MINIMAX R-STAGE STRATEGY FOR THE MULTI-ARMED BANDIT PROBLEM

A.V.Kolnogorov [*,1] **S.V.Melnikova** [*]

[*] *Novgorod State University, B.St-Petersburgskaya str. 41, Velikiy Novgorod, 173003, Russia, Fax: +7 (8162) 62-41-10, e-mail:kav@novsu.ac.ru; kontold@yandex.ru*

Abstract: The r-stage multi-armed bandit problem is considered in minimax setting on the finite sufficiently large time interval T. A sequential control procedure with a priori specified magnitudes of learning stages and thresholds is offered. The value of the minimax risk close to T^{α} with $\alpha = 2^{r-1}/(2^r - 1)$ is obtained. The applications to information transmission and medical treatments are discussed. Considered approach is especially valuable for systems with parallel processing in which the number of stages r mainly influences the total duration of the process.

Keywords: multi-armed bandit, adaptive learning control, adaptive choice of alternatives, sequential decisions, minimax control, parallel processing

1. INTRODUCTION

Let's consider K alternative methods of information transmission numbered by $\ell = 1, \ldots, K$ ($K \geq 2$). It is supposed that information is divided into T small packages, where T is a large number. Each method can be used to transmit information. The usage of the ℓ-th method is either successful (the package is transmitted without error) with a probability p_ℓ or unsuccessful (the package is transmitted with an error) with probability $q_\ell = 1 - p_\ell$, probabilities p_ℓ, $\ell = 1, \ldots, K$ are supposed to be unknown. Therefore, information transmission can be described by a stochastic process ξ_t with $t = 1, \ldots, T$, where $\xi_t = 1$ if a package number t is transmitted correctly and $\xi_t = 0$ otherwise. As total number of packages T is sufficiently large, in the article asymptotic estimates are considered.

The goal is to maximize the total expected value of the sum $\sum_{t=1}^{T} \xi_t$ with the help of some decision procedure based on the current knowledge of the process. This problem is known as the multi-armed bandit problem (Berry and Fristedt, 1985). There are several approaches to the problem depending on its possible applications. For example, optimal strategies can be calculated exactly in Bayes setting for some known prior distributions (Berry and Fristedt, 1985). If a prior distribution is unknown the asymptotically optimal procedure can be used, which makes estimates of probabilities $p_1, \ldots, p_K$ and then preferably applies the method corresponding to the current larger value of these estimates (Sragovich, 1981). To describe the expedient behavior of the simplest organisms in random media, models based on finite automata are used (Tsetlin, 1973).

As for considered approach, the core of the problem is that changes of methods of information transmission cannot be made too often, because such a system would be very slow. Therefore, packages should be divided into a number of sufficiently large groups, so that to use the same

[1] Corresponding author

method to the packages of the same group. Note, that the important property of the approach is that it allows to process packages of the same group simultaneously.

The simplest example can be described as follows. At the beginning of the information transmission each method is applied to τ packages. Then the results of transmission are compared and the most efficient method (with the less number of errors during transmission) is applied to the rest $T - K\tau$ packages. This two-stage bandit problem was investigated in (Kolnogorov, 1999), (Kolnogorov, 2000), (Kolnogorov and Melnikova, 2005) in the minimax setting and the optimal order of τ was proved to be $T^{2/3}$.

