Evsey V. Morozov
Alexander S. Rumyantsev
Oleg V. Lukashenko (Eds.)

Proceedings
of the Fourth International
Workshop **SMARTY'24**

Stochastic
Modeling &
Applied
Research of
TechnologY

Fourth International
Workshop
Petrozavodsk, Karelia
August 26-30, 2024

M. A. Ershov and A. S. Voroshilov

Optimizing the UCB Strategy for Batch Data Processing

Stochastic Modeling and Applied Research of Technology, Vol. 4, Pp. 11-21.
DOI: 10.57753/SMARTY.2026.96.62.002

Citation: M. A. Ershov and A. S. Voroshilov (2026). Optimizing the UCB Strategy for Batch Data Processing. Stochastic Modeling and Applied Research of Technology, 4, 11-21. https://doi.org/10.57753/SMARTY.2026.96.62.002

Optimizing the UCB Strategy for Batch Data Processing*

Ershov M. A.[1] and Voroshilov A. S[2]

[1] Yaroslav-the-Wise Novgorod State University, Veliky Novgorod 173003, Russian Federation
s244525@std.novsu.ru

[2] Yaroslav-the-Wise Novgorod State University, Veliky Novgorod 173003, Russian Federation
s244528@std.novsu.ru

Abstract. For the Gaussian two-armed bandit, which occurs during the analysis of batch data processing, the case of variable batch size is considered. This optimal control problem has a classical interpretation as a game with nature. The player's payment function is the mathematical expectation of the regrets of full income caused by incomplete information. The control goal is considered in a minimax setting, and a UCB strategy is used to achieve it. An invariant description of a strategy with a unit control horizon is constructed for the problem under consideration. According to the invariant description, the normalized regrets depend on the number of batches and do not depend on their size. Increasing the batch size during control allows the player to process more data with a constant number of actions (batches). In the area of "close" distributions, a slight increase in normalized regrets was observed, but in "far" distributions, a significant decrease in these was observed.

Keywords: Two-armed Bandit Problem · UCB strategy · Batch Processing · Invariant Description · Variable Batch Size.

1 Introduction

We consider the process of batch data processing using the two-armed bandit problem [1,2]. The mathematical model in the problem is a slot machine with two arms, which we will further call actions. The whole process consists of N such choices, where N is the control horizon. Each choice of one of the actions at time n gives the player a random one-step income x_n. The distribution of the one-step income depends only on the currently selected action, is fixed during the game and is unknown to the player.

The player's goal is to maximize the mathematical expectation of the full income. To do this, during the game it is necessary to determine the action corresponding to the highest average income and ensure its preferential choice.

* Supported by Russian Science Foundation, project number 23-21-00447, https://rscf.ru/en/project/23-21-00447/.

Since the distribution of income x_n is unknown, the player is forced to solve the "information or control" dilemma. To maximize the full expected income, the player would like to apply only the action that corresponds to the highest expected one-step income. However, to determine such an optimal action, it is necessary to compare all possible actions, which inevitably leads to regrets.

To determine the best action, we use the UCB rule. The UCB (Upper Confidence Bound) strategy is one of the most widely used algorithms in many fields of research, including machine learning and operations research. The idea of the strategy is to choose the best action at time n based on interval estimates of the mathematical expectation of one-step income x_n. It consists of two stages. At the first stage, initial statistics are collected: the player selects each action once and remembers their income. In the second stage, the player chooses the action with the highest current value of the upper bound of the confidence interval U_i. In this case, the upper bounds of U_i depend on the number of selected actions and the player calculates them before each choice.

To maximize full income, the player strives to minimize regrets. This approach is called minimax, as it is quite careful in making decisions and considers all possible risks.

