S©MR

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ
Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru
ISSN 1813-3304

Vol. 20, No. 2, pp. 144–145 (2023)
https://doi.org/10.33048/semi.2023.20.???

УДК 519.244, 519.83
MSC 62C10, 62L05, 91A35

INVARIANT DESCRIPTION OF CONTROL IN THE GAUSSIAN ONE-ARMED BANDIT PROBLEM

A.V. KOLNOGOROV

Communicated by P.P. Petrov

Abstract: We consider the one-armed bandit problem in the application to batch data processing if there are two alternative processing methods with different efficiencies and the efficiency of the second method is a priori unknown. During the processing, it is necessary to determine the most effective method and ensure its preferential use. Processing is performed in batches, so the distributions of incomes are Gaussian. We consider the case of a priori unknown mathematical expectation and variance of income corresponding to the second action. In contrast to the case of known variance, which is considered if the number of batches and the number of data in them are large, this case describes a situation when the batches themselves and their number have moderate or small volumes. We obtain recursive equations for computing the Bayesian risk and the regret, which we then present in an invariant form with a control horizon equal to one. This allows one to obtain the estimates of Bayesian and minimax risk that are valid for all control horizons multiples to the the number of processed batches.

Keywords: one-armed bandit, Bayesian and minimax approaches, invariant description, main theorem of the game theory, batch processing.

Kolnogorov A.V., Invariant description of control in the Gaussian one-armed bandit problem.

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

Received January, 1, 2023, Published December, 31, 2023.

1 Introduction

We consider the one-armed bandit problem, which is a special case of the two-armed bandit (see, e.g., [1, 2]). The name originates from a slot machine with two arms, each choice of which is accompanied by a random income of the player. The player has to play N times against the slot and his/her goal is to maximize the total expected income. Distributions of incomes depend only on the currently selected arms, are fixed during the game but the player does not have a complete a priori information about them. In particular, a one-armed bandit occurs if the characteristics of only one of the arms are a priori known; below we assume that this is the first arm. In what follows, the arms are called actions. So, in order to achieve the goal, during the game, it is necessary to determine a more profitable action and ensure its preferential use. The problem has numerous applications in behavior modelling [3, 4, 5], adaptive control in a random environment [6, 7], medicine, internet technologies, data processing [8, 9].

Formally, Gaussian one-armed bandit is a controlled random process ξ_n, $n = 1, 2, \ldots, N$, which values are interpreted as incomes, depend only on the currently selected actions y_n ($y_n \in \{1, 2\}$) and in the case of choosing the second action (i.e., $y_n = 2$) have a Gaussian distribution density

$$f_D(x|m) = (2\pi D)^{1/2} \exp\left(-(x-m)^2/(2D)\right), \tag{1.1}$$

where $m = \mathbf{E}(\xi_n|y_n = 2)$, $D = \mathbf{D}(\xi_n|y_n = 2)$ are the mathematical expectation and the variance of one-step income provided that the second action is chosen. In the case of choosing the first action, the mathematical expectation m_1 is known and, without loss of generality, is zero (otherwise, one can consider the process $\xi_n - m_1$, $n = 1, 2, \ldots, N$). The knowledge of the variance $D = \mathbf{D}(\xi_n|y_n = 1)$ is not required because it does not affect the achievement of the control goal. So, considered one-armed bandit is completely described by the parameter $\theta = (m, D)$, which value is assumed to be unknown. However, the set of parameters Θ is known, which has the form

$$\Theta = \{(m, D) : |m| \le C,\ \underline{D} \le D \le \overline{D}\},$$

where $0 < C < \infty$, $0 < \underline{D} < \overline{D} < \infty$.

A control strategy σ determines, in general, the random choice of the action y_{n+1} at the time point $n + 1$ depending on the entire known history of the process. However, instead of the whole history, one can use sufficient statistics. Let's assume that by the time point n the first and the second actions have been applied n_1 and n_2 times respectively ($n = n_1 + n_2$). Sufficient statistics are the total income for the application of the second action and the s^2-statistics

$$X = \sum_{i:y_i=2} \xi_i, \quad S = \sum_{i:y_i=2} x_i^2 - X^2/n_2.$$

Note that $X = S = 0$ for $n_2 = 0$ and $S = 0$ for $n_2 = 1$. The total income and s^2-statistics for the application of the first action are not required because

corresponding mathematical expectation of a one-step income is known. Control strategies can be subject to restrictions. In sections 2 and 3 we consider strategies for batch data processing. In particular, UCB strategies are discussed.

Let's define a regret. If the parameter was known then one should always choose the action corresponding to the larger of the mathematical expectations of the one-step incomes 0 and m. The total expected income would thus be $N\max(0,m)$. In the case of choosing the strategy σ, the total expected income is less than the maximum one by an amount

$$L_N(\sigma,\theta) = N\max(0,m) - \mathbf{E}_{\sigma,\theta}\left(\sum_{n=1}^{N}\xi_n\right), \tag{1.2}$$

which is called a regret and is caused by incomplete information. Here $\mathbf{E}_{\sigma,\theta}$ is a sign of the mathematical expectation with respect to the measure generated by the strategy σ and the parameter θ. Note that the regret for the shifted process $\{\xi_n - m_1\}$ is the same as for the original $\{\xi_n\}$.

Let's explain the choice of a normal distribution of incomes. We consider the problem in the application to batch data processing if two alternative methods with different efficiencies can be used for processing and the efficiency of the first method is a priori known. In this case, processing methods correspond to actions, efficiencies characterize, for example, the probabilities of successful data processing, and the goal of the control is to maximize the mathematical expectation of the total number of successfully processed data. In batch processing considered below, the data is divided into equal batches, the same action (processing method) is applied to all the data in the batch and the total incomes in the batches are used for the control. By virtue of the central limit theorem, these incomes, under broad assumptions, have approximately Gaussian distributions if the batch sizes are large enough.

This approach has some advantages. First, control strategies become universal, i.e., the same for a wide class of controlled processes, whose one-step incomes obey the central limit theorem. Second, if it is possible to organize parallel processing of batch data, this approach can significantly reduce the total processing time (by M times if the batch size is M). And an important property of this approach in optimizing big data processing is that it almost does not increase the maximum regret if the number of batches, into which the data is divided, is large enough. For example, it is shown in [10] that in the case of splitting data into 50 batches, the maximum regret grows by only 3% compared to its maximum value if the number of batches grows infinitely.

Note that first batch processing was offered for the treatment of patients with alternative drugs. Since it takes a considerable time for the result of treatment to manifest itself, patients cannot be treated one at a time. Therefore, it was proposed to first give all the drugs to sufficiently large test groups, and then, according to the results of testing, the best drug to all

remaining patients. In this case, the size of the test groups is much smaller than the size of the main one, and the number of stages at which testing is performed is small (one and rarely two or three stages), since patients cannot wait long. For an overview of the results of this approach and references see, e.g., [11, 12].

Let a prior distribution density $\lambda(\theta) = \lambda(m, D)$ be given on the set Θ. We assume that the conditions are met

$$\int_\Theta m^- \lambda(\theta) d\theta > 0, \qquad \int_\Theta m^+ \lambda(\theta) d\theta > 0,$$

otherwise, the more profitable action is a priori known. We use here the standard notations $m^+ = \max(m, 0)$, $m^- = \max(-m, 0)$. The averaged regret is

$$L_N(\sigma, \lambda) = \int_\Theta L_N(\sigma, \theta) \lambda(\theta) d\theta \tag{1.3}$$

and Bayesian risk is defined as

$$R_N^B(\lambda) = \inf_{\{\sigma\}} L_N(\sigma, \lambda), \tag{1.4}$$

the corresponding optimal strategy σ^B is called a Bayesian strategy. Minimax risk is defined as follows

$$R_N^M(\Theta) = \inf_{\{\sigma\}} \sup_{\Theta} L_N(\sigma, \theta), \tag{1.5}$$

the corresponding optimal strategy σ^M is called a minimax strategy.

Bayesian approach is very popular because it allows one to find Bayesian strategy and Bayesian risk for any prior distribution by solving the recursive equation. The disadvantage of the Bayesian approach is the lack of clear criteria for choosing this prior distribution. The advantage of the minimax approach is its robustness, i.e., the fulfillment of the inequality $L_N(\sigma, \theta) \leq R_N^M(\Theta)$ for all $\theta \in \Theta$. Besides, as mentioned above, in the case of batch processing, the minimax approach provides the value of minimax risk close to optimal even with a relatively small number of batches (see, e.g., [10, 12]).

The disadvantage of the minimax approach is the lack of a direct method for finding minimax risk and minimax strategy. On the other hand, an asymptotic estimate of the minimax risk having the order of $N^{1/2}$ is well-known [13]. Note that the maximum values of the regret are achieved in the domain of close distributions, which is characterized by the fact that the difference in mathematical expectations of one-step incomes has the order of $N^{-1/2}$. This is because it is impossible to reliably (i.e., with probability arbitrarily close to 1) statistically determine a more profitable action in this domain. To find minimax risk and minimax strategy, the main theorem of the game theory can be used, according to which the minimax risk is equal to the Bayesian one computed relative to the worst-case prior distribution on which the Bayesian risk reaches its maximum. And the minimax strategy coincides

with the corresponding Bayesian one. In more detail, close distributions and using the main theorem of the game theory are discussed in [14].

The one-armed bandit problem was first considered in [15] in the Bayesian setting for a Bernoullian two-armed bandit which incomes take the values 0 and 1. In [15], a recursive algorithm for finding Bayesian strategy and Bayesian risk is described. The asymptotic properties of the Bayesian strategy for a Bernoullian two-armed bandit were established in [16]. In [15], the following intuitively clear property of the Bayesian strategy was proved: since the application of the first action does not provide additional information, once selected, it will be applied until the end of the control. This property also holds true in the case of a Gaussian one-armed bandit with one or two unknown parameters (see [10, 17, 18]). It also remains true in the statements considered in sections 2 and 3; the proofs are similar to presented in [10, 18] and are therefore omitted.

Let's indicate what is the difference between the considered approach and the one presented in [10, 17]. In [10, 17], the case of a priori known variance is considered. This approach is reasonable if the amount of data is large. Then the variance can be estimated when processing the first batch. Since regret changes little with a small change in variance, the obtained estimate can be used for control. But if the amount of data is moderate or small, then the variance estimation should be carried out in the control process. This is exactly the approach presented below. An analogy with obtaining an interval estimate of the unknown mean of a normal distribution may be useful here. If the sample is large, then a normal distribution is used for this estimate, where the unknown variance is replaced by its point estimate. But if the sample is small or moderate, then one has to use the Student's distribution.

In Bayesian setting, a one-armed bandit with unknown mathematical expectation and variance was considered in [18], where recursive equations for computing the Bayesian risk and the regret were obtained. In [18], the estimate of the variance is updated using incomes received after processing the whole batches. Here we consider also the case, when the estimate of the variance is updated using incomes received when processing data within batches. In both cases, we present recursive equations for calculating Bayesian risk and regret in more computationally convenient ordinary and invariant forms. The invariant forms describe a control on the unit horizon and are obtained using the change of variables. The advantage of the invariant description is that it is the same for all batch sizes and, hence, makes it possible to obtain asymptotic estimates of Bayesian risk and regret.

The rest of the article is as follows. Section 2 describes batch processing, in which the variance estimate is updated based on the income received after processing the whole batch. Section 3 describes batch processing, in which the variance estimate is updated based on the incomes received when processing data within the batch. In both sections 2 and 3, recursive equations are obtained for finding Bayesian strategies, risks and regrets in the ordinary

and invariant forms. Section 4 presents the results of numerical experiments. Section 5 contains the conclusion.

