TWO MODELS FOR ANALYZING THE DYNAMICS OF ADAPTATION ALGORITHMS

D. P. Derevitskii and A. L. Fradkov

UDC 62-50

An approach is proposed to the analysis of the dynamics of adaptation algorithms, based on the construction of approximate models of the algorithms. It is shown that, as a model of an adaptation algorithm described by a stochastic difference equation, one can utilize an ordinary differential equation (the deterministic model) or a stochastic differential equation (the stochastic model). The models described herein are applied to the investigation of the dynamics of concrete adaptation algorithms: a linear algorithm, a sign-dependent algorithm, a perceptron type of algorithm, as well as a random search algorithm.

1. Introduction

In this paper we consider adaptation algorithms described by stochastic difference equations of the form

$$\mathbf{c}_k = \mathbf{c}_{k-1} - \gamma_k \mathbf{f}(\mathbf{x}_k, \mathbf{c}_{k-1}), \qquad (k=1, 2, \ldots), \tag{1}$$

where c_k is the m-dimensional state vector of the adaptive system on the k-th step of the adaptation process; $\mathbf{x}_k$ is the random vector of external stimuli on the k-th step; $\gamma_k > 0$ ($k = 1, 2, \ldots$) are specified numbers determining the magnitude of the steps of the algorithm. Vector function $\mathbf{f}(\mathbf{x}, \mathbf{c})$ may be either deterministic or random.

It is of some interest to investigate the dynamics of the adaptation process and, in particular, to determine the speed of convergence of each algorithm. Of practical importance here are the estimates of the so-called practical speed of convergence, i.e., the mean number of steps required to achieve functional quality of a specified level. However, a direct investigation of the behavior of the solution of stochastic difference Eq. (1) leads to difficulties, while estimates of speed of convergence are far from being known for all algorithms.

In the present paper we propose a general approach to the analysis of the dynamics of algorithms of the form of (1), this approach consisting in the replacement of Eq. (1) by a "model" related to it. This model must be simpler to investigate and, at the same time, must remain close, in some sense or other, to the original Eq. (1).

2. The Deterministic Model

Consider the ordinary differential equation

$$dc/dt = -\mathbf{A}(\mathbf{c}), \tag{2}$$

where vector function $\mathbf{A}(\mathbf{c})$ is obtained by averaging $\mathbf{f}(\mathbf{x}, \mathbf{c})$ in Eq. (1) over all possible input stimuli. It turns out that, under definite conditions, based on the requirement that steps γ_k be small, the vectors $\mathbf{c}_k$ will, with high probability, be close to the solution $\mathbf{c}(t_k)$ of Eq. (2), where $t_k = \sum_{i=1}^{k} \gamma_i$.

Let us render the foregoing more specific.

Let $\mathbf{A}(\mathbf{c}) = M_x \mathbf{f}(\mathbf{x}, \mathbf{c})$, $\mathbf{h}(\mathbf{x}, \mathbf{c}) = \mathbf{f}(\mathbf{x}, \mathbf{c}) - \mathbf{A}(\mathbf{c})$,

Leningrad. Translated from Avtomatika i Telemekhanika, No. 1, pp. 67-75, January, 1974. Original article submitted March 2, 1973.

$$B(\mathbf{c}) = M_x\{\mathbf{h}(\mathbf{x}, \mathbf{c})\mathbf{h}^T(\mathbf{x}, \mathbf{c})\},\ b(\mathbf{c}) = \operatorname{Sp} B(\mathbf{c}) = M_x\|\mathbf{h}(\mathbf{x}, \mathbf{c})\|^2.$$

Here and henceforth, Sp A denotes the trace of matrix A, while A^T denotes the transpose of matrix A, and I_m is the $m \times m$ unit matrix. The symbol R^m denotes the m-dimensional Euclidean space of vectors $\mathbf{c} = (c^{(1)}, \ldots, c^{(m)})$ with the norm

$$\|\mathbf{c}\| = \sqrt{[c^{(1)}]^2 +, \ldots, + [c^{(m)}]^2}.$$

Theorem 1. Assume that, for some $L_1, L_2 > 0$, and $\mathbf{c}, \mathbf{c}' \in R^m$, the following conditions are met

$$\|\mathbf{A}(\mathbf{c}) - \mathbf{A}(\mathbf{c}')\| \leqslant L_1\|\mathbf{c} - \mathbf{c}'\|, \tag{3}$$

$$b(\mathbf{c}) \leqslant L_2(1 + \|\mathbf{c}\|^2) \tag{4}$$

while the vectors x_k $(k = 1, 2, \ldots)$ are independent.

Then, for any $\gamma > 0$ and any integer $N > 0$ with $0 \leq \gamma_k \leq \gamma$ $(k = 1, \ldots, N)$, the following inequality is true

$$M \max_{1 \leqslant k \leqslant N} \|\mathbf{c}_k - \mathbf{c}(t_k)\|^2 \leqslant K_1\gamma. \tag{5}$$

Here, $t_k = \sum_{i=1}^{k} \gamma_i$, while $\mathbf{c}(t)$ is the solution to differential Eq. (2) with the initial condition $\mathbf{c}(0) = \mathbf{c}_0$. The quantity K_1 in (5) depends only on L_1, L_2, $\|\mathbf{c}_0\|$, t_N, m.