As a practical use of the problem, we consider batch data processing with two methods. Processing a large count of data one at a time will take too long. Instead, we will divide the data into K batches of size M and process the data in batches. Processing can be successful ($x_n = 1$) or unsuccessful ($x_n = 0$) [3,4,5]. In this case, choosing one method means choosing the same action M times in a row. Then the income for processing one batch is the sum of the identically distributed random variables with the Bernoulli distribution. If the sizes of the batches are large enough, then under broad assumptions, due to the central limit theorem, the distribution of total incomes in them will be close to Gaussian.

The paper is organized as follows. In Section 2, the main provisions of the two-armed bandit problem and batch data processing are described. In Section 3, the UCB strategy and the regret function are given. Section 4 presents an invariant description on a unit control horizon. According to the description, the maximum normalized regrets depend on the number of batches and do not depend on their size. At the same time, it is worth noting that the invariant description considers the case of "close" distributions, when the difference in mathematical expectations $\Delta = |m_1 - m_2|$ has the order $N^{1/2}$ [6]. The smaller the Δ, the more difficult it is to determine the best action and the greater the maximum regrets. In Section 5, a variable batch size strategy is proposed. We noticed that with a constant number of batches and a gradual increase in their size, there is a slight increase in normalized regrets for "close" distributions and a significant decrease in these for the case of "far" distributions. At the same time, the number of processed data increases significantly. We have considered cases where, after each T actions, the current batch size is increased either by a fixed percentage, or by a fixed count of data. This strategy allows the player to increase the count of processed data while limiting the number of batches. Section 6 concludes the paper.

2 The Two-armed Bandit

The one-step income x_n at time n depends only on the selected action y_n. Formally, it is a controlled random process with density

$$f(x|m_i) = \frac{\exp(-\frac{(x-m_i)^2}{2D_i})}{(2\pi D_i)^{1/2}}, \tag{1}$$

if $y_n = i \in \{1, 2\}$, where $n \in \{1, 2, \ldots, N\}$ and N is the control horizon. The mathematical expectations m_1, m_2 are fixed and unknown to the player. The variances D_1, D_2 are assumed to be known a priori. This requirement may be canceled in the future, since the algorithm is not sensitive to changes in variance. Therefore, the variance can be estimated at an initial stage.

Let's consider the use of a two-armed bandit in the problem of optimizing data processing. Suppose N units of data need to be processed and there are two processing methods with a priori unknown efficiencies. Let the one-step income x_n indicate success or failure in processing one data element, thus having Bernoulli distribution with mathematical expectation $m_i = p_i$ and variance $D_i = p_i\ (1 - p_i)$. If $x_n = 1$, then the element of data was processed successfully, otherwise ($x_n = 0$) – failure. Instead of processing the data one at a time, we divide them into K batches of M elements each. In this case, the income for choosing the method (action) is the number of successfully processed data, which is equal to the sum of random variables with the Bernoulli distribution [7,8]. If the batch sizes are large enough, then, according to the central limit theorem, the total income for one processing of M units of data will have an approximately normal distribution, and this bandit will be equivalent to a Gaussian one. Therefore, control strategies should not depend on the distribution of one-step income and become universal.

Batch processing allows the player to change the action less often, since the total processing time is determined by the number of batches. Also, if parallel data processing can be organized, the total processing time will be significantly reduced. In this case, the regrets will be close to the case when the data is processed one at a time.

Previously, we considered cases where M is const and the control horizon was defined as $N = MK$. However, it will be further proved that M does not have to be fixed.

3 UCB Strategy

Let k batches have been processed by time n, out of which k_i batches have been processed by method i, and the number of successfully processed data is $X_i(k)$ for each method. Then the ratios $X_i(k)/k_i$, $i = 1, 2$, are point estimates of mathematical expectations m_i. However, they are not accurate enough. J. Baser [9] proposed using the upper bounds of their interval estimates (Upper Confidence Bound — UCB) instead of point estimates to select an action [10,11,12,13,14].

