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

Proceedings
of the Third International
Workshop **SMARTY'22**

SMART Y 22
Stochastic Modeling & Applied Research of TechnologY
Third International Workshop
Petrozavodsk, Karelia
August 21-25, 2022

Ershov, M., Kolnogorov, A., & Voroshilov, A.

Customization of the Auer–Cesa-Bianchi–Fisher UCB Strategy for a Gaussian Two-Armed Bandit

Stochastic Modeling and Applied Research of Technology, Vol. 3, Pp. 1-13.
DOI: 10.57753/SMARTY.2023.40.11.001

Citation: Ershov, M., Kolnogorov, A., & Voroshilov, A. (2023). Customization of the Auer–Cesa-Bianchi–Fisher UCB Strategy for a Gaussian Two-Armed Bandit. Stochastic Modeling and Applied Research of Technology, 3, 1-13. https://doi.org/10.57753/SMARTY.2023.40.11.001

Customization of the Auer–Cesa-Bianchi–Fisher UCB Strategy for a Gaussian Two-Armed Bandit

Maxim Ershov, Alexander Kolnogorov, and Albert Voroshilov

Yaroslav-the-Wise Novgorod State University
s244525@std.novsu.ru, Alexander.Kolnogorov@novsu.ru, s244528@std.novsu.ru

Abstract. We consider a two-armed bandit problem in relation to data processing if there are two alternative processing methods with different a priori unknown efficiencies. One has to determine the most efficient method and ensure its preferential use. We consider batch data processing when all the data is divided into batches. For the control, we present the batch version of the UCB strategy which was first introduced by P. Auer, N. Cesa-Bianchi and P. Fisher. We develop two approaches to the invariant description of the control process on the horizon equal to one. The first approach allows us to compute a regret using Monte-Carlo simulations and the second approach provides the analytical formalism for solving a recursive Bellman-type dynamic programming equation. Numerical results show the high efficiency of the presented strategy.
Keywords: Gaussian two-armed bandit, UCB strategies, Bayesian and minimax approaches, batch processing, Monte-Carlo simulations, dynamic programming.

1 Introduction

We consider a two-armed bandit problem (see., e.g., [1,2]). In what follows, we consider a Gaussian two-armed bandit. Formally, it is a controlled random process ξ_n, $n = 1, 2, \ldots, N$, where N is a control horizon. Random variable ξ_n depends only on currently chosen action y_n and is normally distributed with a density

$$f_{D_\ell}(x|m_\ell) = (2\pi D_\ell)^{-1/2} \exp\left(-\frac{(x-m_\ell)^2}{2D_\ell}\right) \tag{1}$$

if $y_n = \ell$, $\ell = 1, 2$. Variances D_1, D_2 are assumed to be known. We also assume that $D_1 \geq D_2$, otherwise, the variances can be renumbered. Note that the assumption of a priori known variances can be removed later. Mathematical expectations m_1, m_2 are assumed to be unknown and are not ordered. This two-armed bandit can be described by a parameter $\theta = (m_1, m_2)$ with the set of possible values Θ, which will be defined below.

Let us explain why Gaussian two-armed bandit is studied. We consider the problem in application to batch data processing. Let ζ_n, $n = 1, 2, \ldots, N$, be a Bernoulli controlled random process described by the distribution

$$\Pr(\zeta_n = 1 | y_n = \ell) = p_\ell, \quad \Pr(\zeta_n = 0 | y_n = \ell) = q_\ell, \tag{2}$$

where $p_\ell + q_\ell = 1$, $\ell = 1, 2$, $n = 1, 2, \ldots, N$. The values of the process 1 and 0 correspond to successively and unsuccessively processed data number n, actions correspond to alternative processing methods and the goal is to maximize the mathematical expectation of the total number of successively processed data. Assume that $N = MK$ and divide all the data into K batches of M data in each. Let us use the same methods to all the data in each batch and cumulative (or total) incomes in batches use for the control. If the sizes of batches are large enough then according to the central limit theorem distributions of cumulative incomes are close to Gaussian. The main property of this approach is that maximal expected loss of the total income corresponding to batch data processing is approximately the same as that corresponding to the optimal processing data one by one if the number of batches is large enough (see, e.g., [3,4]).

Note that initially batch processing was considered in application to treatment of patients with alternative methods. In this case, all the patients are divided into a number of groups: comparatively small test groups to which all the treatments are applied at the initial stage and a large remaining group to which the single treatment is applied that showed the most efficiency at the initial stage (see, e.g., [5,6]).

A strategy σ determines the choice of the action y_{n+1} using available information up to the point of time n. Below we consider a customization of the UCB strategy proposed in [7]. This strategy prescribes to apply all the actions once at the start of the control and then at each instant of time $n + 1$ to choose the action corresponding to the maximum of the values

$$Q_\ell(n) = \frac{X_\ell(n)}{n_\ell} + \left(\frac{2 \ln(n)}{n_\ell} \right)^{1/2}, \tag{3}$$

where $\ell = 1, 2$; $n = 2, \ldots, N - 1$. Here $n_\ell, X_\ell(n)$, $\ell = 1, 2$, are current cumulative counts of both actions' applications and corresponding cumulative incomes. Strategies like this are called UCB (Upper Confidence Bound) rules, they were considered, e.g., in [8,9,10,11].

