MSC 62C10, 62L05, 91A35 **DOI: 10.14529/mmp240103**

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

A. V. Kolnogorov, Yaroslav-the-Wise Novgorod State University, Veliky Novgorod, Russian Federation, kolnogorov53@mail.ru

We consider the one-armed bandit problem in 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 the variance of income corresponding to the second action. 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 regret, which we then present in an invariant form with a control horizon equal to one. This makes it possible to obtain the estimates of Bayesian and minimax risk that are valid for all control horizons multiples to the number of processed batches.

Keywords: one-armed bandit; batch processing; Bayesian and minimax approaches; invariant description.

Introduction

We consider the one-armed bandit problem, which is a special case of the two-armed bandit problem (see, e.g., [1, 2]). The name originates from a slot machine with two arms (in what follows, called actions), each choice of which is accompanied by a random income of the player. The goal of the player is to maximize his/her total expected income. Distributions of incomes depend only on the currently selected actions, are fixed during the game but the player does not have a complete information about them. In particular, a one-armed bandit occurs if the characteristics of only the first action are a priori known. The problem has numerous applications in behavior modelling [3], adaptive control in a random environment [4], medicine, internet technologies, data processing [5, 6].

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

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

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 $D_1 = \mathbf{D}(\xi_n|y_n = 1)$ 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 $\Theta = \{(m, D) : |m| \le C < +\infty,\ 0 < \underline{D} \le D \le \overline{D} < +\infty\}$ is known.

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, which in considered case are the cumulative income and s^2-statistics for the application of the second action. The cumulative 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.

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 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

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

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

Let's explain the choice of a normal distribution of incomes. We consider the problem in the application to batch data processing if there are two alternative processing methods with different efficiencies. In batch processing, the data is divided into equal batches, the same processing method (action) is applied to all the data in the batch and the cumulative numbers of successfully processed data in the batches (incomes) are used for the control. By virtue of the central limit theorem, these incomes have approximately Gaussian distributions if the batch sizes are large enough. 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 is large enough. For example, it is shown in [7] that in the case of splitting data into 50 batches, the maximum regret grows by only 3% compared to its maximum value corresponding to the optimal processing. 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, 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. For an overview of the results of this approach and references see, e.g. [8].

Let a prior distribution density $\lambda(\theta) = \lambda(m,D)$ be given on the set Θ. We assume that the conditions $\int_\Theta m^-\,\lambda(\theta)d\theta > 0$ and $\int_\Theta m^+\,\lambda(\theta)d\theta > 0$ are met; otherwise, the more profitable action is a priori known. We use here the standard notations $m^+ = \max(m,0)$, $m^- = \max(-m,0)$ and denote $d\theta = dmdD$. The averaged regret is

$$L_N(\sigma,\lambda) = \int_\Theta L_N(\sigma,\theta)\lambda(\theta)d\theta. \tag{2}$$

Bayesian and minimax risks are

$$\begin{gathered} R_N^B(\lambda) = \inf_{\{\sigma\}} L_N(\sigma,\lambda), \\ R_N^M(\Theta) = \inf_{\{\sigma\}}\sup_\Theta L_N(\sigma,\theta), \end{gathered} \tag{3}$$

the corresponding optimal strategies σ^B and σ^M are called the Bayesian and the minimax strategy. Bayesian approach allows one to find Bayesian strategy and Bayesian risk for any

prior distribution by solving the recursive Bellman-type equation. Its disadvantage 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 $L_N(\sigma,\theta) \le R_N^M(\Theta)$ for all $\theta \in \Theta$. Moreover, an asymptotic theorem, providing the estimate of the minimax risk of the order of $N^{1/2}$, is well-known [9]. However, there is no a direct method for finding minimax risk and minimax strategy. To find them, the main theorem of 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 at which the Bayesian risk reaches its maximum. And the minimax strategy coincides with the corresponding Bayesian one. In more detail, using the main theorem of the game theory is presented in [7, 10].

