LEARNING TECHNIQUE AND THE STOCHASTIC APPROXIMATION METHOD

Fridrich Sloboda and Jaroslav Fogel
Institute for Engineering Cybernetics
Slovak Academy of Sciences,Bratislava
Dubravska cesta, Czechoslovakia

Summary

In the paper is presented a new method for accelerating the convergence of the stochastic approximation method.which can be used as a mathematical technique to solve some learning problems.

1. Introduction

One of the main aims of contemporary theory of adaptive,learning and self-learning systems is the unification of the problems of adaptation,learning and self-learning from one point of view,i.e.the determination of the appropriate mathematical technique which would be able to solve as large a class of problems as possible.Theoretically it seems that the theory of stochastic approximation could solve this problem[914] .

In the learning problem of dividing input situations into n classes,it is possible in a deterministic case to consider the learning-is is known[12] -as an approximation of the function,which divides two disjoint classes.Let this decision function be written as

$$y = f(\vec{x}) \qquad 0)$$

where x is an 1-dimensional vector and y is a quantity showing the class to which a pattern belongs.In the stochastic formulation of the problem where the condition of disjoint classes is not required, y can be taken as the probability that x belongs to one class and 1-y is the probability that x belongs to another class. The function f in equation (1)can be approximated by the finite expansion

$$f(\vec{x}) = \sum_{i=1}^{n} c_i \phi_i(\vec{x}) = \vec{c}^T \phi(\vec{x})$$

where is an n-dimensional vector of coefficients, *(p(2)iB* an n-dimensional vector of linear independent functions and *T* stands for transposition.Let *P)* be an unknown probability density function of the observed vector "x\A measure of the approximation of *i(t)* through *c2)* can be provided by the mean value of any strictly convex function/" with argumentf£u;*-ftf))** Then we can write the measure of the approximation in the form of

$$J(\vec{c}) = E\{F(y-\vec{c}^T\phi(\vec{x}))\} = \int_{\Omega} F(y-\vec{c}^T\phi(\vec{x}))P(\vec{x})d\vec{x} \qquad (5)$$

where^I is the space of all observed vectors A .The best approximation corresponds to such a choice of vector^ that t?a; attains its minimum.This problem can be solved by the stochastic approximation method where the influence of noise can also be considered.

The practical use of the stochastic approximation method as a mathematical technique for learning,self-learning, and adaptation depends on the speed of convergence of the methods used.This paper deals with this problem.From the practical point of view,the solution of extremumsearching problems has great significance .The stochastic approximation methods,as methods of sequential iterations in the case of extremum-searching problems,exploit various forms of the iterative gradient method.Asymptotic behaviour and the speed of convergence were observed by many authors for various modifications of this method.A common feature of all these algorithms of gradient methods an all these modifications did not,however,change.This is a way in which information is gained during the iterative process.The way of gaining information,the structure of the information and its interconnection provide the basic clue to the construction of the new algorithms and are also the criterion of quality.If there is not sufficient information to make the next step in the iterative scheme,then additional information must be gained by experimental interim steps.From this point of view,those algorithms are important which do not need experimental steps before the next step can be made.The stability of used algorithms against noise depends directly on the structure of gained information and its interconnection.Not every algorithm which solves the problem in the deterministic case is stable against noise. Many control processes can be described only through the mathematical models possessing some uncertainties in the form of a set of random variables with unknown probabilistic descriptions.

In this paper we describe an algorithm which has the above mentioned properties; i.e.,it does not use experimental steps and information is gained by steps during the iterative process,and it solves the extremum-searching problem in the presence of noise.

2. Stochastic approximation methods"

The first methods of stochastic approximation were the algorithms of Robbins / and Monro* ,Blum* ,Kiefer and Wolfowitz / for finding the zero point of the function or the extremum of the one-dimensional function in the presence of noise. A generalization for multidimensional functions was made later by Blum* and these results were unified by Dvoretzky', whose results we shall use.As the speed of convergence of the method given by Robbins and Monro depends on the type of function,often this speed is not satisfactory from the practical point of view.