Let us define the goal of the control. If the parameter $\theta = (m_1, m_2)$ were known, then the optimal strategy would always be to use the action corresponding to the maximum of m_1, m_2, the mathematical expectation of the total income is thus $N \max(m_1, m_2)$. If the strategy σ is applied, the total expected income is less than the maximal one by the value

$$L_N(\sigma, \theta) = N \max(m_1, m_2) - \mathbf{E}_{\sigma,\theta} \left(\sum_{n=1}^{N} \xi_n \right), \tag{4}$$

which is called the regret. Here $\mathbf{E}_{\sigma,\theta}$ denotes mathematical expectation computed over the measure generated by strategy σ and parameter θ. Let

$$R_N^M(\Theta) = \inf_{\{\sigma\}} \sup_{\Theta} L_N(\sigma, \theta) \tag{5}$$

be the minimax risk computed by the regret (4) on the set of parameters Θ for considered strategies. Corresponding strategy σ^M, if it exists, is called the minimax strategy.

The structure of the paper is as follows. In Section 2, we define a batch version of the strategy (3) and obtain its invariant description on the control horizon equal to one. These results allow us to compute the regret by Monte-Carlo simulations. In Section 3, we consider another approach to invariant description which allows us to compute the regret as a solution to the recursive integro-difference equation. Note that invariant descriptions are valid in the domain of "close" distributions where all mathematical expectations of one-step incomes differ by the values of the order of $N^{-1/2}$. Precisely in this domain the regret attains its maximum values. In Section 4, we present numerical results. These results show the high effectiveness of the proposed strategy. For example, in the case of 50 batches it ensures the maximum normalized regret 0.71. As this maximum regret can not be less than approximately 0.637 (see, [4]), this is seen as a good indicator, especially, since the strategy is simple. In Section 4, we also show that small deviations in the variances in the UCB rule do not affect a regret significantly. This allows one to make the estimates of unknown variances at the initial stage, when actions are used by turn, and then to use these estimates on the remaining control horizon. Section 5 contains a conclusion.

2 Invariant Description of the Control

The strategy (3) was proposed in [7] for multi-armed bandits which incomes belong to the segment $[0, 1]$. Let us consider its customization to the case of a Gaussian two-armed bandit. Denote $D = D_1 = \max(D_1, D_2)$ and $\gamma_\ell = D_\ell/D$, $\ell = 1, 2$. Let us define modified upper confidence bounds (3) as follows:

$$Q_\ell(n) = \frac{X_\ell(n)}{n_\ell} + a_\ell(\gamma_1, \gamma_2) \left(\frac{D_\ell \ln(n)}{n_\ell} \right)^{1/2}, \quad \ell = 1, 2, \tag{6}$$

where $a_\ell(\gamma_1, \gamma_2) > 0$, $1 = \gamma_1 \geq \gamma_2$. Factors $(D_\ell/n_\ell)^{1/2}$ characterize the widths of confidence intervals. Functions $a_1(\gamma_1, \gamma_2)$, $a_2(\gamma_1, \gamma_2)$ describe parameters of the UCB strategy, which can be used for ensuring minimax goal of the control (5). Clearly, $a_1(\gamma_1, \gamma_2) = a_2(\gamma_1, \gamma_2) = a$ if $\gamma_1 = \gamma_2 = 1$ because equal variances can be arbitrarily ordered.

Let us move on to the invariant description of the control. Consider the set of parameters $\Theta = \{\theta = (m_1, m_2)\}$, where

$$m_\ell = m + d_\ell (D/N)^{1/2};\ m \in (-\infty, +\infty),\ |d_\ell| \leq c,\ \ell = 1, 2, \tag{7}$$

where $c > 0$ is large enough fixed magnitude. This set of parameters describes "close" distributions, on which the regret attains its maximum values having the order of $N^{1/2}$ [12]. For "distant" distributions the order of the regret is different. For example, it is $\ln(N)$ asymptotically as $N \to \infty$ if $|m_1 - m_2| \geq \delta > 0$. In more detail "close" and "distant" distributions are discussed in [3,4].

