Journal of Cybernetics, 1974, 3, 3, pp. 93-103

Random Search in Problems of Optimization, Identification and Training of Control Systems

L. A. Rastrigin

1. Statement of Problem

The problem of adaptation of control systems (in the wide sense) reduces to finding the extremum of a functional that specifies the performance of the system to be adapted. We shall assume that the problem is parametrized, i.e., it reduces to finding parameters $x_1, \ldots, x_n$ of the adapting plant that extremize the functional

$$Q = Q(X, A, B) \tag{1.1}$$

under given constraints $X \in G$, where $X = (x_1, \ldots, x_n)$ is the vector of the parameters to be optimized, G is the set of admissible values of this vector, $A = (a_1, \ldots, a_m)$ is the vector of input signals, and $B = (b_1, \ldots, b_k)$ is the vector of the system's response to the disturbances A and to $E = (\epsilon_1, \ldots, \epsilon_p)$, which is the vector of uncontrolled parameters of the system (noise, drift, etc.). The function $B = F(A, X, E)$ is not known. The vectors X, A and B are assumed controllable, though the vector A is often characterized by the situation in which the plant operates, i.e., it must be regarded as a random function.

Figure 1 shows the interaction between the plant and the controlling (adapting) device which is assigned the objective of the control, i.e., the form of the extremized functional Q.

In the particular case that $Q(X, A, B)$ is a continuous and differentiable function, the problem of adaptation will be solved by the obvious equation

$$\frac{dX}{dt} = \pm a \operatorname{grad}_X Q\,(X, A, B), \tag{1.2}$$

where the plus sign corresponds to maximization, the minus sign to minimization of the indicator Q; a is a parameter (in general, variable) of the rate of adaptation. For definiteness we shall henceforth consider the case of minimization of Q, when the right-hand side of (1.2) has a minus; this can evidently be done without loss of generality. In discrete form the equation (1.2) is written as follows:

$$X_{N+1} = X_N - a_N \operatorname{grad} Q\,(X_N, A_{N+1}, B_{N+1}) \tag{1.3}$$

(N being the number of iteration).

However in practice the gradient grad Q is usually replaced by its estimate grâd Q, which can be obtained either analytically, or by search.

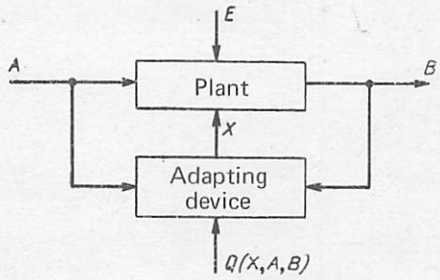

Fig. 1.

In the case that the functional (1.1) has discontinuities, the concept of gradient becomes meaningless and the adaptation Eq. (1.2) must be replaced by another equation, for example:

$$\frac{dX}{dt} = -a\Phi[Q(X,\ A,\ B)], \tag{1.4}$$

where Φ is a vector functional that makes it possible to determine the trend of variation of $Q(X)$. For example,

$$\Phi = (X_2 - X_1)(Q_{max} - Q_{min})/|X_2 - X_1|, \tag{1.5}$$

where

$$Q_{max} = Q(X_2),\ Q_{min} = Q(X_1).$$

Let us consider the principal aspects of the adaptation problem as relating to estimation of the gradient or of its corresponding functional by the random search method. Such an approach to the adaptation problem is not as esoteric as it may appear at first sight. It is found that in the well-known adaptation procedures the randomness element plays an important role and can be regarded as an application of the random search method. Let us show this.

2. Search Optimization

The problem of search optimization is usually formulated for plants of "black box" type (although its formulation for "light" but complex plants is not ruled out). In this case the general scheme of adaptation is simplified (Fig. 2):

$$B = b,\ Q = b = Q(X,\ A,\ E), \tag{2.1}$$

with the functional F, connecting B, A, X and E, being unknown (this "blackens" the plant). With respect to this functional we can a priori state two things: it can be either without lag (without memory), or dynamical. For simplicity we shall assume the plant to be without lag.

