A Method of Optimal Control of a Technical System or Process Using the Neural Network Model and the Thompson Sampling Algorithm

L. S. Chernyshev[a,*]

[a] *OOO Matsoft, Moscow, 115093 Russian Federation*

**e-mail: math4soft@yandex.ru*

Abstract—The paper proposes an approach to solving the problem of optimal management of a technical system or process (control object) in the form of determining a set of values of control parameters on a set time horizon (control horizon) optimizing the target function of the control object on a given horizon. To solve the problem, a method of identifying the model of the control object using an artificial neural network has been developed. For each specified time slice of the required control horizon, the model calculates the predicted values of the object state parameter on the basis of the values of the object control parameter generated for each time slice of the tuples. For each time slice, the tuples of the target function of the object are calculated according to the generated tuples of the control parameters and the forecast calculated by the model on their basis. The solution to the problem of optimal control is the selection in the process of optimization of one control parameter value from the generated tuples for each time slice providing an optimum (maximum or minimum) sum of the values of the target function for all specified time slices of the required control horizon. Optimization of the target function of the object is carried out using a modified Thompson sampling algorithm used in the well-known multiarmed bandit problem. As an example, the problem of building a set of optimal prices for goods, the implementation of which is carried out in the free market, maximizing the profit of the seller for a given management horizon, is solved.

Keywords: neural network, time series, predictive model, Thompson sampling, multiarmed bandit problem, control horizon, control parameters, state parameters, target function, optimization, maximization of profits of the seller in the marketplace

DOI: 10.1134/S1054661823030069

INTRODUCTION

The functioning of any technical system and the evolution of many different processes (hereinafter the control object or simply the object) are associated with a sequential change in the values of the state parameters describing the main characteristics of this control object at a given point in the phase space. For each such point of the phase space, there are a number of parameters that affect the control object and lead to a change in its state parameters, which will be called control parameters.

This work discusses control objects that change state parameters along the time axis. If the control object is subject to the influence of stochastic factors, then in most cases at each time it is impossible to assess the degree of influence of each of them, which makes the task of managing the object in order to obtain a certain result in the future extremely difficult. Let us consider an object with one main parameter of the state Y of dimension m defined on a set of permissible values $D_Y : D_Y \subset R^m$, and one main control parameter U defined on a set of permissible values $D_U : D_U \subset R^n$, where R^m and R^n are arithmetic spaces of dimension m and n, respectively. The influence of the remaining, not directly evaluable, control parameters will be taken into account by using an artificial neural network that models the evolution of the control object at the intervals of the discreteness of the time scale, using only pairs of synchronous historical observations of the dynamics of the main control parameter U and the dynamics of the main parameter of the state of the system Y.

The paper considers the approach to solving the problem of optimizing the sum of the values of the target function $P(Y, U, t)$ on time slices $t \in \{1..T\}$ of the control horizon T, depending on the control parameter U and the predicted values of the object state parameter Y on time slices $t \in \{1..T\}$ of the control horizon T under conditions of a priori uncertainty relative to all factors affecting the object. To do this, at the first stage, the model of the control object is identified using an artificial neural network (ANN), including the choice of the ANN architecture and its training and validation. At the second stage—optimization stage—within the control horizon T, according to a certain sampling algorithm, D_U, one value of the control parameter U is selected from the range of per-

Received February 6, 2023; revised February 6, 2023; accepted February 6, 2023

missible values at the beginning for $t = 1$, then for $t = 2$, and so on, up to $t = T$. The sample (time series) constructed in this way from T values $\{U_t\} = (U_1, U_2 \ldots U_T)$ will be called a sample or a trajectory in the space of possible values of the parameter and the index of the time slice $t \in \{1..T\}$; the sum of all the allowed values of the parameter U_t on the time slice $t \in \{1..T\}$ will be called the tuple of the allowed values U_t on the slice $t \in \{1..T\}$, or simply tuple $\langle U_t \rangle$; the aggregate of all constructed sampling samples (trajectories) $\{U_t\}$ of the control parameter will be called the tuple of trajectories (sample) $\langle\{U_t\}\rangle$; and an incomplete sample representing the vector of the elements $(U_1, U_2, \ldots U_t)$ of dimension t is the vector $\mathbf{U}_t$. Further, on the basis of $\mathbf{U}_t$ as a result of the forecast of the ANN, the values of the system state parameters Y_t are calculated first for $t = 1$, then for $t = 2$, and so on to $t = T$. In this case, in the calculations of the parameters of the state of the object Y_t, as will be shown below, the vectors $\mathbf{Y}_{t-1}$ of the predicted values of the parameters of the state of the object consisting of values calculated for previous time slices are involved: $\mathbf{Y}_{t-1} = (Y_0, Y_1, \ldots Y_{t-1})$. From the values of U_t and Y_t obtained for each $t \in \{1..T\}$, the values of the target function P_t of the control object are calculated, which will also form a time series of $\{P_t\}$ samples. It should be noted that, if the control parameter limits are set for each time slice $U_t \in [U_{\min_t}, U_{\max_t}]$ and the admissible error $\frac{\Delta U_t}{2}$ is given in use of control parameter values U_t, then the maximum possible number Nu_t of different values U_t for each slice $t \in \{1..T\}$ will be equal to $\mathrm{Nu}_t = \frac{U_{\max_t} - U_{\min_t}}{\Delta U_t} + 1$, which is equivalent to setting the sequence (tuple) $\langle U_t \rangle$ values U_t:

$$\langle U_t \rangle = \{U_{\min_t}, U_{\min_t} + \Delta U_t, U_{\min_t} + 2\Delta U_t \ldots U_{\min_t} + (\mathrm{Nu}_t - 2)\Delta U_t, U_{\max_t}\}.$$

To find the optimal set $\{U_t\}_{\text{opt}}$, it is necessary to perform a summarization according to the set $\{P_t\}$ obtained on the basis of $\mathbf{U}_t$ and $\mathbf{Y}_t$ to check for the compliance of this particular set $\{P_t\}$ with the maximum—for the maximization task (or minimum—for the minimization task). The task of optimal object management to maximize the sum of the target function on the control horizon is recorded as

$$\{U_t\}_{\text{opt}} = \underset{\{U_t\}}{\operatorname{Argmax}}(\langle\{P_t\}\rangle) = \underset{\{U_t\}}{\operatorname{Argmax}}\left(\left\langle \sum_{t=0}^{T} P(t, \mathbf{U}_t, \mathbf{Y}_t) \right\rangle\right), \quad (1)$$

where $\mathbf{U}_t$ and $\mathbf{Y}_t$ are vectors of dimension $t : t \in \{1..T\}$, and $\langle\{P_t\}\rangle$ is the tuple of all possible sets (samples) $\{P_t\}$, which can only be built on the basis of the discreteness of the control parameter ΔU_t. Below it will be shown that the number of tuple elements $\langle\{P_t\}\rangle$ will be $\prod_{t=1}^{T} \mathrm{Nu}_t$ or Nu^T if $\mathrm{Nu}_t = \mathrm{Nu}$ for all $t \in \{1..T\}$.

In the work to solve the problem of optimization as a sampling method, both the random selection method and the modified Thompson sampling method, known in application for solving the mutliarmed bandit problem [6, 11, 14], are used. Classical nonlinear optimization methods, such as gradient optimization methods, are not effective for working with the results of forecasting the ANN of the model because of the high "ravineness" of the obtained forecast data. The multiarmed bandit problem is one of the most basic tasks in the science of solutions by the stochastic method for optimal allocation of resources in conditions when it is not possible to apply deterministic optimization methods.

Modifications of the method described in the work are applicable both for the optimal control of complex technical systems, such as gas turbines, oil refining, and systems for generating electrical energy. In this paper, the use of this method is considered by the example of optimal management of the process of selling goods in the free market (trading process) as a solution to the problem of optimal dynamic pricing required by the seller in the marketplace, where the main parameter of managing the trading process is the dynamically changing price of selling goods in the marketplace.

1. PREDICTING THE VALUES OF PARAMETERS OF THE OBJECT STATE

The formulation of the problem of identification of the ANN model of the control object, as well as the method of its solution, is described in detail in works [1–3] based on classical works on the use of the apparatus of neural networks to predict the values of time series [7, 8] and optimal control [5, 9, 10, 12, 13]. In the numerical experiments conducted, the design of a fully connected direct-propagation neural network with one input layer, with the number of neurons $K_y + K_u$, has proven itself well, and on its first K_y neurons, K_y predictions of the parameters of the state of the object $[Y_{t-1} \ldots Y_{t-K_y}]$ are given a step earlier, and on the subsequent K_u neurons, K_u control parameter values at previous time points are given, starting with the current one: $[U_t, \ldots, U_{t-K_u+1}]$. In this case, the activation function of the input layer is one, one hidden layer has G neurons, and the output layer has only one neuron, with activation functions of the hyperbolic tangent form for the hidden and output layers. Then, if the

t index of the time slice of the control horizon T: $t \in \{1..T\}$, then for the predictive model (PM) for the first time slice of the prediction horizon $T(t = 1)$ the forecast function has the following mathematical representation:

$$Y_1(t,U_1) = f\left[\sum_{j=0}^{G}\left[f\left[\sum_{i=0}^{Ky-t}\left(bx_{i,j}\cdot dy\left(L - K_y + t + i\right)\right)\right.\right.\right.$$
$$+\sum_{i=0}^{Ku-t-1}\left(bu_{i,j}\cdot dp\left(L - K_u + t + i + 1\right)\right) \qquad (2)$$
$$\left.\left.\left.+\, bb_j + bu_{Ku,j}\cdot U_1\right]\cdot b2_j\right] + bb2\right],$$