The one-armed bandit problem was first considered in [11, 12] in the Bayesian setting for a Bernoullian two-armed bandit which incomes take the values 0 and 1. In [11], a recursive algorithm for finding Bayesian strategy and Bayesian risk was described. The asymptotic properties were established in [12]. In [11], 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 (see [7, 10, 13]). It also remains true in the statement considered in section 1. The prove is similar to presented in [7, 10, 13] and is therefore omitted.

Let's indicate what is the difference between the considered approach and the one presented in [7, 10]. In [7, 10], the case of a priori known variance is considered, which takes place 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.

The rest of the article is as follows. In section 1, recursive equations are obtained for finding Bayesian strategies, risks and regrets in the ordinary and invariant forms. Presented here ordinary forms of equations are more convenient for computations than obtained in [13]. The advantage of the invariant descriptions is that they depend only on the number of processed batches and, hence, make it possible to obtain asymptotic estimates of Bayesian risk and regret. Section 2 presents numerical results. Section 3 contains the conclusion.

1. Recursive Equations for Computing Bayesian Risks and Regrets

Consider the batch processing. Let the total number of data be $N = MK$, where M is the batch size, K is the number of batches. The same action is applied to each M sequentially incoming data, so that the income for processing the kth batch is $x_k = \sum_{n=(k-1)M+1}^{kM} \xi_n$. The mathematical expectation and the variance of income when processing the batch with the second action are Mm and MD, the mathematical expectation of income when processing the batch with the first action is still 0.

Recall noted in **Introduction** property of the Bayesian strategy: the use of the second action can only start at the beginning of control. Let x_i, $i = 1, \dots, k$, be incomes obtained in response to the application of the second action at the beginning of control. The following sufficient statistics can be used in considered case $X = \sum_{i=1}^{k} x_i$, $S = \sum_{i=1}^{k} x_i^2 - X^2/k$, 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 $k = 0$ and $S = 0$ if $k = 0, 1$.

Let's consider how to update X and S. Assume that $k \geq 1$, and let $x_{k+1} = Y$ be a new income. Then $X_{new} = \sum_{i=1}^{k+1} x_i = X+Y$, $S_{new} = \left(\sum_{i=1}^{k+1} x_i^2\right) - X_{new}^2/(k+1) = \left(\sum_{i=1}^{k} x_i^2\right) + Y^2 - (X+Y)^2/(k+1) = S + M\Delta(X,k,Y)$, where $M\Delta(X,k,Y) = (X-kY)^2\{k(k+1)\}^{-1}$. If $k=0$ then $M\Delta(0,0,Y) = 0$. Hence, X, S are updated according to the rule

$$\begin{gathered} X \leftarrow X+Y, \quad S \leftarrow S + M\Delta(X,k,Y), \\ \text{with } \Delta(0,0,Y) = 0, \quad \Delta(X,k,Y) = (X-kY)^2\{Mk(k+1)\}^{-1}, \quad k \geq 1. \end{gathered} \tag{4}$$

Given a prior distribution density $\lambda(m.D)$, let's describe a posterior distribution density. Consider a chi-squared distribution density with k degrees of freedom $\chi_k^2(x) = \{2^{k/2}\Gamma(k/2)\}^{-1}x^{\frac{k}{2}-1}e^{-\frac{x}{2}}$, $x \geq 0$, $k \geq 1$. We introduce the function $\mathbf{F}(X,S,k|m,D) = f_{kMD}(X|kMm)\psi_{k-1}(S/(MD))$, where $\psi_{k-1}(S/(MD)) = (MD)^{-1}\chi_{k-1}^2(S/(MD))$. Note that defined above cumulative income X and s^2-statistics S after processing k batches have exactly the distribution densities $f_{kMD}(X|kMm)$ and $\psi_{k-1}(S/(MD))$ for $k \geq 1$ and $k \geq 2$ respectively. Since X and S are independent random variables, the posterior distribution density is $\lambda(m,D|X,S,k) = \mathbf{F}(X,S,k|m,D)\lambda(m.D)/P(X,S,k)$ with $P(X,S,k) = \iint_\Theta \mathbf{F}(X,S,k|m,D)\lambda(m.D)dmdD$ for all $k \geq 2$. These approach is used in [13]. However, recursive equation becomes simpler if the posterior distribution is defined in the following equivalent way. Denote $\tilde{\mathbf{F}}(0,0,0|m,D) = 1$, $\tilde{\mathbf{F}}(X,0,1|m,D) = D^{-1/2}\tilde{f}_{kMD}(X|kMm)$ and

