# ADAPTIVE STOCHASTIC APPROXIMATIONS

by

Karel JANÁČ, M.Sc.,Ph.D.

Forschungsbericht Nr. 32
Jänner 1969

# ADAPTIVE STOCHASTIC APPROXIMATIONS

Karel JANÁČ, M.Sc., Ph.D.

Many of the present problems in automatic control economic systems and living organism can be converted to parameter optimization in stochastic systems. Foremost among these problems are questions of the control of systems with incomplete information, learning problems, adaptive control, identification of objects, and the automatic synthesis of objects.

Such problems can be solved by stochastic approximation methods which are, essentially, iterative procedures [1]. For this reason, great attention is paid to these methods in connection with practical applications. They were elaborated as a purely mathematical problem a long time ago and a number of valuable results are now available [2]. Not only the conditions of convergence, but some properties of the asymptotic speed of convergence are also known.

In some cases, however, a disadvantage of stochastic approximations is the slow convergence to the desired extreme of the optimality criterion. At present, utmost attention is devoted to the elimination of these undesirable properties. Unfortunately, practical requirements are often in disagreement with the assumptions from which we start when seeking more effective algorithms [3,4] . One of the assumptions that cannot be satisfied when applying this method is the execution of an infinite number of iterative steps ($n \rightarrow \infty$). This means that we must confine ourselves to a finite n and adopt some method for deciding whether the required optimal parameter values have been attained. Let us assume axiomatically, that such a decision is known. An algorithm leading within the shortest possible time (or, equivalently, at minimum cost) to the decision concerning the finding of the extreme will then be considered to be the most suitable. Such an assessment involves not only demande made on the optimal algorithm, but also on its complexity and thus also on the time required for its implementation. For the time being, the procedure of finding the optimal algorithm for a finite n is theoretically impracticable. It will therefore be advisable to investigate algorithms which retain good asymtotic properties and are in many cases of greater advantage than the algorithms commonly used heretofore. Of course, they must also be easy to implement. Several algorithms of this type will now be discussed.

Let us consider the solution of the following problem. Take a k-dimensional Euclidean space $X = E^k$, over which a k-dimensional regression function R(x) is defined. R(x), where $x \in X$, has only one minimum for x = 0. Known are only estimates of this functions, $\hat{R}(x)$, but not the values of R(x). Our task is to find the minimum of the function R(x).

New stochastic approximation methods have been presented by Dr. Fabian [5], who utilizes the recurrence relations

$$X_{n+1} = X_n + \frac{a_n}{c_n}(h_n - 1)\ |Y_n|\ \text{sign}\ Y_n \qquad (1)$$

and the further modification

$$X_{n+1} = X_n + \frac{a_n}{c_n}(h_n - 1)\ \text{sign}\ Y_n \qquad (2)$$

Both methods converge with probability one under conditions sufficiently general for practical application (for details see [5])and for

$$c_n \rightarrow 0,\ \sum_{n=1}^{\infty} a_n = +\infty\ ,\ \sum_{n=1}^{\infty} a_n c_n < +\infty\ ,\ \sum_{n+1}^{\infty} \left(\frac{a_n}{c_n}\right)^2 < +\infty \qquad (3)$$

In both relations, $h_n$ equals the first unsuccessful step in the direction of the estimated gradient. The method thus utilizes an estimate of the gradient of the function R(x) for several working steps in the same direction. Thus, every estimate of the gradient is utilized to a greater extent for optimizing the function R(x). This is of special advantage in the multidimensional case, where the determination of the estimate of the gradient is highly time-consuming (requiring k or k+1 calculations).

When using the recurrence relation (1), both the length and direction of the working step are determined by the value of the estimated gradient. With the recurrence relation (2), the length of the working step is independent of the estimated gradient and its direction is determined by the diagonal of a k-dimensional parallelepiped. This algorithm is of advantage in case the regression function is flat for values of the parameters x remote from the optimal ones. That is to say, in such a case the algorithm (1) (as well as other known procedures) results in small working steps.

The author presents a new, very easily realizable modification of the algorithm, which combines the advantages of the procedures (1) and (2).