and for each subsequent time slice t ($t = 2..T$), the PM control horizon T has an expanded form:

$$Y_t(t,U_t,U_{t-1},\ldots U_1,Y_{t-1},\ldots Y_1)$$
$$= f\left[\sum_{j=0}^{G}\left[f\left[\sum_{i=0}^{Ky-t}\left(bx_{i,j}\cdot dy\left(L - K_y + t + i\right)\right)\right.\right.\right.$$
$$+\sum_{i=0}^{Ku-t-1}\left(bu_{i,j}\cdot dp\left(L - K_u + t + i + 1\right)\right) + bb_j \qquad (3)$$
$$+\sum_{i=1}^{t-1}\left(bx_{Kx-t+i+1,j}\cdot Y_i\right) + \sum_{i=1}^{t-1}\left(bu_{Ku-t+i,j}\cdot U_i\right)$$
$$\left.\left.\left.+\, bu_{Ku,j}\cdot U_t\right]\cdot b2_j\right] + bb2\right],$$

where for (2) and (3) the f activation function is hyperbolic tangent; the set of PM parameters corresponding to the time slice index t includes in relation to (2) and (3) bx and bu matrices of synaptic transition coefficients to the jth neuron of the hidden layer with neurons of the input layer with numbers 0 to $Ky - 1$ and Ky to $Ky + Ku - 1$, respectively, for bx and bu, defined in the ANN training process using gradient descent method and validated (see [11]); bb is the vector of their displacements; $b2$ is the vector of transition coefficients from hidden to output; $bb2$ is its displacement; $dy(t)$ are values of the training time series of parameters of the state of the object related to the interval of the time series with an index $= L - Ky + t + i$, for $i = 0..Ky - t$; $dp(t)$ are the values of the training time series of control parameters of the object related to the time series interval with index $=$ $L - Ku + t + i + 1$, for $i = 0..Ku - t - 1$; U_t are sets of control parameters corresponding to the time slice t $t \in \{1..T\}$ of the control horizon T, representing sets of parameters of object control generated by the sampling algorithm, having a given discreteness and restriction on values $U_{\min_t} \le U_t \le U_{\max_t}$ and on which the optimization of the target function of the object will be carried out; Y_i is the predicted value of the state parameter calculated using PM for the previous current time slice with index $t > 1$ time slices t_i, $i = (1..t - 1)$ of control horizon T. Thus, if on each time slice t of the control horizon according to a certain algorithm (for example, a uniform distribution on a segment $[U_{\min_t}, U_{\max_t}]$) the values of the control parameters Nu_t are generated: $U_{\min_t} \le U_t \le U_{\max_t}$, the number of all possible predicted values of state parameters Y_t for the time slice t according to (3) will represent a hierarchical structure (tree), the dimension of which grows with growth of t and will correspond to the product $\prod_{i=1}^{t}\mathrm{Nu}_i$. Then, for the case where for each time slice with index t: $t \in \{1..T\}$ the same number of control parameters is generated =Nu, the number of values Y_t calculated for each time slice with index t will increase in proportion to the degree t with the base Nu. The possible variants of the sets $\{P_t\}$ composed of tuples elements $\langle P_t\rangle$ as well as Y_t will represent a hierarchical structure (tree), the dimension of which grows with the growth of t. At the base of the trunk of this tree is the only value of the target function P_1, calculated from the last known value in the past of the training time series of the control parameter with length L: $U_1 = U(L)$ and the last predicted value of the object state parameter Y, calculated only from the known values of K_Y in the past of the training time series of this parameter and the last K_u values of the time series of the control parameter known in the past. The number of possible values of the target function $P_t = F\left(t,\mathbf{Y}_t,\mathbf{U}_t\right)$ (the tuple dimension $\langle P_t\rangle$ for each time slice with the index t would be equal to the product of the dimensions of the tuples $\langle U_t\rangle$ and $\langle Y_t\rangle$ (in the case of the independence of these tuples from each other), i.e., $= \mathrm{Nu}{*}\mathrm{Nu}^t$, but considering that $P_t = F\left(t,\mathbf{Y}_t,\mathbf{U}_t\right)$ and according to (3) $Y_t = f\left(t,\mathbf{Y}_{t-1},\mathbf{U}_t\right)$, the size of the tuple $\langle \mathrm{P}_t\rangle$ on the slice t remains the same as the tuple $\langle Y_t\rangle$ on the slice t, and the tuple dimension of all possible samples $\langle\{P_t\}\rangle$ is equal to the tuple dimension of the last slice $\langle Y_T\rangle = \mathrm{Nu}^T$.

