DOCUMENT RESUME

ED 051 256 TM 000 151

AUTHOR	Wolff, Hans
TITLE	On Approximation of Distribution and Density Functions.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	RB-70-51
PUB DATE	Sep 70
NOTE	13p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Algorithms, *Calculation, *Mathematics, Measurement, Multiple Regression Analysis, Probability, *Research Methodology, *Statistical Analysis, *Statistical Studies, Statistics
IDENTIFIERS	*Stochastic Approximation Theory

ABSTRACT

Stochastic approximation algorithms for least square error approximation to density and distribution functions are considered. The main results are necessary and sufficient parameter conditions for the convergence of the approximation processes and a generalization to some time-dependent density and distribution functions. (Author)

ED051256

RESEARCH BULLETIN

RB-70-51

ON APPROXIMATION OF DISTRIBUTION AND DENSITY FUNCTIONS

Hans Wolff

This Bulletin is a draft for interoffice circulation. Corrections and suggestions for revision are solicited. The Bulletin should not be cited as a reference without the specific permission of the author. It is automatically superseded upon formal publication of the material.

Educational Testing Service
Princeton, New Jersey
September 1970

On Approximation of Distribution and Density Functions

Hans Wolff

Abstract

Stochastic approximation algorithms for least square error approximation to density and distribution functions are considered. The main results are necessary and sufficient parameter conditions for the convergence of the approximation processes and a generalization to some time-dependent density and distribution functions.

On Approximation of Distribution and Density Functions

Hans Wolff

In this paper we deal with the special approach to the estimation of an unknown density or distribution function of a real-valued random variable ξ as developed in [1]-[8]. Using the same notation we briefly describe this approach.

Consider the N-dimensional vector of functions $\Phi(x) = (\phi_1(x),\ldots,\phi_N(x))^T$. The components $\phi_i(x)$, $i = 1,\ldots N$, are assumed to be linearly independent, square-integrable and bounded real functions on an interval $\Omega = [a, b]$ of the real axis. If a sequence of independent observations $\{x_1,x_2,\ldots\}$ from ξ is available, the problem is then to find an approximation

$$\hat{F}(x) = \sum_{i=1}^{N} \alpha_i \phi_i(x) = \underline{\alpha}^T \underline{\Phi}(x)$$

in Ω for the unknown distribution function $F(x)$, such that $\hat{F}(x)$ minimizes the integral-square-error criterion

$$(1) \qquad G_1(\underline{\alpha}) = \int_\Omega [F(x) - \underline{\alpha}^T \underline{\Phi}(x)]^2 \, dx$$

with respect to the vector of coefficients $\underline{\alpha} = (\alpha_1,\alpha_2,\ldots,\alpha_N)^T$. The analogous estimation problem for the unknown density function $f(x)$ consists in determining the estimator $\hat{f}(x)$,

$$\hat{f}(x) = \sum_{i=1}^{N} \beta_i \phi_i(x) = \underline{\beta}^T \underline{\Phi}(x) \quad ,$$

such that again the integral-square-error criterion

$$(2) \qquad G_2(\underline{\beta}) = \int_\Omega [f(x) - \underline{\beta}^T \underline{\Phi}(x)]^2 \, dx$$

is a minimum with respect to $\underline{\beta}$.

As can be easily shown (see e.g., [1]), minimizing (1) and (2) is equivalent to solving the regression equations

$$(3) \qquad E\left[\int_{\Omega} z(\xi, y)\underline{\Phi}(y)\,dy - \underline{A\alpha}\right] = 0$$

and

$$(4) \qquad E[\underline{w}(\xi) - \underline{A\beta}] = 0 \quad ,$$

respectively, where $\underline{A}$ is a known $N \times N$ -matrix,

$$\underline{A} = \int_{\Omega} \underline{\Phi}(y)\underline{\Phi}^T(y)\,dy \quad ,$$

and $z(\xi, y)$ and $\underline{w}(\xi)$ are defined as

$$z(\xi, y) = \begin{cases} 1 & \text{if} \quad \xi \le y \\ 0 & \phantom{\text{if}} \quad \xi > y \end{cases} \quad ,$$

$$\underline{w}(\xi) = \begin{cases} \underline{\Phi}(\xi) & \text{if} \quad \xi \in \Omega \\ 0 & \phantom{\text{if}} \quad \xi \notin \Omega \end{cases} \quad .$$