$$\tilde{\mathbf{F}}(X,S,k|m,D) = D^{-3/2}\tilde{f}_{kMD}(X|kMm)\tilde{\psi}_{k-1}(S/(MD)), \quad if\ k \geq 2, \tag{5}$$

where $\tilde{f}_D(x|m) = \exp(-(x-m)^2/(2D))$, $\tilde{\psi}_{k-1}(s) = s^{\frac{k-1}{2}-1}\exp(-s/2)$. Clearly, the posterior distribution density is

$$\begin{gathered} \lambda(m,D|X,S,k) = \frac{\tilde{\mathbf{F}}(X,S,k|m,D)\lambda(m,D)}{\tilde{P}(X,S,k)}, \\ \text{with } \ \tilde{P}(X,S,k) = \iint_\Theta \tilde{\mathbf{F}}(X,S,k|m,D)\lambda(m,D)dmdD \end{gathered} \tag{6}$$

and this formula is valid for all $k = 0,1,2\ldots$.

Let $R^B(X,S,k)$ denote a Bayesian risk on the remaining control horizon $k+1,\ldots,K$ computed with respect to the posterior distribution $\lambda(m,D|X,S,k)$, i.e., $R^B(X,S,k) = R^B_{M(K-k)}(\lambda(m,D|X,S,k))$. Since the use of the second action can start only at the beginning of control and switching to the first action is performed until the end of the control, a standard recursive equation for computing a Bayesian risk is as follows

$$R^B(X,S,k) = \min\left(R_1^B(X,S,k), R_2^B(X,S,k)\right), \tag{7}$$

where $R_1^B(X,S,k) = R_2^B(X,S,k) = 0$ if $k = K$ and

$$\begin{gathered} R_1^B(X,S,k) = (K-k)\iint_\Theta Mm^+\lambda(m,D|X,S,k)dmdD, \\ R_2^B(X,S,k) = \iint_\Theta \lambda(m,D|X,S,k)\times \\ \times\left(Mm^- + \int_{-\infty}^{\infty} R^B(X+Y,S+M\Delta(X,k,Y),k+1)f_{MD}(Y|Mm)dY\right)dmdD, \end{gathered} \tag{8}$$

if $0 \le k \le K-1$. In the second equation (8), we used (4). Bayesian strategy prescribes, when processing the batch with the number $k+1$, to choose an action corresponding to the smaller of the current values $R_1^B(X,S,k)$, $R_2^B(X,S,k)$. In the case of a draw, the choice can be arbitrary. If the first action is chosen once it will be applied until the end of the control. Bayesian risk (3) is

$$R_N^B(\lambda) = R^B(0,0,0). \tag{9}$$

Let's present another form of recursive equation. We put $R_\ell(X,S,k) = R_\ell^B(X,S,k) \times \tilde{P}(X,S,k)$, $\ell = 1,2$, where $\tilde{P}(X,S,k)$ is defined in (6).