2. APPLICATION OF THE MODIFIED THOMPSON SAMPLING METHOD TO FIND THE OPTIMAL SET OF CONTROL PARAMETERS ON THE CONTROL HORIZON

As a method of optimization for finding the optimal set of control parameters $\langle Uopt_t\rangle$providing the solution of the problem (1), the modified Thompson sampling algorithm described in [4], based on the work [5, 8, 14], is proposed in the work. According to this algorithm, different sets of values $\{U_t\}$ are generated, in each of which one value is selected by the algorithm U_t for each time slice of the control horizon. The obtained sets will be tested for optimality, for which according to (2) for $t = 1$ and according to (3) for $t = 2..T$, the corresponding $\{U_t\}$ sets of values $\{Y_t\}$ are calculated. Then, according to the dependence of the target function $P_t = F(t, Y_t, U_t)$ on the control parameters and state parameters set for this control object, target function sets $\{P_t\}$ are calculated representing trajectories in the space of parameters U, Y, and t.

The range of change in the control parameter of an object $[U_{\min_t}, U_{\max_t}]$ (in the event that it has not been constrained) can be specified synthetically (see, for example, [12]). In this case, the maximum and minimum boundaries are selected from the mean values U_{mean} and from the standard standard standard deviation σ of the existing time series of previously measured values of the control parameter, using the "3 sigma" rule, which means more than 99% probability of covering all possible values of the control parameter in a Gaussian distribution of the random value of its values, i.e.,$U_{\min} = U_{\text{mean}} - 3\boldsymbol{\sigma}$, $U_{\max} = U_{\text{mean}} + 3\boldsymbol{\sigma}$.

The range of change in the control parameter of the object $[U_{\min_t}, U_{\max_t}]$ is divided into a certain number of intervals $\text{Nu}_t$with its boundaries $[U_{\min j,t}, U_{\max_{j,t}}]$, $j = 1..\text{Nu}_t$. The limit on the number of such intervals Nu_t is imposed by the performance of the equipment. If one plans to perform Z sampling operations on the received target function sets, the number Nu_t should be selected as $<Z/100$.

Then, randomly for each time slice t of the control horizon, an index j_t is randomly drawn showing from which range of allowable values $U_{\min j,t} \le U_t \le U_{U_{\max j,t}}$ the control parameters values will be selected. This interval of values is in turn divided into intervals of a certain discreteness and, with accuracy to a given discreteness, the random number generator generates the value of the control parameter within the boundaries of the selected interval. A similar operation is performed for each time slice. A set $\{U_t\}$ of values characterizes a unique trajectory in the space of possible values of the control parameter and time. From them, using (2) and (3), one calculates the sets of values $\{Y_t\}$. In total, the first stage produces $Z1$ of such generations. Then one calculates the sets of the target function $\{P_t\}$, determines the locally optimal trajectory for tuple $\langle\{P_t\}\rangle$ giving the maximum sum for all time slices, and fixes this sum as $S1$.

In the second stage, a modification of the Thompson sampling algorithm is used. To do this, on the first iteration of the sampling ($i = 0$), each of the Nu intervals of the range of possible U values is matched with the same parameters of the beta distribution—α_0 and β_0. Random X has a beta distribution: $X \sim B(\alpha, \beta)$:

$$f_X(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1} (1-x)^{\beta-1},$$
$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1}\, dx, \tag{4}$$

where α, $\beta > 0$ are arbitrary fixed parameters.

Then, for each of$j = 1..\text{Nu}$ intervals for each time slice t of the horizon of the direction from (4), one random value $f_{j,t}$ belonging to the beta distributions of each jth interval is generated. The maximum of the generated values for each time slice will determine the interval number $n_{1,t}$ (index 1 means the first iteration of the sampling), from which the already random one will be selected—according to the law of uniform distribution, from the possible discrete values determined in the previous step, value $U_{\min n_{1,t},t} \le U_{1,t} \le U_{\max n_{1,t},t}$.

The procedure is repeated for each of the time slices in the control horizon. The resulting values $U_{n1,t}$ for each time slice will create a new trace (sample) in the space of possible values of the control parameter and the time slice index t. According to the constructed trajectory according to (2) and (3), trajectories $\{Y_t\}_1$ and then $\{P_t\}_1$ are calculated. We will designate the sum $\{P_t\}_1$ along this trajectory $\{U_{n1,t}\}$ as S_1. It is compared, for example, with $0.5S1$ to determine whether a successful trajectory is obtained or not. If the trajectory is successful, i.e., for example, $S_1 > 0.5S1$, then the parameters of the beta distributions $\alpha_{j,t}$ and $\beta_{j,t}$ intervals of U values participating in this trajectory for each time slice will receive a reward in the form of an adjustment:

$$\hat{\alpha}_{j,t} = \alpha_{j,t} + r, \quad \hat{\beta}_{j,t} = \beta_{j,t} - r, \tag{5}$$