Since the input A is not observed, it can be included in the noise and drift factor E. Thus the problem reduces to finding an admissible vector $X^* \in G$ that minimizes the performance index $Q = Q(X^*, E)$ of the plant irrespective of the values of the noise factor E.

Let us consider the case of additive noise:

$$Q'(X) = Q(X,\ E) = Q(X) + \varepsilon(\sigma), \tag{2.2}$$

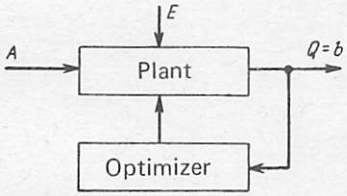

Fig. 2.

where the function $Q(X)$ is not known and $\epsilon(\sigma)$ is a normally distributed random number with zero mean and a variance σ^2. In this case adaptation requires estimation of the gradient of the performance index, this estimation being actively carried out, i.e., by search.

One of the common methods of estimation of the gradient direction is the gradient method, which in fact is a numerical process of determination of partial derivatives (the gradient estimate will be marked by the subscript g):

$$\widehat{\text{grad}}_g\, Q'(X) = \frac{1}{2g}\,[Q'(X + ge_1) - Q'(X - ge_1), \ldots$$

$$\ldots,\ Q'(X + ge_n) - Q'(X - ge_n)], \tag{2.3}$$

where the e_i $(i = 1, \ldots, n)$ are unit vectors.

The gradient can be estimated also in a random manner, for example as follows (the random estimate will be marked by the subscript r):

$$\widehat{\text{grad}}_r\, Q'(X) = \frac{1}{2g}\,\Xi\,[Q'(X + g\Xi) - Q'(X - g\Xi)], \tag{2.4}$$

where $\Xi = \xi_1, \ldots, \xi_n)$ is a unit random vector uniformly distributed in all the directions of the space $\{X\}$ of parameters to be optimized. This is undoubtedly a rougher estimate than (2.3), but on the other hand it has been obtained by simpler means, i.e., by only two determinations of the index Q instead of the $2n$ determinations needed for (2.3). Indeed, for the gradient method we need $2n$ determinations of the performance index, and for random search only two, i.e., by a factor of n less, which is particularly important in the optimization of plants with a large number of parameters to be optimized.

For comparing the two methods let us determine their principal probabilistic characteristics in a linear field.

Thus for the gradient method of estimation of the gradient we have

$$M(\widehat{\text{grad}}_g\, Q) = \text{grad}\, Q, \tag{2.5}$$

$$D(\widehat{\text{grad}}_g\, Q) = \int_0^\infty \rho^2 p(\rho)\, d\rho, \tag{2.6}$$

where $\rho = |\mathrm{E}|$, with E being the noise vector

$$\mathrm{E} = \frac{1}{2g}\,(\varepsilon_2 - \varepsilon_1,\ \varepsilon_4 - \varepsilon_3,\ \ldots,\ \varepsilon_{2n} - \varepsilon_{2n-1}),$$

ϵ_i are realizations of the random noise $\epsilon(\sigma)$, and $p(\rho)$ is the probability density of ρ, equal to a chi-distribution

$$p(\rho) = \rho^{n-1}\, e^{-\rho^2/4\sigma_0^2}/2^{n-1}\sigma_0^n\, \Gamma(n/2), \tag{2.7}$$

whose mean and variance are

$$M(\rho) = 2\tau_0 \sqrt{n}\, \Gamma\left(\frac{n+1}{2}\right) \Big/ \Gamma\left(\frac{n}{2}\right),$$
$$D(\rho) = 2\tau_0^2 n \left[1 - 2\left(\Gamma\left(\frac{n+1}{2}\right) \Big/ \Gamma\left(\frac{n}{2}\right)\right)^2\right]. \tag{2.8}$$

Here we used the notation $\sigma_0 = \sigma/2g$.

For the variance of the estimate (2.3) we hence obtain

$$D(\widehat{\text{grad}}_g Q) = M^2(\rho) + D(\rho) = \tau^2 n/2g^2. \tag{2.9}$$

As we can see, the estimate (2.3) is unbiased, which could be expected, since its variance is proportional to the dimension of the optimized plant and does not depend on the gradient of the performance index.

