S©MR ISSN 1813-3304

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ

Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru

Том 19, №2, стр. 639–650 (2022) УДК 519.244, 519.83
DOI 10.33048/semi.2022.19.053 MSC 62C10, 62L05, 91A35

GAUSSIAN ONE-ARMED BANDIT WITH BOTH UNKNOWN PARAMETERS

A.V. KOLNOGOROV

Abstract. We consider the one-armed bandit problem as applied to data processing. We assume that there are two alternative processing methods and efficiency of the second method is a priory unknown. During control process, one has to determine if the second method is more efficient than the first one and to provide a primary application of the most efficient method. The essential feature of considered approach is that the data is processed in batches and cumulative incomes in batches are used for the control. If the sizes of batches are large enough then according to the central limit theorem incomes in batches are approximately Gaussian. Also if the sizes of batches are large, one can estimate the variances of incomes during the processing initial batches and then use these estimates for the control. However, for batches of moderate sizes it is reasonable to estimate unknown variances throughout the control process. This optimization problem is described by Gaussian one-armed bandit with both unknown parameters. Given a prior distribution of unknown parameters of the second action, we derive a recursive Bellman-type equation for determining corresponding Bayesian strategy and Bayesian risk. Minimax strategy and minimax risk are searched for according to the main theorem of the game theory as Bayesian ones corresponding to the worst-case prior distribution.

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

Kolnogorov, A.V., Gaussian one-armed bandit with both unknown parameters.

The work is supported by RFFI (grant 20-01-00062).
Received April, 20, 2022, published September, 2, 2022.

1. Introduction

The two-armed bandit problem is a famous mathematical optimal control problem (see, e.g., [1, 2, 3]). The name comes from the slot machine with two arms. Choosing any arm, the gambler gets a random income, which distribution is fixed during the game against the slot, depends only on currently chosen arm, but is unknown to the gambler. The gambler has to play N times against the two-armed bandit, the magnitude N is known to him. The goal is to maximize the mathematical expectation of the total income. To achieve the goal, the gambler should, observing the statistics of the game, determine the most profitable arm and provide its preferential usage. In the sequel, the arms are also called actions. This control procedure involves the so-called "information vs control" dilemma, which states that for the gambler it would be better always to choose the action corresponding to the largest mathematical expectation of one-step income. However, in order to determine this preferable action, both actions should be applied for observing a statistics of the game and this reduces the total expected income. The problem is also well-known as the problem of rational behavior in a random environment (see, e.g., [4]) and the problem of adaptive control of stochastic systems (see, e.g., [5]). It has applications in medicine, IT and internet technologies, optimization of research projects, etc [6].

One-armed bandit is often considered as a two-armed bandit with known distribution of income corresponding to the application of the first arm. In the case of Bernoulli incomes, which can take two possible values 1 and 0, it was considered in [7, 8]. In the present article, we consider Gaussian one-armed bandit. Formally, Gaussian one-armed bandit is a controlled random process, which values ξ_n, $n = 1, 2, \ldots, N$, are interpreted as incomes, depend only on currently chosen actions y_n ($y_n \in \{1, 2\}$) and have normal (Gaussian) distribution

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

if $y_n = 2$. Here mathematical expectation m and the variance D of the income are assumed to be unknown. In the case of $y_n = 1$ let's denote the mathematical expectation and the variance of income by m_1, D_1. Then m_1 is assumed to be known and without loss of generality $m_1 = 0$ (otherwise, one can consider the random process $\xi_n - m_1$, $n = 1, 2, \ldots, N$). As for the variance D_1, its value is immaterial within the framework of researched problem.

Considered one-armed bandit can be described by unknown parameter $\theta = (m, D)$. However, the set of possible values of the parameter Θ is known. In what follows, we assume that

$$\Theta = \{(m, D) : |m| \le C_1 < \infty,\ 0 < C_2 \le D \le C_3 < \infty\}.$$

A control strategy σ at the instant of time $n+1$ determines the choice (possibly, randomized) of the action y_{n+1} depending on the whole currently known history of the process $y_1, \xi_1, \ldots, y_n, \xi_n$. However, it follows from equation (17)–(18), which describes computing the Bayesian risk, that actually it is sufficient to know four current values: n_1, n_2 – cumulative application counts of both actions ($n = n_1 + n_2$), and

$$X = \sum_{i: i \le n, y_i = 2} \xi_i, \quad S = \left(\sum_{i: i \le n, y_i = 2} \xi_i^2 \right) - X^2/n_2 \tag{2}$$

– cumulative income and $\chi^2_{n_2-1}$-statistics for applying the second action up to the time point n, which are used to estimate m and D. The cumulative income X_1 and $\chi^2_{n_1-1}$-statistics S_1 for applying the first action are not used because corresponding distribution is known. Therefore,

$$\sigma_\ell(n_1, X, S, n_2) = \Pr(y_{n+1} = \ell | n_1, X, S, n_2),\ \ell = 1, 2,\ n = 0, \ldots, N-1. \tag{3}$$

Note that X and S can not be computed for $n_2 < 1$ and $n_2 < 2$ respectively.