where r is the reward parameter and can be selected, for example, as $r = 1/Z$, where Z is the planned number of trajectories under construction at this stage. In [14], the author proposes to choose the adjustment parameter r on the order of $1/\sqrt{Z\log(Z)}$. The process will be repeated to create samples (trajectories) with numbers $z \in \{1..Z\}$. According to the obtained sample sums $\{P_t\}_z$ compared with $S1$ and upon fulfillment of $S_z > 0.5S1$, the trajectory is considered successful, and the parameters of interval distributions with numbers $n_{z,t} : \alpha_{n_{z,t},t}$ and $\beta_{n_{z,t},t}$ will receive awards according to (5). At the same time, in [4], comparison of S_z instead of comparison with $0.5S1$ can go with a random number f_z selected from the Bernoulli distribution, which was constructed with the participation of sums S_i of trajectories pointed at the previous iteration, i.e., with all i: $i < z$. The intervals participating in the successful trajectory will have a slightly greater advantage over the remaining intervals on the next iteration of the sampling, since the median of their

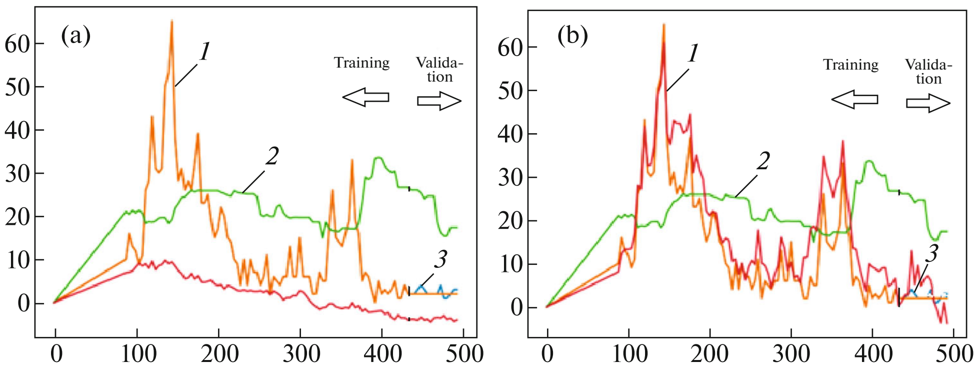

Fig. 1. The result of the INS training process: (a) the initial stage; (b) after 1000 training epochs. Curves: (*1*) dynamics of the training series of the parameter of the state of the control object (number of sales per day), (*3*) dynamics of the validation series of the parameter of the state of the control object (number of sales per day), (*2*) dynamics of the training and validation series of the control parameter (average daily price).

beta distributions will shift a small amount to the right, which will give a slightly higher chance for these intervals to win when comparing the intervals by random values generated by the beta distributions of these intervals in the next round of sampling. Thus, the intervals that participated in the construction of the most effective trajectories will receive greater and greater advantages over the others, which will ensure, with large i numbers of sample iterations, more frequent operation of the most effective ranges of control parameter values for the corresponding time slices. Subsequently, at the next stage, it is possible to discard absolutely inefficient intervals on the corresponding time slices and begin a new stage—the procedure of breaking up only effective intervals, within which their leaders will also be determined. The procedure can be continued until the sampling limit associated with the calculation time limit is reached. The local solution to the optimization problem will be a trajectory (sample) consisting of a set $\{U_t\}$ for which the maximum sum of the set $\{P_t\}$ will be obtained.

3. RESULTS OF APPLICATION OF THE METHOD OF OPTIMAL MANAGEMENT OF THE TRADING PROCESS TO OPTIMIZE THE PROFIT OF THE SELLER IN THE MARKETPLACE

As an example, the problem of dynamic pricing is solved—the construction of a set of optimal prices for goods sold in the marketplace on a given control horizon $T = 36$ days and divided into 9 time slices with a resolution of 4 days. As training series of the ANN model we used the following: a series of the number of daily sales $N(i)$ (as a vector of parameters of the state of the trading process) and a time series of daily price change Price(i) (as a vector of control parameters of the trading process) with a total length of 497 days. The price has an important impact on the number of sales; however, there are many random factors that also affect the studied dependent parameter, the direct impact of which cannot be estimated: for example, a drop in the rating of a product as a result of negative reviews specially ordered by competitors, a sudden glut in the market of this product, etc. The neural network forecast takes into account the impact of these factors in the future. A successful forecast is obtained for prediction horizons on the order of 10% of the length of the training sample.

The construction of the ANN model and its training and validation were carried out according to the methods and algorithms described above (see [2–4]) according to training time series 440 days long and validation series 57 days long.

Figures 1a and 1b show the result of the training procedure of the ANN on the model of the process of sale of goods.