2 Estimating the variance by cumulative incomes in batches

In this section, we consider batch processing of the source data $\{\xi_n\}$. Let the total number of processed data be $N = MK$, where M is the batch size, K is the number of batches. In this case, the same action is applied to each M sequentially incoming data. The income received as a result of processing the kth batch is

$$\zeta_k = \sum_{n=(k-1)M+1}^{kM} \xi_n,$$

In the case of applying the second action, the mathematical expectation of income for processing the batch of data is Mm, the variance is MD. The mathematical expectation of income for the application of the first action is still 0. The batches processed by the first and the second action are numbered k_1, k_2 $(k = k_1 + k_2)$.

Let x_i, $i = 1, \ldots, k_2$, be incomes obtained in response to the application of the second action. The following sufficient statistics can be used in considered case

$$X = \sum_{i=1}^{k_2} x_i, \quad S = \sum_{i=1}^{k_2} x_i^2 - X^2/k_2,$$

where X and S are current values of cumulative income and s^2-statistics for the application of the second action. Note that $X = 0$ if $k_2 = 0$ and $S = 0$ if $k_2 = 0, 1$. Cumulative income and s^2-statistics for the application of the first action are not required because corresponding mathematical expectation of one-step income is known.

Let's consider how to update X and S. Assume that $k_2 \geq 1$, and let $x_{k_2+1} = Y$ be a new income. Then

$$X_{new} = \sum_{i=1}^{k_2+1} x_i = X + Y, \quad S_{new} = \left(\sum_{i=1}^{k_2+1} x_i^2\right) - X_{new}^2/(k_2+1)$$
$$= \left(\sum_{i=1}^{k_2} x_i^2\right) + Y^2 - (X+Y)^2/(k_2+1) = S + M\Delta(X, k_2, Y),$$

where

$$M\Delta(X, k_2, Y) = X^2/k_2 + Y^2 - (X+Y)^2/(k_2+1) = \frac{(X - k_2 Y)^2}{k_2(k_2+1)}.$$

If $k_2 = 0$ then $M\Delta(0,0,Y) = Y^2 - Y^2 = 0$. Hence,

$$\Delta(X,k_2,Y) = \begin{cases} 0, & if\ k_2 = 0, \\ \dfrac{(X-k_2Y)^2}{Mk_2(k_2+1)}, & if\ k_2 \geq 1. \end{cases} \tag{2.1}$$

Therefore, X, S are updated according to the rule

$$X \leftarrow X + Y, \quad S \leftarrow S + M\Delta(X,k_2,Y), \tag{2.2}$$

where $\Delta(X,k_2,Y)$ is given by (2.1).

Given a prior distribution density $\lambda(m.D)$, let's describe a posterior distribution density. Denote by $\chi^2_k(x)$ a chi-squared distribution density with k degrees of freedom

$$\chi^2_k(x) = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1}e^{-\frac{x}{2}}, \quad k \geq 1,$$

and introduce the functions

$$\begin{gathered} f_{kMD}\left(X|kMm\right) \\ = \begin{cases} 1, & \text{if } k = 0, \\ f_{kMD}\left(X|kMm\right), & \text{with } f\left(\cdot\right) \text{ from (1.1) if } k \geq 1, \end{cases} \\ \psi_{k-1}\left((MD)^{-1}S\right) \\ = \begin{cases} 1, & \text{if } k = 0,1, \\ (MD)^{-1}\chi^2_{k-1}((MD)^{-1}S), & \text{if } k \geq 2, \end{cases} \end{gathered} \tag{2.3}$$

Note that defined above cumulative income X and s^2-statistics S after processing k batches have exactly the distribution densities described by (2.3) for $k \geq 1$ and $k \geq 2$ respectively. Since X and S are independent random variables, the posterior distribution density is

$$\lambda(m,D|X,S,k_2) = \frac{f_{k_2MD}\left(X|k_2Mm\right)\psi_{k_2-1}\left((MD)^{-1}S\right)\lambda(m.D)}{P(X,S,k_2)},$$

where

$$P(X,S,k_2) = \iint_\Theta f_{k_2MD}\left(X|k_2Mm\right)\psi_{k_2-1}\left((MD)^{-1}S\right)\lambda(m.D)dmdD.$$

Note that definition (2.3) ensures that this formula remains valid for all $k_2 = 0,1,2,\ldots$ These approach to definition of the posterior distribution is used in [18]. However, recursive equation is simpler if the posterior distribution is defined by the following equivalent way. Denote

$$\begin{gathered} \tilde{\mathbf{F}}(X,S,k|m,D) \\ = \begin{cases} 1, & if\ k = 0, \\ D^{-1/2}\tilde{f}_{kMD}(X|kMm), & if\ k = 1, \\ D^{-3/2}\tilde{f}_{kMD}(X|kMm)\tilde{\psi}_{k-1}\left(S/(MD)\right), & if\ k \geq 2, \end{cases} \end{gathered} \tag{2.4}$$

where

$$\tilde{f}_{kMD}\,(X|kMm) = \begin{cases} 1, & \text{if } k=0, \\ \exp\left(-(X-kMm)^2/(2kMD)\right)., & \text{if } k\geq 1, \end{cases}$$
$$\tilde{\psi}_{k-1}\left(\frac{S}{MD}\right) = \begin{cases} 1, & \text{if } k=0,1, \\ \left(\frac{S}{MD}\right)^{\frac{k-1}{2}-1}\exp\left(-\frac{S}{2MD})\right), & \text{if } k\geq 2, \end{cases} \tag{2.5}$$

Clearly, the posterior distribution density is

$$\lambda(m,D|X,S,k_2) = \frac{\tilde{\mathbf{F}}(X,S,k_2|m,D)\lambda(m,D)}{\tilde{P}(X,S,k_2)}, \qquad \text{with}\quad \tilde{P}(X,S,k_2) = \iint_{\Theta}\tilde{\mathbf{F}}(X,S,k_2|m,D)\lambda(m,D)dmdD. \tag{2.6}$$

Note that (2.6) remains valid if $k_2=0$ and $k_2=1$, too.

Let $R^B(k_1,X,S,k_2)$ denote a Bayesian risk on the remaining control horizon $k+1,\dots,K$ computed with respect to the posterior distribution $\lambda(m,D|X,S,k_2)$, i.e., $R^B(k_1,X,S,k_2) = R^B_{K-k}(\lambda(m,D|X,S,k_2))$, $k=k_1+k_2$. A standard recursive equation for computing a Bayesian risk is as follows

$$R^B(k_1,X,S,k_2) = \min\left(R^B_1(k_1,X,S,k_2), R^B_2(k_1,X,S,k_2)\right), \tag{2.7}$$

where $R^B_1(k_1,X,S,k_2) = R^B_2(k_1,X,S,k_2) = 0$ if $k=K$, and

$$R^B_1(k_1,X,S,k_2) = \iint_{\Theta}\lambda(m,D|X,S,k_2) \times \left(Mm^+ + R^B(k_1+1,X,S,k_2)\right)dmdD$$
$$= M\iint_{\Theta} m^+\lambda(m,D|X,S,k_2)dmdD + R^B(k_1+1,X,S,k_2), \tag{2.8}$$
$$R^B_2(k_1,X,S,k_2) = \iint_{\Theta}\lambda(m,D|X,S,k_2) \times\left(Mm^- + \int_{-\infty}^{\infty} R^B(k_1,X+Y,S+M\Delta,k_2+1)f_{MD}(Y|Mm)dY\right)dmdD,$$

if $0\leq k\leq K-1$. In the second equality (2.8), we used (2.1)–(2.2). One can see that $R^B_\ell(k_1,X,S,k_2)$ is equal to the loss of cumulative expected income at the remaining control horizon $k+1,\dots,K$ if at first the ℓth action was chosen and then the control was optimally performed. Bayesian strategy prescribes, when processing the batch with the number $k+1$, to choose an action corresponding to the smaller of the current values $R^B_1(k_1,X,S,k_2)$, $R^B_2(k_1,X,S,k_2)$. In the case of a draw, the choice can be arbitrary. Bayesian risk (1.4) is

$$R^B_N(\lambda) = R^B(0,0,0,0). \tag{2.9}$$

Let's present another form of recursive equation which is more convenient for computations. We put

$$R_\ell(k_1,X,S,k_2) = R^B_\ell(k_1,X,S,k_2)\times\tilde{P}(X,S,k_2), \quad \ell=1,2, \tag{2.10}$$

where $\tilde{P}(X,S,k_2)$ is defined in (2.6). Note that $R(k_1,X,S,k_2)$ $= \min(R_1(k_1,X,S,k_2),R_2(k_1,X,S,k_2)) = R^B(k_1,X,S,k_2)\times\tilde{P}(X,S,k_2)$ as well.

Theorem 1. *In order to determine the Bayesian risk, one should solve the following recursive equation*

$$R(k_1,X,S,k_2)=\min\left(R_1(k_1,X,S,k_2),R_2(k_1,X,S,k_2)\right), \tag{2.11}$$

where $R_1(k_1,X,S,k_2)=R_2(k_1,X,S,k_2)=0$ *if* $k=K$ *and*

$$\begin{gathered} R_1(k_1,X,S,k_2)=MG_1(X,S,k_2)+R(k_1+1,X,S,k_2),\\ R_2(k_1,X,S,k_2)=MG_2(X,S,k_2)\\ +\int_{-\infty}^{\infty}R(k_1,X+Y,S+M\Delta(X,k_2,Y),k_2+1)H(X,S,k_2,Y)dY, \end{gathered} \tag{2.12}$$

if $0\le k\le K-1$. *Here* $\Delta(X,k_2,Y)$ *is given by* (2.1),

$$\begin{aligned} G_1(X,S,k_2)&=\iint_{\Theta}m^{+}\tilde{\mathbf{F}}(X,S,k_2|m,D)\lambda(m,D)dmdD,\\ G_2(X,S,k_2)&=\iint_{\Theta}m^{-}\tilde{\mathbf{F}}(X,S,k_2|m,D)\lambda(m,D)dmdD, \end{aligned} \tag{2.13}$$

and

$$\begin{aligned} &H(X,S,k,Y)\\ &=\begin{cases} \dfrac{1}{(2\pi M)^{1/2}}, & if\ k=0,\\ \left(\dfrac{\Delta(X,k,Y)}{2\pi M}\right)^{1/2}, & if\ k=1,\\ \dfrac{1}{(2\pi)^{1/2}}\times\dfrac{S^{(k-1)/2-1}}{(S+M\Delta(X,k,Y))^{k/2-1}}, & if\ k\ge 2. \end{cases} \end{aligned} \tag{2.14}$$

When processing the batch bumber $k+1$ *Bayesian strategy prescribes to choose the action corresponding to the smaller value of* $R_1(k_1,X,S,k_2)$, $R_2(k_1,X,S,k_2)$; *in the case of a draw the choice can be arbitrary. Bayesian risk* (1.4) *is*

$$R_N(\lambda)=R(0,0,0,0). \tag{2.15}$$

Proof. Let's multiply (2.7)–(2.8) by $\tilde{P}(X,S,k_2)$ defined in (2.6). We obtain (2.11)–(2.12) with $G_1(X,S,n_2)$, $G_2(X,S,k_2)$ defined in (2.13). Let Δ in (2.16)–(2.18) below be given by (2.1). Denote $D'=MD$, $m'=Mm$. The function $H(X,S,k,Y)$ is

$$\begin{aligned} H(X,S,k,Y)&=\frac{\iint_{\Theta}\tilde{\mathbf{F}}(X,S,k|m,D)f_{D'}(Y|m')\lambda(m,D)dmdD}{\tilde{P}(X+Y,S+M\Delta,k+1)}\\ &=\frac{\tilde{\mathbf{F}}(X,S,k|m,D)f_{D'}(Y|m')}{\tilde{\mathbf{F}}(X+Y,S+M\Delta,k+1|m,D)}. \end{aligned} \tag{2.16}$$