In the present paper such an approach is generalized to the r-stage decision design. The procedure compares alternatives on learning stages and applies the best one on the final stage. More precisely, denote by $t_0 = 0$ initial total number of choices and by $\mathcal{I}_1 = \{1, \dots, K\}$ initial set of indices containing all alternatives. The procedure starts with the 1-st learning stage and then follows the algorithm:

Learning Stages: On the i-th learning stage each of $k(i)$ tested alternatives from the ordered set $\mathcal{I}_i = \{\ell_{i1}, \dots, \ell_{ik(i)}\}$, where $1 \le \ell_{i1} < \dots < \ell_{ik(i)} \le K$, $2 \le k(i) \le K$, is consequently applied τ_i times, $1 \le i \le r-1$. Then total number of choices and total numbers of successes on the i-th stage corresponding to tested alternatives are calculated as

$$t_i = t_{i-1} + k(i)\tau_i,$$

$$X_i(\ell_{ij}) = \sum_{t=t_{i-1}+(j-1)\tau_i+1}^{t_{i-1}+j\tau_i} \xi_t,$$

$$j = 1, \dots, k(i).$$

The maximal current total number of successes is determined

$$X_i(\ell_{ij_0}) = \max_{1 \le j \le k(i)} X_i(\ell_{ij}).$$

If there are more than one maximal elements, the choice can be arbitrary made. Then $X_i(\ell_{ij_0})$ is compared with all the rest $\{X_i(\ell_{ij})\}$ and the set of indices

$$\mathcal{J}_i = \{\ell_{ij} : X_i(\ell_{ij}) \le X_i(\ell_{ij_0}) - \Delta_i\}$$

is determined. The alternatives of the set $\mathcal{J}_i$ are removed from the set $\mathcal{I}_i$ at the end of the stage, so as $\mathcal{I}_{i+1} = \mathcal{I}_i \setminus \mathcal{J}_i$.

If at the end of the i-th stage the set $\mathcal{I}_{i+1}$ contains the only alternative ℓ_{ij_0} then we apply $\mathcal{I}_j = \mathcal{I}_{i+1}$, $j = i+2, \dots, r-1$, and the strategy prescribes immediately to go to the **Final Stage**, otherwise the following $(i+1)$-th learning stage is implemented. We suppose that $\Delta_{r-1} = 0$, so the transition to the **Final Stage** at the end of the $(r-1)$-th stage is obligatory.

Final Stage: On this stage only the alternative corresponding to the larger value $X_i(\ell_{ij_0})$ on the previous i-th stage is implemented. This alternative is considered to be the better one according to learning stages.

The total number T of packages is assumed to be sufficiently large. In Bayes setting this problem (especially for $r = 2$ and $r = 3$) was considered by several authors (see, e.g., (Witmer, 1986), (Cheng, 1994)). Note, that orders of magnitudes of learning stages in these papers differ from represented below and factors depend on the prior distribution.

The structure of the paper is the following. In section 2 optimal asymptotic expressions for learning stages $\{\tau_i\}$ and thresholds $\{\Delta_i\}$ are found in general case, 3-stage and 4-stage procedures are considered as examples. In section 3 schemes of proves of theorems are given. Section 4 contains a discussion of results.

2. ASYMPTOTIC ESTIMATES OF OPTIMAL MAGNITUDES OF LEARNING STAGES AND THRESHOLDS

Let $\theta = (p_1, p_2, \dots, p_K)$ denote a parameter describing the multi-armed bandit and $w_\ell(\theta) = p_\ell$ denote the mathematical expectation if the ℓ-th alternative is chosen. We consider the set of parameters

$$\Theta = \{\theta : |p_l - p_k| \le \Delta W \le 0.5;\ l, k = 1, \dots, K\},$$

excluding the case $\Delta W = 0$ where all strategies are optimal.

If we know θ, the optimal strategy is to choose the alternative corresponding to $\max_{\ell=1,\dots,K} w_\ell(\theta)$. Then the expected total income is equal to $T \max_{\ell=1,\dots,K} w_\ell(\theta)$. If θ is unknown, the strategy prescribes to use sequential learning procedure defined above which can be described by the sequence $\pi = \{\tau_1, \dots, \tau_{r-1}; \Delta_1, \dots, \Delta_{r-1}\}$. If $\mathbf{E}_{\pi,\theta}$ denotes the mathematical expectation provided that the strategy π and the parameter θ are fixed, then the loss function due to the lack of information on θ is equal to

$$L_T(\pi, \theta) = T \max_{\ell=1,\dots,K} w_\ell(\theta) - \sum_{t=1}^{T} \mathbf{E}_{\pi,\theta}\ \xi_t.$$

