, X_{t+1}) - f(\cdot)$, et $K_t(x, y) = \frac{1}{h_t^m} K\left(\frac{x-y}{h_t}\right)$. On retrouve donc en posant $\gamma_t = 1/(t+1)$ que f_t défini par l'algorithme (2) et g_t défini par l'algorithme de gradient coïncident. En utilisant le Théorème 2.2, nous allons montrer la convergence de l'algorithme de gradient vers la solution du problème (9), ce qui prouvera le théorème. L'application J est fortement convexe, et f est la solution du problème (9) ; En définissant pour tout $t \in \mathbb{N}$, $\mathcal{F}_t$ la tribu engendrée par $(X_1, \ldots, X_t)$, s_t et f_t sont bien $\mathcal{F}_t$-mesurables. Le gradient de J est donné par $\nabla J(g) = g - f$ pour tout $g \in L^2$. Il est donc lipschitzien. Par définition, $\langle s_t, g_t - f\rangle = -\|g_t - f\|^2$ et $\|s_t\| = \|\nabla J(g_t)\|$. En d'autres termes, s_t est bien une direction de descente.

Il ne reste plus qu'à vérifier les hypothèses sur les bruits w_t :

$$\begin{aligned}
\mathbb{E}\left(w_t|\mathcal{F}_t\right)(x) &= \int_{\mathbb{R}^m} K_{t+1}(x, y) f(y) dy - f(x) = \int_{\mathbb{R}^m} \frac{1}{h_{t+1}^m} K\left(\frac{x-y}{h_{t+1}}\right) f(y) dy - f(x), \\
&= \int_{\mathbb{R}^m} K(z) f(x - z h_{t+1}) dz - f(x) = \int_{\mathbb{R}^m} K(z) \left(f(x - z h_{t+1}) - f(x)\right) dz.
\end{aligned}$$

D'où, par le caractère lipschitzien de f, et grâce à l'hypothèse (4a), $\|\mathbb{E}\left(w_t|\mathcal{F}_t\right)\| \leq h_{t+1} L B$, ce qui permet de vérifier la première partie de l'hypothèse (7a) avec $\eta_t = h_{t+1}$. Par ailleurs :

$$\begin{aligned}
\mathbb{E}\left(\|w_t\|^2|\mathcal{F}_t\right) &= \int_{\mathbb{R}^m \times \mathbb{R}^m} \left(\frac{1}{h_{t+1}^m} K\left(\frac{x-y}{h_{t+1}}\right) - f(x)\right)^2 f(x) f(y) dx dy, \\
&= \int_{\mathbb{R}^m \times \mathbb{R}^m} \frac{1}{h_{t+1}^{2m}} K\left(\frac{x-y}{h_{t+1}}\right)^2 f(x) f(y) dx dy + \int_{\mathbb{R}^m} f(x)^3 dx \\
&\quad - \frac{2}{h_{t+1}^m} \int_{\mathbb{R}^m \times \mathbb{R}^m} K\left(\frac{x-y}{h_{t+1}}\right) f(x)^2 f(y) dx dy, \\
&\leq \int_{\mathbb{R}^m} f(x)^3 dx + \frac{A}{h_{t+1}^m} \int_{\mathbb{R}^m \times \mathbb{R}^m} K(z) f(x) f(x - z h_{t+1}) dx dz.
\end{aligned}$$

Dès lors, comme f est une densité de carré et cube intégrables, et grâce aux hypothèses (4a), il existe deux constantes $C, C' > 0$ telles que : $\mathbb{E}\left(\|w_t\|^2|\mathcal{F}_t\right) \leq C + \frac{C'}{h_{t+1}^m}$, ce qui donne la deuxième partie de l'hypothèse (7a) du Théorème 2.2. Enfin, les hypothèses sur les suites (3) impliquent dans notre cas les hypothèses (8). On peut donc appliquer le Théorème 2.2, ce qui prouve le théorème de Wagner et Wolverton.

4. Généricité de l'approche

Nous donnons ici à travers deux autres exemples d'estimateurs récursifs une idée du caractère générique de notre approche par gradient stochastique généralisé de l'estimation récursive.