We shall use the two following ways of accelerating the convergence .Let *(l) be a fixed but unknown function,which has a unique rootx= .Let 2 is a random variable with distribution function *W(Vt)* 9 which depends on the real parameter £ ♦ The Robbins-Monro procedure is defined for searching the zero point of regresion function

$$y(x) = \int_{-\infty}^{\infty} z \, dH(z/x)$$

as a sequence $\{x_n\}$ given by relation

$$x_{n+1} = x_n - a_n Z_n \qquad (4)$$

where $\{a_n\}$ is a sequence of positive constants such that

$$a_n > 0 \, , \quad \sum_{n=1}^{\infty} a_n = \infty \, , \quad \sum_{n=1}^{\infty} a_n^2 < \infty \qquad (5)$$

Fabian[5] gives the modification of the algorithm which makes it posseble to accelerate the convergence to within certain limits of the probability density function $f(z/\theta)$ .This limitation for the probability density function results from the condition that the regresion function $\bar{y}(x)$ corresponding to the random variable $\bar{z} = sign\, z$ should satisfy

$$\bar{y}(\theta) = \int_{-\infty}^{\infty} sign(z) f(z/\theta) dz = -\int_{-\infty}^{0} f(z/\theta) dz + \int_{0}^{\infty} f(z/\theta) dz = 0 \qquad (6)$$

as was shown by Avedjan[15] .

The other method of accelerating the convergence of the Stochastic approximation method was given by Kesten[6] .The idea behind this method is that if the expression 4«y*CVVi) retains its signore, then the value of o in the algorithm does not decrease /as it does in the algorithm of Robbins and Monro/[4]but remains unchanged until the first change of signum.On the condition that a*, is a nonincreasing sequence,Kesten shows the convergence of this algorithm with probability one.The advantage of this algorithm lies in the fact that,until the iterative process reaches the domain of its extremum,the algorithm retains a greater value of step and so decreases the influence of noise, as this is indirectly dependent on the value of the step, i.e.,the greater the value of the step the smaller the influence of the random component.Kesten examined the possibility of also utilizing this idea also for the algorithm of Kieffer-Wolfowitz

$$x_{n+1} = x_n - a_n \left[\frac{z(x_n + c_n) - z(x_n - c_n)}{2 c_n} \right] \qquad a)$$

in order to find the extremum of the regression function.Kesten shows the convergence of this algorithm only with an additional strong condition for sequence C. Instead of the original condition C o f the condition *c+*coast,* is essential .And so, in this case,*, in general does not have to converge to the point in which the regression function achieves its extremum.This leads to two opposing demands:

on one hand,the condition of convergence itself,i.e.that the value of should be as low as possible;on the other hand,the question of the speed of convergence, because, the greater the value of the smaller the influence of noise. Obviously, this is valid only so long as the algorithm does not achieve its extremum domain.The idea behind these algorithms for extremum-searching of the k-dimensional function is the way of determination of the gradient.Up to now used standard method for searching the gradient is a static one and the gradient is determined using the experimental steps.In the case *of* stochastic approximation the symmetric gradient is the most suitable,and this needs 2k experimental steps for the determination of direction of the gradient.Further,it will be given the algorithm which creates a sequential process during which the gradient is searched without using the experimental step8.This sequential multidimensional search algorithm creates a sequential process of k onedimensional search algorithms.This algorithm does not have the above drawbacks and its speed of convergence is better than the speed of convergence of the above-mentioned modifications.

3. Multidimensional stochastic approximation method

Let $\{Z(\vec{x})\}$ be a vell-defined family of random variables with unimodal regression function $y(\vec{x})$ which achieves its maximum at the point $\vec{x}=\theta$ .Let E_k be k-dimensional Euclidean space.If $\vec{x}=(x_1,\dots,x_k)\in E_k$ then we denote its norm as $\|\vec{x}\|$ .Let $f(u/x)$ be a probability density function which satisfies the condition

$$\int_0^{\infty} f(u/\theta)\,du - \int_{-\infty}^{0} f(u/\theta)\,du = 0 \qquad (8)$$

where u will be defined latter.

Let $z(\vec{x})=y(\vec{x})+\delta$,where $y(\vec{x})$ is an unknown comtinuous fuction,possessing continuous-partial derivativives of the first order,and δ is a random noise

Let

$$\left|\frac{\partial y(\theta_1,\theta_2,\dots,x_i,\dots,\theta_k)}{\partial x_i}\right| < A|x_i-\theta_i|+B<\infty \qquad (9)$$

$i=1,2,\dots,k$

where A,B are constants.