Therefore, for $k \geq 2$ we have

$$H(X,S,k,Y) = \frac{\tilde{f}_{kD'}(X|km')f_{D'}(Y|m')}{\tilde{f}_{(k+1)D'}(X+Y|(k+1)m')} \times \frac{\tilde{\psi}_{k-1}(S/D')}{\tilde{\psi}_k((S+M\Delta)/D')}.$$

Since

$$\frac{\tilde{f}_{kD'}(X|km')f_{D'}(Y|m')}{\tilde{f}_{(k+1)D'}(X+Y|(k+1)m')} = \left(\frac{1}{2\pi D'}\right)^{1/2} \times \exp\left(-\frac{M\Delta}{2D'}\right) \tag{2.17}$$

and

$$\begin{aligned}\frac{\tilde{\psi}_{k-1}(S/D')}{\tilde{\psi}_k((S+M\Delta)/D')} &= \frac{(S/D')^{(k-1)/2-1}}{((S+M\Delta)/D')^{k/2-1}} \times \frac{\exp\left(-S/(2D')\right)}{\exp\left(-(S+M\Delta)/(2D')\right)} \\ &= (D')^{1/2} \times \frac{S^{(k-1)/2-1}}{(S+M\Delta)^{k/2-1}} \times \exp\left(\frac{M\Delta}{2D'}\right), \end{aligned} \tag{2.18}$$

it follows from (2.16)–(2.18) that $H(X,S,k,Y)$ is given by (2.14) for $k \geq 2$. If $k=1$ then $S=0$, $\tilde{\psi}_{k-1}(S|D')=1$ and according to (2.16)

$$\begin{aligned} H(X,0,1,Y) &= \left.\frac{D^{-1/2}\tilde{f}_{kD'}(X|km')f_{D'}(Y|m')}{D^{-3/2}\tilde{f}_{(k+1)D'}(X+Y|(k+1)m')\tilde{\psi}_1(M\Delta/D')}\right|_{k=1} \\ &= \left(\frac{1}{2\pi D'}\right)^{1/2} \times \exp\left(-\frac{M\Delta}{2D'}\right) \times \frac{D}{\tilde{\psi}_1(M\Delta/D')} \\ &= \left(\frac{1}{2\pi D'}\right)^{1/2} \times \exp\left(-\frac{M\Delta}{2D'}\right) \times \frac{D}{(M\Delta/D')^{1/2-1}\exp(-M\Delta/(2D'))} \\ &= \left(\frac{\Delta}{2\pi M}\right)^{1/2} \end{aligned}$$

and this is given by (2.14) for $k=1$. If $k=0$ then $X=0$, $S=0$, $\tilde{f}_{kD'}(X|km')=1$, $\tilde{\psi}_{k-1}(S/D')=\tilde{\psi}_k((S+M\Delta)/D')=1$ and according to (2.16)

$$H(0,0,0,Y) = \frac{f_{D'}(Y|m')}{D^{-1/2}\tilde{f}_{D'}(Y|m')} = \frac{1}{(2\pi M)^{1/2}},$$

and this corresponds to (2.14) for $k=0$. Formula (2.15) follows from (2.9) and the fact that $\tilde{P}(0,0,0)=1$. □

Let's present a recursive equation for computing the regret (1.3). The control strategy σ is described by a set of probabilities

$$\sigma_\ell(k_1,X,S,k_2) = \Pr(y_{k+1}=\ell|k_1,X,S,k_2),$$

$\ell = 1,2$; $k_1+k_2=k$, $k=0,\dots,K-1$; $X \in (-\infty,+\infty)$, $S \in (0,+\infty)$. Denote by $L^B(k_1,X,S,k_2)$ the regret on the last $K-k$ steps computed with

respect to the density $\lambda(m|X,S,k_2)$. This regret can be computed recursively by solving the following standard Bellman-type equation

$$L^B(k_1,X,S,k_2)=\sum_{\ell=1}^{2}\sigma_\ell(k_1,X,S,k_2)L^B_\ell(k_1,X,k_2), \tag{2.19}$$

where $L^B_1(k_1,X,S,k_2)=L^B_2(k_1,X,S,k_2)=0$ if $k=K$ and

$$\begin{aligned} L^B_1(k_1,X,S,k_2)&=\iint_\Theta \lambda(m,D|X,S,k_2)\\ &\quad\times\left(Mm^+ + L^B(k_1+1,X,S,k_2)\right)dmdD\\ &= M\iint_\Theta m^+\lambda(m,D|X,S,k_2)dmdD + L^B(k_1+1,X,S,k_2), \end{aligned} \tag{2.20}$$

$$\begin{aligned} L^B_2(k_1,X,S,k_2)&=\iint_\Theta \lambda(m,D|X,S,k_2)\\ &\quad\times\left(Mm^- + \int_{-\infty}^{\infty} L^B(k_1,X+Y,S+M\Delta,k_2+1)f_{MD}(Y|Mm)dY\right)dmdD, \end{aligned}$$

if $0\le k\le K-1$. In equation (2.19)–(2.20), $L^B_\ell(k_1,X,k_2)$ stands for cumulative loss of expected income on the remaining control horizon $K-k$ if the ℓth action is chosen first and then control is performed according to strategy σ. A regret (1.3) is

$$L^B_N(\sigma,\lambda)=L^B(0,0,0,0). \tag{2.21}$$

Note that to compute the regret (1.2), a prior density $\lambda(m,D)$ must be considered degenerate and concentrated at the point $\theta=(m,D)$. Let's present a more computationally convenient form of the equation (2.19)–(2.20) for computing the regret (1.2). Let's put

$$L_\ell(k_1,X,S,k_2)=L^B_\ell(k_1,X,S,k_2)]\tilde P(X,S,k_2),\quad \ell=1,2.$$

where $\tilde P(X,S,k_2)$ is defined in (2.6). Similarly to theorem 1, the following theorem holds true.

Theorem 2. *Consider a recursive equation*

$$L(k_1,X,S,k_2)=\sum_{\ell=1}^{2}\sigma_\ell(k_1,X,S,k_2)L_\ell(k_1,X,S,k_2), \tag{2.22}$$

where $L_1(k_1,X,k_2)=L_2(k_1,X,k_2)=0$ *if* $k=K$ *and*

$$\begin{aligned} L_1(k_1,X,S,k_2)&=MG_1(X,S,k_2)+L(k_1+1,X,S,k_2),\\ L_2(k_1,X,S,k_2)&=MG_2(X,S,k_2)\\ &+\int_{-\infty}^{\infty}L(k_1,X+Y,S+M\Delta(X,k_2,Y),k_2+1)H(X,S,k_2,Y)dY, \end{aligned} \tag{2.23}$$

if $0 \le k \le K-1$. *Here*

$$G_1(X,S,k_2) = m^+\tilde{\mathbf{F}}(X,S,k_2|m,D),$$
$$G_2(X,S,k_2) = m^-\tilde{\mathbf{F}}(X,S,k_2|m,D),$$

$H(X,S,k,Y)$ *is given by* (2.14). *Then a regret* (1.2) *is*

$$L_N(\sigma,\theta) = L(0,0,0,0). \tag{2.24}$$

Proof. The proof is similar to the proof of theorem 1. One should transform equation (2.19)–(2.20) under assumption that $\lambda(m,D)$ is degenerate and concentrated at the point $\theta = (m,D)$. □

Let's give an invariant form of formulas (2.11)–(2.15). We choose the following set of parameters $\Theta_N = \{(m,D) : \underline{D} \le D \le \overline{D}, |m| \le c(D/N)^{1/2}\}$, where $c > 0$, $0 < \underline{D} \le D \le \overline{D} < \infty$. If we put $D = \beta\overline{D}$, $m = \alpha(\overline{D}/N)^{1/2} = \alpha(\beta^{-1}D/N)^{1/2}$, then it takes the form $\Theta_N = \{(\alpha,\beta) : \underline{D}/\overline{D} = \beta_0 \le \beta \le 1, |\alpha| \le c\beta^{1/2}\}$.

Consider the change of variables $X = x(\overline{D}N)^{1/2}$, $Y = y(\overline{D}N)^{1/2}$, $S = s\overline{D}M$, $k = tK$, $k_1 = t_1K$, $k_2 = t_2K$, $M/N = K^{-1} = \varepsilon$, $m = \alpha(\overline{D}/N)^{1/2}$, $D = \beta\overline{D}$, $\lambda(m,D) = (N/\overline{D}^{\,3})^{1/2}\varrho(\alpha,\beta)$. Let

$$\begin{aligned} R_\ell(k_1,0,0,0) &= (\overline{D}N)^{1/2} r_\ell(t_1,0,0,0), \\ R_\ell(k_1,X,0,1) &= (\overline{D}\,N)^{1/2}(\overline{D})^{-1/2} r_\ell(t_1,x,0,\varepsilon), \\ R_\ell(k_1,X,S,k_2) &= (\overline{D}\,N)^{1/2}(\overline{D})^{-3/2} r_\ell(t_1,x,s,t_2), \text{ if } k_2 \ge 2, \end{aligned} \tag{2.25}$$

$\ell = 1,2$. The following theorem is valid.

Theorem 3. *To find the Bayesian risk, the recursive equation should be solved*

$$r(t_1,x,s,t_2) = \min\left(r_1(t_1,x,s,t_2), r_2(t_1,x,s,t_2)\right), \tag{2.26}$$

where $r_1(t_1,x,s,t_2) = r_2(t_1,x,s,t_2) = 0$ *if* $t = 1$ *and*

$$\begin{aligned} r_1(t_1,x,s,t_2) &= \varepsilon g_1(x,s,t_2) + r(t_1+\varepsilon,x,s,t_2), \\ r_2(t_1,x,s,t_2) &= \varepsilon g_2(x,s,t_2) \\ &+ \int_{-\infty}^{\infty} r(t_1,x+y,s+\delta(x,t_2,y),t_2+\varepsilon)h(x,s,t_2,y)dy, \end{aligned} \tag{2.27}$$

if $0 \le t \le 1-\varepsilon$. *Here*

$$\begin{aligned} g_1(x,s,t_2) &= \iint_{\Theta_N} \alpha^+\tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta)\varrho(\alpha,\beta)d\alpha d\beta, \\ g_2(x,s,t_2) &= \iint_{\Theta_N} \alpha^-\tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta)\varrho(\alpha,\beta)d\alpha d\beta, \end{aligned} \tag{2.28}$$

with

$$\tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta) = \begin{cases} 1, & if\ t_2 = 0, \\ \beta^{-1/2}\tilde{f}_{t_2\beta}(x|t_2\alpha), & if\ t_2 = \varepsilon, \\ \beta^{-3/2}\tilde{f}_{t_2\beta}(x|t_2\alpha)\tilde{\psi}_{k_2-1}(s/\beta), & if\ t_2 \ge 2\varepsilon, \end{cases}$$

$$h(x,s,t,y)=\begin{cases}\dfrac{1}{(2\pi\varepsilon)^{1/2}}, & if\ k=0,\\ \left(\dfrac{\delta(x,t,y))}{2\pi\varepsilon}\right)^{1/2}, & if\ k=1,\\ \dfrac{1}{(2\pi\varepsilon)^{1/2}}\times\dfrac{s^{(k-1)/2-1}}{(s+\delta(x,t,y))^{k/2-1}}, & if\ k\geq 2,\end{cases}\tag{2.29}$$

and

$$\delta(x,t_2,y)=\begin{cases}0, & if\ t_2=0,\\ \dfrac{(\varepsilon x-t_2y)^2}{\varepsilon t_2(t_2+\varepsilon)}, & if\ t_2\geq\varepsilon.\end{cases}\tag{2.30}$$

At the point of time $t+\varepsilon$ (or, equivalently, $k+1$) Bayesian strategy prescribes to choose the action corresponding to the smaller value of $r_1(t_1,x,s,t_2)$, $r_2(t_1,x,s,t_2)$; in the case of a draw the choice can be arbitrary. Bayesian risk (1.4) is

$$R_N(\lambda)=(\overline{D}N)^{1/2}r(0,0,0,0).\tag{2.31}$$

This description of control on the unit horizon is invariant in the sense that it does not depend on the total amount of data N but only on the number of batches K.