Now let us consider the statistical properties of the random estimate (2.4):

$$M(\widehat{\text{grad}}_r Q) = \frac{1}{2g} M_{\Xi}\{\Xi[Q(X + g\Xi) - Q(X - g\Xi)]\} =$$
$$= B_n \text{grad}\, Q \int_0^{\pi/2} \cos^2\varphi \sin^{n-2}\varphi d\varphi = \frac{1}{n} \text{grad}\, Q,$$

where [1]

$$B_n = \frac{1}{2}\left[\int_0^{\pi/2} \sin^{n-2}\varphi\, d\varphi\right]^{-1} = \Gamma\left(\frac{n}{2}\right) \Big/ \sqrt{\pi}\, \Gamma\left(\frac{n-1}{2}\right). \tag{2.10}$$

Next,

$$D(\widehat{\text{grad}}_r Q) = \frac{1}{4g^2} D_{\Xi}\{\Xi[\Xi, \text{grad}\, Q]\} +$$
$$+ \frac{1}{4g^2} D_{\Xi, \varepsilon}\{\Xi\varepsilon(1/\sqrt{2}\tau)\} = D_1 + D_2. \tag{2.11}$$

The first term of this expression specifies the variance due to the randomness of the search,

$$D_1 = 2B_n k^2 \int_0^{\pi/2} \cos^4\varphi \sin^{n-2}\varphi d\varphi - \frac{k^2}{n^2} = 2k^2(n-1)/n^2(n-2), \tag{2.12}$$

where

$$k = |\text{grad}\, Q|.$$

This signifies that when the dimension of the plant increases, the variance due to randomness of the search decreases, tending to zero as $1/n^2$. The variance of random search due to noise is equal to

$$D_2 = \sigma^2/2g^2. \tag{2.13}$$

As we can see, the random estimate of the gradient (2.4) is directed along the gradient vector and biased in absolute value by a factor of $1/n$. This makes it possible to construct an

estimate which does not have such a shortcoming, i.e., which is not biased:

$$\widehat{\text{grad}}_{r_0} Q = \frac{n}{2g} \Xi [Q(X + g\Xi) - Q(X - g\Xi)]. \tag{2.14}$$

As is easy to see, it has the following properties:

$$\begin{aligned} M(\widehat{\text{grad}}_{r_0} Q) &= \text{grad } Q, \\ D(\widehat{\text{grad}}_{r_0} Q) &= 2k^2 \frac{n-1}{n+2} + \frac{n^2\sigma^2}{2g^2}. \end{aligned} \tag{2.15}$$

The variance of this unbiased estimate increases with the dimension of the plant as n^2. Hence we can assert that the variance of a gradient estimate for $n > 1$ is always smaller than that of an unbiased random estimate, i.e.,

$$D(\widehat{\text{grad}}_g Q) < D(\text{grad}_{r_0} Q) \tag{2.16}$$

(let us note that for small k the biased random estimate (2.4) has the converse property).

The larger variance of an unbiased random estimate is the "price" for reducing the number of samples (two instead of $2n$). By taking $2m$ samples in a random manner, we can use the following estimate:

$$\widehat{\text{grad}}_{r,g} Q = \frac{n}{2gm} \sum_{i=1}^{m} \Xi_i [Q(X + g\Xi_i) - Q(X - g\Xi_i)], \tag{2.17}$$

where the Ξ_i are independent unit random vectors.

As is easy to see, this estimate is based on the statistical gradient method. Its statistical properties are obvious:

$$\begin{aligned} M(\widehat{\text{grad}}_{rg} Q) &= \text{grad } Q, \\ D(\widehat{\text{grad}}_{rg} Q) &= \frac{1}{m} \left[2k^2 \frac{n-1}{n+2} + \frac{n^2\sigma^2}{2g^2} \right]. \end{aligned} \tag{2.18}$$

Hence it is easy to see that with increasing number of samples (i.e., with $m \to \infty$) the estimate tends to the exact value of the gradient.