The proof of Theorem 1 is provided in Appendix 1.*

Corollary 1. Under the conditions of Theorem 1, we obtain from (5) and the Chebyshev inequality, an estimate of the probability that the solution to Eq. (1) goes beyond an ε cylinder about the solution to Eq. (2)

$$P\{\max_{1 \leqslant k \leqslant N} \|\mathbf{c}_k - \mathbf{c}(t_k)\| \geqslant \varepsilon\} \leqslant \beta \text{ for } \gamma \leqslant \frac{\beta\varepsilon^2}{K_1}. \tag{6}$$

Theorem 1 provides an estimate of the closeness of the vectors $\mathbf{c}_k$ and $\mathbf{c}(t_k)$ only on a finite interval of time $(0 \leq t_k \leq t < \infty)$. In case Eq. (2) is stable it is possible to obtain estimates which are uniform over $0 \leq t_k < \infty$.

Theorem 2. Let the conditions of Theorem 1 hold and, in addition, for some vector $\mathbf{c}_* \in R^m$, matrix $H = H^T > 0$, and number $\delta > 0$, let the following condition† hold

$$\mathbf{A}(\mathbf{c})^T - H(\mathbf{c} - \mathbf{c}_*) \geqslant \delta\|\mathbf{c} - \mathbf{c}_*\|^2. \tag{7}$$

Then, the following inequalities will hold for some $\gamma_0 > 0$, $K_2 > 0$, and $\alpha > 0$, when $0 \leq \gamma_k \leq \gamma < \gamma_0$

$$M\|\mathbf{c}_k - \mathbf{c}(t_k)\|^2 \leqslant K_2\gamma^\alpha \quad (k = 1, 2, \ldots), \tag{8}$$

with the quantity K_2 depending only on L_1, L_2, H, δ, m, $\|\mathbf{c}_0 - \mathbf{c}_*\|$.

The proof of Theorem 2 is provided in Appendix 2.

Remark 1. If function $\mathbf{A}(\mathbf{c})$ is linear, condition (7) is equivalent to asymptotic stability in the large of Eq. (2). In the general case, asymptotic stability in the large of Eq. (2) follows from (7).

Remark 2. In many algorithms of the form of (1), the right side is, for each x, a function of the mismatch $\mathbf{v}_k = \mathbf{c}_k - \mathbf{c}_*$. Then, instead of (2), it is natural to consider the corresponding equation for the mismatch

$$d\mathbf{v} / dt = -\mathbf{A}(\mathbf{v}). \tag{9}$$

With this, the appropriate obvious changes must be made in the formulation of Theorems 1 and 2.

* The proof of Theorem 1 used ideas of S. N. Bernshtein [1] (in [1] it was proven for the scalar case of the assertion close to Corollary 1). In [2], it was suggested that, as the model of Eq. (1), one take the "averaged" difference equation $c_k = c_{k-1} - \gamma_k A(c_{k-1})$. An assertion close to Theorem 1 of [2] was obtained as an intermediate result in proving Theorem 1 (cf., Appendix 1).

† The notation $H > 0$ means that matrix H is positive definite, i.e., that $c^T H c > 0$ when $c \neq 0$.

3. The Stochastic Model

The results of section 2 show that the deterministic model of (2) is a good approximation to Eq. (1) only in the case when the variance of the increment* of vectors c_k is small, i.e., when it is possible to neglect the influence of the random contributions $\gamma_k \mathbf{h}(\mathbf{x}_k, c_{k-1})$. We now construct a stochastic model of process (1) which takes account of the role of the random contributions. This model will be comprised of a stochastic differential equation whose solution is a Markov diffusion process. We present the original Eq. (1) in the form

$$\mathbf{c}_k=\mathbf{c}_{k-1}-\gamma_k\mathbf{A}(\mathbf{c}_{k-1})-\sqrt{\gamma_k}[\sqrt{\gamma_k'}\mathbf{h}(\mathbf{x}_k,\mathbf{c}_{k-1})] \tag{10}$$

and we let $\gamma_k \to 0$, while fixing the quantity $t_k=\sum_{i=1}^{k}\gamma_i$ and keeping the quantities γ_k' unchanged, equal to the steps of the original algorithm. It is known [3] that the vector c_k will, in this case, converge by probability distribution to the solution $\tilde{c}(t_k)$ of the stochastic differential equation

$$d\tilde{c}(t)=-A(\tilde{c}(t))dt+\sqrt{\gamma(t)B(\tilde{c}(t))}\,dW(t), \tag{11}$$

where $\gamma(t) = \gamma_k'$ when $t \in [t_{k-1}, t_k)$ and $W(t)$ is an m-dimensional Wiener process with unknown components. The solution to Eq. (11) can serve as an approximation to the solution of Eq. (1) in the sense that the conditional first and second moments of the vectors c_k-c_{k-1} and $\tilde{c}(t_k)-\tilde{c}(t_{k-1})$ coincide to within $o(\gamma_k)$. We note that the vectors c_k-c_{k-1} and $c(t_k)-c(t_{k-1})$, where $c(t)$ is the solution to (2), coincide only for their first moments, i.e., the mathematical expectations. We formulate as a theorem what we have just said.

Theorem 3. Let the functions $A(c)$ and $h(x, c)$ be triply differentiable and also satisfy a Lipschitz condition with respect to c. Let $0 \le \gamma_k \le \gamma$ and the quantity $t_k=\sum_{i=1}^{k}\gamma_i$ be fixed. Then, the solution c_k to stochastic difference Eq. (10), for any distribution $P_0(c)$ of initial vector c_0, converges by probability distribution as $\gamma \to 0$ to the solution $\tilde{c}(t_k)$ of stochastic differential Eq. (11).