Let's define the goal of the control. If the parameter of the process was known, one should always apply the action corresponding to $\max(0, m)$, the total expected income is thus $N \max(0, m)$. But if the parameter was unknown and, hence, the strategy σ was applied, the total expected income is less than the maximal one by the value

$$L_N(\sigma, \theta) = N \max(0, m) - \mathbf{E}_{\sigma,\theta} \left(\sum_{n=1}^{N} \xi_n \right), \tag{4}$$

which is called the regret and is caused by incomplete information. Here $\mathbf{E}_{\sigma,\theta}$ denotes mathematical expectation with respect to the measure generated by strategy σ and parameter θ.

Let's assign a prior distribution density $\lambda(\theta) = \lambda(m, D)$ on Θ. A regret averaged over $\lambda(\theta)$ is defined as

$$L_N(\sigma, \lambda) = \int_\Theta L_N(\sigma, \theta)\lambda(\theta) d\theta = \iint_\Theta L_N(\sigma, (m, D))\lambda((m, D)) dm dD. \tag{5}$$

Bayesian risk computed with respect to a prior distribution density $\lambda(\theta)$ is

$$R_N(\lambda) = \inf_{\{\sigma\}} L_N(\sigma, \lambda), \tag{6}$$

corresponding optimal strategy σ^B is called the Bayesian strategy. Bayesian approach is very popular (see, e.g., [1]) because it allows to determine Bayesian strategy and Bayesian risk using dynamic programming technique. A disadvantage of this approach is the necessity to assign a prior distribution, for which there are no clear criteria. Minimax risk is

$$R_N^M(\Theta) = \inf_{\{\sigma\}} \sup_{\Theta} L_N(\sigma, \theta), \tag{7}$$

corresponding optimal strategy σ^M is called the minimax strategy. Minimax approach is robust, i.e., the application of the minimax strategy ensures the fulfillment of the inequality

$$L_N(\sigma^M, \theta) \le R_N^M(\Theta)$$

for all $\theta \in \Theta$. A disadvantage of the minimax approach is that there is no a direct method to determine minimax strategy and minimax risk. However, they can be found using the main theorem of the game theory according to the equality

$$R_N^M(\Theta) = R_N(\lambda^0) = \sup_\lambda R_N(\lambda), \tag{8}$$

i.e., minimax risk (7) is equal to Bayesian risk (6) computed over the worst-case prior distribution, at which the Bayesian risk attains its maximum value, and minimax strategy is equal to corresponding Bayesian strategy, too.

Let's explain why Gaussian one-armed bandit is considered. We study the problem in application to batch processing of incomes. For instance, it can be applied to batch data processing. Let's assume that there are two alternative methods of

data processing with a priori unknown efficiency of the second method. In this case, incomes correspond to successfully processed data and methods correspond to actions. In batch processing, the same actions are applied to batches of data and then cumulative incomes (cumulative numbers of successfully processed data in batches) are used for the control. According to the central limit theorem distributions of cumulative incomes are close to Gaussian. Note, that batch processing is more convenient than one-by-one processing because it allows to change actions more rarely. And if parallel processing is available then it allows considerably to reduce the total processing time because parallel processing is determined by the number of processed batches rather than by the total number of data.

An important property of batch processing is that it almost does not enlarges the minimax risk (7) if the total number of batches is large enough. Gaussian one-armed bandit was earlier considered in [9, 10], where the variance D was assumed to be a priori known. This assumption is due to the fact that the minimax risk changes little with a significant changes of the variance up to 5–10 %. If sizes of batches are large enough, this means that the variance can be estimated during processing the initial batch and then the obtained estimate be used for the control. However, for moderate sizes of batches it is reasonable to assume the prior uncertainty of the variance as it is done in the present article.

Note that the idea of batch processing was initially proposed for medical treatments of patients by alternative drugs. According to this approach two alternative treatments are used initially for two sufficiently large test groups of patients and then more efficient treatment (providing the maximum recovered patients in the initial groups) is used for the remaining patients (see, e.g., [11, 12]).

The rest of the article is organized as follows. In Section 2, we present standard recursive equations for computing a regret and Bayesian risk. In Section 3, these equations are presented in another, more convenient for calculations form.

In Section 4, we prove the thresholding property of Bayesian strategy, which was first established in [7] for Bernoulli one-armed bandit. According to this property, once the first action is chosen for application, it will be used until the end of the control. This follows from the fact that the usage of the first action does not give any more information. Therefore, additional applications of the second action are not required. This property allows to simplify recursive equation for computing Bayesian risk and Bayesian strategy.

In Section 5, we compare recursive equation for computing Bayesian risk with its previously obtained version with a priory known variance D and show that both versions are approximately the same if n is large enough.