At the beginning of the control, the player selects each method M times and remembers the number of successfully processed data. Next, before processing the $k+1$-st batch, the player calculates the value of the upper bound of the confidence interval $U_i(k)$

$$U_i(k) = \frac{X_i(k)}{k_i} + a\left(\frac{MD_i \ln k}{k_i}\right)^{1/2}, \tag{2}$$

where $i \in \{1,2\}$, $X_i(k)$ is the number of successfully processed data (current total income) when selecting the action i, k_i – the number of processed batches when selecting action i, $k = k_1 + k_2$ – total number of processed batches, $a > 0$ – strategy parameter.

Note that rule (2) is valid only when M is constant throughout the entire control. Instead, the upper bounds of U_i can be found based on the total count of processed data n

$$U_i(n) = \frac{X_i(n)}{n_i} + a\left(\frac{D_i \ln n}{n_i}\right)^{1/2}, \tag{3}$$

where n_i is the number of processed data when selecting method i, $n = n_1 + n_2$ – the total number of processed data. For further control, we will use the formula (3).

Let's move on to the formulation of the control goal. If the best action were known, then the total income would be defined as $N \cdot \max(m_1, m_2)$. However, in practice, the total income is almost always less than the maximum, and on average the income is equal to $\mathbf{E}(X_1(K) + X_2(K))$. Note that the player can get the maximum income only when $m_1 = m_2$. Let's normalize the income difference and get the regret function

$$L_N(d,a) = \frac{N\max(m_1,m_2) - \mathbf{E}(X_1(K) + X_2(K))}{(DN)^{1/2}}, \tag{4}$$

where $m_1 = m + d(D/N)^{1/2}$, $m_2 = m - d(D/N)^{1/2}$, $d \geq 0$ is the deviation, and $D = \max(D_1, D_2)$. With such values of mathematical expectations, the difference $|m_1 - m_2| = cN^{-1/2}$, that is, it will have the order $N^{-1/2}$, where c is a constant [6]. The closer the value of c is to zero, the more difficult it is to determine the optimal action. These values m_1, m_2 characterize "close" distributions. Therefore, we will limit ourselves to this set of parameters and minimize the maximum regrets.

It was found in [8] that the regret function $L_N(d,a)$ is not very sensitive to changes in variance. In the case of batch processing, D_i can be evaluated once at the statistics collection stage, when the player selects each action once and processes M data at a time. Therefore, the requirement of a priori knowledge of variances can be abolished.

4 Invariant Description

As suggested in [8], we introduce the function $I_i(k)$ – an indicator of the selected action for batch k, provided $k > 2$

$$I_i(k) = \begin{cases} 1 & \text{if } U_i(k) = \max(U_1(k), U_2(k)) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $i = \{1, 2\}$ is action number. For $k \leq 2$, each action will be applied one batch at a time. Applying the introduced function, the total income for the application of each action can be written as the sum of the income received for the application of action i:

$$X_i(k) = k_i M m_i + \sum_{j=1}^{k} I_i(j)\eta_i(MD_i; j) = \\ = k_i M \left(m + d_i \left(\frac{D}{MK} \right)^{1/2} \right) + \sum_{j=1}^{k} I_i(j)\eta_i(MD_i; j), \tag{6}$$

where $d_1 = d, d_2 = -d$ from (4), $\eta_i(MD_i; j) \sim N(0, (MD_i)^{1/2})$ are independent normally distributed random variables with zero mathematical expectation and variance MD_i. Since $I_i(k)$ takes the value 1 exactly k_i times, $\sum_{j=1}^{k} I_i(j)\eta_i(MD_i; j)$ is the sum of k_i normally distributed quantities, each of which can be represented as the product of a standard random variable $\eta \sim N(0, 1)$ by the standard deviation of $\eta_j(MD_i)$,

$$X_i(k) = k_i M \left(m + d_i \left(\frac{D}{MK} \right)^{1/2} \right) + \eta (k_i MD_i)^{1/2}. \tag{7}$$