Process $\tilde{c}(t)$ satisfies the relationships

$$M\{\tilde{c}(t_k)-\tilde{c}(t_{k-1})/\tilde{c}(t_{k-1})\}=-\gamma_k A(\tilde{c}(t_{k-1}))+o(\gamma_k), \tag{12}$$

$$M\{[\tilde{c}(t_k)-\tilde{c}(t_{k-1})][\tilde{c}(t_k)-\tilde{c}(t_{k-1})]^T\tilde{c}(t_{k-1})\}=\gamma_k\gamma_k'B(\tilde{c}(t_{k-1}))-o(\gamma_k). \tag{13}$$

The proof of Theorem 3 is analogous to that given in [3, 4] and, due to its great cumbersomeness, we do not adduce it here.

Remark 1. The probability density $p(c, t)$ of vector $\tilde{c}(t)$ satisfies the Kolmogorov equation [4] with the initial condition $p(c, 0) = p_0(c)$.

Remark 2. In [5], stochastic differential Eq. (11) was used in the analysis of the simplest algorithm for random search of the extremum of a function of one variable.

It follows from Theorem 2 that, as an approximation to vector c_k, one could choose $\tilde{c}(t_k)$, where $\tilde{c}(t)$ is a diffusion process with flux vector $-A(c)$ and diffusion matrix $\gamma_k B(c)$ with $t_{k-1} \le t < t_k$.

It is of interest to estimate the probabilistic characteristics of process $\tilde{c}(t)$ and, in particular, its first and second moments. It can be shown [6] that the quantities $\mu(t) = M\tilde{c}(t)$ and $\rho(t) = M\|\tilde{c}(t)-c\|^2$, where $c \in R^m$, satisfy the relationships

$$d\mu(t)/dt=-MA(\tilde{c}(t)), \tag{14}$$

$$d\rho(t)/dt=-2M[A(\tilde{c}(t))]^T[\tilde{c}(t)-c]+\gamma(t)Mb(\tilde{c}(t)). \tag{15}$$

4. Application of the Models to the Analysis of the Dynamics of Certain Adaptation Algorithms

As an example we shall now consider the problem of the adaptive reconstruction of an unknown function $y(x) = c_*^T\varphi(x) + \eta$, where c_* and $\varphi(x)$ are m-dimensional vectors, η is random noise ($M\eta = 0$, $M\eta^2 = \sigma^2$) with respect

---

* By the variance of the increment of c_k we understand the quantity $M_x\|\gamma_k \mathbf{h}(\mathbf{x}_k, c_k)\|^2 = \gamma_k^2 b(c_k)$.

to its values at the randomly observed points x_k ($k = 1, 2, \ldots$). We shall assume vectors x_k and "noise" η to be independent and identically distributed.

Typical examples of algorithms for solving this problem [7, 8] are the linear algorithm

$$c_k = c_{k-1} - \gamma_k [y(x_k) - c_{k-1}^T \varphi(x_k)] \varphi(x_k) \tag{16}$$

and the sign-dependent algorithm

$$c_k = c_{k-1} - \gamma_k \operatorname{sign} [y(x_k) - c_{k-1}^T \varphi(x_k)] \varphi(x_k). \tag{17}$$

In constructing adaptive systems one frequently makes use of perceptron-type algorithms

$$c_k = c_{k-1} - \gamma_k [\operatorname{sign} y(x_k) - \operatorname{sign} c_{k-1}^T \varphi(x_k)] \varphi(x_k) \tag{18}$$

for reconstituting the function sign $y(x)$.

We initially consider the linear algorithm of (16). Remark 2 to Theorem 2 applies to it. Obviously, $A(v) = Rv$, where $R = M\varphi(x)\varphi^T(x)$, i.e., condition (3) holds for the algorithm of (16). Condition (4) will hold if vector $\varphi(x)$ has finite fourth moments (for example, if vector $\varphi(x)$ is almost certainly bounded, or has a normal distribution). Thus, we can use Theorem 1, by virtue of which the mismatch $v_k = c_k - c_*$ converges as $\gamma_k \to 0$ to the solution to Eq. (9), defined by the formula $v(t_k) = \exp\left(-R \sum_{i=1}^{k} \gamma_i\right) v_0$. Let L be the largest, and l the smallest, eigenvalues of matrix R. We have the following estimate for the square of the norm of the mismatch

$$\| v(t_k) \|^2 \leqslant \exp\left(-2l \sum_{i=1}^{k} \gamma_i\right) \| v_0 \|^2. \tag{19}$$

For the application of the stochastic model it is necessary to compute $b(v)$. We shall assume that* $\varphi(x) \in N_m(0, R)$, $R > 0$. It is easy to show that, in this case, the following equation is true

$$M[\varphi(x)\varphi^T(x)]^2 = R \operatorname{Sp} R + 2R^2. \tag{20}$$

By virtue of (20), $b(v) = v^T (R \operatorname{Sp} R + R^2) v + \sigma^2 \operatorname{Sp} R$, i.e., the conditions of Theorem 3 hold. We estimate the mean square mismatch $\rho(t) = M \|\tilde{v}(t_k)\|^2$ by using relationship (15). We have $d\rho(t)/dt \leq -2l\rho(t) + \gamma(m+1) L^2 \rho(t) + m\gamma L\sigma^2$ with $0 \leq \gamma_k \leq \gamma$. Hence

$$M \| v(t_k) \|^2 \leqslant \exp[(-2l + (m+1)\gamma L^2) t_k] \| v_0 \|^2 + \frac{m\gamma\sigma^2 L}{2l - (m+1)\gamma L^2}. \tag{21}$$