Section 6 contains a conclusion.

2. Recursive Equations for Computing the Regret and Bayesian Risk

Let $\chi^2_n(x)$ denote chi-squared distribution density with n degrees of freedom:

$$\chi^2_n(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$$

and introduce the distribution density

$$\psi_n(S|D) = D^{-1}\chi^2_{n-1}(D^{-1}S). \tag{9}$$

Note that defined in (2) statistics S has exactly the distribution density $\psi_{n_2}(S|D)$. Given a prior distribution density $\lambda(m, D)$, the posterior distribution

density is

$$\lambda(m, D|X, S, n_2) = \frac{f_{n_2 D}(X|n_2 m)\psi_{n_2}(S|D)\lambda(m, D)}{P(X, S, n_2)}, \tag{10}$$

where

$$P(X, S, n_2) = \iint_{\Theta} f_{n_2 D}(X|n_2 m)\psi_{n_2}(S|D)\lambda(m, D)dmdD. \tag{11}$$

The cases $n_2 = 0$ and $n_2 = 1$ need a special consideration. If $n_2 = 0$ then we put $X = S = 0$ and $f_{n_2 D}(X|n_2 m) = 1$, $\psi_{n_2}(S|D) = 1$. If $n_2 = 1$ then we put $S = 0$ and $\psi_{n_2}(S|D) = 1$. Therefore, (10)–(11) remain true if $n_2 = 0$ and $n_2 = 1$, too.

Now let's consider how to update X, S. Let

$$X = \sum_{i=1}^{n_2} x_i, \quad S = \sum_{i=1}^{n_2} x_i^2 - X^2/n_2$$

be current values X, S, where all $\{x_i\}$ are incomes obtained in response to the application of the second action, $n_2 \geq 1$, and let $x_{n_2+1} = Y$ be a new such income. Then

$$X_{new} = \sum_{i=1}^{n_2+1} x_i = X + Y, \quad S_{new} = \left(\sum_{i=1}^{n_2+1} x_i^2\right) - X_{new}^2/(n_2 + 1)$$
$$= \left(\sum_{i=1}^{n_2} x_i^2\right) + Y^2 - (X + Y)^2/(n_2 + 1) = S + \Delta(X, n_2, Y),$$

where

$$\Delta(X, n_2, Y) = Y^2 + X^2/n_2 - (X + Y)^2/(n_2 + 1) = \frac{(X - n_2 Y)^2}{n_2(n_2 + 1)}.$$

If $n_2 = 0$ then $\Delta(0, 0, Y) = Y^2 - Y^2 = 0$. Hence,

$$\Delta(X, n_2, Y) = \begin{cases} 0, & if\ n_2 = 0, \\ \dfrac{(X - n_2 Y)^2}{n_2(n_2 + 1)}, & if\ n_2 \geq 1. \end{cases} \tag{12}$$

Therefore, X, S are updated according to the rule

$$X \leftarrow X + Y, \quad S \leftarrow S + \Delta(X, n_2, Y), \tag{13}$$

where $\Delta(X, n_2, Y)$ is given by (12).

Denote $n = n_1 + n_2$, $\Theta^+ = \Theta \cap \{m \geq 0\}$, $\Theta^- = \Theta \cap \{m \leq 0\}$. Let $L(\sigma; n_1, X, S, n_2)$ denote a regret (5) on the remaining control horizon $n + 1, \ldots, N$ computed with respect to the posterior distribution $\lambda(m, D|X, S, n_2)$, i.e., $L(\sigma; n_1, X, S, n_2) = L_{N-n}(\sigma, \lambda(m, D|X, S, n_2))$. A standard recursive equation for computing a regret (5) is as follows

$$L(\sigma; n_1, X, S, n_2) = \sum_{\ell=1}^{2} \sigma_\ell(n_1, X, S, n_2) L^{(\ell)}(\sigma; n_1, X, S, n_2), \tag{14}$$

where $L^{(1)}(\sigma; n_1, X, S, n_2) = L^{(2)}(\sigma; n_1, X, S, n_2) = 0$ if $n = N$, and

$$\begin{aligned} L^{(1)}(\sigma; n_1, X, S, n_2) &= \iint_{\Theta^+} \lambda(m, D|X, S, n_2) \\ &\quad \times (m + L(\sigma; n_1 + 1, X, S, n_2))\, dmdD \\ &= \iint_{\Theta^+} m\lambda(m, D|X, S, n_2) dmdD + L(\sigma; n_1 + 1, X, S, n_2), \end{aligned} \tag{15}$$

$$\begin{aligned} L^{(2)}(\sigma; n_1, X, S, n_2) &= \iint_{\Theta^-} \lambda(m, D|X, S, n_2) \\ &\times \left(|m| + \int_{-\infty}^{\infty} L(\sigma; n_1, X + Y, S + \Delta(X, n_2, Y), n_2 + 1) f_D(Y|m) dY \right) dmdD, \end{aligned}$$