Proof. One should perform presented above change of variables in (2.11)–(2.15). Clearly, the equality is fulfilled

$$\lambda(m,D)dmdD=(N/\overline{D}^{\,3})^{1/2}\varrho(\alpha,\beta)\partial(m,D)/\partial(\alpha,\beta)d\alpha d\beta=\varrho(\alpha,\beta)d\alpha d\beta,$$

where $\partial(m,D)/\partial(\alpha,\beta)=(\overline{D}^{\,3}/N)^{1/2}$ is the Jacobian of the variable transformation. Therefore,

$$G_\ell(X,S,k_2)=\begin{cases}(\overline{D}/N)^{1/2}g_\ell(x,s,t_2), & if\ k_2=0,\\ (\overline{D}/N)^{1/2}(\overline{D})^{-1/2}g_\ell(x,s,t_2), & if\ k_2=1,\\ (\overline{D}/N)^{1/2}(\overline{D})^{-3/2}g_\ell(x,s,t_2), & if\ k_2\geq 2.\end{cases}$$

Also, the argument $k_\ell+1$ should be replaced by $K^{-1}(k_\ell+1)=t_\ell+\varepsilon$ and $S+M\Delta$ should be replaced by $(S+M\Delta)(\overline{D}M)^{-1}$, i.e.,

$$s+\frac{(X-k_2Y)^2}{\overline{D}Mk_2(k_2+1)}=s+\frac{\overline{D}NK^2(\varepsilon x-t_2y)^2}{\overline{D}MK^2t_2(t_2+\varepsilon)}=s+\delta(x,t_2,y),$$

where $\delta(x,t_2,y)$ is given by (2.30). Becides, $H(X,S,k,Y)dY$ should be replaced by

$$H(X,S,k,Y)dY=\begin{cases}(\overline{D})^{1/2}h(x,s,t,y)dy, & if\ k=0,\\ \overline{D}\,h(x,s,t,y)dy, & if\ k=1,\\ h(x,s,t,y)dy, & if\ k\geq 2,\end{cases}$$

where $H(X,S,k,Y)$, $h(x,s,t,y)$ are given by (2.14), (2.29).

Let's check the validity of (2.27). It's sufficient to check the second equation. Taking into account (2.25) and the comments made above, after substituting

the transformed variables and functions into the second equation (2.12), we obtain

$$(\overline{D}N)^{1/2}r_2(t_1,0,0,0) = M(\overline{D}/N)^{1/2}g_2(0,0,0)$$
$$+\int_{-\infty}^{\infty}(\overline{D}\,N)^{1/2}(\overline{D})^{-1/2}r(t_1,y,0,\varepsilon)(\overline{D})^{1/2}h(0,0,0,y)dy,$$
$$(\overline{D}N)^{1/2}(\overline{D})^{-1/2}r_2(t_1,x,0,\varepsilon) = M(\overline{D}/N)^{1/2}(\overline{D})^{-1/2}g_2(x,0,\varepsilon)$$
$$+\int_{-\infty}^{\infty}(\overline{D}N)^{1/2}(\overline{D})^{-3/2}r(t_1,x+y,\delta(x,\varepsilon,y),2\varepsilon)\overline{D}\,h(x,0,\varepsilon,y)dy,$$
$$(\overline{D}N)^{1/2}(\overline{D})^{-3/2}r_2(t_1,x,s,t_2) = M(\overline{D}N)^{1/2}(\overline{D})^{-3/2}g_2(x,s,t_2)$$
$$+\int_{-\infty}^{\infty}(\overline{D}N)^{1/2}(\overline{D})^{-3/2}r(t_1,x+y,s+\delta(x,t_2,y),t_2+\varepsilon)h(x,s,t_2,y)dy,$$

if $t_2 \geq 2\varepsilon$, and this gives the second equation (2.27). Formula (2.31) follows from (2.15) and performed change of variables. □

It was noted in section 1 that the optimal Bayesian strategy in the one-armed bandit problem applies the second action at the start of the control. After that, if the first action is applied once, it will be used until the end of the control. This follows from the fact that the application of the first action does not provide additional information about the process. Therefore, its application will not lead to the decision that the choice of the second action will become more preferable in the future. The property was first proved in [15] for a Bernoullian one-armed bandit and then in [10, 18] for a Gaussian one-armed bandit with one and both unknown parameters respectively.

This property makes it easier to find Bayesian strategy and risk. We immediately present the corresponding recursive equation in invariant form.

Theorem 4. *Consider a recursive equation*

$$r(0,x,s,t_2) = \min(r_1(0,x,s,t_2), r_2(0,x,s,t_2)), \tag{2.32}$$

where $r_1(0,x,s,t_2) = r_2(0,x,s,t_2) = 0$ *if* $t_2 = 1$ *and*

$$r_1(0,x,s,t_2) = (1-t_2)g_1(x,s,t_2),$$
$$r_2(0,x,s,t_2) = \varepsilon g_2(x,s,t_2) \tag{2.33}$$
$$+\int_{-\infty}^{\infty} r(0,x+y,s+\delta(x,t_2,y),t_2+\varepsilon)h(x,s,t_2,y)dy,$$

if $0 \leq t \leq 1-\varepsilon$. *Here* $g_1(x,s,t_2)$, $g_2(x,s,t_2)$ *are given by* (2.28), $h(x,s,t,y)$ *and* $\delta(x,t,y)$ *are given by* (2.29)–(2.30). *Bayesian strategy prescribes to choose the action corresponding to the smaller value of* $r_1(0,x,s,t_2)$, $r_2(0,x,s,t_2)$; *in the case of a draw, the choice can be arbitrary. Once the first action is chosen, it will be applied until the end of the control. Bayesian risk* (1.4) *is given by* (2.31).

The proof of theorem 4 is similar to the one presented in [18] and is therefore omitted.

To present invariant form of the equation for computing the regret, let's make additional change

$$\begin{gathered}\sigma_\ell(k_1,X,S,k_2)=\sigma_\ell(t_1,x,s,t_2),\\ L_\ell(k_1,0,0,0)=(\overline{D}N)^{1/2}l_\ell(t_1,0,0,0),\\ L_\ell(k_1,X,0,1)=(\overline{D}N)^{1/2}(\overline{D})^{-1/2}l_\ell(t_1,x,0,\varepsilon),\\ L_\ell(k_1,X,S,k_2)=(\overline{D}N)^{1/2}(\overline{D})^{-3/2}l_\ell(t_1,x,s,t_2),\end{gathered}\tag{2.34}$$

if $k_2\geq 2$, $\ell=1,2$.

Theorem 5. *To find the regret, one should solve the recursive equation*

$$l(t_1,x,s,t_2)=\sum_{\ell=1}^{2}\sigma_\ell(t_1,x,s,t_2)l_\ell(t_1,x,s,t_2),\tag{2.35}$$

where $l_1(t_1,x,s,t_2)=l_2(t_1,x,s,t_2)=0$ *if* $t=1$ *and*

$$\begin{gathered}l_1(t_1,x,s,t_2)=\varepsilon g_1(x,s,t_2)+l(t_1+\varepsilon,x,s,t_2),\\ l_2(t_1,x,s,t_2)=\varepsilon g_2(x,s,t_2)\\ +\int_{-\infty}^{\infty}l(t_1,x+y,s+\delta(x,t_2,y),t_2+\varepsilon)h(x,s,t_2,y)dy,\end{gathered}\tag{2.36}$$

if $0\leq t\leq 1-\varepsilon$. *Here*

$$\begin{aligned}g_1(x,s,t_2)&=\alpha^{+}\tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta),\\ g_2(x,s,t_2)&=\alpha^{-}\tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta),\end{aligned}\tag{2.37}$$

$h(x,s,t,y)$ *and* $\delta(x,t,y)$ *are given by* (2.29)–(2.30). *A regret* (1.2) *is*

$$L_N(\sigma,\theta)=(\overline{D}N)^{1/2}l(0,0,0,0).\tag{2.38}$$

This description of control on the unit horizon is invariant in the sense that it does not depend on the total amount of data N *but only on the number of batches* K.

Corollary 1. *Given* $\theta=(m,D)$, *consider the change of variables* $X=x(DN)^{1/2}$, $Y=y(DN)^{1/2}$, $S=sDM$, $k=tK$, $k_\ell=t_\ell K$, $M/N=K^{-1}=\varepsilon$, $m=\alpha(D/N)^{1/2}$, $L_\ell(k_1,0,0,0)=(DN)^{1/2}l_\ell(t_1,0,0,0)$, $L_\ell(k_1,X,0,1)=(DN)^{1/2}D^{-1/2}l_\ell(t_1,x,0,\varepsilon)$ *and* $L_\ell(k_1,X,S,k_2)=(DN)^{1/2}D^{-3/2}l_\ell(t_1,x,s,t_2)$ *if* $k_2\geq 2$, $\ell=1,2$. *Let also*

$$\sigma_\ell(k_1,X,S,k_2)=\sigma_\ell(t_1,x,s,t_2),\quad \ell=1,2.\tag{2.39}$$

Then, for finding the regret, one should solve equation (2.35)–(2.36), *where* $g_1(x,s,t_2)$, $g_1(x,s,t_2)$ *are given by* (2.37) *at* $\beta=1$ *and* $h(x,s,t,y)$, $\delta(x,t,y)$ *are given by* (2.29)–(2.30). *A regret* (1.2) *is*

$$L_N(\sigma,\theta)=(DN)^{1/2}l(0,0,0,0).\tag{2.40}$$

It follows from (2.40) that, when using strategies (2.39), the maximum values of the regret are attained at that parameters $\theta = (m, \overline{D})$, which correspond to maximal D. In general, the transformation of strategy according to (2.39) is impossible, because actual value of D is not known. Let's give an example, when strategy satisfying (2.39) can be used.

Upper Confidence Bound (UCB) strategy. This is a widely used strategy, some examples are given in [9, 19, 20]. Let's define the following values after processing k_2 batches ($k_2 \geq 2$) using the second action

$$Q_1(k_2) = 0, \quad Q_2(k_2) = \frac{X}{k_2} + a\,\gamma(k_2)\left(\frac{S/(k_2-1)}{k_2}\right)^{1/2}.$$

Here X/k_2 and $S/(k_2-1)$ are current estimates of the mathematical expectation and the variance of the income obtained by applying the second action to one batch, $(S/(k_2-1)/k_2)^{1/2}$ characterizes the mean-squared deviation of the estimate X/k_2. Note that $(S/(k_2-1)/k_2)^{1/2} \to 0$ as $k_2 \to \infty$. Parameters of the strategy are $a > 0$ and slow-growing function $\gamma(k_2) > 0$, e.g., $\gamma(k_2) = \ln(k_2)$. Thus, $Q_2(k_2)$ is the upper bound of the confidence interval for the interval estimate of mathematical expectation of the income obtained by applying the second action to one batch. For the first action a complete information is available and, hence, $Q_1(k_2) = 0$.