Then we write down the value of the upper bound (2) considering (7)

$$U_i(k) = Mm + d_i \left(\frac{MD}{K} \right)^{1/2} + \eta \left(\frac{MD_i}{k_i} \right)^{1/2} + a \left(\frac{MD_i \ln k}{k_i} \right)^{1/2} \tag{8}$$

Next, we apply the following transformation to (8), which does not affect the order of the arrangement

$$u_i(t) = (U_i(k) - Mm) \left(\frac{K}{MD} \right)^{1/2}. \tag{9}$$

To transform the strategy on the control horizon $N = 1$, we introduce the following notation

$$t = \frac{k}{K},\ t_i = \frac{k_i}{K},\ \varepsilon = \frac{1}{K},\ \gamma_i = \frac{D_i}{D}. \tag{10}$$

As a result, we obtain an expression for the upper bounds of confidence intervals (8) in invariant form, depending on the relative size of the batch ε,

$$u_i(t) = d_i + \eta \left(\frac{\gamma_i}{t_i}\right)^{1/2} + a \left(\gamma_i \frac{\ln(t/\varepsilon)}{t_i}\right)^{1/2}. \tag{11}$$

Next, we describe the regret function in an invariant form. Let $d_0 = \max(d_1, d_2)$. Then the expected regrets of the strategy can be written as the sum of the regrets from the application of the worst:

$$L_N(d,a) = \frac{\sum_{i=1}^{2}(d_0 - d_i)(D/N)^{1/2}\mathbf{E}(\sum_{k=1}^{K} MI_i(k))}{(DN)^{1/2}} = \\ = \frac{\sum_{i=1}^{2}(d_0 - d_i)\mathbf{E}(Mk_i)}{N} = \frac{\sum_{i=1}^{2}(d_0 - d_i)\mathbf{E}(MKt_i)}{N} = \sum_{i=1}^{2}(d_0 - d_i)\mathbf{E}(t_i). \tag{12}$$

From (12) it becomes obvious that the regret function does not depend on the batch size M.

5 Numerical Results

Earlier in [15] it was found that with a decrease in the batch size at the stage of collecting statistics, the local maximum of the regret function $L_N(d,a)$ increases slightly but decreases significantly with an increase in the difference in mathematical expectations of actions.

Let's introduce the following rule for batch processing. Suppose there is a starting batch size M_0, which is constant at the stage of collecting statistics. Next, after each T-th processed batch, let the current batch size $M(k)$ be multiplied by $\alpha > 1$. As a result, the formula for the dependence of the batch size on the number of actions is obtained

$$M(k) = \left[\alpha^{[k/T]} M_0\right], \tag{13}$$

where $[x]$ gives the integer part of x (it is needed since the batch size is a natural number). Then the control horizon N will be defined as

$$N(K) = J \cdot M_0 + \sum_{k=J+1}^{K} \left[\alpha^{[k/T]} M_0\right], \tag{14}$$

where $J \geq 2$ is the number of the bandit's arms.

Next, let's consider how each parameter of formula (13) affects the regrets $L_N(d,a)$. Let $K = 50$, $M_0 = 50$, $T = 10$. With increasing α, the local maximum increases slightly. However, with large differences in the mathematical expectations of the actions, the regrets are significantly reduced (Fig. 1). Note that for