In the particular case $R = r^2 I_m$, $\gamma_k \equiv \gamma$, we have an exact expression for the model's square mismatch

$$\| v(t_k) \|^2 = \exp(-2k\gamma r^2) \| v_0 \|^2, \tag{22}$$

$$M \| \tilde{v}(t_k) \|^2 = \exp[-k\gamma r^2 (2 - (m+1)\gamma r^2)] [\| v_0 \|^2 - m\gamma\sigma^2 / (2 - (m+1)\gamma r^2)] + m\gamma\sigma^2 / (2 - (m+1)\gamma r^2). \tag{23}$$

In this case it is also easy to obtain the exact expression for the mean square of the mismatch, $M\|v_k\|^2$, for the original algorithm of (16)

$$M \| v_k \|^2 = (1 + 2\gamma r^2 + (m+2)\gamma^2 r^4)^k \left(\| v_0 \|^2 - \frac{m\gamma\sigma^2}{2 - (m+2)\gamma r^2} \right) + \frac{m\gamma\sigma^2}{2 - (m+2)\gamma r^2}. \tag{24}$$

A comparison of expressions (22)-(24) shows that the following inequality holds for small values of γ

* The expression $z \in N_m(a, R)$, where $a \in R^m$ and R is an $m \times m$ matrix, means that the m-dimensional random vector z is normally distributed, with $Mz = a$, $M(z-a)(z-a)^T = R$.

$$\|v(t_k)\|^2 \leqslant M\|v_k\|^2 \leqslant M\|\tilde{v}(t_k)\|^2, \tag{25}$$

i.e., the model's errors are upper and lower bounds of the error of the initial algorithm. Expressions analogous to (22) and (23) can also be written for another choice of the sequence γ_k, for example, for $\gamma_k = \gamma/k$. In this case, $t_k = \gamma \ln k + \gamma C + \eta_k$, where $C = 0.577, \ldots$, is the Euler constant, and $\eta_k \to 0$ as $k \to \infty$. Expressions (22) and (23) assume the forms

$$\|v(t_k)\|^2 \sim t^{-2\gamma r^2}\|\tilde{v}_0\|^2, \tag{26}$$

$$M\|v(t_k)\|^2 \sim k^{-2\gamma r^2}\|v_0\|^2 + \xi_k, \text{where } \xi_k \leqslant \begin{cases} C_1 k^{-1} \text{ when } 2\gamma r^2 > 1 \\ C_2 k^{-1}\ln k \text{ when } 2\gamma r^2 = 1 \\ C_3 k^{-2\gamma r^2} \text{ when } 2\gamma r^2 < 1. \end{cases} \tag{27}$$

The result in (27) accords with the known estimates of [8]. We note that condition (7) holds for the algorithms of (16) with $H = I_m$ and $\delta = l$. Therefore, Theorem 2 is applicable, from which follows the closeness (for small values of the γ_k) of the trajectories of c_k and $c(t_k)$ for all $k = 1, 2, \ldots$.

We now turn to the sign-dependent algorithm of (17), limiting ourselves to the case $\eta = 0$, $\varphi(x) \in N_m'(0, R)$, $R > 0$. It is easy to show that, with this,

$$M\{\text{sign}[c^T\varphi(x)]\varphi(x)\} = \sqrt{\frac{2}{\pi}}\frac{Rc}{\sqrt{c^TRc}}. \tag{28}$$

It follows, from (28) and from remark 2 to Theorem 2, that, for the algorithm of (17),

$$A(v) = \sqrt{\frac{2}{\pi}}\frac{Rv}{\sqrt{v^TRv}}, \quad b(v) = \text{Sp}\,R - \frac{2}{\pi}\frac{\|Rv\|^2}{v^TRv}, \tag{29}$$

i.e., condition (4) holds while condition (3) is violated in the neighborhood of point $v = 0$. Nonetheless, Theorem 1 can also be applied to this case. Simple computations show that the solution to Eq. (2) for algorithm (17) with initial condition v_0 attains point $v = 0$ in a finite time $\bar{t}$ satisfying the inequality $\|v_0\|\sqrt{\pi/2L} \leq \bar{t} \leq \|v_0\|\sqrt{\pi/2l}$. Thus, for any $t < \bar{t}$ the trajectory $v(s)$, for $s \leq t$, lies outside some neighborhood of point $v = 0$, i.e., in a region where condition (3) holds. It is not difficult to show that, for small values of γ_k, the vectors v_k, for $t_k \leq t < \bar{t}$, with probability close to unity, will also lie in this region, i.e., the assertion of Theorem 1 is true for $t_N < \bar{t}$. In the case when $R = r^2 I_m$, the time $\bar{t}$ can be computed exactly: $\bar{t} = \sqrt{\pi/2}\|v_0\|/r$.