The purpose of the mentioned papers consisted in solving the parameter-dependent regression equations (3) and (4) by the application of the stochastic approximation theory as an appropriate method. A further goal was to give an iterative solution in order to avoid computer storage problems. But because of the linear independence of the $\Phi_i(x)$, $i = 1,\ldots,N$, $\underline{A}^{-1}$ exists and we can solve (3) and (4) directly:

$$(5) \qquad \underline{\alpha}^* = \underline{A}^{-1}E\left[\int_{\Omega} z(\xi, y)\underline{\Phi}(y)\,dy\right] \quad ,$$

$$(6) \qquad \underline{\beta}^* = \underline{A}^{-1}E[\underline{w}(\xi)] \quad .$$

Therefore we have only to estimate the expectations of the parameter-independent random variables $\zeta_1 = \int_{\Omega} z(\xi, y)\Phi(y)\, dy$ and $\zeta_2 = \underline{w}(\xi)$. So simplifying the statement of the problem we can expect stronger limiting theorems for those procedures considered in [1]-[8]. In previous papers ([9], [10]) the author has dealt with such iterative approximations of the expectation of a random variable. The following process was considered.

Let $\{a_n\}$ be any sequence of real numbers restricted to $0 < a_n < 1$ for all n and let $\underline{Y}_n = (y_1, \ldots, y_N)^T$ denote the n-th observation of a real-valued N-dimensional random variable $\underline{\eta} = (\eta_1, \ldots, \eta_N)^T$. Then the approximation procedure $\{\underline{X}_n\}$ is defined by the iteration formula

(7) $$\underline{X}_{n+1} = (1 - a_{n+1})\, \underline{X}_n + a_{n+1}\, \underline{Y}_{n+1} \quad , \quad n = 0, 1, 2, \ldots \quad ,$$

with an arbitrary but fixed starting point $\underline{X}_0 = \underline{a} \in R^N$. Theorem 1 gives necessary and sufficient parameter conditions for the convergence of this process.

<u>Theorem 1</u>: The process (7) converges under the assumption

$$0 < \max_{1 \leq i \leq N} \operatorname{Var} \eta_i < \infty$$

with probability one and in the mean to the expectation $\underline{M}$ of $\underline{\eta}$,

$$\underline{X}_n \to \underline{M} \text{ w.p.1} \quad , \quad E(\underline{X}_n - \underline{M})^2 \to 0 \quad (n \to \infty) \quad ,$$

if and only if

(8) $$a_n \to 0 \quad , \quad \sum_{i=1}^{n} a_i \to \infty \qquad (n \to \infty) \quad .$$

The parameter condition (8) is only sufficient if we admit the degenerated and trivial case $\mathrm{Var}\ \eta_i = 0$, $i = 1,\ldots,N$. The proof of Theorem 1 is given in [10].

The application of Theorem 1 to the random variables $\underline{A}^{-1}\underline{\zeta}_1$ and $\underline{A}^{-1}\underline{\zeta}_2$ yields at once those estimation procedures $\{\underline{\alpha}_n\}$ and $\{\underline{\beta}_n\}$ for the sought vectors $\underline{\alpha}^*$ and $\underline{\beta}^*$ considered in [1]-[8]:

$$(9) \qquad \underline{\alpha}_{n+1} = (1 - a_{n+1})\underline{\alpha}_n + a_{n+1}\underline{A}^{-1}\underline{z}_{1,n+1} \ , \quad \underline{\alpha}_0 = \underline{b} \in R^N \text{ w.p.1} \quad ,$$

$$(10) \qquad \underline{\beta}_{n+1} = (1 - a_{n+1})\underline{\beta}_n + a_{n+1}\underline{A}^{-1}\underline{z}_{2,n+1} \ , \quad \underline{\beta}_0 = \underline{c} \in R^N \text{ w.p.1} \quad ,$$