The model resulting from training after 1000 iterations is validated and suitable for use in the subsequent optimization procedure. The synaptic coefficients of this model will be used in formulas (2) and (3) to predict the parameters of the state of the control object, in our case to predict the number of daily sales on time slices of a given control horizon.

The target function of the control object (in our case, the trading process of selling goods in the free market) is the profit function, which was calculated for each of the time slices t on the control horizon consisting of T time slices. For each time slice t of the control horizon T, the control parameters were restored—price—($\text{Price}_{j,t}$)—according to the corresponding deviations in price $U_{j,t}$ generated on the jth interval of discreteness (j = 1..Nu) and state parameters—the number of daily sales $Q_{j,t}$—according to the forecast values of increments received from the ANN, taking into account the reverse recovery. Also, a recovery from the conducted rationings necessary for proper training of the ANN was carried out, which includes multiplication by the norming multiplier Normir

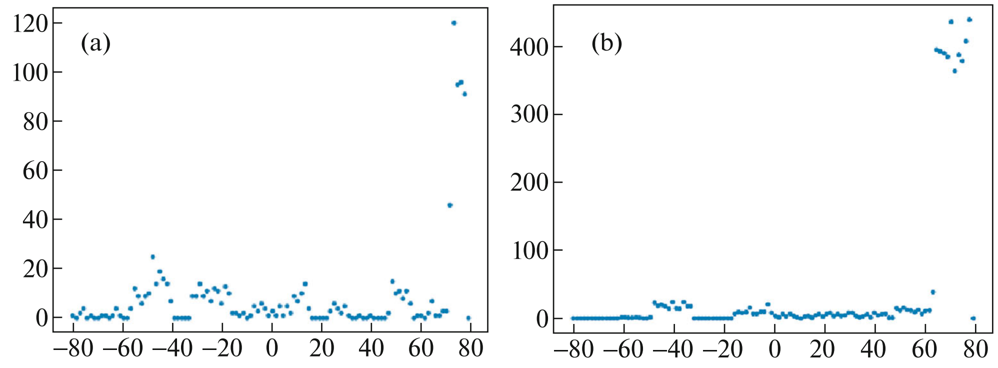

Fig. 2. Distribution of the number of samples according to the ranges of control parameter values: (a) $Z = 1000$, (b) $Z = 5000$.

(before training the ANN, the values of the differences were divided by Normir = 40 in order to guarantee the values of the increments of parameters submitted to the neural network in the range [−1, 1]) and summing with the last known values of yfin and pfin and calculated forecast values for time slices of the management horizon—sales quantities $Q_{j,t}$ and prices $\text{Price}_{j,t}$,

$$Q_{j,t} = \text{yfin} + \left(\sum_{u=0}^{t} Y_{j,t}\right)\text{Normir}$$
$$\text{Price}_{j,t} = \text{pfin} + \left(\sum_{u=0}^{t} U_{j,t}\right)\text{Normir} \quad (6)$$
$$P_t = Q_t(\text{Price}_t - \text{NetCost}).$$

The optimization task is to find the set U_t:

$$\{U_t\}\text{opt} = \underset{\{U_t\}}{\text{Argmax}} \sum_{t=1}^{T} \langle\{P_t\}\rangle, \quad (7)$$

where $\langle\{P_t\}\rangle$ is the profit function values of the time slices t of the control horizon calculated from the restored forecast values of the average daily sales Q_t and the restored values of the average daily price Price_t on time slices t of the management horizon, $t \in \{1..T\}$; NetCost is the cost of the goods.

Optimization was carried out in two stages. At the first stage, the entire range of possible values of the control parameter (values of differences in average daily sales) for each time slice of the control horizon $[U_{\min_t}, U_{\max_t}]$ was divided into a certain number of discreteness intervals Nu1 (in our case Nu1 = 10). Then, from the obtained discrete values U_{j_t} for each time slice t, the integer generator on the iteration with the number $z = 1..Z1$ generated the number $n_{z,t}$ of the discreteness interval from the set of numbers {1, 2..Nu1} from which a set of value increments of price $U_{n_{z,t}}$ was randomly selected, one for each time slice of the control horizon. For this set, using (2) and (3), the corresponding values of the state parameters $Y_{z,t}$ (increase in the number of sales per day) were calculated for which by (6) sets of values of the target function—the profit function $P_{z,t}$—were calculated. The range of values range of the control parameter U (price increment) calculated using the 3 sigma rule was –80.371 to 80.371. From the sets $P_{z,t}$ obtained by random sampling with the number of attempts $Z1 = 1000$, a locally optimal trajectory was calculated, having the maximum sum $P_{z,t}$ for all time slices t. A random sampling of the 1000 acquired paths resulted in one with the maximum sum, which is denoted as $S1$. As a result, for the used ANN model, a locally optimal trajectory was selected, for which the value of the sum of the target function (profit function of the enterprise) for all time slices of the control horizon was obtained: $S1$ = 48 518.902.