4.1. *Autre estimateur récursif de la densité*

Dans les articles [3,4], Deheuvels a proposé et montré la convergence d'un estimateur récursif de la densité d'une autre forme, donné par la formule suivante :

$$\forall x \in \mathbb{R}^m,\ \hat{f}_{t+1}(x) = \frac{1}{\sum_{i=1}^{t+1} h_i^m} \sum_{i=1}^{t+1} K\left(\frac{x - X_i}{h_i}\right). \tag{10}$$

Cette formule s'écrit aisément sous forme récursive :

$$\forall x \in \mathbb{R}^m,\ \hat{f}_{t+1}(x) = \hat{f}_t(x) - \frac{h_{t+1}^m}{\sum_{i=1}^{t+1} h_i^m}\left(\hat{f}_t(x) - \frac{1}{h_{t+1}^m} K\left(\frac{x - X_{t+1}}{h_{t+1}}\right)\right). \tag{11}$$

Par comparaison avec l'algorithme de Wagner-Wolverton (2), la seule différence est donc que l'on a remplacé le pas de moyennisation en $\frac{1}{t+1}$ par le pas dépendant des fenêtres en $\frac{h_{t+1}^m}{\sum_{i=1}^{t+1} h_i^m}$. En faisant les mêmes hypothèses que celles du Théorème 1.1 sur la densité à estimer et le noyau K, et en modifiant les hypothèses sur les pas (3), afin de correspondre au cadre générique du Théorème 2.2, on obtient donc facilement une preuve de convergence forte de l'estimateur de Deheuvels. De plus, en prenant par exemple pour la fenêtre $h_t = 1/t^\alpha$, on prouve facilement que les hypothèses nécessaires à la convergence sont vérifiées pour tout $\alpha \in (0, 1/m)$.

4.2. *Fonctions de régression*

Dans l'article [12], Révész introduit un estimateur récursif de la fonction de régression $r(x) = \mathbb{E}\left(Y|X = x\right)$, avec X, Y deux variables aléatoires à valeurs respectivement dans $\mathbb{R}^m$, et $\mathbb{R}^p$. Le Théorème 2.2 permet facilement de montrer la convergence de l'algorithme suivant, trés proche de celui de Révész :

$$\forall x \in \mathbb{R}^m,\ r_{t+1}(x) = r_t(x) + \gamma_t\left(Y_{t+1} - r_t(X_{t+1})\right)\frac{1}{h_{t+1}^m} K\left(\frac{X_{t+1} - x}{h_{t+1}}\right). \tag{12}$$

Le Théorème 2.2 permettrait de plus de tenir compte de contraintes de bornes sur les coefficients d'un problème de régression.

Pour conclure, l'approche présentée ici a pour intérêt majeur de pouvoir être utilisée facilement afin de prouver la convergence d'estimateurs fonctionnels récursifs.

Références

[1] K. Barty, J.-S. Roy, and C. Strugarek. Hilbert valued perturbed subgradient algorithms. *to appear in Mathematics of Operations Research*, 2005.
[2] P. Deheuvels. Sur l'estimation séquentielle de la densité. *C. R. Acad. Sci. Paris Ser. A-B*, 276 :1119–1121, 1973.
[3] P. Deheuvels. Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité. *C. R. Acad. Sci. Paris Ser. A*, 278 :1217–1220, 1974.
[4] P. Deheuvels. Sur une famille d'estimateurs de la densité d'une variable aléatoire. *C. R. Acad. Sci. Paris Ser. A*, 276 :1013–1015, 1974.
[5] P. Deheuvels. Estimation séquentielle de la densité. In *Contribuciones en Probabilidad y Estadistica Matematica Enseñanza de la Matematica y Analysis*, pages 156–168. Grindley, Granada, Espagne, 1979.
[6] L. Devroye. On the pointwise and the integral convergence of recursive kernel estimates of probability densities. *Utilitas Mathematica*, 15 :113–128, 1979.
[7] L. Devroye. *A Course In Density Estimation*. Birkhäuser, Boston, 1987.
[8] J.-B. Hiriart-Urruty, Algorithmes de résolution d'équations et d'inéquations variationnelles, Z. Wahrscheinlichkeitstheorie verw. Gebiete, 33 : 167–186, 1975.
[9] E. Parzen. On estimating of a probability density and mode. *Annals of Mathematical Statistics*, 35 :1065–1076, 1962.
[10] P. Révész. Robbins-Monro procedure in a Hilbert space and its application in the theory of learning processes, I. *Studia Sci. Math. Hungar.*, 8 :391–398, 1973.
[11] P. Révész. Robbins-Monro procedure in a Hilbert space, II. *Studia Sci. Math. Hungar.*, 8 :469–472, 1973.
[12] P. Révész. How to apply the method of stochastic approximation in the non-parametric estimation of a regression function. *Math. Operationsforsch. Statist. Ser. Statistics*, 8(1) :119–126, 1977.
[13] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27 :832–837, 1956.
[14] A. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer Verlag, 2004.
[15] T.J. Wagner and C.T. Wolverton. Recursive estimates of probability densities. *IEEE Trans. Syst. Man. Cybern.*, 5 :307, 1969.
[16] G. Yin and Y.M. Zhu. On H-valued Robbins-Monro processes. *J. Multivariate Anal.*, 34 :116–140, 1990.