Theorem /of Dvoretzky/[8]

Let α_n, β_n be non-negative numbers defined for $n=1,2,\dots$;which satisfy the conditions

$$\lim_{n\to\infty}\alpha_n=0 \qquad (10)$$

$$\sum_{n=1}^{\infty}\beta_n=\infty \qquad (11)$$

Let $\{T_n\}$ be Borel measurable vector mappings which satisfy the condition

$$\|T_n(x_n)-\theta\| \le \max\{\alpha_n, \|x_n-\theta\|-\beta_n\} \qquad (12)$$

Let x_1, r_n be random variables and let the following sequence be defined as

$$x_{n+1}=T(x_n)+r_n \qquad (13)$$

Then,if the following conditions are satisfied

$$\sum_{n=1}^{\infty}E\|r_n^2\|<\infty\ ,\quad E\|x_1\|^2<\infty\ ,\quad E\|r_n\|=0 \qquad (14)$$

where E denotes a mathematical expectation,the sequence $\{x_n\}$ converges with probability one and in the mean square to θ i.e.

$$Pr\left\{\lim_{n\to\infty}x_n=\theta\right\}=1 \qquad (15)$$

$$\lim_{n\to\infty}E\left\{\|x_n-\theta\|^2\right\}=0 \qquad (16)$$

Proof:

To prove the convergence x_n in E_k space it is sufficient to prove the convergence of every component of vector $x_n=(x_n^{(1)},x_n^{(2)},\dots,x_n^{(k)})$.

Let us define the algorithm

$$x_{n+1}^{(i)}=x_n^{(i)}-a_n^{(i)}\operatorname{sign}\{\Delta z_n^{(i)}\}\operatorname{sign}\{x_n^{(i)}-x_{n-1}^{(i)}\} \qquad (17)$$

where i is changing cyclically from 1 to k and

$$\Delta Z_n^{(i)}=Z(x_n^{(1)},x_n^{(2)},\dots,x_n^{(i)},x_{n-1}^{(i+1)},\dots,x_{n-1}^{(k)})-$$
$$-Z(x_n^{(1)},\dots,x_n^{(i-1)},x_{n-1}^{(i)},\dots,x_{n-1}^{(k)})$$

Let

$$\sum_{n=1}^{\infty}a_n^{(i)}=\infty\ ,\quad \sum_{n=1}^{\infty}(a_n^{(i)})^2<\infty\ ,\quad a_{n+1}^{(i)}\le a_n^{(i)} \qquad (18)$$

for $i=1,2,\dots,k$

To find an extremum of a function is formally equal to finding the roots of its derivative.Thus instead of equation *(11)* we can write

$$x_{n+1}^{(i)}=x_n^{(i)}-a_n^{(i)}\operatorname{sign}u_n^{(i)} \qquad (19)$$

where i is changing cyclically from 1 to k and

$$u_n^{(i)}=\bar{y}_n^{(i)}+\delta_n \qquad (20)$$

where

$$\bar{y}_n^{(i)}=\frac{\partial y(x_n^{(1)},x_n^{(2)},\dots,x^{(i)},x_{n-1}^{(i+1)},\dots,x_{n-1}^{(k)})}{\partial x^{(i)}} \qquad (21)$$

and $\bar{y}_n^{(i)}=0$ for $x^{(i)}=\theta_n^{(i)}$.

Let

$$E\|u\|=\bar{y}\ ,\quad \sigma^2(x)=E\{\|u-\bar{y}\|^2\}<\sigma^2<\infty \qquad (22)$$

be valid.

From the equation (20) we have

$$\operatorname{sign} u_n^{(i)} = \frac{|\bar{y}_n^{(i)}|}{|u_n^{(i)}|} \operatorname{sign} \bar{y}_n^{(i)} + \frac{|\delta_n|}{|u_n^{(i)}|} \operatorname{sign} \delta_n$$

Let us denote

$$a_n^{(i)} \frac{|\bar{y}_n^{(i)}|}{|u_n^{(i)}|} = \bar{a}_n^{(i)}$$

and

$$\frac{|\delta_n|}{|u_n^{(i)}|} \operatorname{sign} \delta_n = \bar{\delta}_n$$

where $\bar{a}_n^{(i)}$ satisfy the condition (18).