In the sequel, we consider strategies which can change actions only after they have been used M times in succession. These strategies allow one to implement batch processing of data. Let us assume that $N = MK$, where K is the number of batches. For batch processing upper confident bounds are as follows

$$\hat{Q}_\ell(k) = \frac{\hat{X}_\ell(k)}{k_\ell} + a_\ell(\gamma_1, \gamma_2) \left(\frac{MD_\ell \ln(k)}{k_\ell} \right)^{1/2}, \ \ell = 1, 2, \tag{8}$$

where k is the cumulative number of processed batches, k_ℓ is the number of batches to which the ℓth action was applied, $\hat{X}_\ell(k)$ is the cumulative income corresponding to application of ℓth action after processing k batches ($k = 2, \ldots, K-1$). Let us denote by

$$\mathbf{I}_\ell(k) = \begin{cases} 1, & \text{if } \hat{Q}_\ell(k-1) = \max\left(\hat{Q}_1(k-1), \hat{Q}_2(k-1)\right), \\ 0, & \text{otherwise,} \end{cases}$$

an indicator function of chosen action while processing the kth batch according to considered rule if $k > 2$. Note that with probability 1 at each k only one $\mathbf{I}_\ell(k)$ is equal to 1. If $k \leq 2$ then each action is applied to precisely one batch and we can put $\mathbf{I}_\ell(k) = \delta_{\ell k}$, where $\delta_{\ell k}$ is Kronecker symbol. The cumulative income for each action application is

$$\hat{X}_\ell(k) = k_\ell M(m + d_\ell (D/N)^{1/2}) + \sum_{i=1}^{k} \mathbf{I}_\ell(i) \eta_\ell(MD_\ell; i), \tag{9}$$

where $\eta_\ell(MD_\ell; i) \sim \mathcal{N}(0, \sqrt{MD_\ell})$ are independent normally distributed random variables with zero mathematical expectation and variance MD_ℓ. Denote $t = kK^{-1}$, $t_\ell = k_\ell K^{-1}$, $\varepsilon = K^{-1}$. Note that t_1, t_2, t describe the relative usage times of actions computed with respect to the control horizon N, $\varepsilon = K^{-1} = M/N$ is a relative size of the batch. Taking into account (9), let us present upper confidence bounds (8) as

$$\hat{Q}_\ell(k) = Mm + \left(\frac{MD}{K} \right)^{1/2} \tag{10}$$
$$\times \left(d_\ell + \frac{\sum_{i=1}^{k} \mathbf{I}_\ell(i) \eta_\ell(\gamma_\ell \varepsilon; i)}{t_\ell} + a_\ell(\gamma_1, \gamma_2) \left(\frac{\gamma_\ell \ln(t\varepsilon^{-1})}{t_\ell} \right)^{1/2} \right), \ \ell = 1, 2.$$

Next, let us apply the following linear transformation to bounds (10), which does not affect their order:

$$\hat{q}_\ell(t) = (\hat{Q}_\ell(k) - Mm) \left(\frac{K}{MD} \right)^{1/2}.$$

As a result, we obtain the expressions for upper confident bounds of the UCB strategy with a control horizon equal to one, i.e., in invariant form depending on the relative size of the batch ε rather than on the magnitude of horizon N:

$$\hat{q}_\ell(t) = d_\ell + \frac{\sum_{i=1}^{k} \mathbf{I}_\ell(i)\eta_\ell(\gamma_\ell\varepsilon; i)}{t_\ell} + a_\ell(\gamma_1, \gamma_2)\left(\frac{\gamma_\ell \ln(t\varepsilon^{-1})}{t_\ell}\right)^{1/2}, \qquad (11)$$

$\ell = 1, 2$, where d_1, d_2 are determined according to (7). Now let us obtain the expression for the regret (4) in invariant form. Let us put $d_{l_0} = \max(d_1, d_2)$. Then using (7) we obtain

$$L_N(\sigma, \theta) = \sum_{\ell=1}^{2}(m_{l_0} - m_\ell)\mathbf{E}_{\sigma,\theta}\left(\sum_{i=1}^{K} M\mathbf{I}_\ell(i)\right)$$

$$= (D/N)^{1/2}\sum_{\ell=1}^{2} M(d_{l_0} - d_\ell)\mathbf{E}_{\sigma,\theta}\,(k_\ell) = (DN)^{1/2}\sum_{\ell=1}^{2}(d_{l_0} - d_\ell)\mathbf{E}_{\sigma,\theta}(t_\ell).$$

Therefore, the following normalized (with respect to $(DN)^{1/2}$) expression can be obtained for the regret

$$(DN)^{-1/2}L_N(\sigma, \theta) = \sum_{\ell=1}^{2}(d_{l_0} - d_\ell)\mathbf{E}_{\sigma,\theta}(t_\ell). \qquad (12)$$

The results can be stated as a theorem.