if $0 \le n \le N - 1$, where $\Delta(X, n_2, Y)$ is given by (12). One can see that regret $L^{(\ell)}(\sigma; n_1, X, S, n_2)$ is equal to the loss of cumulative expected income at the remaining control horizon $n + 1, \dots, N$ if at first the ℓ-th action was chosen and then the control was implemented according to the strategy σ. The averaged regret (5) is

$$L_N(\sigma, \lambda) = L(\sigma; 0, 0, 0, 0). \tag{16}$$

Equation (14)–(15) provides the algorithm for how to minimize the regret (5), i.e., how to find the Bayesian risk (6) and appropriate Bayesian strategy. At the point of time $n+1$ one should choose the action corresponding to the smaller value of $L^{(1)}(\sigma; n_1, X, S, n_2)$, $L^{(2)}(\sigma; n_1, X, S, n_2)$, in the case of a draw the choice can be arbitrary. This algorithm provides a standard equation for finding Bayesian risk (6):

$$R(n_1, X, S, n_2) = \min_{\ell=1,2} R^{(\ell)}(n_1, X, S, n_2), \tag{17}$$

where $R^{(1)}(n_1, X, S, n_2) = R^{(2)}(n_1, X, S, n_2) = 0$ if $n = N$, and

$$\begin{aligned} R^{(1)}(n_1, X, S, n_2) &= \iint_{\Theta^+} \lambda(m, D|X, S, n_2) \\ &\quad \times (m + R(n_1 + 1, X, S, n_2))\, dmdD \\ &= \iint_{\Theta^+} m\lambda(m, D|X, S, n_2) dmdD + R(n_1 + 1, X, S, n_2), \end{aligned} \tag{18}$$

$$\begin{aligned} R^{(2)}(n_1, X, S, n_2) &= \iint_{\Theta^-} \lambda(m, D|X, S, n_2) \\ &\times \left(|m| + \int_{-\infty}^{\infty} R(n_1, X + Y, S + \Delta(X, n_2, Y), n_2 + 1) f_D(Y|m) dY \right) dmdD, \end{aligned}$$

if $0 \le n \le N - 1$, where $\Delta(X, n_2, Y)$ is given by (12). One can see that $R^{(\ell)}(n_1, X, S, n_2)$ is equal to the loss of cumulative expected income at the remaining control horizon $n+1, \dots, N$ if at first the ℓ-th action was chosen and then the control was optimally implemented. Bayesian risk (6) is

$$R_N(\lambda) = R(0, 0, 0, 0). \tag{19}$$

3. Another Form of Recursive Equations

In this section, we present another forms of recursive equations which are more convenient for computations. Let's put

$$\begin{aligned}\tilde{L}^{(1)}(\sigma;n_1,X,S,n_2)&=L^{(1)}(\sigma;n_1,X,S,n_2)\times P(X,S,n_2),\\ \tilde{L}^{(2)}(\sigma;n_1,X,S,n_2)&=L^{(2)}(\sigma;n_1,X,S,n_2)\times P(X,S,n_2),\\ \tilde{L}(\sigma;n_1,X,S,n_2)&=L(\sigma;n_1,X,S,n_2)\times P(X,S,n_2),\end{aligned}\tag{20}$$

where $P(X,S,n_2)$ is defined in (11).

Theorem 1. *In order to determine the regret* (5) *one should solve the following recursive equation*

$$\tilde{L}(\sigma;n_1,X,S,n_2)=\sum_{\ell=1}^{2}\sigma_\ell(n_1,X,S,n_2)\tilde{L}^{(\ell)}(\sigma;n_1,X,S,n_2),\tag{21}$$

where $\tilde{L}^{(1)}(\sigma;n_1,X,S,n_2)=\tilde{L}^{(2)}(\sigma;n_1,X,S,n_2)=0$ *if* $n=N$, *and*

$$\begin{aligned}\tilde{L}^{(1)}(\sigma;n_1,X,S,n_2)&=g^{(1)}(X,S,n_2)+\tilde{L}(\sigma;n_1+1,X,S,n_2),\\ \tilde{L}^{(2)}(\sigma;n_1,X,S,n_2)&=g^{(2)}(X,S,n_2)\\ &+\int_{-\infty}^{\infty}\tilde{L}(\sigma;n_1,X+Y,S+\Delta(X,n_2,Y),n_2+1)h(X,S,n_2,Y)dY,\end{aligned}\tag{22}$$

if $0\le n\le N-1$, *where* $\Delta(X,n_2,Y)$ *is given by* (12). *Here*

$$\begin{aligned}g^{(1)}(X,S,n_2)&=\iint_{\Theta^+}mf_{n_2D}(X|n_2m)\psi_{n_2}(S|D)\lambda(m,D)dmdD,\\ g^{(2)}(X,S,n_2)&=\iint_{\Theta^-}|m|f_{n_2D}(X|n_2m)\psi_{n_2}(S|D)\lambda(m,D)dmdD,\end{aligned}\tag{23}$$