For this purpose, use is made of the non-linear recurrence relation

$$X_{n+1} = X_n + \frac{a_n}{c_n} (h_n - 1) f(Y_n) \qquad (4)$$

where f is an odd, non-decreasing function, the absolute value of which has positive constants for its upper and lower bounds. Thus, $f(-Y) = -f(Y)$; then $f(Y_2) \geq f(Y_1)$ for every $Y_2 \geq Y_1$. Furthermore, $K_2 < |f(Y)| < K_1$. (5)

For practical reasons it is advisable to choose the function f as follows:

$$\begin{aligned} f_1 &= K \operatorname{sign} Y && \text{for } 0 \leq |Y| \leq K \\ f_1 &= Y && \text{for } K \leq |Y| \leq 1 \qquad (6) \\ f_1 &= \operatorname{sign} Y && \text{for } 1 \leq |Y| \end{aligned}$$

The form of such a function is shown in Fig.1. Obviously, we move in the direction of the gradient as long as the estimates of the derivatives in the individual directions of $Y_n^{(i)}$ lie between K and unity. In all other cases it is certain that no parameter will change by less than $K \frac{a_n}{c_n}$ and by more than $\frac{a_n}{c_n}$.

The fact that the method (4) retains its properties of convergence follows from the two considerations presented below and from the proof given in [5] . Let us first consider a recurrence relation that may be regarded as a modification of the Blum method,

$$X_{n+1} = X_n + \frac{a_n}{c_n} f(Y_n) \qquad (7)$$

Then consider the relation given in [5] , for which convergence is ensured, i.e.

$$X_{n+1} = X_n + \frac{a_n}{c_n} K \text{ sign } Y_n \qquad (8)$$

Let ${}^1X_n$ be a vector for which the relation (8) holds good when $K = K_1 > |f(Y)|$ , and ${}^2X_n$ a vector for which (8) holds good when $K = K_2 < |f(Y)|$ . For every step we have

$$\left\|{}^2X_{n+1} - X_n\right\| < \left\|X_{n+1} - X_n\right\| = \left\|\frac{a_n}{c_n} f(Y_n)\right\| < \left\|{}^1X_{n+1} - X_n\right\| \qquad (9)$$

and the convergence of the relation (6) then follows from the procedure used to prove the convergence of the relation (8).

The proof relating to the convergence of the method involving several working steps [5] also applies to the properties of convergence of the relation (4).

The method (4) combines the advantages of the two procedures (1) and (2) and it is very easy to realize technically. When solving problems by hybrid computation techniques, it can be implemented as shown in Fig.2, i.e. by adding a single nonlinear function generator which produces the function f(Y).

An unpleasant feature appears when practically applying the method which utilizes several working steps in the direction of the estimated gradient. Let us consider the situation occuring when the sensitivity of the regression function R(x), expressed by the partial derivatives $\frac{\partial R(x_i)}{\partial x_i}$, is in some region considerably larger for several parameters $x_1, x_2, \ldots, x_r$

than its sensitivity with respect to the parameters $x_{r+1}$, $x_{r+2}$, ..., $x_k$. In this case there is a large probability that the gradient of the regression function R(x) will be correctly estimated for the parameters $x_1$, $x_2$, ..., $x_r$, but incorrectly (with opposite sign) for some of the parameters $x_{r+1}$, $x_{r+2}$, ..., $x_k$. In consequence, even though the function R(x) is optimized in the course of the working steps, some parameters will recede considerably from the point of extremes of R(x). This case cannot be considered as being favourable. In order not to loose information on the form of the gradient of the function R(x) for too long a period, it is possible to use a modified recurrence relation which limits the number of working steps, namely

$$X_{n+1} = X_n + \frac{a_n}{c_n} (H_n - 1) f(Y_n) \qquad (10)$$

where $H_n = h_n$ for $h_n \leq A$

and $H_n = A$ for $h_n > A$,

A being a natural number ($A > 2$) suitably chosen with regard to the number of parameters $x_1$, ..., $x_i$, ..., $x_k$, so that a certain equilibrium is reached between the number of calculations needed for estimating the gradient and the number of working steps.