Theorem 1. *For a Gaussian two-armed bandit with fixed a priory known variances D_1, D_2 and unknown mathematical expectations m_1, m_2 belonging to the set of parameters* (7)*, the batch version of UCB strategy with upper confidence bounds* (8) *has invariant description with upper confidence bounds* (11)*. Normalized (with respect to the value $(DN)^{1/2}$) regret* (4) *is described by* (12)*.*

3 Another Approach to Invariant Description

In this section, we consider another approach to invariant description, which uses the results of [13]. Let us introduce a prior distribution density $\lambda(\theta)$ and consider the averaged regret

$$L_N(\sigma, \lambda) = \int_\Theta L_N(\sigma, \theta)\lambda(\theta)d\theta, \qquad (13)$$

where $L_N(\sigma, \theta)$ is defined in (4). It is convenient to present upper confidence bounds as

$$Q_\ell(X_\ell, n_\ell) = \frac{X_\ell}{n_\ell} + a_\ell(\gamma_1, \gamma_2)\left(\frac{D_\ell \ln(n)}{n_\ell}\right)^{1/2}, \qquad (14)$$

where n_ℓ, X_ℓ are current cumulative number of the usage of the ℓth action and corresponding cumulative income, $\gamma_\ell = D_\ell/D_1$, $\ell = 1, 2$. Let us use a notation $n_\ell^* = n_\ell D_\ell$. Given a history of the process X_1, n_1, X_2, n_2, the posterior distribution density is

$$\lambda(m_1, m_2|X_1, n_1, X_2, n_2) \tag{15}$$
$$= \frac{f_{n_1^*}(X_1|n_1 m_1) f_{n_2^*}(X_2|n_2 m_2)\lambda(m_1, m_2)}{p(\lambda; X_1, n_1, X_2, n_2)}$$

with

$$p(\lambda; X_1, n_1, X_2, n_2) \tag{16}$$
$$= \iint_\Theta f_{n_1^*}(X_1|n_1 m_1) f_{n_2^*}(X_2|n_2 m_2)\lambda(m_1, m_2) dm_1 dm_2.$$

We again assume that processing is implemented in K stages by batches of M data, and the first two batches are processed using both actions by turn. So, $N = KM$. Given the history of the process X_1, n_1, X_2, n_2, a strategy σ describes probabilities of choosing the ℓth action at the time interval $n+1, \ldots, n+M$, i.e., consists of probabilities $\sigma_\ell(X_1, n_1, X_2, n_2) = \Pr(y_{n+\nu} = \ell|X_1, n_1, X_2, n_2)$, where $n = n_1+n_2 = kM$, $k = 2, \ldots, K-1$, $\nu = 1, \ldots, M$. Denote by $L_{N-n}(\lambda; X_1, n_1, X_2, n_2)$ a regret on the remaining control horizon $N - n$ computed with respect to the posterior distribution density (15) ($n = n_1 + n_2$). Then for computing a regret (13) one has to solve the following recursive equation

$$L_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \sigma_1(X_1, n_1, X_2, n_2) L^{(1)}_{N-n}(\lambda; X_1, n_1, X_2, n_2)$$
$$+\sigma_2(X_1, n_1, X_2, n_2) L^{(2)}_{N-n}(\lambda; X_1, n_1, X_2, n_2), \tag{17}$$

where $L^{(1)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = L^{(2)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = 0$ if $n = N$ and

$$L^{(1)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint_\Theta \lambda(m_1, m_2|X_1, n_1, X_2, n_2)$$
$$\times \Big(M(m_2 - m_1)^+ + \mathbf{E}^{(1)}_x L_{N-n-M}(\lambda; X_1 + x, n_1 + M, X_2, n_2) \Big) dm_1 dm_2,$$
$$L^{(2)}_{N-n}(\lambda; X_1, n_1, X_2, n_2) = \iint_\Theta \lambda(m_1, m_2|X_1, n_1, X_2, n_2) \tag{18}$$
$$\times \Big(M(m_1 - m_2)^+ + \mathbf{E}^{(2)}_x L_{N-n-M}(\lambda; X_1, n_1, X_2 + x, n_2 + M) \Big) dm_1 dm_2$$

if $2M \le n < N$. Here $x^+ = \max(x, 0)$ and

$$\mathbf{E}^{(\ell)}_x L(x) = \int_{-\infty}^{\infty} L(x) f_{MD_\ell}(x|M m_\ell) dx$$

denotes mathematical expectation with respect to the probability density describing the application of the ℓth action to the batch of data ($\ell = 1, 2$). One

can see that $L^{(\ell)}_{N-n}(\lambda; X_1, n_1, X_2, n_2)$ is a mathematical expectation of the loss of cumulative income on the remaining control horizon $N-n$ if the ℓth action is applied to the first batch and then a control is implemented according to the strategy σ. UCB strategy is described as follows

$$\sigma_1(X_1, n_1, X_2, n_2) = \begin{cases} 1, & \text{if } Q_1(X_1, n_1) > Q_2(X_2, n_2), \\ 0, & \text{if } Q_1(X_1, n_1) < Q_2(X_2, n_2), \end{cases} \tag{19}$$