At the second stage, the sampling algorithm begins with dividing the range of possible values $[U_{\min_t}, U_{\max_t}]$ by Nu = 10 of the same intervals and constructing the same beta distributions for these intervals with parameters $a = 2.01$ and $b = 3.01$. For each zth iteration of sampling ($z = 1..Z$) in accordance with the modified Thompson algorithm described above, an interval with a maximum value is selected for each time slice $f_{j,t}$, calculated according to (4), taking into account parameters of the beta distributions $\alpha_{j,t}$ and $\beta_{j,t}$ from which $U_{n_{z,t}}$ was selected; then, according to (2) and (3), $Y_{z,t}$ was calculated; according to (6), the tuple $\langle\{P_t\}\rangle$ was calculated; according to (7), the amount of profit on the trajectory S_z was calculated. If $S_z > 0.5S1$, then according to (5) $\alpha_{j,t}$ and $\beta_{j,t}$ were adjusted. In this case, it is proposed to choose the adjustment parameter, as in [14], on the order of $1/\sqrt{Z\log(Z)}$.

After 1000 consecutive sampling iterations, the following pattern of the distribution of the number of samples per interval is obtained (see Fig. 2a). In Fig. 2b, the distribution of the number of samples by intervals for Nu = 10 and $Z = 5000$ iterations of sampling is given.

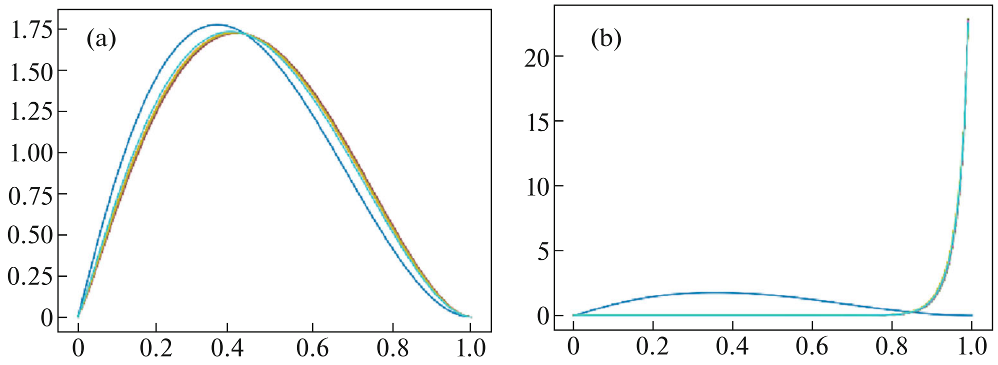

Fig. 3. Offset of the median of distributions: (a) at $r = 1/Z$, (b) at $r = 50/Z$.

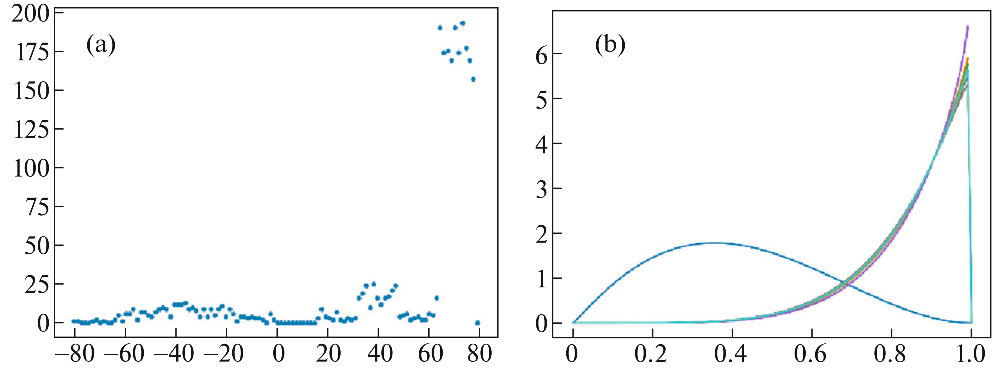

Fig. 4. Distribution of samples by intervals (a); evolution of beta-distribution density functions (b).

The result of sampling according to the proposed modified Thompson algorithm with a total number of 1000 attempts was a trajectory that leads to the value of the sum of the target function (profit of the seller) $S2 = 88776.489$, i.e., twice the amount of profit obtained for the optimal trajectory selected by random sampling for 1000 attempts. Figure 3 shows the offset of the median distributions for the most successful intervals with a soft change in the distribution parameters due to the reward $r = 1/Z$ (Fig. 3a) and a more stringent change in the distribution parameters $r = 50/Z$ (Fig. 3b).