When using stochastic approximations, the sequences $a_n$ and $c_n$ must be chosen so as to satisfy the conditions for the convergence of the method and to obtain, at the same time, the maximum speed of convergence. For this purpose it is possible to make use of a number of papers ( [6] [7] [8] ) which treat this problem and seek sequences optimal in the asymptotic sense as well. Attempts have also been made to find optimal sequences for a finite n. At present, however, it seems that in the general case there is no possibility of determining the optimal length of the step $a_n/c_n$ for a finite n, and therefore it appears reasonable to consider algorithms which would adapt the lenght of the step to the course of the optimization process.

This means using the principle of experience. Such an approach has the advantage of not requiring any further estimates of the function R(x) (e.g., when considering the form of the second derivative).

The procedure outlined below derives the information needed for adaptation from the requirement of a certain balance between the number of working and trial steps and the mean value of the variation in the direction of the parameters during the last q steps.

Let the length of the working steps be

$$\frac{a_n^x}{c_n^x} = \varphi_n \frac{a_n}{c_n} \qquad (11)$$

where $1 \leq \varphi \leq c$.

In addition, let us assume that

a) $\varphi_{n+1} = \varphi_n + d$

for $h_n > A_1$ and $\sum_{l=0}^{q} \sum_{i=1}^{k} \left| \text{sign } Y_{n-l-1}^{(i)} - \text{sign } Y_{n-l}^{(i)} \right| < \frac{2qk}{b_1}$

where $b_1 \leq k$; $d < c - 1$ is a positive number, (12)

b) $\varphi_{n+1} = \varphi_n - d$

for $h_n < A_2$ and $\sum_{l=0}^{q} \sum_{i=1}^{k} \left| \text{sign } Y_{n-l-1}^{(i)} - \text{sign } Y_{n-l}^{(i)} \right| > \frac{2qk}{b_2}$

where $b_2 < b_1$, $A_2 < A_1$ are natural numbers;

c) in all other cases,

$$\varphi_{n+1} = \varphi_n$$

The values of $\varphi_n$ are, for example, integers between one and ten.

The length of the steps is varied in the aforesaid manner so that the number of working steps should not be larger than $A_1$ (increasing the length of the steps), but notsmaller than $A_2$ (reducing the length of the steps). At the same time, the mean number of variations in the direction of all parameters during the last q steps is taken into account. If we have moved predominantly in the same direction and if $h_n > A_1$, the length of the steps will have to be increased. If we have moved predominantly in various directions and if $h_n < A_2$, the length of the steps will have to be reduced. In all other cases, $\varphi_n$ will not be changed and we proceed according to the original sequence $a_n/c_n$. $A_1$, $A_2$, q, $b_1$, and $b_2$ must be chosen so as to utilize the adaptive properties of the procedure, but so as to make the adaptation less dependent on random observational errors of the function R(x). For example, if we choose $b_1 = k$ and $b_2 = 0$, we will have to increase the working steps only if, during all q preceding steps, all the parameters varied in the same direction, and we will have to reduce them if the direction of the parameters has alternated. (Analogously, $c_n^x = \psi_n c_n$ can also be varied by the same principle as $a_n/c_n$). Fig.3., where the sums are evaluated approximately by a simple analogue-digital element with memory, shows how simple it is to implement the procedure presented above. The change of $\varphi$ (or $\psi$ ) is produced by a digitally switched voltage divider.

The function $\varphi$ (or $\psi$ ) has positive constants for its upper and lower bounds, and for this reason the properties of convergence of the recurrence relation (10) are not disturbed. In the vicinity of the optimum, $\varphi_{min}$ will obviously be treated in a manner conforming to the method of adaptation (12).

When choosing $a_n$ and $c_n$ by the criteria of maximum speed of convergence, the value of $a_n/c_n$ drops in the asymptotic case very rapidly from the original value $a_1/c_1$. This circumstance can lead to a situation where the optimum of the function R(x) will be approached by steps too small even of the length of the step is modified. Such a situation is indicated by the fact that $\varphi = \varphi_{\max}$ for a number of steps. This information can be utilized for a hierarchic control of the adaptive properties of stochastic approximations.