UCB strategy applies the second action to the two former batches. Then to the batch number $k_2 + 1$ it applies the action corresponding to the largest value of $Q_1(k_2)$, $Q_2(k_2)$, ($k_2 \geq 2$). Once the first action is selected, it will be applied until the end of the control. The second term in $Q_2(k_2)$ makes it possible to avoid the immediate transition to applying the first action if at $m > 0$ and some small k_2 the condition $X/k_2 < 0$ takes place by chance. Note that if $m < 0$, then the inequality $Q_2(k_2) < Q_1(k_2)$ is fulfilled for some k_2 with probability 1 .

Now let's show that for UCB strategy a condition (2.39) is valid. Indeed, after the change of variables we obtain

$$\begin{aligned} Q_2(k_2) &= \frac{x(DN)^{1/2}}{t_2 K} + a\,\gamma(k_2)\left(\frac{sDM}{t_2 K(t_2 K - 1)}\right)^{1/2} \\ &= \frac{(DM)^{1/2}}{K}\left(\frac{xK^{1/2}}{t_2} + a\,\gamma(k_2)\left(\frac{s}{t_2(t_2-\varepsilon)}\right)^{1/2}\right). \end{aligned}$$

Therefore, the mutual order of bounds $Q_1(k_2)$, $Q_2(k_2)$ is the same as that of bounds

$$q_1(k_2) = 0, \quad q_2(k_2) = \frac{xK^{1/2}}{t_2} + a\,\gamma(k_2)\left(\frac{s}{t_2(t_2-\varepsilon)}\right)^{1/2}.$$

So, mutual order of bounds $q_1(k_2)$, $q_2(k_2)$ does not depend on what D was used for the change of variables.

3 Estimating the variance by incomes within batches

Let's now consider batch data processing, in which the variance estimation is performed during data processing within the batch. To do this, we assume that the processing is carried out in batches of the size $M = M_1 M_2$. In turn, these batches are divided into M_2 small packets, each of which includes M_1 of data. This makes it possible to estimate the variance when processing the next large batch based on the observed incomes of small packets by computing the corresponding s^2-statistics. It is clear that small packets and also large batches themselves still allow parallel processing. The number of large batches and, accordingly, the number of processing stages is K. Thus, the total number of data is $N = KM_1M_2 = KM$. Note that although some notations below are similar as in the previous section (i.e., $R_\ell(k_1, X, S, k_2)$, $\tilde{\mathbf{F}}(X, S, k_2|m', D')$, etc) they are not exactly the same and should be interpreted as independent from used in section 2.

Let's consider how to recalculate the total income X and s^2-statistics S after processing the next large batch. Let k be the current number of large batches processed and, therefore, $n = kM_2$ be the current total number of small packets included in them. Then the current total income and s^2-statistics are

$$X = \sum_{i=1}^{n} x_i, \quad S = \sum_{i=1}^{n} x_i^2 - X^2/n,$$

where $x_1, \ldots, x_n$ are the incomes of small packets. For the next $(k+1)$th large batch, one can compute its total income and s^2-statistics on the small packets included in it with incomes $x_{n+1}, \ldots, x_{n+M_2}$:

$$Y = \sum_{i=n+1}^{n+M_2} x_i, \quad U = \sum_{i=n+1}^{n+M_2} x_i^2 - Y^2/M_2.$$

Then the new values of total income and s^2-statistics are recalculated using the old ones according to the following formulas

$$X_{new} = \sum_{i=1}^{n+M_2} x_i = X + Y,$$

$$S_{new} = \left(\sum_{i=1}^{n+M_2} x_i^2 \right) - (X+Y)^2/(n+M_2) = S + U + M_1\Delta,$$

where

$$M_1\Delta = Y^2/M_2 + X^2/n - (X+Y)^2/(n+M_2) = \frac{(M_2X - nY)^2}{nM_2(n+M_2)}.$$

Thus, the recalculation of statistics after the receipt of the next large data batch is carried out according to the formulas

$$X \leftarrow X + Y, \quad S \leftarrow S + U + M_1\Delta, \tag{3.1}$$

where

$$\Delta = \frac{(M_2X - kM_2Y)^2}{M_1M_2^3k(k+1)} = \frac{(X-kY)^2}{Mk(k+1)}. \tag{3.2}$$

Denote by $D' = M_1D$ and $m' = M_1m$ the variance and the mathematical expectation of income for processing the small packet. Let's introduce the functions

$$f_{k_2M_2D'}(X|k_2M_2m') = \begin{cases} 1, & \text{if } k_2 = 0, \\ f_{k_2M_2D'}(X|k_2M_2m'), \text{ with } f(\cdot) \text{ from (1.1)} & \text{if } k_2 \geq 1, \end{cases}$$

$$\psi_{k_2M_2-1}(S/D') = \begin{cases} 1, & \text{if } k_2 = 0, \\ (D')^{-1}\chi^2_{k_2M_2-1}(S/D'), & \text{if } k_2 \geq 1. \end{cases}$$

If $k_2 \geq 1$, these functions describe the probability density functions (pdf) of cumulative income X and s^2-statistics S computed after processing k_2 large batches or, equivalently, after processing k_2M_2 small packets. Since X and S are independent random variables, the joint pdf

$$\mathbf{F}(X, S|m', D') = f_{M_2D'}(X|M_2m')\psi_{M_2-1}\left(S/D'\right) \tag{3.3}$$

describes the pdf of X, S, corresponding to processing one large batch.

Like in section 2, given a prior distribution density $\lambda(m, D)$, the posterior distribution density is

$$\lambda(m, D|X, S, k_2) = \frac{f_{k_2M_2D'}\left(X|k_2M_2m'\right)\psi_{k_2M_2-1}\left(S/D'\right)\lambda(m.D)}{P(X, S, k_2)},$$

where

$$P(X, S, k_2) = \iint_\Theta f_{k_2M_2D'}\left(X|k_2M_2m'\right)\psi_{k_2M_2-1}\left(S/D'\right)\lambda(m.D)dmdD.$$

where $k_2 = 1, 2, \ldots$ and $\lambda(m, D|0, 0, 0) = \lambda(m, D)$. However, recursive equation is simpler if the posterior distribution is defined in an equivalent way. Denote

$$\tilde{\mathbf{F}}(X, S, k_2|m', D') \tag{3.4}$$
$$= \begin{cases} 1, & if\ k_2 = 0, \\ (D')^{-3/2}\tilde{f}_{k_2M_2D'}(X|k_2M_2m')\tilde{\psi}_{M_2k_2-1}\left(S/D'\right), & if\ k_2 \geq 1, \end{cases}$$

where

$$\tilde{f}_{k_2M_2D'}(X|k_2M_2m') = \begin{cases} 1, & \text{if } k_2 = 0, \\ \exp\left(-\dfrac{(X - k_2M_2m')^2}{2k_2M_2D'}\right), & \text{if } k_2 \geq 1, \end{cases}$$

$$\tilde{\psi}_{k_2M_2-1}(S/D') = \begin{cases} 1, & \text{if } k_2 = 0, \\ (S/D')^{\frac{k_2M_2-1}{2}-1}e^{-S/(2D')}, & \text{if } k_2 \geq 1, \end{cases} \tag{3.5}$$

Then, given a prior distribution density $\lambda(m, D)$, the posterior distribution density is

$$\lambda(m, D|X, S, k_2) = \frac{\tilde{\mathbf{F}}(X, S, k_2|m', D')\lambda(m, D)}{\tilde{P}(X, S, k_2)},$$
$$\text{with} \quad \tilde{P}(X, S, k_2) = \iint_{\Theta} \tilde{\mathbf{F}}(X, S, k_2|m', D')\lambda(m, D)dmdD. \tag{3.6}$$

Note that (3.6) remains valid if $k_2 = 0$, too.

Denote by $R^B(k_1, X, S, k_2) = R^B_{K-k}(\lambda(m, D|X, S, k_2)$ the Bayesian risk computed on the control horizon $K - k$ with respect to a prior distribution density $\lambda(m, D|X, S, k_2)$. Taking into account (3.1)–(3.2), the standard dynamic programming equation has the form

$$R^B(k_1, X, S, k_2) = \min\left(R^B_1(k_1, X, S, k_2), R^B_2(k_1, X, S, k_2)\right), \tag{3.7}$$

where $R^B_1(k_1, X, S, k_2) = R^B_2(k_1, X, S, k_2) = 0$ if $k_1 + k_2 = K$ and

$$R^B_1(k_1, X, S, k_2) = \iint_{\Theta} \lambda(m, D|X, S, k_2)$$
$$\times \left(M_2(m')^+ + R^B(k_1 + 1, X, S, k_2)\right) dmdD, \tag{3.8}$$
$$R^B_2(k_1, X, S, k_2) = \iint_{\Theta} \lambda(m, D|X, S, k_2) \times \Big(M_2(m')^-$$
$$+ \int_0^{\infty} \int_{-\infty}^{\infty} R^B(k_1, X + Y, S + U + M_1\Delta, k_2 + 1)\mathbf{F}(Y, U|m', D')dY\,dU\Big)dmdD$$

if $0 \le k_1 + k_2 < K$. Bayesian risk (1.4) is

$$R_N(\lambda) = R(0, 0, 0, 0). \tag{3.9}$$

Here $R^B_\ell(k_1, X, S, k_2)$ characterizes the expected loss on the control horizon $K - k$ if the ℓth action is applied first and then the control is carried out optimally. When processing $(k + 1)$th the large batch, the Bayesian strategy prescribes choosing an action corresponding to the current smaller value $R^B_1(k_1, X, S, k_2)$, $R^B_2(k_1, X, S, k_2)$; in the case of a draw, the choice can be arbitrary.

Let's present equation (3.7)–(3.8) in a more convenient for computations form. We put

$$R_\ell(k_1, X, S, k_2) = R^B_\ell(k_1, X, S, k_2) \times \tilde{P}(X, S, k_2), \tag{3.10}$$

$\ell = 1, 2$. The following theorem is valid.