With a number of random samples equal to

$$m^* = \frac{4g^2k^2}{n\sigma^2} \frac{n-1}{n+2} + n, \tag{2.19}$$

the statistical properties (i.e., the first two moments) of a random estimate and of a gradient estimate will be equal. As we can see, we always have

$$m^* > n \tag{2.20}$$

and they coincide only for $n \to \infty$. This signifies that with a fixed number of samples $m = n$ the gradient estimate (2.3) will be the more exact. But if $m > m^*$, it is more convenient to use the random estimate (2.17), just as in the case of $m < n$, when the gradient estimate (2.3) is simply unfeasible.

The foregoing prompts us to refer to the method of orthogonalized statistical gradient [2], which differs from (2.17) by the fact that $m \leqslant n$ and the samples are orthogonal, i.e.,

$$[\Xi_i, \Xi_j] = \begin{cases} 1 \text{ for } i \neq j, \\ 0 \text{ for } i \neq j. \end{cases} \tag{2.21}$$

In this case we can expect a further decrease in the variance of the estimate (2.17).

Summing up this section, we can say that by solving the problem of multiparametric optimization under conditions of noise, random search makes it possible to estimate faster (though less exact) the gradient of the performance index as compared to the use of the gradient method. For increasing the estimation accuracy it is necessary to increase the number of samples, or to orthogonalize the random samples. However, usually for a multistep optimization process the profit gained by reducing the number of samples is much more important than a larger variance of the final result. This constitutes a certain advantage of random search.

3. Identification

The identification of plants with an assigned structure is based on the use of the well-known method of self-adaptive models, where the control device is a multichannel optimizer. However for a sufficiently large class of plants it is possible to determine the gradient of the residual, and hence construct a controller that operates in accordance with Eq. (1.2). Figure 3 shows the mutual relations between plant, model and controller during the identification process.

For definiteness let us consider the case that the plant is an $(n \times 1)$-network [3], i.e.,

$$\begin{aligned} A(t) &= [a_1(t), \ldots, a_n(t)], \\ \mathrm{E} &= (\varepsilon_1, \ldots, \varepsilon_n), \\ B = b &= [A, \mathrm{E}] + \varepsilon(\sigma), \\ B' = b' &= [A, X], \\ Q &= (b' - b)^2, \end{aligned} \tag{3.1}$$

where E is the vector of unknown parameters and $\epsilon(\sigma)$ is random noise at the plant output. This noise is usually a normal discrete random process with zero mean and a variance σ^2.

Let us consider the dynamics of the identification process in this case. By substituting (3.1) into (1.2), we obtain the equation for this process in the continuous version:

$$\frac{dX}{dt} = -2aA(b' - b). \tag{3.2}$$

(The discrete version can be analyzed in the same way).

For analyzing the behavior of the identification system we shall introduce a field $Q^0(X)$ with a gradient

$$\operatorname{grad} Q^0(X) = X - \mathrm{E}. \tag{3.3}$$

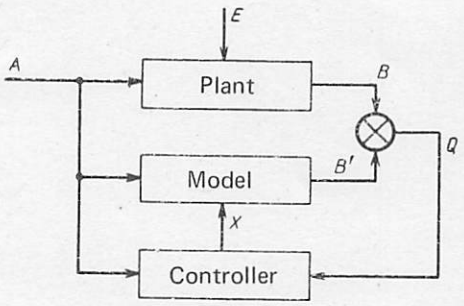

Fig. 3.

This field is central, i.e., it has a center at the point $X^* = E$. Identification involves finding the center of this field for which the gradient is equal to zero.

As is easy to see, the right-hand side of equation (3.2) is an estimate of the gradient of this field:

$$\widehat{\operatorname{grad}}_r Q^0 = \frac{A}{|A|^2}\{[A, \operatorname{grad} Q^0] - \varepsilon\}, \tag{3.4}$$

taken in the direction of the vector A. Indeed, since for sufficiently small absolute values of $|\Xi|$ we have the expression

$$Q(X+g\Xi) - Q(X-g\Xi) \approx 2g[\Xi, \operatorname{grad} Q(X)], \tag{3.5}$$

the formula (2.4) will yield

$$\widehat{\operatorname{grad}}_r Q \approx \frac{1}{2g}\Xi\,[\Xi, \operatorname{grad} Q]. \tag{3.6}$$