where $\underline{z}_{1,n}$ and $\underline{z}_{2,n}$ denote the n -th observation of the random variables $\underline{\zeta}_1$ and $\underline{\zeta}_2$, respectively:

$$\underline{z}_{1,n} = \int_\Omega z(x_n,y)\underline{\Phi}(y)\,dy = \begin{cases} \int_a^b \underline{\Phi}(y)\,dy & x_n < a \\ \int_{x_n}^b \underline{\Phi}(y)\,dy \quad \text{if} & a \le x_n \le b \\ 0 & x_n > b \end{cases} ,$$

$$\underline{z}_{2,n} = \begin{cases} \underline{\Phi}(x_n) & x_n \in \Omega \\ & \text{if} \\ 0 & x_n \notin \Omega \end{cases} .$$

From Theorem 1 follows immediately,

Theorem 2: The stochastic process defined by (9) and (10) converges with probability one and in the mean to $\underline{\alpha}^*$ and β^* , respectively, if and only if the sequence of parameters $\{a_n\}$ fulfills condition (8).

We mention that the following modifications of (9) and (10) suggested, for example in [1], [6], [7],

$$\underline{\alpha}_{n+1} = (1 - a_{n+1})\,\underline{\alpha}_n + a_{n+1}\underline{A}^{-1} \cdot \frac{1}{n+1}\sum_{i=1}^{n+1} \underline{z}_{1,i} \quad ,$$

$$\underline{\beta}_{n+1} = (1 - a_{n+1})\,\underline{\beta}_n + a_{n+1}\underline{A}^{-1} \cdot \frac{1}{n+1}\sum_{i=1}^{n+1} \underline{z}_{2,i} \quad ,$$

do not have a faster rate of convergence than (9) and (10) themselves as was erroneously asserted in [6] and [7]. The error consisted essentially in taking $\underline{\alpha}_n$ and $\frac{1}{n+1}\sum_{i=1}^{n} \underline{z}_{1,i}$ (or $\underline{\beta}_n$ and $\frac{1}{n+1}\sum_{i=1}^{n+1} \underline{z}_{2,i}$, respectively) as independent random variables (e.g. [6], p. 133, equation (7)).

Time-dependent Density and Distribution Functions

Instead of identically distributed values x_i , $i = 1,2,\ldots$ from ξ we deal now with a sample $\{x_1,x_2,\ldots\}$ corresponding to a sequence of random variables $\{\xi_1,\xi_2,\ldots\}$ where ξ_i is distributed with $F_i(x)$, $i = 1,2,\ldots$, representing, e.g. successive time periods. Since we want to derive an analogous limiting theorem to that given in Theorem 2 we restrict ourselves to the case where $\{F_i(x)\}$ converges to a limiting distribution $F(x)$ and $\{f_i(x)\}$ converges to a limiting density function $f(x)$. For this situation we have the following corollary to Theorem 2.

<u>Corollary:</u> Theorem 2 holds even in the case where the observations x_i , $i = 1,2,\ldots$, are drawn from a population with a distribution function $F_i(x)$ and a density function $f_i(x)$, if we assume

$$F_i(x) \to F(x)\ ,\quad f_i(x) \to f(x) \quad (i \to \infty)\ ^1$$

$[F(x)$ distribution function, $f(x)$ density function$]$.

This corollary follows immediately from (5) and (6) and from a generalized version of Theorem 1 given below.

Let $\{\underline{y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,N})^T\}$ be a sequence of independent N -dimensional real-valued observations distributed with $\{F_i(y_1, \ldots, y_N)\}$, respectively, and where $F_i(y_1, \ldots, y_N)$ converges to a nondegenerated limiting distribution $F(y_1, \ldots, y_N)$. Th n we have

Theorem 3: The process (7)

$$\underline{X}_{n+1} = (1 - a_{n+1})\,\underline{X}_n + a_{n+1}\underline{y}_{n+1} \qquad \underline{X}_0 = \underline{a} \in R^N \quad ,$$

converges under the assumption

$$\max_{1 \le j \le N} \operatorname{Var} y_{i,j} \le C < \infty \quad , \quad i = 1,2,\ldots$$

with probability one and in the mean to the expectation $\underline{M}$ of $F(y_1, \ldots, y_N)$,

$$\underline{X}_n \to \underline{M} \text{ w.p.1}\ ,\quad E(\underline{X}_n - \underline{M})^2 \to 0 \quad (n \to \infty)$$

if and only if $\{a_n\}$ fulfills condition (8).