$\sigma_2(X_1, n_1, X_2, n_2) = 1 - \sigma_1(X_1, n_1, X_2, n_2)$. Probabilities $\sigma_1(X_1, n_1, X_2, n_2)$, $\sigma_2(X_1, n_1, X_2, n_2)$ can be arbitrarily chosen if $Q_1(X_1, n_1, s_1) = Q_2(X_2, n_2, s_2)$. A regret (13) is

$$L_N(\sigma, \lambda) = \iint_{\Theta} M|m_1 - m_2|\lambda(m_1, m_2)dm_1dm_2 \tag{20}$$
$$+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} L_{N-2M}(\lambda; X_1, M, X_2, M)p(\lambda; X_1, M, X_2, M)dX_1dX_2,$$

where $p(\lambda; X_1, M, X_2, M)$ is defined in (16). The first term in (20) describes the loss of income at the initial stage of the control when actions are applied by turn. The second term describes the loss of income at the remaining horizon. Note that if one has to determine a regret (4) then a degenerate prior distribution density, concentrated at a single parameter θ, should be chosen. In this case, all the posterior distribution densities will be degenerate, too.

Formulas (17)–(20), in principle, allow one to compute a regret but require a very large amount of computations. One can drastically reduce the amount of computations using the following properties of the UCB strategy.

Lemma 1. *Let the strategy σ be such that*

$$\sigma_\ell(X_1, n_1, X_2, n_2) = \sigma_\ell(X_1 + n_1 z, n_1, X_2 + n_2 z, n_2) \tag{21}$$

for all histories X_1, n_1, X_2, n_2, $\ell = 1, 2$, and for some fixed z. Then the following equality holds true

$$L_N(\sigma, \lambda(m_1, m_2)) = L_N(\sigma, \lambda(m_1 + z, m_2 + z)). \tag{22}$$

Lemma 2. *Put $\overline{\ell} = 3 - \ell$. Let the strategy σ be such that*

$$\sigma_\ell(X_1, n_1, X_2, n_2) = \sigma_{\overline{\ell}}(X_2, n_2, X_1, n_1), \tag{23}$$

for all histories X_1, n_1, X_2, n_2 and $\ell = 1, 2$. If $D_1 = D_2$ then the following equality holds true

$$L_N(\sigma, \lambda(m_1, m_2)) = L_N(\sigma, \lambda(m_2, m_1)). \tag{24}$$

Lemma 3. *A strategy σ defined by* (14) *satisfies* (21) *and* (23) *and for all n_1, n_2 depends only on current statistics (U, n_1, n_2), where $U = (X_1 n_2 - X_2 n_1)/n'$, $n' = n'_1 + n'_2$, $n'_1 = n_1/D_1$, $n'_2 = n_2/D_2$. For fixed n_1, n_2 function $\sigma_1(U, n_1, n_2)$ monotonously increases in U.*

To prove lemmas 1, 2, one has to use formulas (17)–(20). These proofs are similar to those presented in [14]. Lemma 3 follows from (14) and (19). For further it is convenient to change parametrization as follows: $m_1 = m + v$, $m_2 = m - v$. Denote by $\mu(m, v)$ a prior distribution density in the new variables. Then for strategies satisfying a condition of lemma 1 the equality $L_N(\sigma, \mu(m, v)) = L_N(\sigma, \mu(m+z, v))$ holds true for any fixed z. For strategies satisfying a condition of lemma 2 the equality $L_N(\sigma, \mu(m, v)) = L_N(\sigma, \mu(m, -v))$ holds true if $D_1 = D_2$.

Let us consider distribution density $\mu(m, v) = \rho(v)\beta(m)$, where $\beta(m)$ is an arbitrary density. Clearly, if σ is defined by (14) then the equality holds

$$L_N(\sigma, \rho(v)\beta(m)) = \int_{-\infty}^{\infty} L_N(\sigma, \rho(v)\delta(m - z))\beta(z)dz \tag{25}$$

where $\delta(m - z)$ is a Dirac delta function. Since, according to lemmas 1 and 3, all $L_N(\sigma, \rho(v)\delta(m - z))$ are equal to each other, it follows from (25) that $L_N(\sigma, \rho(v)\beta_1(m)) = L_N(\sigma, \rho(v)\beta_2(m)) = L_N(\sigma, \rho(v))$ for all $\beta_1(m), \beta_2(m)$. Here a notation $L_N(\sigma, \rho(v))$ emphasizes that a regret is determined only by a density $\rho(v)$. In what follows, we choose a prior distribution density in the form

$$\mu(m, v) = \rho(v)\kappa_a(m), \tag{26}$$

where $\kappa_a(m)$ is a uniform density on the segment $m \in [-a, a]$ and a is large enough.