Hence if the vector A is random, the estimate (3.4) will also be random. This signifies that in the identification process under consideration we have in fact random search for the minimum of a performance function of the form

$$Q^0(X) = |X - \mathrm{E}|^2/2, \tag{3.7}$$

whose derivative is obtained not by numerical differentiation, but estimated by the formula

$$\widehat{\operatorname{grad}}_r Q^0(X) = \frac{A}{|A|^2}(b' - b), \tag{3.8}$$

which enables us to avoid the search.

The identification process is strongly dependent on the statistical properties of the random vector $A(t)$. Let us consider some cases. Let $A(t)$ be a random independent process of the form

$$A(t) = \Xi_t, \quad t = 1, 2, \ldots, \tag{3.9}$$

where the Ξ_t are unit random vectors. Then the identification process will take place in accordance with the method of random descent whose convergence has been studied in detail in [1] and [3]. (Let us note that although in [3] this process is considered not from the viewpoint of random search, the obtained results are in full agreement with [1], where the convergence of random search is studied).

In the space of the parameters to be identified the identification process is a broken line formed by random directions Ξ_t; in the absence of noise this line does not move away from the objective $X^* = E$, i.e., on the average it gets closer to it. Let us show this.

It suffices to consider one section of this broken line, when $\Xi = \text{const}$. In this case

$$\frac{dQ^0}{dt} = \left[\operatorname{grad} Q^0, \frac{dX}{dt}\right] = a\,[\operatorname{grad} Q^0, \widehat{\operatorname{grad}}_r Q^0]. \tag{3.10}$$

By substituting (3.4) into this formula and using the relation $\sigma = 0$, we obtain

$$\frac{dQ^0}{dt} = -\frac{a}{|A|^2}[A, \operatorname{grad} Q^0]^2 \leqslant 0, \tag{3.11}$$

which signifies that Q^0 decreases monotonically, i.e., convergence of the identification process. If $A(t)$ is a continuous random process, then under these same conditions the

identification trajectory $X(t)$ will be a curve that does not move away from the objective, but monotonically approaches it if $A(t)$ is sufficiently diverse.

In the presence of noise the identification trajectory $X(t)$ can move away from the objective and in the limit not converge to it, but perform a random walk in a region that depends on the noise variance. For convergence it suffices to require that the coefficient a decrease with time, i.e.,

$$\lim_{t\to\infty} a(t) = 0, \tag{3.12}$$

at a rate not smaller than $1/t$ [4]. The conditions of convergence for a discrete process under noise conditions were studied in detail in [3].

Thus in the case of a random input $A(t)$ the identification process reduces to random search for the minimum of a given performance index.

4. Perception Training

For definiteness let us consider the training of a linear perceptron that realizes a linear decision rule

$$b = \operatorname{sign}[(A, X) - \eta], \tag{4.1}$$

which refers the pattern A to one of the two classes $b = +1$. Here A is the pattern vector, X the vector of perceptron parameters, and η a threshold.

During the training process we apply to the perceptron input different patterns of the training sequence

$$A_1, A_2, \ldots, A_k, A_{k+1}, \ldots, A_p, \tag{4.2}$$

for which membership in different classes is defined:

$$b(A_i) = \begin{cases} +1 & \text{for } i = 1, \ldots, k \\ -1 & \text{for } i = k+1, \ldots, p. \end{cases} \tag{4.3}$$

The perceptron output $b(X, A_i)$ (Fig. 4) is compared with a given value b_i, and if they differ we change the parameter vector X. By formulating this problem as an extremal problem, we can say that here we are minimizing a residual of the form

$$Q(X, A_i, b_i) = [(A_i, X) - \eta] \{b_i - \operatorname{sign}[(A_i, X) - \eta]\}, \tag{4.4}$$

which is linear for $b_i \neq b(A_i)$ and equal to zero for $b_i = b(A_i)$.