Because of the length of the proof of this theorem, the reader is referred to [9] or [10]. Some problems arise if we consider the case where Ω is the whole probability space, especially the entire real axis. In this case it is natural to require that the approximation $\hat{f}(x)$ should satisfy the normalization condition

---

[1] The assumption $f_i(x) \to f(x)$, where $f(x)$ is a density function, is sufficient for $F_i(x) \to F(x)$, and $F(x)$ distribution function (see e.g., [11]).

$$\int_{\Omega} \hat{f}(x)\, dx = 1 \quad .$$

Unfortunately this is not true in general. To avoid this we can use Lagrange's coefficients method as was done for orthonormal functions $\phi_i(x)$ by Laski [5] and for a similar problem by Nikolic and Fu [6].

Instead of (2) we now minimize the criterion

$$G_3 = \int_{\Omega} [f(x) - \sum_{i=1}^{N} \beta_i \phi_i(x)]^2\, dx - 2\lambda(\sum_{i=1}^{N} \beta_i d_i - 1) \quad ,$$

where λ is a Lagrange coefficient and

$$d_i = \int_{\Omega} \phi_i(x)\, dx \ , \quad 0 < |d_i| < \infty \ , \quad i = 1,2,\ldots,N \quad .$$

The minimization conditions

$$\frac{\partial G_3}{\partial \beta_i} = 0 \ , \quad i = 1,\ldots,N \ ; \quad \frac{\partial G_3}{\partial \lambda} = 0$$

yield the system of linear equations

$$\sum_{i=1}^{N} d_i \beta_i = 1$$

$$\sum_{i=1}^{N} a_{ik} \beta_i + d_k \lambda = E(\phi_k) \ , \quad k = 1,\ldots,N \quad ,$$

where $\underline{A} = (a_{ik})$ means the same $N \times N$ -matrix as given in (4).

From this we obtain the solution

$$\beta_j^{**} = \frac{1}{|\underline{A}|} \sum_{i=1}^{N} A_{ij} \left[E\Phi_i(x) + d_i \frac{|\underline{A}| - \sum_{k=1}^{N} E\Phi_k(x) \sum_{\ell=1}^{N} d_\ell A_\ell}{\sum_{k=1}^{N} d_k \sum_{\ell=1}^{N} d_\ell A_{k\ell}} \right] , \tag{11}$$

where A_{ij} is the adjunct of a_{ij} .

With the abbreviations

$$D_{ij} = \frac{\sum_{\ell=1}^{N} d_\ell A_{i\ell} \sum_{k=1}^{N} d_k A_{kj}}{\sum_{k=1}^{N} d_k \sum_{\ell=1}^{N} d_\ell A_{k\ell}} , \quad D_j = \frac{\sum_{i=1}^{N} d_i A_{ij}}{\sum_{k=1}^{N} d_k \sum_{\ell=1}^{N} d_\ell A_{k\ell}} ,$$

$$c_{ij} = \frac{1}{|\underline{A}|} (A_{ij} - D_{ij}) ,$$

we can rewrite (11):

$$\beta_j^{**} = D_j + \sum_{i=1}^{N} c_{ij} E\Phi_i(x) .$$

From Theorem 1 it follows at once that the stochastic processes defined by

$$Y_{n+1} = (1 - a_{n+1})Y_n + a_{n+1}[D_j + \sum_{i=1}^{N} c_{ij}\Phi_i(x_n)] , \tag{12}$$

$$Y_0 = b_j \in R' , \quad j = 1,\dots,N$$

converge to β_j^{**} , $j = 1,\dots,N$, with probability one and in the mean if and only if the parameter condition (8) is fulfilled. To avoid unnecessary computations we estimate the parameters $B_j = \beta_j^{**} - D_j$. The final form of the sequential estimation of the unknown vector of parameters $\underline{B}^T = (\beta_1^{**} - D_1, \ldots, \beta_N^{**} - D_N)^T$ is then

$$(13) \qquad \underline{Y}_{n+1} = (1 - a_{n+1})Y_n + a_n \, \underline{C} \, \underline{\Phi}(x_n) \quad , \quad Y_0 = \underline{b} \in R^N \quad ,$$

where $\underline{C}$ is the $N \times N$ -matrix $\underline{C} = (c_{ij})$.