and

$$h(X,S,n,Y)\tag{24}$$

$$=\begin{cases}1, & if\ n=0,\\ (2\Delta(X,n,Y))^{1/2}, & if\ n=1,\\ \left(\dfrac{n+1}{\pi n}\right)^{1/2}\dfrac{\Gamma(n/2)S^{(n-1)/2-1}}{\Gamma((n-1)/2)(S+\Delta(X,n,Y))^{n/2-1}}, & if\ n\ge 2.\end{cases}$$

The regret (5) *is*

$$L_N(\sigma,\lambda)=\tilde{L}(\sigma;0,0,0,0).\tag{25}$$

Proof. Let's multiply (14) by $P(X,S,n_2)$ defined in (11). We obtain (21)–(22) with $g^{(1)}(X,S,n_2)$, $g^{(2)}(X,S,n_2)$ defined in (23). Let Δ in (26)–(28) below be given by (12). For $n\ge 2$ the function $h(X,S,n,Y)$ is

$$\begin{aligned}h(X,S,n,Y)&=\frac{\iint_{\Theta}f_{nD}(X|nm)\psi_n(S|D)f_D(Y|m)\lambda(m,D)dmdD}{P(X+Y,S+\Delta,n+1)}\\ &=\frac{f_{nD}(X|nm)\psi_n(S|D)f_D(Y|m)}{f_{(n+1)D}(X+Y|(n+1)m)\psi_{n+1}(S+\Delta|D)}\\ &=\frac{f_{nD}(X|nm)f_D(Y|m)}{f_{(n+1)D}(X+Y|(n+1)m)}\times\frac{\psi_n(S|D)}{\psi_{n+1}(S+\Delta|D)}.\end{aligned}\tag{26}$$

Since

$$\frac{f_{nD}(X|nm)f_D(Y|m)}{f_{(n+1)D}(X+Y|(n+1)m)} = \left(\frac{n+1}{2\pi nD}\right)^{1/2} \times \exp\left(-\frac{\Delta}{2D}\right), \tag{27}$$

and

$$\begin{aligned}\frac{\psi_n(S|D)}{\psi_{n+1}(S+\Delta|D)} &= \frac{2^{1/2}\Gamma(n/2)}{\Gamma((n-1)/2)} \times \frac{(S/D)^{(n-1)/2-1}}{((S+\Delta)/D)^{n/2-1}} \\ &\times \frac{\exp\left(-S/(2D)\right)}{\exp\left(-(S+\Delta)/(2D)\right)} \\ &= \frac{(2D)^{1/2}\Gamma(n/2)}{\Gamma((n-1)/2)} \times \frac{S^{(n-1)/2-1}}{(S+\Delta)^{n/2-1}} \times \exp\left(\frac{\Delta}{2D}\right),\end{aligned} \tag{28}$$

it follows from (26)–(28) that $h(X,S,n,Y)$ is given by (24) if $n \geq 2$. If $n=1$ then $\psi_n(S|D)=1$ and according to (26)

$$\begin{aligned}h(X,S,n,Y) &= \frac{f_{nD}(X|nm)\psi_n(S|D)f_D(Y|m)}{f_{(n+1)D}(X+Y|(n+1)m)\psi_{n+1}(S+\Delta|D)} \\ &= \left(\frac{n+1}{2\pi nD}\right)^{1/2} \times \exp\left(-\frac{\Delta}{2D}\right) \times \frac{1}{\psi_{n+1}(\Delta|D)} \\ &= \left(\frac{n+1}{2\pi nD}\right)^{1/2} \times \exp\left(-\frac{\Delta}{2D}\right) \times \frac{2^{n/2}\Gamma(n/2)D}{(\Delta/D)^{n/2-1}\exp(-\Delta/(2D))} \\ &= \left(\frac{2\Delta}{\pi}\right)^{1/2} \Gamma\left(\frac{1}{2}\right) = (2\Delta)^{1/2}.\end{aligned}$$

If $n=0$ then $f_{nD}(X|nm)=1$, $\psi_n(S|D)=1$, $\psi_{n+1}(S+\Delta|D)=1$ and according to (26)

$$h(X,S,n,Y) = \frac{f_{nD}(X|nm)\psi_n(S|D)f_D(Y|m)}{f_{(n+1)D}(X+Y|(n+1)m)\psi_{n+1}(S+\Delta|D)} = 1.$$

Formula (25) follows from (16) and the fact that $P(X,S,n_2)=1$ if $n_2=0$. This finishes the proof of theorem 1. □