Therefore from equation (19) we obtain

$$x_{n+1}^{(i)} = [x_n^{(i)} - \bar{a}_n^{(i)} \operatorname{sign} \bar{y}_n^{(i)}] - a_n^{(i)} \bar{\delta}_n$$

Let us denote

$$T_n^{(i)}(x_n^{(i)}) = x_n^{(i)} - \bar{a}_n^{(i)} \operatorname{sign} \bar{y}_n^{(i)} \qquad (23)$$

which represents the deterministic part of the algorithm, and

$$r_n^{(i)} = -a_n^{(i)} \bar{\delta}_n \qquad (24)$$

is the stochastic component.
Then the equation (19) can be written

$$x_{n+1}^{(i)} = T_n^{(i)}(x_n^{(i)}) + r_n^{(i)}$$

for $i = 1, 2, \ldots, K$; and

$$T_n(x_n) = (T_n^{(1)}(x_n^{(1)}), T_n^{(2)}(x_n^{(2)}), \ldots, T_n^{(K)}(x_n^{(K)}))$$

$$r_n = (r_n^{(1)}, r_n^{(2)}, \ldots, r_n^{(K)}).$$

Now we prove that conditions of Dvoretzky's theorem are satisfied for every

$T_n^{(i)}(x_n^{(i)})$, $r_n^{(i)}$.

3.1. Stochastic part

It must be proved that the following conditions are satisfied for our algorithm:

$$\sum_{n=1}^{\infty} E\|(r_n^{(i)})^2\| < \infty \,, \quad E\|r_n^{(i)}\| = 0$$

for $i = 1, 2, \ldots, K$.

$$E|r_n^{(i)}| = E|-a_n^{(i)} \bar{\delta}_n| = a_n^{(i)} E|\bar{\delta}_n| = 0 \qquad (25)$$

for $i = 1, 2, \ldots, K$.

If the noise makes a sequence of independent random variables, then we can write

$$E|(\sum_{n=1}^{\infty} r_n^{(i)})^2| = \sum_{n=1}^{\infty} E|(r_n^{(i)})^2|$$

Now we denote the sum of the right side of the above equation as $S^{(i)}$ for every $i = 1, \ldots, K$.

$$S^{(i)} = \sum_{n=1}^{\infty} E|(r_n^{(i)})^2| = \sum_{n=1}^{\infty} E|(a_n^{(i)} \bar{\delta}_n)^2| =$$

$$= \sum_{n=1}^{\infty} (a_n^{(i)})^2 E|\bar{\delta}_n^2| \qquad (26)$$

The mathematical expectation of $\bar{\delta}^2$ is the error dispersion which, according to the assumption, is limited

$$E|\bar{\delta}_n^2| < K_0 \sigma^2 < \infty \,, \quad K_0 > 0$$

Then

$$S^{(i)} < K_0 \sigma^2 \sum_{n=1}^{\infty} (a_n^{(i)})^2$$

and this sum is finite then and only then, when

$$\sum_{n=1}^{\infty} (u_n^{(i)})^2 < \infty$$

for every $i = 1, 2, \ldots K$;
which is satisfied in accordance with the assumption.

3.2. Deterministic part

Let us consider a family of functions

$$\bar{y}_n^{(i)} = \bar{y}(x_n^{(1)}, x_n^{(2)}, \ldots, x_1^{(i)}, \ldots, x_{n-1}^{(K)})$$

whereby

$$\bar{y}_n^{(i)}(\theta_n^{(i)}) = 0 .$$

Let $|x_n^{(i)} - \theta_{n+1}| \geqslant K_1 \bar{a}_n^{(i)}$ for $K_1 > 1$
and denote

$$\nu_n^{(i)} = \theta_{n+1}^{(i)} - \theta_n^{(i)} .$$

Then

$$|T_n^{(i)}(x_n^{(i)}) - \theta_{n+1}^{(i)}| = |x_n^{(i)} - \bar{a}_n^{(i)} \operatorname{sign} \bar{y}_n^{(i)} - \theta_{n+1}^{(i)}| = |x_n^{(i)} - \theta_{n+1}^{(i)}| -$$
$$- \bar{a}_n^{(i)} |\operatorname{sign} \bar{y}_n^{(i)}| \leqslant |x_n^{(i)} - \theta_n^{(i)}| - \bar{a}_n^{(i)} + |\nu_n^{(i)}|$$

(27)

$$\nu_n = \nu_n^{(l)}(\bar{a}_n^{(1)}, \bar{a}_n^{(2)}, \dots, \bar{a}_n^{(i-1)}, \bar{a}_{n-1}^{(i+1)}, \dots, \bar{a}_{n-1}^{(k)})$$