Theorem 1. *Consider the recursive equation*

$$R(X,S,k) = \min\left(R_1(X,S,k), R_2(X,S,k)\right), \tag{10}$$

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

$$\begin{gathered} R_1(X,S,k) = (K-k)MG_1(X,S,k), \\ R_2(X,S,k) = MG_2(X,S,k) + \\ + \int_{-\infty}^{\infty} R(X+Y, S+M\Delta(X,k,Y), k+1) H(X,S,k,Y)dY, \end{gathered} \tag{11}$$

if $0 \le k \le K-1$. *Here* $\Delta(X,k,Y)$ *is given by* (4),

$$\begin{aligned} G_1(X,S,k) &= \iint_\Theta m^+ \tilde{\mathbf{F}}(X,S,k|m,D)\lambda(m,D)dmdD, \\ G_2(X,S,k) &= \iint_\Theta m^- \tilde{\mathbf{F}}(X,S,k|m,D)\lambda(m,D)dmdD, \end{aligned} \tag{12}$$

and $H(0,0,0,Y) = (2\pi M)^{-1/2}$, $H(X,0,1,Y) = (2\pi M)^{-1/2}\Delta^{1/2}(X,1,Y)$,

$$H(X,S,k,Y) = \frac{1}{(2\pi)^{1/2}} \times \frac{S^{(k-1)/2-1}}{(S+M\Delta(X,k,Y))^{k/2-1}}, \qquad if\ k \ge 2. \tag{13}$$

When processing the batch number $k+1$, *Bayesian strategy prescribes to choose the action corresponding to the smaller value of* $R_1(X,S,k)$, $R_2(X,S,k)$; *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* (3) *is*

$$R_N(\lambda) = R(0,0,0). \tag{14}$$

Proof. Let's multiply (7), (8) by $\tilde{P}(X,S,k)$ defined in (6). We obtain (10), (11) with $G_1(X,S,n_2)$, $G_2(X,S,k)$ defined in (12). Let Δ in formulas below be given by (4). Denote $D' = MD$, $m' = Mm$. The function $H(X,S,k,Y)$ is

$$\frac{\iint_\Theta \tilde{\mathbf{F}}(X,S,k|m,D) f_{D'}(Y|m')\lambda(m,D)dmdD}{\tilde{P}(X+Y, S+M\Delta, k+1)} = \frac{\tilde{\mathbf{F}}(X,S,k|m,D) f_{D'}(Y|m')}{\tilde{\mathbf{F}}(X+Y, S+M\Delta, k+1|m,D)}$$

with $\tilde{\mathbf{F}}(\cdot)$ given by (5). Consider the case $k \ge 2$, We have

$$H(X,S,k,Y) = \frac{\tilde{f}_{kD'}(X|km') f_{D'}(Y|m')}{\tilde{f}_{(k+1)D'}(X+Y|(k+1)m')} \times \frac{\tilde{\psi}_{k-1}(S/D')}{\tilde{\psi}_k((S+M\Delta)/D')}. \tag{15}$$

Since

$$\frac{\tilde{f}_{kD'}(X|km')f_{D'}(Y|m')}{\tilde{f}_{(k+1)D'}(X+Y|(k+1)m')} = \left(\frac{1}{2\pi D'}\right)^{1/2} \times \exp\left(-\frac{M\Delta}{2D'}\right), \tag{16}$$

$$\begin{aligned}\frac{\tilde{\psi}_{k-1}(S/D')}{\tilde{\psi}_k((S+M\Delta)/D')} &= \frac{(S/D')^{(k-1)/2-1}}{((S+M\Delta)/D')^{k/2-1}} \times \frac{\exp\left(-S/(2D')\right)}{\exp\left(-(S+M\Delta)/(2D')\right)} \\ &= (D')^{1/2} \times \frac{S^{(k-1)/2-1}}{(S+M\Delta)^{k/2-1}} \times \exp\left(\frac{M\Delta}{2D'}\right),\end{aligned} \tag{17}$$

it follows from (15) – (17) that $H(X,S,k,Y)$ is given by (13) for $k \geq 2$. The cases $k=1$ and $k=0$ are similarly considered. Since, $\tilde{P}(0,0,0)=1$ then (14) follows from (9).