Theorem 6. *To determine the Bayesian risk, one should solve a recursive equation*

$$R(k_1, X, S, k_2) = \min\left(R_1(k_1, X, S, k_2), R_2(k_1, X, S, k_2)\right), \tag{3.11}$$

where $R_1(k_1, X, S, k_2) = R_2(k_1, X, S, k_2) = 0$ *if* $k_1 + k_2 = K$ *and*

$$R_1(k_1, X, S, k_2) = MG_1(X, S, k_2) + R(k_1 + 1, X, S, k_2),$$
$$R_2(k_1, X, S, k_2) = MG_2(X, S, k_2) \tag{3.12}$$
$$+ \int_0^\infty \int_{-\infty}^\infty R(k_1, X + Y, S + U + M_1\Delta, k_2 + 1)H(X, S, k_2, Y, U)dY\,dU$$

if $0 \le k_1 + k_2 < K$. *Here*

$$G_1(X, S, k_2) = \iint_\Theta m^+ \tilde{\mathbf{F}}(X, S, k_2|m', D')\lambda(m, D)dmdD,$$
$$G_2(X, S, k_2) = \iint_\Theta m^- \tilde{\mathbf{F}}(X, S, k_2|m', D')\lambda(m, D)dmdD \tag{3.13}$$

and

$$H(X, S, k, Y, U)$$
$$= \begin{cases} C(M_2), & if\ k = 0, \\ C(M_2) \times \dfrac{S^{(kM_2-1)/2-1}U^{(M_2-1)/2-1}}{(S + U + M_1\Delta)^{((k+1)M_2-1)/2-1}}, & if\ k \ge 1 \end{cases} \tag{3.14}$$

with

$$C(M_2) = \left(\frac{1}{2^{M_2}M_2\pi}\right)^{1/2} \times \frac{1}{\Gamma((M_2 - 1)/2)}.$$

Bayesian risk (1.4) *is*

$$R_N(\lambda) = R(0, 0, 0, 0). \tag{3.15}$$

When processing the $(k+1)$*th large batch, the Bayesian strategy prescribes to choose an action corresponding to the current smaller value* $R_1(k_1, X, S, k_2)$, $R_2(k_1, X, S, k_2)$*; in the cade of a draw, the choice can be arbitrary.*

Proof. Let's multiply the left-hand and right-hand sides of the equation (3.7)–(3.8) by $\tilde{P}(X, S, k_2)$ in (3.6). Taking into account (3.10) and (3.6), we get (3.11)–(3.12), where $G_1(X, S, k_2)$, $G_2(X, S, k_2)$ are described by (3.13), and

$$H(X, S, k, Y, U) = \frac{\tilde{\mathbf{F}}(X, S, k|m', D')\mathbf{F}(Y, U|m', D')}{\tilde{\mathbf{F}}(X + Y, S + U + M_1\Delta, k + 1|m', D')}. \tag{3.16}$$

The cases $k \ge 1$ and $k = 0$ should be considered separately. For $k \ge 1$, taking into account (3.3)–(3.4), it follows from (3.16) that

$$H(X, S, k, Y, U) = \frac{\tilde{f}_{kM_2D'}(X|kM_2m')f_{M_2D'}(Y|M_2m')}{\tilde{f}_{(k+1)M_2D'}(X + Y|(k + 1)M_2m')}$$
$$\times \frac{\tilde{\psi}_{kM_2-1}(S/D')\psi_{M_2-1}(U/D')}{\tilde{\psi}_{(k+1)M_2-1}((S + U + M_1\Delta)/D')}.$$

Here

$$\frac{\tilde{f}_{kM_2D'}(X|kM_2m')f_{M_2D'}(Y|M_2m')}{\tilde{f}_{(k+1)M_2D'}(X+Y|(k+1)M_2m')} = \left(\frac{1}{2\pi M_2D'}\right)^{1/2}\exp\left(-\frac{M_1\Delta}{2D'}\right),$$

and

$$\frac{\tilde{\psi}_{kM_2-1}(S/D')\psi_{M_2-1}(U/D')}{\tilde{\psi}_{(k+1)M_2-1}((S+U+M_1\Delta)/D')} = \frac{1}{D'\times 2^{(M_2-1)/2}\Gamma((M_2-1)/2)}$$
$$\times\frac{(S/D')^{(kM_2-1)/2-1}(U/D')^{(M_2-1)/2-1}}{((S+U+M_1\Delta)/D')^{((k+1)M_2-1)/2-1}}\times\frac{\exp\left(-S/(2D')\right)\exp\left(-U/(2D')\right)}{\exp\left(-(S+U+M_1\Delta)/(2D')\right)}$$
$$=\frac{(D')^{1/2}\exp\left(M_1\Delta/(2D')\right)}{2^{(M_2-1)/2}\Gamma((M_2-1)/2)}\times\frac{S^{(kM_2-1)/2-1}U^{(M_2-1)/2-1}}{(S+U+M_1\Delta)^{((k+1)M_2-1)/2-1}}.$$

Hence, $H(X,S,k,Y,U)$ satisfies (3.14) if $k\geq 1$. If $k=0$ then $X=0$, $S=0$ and (3.16) takes the form

$$H(0,0,0,Y,U) = \frac{f_{M_2D'}(Y|M_2m')\psi_{M_2-1}(U/D')}{(D')^{-3/2}\tilde{f}_{M_2D'}(Y|M_2m')\tilde{\psi}_{M_2-1}(U/D')}$$
$$=\left(\frac{1}{2\pi M_2}\right)^{1/2}\frac{1}{2^{(M_2-1)/2}\Gamma((M_2-1)/2)} = C(M_2).$$

Hence, $H(X,S,k,Y,U)$ satisfies (3.14) if $k=0$. Formula (3.15) follows from (3.9) and equality $\tilde{P}(0,0,0)=1$. □

Let's present a recursive equation for computing the regret (1.2) and right now in a more convenient form for computations. Let the control strategy σ be described by a set of probabilities

$$\sigma_\ell(k_1,X,S,k_2) = \Pr(y_{k+1}=\ell|k_1,X,S,k_2),$$

$\ell=1,2$; $k_1+k_2=k$, $k=0,\dots,K-1$; $X\in(-\infty,+\infty)$, $S\in(0,+\infty)$. Similarly to theorem 2 the following theorem holds true.

Theorem 7. *Consider a recursive equation*

$$L(k_1,X,S,k_2) = \sum_{\ell=1}^{2}\sigma_\ell(k_1,X,S,k_2)L_\ell(k_1,X,S,k_2), \tag{3.17}$$

where $L_1(k_1,X,k_2)=L_2(k_1,X,k_2)=0$ *if* $k=K$ *and*

$$L_1(k_1,X,S,k_2) = MG_1(X,S,k_2)+L(k_1+1,X,S,k_2),$$
$$L_2(k_1,X,S,k_2) = MG_2(X,S,k_2) \tag{3.18}$$
$$+\int_0^\infty\int_{-\infty}^\infty L(k_1,X+Y,S+U+M_1\Delta,k_2+1)H(X,S,k_2,Y,U)dY\,dU,$$

if $0\leq k\leq K-1$. *Here*

$$G_1(X,S,k_2) = m^+\tilde{\mathbf{F}}(X,S,k_2|m',D'),$$
$$G_2(X,S,k_2) = m^-\tilde{\mathbf{F}}(X,S,k_2|m',D'),$$

$H(X,S,k,Y,U)$ is given by (3.14) and $\Delta = \Delta(X,k,Y)$ is given by (3.2). Then a regret (1.2) is

$$L_N(\sigma,\theta) = L(0,0,0,0). \tag{3.19}$$

Proof. One should write a standard equation for computing the regret, which is similar to given by formulas (2.19)–(2.20). Then one should transform this equation similarly to theorem 2 and take the degenerate prior pdf $\lambda(m,D)$ concentrated at the parameter $\theta = (m,D)$. □

Let's obtain an invariant form of formulas (3.11)–(3.15). We take the set of parameters $\Theta_N = \{(m,D) : \underline{D} \le D \le \overline{D}, |m| \le c(D/N)^{1/2}\}$, where $c > 0$, $0 < \underline{D} \le D \le \overline{D} < \infty$. If one puts $D = \beta\overline{D}$, $m = \alpha(\overline{D}/N)^{1/2} = \alpha(\beta^{-1}D/N)^{1/2}$, then the set of parameters takes the form $\Theta_N = \{(\alpha,\beta) : \underline{D}/\overline{D} = \beta_0 \le \beta \le 1, |\alpha| \le c\beta^{1/2}\}$.

Consider the change of variables: $X = x(\overline{D}N)^{1/2}$, $Y = y(\overline{D}N)^{1/2}$, $S = s\overline{D}M_1$, $U = u\overline{D}M_1$, $k = tK$, $k_1 = t_1K$, $k_2 = t_2K$, $M/N = K^{-1} = \varepsilon$, $m = \alpha(\overline{D}/N)^{1/2}$, $D = \beta\overline{D}$, $\lambda(m,D) = (N/\overline{D}^{\,3})^{1/2}\varrho(\alpha,\beta)$. Let

$$\begin{gathered} R_\ell(k_1,0,0,0) = (\overline{D}N)^{1/2} r_\ell(t_1,0,0,0), \\ R_\ell(k_1,X,S,k_2) = (\overline{D}\,N)^{1/2}(\overline{D}\,M_1)^{-3/2} r_\ell(t_1,x,s,t_2), \text{ if } k_2 \ge 1, \end{gathered} \tag{3.20}$$

$\ell = 1,2$. Then the following theorem is valid.

Theorem 8. *To determine a Bayesian risk, one should solve a recursive equation*

$$r(t_1,x,s,t_2) = \min\left(r_1(t_1,x,s,t_2), r_2(t_1,x,s,t_2)\right), \tag{3.21}$$

where $r_1(t_1,x,s,t_2) = r_2(t_1,x,s,t_2) = 0$ if $t = 1$ and

$$\begin{gathered} r_1(t_1,x,s,t_2) = \varepsilon g_1(x,s,t_2) + r(t_1+\varepsilon,x,s,t_2), \\ r_2(t_1,x,s,t_2) = \varepsilon g_2(x,s,t_2) \\ + \int_0^\infty \int_{-\infty}^\infty r(t_1,x+y,s+u+\delta(x,t_2,y),t_2+\varepsilon) h(x,s,t_2,y,u)\,dy\,du, \end{gathered} \tag{3.22}$$

if $0 \le t \le 1-\varepsilon$. Here

$$\begin{aligned} g_1(x,s,t_2) &= \iint_{\Theta_N} \alpha^+ \tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta)\varrho(\alpha,\beta)\,d\alpha\,d\beta, \\ g_2(x,s,t_2) &= \iint_{\Theta_N} \alpha^- \tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta)\varrho(\alpha,\beta)\,d\alpha\,d\beta, \end{aligned} \tag{3.23}$$

with

$$\tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta) = \begin{cases} 1, & if\ t_2 = 0, \\ \beta^{-3/2}\tilde{f}_{t_2\beta}(x|t_2\alpha)\tilde{\psi}_{k_2M_2-1}(s/\beta), & if\ t_2 \ge \varepsilon \end{cases}$$

and

$$h(x,s,t,y,u) \tag{3.24}$$
$$= \begin{cases} c(M_2), & if\ t=0, \\ c(M_2)\times \dfrac{s^{(kM_2-1)/2-1}u^{(M_2-1)/2-1}}{(s+u+\delta(x,t,y))^{((k+1)M_2-1)/2-1}}, & if\ t\geq\varepsilon \end{cases}$$

with

$$c(M_2)=\left(\frac{1}{2^{M_2}\pi\varepsilon}\right)^{1/2}\times\frac{1}{\Gamma((M_2-1)/2)}$$

and

$$\delta(x,t_2,y)=\begin{cases} 0, & if\ t_2=0, \\ \dfrac{(\varepsilon x-t_2y)^2}{\varepsilon t_2(t_2+\varepsilon)}, & if\ t_2\geq\varepsilon. \end{cases} \tag{3.25}$$

When processing the $(k+1)$th large batch (respective to $(t+\varepsilon)$ point of time) the Bayesian strategy prescribes choosing an action corresponding to a smaller value $r_1(t_1,x,s,t_2)$, $r_2(t_1,x,s,t_2)$; in the case of a draw, the choice can be arbitrary. Bayesian risk (1.4) *is*

$$R_N(\lambda)=(\overline{D}\,N)^{1/2}r(0,0,0,0). \tag{3.26}$$

This description of control on the unit horizon is invariant in the sense that it does not depend on the total amount of data N but only on the number of large batches K and the number of small packets M_2 as parts of large ones.