In case that

$$\sum_{k=0}^{r} \varphi_{n-1} = r\, \varphi_{\max} \qquad (13)$$

we take

$$\frac{a_n^{x\,x}}{c_n^{x\,x}} = \frac{a_{n-N}^{x}}{c_{n-N}^{x}} \; ; \; c_n^{x\,x} = c_{n-N}^{x}$$

where $N \gg 1$.

This means that we return by N steps in the sequence of the coefficients $a_n$ and $c_n$. In this way it is possible to pass on to considerably larger steps while the algorithm is implemented in a rather simple manner. Stochastic approximations controlled in this way can also be used for following the point of optimum in slightly non-stationary problems.

All the modifications presented above are characterized by their simple realizability. If the length of the step is varied separately for each individual parameter, the realization becomes far more complicated. Using the principle of experience, it is possible to implement the following method of adaptation:

$$\frac{a_n^{(i)^x}}{c_n^{(i)^x}} = \varphi_n^{(i)} \frac{a_n}{c_n} \qquad (14)$$

where $1 \leq \varphi^{(i)} \leq c$, and let

a) $\varphi_{n+1}^{(i)} = \varphi_n^{(i)} + d$ for $\sum_{l=0}^{q} \left| \text{sign } Y_{n-1-l}^{(i)} - \text{sign } Y_{n-l}^{(i)} \right| < \frac{2q}{b_1}$

b) $\varphi^{(i)}_{n+1} = \varphi^{(i)}_{n} - d$ for $\sum_{l=0}^{q} \left|\mathrm{sign} Y^{(i)}_{n-1-l} - \mathrm{sign}\ Y^{(i)}_{n-l}\right| > \frac{2q}{b_2}$

c) $\varphi n+1 = 1$ for $\sum_{l=0}^{q} \left|\mathrm{sign}\left|Y^{(i)}_{n-1-l} - \mathrm{sign}\ Y^{(i)}_{n-l}\right| > \frac{2q}{b_3}\right.$

where $1<b_3<b_2<b_1$; for $i = 1, 2, \ldots, k^x$.

In cases where the change of parameters during the q last steps does not proceed predominantly in the same direction, condition c) is intended to help us quickly to progress again in the direction of the estimated gradient.

The iterative procedures presented above cannot be considered as optimal, but the principle of adaptation of the step length has a number of advantages, since it utilizes to a greater extent the information on the form of the regression R(x). Their application is of special advantage in cases where it is impossible to execute a large number of steps n and thus to make full use of the work done by various authors on the asymtotic properties of stochastic approximations. The new methods described here are characterized by the ease with which they can be implemented. Even in cases where the classical Blum method would appear to be the best, the aforementioned methods exhibit properties that are only slightly worse (especially in the multidimensional case). Their chief advantage thus lies in their greater versatility.

1 Tsypkin Ya.E.: Adaptivity, Learning and Self-learning in Automatic Systems (in Russian). Institut Avtomatiki i Telemekhaniki, Moscow, 1965.

2 Schmetterer L.: Stochastic Approximation. Proc. of the 4th Berkeley Sympos.Math.Statist. 1961, Probab.1, pp. 587-609.

3 Dvoretzky A.: On Stochastic Approximation. Proc. of the 3rd Berkeley Sympos.Math.Statist. 1956, Probab.1, pp. 39-55.

4 Tsypkin Ya.E.: There does exist a theory of the synthesis of optimal adaptation systems. (in Russian). Avtomatika i Telemekharika, 1968, No.1, pp. 108-115.

5 Fabian V.: Stochastic Approximation Methods. Czech.Math. Journal, 10, 1960, pp. 123-159.

6 Sakrison D.J.: Application of Stochastic Approximation Methods to System Optimization. MIT Techn.Rep.391, July 10, 1962.

7 Fu K.S., Chien Y.T., Nikolic Z.J., Wee W.G.: On the Stochastic Approximation and Related Learning Techniques. Purdue University, April 1966.

# Adaptive Stochastic Approximations

Karel Janaśč, M.Sc. Ph. D.

Many of the present problems in automatic control can be converted to the parameter optimization in stochastic systems.