We apply the results we have obtained to the comparison of the several algorithms with respect to their speeds of convergence. We compare, for example, the algorithms of (16) and (17), making use of the deterministic model and assuming, for simplicity, that $\eta = 0$, $R = r^2 I_m$, $\gamma_k \equiv \gamma$. Let k_ε be the number of steps required to attain the specified accuracy, i.e., the inequality $\|v(t_k)\| < \varepsilon$ holds. We have

$$k_{\varepsilon\,\text{lin}} = \frac{1}{\gamma r^2}\ln\frac{\|v_0\|}{\varepsilon}, \quad k_{\varepsilon\,\text{si}} = \sqrt{\frac{\pi}{2}}\frac{\|v_0\| - \varepsilon}{\gamma r}. \tag{30}$$

Thus, $k_{\varepsilon_{\text{lin}}}$ depends logarithmically on the quantity $\|v_0\|$ while the dependence of $k_{\varepsilon_{\text{si}}}$ is linear, so that, for large initial errors $\|v_0\|$, the linear algorithm converges more rapidly. It follows from (30) that, for sufficiently large ε, namely, for $\varepsilon > \sqrt{2}/(r\sqrt{\pi})$, whatever the initial mismatch it is preferable to use the linear algorithm. Otherwise, i.e., if ε is small, it is preferable to use the linear algorithm only for sufficiently large values of $\|v_0\|$, namely, when the inequality $\|v_0\| \geq p$ holds. From (30) we derive the asymptotic expression for the quantity p for small values of ε: $p = r\sqrt{\pi/2}\,[\ln(1/\varepsilon) + \ln\ln(1/\varepsilon) + \ln(r\sqrt{\pi/2})] + \beta_\varepsilon$, where $\beta_\varepsilon \to 0$ as $\varepsilon \to 0$. By virtue of Theorem 1, the conclusions we have arrived at can also be extended to the original algorithms of (16) and (17) for small values of the step γ.

We now consider the perceptron-type algorithm of (18), assuming for simplicity that the distribution of vector $\varphi(x)$ is spherically symmetric, and that noise lacks, i.e., that $\eta = 0$.

In this case Eq. (2) will have the form

$$dc/dt=\varkappa(\mathbf{c}/\|\mathbf{c}\|-\mathbf{c}_*/\|\mathbf{c}_*\|), \tag{31}$$

where the constant $\varkappa$ depends on the form of the distribution of $\varphi(\mathbf{x})$.* The stationary solutions to (31) have the form $\lambda \mathbf{c}_*$, $\lambda > 0$. Any solution to (31) lies in the two-dimensional plane defined by the vectors $\mathbf{c}(0)$ and $\mathbf{c}_*$.

We now introduce a measure of accuracy of approximation

$$w(\mathbf{c})=\|\mathbf{c}/\|\mathbf{c}\|-\mathbf{c}_*/\|\mathbf{c}_*\|\;\|^2=4\sin^2(\theta(\mathbf{c})/2),$$

where $\theta(\mathbf{c})$ is the angle between the vectors $\mathbf{c}$ and $\mathbf{c}_*$.

Theorem 4. Let $\mathbf{c}_0 \neq \lambda \mathbf{c}_*$, $\lambda \leq 0$. Then, for any $t \geq 0$, the following inequality is valid

$$w(\mathbf{c}(t))\leqslant w(\mathbf{c}_0)\exp(-\varkappa t/2\|\mathbf{c}_0\|), \tag{32}$$

where $\mathbf{c}(t)$ is the solution to (31) for the initial condition $\mathbf{c}(0) = \mathbf{c}_0$. The proof of Theorem 4 is provided in Appendix 3.

Analogously to the case of the algorithm of (17), it is shown that, in the given case, Theorem 1 is applicable, i.e., formula (32) provides an estimate of the speed of convergence of the initial algorithm of (18) for small γ_k. It is of interest to compare (32) with the estimates obtained by the methods of the theory of finitely-converging algorithms for the solution of inequalities [9]. Let $\gamma_k \equiv \gamma$, let vector $\varphi(\mathbf{x})$ be uniformly distributed on the sphere of radius r, and let k_ε be the number of steps required to attain the specified precision $w(\mathbf{c}_k) < \varepsilon$. It follows from the results of [9] that $k_\varepsilon \leq w(\mathbf{c}_0)/(2\gamma r\sqrt{\varepsilon})$, while, from (32), we have $k_\varepsilon \leq 2\ln(w(\mathbf{c}_0)/\varepsilon)(\varkappa\gamma)$. Thus, taking into account the independence of the input stimuli $\varphi(\mathbf{x}_k)$ $(k = 1, 2, \ldots)$ allows us, for small values of ε, to improve the estimate of practical speed of convergence of the algorithm of (18).

In conclusion, we use the models we have been considering for the analysis of the dynamics of search-type adaptation algorithms with paired probes [10, 11]. This algorithm, in the problem of seeking a minimum of the functional $I(\mathbf{c}) = M_{\mathbf{x}}Q(\mathbf{x}, \mathbf{c})$, where $Q(\mathbf{x}, \mathbf{c})$ is the quality function, has the form

$$\mathbf{c}_k = \mathbf{c}_{k-1} - \frac{\gamma_k}{2\alpha_k}\left[Q(\mathbf{x}_k, \mathbf{c}_{k-1} + \alpha_k \mathbf{z}_k) - Q(\mathbf{x}_k, \mathbf{c}_{k-1} - \alpha_k \mathbf{z}_k)\right]\mathbf{z}_k. \tag{33}$$

Here, $\mathbf{z}_k$ are independent vectors, identically distributed with density $q(\mathbf{z})$ of directions of probe steps, and the α_k are the lengths of the probe steps.