□

Let's give an invariant form of equation for computing Bayesian strategy and risk. We choose the following set of parameters $\Theta_N = \{(m,D) : |m| \leq c(D/N)^{1/2}, \underline{D} \leq D \leq \overline{D}\}$, where $c>0$, $0<\underline{D} \leq \overline{D}<\infty$. 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 $\Theta_N=\{(\alpha,\beta) : \underline{D}/\overline{D}=\beta_0 \leq \beta \leq 1, |\alpha| \leq c\beta^{1/2}\}$.

Consider the change of variables $X=x(\overline{D}N)^{1/2}$, $Y=y(\overline{D}N)^{1/2}$, $S=s\overline{D}M$, $k=tK$, $M/N=K^{-1}=\varepsilon$, $m=\alpha(\overline{D}/N)^{1/2}$, $D=\beta\overline{D}$, $\lambda(m,D)=(N/\overline{D}^{\,3})^{1/2}\varrho(\alpha,\beta)$. Let's put $R_\ell(0,0,0)=(\overline{D}N)^{1/2}r_\ell(0,0,0)$, $R_\ell(X,0,1)=(\overline{D}\,N)^{1/2}(\overline{D})^{-1/2}r_\ell(x,0,\varepsilon)$ and $R_\ell(X,S,k)=(\overline{D}\,N)^{1/2}(\overline{D})^{-3/2}r_\ell(x,s,t)$ if $k\geq 2$, $\ell=1,2$. The following theorem is valid.

Theorem 2. *To find the Bayesian risk, the recursive equation should be solved*

$$r(x,s,t)=\min\left(r_1(x,s,t),r_2(x,s,t)\right), \tag{18}$$

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

$$\begin{aligned}r_1(x,s,t)&=(1-t)g_1(x,s,t),\\ r_2(x,s,t)&=\varepsilon g_2(x,s,t)+\int_{-\infty}^{\infty} r(x+y,s+\delta(x,t,y),t+\varepsilon)h(x,s,t,y)dy,\end{aligned} \tag{19}$$

if $0 \leq t \leq 1-\varepsilon$. *Here* $g_1(x,s,t) = \iint_{\Theta_N} \alpha^{+}\tilde{\mathbf{f}}(x,s,t|\alpha,\beta)\varrho(\alpha,\beta)d\alpha d\beta$, $g_2(x,s,t) = \iint_{\Theta_N} \alpha^{-}\tilde{\mathbf{f}}(x,s,t|\alpha,\beta)\varrho(\alpha,\beta)d\alpha d\beta$ *with* $\tilde{\mathbf{f}}(0,0,0|\alpha,\beta)=1$, $\tilde{\mathbf{f}}(x,0,\varepsilon|\alpha,\beta)=\beta^{-1/2}\tilde{f}_{t\beta}(x|t\alpha)$ *and* $\tilde{\mathbf{f}}(x,s,t|\alpha,\beta)=\beta^{-3/2}\tilde{f}_{t\beta}(x|t\alpha)\tilde{\psi}_{k_2-1}(s/\beta)$ *if* $t\geq 2\varepsilon$. *The function* $h(x,s,t,y)$ *is as follows:* $h(0,0,0,y)=(2\pi\varepsilon)^{-1/2}$, $h(x,0,\varepsilon,y)=(2\pi\varepsilon)^{-1/2}\delta^{1/2}(x,\varepsilon,y)$ *and* $h(x,s,t,y)=(2\pi\varepsilon)^{-1/2}s^{(k-1)/2-1}/(s+\delta(x,t,y))^{k/2-1}$ *if* $t\geq 2\varepsilon$ *with* $\delta(x,t,y)=(\varepsilon x-ty)^2\{\varepsilon t(t+\varepsilon)\}^{-1}$. *Bayesian risk* (3) *is*

$$R_N(\lambda)=(\overline{D}N)^{1/2}r(0,0,0). \tag{20}$$

Bayesian strategy prescribes to choose the action corresponding to the currently smaller value of $r_1(x,s,t)$, $r_2(x,s,t)$*; 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.*

Proof. The theorem is proved by performing the presented above change of variables in (10) – (14).