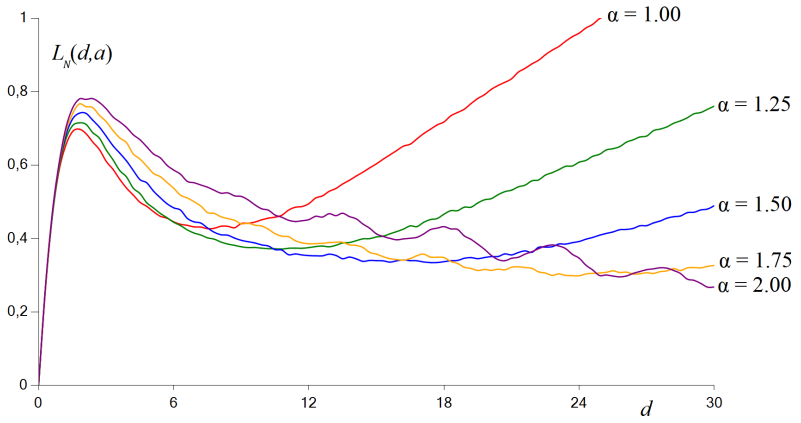

Fig. 1. Changing the coefficient α.

$\alpha = 1$, the batch size $M(k)$ is constant. Based on the data, it is optimal to take $\alpha = 1.50$ or $\alpha = 1.75$.

Next, let $K = 50, M_0 = 50, \alpha = 1.5$. As T decreases, the local maximum increases slightly. Similarly, with large differences in the mathematical expectations of actions, significant reductions in regrets are visible (Fig. 2). Based on the data, it is optimal to take $T = 10$.

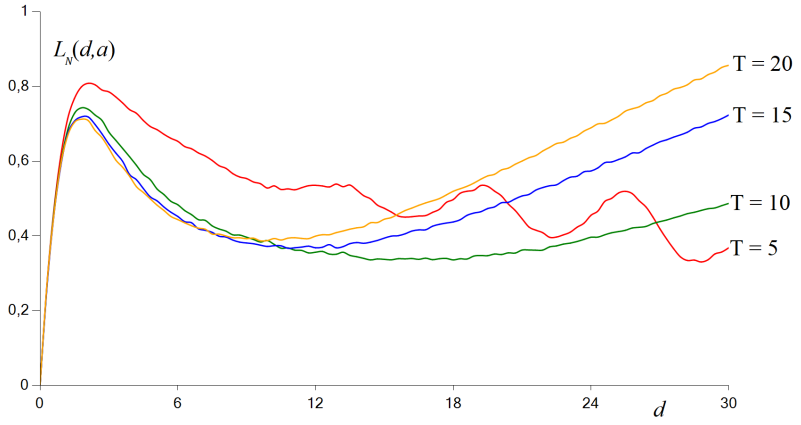

Fig. 2. Changing the number of processed batches T before changing the batch size M.

We will also confirm by modeling that the regrets of $L_N(d, a)$ do not depend on the batch size. Let $K = 50, T = 10, \alpha = 1.5$. The starting batch size M_0 does not affect the normalized regrets (Fig. 3).

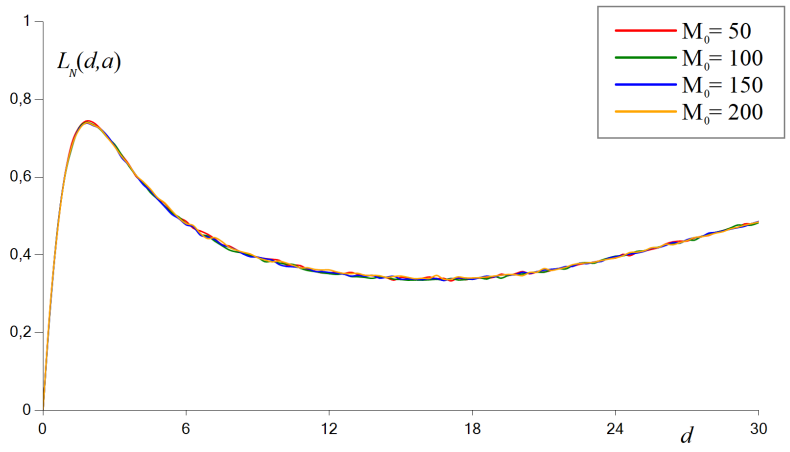

Fig. 3. Changing the initial batch size M_0.