For strategies $\{\sigma_\ell(U, n_1, n_2)\}$ and a prior distribution density (26), provided that $a \to +\infty$, in [13], theorem 5, a recursive dynamic programming equation is obtained in invariant form with a control horizon equal to one in the domain of "close" distributions. This equation uses much less calculations than (17)–(18). Let us present this equation. We need the following change of variables

$$\begin{gathered} C = cN^{-1/2}, w = N^{1/2}v, \varrho(w) = N^{-1/2}\rho(v), \\ x_\ell = X_\ell N^{-1/2}, u = UN^{-1/2}, \\ t_\ell = n_\ell N^{-1}, t_\ell^* = n_\ell^* N^{-1}, t'_\ell = n'_\ell N^{-1}, t = t_1 + t_2, t' = t'_1 + t'_2, \\ \varepsilon = MN^{-1}, \varepsilon_\ell^* = \varepsilon D_\ell,\ \varepsilon'_\ell = \varepsilon/D_\ell,\ \ell = 1, 2. \end{gathered}$$

It follows from (19) that $\sigma_\ell(X_1, n_1, X_2, n_2) = \sigma_\ell(x_1, t_1, x_2, t_2)$ for all histories (X_1, n_1, X_2, n_2) and $\ell = 1, 2$. Therefore, $\sigma_\ell(U, n_1, n_2) = \sigma_\ell(u, t_1, t_2)$ for all histories (U, n_1, n_2) and $\ell = 1, 2$. Let us put $D_g^2 = D_1 D_2$, $D_h^{-1} = 0.5(D_1^{-1} + D_2^{-1})$, $f_D(x) = f_D(x|0)$ and use a standard notation for convolution of functions $F(x) * G(x) = \int_{-\infty}^{\infty} F(x - y)G(y)dy$. The following equation must be solved recursively

$$l_\varepsilon(u, t_1, t_2) = \sigma_1(u, t_1, t_2)l_\varepsilon^{(1)}(u, t_1, t_2) + \sigma_2(u, t_1, t_2)l_\varepsilon^{(2)}(u, t_1, t_2), \tag{27}$$

backwards where $l_\varepsilon^{(1)}(u, t_1, t_2) = l_\varepsilon^{(2)}(u, t_1, t_2) = 0$ if $t_1 + t_2 = 1$ and then

$$\begin{gathered} l_\varepsilon^{(1)}(u, t_1, t_2) = \varepsilon g^{(1)}(u, t_1, t_2) \\ + l_\varepsilon(u, t_1 + \varepsilon, t_2) * f_{\varepsilon_1^* t_2^2 (t')^{-1}(t' + \varepsilon'_1)^{-1}}(u), \\ l_\varepsilon^{(2)}(u, t_1, t_2) = \varepsilon g^{(2)}(u, t_1, t_2) \\ + l_\varepsilon(u, t_1, t_2 + \varepsilon) * f_{\varepsilon_2^* t_1^2 (t')^{-1}(t' + \varepsilon'_2)^{-1}}(u), \end{gathered} \tag{28}$$

if $t_1 + t_2 < 1$, $t_1 \geq \varepsilon$ and $t_2 \geq \varepsilon$. A regret (13) is

$$L_N(\sigma; \rho(v)) = N^{1/2} l_\varepsilon(\sigma; \varrho(w)), \quad \text{where} \tag{29}$$
$$l_\varepsilon(\sigma; \varrho(w)) = \varepsilon \int_{-c}^{c} 2|w| \varrho(w) dw + \int_{-\infty}^{\infty} f_{0.5\varepsilon D_g^2 D_h}(u) l_\varepsilon(u, \varepsilon, \varepsilon) du.$$

Here

$$g^{(1)}(u, t_1, t_2) = \int_{-c}^{0} 2|w| g(w; u, t_1, t_2) \varrho(w) dw,$$
$$g^{(2)}(u, t_1, t_2) = \int_{0}^{c} 2w g(w; u, t_1, t_2) \varrho(w) dw,$$
$$g(w; u, t_1, t_2) = \exp\left(2D_g^{-2}(uw - w^2 t_1 t_2 (t')^{-1})\right).$$

Note that ε characterizes relative sizes of batches.

4 Numerical Results

In this section, we present some numerical results. In figure 1, we present normalized regrets for different control horizons. Thick solid lines 1, 2, 3 correspond to normalized regrets

$$l_N(\sigma, \theta) = (DN)^{-1/2} L_N(\sigma, \theta)$$

computed on the control horizons $N = 50$, 250, 1000. Thin dotted lines 4, 5, 6 correspond to normalized regrets on the same control horizons $N = 50$, 250, 1000 but computed without initial stage of the control when actions are applied by turn. Here $\theta = (d(D/N)^{1/2}, -d(D/N)^{1/2})$ with $D = 1$ a common variance. The step of changing d was always 0.3 from 0 to 7.5. Lines 1 and 4 were computed by both Monte-Carlo simulations (averaged over 400000 simulations) and analytically using recursive equation. Computed by these two methods lines turned out to be very close and visually coincide. However, the speed of the methods varies greatly. Monte-Carlo simulations were about 38 times faster. Since the operating time of Monte-Carlo simulations is approximately proportional to N and operating time of analytic method is approximately proportional to N^2, the difference in operating times of two methods grows significantly with growing N. That is why all other computations were performed using 400000 Monte-Carlo simulations.