Such problems can be solved by stochastic approximation methods which consist, essentially, if iterative procedures. However, in a number of cases they suffer from the disadvantage of a low speed of convergence to the desired extreme of the optimality criterion. In a number of practical cases, the properties of asymptotic convergence of stochastic approximations cannot be utilized.

In this paper the author presents several new algorithm modifications which in some cases may be more advantageous as regards the finite number of steps n and still retain good asymptotic properties.

The first method makes use of the non-linear recurrence relation

$$x_{n+1} = x_n + \frac{a_n}{c_n} (h_n - 1) f(Y_n) \qquad (1)$$

where f is a special function. The method involves $(h_n - 1)$ working steps in the direction of the estimated gradient. $h_n$ is a natural number determining the order of the first unsuccessful step.

Another modification limits the number of working step in order to prevent loss of information concerning the form of the gradient for excessive periods of time.

A further method consists in adapting the length of the step (multiplying by $\varphi_n$) on the basis of the number of working steps and of experience gained during preceding steps.

$$\frac{a_n^x}{c_n^x} = \varphi_n \frac{a_n}{c_n} , \quad \text{where} \quad 1 \leq \varphi \leq c \qquad (2)$$

The hierarchic principle of adaptation can be used in cases where a number of steps is performed with $\varphi = \varphi_{max}$. We then assume that

$$\frac{a_n^{x\ x}}{c_n^{x\ x}} = \frac{a_{n-N}^x}{c_{n-N}^x} , \quad \text{where } N \gg 1 \qquad (3)$$

This means that we return N steps in the sequence of coefficients.

The last modification consists in adapting the length of the step separately for each individual parameter,

$$\frac{a_n^{(i)}}{c_n^{(i)}} = \varphi_n^{(i)} \frac{a_n}{c_n} \qquad (4)$$

where $\varphi_n^{(i)}$ is changed on the basis of experience obtained in the foregoing steps.

The stochastic apprcximations presented in this paper are characterized by the ease with which they can be implemented. They are of special advantage in cases where it is impossibel to execute a sufficiently large number of steps n. The procedures outlined here retain the good asymptotic properties of known stochastic approximations.

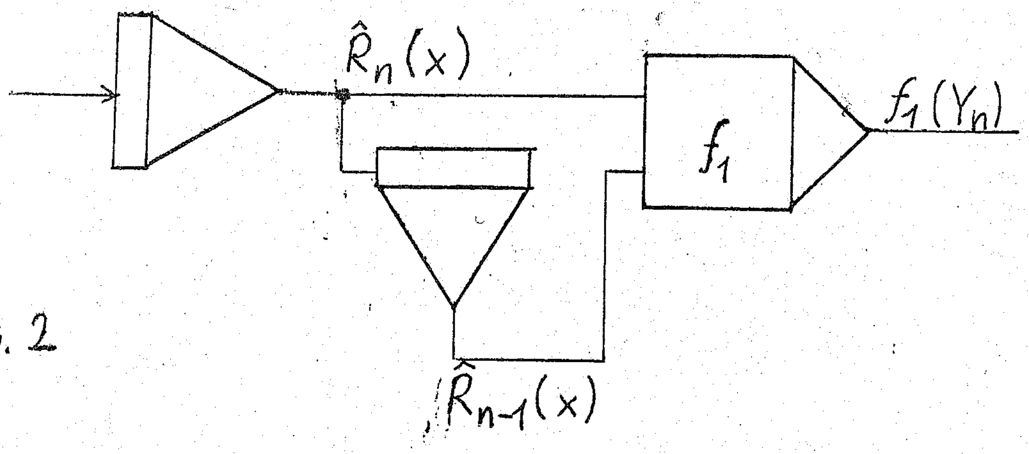
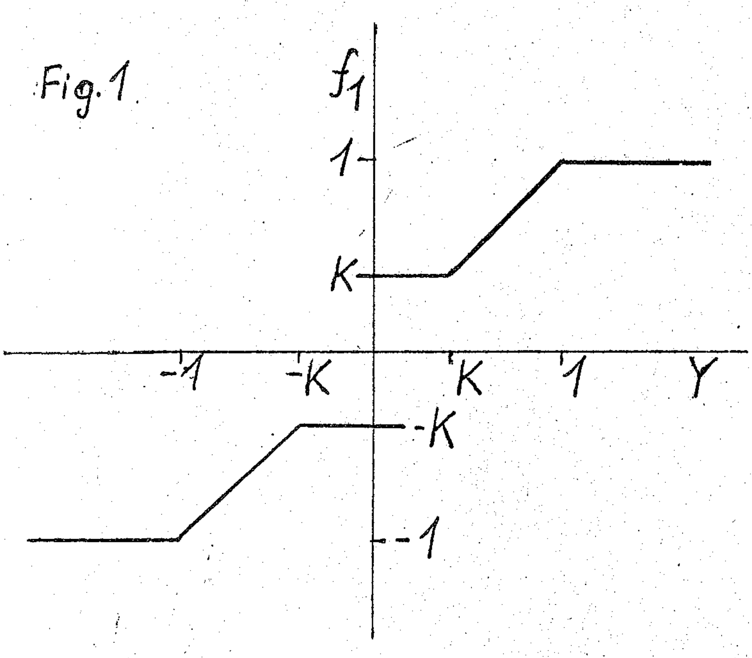