Next, let's put the number of batches $K = 50$, the initial batch size $M_0 = 50$, the number of processed batches before updating the size $T = 10$, the coefficient $\alpha = 1.5$. Let's compare regrets with a constant size $M = M_0$ and regrets with a variable size $M(k)$. In this case, with a constant size, the player will process $N = 2500$ data, with a variable size $N = 6909$ (Fig. 4). One can see that with small differences in mathematical expectations of actions, the player regrets a little more, but with large deviations, the player loses much less. In both cases, there are exactly 50 batches, but with a variable batch size, the player processes about 2.75 times more data.

Let's introduce another rule for batch processing. Suppose there is still a starting batch size M_0, which is constant at the stage of collecting statistics. Next, let's assume that after each T-th processed batch, the current size $M(k)$ increases by β data. Then the batch size will be equal to

$$M(k) = M_0 + [k/T] \cdot \beta \tag{15}$$

The control horizon N will be defined as

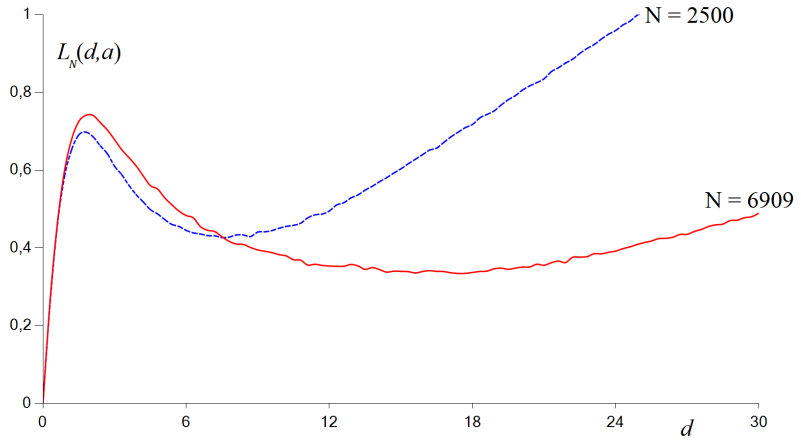

Fig. 4. Comparing a constant batch size with a variable.

$$N(K) = J \cdot M_0 + \sum_{k=J+1}^{K} (M_0 + [k/T] \cdot \beta) =$$

$$= J \cdot M_0 + (K - (J+1) + 1) \cdot M_0 + \beta \sum_{k=J+1}^{K} [k/T] = M_0 K + \beta \sum_{k=J+1}^{K} [k/T], \tag{16}$$

where $J \geq 2$ is the number of the bandit's arms.

Let's consider how the parameter β affects the regrtes $L_N(d, a)$. Let $K = 50, M_0 = 50, T = 10$. As in the case of α with an increase in β, the local maximum increases slightly. However, with large differences in the mathematical expectations of actions, regrets are significantly reduced (Fig. 5). Based on the data, it is optimal to take $\beta = 50$.

Note that at $\beta = 0$ the batch size $M(k)$ is constant and $N = 2500$, at $\beta = 50$ the horizon is $N = 7750$. In both cases, there are exactly 50 batches, but with a variable batch size, the player processes almost 3 times more data.

6 Conclusion

We have considered a modification of batch data processing using the two-armed bandit problem. Using invariant description and modeling, it was found that regrets do not depend on the size of the batch, therefore, it is possible to change it, thereby increasing the count of processed data. We have proposed two rules for changing the batch size $M(k)$. The first rule is multiplication $\alpha > 1$, that is,