Minimax risk on the class of strategies $\{\pi\}$ and the set of parameters Θ is equal to

$$R_T(\Theta) = \inf_{\{\pi\}} \sup_{\Theta} L_T(\pi, \theta).$$

Note, that maximal loss does not exceed $R_T(\Theta)$ if the minimax strategy π^* is chosen. On the other hand, $R_T(\Theta)$ is achieved for any strategy π and some $\theta = \theta(\pi)$. Let V denote the maximal

variance of income $V_\ell(\theta) = p_\ell q_\ell$ on the set of parameters Θ

$$V = \max_{\Theta} \max_{\ell=1,\dots,K} V_\ell(\theta) = 0.25.$$

The following theorems describe asymptotic dependence of minimax risk $R_T(\Theta)$ on T for cases $r = 2$ and $r \geq 3$.

Theorem 2.1. If $r = 2$ then $\Delta_1 = 0$ and asymptotically optimal learning stage satisfies condition

$$\Delta W \tau_1 \sim C_2 (V/\tau_1)^{1/2} T.$$

Consequently

$$\tau_1 \sim C_2^{2/3} \kappa^{1/3} T^{2/3},$$

where $C_2 \approx 0.240$, $\kappa = V(\Delta W)^{-2}$ and minimax risk $R_T(\Theta)$ satisfies the limiting equality

$$\lim_{T\to\infty} R_T(\Theta)\tau_1^{-1} = (K-1)\Delta W.$$

Theorem 2.2. If $r \geq 3$ then asymptotically optimal learning stages and thresholds satisfy conditions

$$\tau_i \sim (a_i T^{\alpha_i})^2, \quad \Delta_i \sim b_i (2V\tau_i)^{1/2},$$
$$i = 1, \dots, r-1,$$

where $\alpha_1, \dots, \alpha_{r-1}$ satisfy the system of equations

$$2\alpha_1 = 2\alpha_{i+1} - \alpha_i = 1 - \alpha_{r-1}, \qquad (1)$$
$$i = 1, \dots, r-2,$$

$b_1, \dots, b_{r-1}$ are expressed by

$$b_i = (\beta_i \ln T)^{1/2}, \quad \beta_i = 2(1 - \alpha_i - 2\alpha_1), \qquad (2)$$
$$i = 1, \dots, r-1,$$

$a_1, \dots, a_{r-1}$ satisfy asymptotic (as $T \to \infty$) relations

$$\Delta W a_1^2 \sim \frac{b_i a_{i+1}^2 (2V)^{1/2}}{a_i} \sim \frac{C_2 V^{1/2}}{a_{r-1}}, \qquad (3)$$
$$i = 1, \dots, r-2.$$

Minimax risk $R_T(\Theta)$ satisfies the limiting equality

$$\lim_{T\to\infty} \{a_1 T^{\alpha_1}\}^{-2} R_T(\Theta) = (K-1)\Delta W, \qquad (4)$$

where $\alpha_1 = 2^{r-2}(2^r - 1)^{-1}$, $a_1 \sim \gamma_1 \{\ln T\}^{\delta_1/2}$, $\delta_1 = (2^{r-2} - 1)(2^r - 1)^{-1}$.