□

Let's present a recursive equation for computing the regret (2). We restrict consideration to strategies σ which can start the use of the second action only at the beginning of the control and are described by a set of probabilities $\sigma_\ell(X,S,k) = \Pr(y_{k+1} = \ell|X,S,k)$, $\ell = 1,2$; $k = 0,\dots,K-1$. Similarly to theorem 1, the following theorem is valid.

Theorem 3. *Consider a recursive equation*

$$L(X,S,k) = \sum_{\ell=1}^{2} \sigma_\ell(X,S,k) L_\ell(X,S,k), \tag{21}$$

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

$$\begin{gathered} L_1(X,S,k) = (K-k)MG_1(X,S,k), \\ L_2(X,S,k) = MG_2(X,S,k) + \\ + \int_{-\infty}^{\infty} L(X+Y, S+M\Delta(X,k,Y), k+1) H(X,S,k,Y)dY, \end{gathered} \tag{22}$$

if $0 \le k \le K-1$. *Here* $G_1(X,S,k)$, $G_2(X,S,k)$ *are given by* (12), $H(X,S,k,Y)$ *is given by* (13). *A regret* (2) *is*

$$L_N(\sigma,\lambda) = L(0,0,0). \tag{23}$$

To obtain the regret (1) *one should choose a degenerate prior distribution density concentrated at the parameter* θ.

To present invariant form of the equation for computing the regret, let's make additional change $\sigma_\ell(X,S,k) = \sigma_\ell(x,s,t)$, $L_\ell(0,0,0) = (\overline{D}N)^{1/2} l_\ell(0,0,0)$, $L_\ell(X,0,1) = (\overline{D}N)^{1/2}(\overline{D})^{-1/2} l_\ell(x,0,\varepsilon)$ and $L_\ell(X,S,k) = (\overline{D}N)^{1/2}(\overline{D})^{-3/2} l_\ell(x,s,t)$ if $k \ge 2$, $\ell = 1,2$.

Theorem 4. *To find the regret, one should solve the recursive equation*

$$l(x,s,t) = \sum_{\ell=1}^{2} \sigma_\ell(x,s,t) l_\ell(x,s,t), \tag{24}$$

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

$$\begin{gathered} l_1(x,s,t) = (1-t)g_1(x,s,t), \\ l_2(x,s,t) = \varepsilon g_2(x,s,t) + \int_{-\infty}^{\infty} l(x+y, s+\delta(x,t,y), t+\varepsilon) h(x,s,t,y)dy, \end{gathered} \tag{25}$$

if $0 \le t \le 1-\varepsilon$. *Here* $g_1(x,s,t)$, $g_2(x,s,t)$, $h(x,s,t,y)$ *and* $\delta(x,t,y)$ *are described in theorem 2. A regret* (2) *is*

$$L_N(\sigma,\theta) = (\overline{D}N)^{1/2} l(0,0,0). \tag{26}$$

2. Numerical Results

We computed Bayesian risk for $K = 18$ and $M = 1$ using formulas (10) – (14). 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 $R(X,S,k)$,

at the point $S = 0$ if $k = 2$. The set of parameters was $\Theta = \{\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22}\}$ with $\theta_{11} = (m_p, \overline{D})$, $\theta_{12} = (m_n, \overline{D})$, $\theta_{21} = (m_p, \underline{D})$, $\theta_{22} = (m_n, \underline{D})$, where $\overline{D} = 1$, $\underline{D} = 0,7$, $m_p = 1,5(\overline{D}/N)^{1/2}$, $m_n = -2,5(\overline{D}/N)^{1/2}$. For prior distributions $\lambda = (\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22})$, where $\lambda_{ij} = \Pr(\theta = \theta_{ij})$, values of normalized Bayesian risks $r_N^B(\lambda) = (\overline{D}N)^{-1/2} R_N^B(\lambda)$ are presented in the Table. Then we approximated risks from the Table by risks and