Now we can write down the adaptation equation

$$\frac{dX}{dt} = \pm a \operatorname{grad}_X Q(X, A_i, b_i) = \begin{cases} \pm aA_i & \text{for } b_i \neq b(A_i), \\ 0 & \text{for } b_i = b(A_i), \end{cases} \tag{4.5}$$

where the plus sign is taken for $b_i > 0$, and the minus sign for $b_i < 0$. This is the adaptation algorithm.

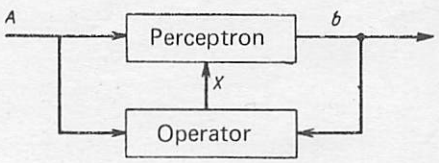

Fig. 4.

Let us examine what takes place in the meantime in the parameter space $\{X\}$. Each pattern A_i partitions the space $\{X\}$ of training parameters with the aid of the hyperplane

$$[A_i, X]=\eta \tag{4.6}$$

into two parts, in one of which the derivative (4.5) is equal to zero. Our problem consists in determining a value $X = X^*$ such that $Q(X^*, A_i, b_i) = 0$ for any $i = 1, \ldots, p$.

For any point of the perceptron parameter space $\{X\}$ the adaptation algorithm (4.6) specifies a motion that depends on the displayed pattern A_i (Fig. 5). Hence the character of variation of X is entirely determined by the strategy of the displays, i.e., by the sequence of patterns introduced into the perceptron during training.

Figure 5 shows an example of a space partitioned with the aid of four hyperplanes corresponding to four displays $A_1, \ldots, A_4$. Here the arrows denote the vectors of the rates of parameter variation for different patterns ($\dot{X}_i$ corresponds to A_i), and the hatched area is the region of sought parameters X^* that satisfy our problem. As we can see, in each region there exist several feasible directions of the adaptation process whose number is equal to the number of elements of the training sequence (the trajectory of the process for the initial points selected by us is shown in Fig. 5 by dashed lines).

If the strategy of displays during the training process is random, i.e., if the patterns A_i are displaying by sampling, then the trajectory $X(t)$ will also be random, which is characteristic for random search.

Thus perceptron adaptation in the case of random displays is in fact realized by the random search method. Indeed, in each region of parameter space we have feasible directions of adaptation. The selection of one of them is carried out by introducing the corresponding pattern into the perceptron. As a rule, the operator has no a priori knowledge how to select the patterns; therefore he must use sampling. This is the reason for randomness in random search for the state X^*.

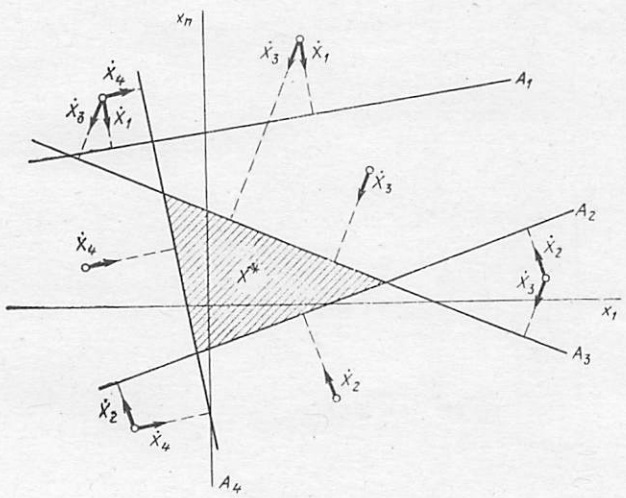

Fig. 5.

5. Training of Converters

Let us consider a more complex case of training, when the output B of the plant to be adapted is a vector that can assume a countable set of values. The plant has a controllable input X (Fig. 6) on which the controller acts. The problem is formulated as follows. It is necessary to determine a controller X such that the plant output B has the smallest deviation from a given value B' assigned by the simulator. For this purpose the same disturbance $A(t)$ is applied to both the plant and the simulator. The plant output B is compared with the simulator output B', and we form the residual

$$Q(X, A, \mathrm{E}) = \Phi(B - B'), \tag{5.1}$$

where Φ is a monotonic scalar function of a vector argument that is minimized only for $B = B'$. Thus our problem involves the minimization of Q for all inputs $A(t)$.