A similar theorem holds for a Bayesian risk. Let's put

$$\begin{aligned}\tilde{R}^{(1)}(n_1,X,S,n_2) &= R^{(1)}(n_1,X,S,n_2) \times P(X,S,n_2), \\ \tilde{R}^{(2)}(n_1,X,S,n_2) &= R^{(2)}(n_1,X,S,n_2) \times P(X,S,n_2), \\ \tilde{R}(n_1,X,S,n_2) &= R(n_1,X,S,n_2) \times P(X,S,n_2).\end{aligned} \tag{29}$$

Theorem 2. *In order to determine the Bayesian risk* (6) *one should solve the following recursive equation*

$$\tilde{R}(n_1,X,S,n_2) = \min\left(\tilde{R}^{(1)}(n_1,X,S,n_2), \tilde{R}^{(2)}(n_1,X,S,n_2)\right), \tag{30}$$

where $\tilde{R}^{(1)}(n_1,X,S,n_2) = \tilde{R}^{(2)}(n_1,X,S,n_2) = 0$ *if* $n=N$ *and*

$$\begin{aligned}\tilde{R}^{(1)}(n_1,X,S,n_2) &= g^{(1)}(X,S,n_2) + \tilde{R}(n_1+1,X,S,n_2), \\ \tilde{R}^{(2)}(n_1,X,S,n_2) &= g^{(2)}(X,S,n_2) \\ &+ \int_{-\infty}^{\infty} \tilde{R}(n_1,X+Y,S+\Delta(X,n_2,Y),n_2+1)h(X,S,n_2,Y)dY,\end{aligned} \tag{31}$$

if $0 \le n \le N-1$. *Here* $\Delta(X, n_2, Y)$ *is given by* (12), $g^{(1)}(X, S, n_2)$, $g^{(2)}(X, S, n_2)$ *are given by* (23) *and* $h(S, X, Y, n)$ *is given by* (24). *At the point of time* $n+1$ *Bayesian strategy prescribes to choose the action corresponding to the smaller value of* $\tilde{R}^{(1)}(n_1, X, S, n_2)$, $\tilde{R}^{(2)}(n_1, X, S, n_2)$; *in the case of a draw the choice can be arbitrary. Bayesian risk* (6) *is*

$$R_N(\lambda) = \tilde{R}(0, 0, 0, 0). \tag{32}$$

Proof of theorem 2 is similar to the proof of theorem 1.

4. Thresholding Property of the Strategy

In this section, we prove the following property of Bayesian strategy, which was first established in [7]. Since applying the first action does not give any additional information, it means that once being chosen, the first action will be used until the end of the control. Namely, the following lemma holds.

Lemma 1. *Let*

$$\tilde{R}^{(1)}(n_1, X, S, n_2) < \tilde{R}^{(2)}(n_1, X, S, n_2), \quad n_1 + n_2 \le N-1. \tag{33}$$

Then $\tilde{R}^{(1)}(n_1+1, X, S, n_2) < \tilde{R}^{(2)}(n_1+1, X, S, n_2)$. *This means, that once the first action is chosen, it will be used until the end of the control.*

Proof. Proof is done by induction. Obviously, theorem holds if $n = N-1$. Suppose that it holds if $n = M < N-1$ and check that it holds if $n = M-1$. It follows from (31) that

$$\begin{gathered}\tilde{R}^{(1)}(n_1, X, S, n_2) = g^{(1)}(X, S, n_2) \\ + \min(\tilde{R}^{(1)}(n_1+1, X, S, n_2), \tilde{R}^{(2)}(n_1+1, X, S, n_2)), \\ \tilde{R}^{(2)}(n_1, X, S, n_2) \le \tilde{R}^{(2)}(n_1+1, X, S, n_2) + \mathbf{E}\tilde{R}(n_1', X', S', n_2'),\end{gathered} \tag{34}$$

where $n_1' + n_2' = N-1$ and $n_2' \ge n_2$. In inequality the expression on the right-hand side is equal to expected loss if at first the Bayesian strategy is used on the control horizon of the length $N-M-1$ and then the optimal strategy on the latter step at the time point $N-1$ is applied. Since $g^{(1)}(X, S, n_2)$ is one-step expected income for the use of the first action and $n_2' \ge n_2$ then $\mathbf{E}\tilde{R}(n_1', X', S', n_2') \le \min(g^{(1)}(X, S, n_2), g^{(2)}(X, S, n_2)) \le g^{(1)}(X, S, n_2)$ and, hence,

$$\tilde{R}^{(2)}(n_1, X, S, n_2) \le g^{(1)}(X, S, n_2) + \tilde{R}^{(2)}(n_1+1, X, S, n_2). \tag{35}$$

Suppose that the assertion of lemma does not hold. Then it follows from the first equality (34) that

$$\tilde{R}^{(1)}(n_1, X, S, n_2) = g^{(1)}(X, S, n_2) + \tilde{R}^{(2)}(n_1+1, X, S, n_2), \tag{36}$$

One can see that (36) and (35) contradict (33). This proves the lemma. □

Lemma 1 allows us to present recursive equation for computing Bayesian strategy and Bayesian risk in a more simple form.