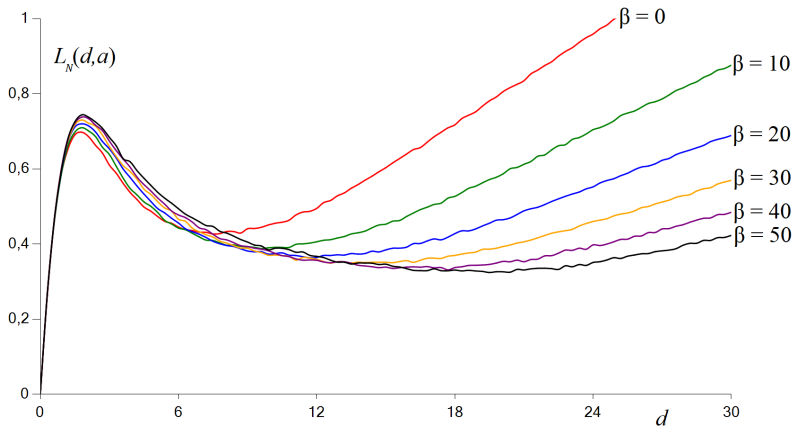

Fig. 5. Changing the coefficient β.

an increase by a constant percentage. The second rule is to increase the data by β. For the first rule, it is optimal to take $\alpha = 1.5$, for the second - take $\beta = 50$. In the case of "close" distributions, a slight increase in regrets was observed, but a significant decrease in these for "far" distributions.

References

1. Presman, E., Sonin, I.: Sequential control with incomplete information. Academic Press, New York (1990)
2. Berry, D., Fristedt, B.: Bandit Problems: Sequential Allocation of Experiments. Chapman and Hall, London, New York (1985)
3. Varshavsky, V.: Kollektivnoe povedenie avtomatov (Collective Behavior of Automata). Nauka, Moscow (1973) (in Russian)
4. Poznyak. A., Nazin A.:Adaptivnyi vybor variantov (Adaptive choice of options). Nauka, Moscow (1986) (in Russian)
5. Tsetlin, M.: Automation theory and modeling of biological systems. Academic Press, New York (1973) https://doi.org/10.1016/s0076-5392(08)x6049-7
6. Vogel, W.: An asymptotic minimax theorem for the two-armed bandit problem. Annals of Mathematical Statistics **31**(2), 444–451 1960 https://doi.org/10.1214/aoms/1177705907
7. Kolnogorov, A.: Gaussian Two-Armed Bandit and Optimization of Batch Data Processing. Problems of Information Transmission **54**(1), 84–100 (2018) https://doi.org/10.1134/S0032946018010076
8. Kolnogorov, A.: Gaussian Two-Armed Bandit: Limiting Description. Problems of Information Transmission **56**(3), 278–301 (2020) https://doi.org/10.1134/S0032946020030059
9. Bather, J.: The Minimax Risk for the Two-Armed Bandit Problem. In: Herkenrath, U., Kalin, D., Vogel, W. (eds) Mathematical Learning Models — Theory and Algorithms. Lecture Notes in Statistics, vol. 20, pp 1–11. Springer, New York. (1983) https://doi.org/10.1007/978-1-4612-5612-0_1

10. Smirnov, D., Gromova, E.: Decision-making model under presence of experts as a modified multi-armed bandit problem. Mathematical Game Theory and Applications **9**(4), 69–87 (2017) (in Russian)
11. Auer, P.: Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research **3**, 397–422 (2002)
12. Kaufmann E.: On Bayesian Index Policies for Sequential Resource Allocation. Annals of Statistics **46**, 842–865 (2018) https://doi.org/10.1214/17-AOS1569
13. Lai, T.: Adaptive Treatment Allocation and the Multi-Armed Bandit Problem. Annals of Statistics **15**(3), 1091–1114 (1987) https://doi.org/10.1214/aos/1176350495
14. Lattimore T., Szepesvari C.: Bandit Algorithms. Cambridge University Press, Cambridge (2020) https://doi.org/10.1017/9781108571401
15. Ershov, M.,Kolnogolov, A., Voroshilov, A. : Machine learning, customization of the UCB strategy with reducing the amount of data. p. 9, in press.