Fig. 2

Fig. 1



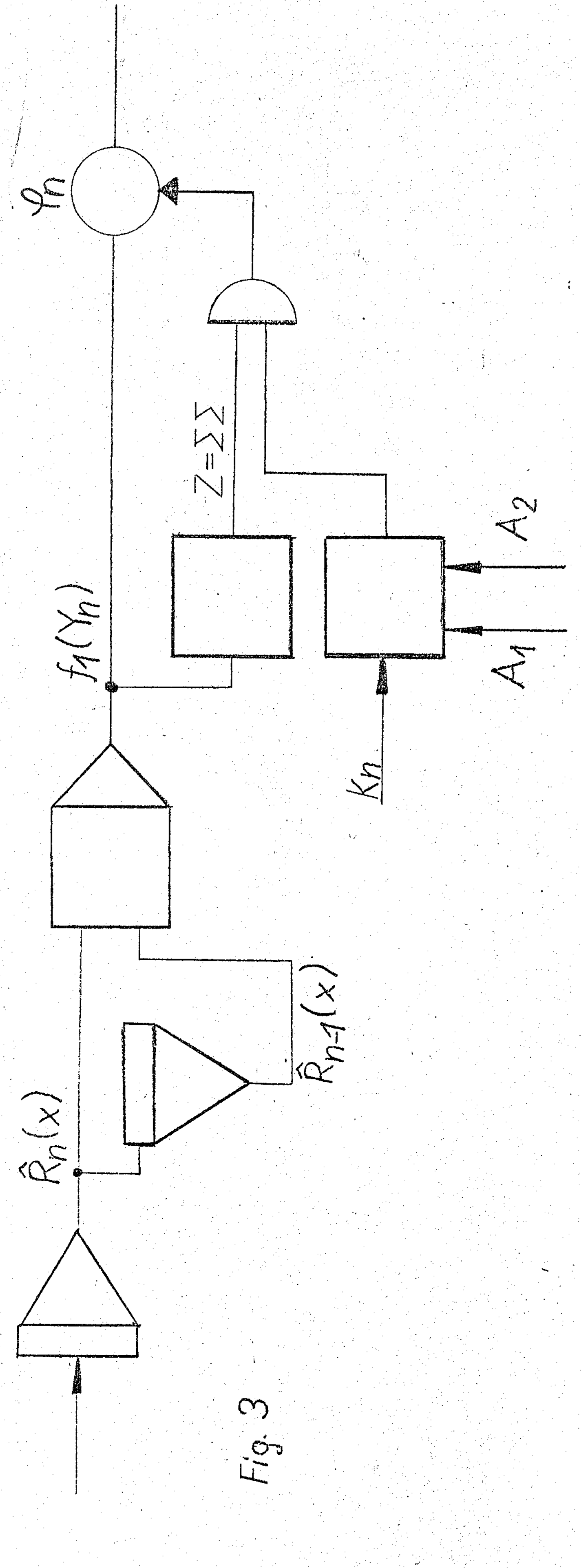


Fig. 3