The algorithm of operation of the controller effecting the training is obvious:

$$\frac{dX}{dt} = -\alpha \operatorname{grad}_X Q(X, A, \mathrm{E}). \tag{5.2}$$

Let $A(t)$ and $E(t)$ be random functions. Then the mean of the residual will be

$$M_{AE}[Q(X, A, \mathrm{E})] = \bar{Q}(X). \tag{5.3}$$

The minimum of this quantity optimizes the control X^*:

$$\bar{Q}(X^*) = \min_{X \in G} \bar{Q}(X). \tag{5.4}$$

However it is very difficult to determine the function $\bar{Q}(X)$, so that one must use its statistical estimate (5.1):

$$\hat{\bar{Q}}(X) = Q(X, A, \mathrm{E}). \tag{5.5}$$

Let us consider the dynamics of the training process in the control parameter space $\{X\}$. At the i-th stage let $A(t_i) = A_i$ and $E(t_i) = E_i$. Then the process will tend at this stage to the state X specified by the equation

$$\operatorname{grad}_X Q(X, A_i, \mathrm{E}_i) = 0. \tag{5.6}$$

The set $\{X^*\}_i$ of solutions of this equation contains also the optimal solution X^*, i.e.,

$$X^* \in \{X^*\}_i. \tag{5.7}$$

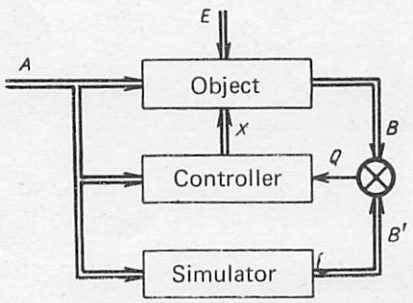

Fig. 6.

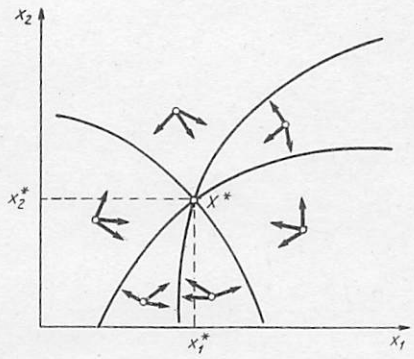

Fig. 7.

The geometrical interpretation of this set in parameter space is in the form of the bottom of a trough (Fig. 7, where each line represents the bottom of the corresponding trough). There are as many such troughs as there are situations characterized by the pair A_i, E_i.

The adaptation process involves moving towards the bottom of the trough specified by a given situation. The spectrum of feasible directions of adaptation is entirely dependent on the number of troughs (in Fig. 7 this spectrum is indicated for three troughs). But since the situation A_i, E_i is random, it follows that the direction of motion in parameter space is also random. This signifies that the process of training realizes an algorithm of random search in which the element of randomness is introduced as a result of the random character of the functions $A(t)$ and $E(t)$.

Conclusion

The above analysis shows that the most typical adaptation processes (optimization, identification and training) are in fact realizing random algorithms of search for optimal plant parameters. The randomness is introduced into the search process either as a result of the measurement errors (in the case of optimization), or due to random disturbances of the plant (in the case of identification and training), or by the operator himself (as in the case of perceptron training). Therefore these adaptation processes can be regarded as different aspects of random search in which often the randomness is introduced not from outside, but "generated" by the plant itself, or it occurs as a result of the peculiar features of operation of the plant.

References

[1] L. A. Rastrigin, *Statistical Methods of Search*, Nauka Press, 1968.
[2] Ye. G. Nikolayev, "The random *m*-gradient method," *Avtomatika i Vychislitel'naya Tekhnika*, Riga, no. 1, 1969.
[3] N. S. Raybman and V. M. Chadeyev, *Adaptive Methods in Control Systems*. Sovetskoye Radio, 1966.
[4] Ya. Z. Tsypkin, *Adaptation and Training in Control Systems*. Nauka Press, 1968.
[5] L. A. Rastrigin, *Random Search with Linear Strategy*. Zinatne, Riga, 1971.