If $\lim \bar{a}_n^{(i)} = 0$ for every $i = 1, 2, \dots, k$, then

$$\lim_{n\to\infty} |\nu_n^{(i)}| = 0$$

which results from the fact that the function is continuous and unimodal. Let us denote

$$\bar{a}_n^{(i)} - |\nu_n^{(i)}| = \gamma_n^{(i)}$$

then

$$\lim_{n\to\infty} \gamma_n^{(i)} = 0$$

and

$$\sum_{n=1}^{\infty} \gamma_n^{(i)} = \infty$$

for every $i = 1, 2, \dots, k$ /$\sum_{n=1}^{\infty} |\nu_n^{(i)}| < \infty$ it follows from Cauchy-Bolzano criterion/

Thus the conditions of the theorem are satisfied.

Let $|x_n^{(i)} - \theta_{n+1}^{(i)}| < \bar{a}_n^{(i)}$

Let $\alpha_n^{(i)}$ be the maximum overstepping of the zero point in the n-th step for i-th parameter.

Then we can write

$$|T_n^{(i)}(x_n^{(i)}) - \theta_{n+1}^{(i)}| = |x_n^{(i)} - \bar{a}_n^{(i)} \operatorname{sign} \bar{y}_n^{(i)} - \theta_{n+1}^{(i)}| =$$
$$= \bar{a}_n^{(i)} |\operatorname{sign} \bar{y}_n^{(i)}| - |x_n^{(i)} - \nu_n^{(i)} - \theta_n^{(i)}| \le \bar{a}_n^{(i)} |\operatorname{sign} \bar{y}_n^{(i)}| -$$
$$- |x_n^{(i)} - \theta_n^{(i)}| + |\nu_n^{(i)}| \qquad (28)$$

According to the assumption for $n \geq n_c$ we can write that

$$|\operatorname{sign} \bar{y}_n^{(i)}| < A |x_n^{(i)} - \theta_n^{(i)}| + B$$

and

$$|T_n^{(i)}(x_n^{(i)}) - \theta_{n+1}^{(i)}| < (\bar{a}_n^{(i)} A - 1) |x_n^{(i)} - \theta_n^{(i)}| + \bar{a}_n^{(i)} B + |\nu_n^{(i)}|$$

as $\lim \bar{a}_n^{(i)} = 0$ then $\bar{a}_n^{(i)} A < 1$

and therefore

$$|T_n^{(i)}(x_n^{(i)}) - \theta_{n+1}^{(i)}| < \bar{a}_n^{(i)} B + |\nu_n^{(i)}| \qquad (29)$$

for every $i = 1, 2, \dots k$.

Let us now denote

$$\bar{a}_n^{(i)} B + |\nu_n^{(i)}| = \alpha_n^{(i)}$$

As $\lim_{n\to\infty} \alpha_n^{(i)} = 0$ the assumptions of the theorem are fulfilled.

As we can see from the above $\lim_{n\to\infty} \theta_n^{(i)} = \theta^{(i)}$

and thus $\max y(x_1, \dots, x_k) = y(\theta_1, \dots, \theta_k)$

where $\theta_i = \theta^{(i)}$.

Up to now we have assumed that the function $y(x_1, \dots, x_k)$ has only one extremum on an arbitrary straight line.The algorithm is not limited only to this class of functions,but satisfies also the class of functions which have more than one extremum on this straight line.This results from the unimodality of the function.The algorithm is shown in fig.1, where the sequence $\{\alpha_i^{(0)}, \beta_i^{(0)}, \dots, \alpha_i^{(m)}, \beta_i^{(m)}\}$ corresponds to those $x_i^{(n)}$ in which the corresponding partial derivative changes its signum.

The conditions given for $\{a_n\}$ are satisfied by the harmonic sequence $\{1/n\}$ or $\{c/n\}$, $c = \text{const}$.

And also by Kesten's sequence which is constructed with the help of harmonic sequence according to the following definition,as the sequence $\{d_n\}$ whose terms are

$$d_1 = a_1,\ d_2 = a_2,\ \dots,\ d_n = a_{s(n)}$$
$$s(n) = 2 + \sum^{n} \phi[(x_i - x_{i-1})(x_{i-1} - x_{i-2})]$$
$$\phi(x) = \begin{cases} 1 & x \le 0 \\ 0 & x > 0 \end{cases} \qquad (30)$$

Using Kesten's sequence,the convergence is accelerated and we avoid the difficulties which occured in part 2.In the domain of the extremum,the value of the step does not affect the noise level and then only the law of large numbers is valid.Therefore,in the domain of the extremum,Kesten's sequence does not accelerate the convergence,and it is better to use the harmonic sequence,i.e., to lessen the value of the step as quickly as possible.This modification gives better results than other modifications.