It is known [11, 12] that $(2\alpha)^{-1}M\{[Q(\mathbf{x}_k, \mathbf{c} + \alpha\mathbf{z}_k) - Q(\mathbf{x}_k, \mathbf{c} - \alpha\mathbf{z}_k)]\mathbf{z}_k\} = r^2\nabla\tilde{I}(\mathbf{c})$ when $\mathbf{z}_k \in N_m(0, r^2 I_m)$,

where $\tilde{I}(\mathbf{c}) = \int_{R^m} I(\mathbf{c}+\alpha\mathbf{z})\,q(\mathbf{z})\,d\mathbf{z}$ is the smoothed functional. Equation (7) will have the form $d\mathbf{c}/dt = -r^2\nabla\tilde{I}(\mathbf{c})$ and its solutions will be lines of steepest descent for the smoothed functional $\tilde{I}(\mathbf{c})$.

With the usual requirements of smoothness and boundedness of growth on function $Q(\mathbf{x}, \mathbf{c})$, the conditions of Theorem 1 are met, i.e., for small values of the γ_k the well-known estimates of speed of convergence of gradient methods of minimization [13] are valid. Let us consider in more detail the use of random search for reconstituting an unknown function. With $Q(\mathbf{x}, \mathbf{c}) = [y(\mathbf{x}) - \mathbf{c}^T\varphi(\mathbf{x})]^2$, the algorithm of (33) assumes the form

$$\mathbf{c}_k = \mathbf{c}_{k-1} - \gamma_k\left[y(\mathbf{x}_k) - \mathbf{c}_{k-1}^T\varphi(\mathbf{x}_k)\right]\mathbf{z}_k^T\varphi(\mathbf{x}_k)\,\mathbf{z}_k. \tag{34}$$

If $\mathbf{z}_k \in N_m(0, r^2 I_m)$ and $\varphi(\mathbf{x}_k) \in N_m(0, R)$ we then obtain $A(\mathbf{v}) = R\mathbf{v}$, $b(\mathbf{v}) = (m + 2)r^4\mathbf{v}^T(R\,\mathrm{Sp}\,R + 2R^2)\mathbf{v} - r^4\mathbf{v}^T R^2\mathbf{v} + m(m + 2)\sigma^2 r^4 L$. Thus, for small values of the γ_k the behavior of the algorithm of (34) is close to that of the linear algorithm of (16). For the stochastic model of relationship (15), when $\gamma_k \equiv \gamma$ we obtain the estimate

---

* For example, if $\varphi(\mathbf{x}) \in N_m(0, r^2 I_m)$, then $\varkappa = r\sqrt{2/\pi}$. If, however, $\varphi(\mathbf{x})$ is uniformly distributed on a sphere of radius r then $\varkappa = 2r\Gamma(m-1)/2$ and, as $m \to \infty$, $\varkappa \sim r\sqrt{2/m\pi}$. Here, $\Gamma(z)$ is the gamma function and m is the dimensionality of vector $\varphi(\mathbf{x})$.

$$M\|\tilde{v}(t_k)\|^2 \leqslant \exp[-r^2k\gamma(2l-(m+1)(m+3)\gamma r^2L^2)]\|v_0\|^2 + \frac{m(m+2)\gamma\sigma^2r^2L}{2l-(m+1)(m+3)\gamma r^2L^2}, \quad (35)$$

i.e., the estimated error for the stochastic model is larger (as also in the case of the algorithm of (16)).

APPENDIX 1

We write Eq. (1) in the form $c_k = c_{k-1} - \gamma_k A(c_{k-1}) - \gamma h_k$, where $h_k = h(x_k, c_{k-1})$. We define the sequence of vectors $\{d_k\}_{k=0}^{N}$ by the relationships $d_0 = c_0$, $d_k = d_{k-1} - \gamma_k A(d_{k-1})$ and use the notation $\mu_k = \max_{0\le i\le k}\|c_i - d_i\|^2$, $\nu_k = \max_{0\le i\le k}\|d_i - c(t_i)\|^2$, $k = 0, 1, \ldots, N$. It is clear that $M\max_{0\le i\le k}\|c_k - c(t_k)\|^2 \le 2M\mu_N + 2\nu_N$. The quantity ν_N is the error in the approximate solution to differential Eq. (2) by Euler's method. Standard computations [14] give the following estimate of this error:

$$\nu_N \leqslant \|A(c_0)\|^2 \exp(4L_1t_N)[\exp(L_1\gamma)-1]^2L_1^{-2}. \quad (A.1)$$

For estimating the quantity $M\mu_N$ we use the following lemmas.

Lemma 1. Let $z_1, \ldots, z_N$ be a sequence of m-dimensional random vectors, and $F_0, \ldots, F_N$ an expanding family σ algebras, with z_k being measurable with respect to F_k and $M\{z_k|F_{k-1}\} = 0$, $k = 1, \ldots, N$. Then,

$M\max_{1\leqslant k\leqslant N}\left\|\sum_{i=1}^{k} z_i\right\|^2 \leqslant 4m\sum_{k=1}^{N} M\|z_k\|^2$. The proof of the lemma is analogous to that given in [4].

Lemma 2. Under the conditions of Theorem 1, the following inequalities hold

$$M\|c_k\|^2 \leqslant C_1\exp(2t_k\sqrt{2L_1^2+L_2}), \quad k=0, 1, \ldots, N,$$

where C_1 depends on L_1, L_2, and $\|c_0\|$.