At the same time, with a larger number of samples $Z = 5000$, but a more rigid method of applying remuneration, the difference in the efficiency of random sampling and the modified Thompson sampling algorithm was reduced: $S1 = 54489.066$, $S2 = 69028.902$. The use of intermediate parameters increased efficiency compared to a rigid approach, but it was still less effective than the soft approach: $S1 = 48552.101$, $S2 = 70581.827$ at $Z = 2500$, Nu = 10, and $r = 10/Z$. The interval distribution of the number of samples and the evolution of the density functions of the beta distributions are shown in Figs. 4a and 4b, respectively.

For this variant, the locally optimal trajectory of increments of the guide parameter (price) is as follows: [80.3 74.7 76.2 79.3 72.8 70.7 78.8 –46.8 77]. On time slices 1–7, the system proposes to choose price increments close to the maximum possible. On the eighth slice, it proposes to reduce the price and raise it again on the ninth time slice.

CONCLUSIONS

The methods and algorithms developed by the author of construction of a predictive model of a control object based on an artificial neural network, as well as a modified Thompson sampling algorithm to determine the optimal set of control parameters that deliver the maximum amount of the target function across all slices of the control horizon, are considered in the work. They have well proved themselves by a specific example of predicting the optimal prices for the sale of goods in the marketplace. The optimal prices are calculated for nine time slices on the management horizon for more than a month using time series of data on the average daily price for a product and on the average daily sales of this product in the marketplace recorded for 497 days.

FUNDING

This work was carried out within the framework of the grant agreement from the Innovation Promotion Foundation 122GS1IIS12-D7/76682 dated June 9, 2022.

CONFLICT OF INTEREST

The author declares that he has no conflicts of interest.

REFERENCES

1. L. S. Chernyshev, “Method and a system for predicting time series values using an artificial neural network,” RF Patent No. 2744041 S1 (2021).
2. L. S. Chernyshev, “Method and control system for engineering system or process control using an artificial neural network optimizing a goal function,” Invention Appl. 2022123701 (2022).
3. L. S. Chernyshev, “Forecasting the values of a time series using neural network in solving the problem of dynamic pricing,” in *24th Int. Sci.-Tech. Conf. Neiroinformatics-2022, Dolgoprudny, Moscow oblast, 2022.*
4. L. S. Chernyshev, “Method and optimal control system for engineering system or process control using the Tompson sampling,” Invention Appl. 2023102321 (2022).
5. N. D. Egupov, *Methods of Robust, Neuro-Fuzzy, and Adaptive Control: Textbook*, Ed. by N. D. Egupov, 2nd ed. (Mosk. Gos. Tekh. Univ. im. N.E. Baumana, Moscow, 2002).
6. K. J. Ferreira, D. Simchi-Levi, and H. Wang, “Online Network Revenue Management Using Thompson Sampling,” Oper. Res. (2017). https://doi.org/10.2139/ssrn.2588730
7. A. I. Galushkin, *Theory of Neural Networks* (IPRZhR, Moscow, 2000).
8. S. Haykin, *Neural Networks: A Comprehensive Foundation* (Prentice Hall, 1998).
9. I. M. Makarov, V. M. Lokhin, S. V. Man’ko, and M. P. Romanov, *Artificial Intelligence and Smart Control Systems* (Nauka, Moscow, 2006).
10. K. A. Pupkov, “Problems of theory and practice of intelligent systems,” in *Mechanical, Instrumental, and Power Engineering*, Ed. by A. N. Tikhonov, V. A. Sadovnichii (Izd-vo Mosk. Univ., Moscow, 1994), pp. 263–266.
11. D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Zh. Wen, “A tutorial on thompson sampling,” Found. Trends Mach. Learn. **11** (1), 1–98 (2018). https://doi.org/10.1561/2200000070
12. V. A. Terekhov, D. V. Efimov, I. Yu. Tyukin, and V. N. Antonov, *Neural-Network Control Systems* (S.-Peterb. Univ., St. Petersburg, 1999).
13. A. V. Timofeev and R. M. Yusupov, “Intelligent control systems,” Izv. Ross. Akad. Nauk. Tekh. Kibern., No. 5, 209–224 (1994).
14. W. R. Thompson, “On the theory of apportionment,” Am. J. Math. **57** (2), 450–456 (1935). https://doi.org/10.2307/2371219

L.S. Chernyshev. Born in Moscow in 1972. In 1996, he graduated from the Faculty of Theoretical and Experimental Physics of MEPhI.

From 1996 to 2006, postgraduate student, software engineer, Head of the Laboratory of the State Oceanographic Institute, Moscow. In 2001, he defended his dissertation for the degree of Candidate of physical and mathematical sciences titled “Modeling and Analysis of Variability of Hydrological-Hydrochemical Characteristics of Water Masses of the North Atlantic.” Since 2006, director at commercial companies. Since 2022, General Director of the Matsoft scientific company, Moscow. Author of 13 published works (including in co-authorship), including one patent for an invention.