Corollary 1. *In order to determine the Bayesian risk* (6) *one should solve the following recursive equation*

$$\tilde{R}(0, X, S, n_2) = \min\left(\tilde{R}^{(1)}(0, X, S, n_2), \tilde{R}^{(2)}(0, X, S, n_2)\right), \tag{37}$$

where $\tilde{R}^{(1)}(0,X,S,n_2) = \tilde{R}^{(2)}(0,X,S,n_2) = 0$ *if* $n_2 = N$ *and*

$$\tilde{R}^{(1)}(0,X,S,n_2) = (N-n_2)g^{(1)}(X,S,n_2),$$

$$\tilde{R}^{(2)}(0,X,S,n_2) = g^{(2)}(X,S,n_2) \tag{38}$$

$$+\int_{-\infty}^{\infty} \tilde{R}(0,X+Y,S+\Delta(X,n_2,Y),n_2+1)h(X,S,n_2,Y)dY,$$

if $0 \le n_2 \le N-1$. *Here* $\Delta(X,n_2,Y)$ *is given by* (12), $g^{(1)}(X,S,n_2)$, $g^{(2)}(X,S,n_2)$ *are given by* (23) *and* $h(X,S,n_2,Y)$ *is given by* (24). *At the point of time* n_2+1 *Bayesian strategy prescribes to choose the action corresponding to the smaller value of* $\tilde{R}^{(1)}(0,X,S,n_2)$, $\tilde{R}^{(2)}(0,X,S,n_2)$, *in the case of a draw the choice can be arbitrary. Once the first action has been chosen, it will be used until the end of the control. Equation* (37)–(38) *follows from* (30)–(31) *and lemma 1. Bayesian risk* (6) *is given by* (32).

5. Comparison with the Case of a Known Variance

In this section, we present recursive equation for computing Bayesian risk and Bayesian strategy if the variance D is known. The set of parameters is thus $\Theta = \{m : |m| \le C_1 < \infty\}$. This equation can be used if sizes of processed batches are large enough. In this case the unknown variance can be estimated while processing initial batch of data and then obtained estimate can be used for the control. According to [13], theorem 2.1, one should solve a recursive equation

$$\tilde{R}(0,X,n_2) = \min(\tilde{R}^{(1)}(0,X,n_2), \tilde{R}^{(2)}(0,X,n_2)), \tag{39}$$

where $\tilde{R}^{(1)}(0,X,n_2) = \tilde{R}^{(2)}(0,X,n_2) = 0$ if $n_2 = N$ and

$$\tilde{R}^{(1)}(0,X,n_2) = (N-n_2)g_D^{(1)}(X,n_2),$$

$$\tilde{R}^{(2)}(0,X,n_2) = g_D^{(2)}(X,n_2) \tag{40}$$

$$+\int_{-\infty}^{+\infty} \tilde{R}(0,X+Y,n_2+1)h_D(X,n_2,Y)\,dY,$$

if $0 \le n_2 \le N-1$. Here

$$g_D^{(1)}(X,n_2) = \int_0^{C_1} m f_{Dn_2}(X-n_2m)\lambda(m)dm,$$

$$g_D^{(2)}(X,n_2) = \int_{-C_1}^{0} |m| f_{Dn_2}(X-n_2m)\lambda(m)dm,$$

and

$$h_D(X,n,Y) = \begin{cases} 1, & if\ n=0, \\ \left(\dfrac{n+1}{2\pi Dn}\right)^{1/2} \exp\left(-\dfrac{\Delta(X,n,Y)}{2D}\right), & if\ n>0, \end{cases} \tag{41}$$

where $\Delta(X,n,Y)$ is given by (12). At the point of time n_2+1 Bayesian strategy prescribes to choose the action corresponding to the smaller value of $\tilde{R}^{(1)}(0,X,n_2)$, $\tilde{R}^{(2)}(0,X,n_2)$, in the case of a draw the choice can be arbitrary. Once the first action has been chosen, it will be used until the end of the control. Bayesian risk (6) is $\tilde{R}(0,0,0)$.

Let's consider approximations of $g^{(\ell)}(X,S,n_2)$ and $h(X,S,n_2,Y)$ given by (23)–(24) for large n_2.