To prove the lemma, we estimate the size of $M\|c_k\|^2$ by virtue of Eq. (1) in terms of $M\|c_{k-1}\|^2$, after which we carry out induction on k.

Lemma 3 [1]. Let the sequence of numbers $\mu_k \ge 0$, $k = 0, \ldots, N$, satisfy the inequalities $\mu_k \le r_1 + r_2\sum_{i=1}^{k}\gamma_i\mu_{i-1}$, $r_1, r_2 \geqslant 0$. Then, $\mu_k \leqslant r_1\exp\left(r_2\sum_{i=1}^{k}\gamma_i\right)$.

Proof of Theorem 1. We apply Lemma 1 to the sequence of random vectors $\gamma_k h_k$, $k = 1, \ldots, N$, having taken as F_k the σ algebra generated by the random vectors $c_0, \ldots, c_k$. From Lemmas 1 and 2 we have

$$M\max_{1\leqslant k\leqslant N}\left\|c_k + \sum_{i=1}^{k}\gamma_iA(c_{i-1})\right\|^2 \leqslant 4m\sum_{k=1}^{N}\gamma_k^2M\|h_k\|^2 \leqslant 4m\gamma t_NL_2[1 + C_1\exp(2t_N\sqrt{2L_1^2+L_2})]. \quad (A.2)$$

We now estimate the quantities $\|c_k - d_k\|^2$, $k = 1, \ldots, N$:

$$\|c_k - d_k\|^2 = \left\|\left[c_k + \sum_{i=1}^{k}\gamma_iA(c_{i-1})\right] - \sum_{i=1}^{k}\gamma_i[A(c_{i-1}) - A(d_{i-1})]\right\|^2 \leqslant$$

$$\leqslant 2\left\|c_k - \sum_{i=1}^{k}\gamma_iA(c_{i-1})\right\|^2 + 2L_1t_k\sum_{i=1}^{k}\gamma_i\|c_{i-1} - d_{i-1}\|^2 \leqslant$$

$$\leqslant 2\max_{1\leqslant k\leqslant N}\left\|c_k + \sum_{i=1}^{k}\gamma_iA(c_{i-1})\right\|^2 + 2L_1t_N\sum_{i=1}^{k}\gamma_i\mu_{i-1}.$$

We turn to mathematical expectation and, using (A.2) and Lemma 3, we obtain

$$M\mu_N \leqslant 8mt_NL_2[1+C_1\exp(2t_N\sqrt{2L_1^2+L_2})]\exp(2L_1^2t_N)\gamma. \quad (A.3)$$

The assertion of the theorem follows from (A.1) and (A.3).

APPENDIX 2

With no loss of generality we can take $c_* = 0$. It follows from (7) that $A(0) = 0$. We denote by $\lambda', \lambda'' > 0$ the smallest and largest eigenvalues of matrix H. In addition, we use the notation

$$(c, c')_H = c^T H c', \quad \|c\|_H^2 = c^T H c, \quad t_k = \sum_{i=1}^{k} \gamma_i.$$

We shall write K = const if the quantity K depends only on L_1, L_2, H, δ, c_0, and m. For the proof of the theorem we need the following lemmas.

Lemma 4. If the numbers $\mu_k \ge 0$ satisfy the inequalities $\mu_k \le (1 + r_1\gamma_k)\mu_{k-1} + r_2\gamma_k$, k = 1, 2, ..., where $r_1, r_2 > 0$, then $\mu_k \le (\mu_0 + r_2/r_1)\exp(r_1 t_k)$.

If, however, $r_1 < 0$ and $0 \le \gamma_k \le \gamma < -1/r_1$, then $\mu_k \le -3r_2/r_1 + \mu_0 \exp(r_1 t_k)$.

The proof of Lemma 4 is carried through by means of standard estimates close to those in [4].

Lemma 5. When $0 \le \gamma_k \le \gamma < \varkappa_1 = 2\delta\lambda'/[\lambda''^2(L_1^2 + L_2)]$, the following inequalities are valid

$$M\|c_k\|_H^2 \leqslant \|c_0\|_H^2 \exp(-\varkappa_2 t_k) + K_3\gamma,$$

$\|d_k\|_H^2 \le \|c_0\|_H^2 \exp(-\varkappa_3 t_k)$, where K_3 = const, $\varkappa_2 = 2\delta/\lambda'' - \gamma(L_1^2 + L_2)\lambda''/\lambda'$, $\varkappa_3 = 2\delta/\lambda'' - \gamma L_1^2 \lambda''/\lambda'$, and the sequence of vectors d_k, k = 0, 1, ..., was introduced in the proof of Theorem 1.

Proof: $M\{\|c_k\|_H^2 | c_{k-1}\} = \|c_{k-1}\|_H^2 - 2\gamma_k(A(c_{k-1}), c_{k-1})_H + \gamma_k^2 M\{\|h(x_k, c_{k-1})\|_H^2 \,|\, c_{k-1}\} \leqslant (1 - 2\gamma_k\delta/\lambda'' + \gamma_k^2\lambda'' L_2/\lambda' + \gamma_k^2 L_1^2\lambda''/\lambda')\|c_{k-1}\|_H^2 + \gamma_k^2 L_2$.

By averaging this inequality and using Lemma 4, we obtain the required estimate for $M\|c_k\|_H^2$. The estimate for $\|d_k\|_H^2$ is obtained analogously.