Table

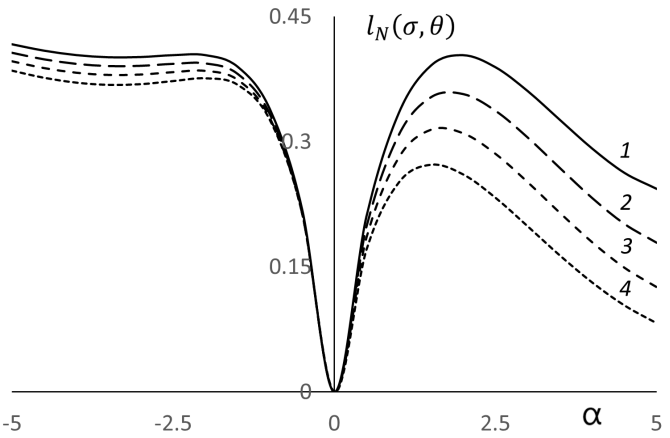
Normalized Bayesian risks and their estimates

λ	λ_{11}	λ_{12}	λ_{21}	λ_{22}	$r_N^B(\lambda)$	$r_N^B(\lambda')$	$l_N^B(\sigma(\lambda'), \lambda'')$	$r_N^B(\lambda'')$	$l_N^B(\sigma(\lambda''), \lambda')$	$l_N(\lambda)$
1	0,2	0,3	0,2	0,3	0,36	0,39	0,33	0,33	0,39	0,36
2	0,2	0,3	0,3	0,2	0,34	0,39	0,31	0,33	0,39	0,35
3	0,1	0,4	0,4	0,1	0,33	0,30	1,20	0,18	0,60	0,57

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 using (21)–(23). 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 $l_N(\lambda) = \mu_1 \left(\mu_1 r_N^B(\lambda') + \mu_2 l_N^B(\sigma(\lambda'), \lambda'')\right) + \mu_2 \left(\mu_1 l_N^B(\sigma(\lambda''), \lambda') + \mu_2 r_N^B(\lambda'')\right)$.

The results are also presented in the Table. 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 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 strategy and risk coincide with the Bayesian ones computed with respect to the worst-case prior distribution at which



Normalized regrets for different variances

the Bayesian risk is maximal. As an example, consider the approximate finding the minimax risk at $K = 18$, $M = 1$ on the set $\Theta = \{(m, D) : 0,7 = \underline{D} \le D \le 1 = \overline{D}, m = \alpha(\overline{D}/N)^{1/2}, |\alpha| \le 5\}$. In this case, approximately the worst-case prior distribution was found as $\Pr(D = 1, \alpha = 1,9) = 0,3$, $\Pr(D = 1, \alpha = -2,2) = 0,15$, $\Pr(D = 1, \alpha = -5) = 0,55$, the corresponding Bayesian risk is approximately 0,41. Then, regrets were calculated for the strategy found. In Figure, lines 1, 2, 3, 4 correspond to regrets, calculated in increments of 0,5, at variance values of $D = 1$, $0,9$, $0,8$, $0,7$. 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.

Conclusion

We obtained recursive equations for computing Bayesian risk and regret in the usual and invariant form which make it possible to compute them for any number of data multiples of the number of batches. To find minimax strategy and risk, one should determine them as Bayesian ones on the worst-case prior distribution.