Proof. The proof is similar to the proof of theorem 3. One should perform the above change of variables in (3.11)–(3.15). Again, $\lambda(m,D)dmdD = \varrho(\alpha,\beta)d\alpha d\beta$. Therefore,

$$G_\ell(X,S,k_2)=\begin{cases} (\overline{D}/N)^{1/2}g_\ell(x,s,t_2), & if\ k_2=0, \\ (\overline{D}/N)^{1/2}(\overline{D}\,M_1)^{-3/2}g_\ell(x,s,t_2), & if\ k_2\geq 1. \end{cases}$$

Next, argument $k_\ell+1$ must be replaced by $K^{-1}(k_\ell+1)=t_\ell+\varepsilon$, $S+U+M_1\Delta$ must be replaced by $(S+U+M_1\Delta)(\overline{D}M_1)^{-1}$, i.e.,

$$s+u+\frac{(X-k_2Y)^2}{\overline{D}Mk_2(k_2+1)}=s+u+\frac{NK^2(\varepsilon x-t_2y)^2}{MK^2t_2(t_2+\varepsilon)}=s+u+\delta(x,t_2,y),$$

where $\delta(x,t_2,y)$ is given by (3.25). Besides, $H(X,S,k,Y,U)dYdU$ must be replaced as

$$H(X,S,k,Y,U)dYdU=\begin{cases} (\overline{D}\,M_1)^{3/2}h(x,s,t,y,u)dydu, & if\ k=0, \\ h(x,s,t,y,u)dydu, & if\ k\geq 1. \end{cases}$$

Let's check the validity of (3.22). It's sufficient to check the second equation. Taking into account (3.20) and the comments made above, after substituting

the transformed variables and functions into the second equation (3.12), we obtain

$$(\overline{D}N)^{1/2}r_2(t_1,0,0,0) = M(\overline{D}/N)^{1/2}g_2(0,0,0)$$
$$+\int_0^\infty\int_{-\infty}^\infty (\overline{D}\,N)^{1/2}(\overline{D}\,M_1)^{-3/2}r(t_1,y,u,\varepsilon)(\overline{D}\,M_1)^{3/2}h(0,0,\varepsilon,y,u)dydu,$$
$$(\overline{D}N)^{1/2}(\overline{D}\,M_1)^{-3/2}r_2(t_1,x,s,t_2)$$
$$= M(\overline{D}/N)^{1/2}(\overline{D}\,M_1)^{-3/2}g_2(x,s,t_2) + (\overline{D}\,N)^{1/2}(\overline{D}\,M_1)^{-3/2}$$
$$\times\int_0^\infty\int_{-\infty}^\infty r(t_1,x+y,s+u+\delta(x,t_2,y),t_2+\varepsilon)h(x,s,t_2,y,u)dydu,$$

if $t_2 \geq \varepsilon$, which gives the second equality (3.22). Formula (3.26) follows from (3.15) and change of variables made above. □

Now let's present in invariant form an equation that takes into account the nature of the optimal strategy, i.e., if the first action is applied once, it will be used until the end of the control. Like theorem 4, the following one is valid.

Theorem 9. *Consider a recursive equation*

$$r(0,x,s,t_2) = \min(r_1(0,x,s,t_2), r_2(0,x,s,t_2)), \tag{3.27}$$

where $r_1(0,x,s,t_2) = r_2(0,x,s,t_2) = 0$ *if* $t_2 = 1$ *and*

$$r_1(0,x,s,t_2) = (1-t_2)g_1(x,s,t_2),$$
$$r_2(0,x,s,t_2) = \varepsilon g_2(x,s,t_2) \tag{3.28}$$
$$+\int_0^\infty\int_{-\infty}^\infty r(0,x+y,s+u+\delta(x,t_2,y),t_2+\varepsilon)h(x,s,t_2,y,u)dydu,$$

if $0 \leq t \leq 1-\varepsilon$. *Here* $g_1(x,s,t_2)$, $g_2(x,s,t_2)$ *are given by* (3.23), $h(x,s,t,y,u)$ *and* $\delta(x,t,y)$ *are given by* (3.24)–(3.25). *Bayesian strategy prescribes choosing an action corresponding to the current smaller value of* $r_1(0,x,s,t_2)$, $r_2(0,x,s,t_2)$; *in the case of a draw, the choice can be arbitrary. Being selected once, the first action will be applied until the end of the control. Bayesian risk is given by* (3.26).

The proof of theorem 9 is similar to that given in [18] and is therefore omitted.

For an invariant representation of the equation for computing the regret, we make an additional replacement

$$\sigma_\ell(k_1,X,S,k_2) = \sigma_\ell(t_1,x,s,t_2),$$
$$L_\ell(k_1,0,0,0) = (\overline{D}N)^{1/2}l_\ell(t_1,0,0,0), \tag{3.29}$$
$$L_\ell(k_1,X,S,k_2) = (\overline{D}\,N)^{1/2}(\overline{D}\,M_1)^{-3/2}l_\ell(t_1,x,s,t_2), \text{ if } k_2 \geq 1,$$

$\ell = 1,2$. Then the following theorem is valid.

Theorem 10. *To determine a regret, one should solve a recursive equation*

$$l(t_1,x,s,t_2)=\sum_{\ell=1}^{2}\sigma_\ell(t_1,x,s,t_2)l_\ell(t_1,x,s,t_2), \tag{3.30}$$

where $l_1(t_1,x,s,t_2)=l_2(t_1,x,s,t_2)=0$ *if* $t=1$ *and*

$$\begin{gathered} l_1(t_1,x,s,t_2)=\varepsilon g_1(x,s,t_2)+l(t_1+\varepsilon,x,s,t_2),\\ l_2(t_1,x,s,t_2)=\varepsilon g_2(x,s,t_2)\\ +\int_0^\infty\int_{-\infty}^\infty l(t_1,x+y,s+u+\delta(x,t_2,y),t_2+\varepsilon)h(x,s,t_2,y,u)dydu, \end{gathered} \tag{3.31}$$

if $0\le t\le 1-\varepsilon$. *Here*

$$\begin{aligned} g_1(x,s,t_2)&=\alpha^+\tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta),\\ g_2(x,s,t_2)&=\alpha^-\tilde{\mathbf{f}}(x,s,t_2|\alpha,\beta), \end{aligned} \tag{3.32}$$

$h(x,s,t,y,u)$ *and* $\delta(x,t,y)$ *are given by* (3.24)–(3.25). *A regret* (1.2) *is*

$$L_N(\sigma,\theta)=(\overline{D}N)^{1/2}l(0,0,0,0). \tag{3.33}$$

This description of control on the unit horizon is invariant in the sense that it does not depend on the total amount of data N *but only on the number of large batches* K *and the number of small packets* M_2.

Corollary 2. *Given* $\theta=(m,D)$, *consider a change of variables* $X=x(DN)^{1/2}$, $Y=y(DN)^{1/2}$, $S=sDM_1$, $U=uDM_1$, $k=tK$, $k_\ell=t_\ell K$, $M/N=K^{-1}=\varepsilon$, $m=\alpha(D/N)^{1/2}$, $L_\ell(k_1,0,0,0)=(DN)^{1/2}l_\ell(t_1,0,0,0)$ *and* $L_\ell(k_1,X,S,k_2)=(DN)^{1/2}(DM_1)^{-3/2}l_\ell(t_1,x,s,t_2)$ *if* $k_2\ge 1$, $\ell=1,2$. *Let also*

$$\sigma_\ell(k_1,X,S,k_2)=\sigma_\ell(t_1,x,s,t_2),\quad \ell=1,2. \tag{3.34}$$

Then for finding the regret, one should solve the equation (3.30)–(3.31), *where* $g_1(x,s,t_2)$, $g_1(x,s,t_2)$ *are given by* (3.32) *with* $\beta=1$, $h(x,s,t,y,u)$ *and* $\delta(x,t,y)$ *are given by* (3.24)–(3.25). *A regret* (1.2) *is*

$$L_N(\sigma,\theta)=(DN)^{1/2}l(0,0,0,0). \tag{3.35}$$

It follows from (3.35) that when using strategies (3.34), the largest values of regret are achieved at parameters $\theta=(m,\overline{D})$ that correspond to the largest values of D. Although in general it is impossible to perform the transformation (3.34) due to the unknown D, for some strategies the property (3.34) is valid.

UCB strategy. After processing k_2 batches using the second action ($k_2\ge 1$), we determine the following statistics

$$Q_1(k_2)=0,\quad Q_2(k_2)=\frac{X}{k_2}+a\,\gamma(k_2)\left(\frac{M_2S/(M_2k_2-1)}{k_2}\right)^{1/2}.$$

Here X/k_2 and $M_2S/(M_2k_2-1)$ are the estimates of the mathematical expectation and the variance of income obtained when processing the batch

Таблица 1. Normalized Bayesian risks

λ	λ_{11}	λ_{12}	λ_{21}	λ_{22}	$r_N^B(\lambda)$
1	0.2	0.3	0.2	0.3	0.36
2	0.2	0.3	0.3	0.2	0.34
3	0.1	0.4	0.4	0.1	0.33

Таблица 2. Estimates of normalized Bayesian risks

λ	$r_N^B(\lambda')$	$l_N^B(\sigma(\lambda'),\lambda'')$	$r_N^B(\lambda'')$	$l_N^B(\sigma(\lambda''),\lambda')$	$l_N(\lambda)$
1	0.39	0.33	0.33	0.39	0.36
2	0.39	0.31	0.33	0.39	0.35
3	0.30	1.20	0.18	0.60	0.57

of M data with the use of the second action. Next, $(M_2S/(M_2k_2-1)/k_2)^{1/2}$ characterizes the mean-squared deviation of the estimate X/k_2, and $(M_2S/(M_2k_2-1)/k_2)^{1/2} \to 0$ as $k_2 \to \infty$. The strategy parameters are again $a > 0$ and the slowly growing function $\gamma(k_2) > 0$.

The UCB strategy applies the second action to the first batch, and then applies to the batch with the number k_2+1 the action, which corresponds to the largest of the values $Q_1(k_2)$, $Q_2(k_2)$, $(k_2 = 1, 2, \dots)$. If the first action is selected once, it will be applied until the end of the control.

Let's show that the condition (3.34) is met for the UCB strategy. Indeed, after performing the change of variables, we obtain

$$Q_2(k_2) = \frac{x(DN)^{1/2}}{t_2K} + a\,\gamma(k_2)\left(\frac{M_2sDM_1}{t_2K(M_2t_2K-1)}\right)^{1/2}$$
$$= \frac{(DM)^{1/2}}{K}\left(\frac{xK^{1/2}}{t_2} + a\,\gamma(k_2)\left(\frac{s}{t_2(M_2t_2-\varepsilon)}\right)^{1/2}\right).$$

Therefore, the mutual order of the bounds $Q_1(k_2)$, $Q_2(k_2)$ is the same as that of the bounds

$$q_1(k_2) = 0, \quad q_2(k_2) = \frac{xK^{1/2}}{t_2} + a\,\gamma(k_2)\left(\frac{s}{t_2(M_2t_2-\varepsilon)}\right)^{1/2}.$$

So, mutual order of bounds $q_1(k_2)$, $q_2(k_2)$ does not depend on which D was used when changing the variables.

4 Numerical results

Let's describe the results of numerical experiments. In the case of estimating the variance using cumulative incomes in batches, we computed Bayesian risk for the number of batches $K = 18$ with the batch size $M = 1$, so $N = K$. Calculations of Bayesian risk were performed using formulas (2.11)–(2.15) which can be used directly. But we simplified them taking into account

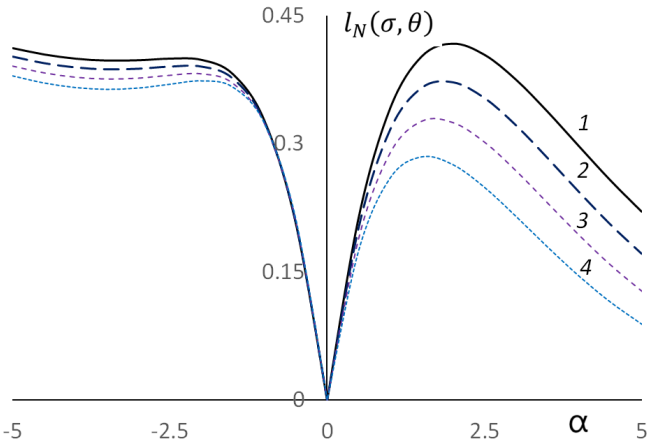

FIG. 4.1. Normalized regrets for different variances - 1.

the fact that if the strategy switches to the second action, it will apply it until the end of control. In invariant form. these simplified formulas are presented in theorem 4. When performing numerical integration, X varied in the range from -7 to 7 in increments of 0.07, and S varied from 0.005 to 20.005 in increments of 0.01. A small increment in S is due to the singularity of $H(X,S,k,Y)$, and accordingly $R(k_1,X,S,k_2)$, at the point $S=0$ if $k=2$.