4. Coclusion

The above-mentioned method of stochastic approximation accelerates the convergence and therefore has an influence on the quality of learning.This method satisfies the condition so-called "An ideal stochastic approximation",i.e.,steps of large value are used far from the extremum; and the nearer the extremum,the smaller the step becomes*The fact that this method does not use the experimental steps means that,in this case,the convergence is accelerated and the method gives considerable better results than other methods .The method was used in pattern recognition for disjoint classes.

Fig.I. The multidimesional search algorithm

<u>Bibliography:</u>

1. H.Robbins,S.Monro, "A Stochastic Approximation Method" A.M.S.22/1951/ pp.400-407

2. J.Kifer,J.Wolfowitz, "Stochastic Estimation of the Maximum of a Regression Function"A.M.S. 23 /1952/, pp.462-466

3. J.R.Blum, "Approximation Methods which Convergence with Probability One" A.M.S. 25 /1954/,pp.382-386

4. J.R.Blum, "Multidimensional Stochastic Approximation Methods" A.M.S. 25 /1954/,pp.463-483

5. V.Fabian, "Stochastic Approximation Methods" Cech.Mat.Zurn. 10/85/ No.l, /1960/,pp.123-158

6. H.Kesten, "Accelerated Stochastic Approximation" A.M.S. 29/1958/ pp.41-59

7. V.Dupac, "0 Kiefer-Wolfowitzove Approximacni Metode" Cas.Pest.Mat. 82/1957/,pp.47-75

8. A.Dvoretzky, "On Stochastic Approximation"Proc.of the Third Berkeley Symposium on Math.Statistics and Probability,Vol.1,Berkeley,1956 pp.39-55

9. D.J.Wilde, "Optimum Seeking Methods", Prentice Hall,1964

10. Z.Cypkin, "Adaptation,Learning and Self-learning in Control Systems" Third Congress of IFAC,London 1966

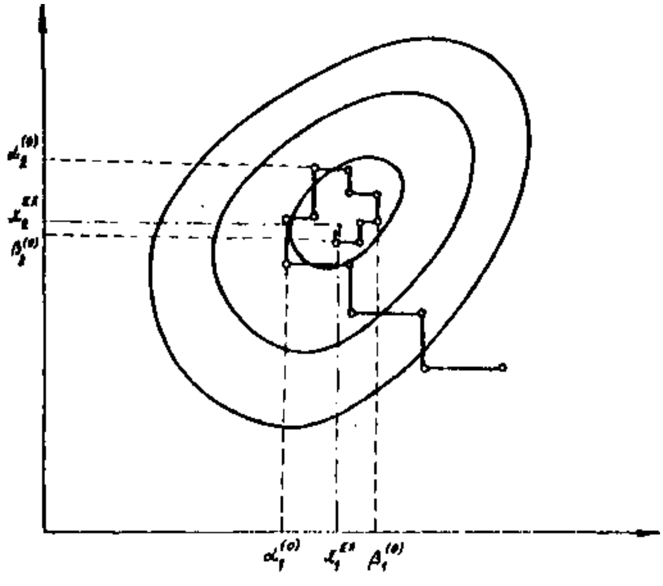
11. F.Rosenblatt, "Principles of Neurodynamics Perception and the Theory of Brain Mechanisms" Spartan Books, Washington D.C.1961

12. Aizerman,Braverman,Ro2onoer, "Veroiatnostnaia Zadaca Ob Obucenii Automatov Raspoznavaniu Klasov i Metod Potencialnych Funkcii".Automatika i Telemechanika,Moskva, T.25,No.9,1964

13. Z.Cypkin, "A vse ze suscestvuet li teoria sinteza optimalnych adaptivnych sistem?"Automatika i Telemechanika, No.l,1968,Moskva

14. I.Stratonovic, "Suscestvuet teoria sinteza optimalnych adaptivnych, samoobucaiuscichsa samonastraivavaiuscichsd sistem?"Automatika i Telemechanika,No.1,1968,Moskva

15. E.D.Avedian, "K odnoj modifikacii algoritma Robbinsa i Monro",Automatika i Telemechanika,No.4,1967



Figurc 1. The Multidimensjonal Search Algorithm