Acknowledgements. *The research was supported by Russian Science Foundation, project number 23-21-00447, https://rscf.ru/en/project/23-21-00447/.*

References

1. Berry D.A., Fristedt B. *Bandit Problems: Sequential Allocation of Experiments*. London, New York, Chapman and Hall, 1985.
2. Presman E.L., Sonin I.M. *Sequential Control with Incomplete Information*. New York, Academic Press, 1990.
3. Tsetlin M.L. *Automaton Theory and Modeling of Biological Systems*. New York, Academic Press, 1973.
4. Sragovich V.G. *Mathematical Theory of Adaptive Control*. Singapore, World Scientific, 2006.
5. Gittins J.C. *Multi-Armed Bandit Allocation Indices*. Chichester, John Wiley and Sons, 1989.
6. Lattimore T., Szepesvari C. *Bandit Algorithms*. Cambridge, Cambridge University Press, 2020.
7. Kolnogorov A.V. One-Armed Bandit Problem for Parallel Data Processing Systems. *Problems of Information Transmission*, 2015, vol. 51, no. 2, pp. 177–191. DOI: 10.1134/S0032946015020088
8. Perchet V., Rigollet P., Chassang S., Snowberg E. Batched Bandit Problems. *The Annals of Statistics*, 2016, vol. 44, no. 2, pp. 660–681. DOI: 10.1214/15-AOS1381
9. Vogel W. An Asymptotic Minimax Theorem for the Two-Armed Bandit Problem. *The Annals of Mathematical Statistics*, 1960, vol. 31, no. 2, pp. 444–451.
10. Kolnogorov A. Gaussian One-Armed Bandit Problem. *2021 XVII International Symposium "Problems of Redundancy in Information and Control Systems"*. Moscow, Institute of Electrical and Electronics Engineers, 2021, pp. 74–79. DOI: 10.1109/REDUNDANCY52534.2021.9606464
11. Bradt R.N., Johnson S.M., Karlin S. On Sequential Designs for Maximizing the Sum of n Observations. *The Annals of Mathematical Statistics*, 1956, vol. 27, pp. 1060–1074. DOI: 10.1214/aoms/1177728073

12. Chernoff H., Ray S.N. A Bayes Sequential Sampling Inspection Plan. *The Annals of Mathematical Statistics*, 1965, vol. 36, pp. 1387–1407. DOI: 10.1214/aoms/1177699898

13. Kolnogorov A.V. Gaussian One-Armed Bandit with Both Unknown Parameters. *Siberian Electronic Mathematical Reports*, 2022, vol. 19, no. 2, pp. 639–650. Available at: http://semr.math.nsc.ru/v19n2ru.html.

Received November 22, 2023

УДК 519.244, 519.83 **DOI: 10.14529/mmp240103**

ИНВАРИАНТНОЕ ОПИСАНИЕ УПРАВЛЕНИЯ В ЗАДАЧЕ О ГАУССОВСКОМ ОДНОРУКОМ БАНДИТЕ

А.В. Колногоров, Новгородский государственный университет им. Ярослава Мудрого, г. Великий Новгород, Российская Федерация

Рассматривается задача об одноруком бандите в приложении к пакетной обработке данных, если имеются два альтернативных метода обработки с разной эффективностью, причем эффективность второго метода априори неизвестна. В процессе обработки необходимо определить наиболее эффективный метод и обеспечить его преимущественное использование. Обработка выполняется пакетами, поэтому распределение доходов является гауссовским. Мы рассматриваем случай априори неизвестных математического ожидания и дисперсии одношагового дохода, соответствующих второму действию. Этот случай описывает ситуацию, когда сами пакеты и их количество имеют умеренные или небольшие объемы. Получены рекуррентные уравнения для вычисления байесовского риска и функции потерь, которые затем представлены в инвариантном виде с горизонтом управления, равным единице. Это позволяет получить оценки байесовского и минимаксного рисков, которые справедливы для всех горизонтов управления, кратных количеству обработанных пакетов.

Ключевые слова: однорукий бандит; пакетная обработка; байесовский и минимаксный подходы; инвариантное описание.

Александр Валерианович Колногоров, доктор физико-математических наук, профессор, кафедра «Прикладная математика и информатика», Новгородский государственный университет им. Ярослава Мудрого (г. Великий Новгород, Российская Федерация), kolnogorov53@mail.ru.

Поступила в редакцию 22 ноября 2023 г.