A set of parameters Θ was characterized by different variance values and was chosen concentrated at four parameters $\theta_{11}=(m_p,D_1)$, $\theta_{12}=(m_n,D_1)$, $\theta_{21}=(m_p,D_2)$, $\theta_{22}=(m_n,D_2)$, where $D_1=\overline{D}=1$, $D_2=\underline{D}=0.7$, $m_p=1.5(\overline{D}/N)^{1/2}$, $m_n=-2.5(\overline{D}/N)^{1/2}$. For prior distributions $\lambda=(\lambda_{11},\lambda_{12},\lambda_{21},\lambda_{22})$, where $\lambda_{ij}=\Pr(\theta=\theta_{ij})$, $i,j=1,2$, values of normalized Bayesian risks $r_N^B(\lambda)=(\overline{D}N)^{-1/2}R_N^B(\lambda)$ are presented in the table 1.

Then we approximated risks from the table 1 by risks and regrets, computed on the sets of parameters $\{\theta_{11},\theta_{12}\}$ and $\{\theta_{21},\theta_{22}\}$, each of which is characterized by a single value of variance.To this end, on the sets $\{\theta_{11},\theta_{12}\}$ and $\{\theta_{21},\theta_{22}\}$ prior distributions $\lambda'=(\lambda_{11}/\mu_1,\lambda_{12}/\mu_1)$ and $\lambda''=(\lambda_{21}/\mu_2,\lambda_{22}/\mu_2)$ were assigned with $\mu_1=\lambda_{11}+\lambda_{12}$, $\mu_2=\lambda_{21}+\lambda_{22}$. Then for a prior distribution λ' a normalized Bayesian risk $r_N^B(\lambda')=(\overline{D}N)^{-1/2}R_N^B(\lambda')$ and a Bayesian strategy $\sigma^B(\lambda')$ were determined. Then the strategy $\sigma^B(\lambda')$ was applied on a prior distribution λ'' and the normalized regret $l_N^B(\sigma(\lambda'),\lambda'')=(\overline{D}N)^{-1/2}L_N^B(\sigma(\lambda'),\lambda'')$ was computed. Similarly, $r_N^B(\lambda'')$ and $l_N^B(\sigma(\lambda''),\lambda')$ were computed. Finally, the estimate of Bayesian risk $r_N^B(\lambda)$ is as follows

$$l_N(\lambda)=\mu_1\left(\mu_1 r_N^B(\lambda')+\mu_2 l_N^B(\sigma(\lambda'),\lambda'')\right)+\mu_2\left(\mu_1 l_N^B(\sigma(\lambda''),\lambda')+\mu_2 r_N^B(\lambda'')\right).$$

The results corresponding to the prior distributions from the table 1 are presented in the table 2. Everywhere $\mu_1=\mu_2=0.5$. One can see that if the distributions λ' and λ" are close (cases 1 and 2), then the estimate $l_N(\lambda)$ is

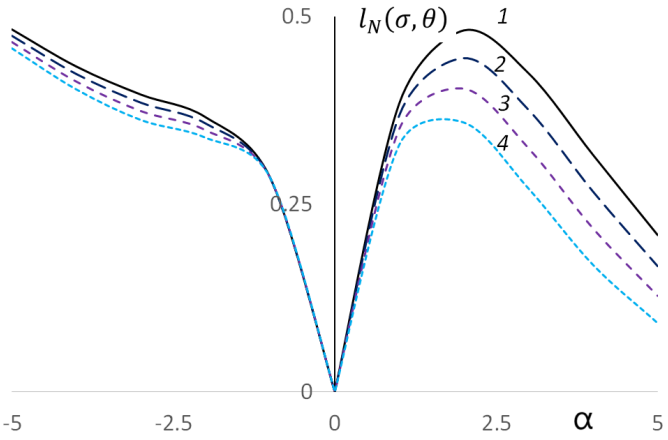

FIG. 4.2. Normalized regrets for different variances - 2.

close to the value of the risk $r_N^B(\lambda)$. If the distributions λ' and λ" are very different, then the estimate $l_N(\lambda)$ is very different from $r_N^B(\lambda)$.

For approximate finding the minimax strategy and risk, the main theorem of game theory is used, according to which the minimax risk coincides with the Bayesian risk computed with respect to the worst-case prior distribution on which the Bayesian risk is maximal. At the same time, the minimax strategy coincides with the corresponding Bayesian one. As an example, consider the approximate finding the minimax risk at $K = 18$, $M = 1$ on the set of parameters $\Theta = \{(m, D) : 0.7 = \underline{D} \le D \le 1 = \overline{D}, m = \alpha(\overline{D}/N)^{1/2}, |\alpha| \le 5\}$. In this case, approximately the worst-case prior distribution is concentrated on three parameters and has the form: $\Pr(D = 1, \alpha = 1.9) = 0.3$, $\Pr(D = 1, \alpha = -2.2) = 0.15$, $\Pr(D = 1, \alpha = -5) = 0.55$, the corresponding Bayesian risk is approximately 0.41. Then, regrets were calculated for the strategy found. In Fig. 4.1, lines 1, 2, 3, 4 correspond to regrets at variance values of $D = 1$, 0.9, 0.8, 0.7, calculated in increments of 0.5. One can see that the maximum values of the regret are approximately the same as the Bayesian risk calculated with respect to the worst-case prior distribution.

Finally, in Fig. 4.2, we present approximate finding minimax strategy and minimax risk in the case of estimating the variance by incomes within batches. In considered case, $K = 12$, $M_2 = 5$, $M_1 = 1$ and, therefore, the total number of batches is $N = 60$. The set of parameters is again $\Theta = \{(m, D) : 0.7 = \underline{D} \le D \le 1 = \overline{D}, m = \alpha(\overline{D}/N)^{1/2}, |\alpha| \le 5\}$. The results are presented for a Bayesian strategy computed with respect to a prior distribution $\Pr(D = 1, \alpha = 3.5) = 0.16$, $\Pr(D = 1, \alpha = -5) = 0.84$, corresponding normalized Bayesian risk is approximately 0.47. For determined strategy, the regrets corresponding to variance values of $D = 1$, 0.9, 0.8, 0.7, are presented by lines 1, 2, 3, 4 respectively, their maximum

is approximately 0.48. Calculations of Bayesian risk were performed using formulas (3.12)–(3.15), which were ssimplified taking into account the fact that if the strategy switches to the second action, it will apply it until the end of control. In invariant form, these simplified formulas are presented in theorem 9. When performing numerical integration, X varied in the range from -18 to 18 in increments of 0.15, and S varied from 0.5 to 120.5 in increments of 1. Since, a function $H(X, S, k, Y, U)$ has no singularities if $M_2 = 5$, now there is no need to provide a small increment in S.

5 Conclusion

We considered a Gaussian one-armed bandit problem with both unknown the mathematical expectation and the variance. Such a problem arises when optimizing batch data processing if the number of data batches and their sizes have small or moderate volumes. We obtained recursive equations for computing Bayesian risk and regret in the usual and invariant form with a control horizon equal to one. This makes it possible to compute Bayesian strategy and risk for any number of data multiples of the number of batches processed. To find minimax strategy and risk, one should first find the worst-case prior distribution at which the Bayesian risk is maximal. Minimax strategies and risk coincide with Bayesian ones calculated with respect to the worst-case prior distribution. The presented results of numerical experiments confirm the theoretical results.

References

[1] D.A. Berry, B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*, Chapman and Hall, London, New York, 1985.
[2] E.L. Presman, I.M. Sonin, *Sequential Control with Incomplete Information*, Academic, New York, 1990.
[3] M.L. Tsetlin, *Automaton Theory and Modeling of Biological Systems*, Academic, New York, 1973.
[4] V.I. Varshavsky, *Kollektivnoe povedenie avtomatov (Collective Behavior of Automata)*, Nauka, Moscow, 1973. Translated under the title Kollektives Verhalten von Automaten, Warschawski, W.I., Akademie, Berlin, 1978.
[5] M. E. Hellman, T. M. Cover, *Learning with finite memory*, Ann. Math. Statist., **41**:3 (1970), 765—782
[6] V.G. Sragovich, *Mathematical Theory of Adaptive Control*, World Sci., Singapore, 2006.
[7] A.V. Nazin, A.S. Poznyak, *Adaptivnyi vybor variantov: rekurrentnye algoritmy (Adaptive Choice between Alternatives: Recursive Algorithms)*, Nauka, Moscow, 1986.
[8] J.C. Gittins, *Multi-Armed Bandit Allocation Indices*, Wiley-Interscience Series in Systems and Optimization, John Wiley & Sons, Ltd., Chichester, 1989.
[9] T. Lattimore, C. Szepesvari, *Bandit Algorithms*, Cambridge University Press, Cambridge, 2020.
[10] A.V. Kolnogorov, *One-armed bandit problem for parallel data processing systems*, Problems of Information Transmission, **51**:2 (2015), 177–191.
[11] T.L. Lai, B. Levin, H. Robbins, D. Siegmund, *Sequential medical trials*, Proc. Natl. Acad. Sci. USA, **77**:6 (1980), 3135–3138.

[12] V. Perchet, P. Rigollet, S. Chassang, E. Snowberg, *Batched bandit problems*, Ann. Statist., **44**:2 (2016), 660–681.
[13] W. Vogel, *An Asymptotic minimax theorem for the two-armed bandit problem*, Ann. Math. Statist., **31**: 2 (1960), 444–451.
[14] A.V. Kolnogorov, *Gaussian two-armed bandit: limiting description*, Probl. Inf. Transm., **56**:3 (2020), 278–301.
[15] R.N. Bradt, S.M. Johnson, S. Karlin, *On sequential designs for maximizing the sum of n observations*, Ann. Math. Statist., **27** (1956), 1060–1074.
[16] H. Chernoff, S.N. Ray, *A Bayes sequential sampling inspection plan*, Ann. Math. Statist., **36** (1965), 1387–1407.
[17] A. Kolnogorov, *Gaussian one-armed bandit problem*, In: 2021 XVII International symposium "Problems of redundancy in information and control systems" (REDUNDANCY), (2021), 74–79.
[18] A.V. Kolnogorov, *Gaussian one-armed bandit with both unknown parameters*, Siberian Electronic Mathematical Reports, **19** (2022), 639–650.
[19] T.L. Lai, *Adaptive treatment allocation and the multi-armed bandit problem*, Ann. Statist., **25** (1987), 1091–1114.
[20] P. Auer, N. Cesa-Bianchi, P. Fisher, *Finite-time analysis of the multi-armed bandit problem*, Machine Learning, **47**: 2–3 (2002), 235–256.

Alexander Valerianovich Kolnogorov
Yaroslav-the-Wise Novgorod State University,
ul. Bolshaya St.-Petersburgskaya, 41,
173003, Velikiy Novgorod, Russia
E-mail address: kolnogorov53@mail.ru