Lemma 6. If the conditions of Lemma 5 are met, then $M\|c_k - d_k\|^2 \le \gamma \exp(\varkappa_4 t_k) K_4/\varkappa_4$, where $\varkappa_4 = 2L_1 + \gamma L_1^2$, K_4 = const.

Proof. We have $M\{\|c_k - d_k\|^2 | c_{k-1}\} \le (1 + \gamma_k\varkappa_4)\|c_{k-1} - d_{k-1}\|^2 + \gamma_k^2 L_2(1 + \|c_{k-1}\|^2)$. Averaging this inequality, and using Lemmas 5 and 4, completes the proof.

Proof of Theorem 2. Let $\varkappa_5 = 2\delta/\lambda'' - \gamma_0(L_1^2 + L_2)\lambda''/\lambda' > 0$. We select number α on the basis of the condition $1 - \alpha\varkappa_4/\varkappa_5 = \alpha$, i.e., $\alpha = \varkappa_5/(\varkappa_4 + \varkappa_5) > 0$, and we set $t_\gamma = \varkappa_5^{-1}\ln(\|c_0\|_{H^2}/\gamma^\alpha)$. It follows from condition (7) that $\|c(t)\|_{H^2} \le \gamma^\alpha$ when $t \ge t_\gamma$, and it follows from Lemma 5 that $M\|c_k\|_{H^2} \le \gamma^\alpha + K_3\gamma$ when $t_k \le t_\gamma$. Therefore, when $t_k \ge t_\gamma$ the inequality $M\|c_k - c(t_k)\|^2 \le 2M\|c_k\|_{H^2}/\lambda' + 2\|c(t_k)\|_{H^2}/\lambda' \le K_6\gamma^\alpha$ is true, where K_6 = const. If, however, $t_k \le t_\gamma$ then, by using Lemma 6 and Ineq. (A.1), we obtain $M\|c_k - c(t_k)\|^2 \le 2M\|c_k - d_k\|^2 + 2\|d_k - c(t_k)\|^2 \le 2\gamma K_4 \exp(\varkappa_4 t_\gamma)/\varkappa_4 - 2\gamma^2 \exp[4L_1 t_\gamma][\exp(L_1 K_5) - 1]^2/L_1^2 \le K_7\gamma\exp(\varkappa_4 t_\gamma) + K_8[\gamma\exp(\varkappa_4 t_\gamma)]^2 \le K_9\gamma^{1-\alpha\varkappa_4/\varkappa_5} = K_9\gamma^\alpha$, where K_7, K_8, K_9 = const.

Thus, $M\|c_k - c(t_k)\|^2 \le K_2\gamma^\alpha$, k = 1, 2, ..., where $K_2 = \max\{K_6, K_9\}$, q.e.d.

APPENDIX 3

Proof of Theorem 4. We compute the total derivative of function w(c) by virtue of system (31), directing one of the coordinate axes along the vector c_*: $dw(c)\,dt = \nabla w^T(dc/dt) = -2\varkappa\|c\|^{-1}\sin\theta[-\sin\theta(\cos\theta - 1) + \cos\theta\sin\theta] = -\varkappa w(c)/2\|c\|$. But, as we easily convince ourselves, $d(\|c\|)/dt \le 0$, so that $dw(c)/dt \le -\varkappa w(c)/2\|c_0\|$ whence follows the assertion of Theorem 4.

LITERATURE CITED

1. S. N. Bernshtein, "Stochastic finite difference equations and stochastic differential equations," in: Collected Works [in Russian], Vol. 4, Nauka (1964).
2. S. M. Meerkov, "On simplifying the description of slow Markov walks, I, II," Avtomat. i Telemekhan., 33, Nos. 3, 5 (1972).
3. S. N. Bernshtein, "Stochastic differential equations," in: Collected Works [in Russian], Vol. 4, Nauka (1964).

I. I. Gikhman and A. V. Skorokhod, Stochastic Differential Equations [in Russian], Naukova Dumka, Kiev (1968).
A. N. Kabalevskii, "Analysis of search under noisy conditions using the theory of Markov processes," Avtomat. i Telemekhan., 26, No. 11 (1965).
E. B. Dynkin, Markov Processes [in Russian], Fizmatgiz (1963).
Ya. Z. Tsypkin, Adaptation and Learning in Automatic Systems [in Russian], Nauka (1968).
M. A. Aizerman, É. M. Braverman, and L. I. Rozonoér, The Method of Potential Functions in the Theory of Learning Machines [in Russian], Nauka (1970).
V. A. Yakubovich, "Recursive finitely-converging algorithms for solving systems of inequalities," Dokl. Akad. Nauk SSSR, 166, No. 6 (1966).
L. A. Rastrigin, Statistical Methods of Search [in Russian], Nauka (1968).
Ya. Z. Tsypkin, "Smoothed randomized functionals and algorithms in the theory of adaptation and learning," Avtomat. i Telemekhan., 32, No. 8 (1971).
G. E. Antonov and V. Ya. Katkovnik, "Method of synthesizing one class of algorithms for random search," Avtomat. i Telemekhan., 32, No. 6 (1971).
B. T. Polyak, "Gradient methods of minimizing functionals," Zh. Vychisl. Matem. i Matem. Fiz., 3, No. 4 (1963).
I. S. Berezin and N. P. Zhidkov, Computational Methods [in Russian], Vol. 2, Fizmatgiz (1960).