<u>Theorem 4</u>: The process (13) converges to the vector $\underline{B}^T$ with probability one and in the quadratic mean iff the parameter sequence $\{a_n\}$ satisfies condition (8).

We give a simple application. Consider a mixture

$$p(x) = \sum_{i=1}^{N} \beta_i \phi_i(x) \quad , \quad \sum_{i=1}^{N} \beta_i = 1 \quad ,$$

of density functions $\phi_i(x)$, $i = 1,\ldots,N$. The set of functions $\phi_i(x)$ is assumed to be known and to be linearly independent on Ω . Furthermore a sequence of independent observations $\{x_1,\ldots,x_n\}$--identically distribu+ with $p(x)$--may be available from which we want to estimate the parameters β_i , $i = 1,\ldots,N$. This decomposition of a mixture can be done by our sequential estimation procedure (12) or (13). Because d_i equals 1 , $i = 1,\ldots,N$, we get simpler formulas for the D_{ij} and D_j :

$$D_{ij} = \frac{\sum_{\ell=1}^{N} A_{i\ell} \sum_{k=1}^{N} A_{k\ell}}{\sum_{k=1}^{N} \sum_{\ell=1}^{N} A_{k\ell}} \quad , \quad D_j = \frac{\sum_{i=1}^{N} A_{ij}}{\sum_{k=1}^{N} \sum_{\ell=1}^{N} A_{k\ell}} \quad , \quad i,j = 1,\ldots,N \quad .$$

The stochastic processes (12) and (13) converge to the unknown parameters β_j , $j = 1,\ldots,N$, and $B_j = \beta_j - D_j$, respectively.

References

[1] C. C. Blaydon, "Approximation of distribution and density functions," Proc. IEEE (Letters), Vol. 55, pp. 231-232, February 1967.

[2] K. S. Fu, Sequential Methods in Pattern Recognition and Machine Learning. New York: Academic Press, 1968.

[3] R. L. Kashyap and C. C. Blaydon, "Recovery of functions from noisy measurements taken at randomly selected points and its application to pattern classification," Proc. IEEE (Letters), Vol. 54, pp. 1127-1129, August 1966.

[4] R. L. Kashyap and C. C. Blaydon, "Estimation of probability density and distribution functions," IEEE Trans. Information Theory, Vol. IT-14, pp. 549-556, July 1968.

[5] J. Laski, "On the probability density estimation," Proc. IEEE (Letters), Vol. 56, pp. 866-867, May 1968.

[6] Z. J. Nikolic and K. S. Fu, "On the estimation and decomposition of mixtures using stochastic approximation," IEEE - Southwestern Conference Record, pp. 131-138, 1967.

[7] Z. J. Nikolic and K. S. Fu, "A mathematical model of learning in an unknown random environment," Proc. 1966 Nat'l Electronics Conf. (Chicago, Ill.), Vol. 22, pp. 607-612.

[8] Y. Z. Tsypkin, "Use of the stochastic approximation method in estimating unknown distribution densities from observations," Automation and Remote Control, Vol. 27, pp. 432-434, 1966.

[9] H. Wolff, "Zur Konvergenz von Lernprozessen," Unpublished doctoral dissertation, Technical University of Braunschweig, Germany, 1968.

[10] H. Wolff, "Limiting theorems for some generalized Bush-Mosteller Models," Research Bulletin RB-70-23, Educational Testing Service, Princeton, New Jersey. (Also submitted to the Journal of Mathematical Psychology.)

[11] H. Scheffé, "A useful convergence theorem for probability distributions," Ann. Math. Stat., Vol. 18, pp. 434-438, 1947.