In table 1, approximate values of the parameters of the optimal strategy for control horizons $N = 50$, 250, 1000 are presented. Here a^* are optimal values of a and d^* correspond to d at which normalized regrets attain their maximum values

$$r_N^M(\Theta) = (DN)^{-1/2} \max_\theta L_N(\sigma, \theta),$$

which are equal to normalized minimax risks (5). Note that the strategy is not very sensitive to the varying of a in the vicinity of a^*. If a changes by 0.1 then maximum of the regret changes approximately by 0.01.

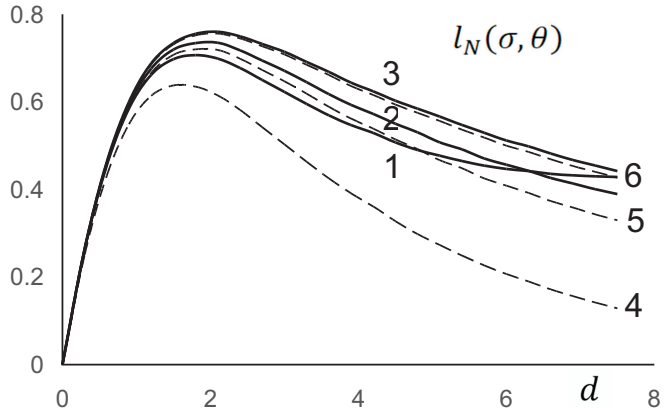

Fig. 1. Normalized regrets for different control horizons $N = 50$, 250, 1000.

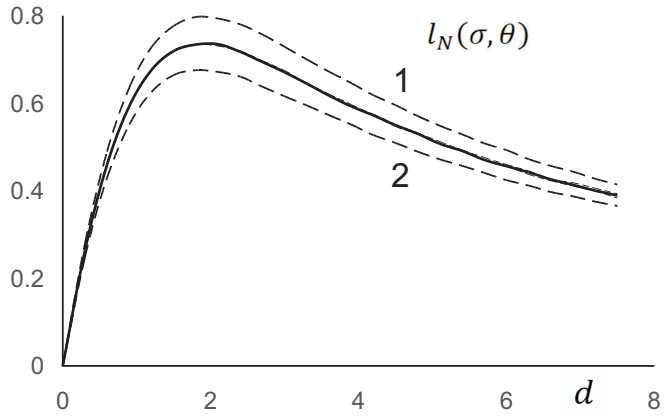

Fig. 2. Effect on the regret of inaccurate values of variances in the strategy.

Table 1. Parameters of the strategy.

N	a^*	d^*	$r_N^M(\Theta)$
50	0.94	1.8	0.71
250	1.0	2.1	0.69
1000	1.0	2.1	0.64

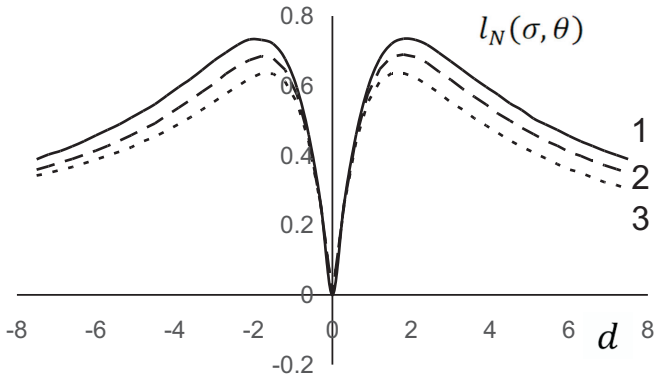

Fig. 3. Normalized regrets in the case of different variances, $N = 50$.

In figure 2, we present an effect on the regret of using inaccurate values of variances in the bounds (8) if $D_1 = D_2 = 1$ and $N = 250$. Thick solid line corresponds to the case of accurately indicated variances, i.e., $\gamma_1 = \gamma_2 = 1$. Thin dotted lines 1 and 2 correspond to the cases of $\gamma_1 = 0.95$, $\gamma_2 = 1.05$ and $\gamma_1 = 1.05$, $\gamma_2 = 0.95$ respectively. Thin dotted lines corresponding to the cases $\gamma_1 = \gamma_2 = 0.95$ and $\gamma_1 = \gamma_2 = 1.05$ are also presented in figure 2 but they almost coincide with the thick solid line. If sizes of batches are large enough, this means that estimates of the variances can be made at the initial stage, when actions are applied by turn, and then used for the control.