It is sufficient to consider $g^{(\ell)}(X,S,n_2)$ for $\ell=1$. For the sake of simplicity let's assume that $\lambda(m,D)=\lambda_1(m)\lambda_2(D)$. Let's put $S=(n_2-1)\hat{D}$, where $\hat{D}$ is the estimate of D. Since $\mathbf{E}(S)=(n_2-1)D$, $\mathbf{D}S=2(n_2-2)D^2$ then $\psi_{n_2}(S|D)$ is almost completely concentrated in the interval

$$\left((n_2-1)D-3(2n_2-4)^{1/2}D,(n_2-1)D+3(2n_2-4)^{1/2}D\right),$$

and, hence,

$$\begin{aligned} g^{(1)}(X,S,n_2) &= \iint_{\Theta^+} mf_{n_2D}(X|n_2m)\psi_{n_2}(S|D)\lambda_1(m)\lambda_2(D)dmdD \\ &\sim \int_0^{C_1} mf_{n_2\hat{D}}(X|n_2m)\lambda_1(m)dm, \end{aligned}$$

and this means that $g^{(1)}(X,S,n_2)$ approximates $g_D^{(1)}(X,n_2)$ for large n_2.

Consider $h(X,S,n,Y)$. Using the approximation $\Gamma(n)\sim(2\pi/n)^{1/2}(n/e)^n$ we obtain

$$\frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\sim\left(\frac{n-1}{n}\right)^{1/2}\left(\frac{n-1}{2e}\right)^{1/2}\left(\frac{n}{n-1}\right)^{n/2}\sim\frac{n-1}{(2n)^{1/2}}. \tag{42}$$

Denote $\hat{D}=(n-1)^{-1}S$. Then

$$\begin{aligned} \frac{S^{(n-1)/2-1}}{(S+\Delta(X,n,Y))^{n/2-1}} &= \frac{(S+\Delta(X,n,Y))^{1/2}}{S}\left(1+\frac{\Delta(X,n,Y)}{S}\right)^{-(n-1)/2} \\ &\sim \frac{1}{(n-1)^{1/2}\hat{D}^{1/2}}\exp\left(-\frac{\Delta(X,n,Y))}{2\hat{D}}\right) \end{aligned} \tag{43}$$

From (24), (42), (43) we obtain

$$h(X,S,n,Y)\sim\left(\frac{n-1}{n}\right)^{1/2}h_{\hat{D}}(X,n,Y)\sim h_{\hat{D}}(X,n,Y),$$

and this means that $h(X,S,n_2,Y)$ approximates $h_{\hat{D}}(X,n_2,Y)$ for large n_2.

6. Conclusion

We considered Gaussian one-armed bandit with both unknown mathematical expectation and the variance. We obtained recursive Bellman-type equation for computing a regret, Bayesian risk and Bayesian strategy. We show that this equation is equvivalent to that in the case of known variance if cumulative number of applications of the action with unknown parameters is large enough. As a direction of future research the one-armed bandit with updating the estimate of the variance within each processed batch of data can be considered.

References

[1] D.A. Berry, B. Fristedt, *Bandit problems. Sequential allocation of experiments*, Chapman and Hall, London - New York, 1985. Zbl 0659.62086

[2] E.L. Presman, I.N. Sonin, *Sequential control with incomplete information. The Bayesian approach to multi-armed bandit problems*, Academic Press, Inc., London etc., 1990. Zbl 0721.93003

[3] T. Lattimore, C. Szepesvári, *Bandit algorithms*, Cambridge University Press, Cambridge, 2020. Zbl 1439.68002
[4] M.L Tsetlin, *Automaton theory and modelling of biological systems*, Academic Press, London etc., 1973. Zbl 0297.68050
[5] V.G. Sragovich, *Mathematical theory of adaptive control*, World Sci., Hackensack, 2006. Zbl 1204.93005
[6] J.C. Gittins, *Multi-armed bandit allocation indices*, Wiley-Interscience Series in Systems and Optimization, John Wiley & Sons, Chichester, 1989. Zbl 0699.90068
[7] R.N. Bradt, S.M. Johnson, S. Karlin, *On sequential designs for maximizing the sum of n observations*, Ann. Math. Stat., **27** (1956), 1060–1074. Zbl 0073.14203
[8] H. Chernoff, S.N. Ray, *A Bayes sequential sampling inspection plan*, Ann. Math. Stat., **36** (1965), 1387–1407. Zbl 0203.21601
[9] A.V. Kolnogorov, *One-armed bandit problem for parallel data processing systems*, Probl. Inf. Transm., **51**:2 (2015), 177–191. Zbl 1333.62033
[10] A. Kolnogorov, *Gaussian one-armed bandit problem*, In: 2021 XVII International symposium *Problems of redundancy in information and control systems* (REDUNDANCY), (2021), 74–79,
[11] T.L. Lai, B. Levin, H. Robbins, D. Siegmund, *Sequential medical trials*, Proc. Natl. Acad. Sci. USA, **77**:6 (1980), 3135–3138. Zbl 0453.62066
[12] V. Perchet, P. Rigollet, S. Chassang, E. Snowberg, *Batched bandit problems*, Ann. Stat., **44**:2 (2016), 660–681. Zbl 1338.62180
[13] A.V. Kolnogorov, *Gaussian one-armed bandit and optimization of batch data processing*, 2019, arXiv:1901.08845.

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