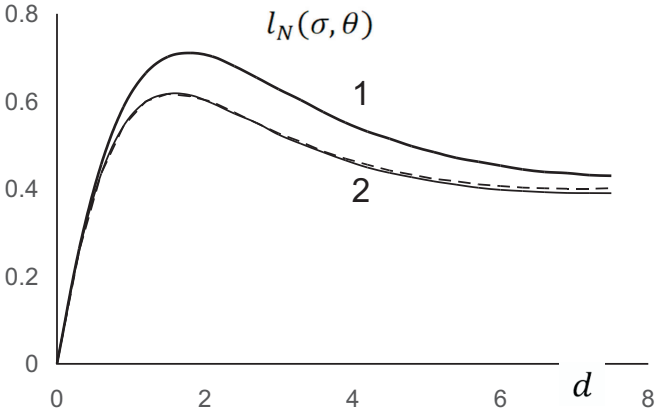

Fig. 4. Normalized regrets for batch processing of Bernoulli incomes.

Table 2. Parameters of the strategy in the case of different variances, $N = 50$.

γ_2	a_1^*	d_1^*	a_2^*	d_2^*	$r_N^M(\Theta)$
1	0.94	1.8	0.94	-1.8	0.71
0.75	0.90	1.8	0.95	-1.5	0.66
0.5	0.88	1.5	1.01	-1.5	0.61

In figure 3, we present the normalized regrets in the case of different variances and $N = 50$. Lines 1, 2, 3 correspond to $\gamma_2 = 1$, 0.75, 0.5 respectively. Approximate values of the parameters of the optimal UCB strategy are presented in table 2.

Finally, in figure 4, we present normalized regrets for Bernoulli incomes described by probability distribution (2). Here total number of data is $N = 5000$, batch size is $M = 100$, and, therefore, the number of control stages is $K = 50$. Parameter is as follows

$$\theta = (p_1, p_2) = \left(p + d(D/N)^{1/2}, p - d(D/N)^{1/2}\right),$$

where $D = 0.25$ is the maximum variance of one-step income attained at $p = 0.5$. In figure 4, a thick solid line 1 corresponds to the case of $p = 0.5$, this line is visually identical with line 1 in figure 1. Thin solid and dotted lines correspond to the cases of $p = 0.25$ and $p = 0.75$, which have equal variances. This lines are almost the same and provide less maximum regret than line 1. For all different p maximum regrets do not exceed that attained at $p = 0.5$.

5 Conclusion

We considered a batch version of a UCB strategy for controlling a Gaussian two-armed bandit and developed two approaches to invariant description of the control on the horizon equal to one. The first approach allows one to compute a regret using Monte-Carlo simulations. The second approach provides analytical method to determine a regret by solving a recursive Bellman-type dynamic programming equation. Although both approaches give the same results, Monte-Carlo simulations are much faster. Using Monte-Carlo simulations, we obtained optimal parameters of the UCB strategy which allow one to minimize the maximum regret, i.e., to ensure the minimax goal of the control for considered strategies. The results show the high effectiveness of considered UCB strategy.

6 Acknowledgements

This research was supported by RFBR, project number 20-01-00062.

References

1. Berry, D. A., Fristedt, B.: Bandit problems: sequential allocation of experiments. Chapman and Hall, London, New York (1985)

2. Presman, E.L., Sonin, I.M.: Sequential control with incomplete information. Academic, New York (1990)
3. Kolnogorov, A. V.: Robust parallel control in a random environment and data processing optimization. Automation and Remote Control. **75** (12) 2124–2134 (2014)
4. Kolnogorov, A.V.: On a limiting description of robust parallel control in a random environment. Automation and Remote Control. **76** (7) 1229–1241 (2015)
5. Lai, T. L., Levin, B., Robbins, H., Siegmund, D.: Sequential medical trials, Proc. Natl. Acad. Sci. USA. **77** (6) 3135–3138 (1980)
6. Perchet, V., Rigolletm P., Chassang, S., Snowberg, E.: Batched bandit problems. Annals of Statistics. **44** (2) 660–681 (2016)
7. Auer, P., Cesa-Bianchi, N., Fisher P.: Finite-time analysis of the multi-armed bandit problem. Machine Learning. **47** (2–3) 235–256 (2002)
8. Bather, J.A.: The minimax risk for the two-armed bandit problem. Mathematical Learning Models—Theory and Algorithms. Lect. Notes Statist. Springer, New York **20** 1–11 (1983)
9. Lattimore, T., Szepesvari, C.: Bandit algorithms. Cambridge University Press, Cambridge (2020)
10. Lugosi, G., Cesa-Bianchi, N.: Prediction, learning and games. Cambridge University Press, Cambridge (2006)
11. Lai, T. L.: Adaptive treatment allocation and the multi-armed bandit problem. Ann. Statist. **25** 1091–1114 (1987)
12. Vogel, W.: An asymptotic minimax theorem for the two-armed bandit problem. Ann. Math. Statist. **31** (2) 444–451 (1960)
13. Kolnogorov, A.V.: Gaussian two-armed bandit: limiting description. Problems of Information Transmission. **56** (3) 278–301 (2020)
14. Garbar, S.V., Kolnogorov, A.V.: Customization of J. Bather UCB strategy for a Gaussian multi-armed bandit. Mat. Teor. Igr Prilozh. **14** (2) 3–30 (2022) (Russian. English summary)