It can be found from (2) and (3) that

$$a_i \sim \gamma_i (\ln T)^{\delta_i/2}, \ i = 1, \dots, r-1,$$

and $\delta_1, \dots, \delta_{r-1}$, $\gamma_1, \dots, \gamma_{r-1}$, satisfy systems of equations

$$\delta_1 = \frac{1 + 2\delta_{i+1} - \delta_i}{2} = -\frac{\delta_{r-1}}{2}, \qquad (5)$$

$$\frac{\gamma_1^2}{\kappa^{1/2}} = \frac{(2\beta_i)^{1/2}\gamma_{i+1}^2}{\gamma_i} = \frac{C_2}{\gamma_{r-1}}, \qquad (6)$$
$$i = 1, \dots, r-2.$$

The values $\alpha_1, \dots, \alpha_{r-1}, \beta_1, \dots, \beta_{r-1}, \delta_1, \dots, \delta_{r-1}$ can be found from (1), (2) and (5) by induction or as the solutions of appropriate difference equations and are equal to:

$$\alpha_i = \frac{2^{r-1} - 2^{r-i-1}}{2^r - 1}, \ \beta_i = 2\frac{2^{r-i-1} - 1}{2^r - 1},$$

$$\delta_i = \frac{2^{r-i-1} \cdot (3 - 2^i) - 1}{2^r - 1}, \ i = 1, \dots, r-1.$$

It's easy to see that $\alpha_i > 0$, $\beta_i > 0$ for $i = 1, \dots, r-1$; $\delta_1 > 0$ and $\delta_i < 0$ for $i = 2, \dots, r-1$. Obviously, $\gamma_i > 0$ for all $i = 1, \dots, r-1$ but their determination is a bit more difficult. From (6) the following expressions can be obtained

$$\ln \gamma_i = \frac{2^i - 1}{2^{i-1}} \ln \gamma_1 + \sum_{j=1}^{i-1} \frac{2^{i-j} - 1}{2^{i-j-1}} h_j \qquad (7)$$

for $i = 2, \dots, r$ with

$$h_1 = -\frac{\ln(2\kappa\beta_1)}{4}, \ h_{r-1} = \frac{\ln(2\beta_{r-2}) - 2\ln C_2}{4},$$

$$h_i = \frac{\ln \beta_{i-1} - \ln \beta_i}{4}, \ i = 2, \dots, r-2$$

and $\ln \gamma_r = 0$. Hence, $\ln \gamma_1$ can be determined with the help of (7) as

$$\ln \gamma_1 = -\frac{2^{r-1}}{2^r - 1} \sum_{j=1}^{r-1} \frac{2^{r-j} - 1}{2^{r-j-1}} h_j$$

and then all the rest $\ln \gamma_2, \dots, \ln \gamma_{r-1}$ can be calculated. Examples of parameters in cases $r = 3$ and $r = 4$ are given in Table 1 and Table 2.

Table 1. Parameters in case $r = 3$

i	α_i	β_i	γ_i	δ_i
1	$\frac{2}{7}$	$\frac{2}{7}$	$\left(\frac{4C_2^4\kappa^3}{7}\right)^{1/14} \approx 0.639\kappa^{3/14}$	$\frac{1}{7}$
2	$\frac{3}{7}$	0	$\left(\frac{49C_2^6\kappa}{16}\right)^{1/14} \approx 0.588\kappa^{1/14}$	$-\frac{2}{7}$

Table 2. Parameters in case $r = 4$

i	α_i	β_i	γ_i	δ_i
1	$\frac{4}{15}$	$\frac{6}{15}$	$\left(\frac{4^3C_2^8\kappa^7}{3^2 5^3}\right)^{1/30} \approx 0.621\kappa^{7/30}$	$\frac{1}{5}$
2	$\frac{6}{15}$	$\frac{2}{15}$	$\left(\frac{5C_2^4\kappa}{12}\right)^{1/10} \approx 0.518\kappa^{1/10}$	$-\frac{1}{5}$
3	$\frac{7}{15}$	0	$\left(\frac{5^6 3^4 C_2^{14}\kappa}{2^{12}}\right)^{1/30} \approx 0.622\kappa^{1/30}$	$-\frac{2}{5}$

The idea of the proof is that there exist r parameters $\theta_1, \dots, \theta_r$, depending on T, such that expected losses $L_T(\pi, \theta_i)$, $i = 1, \dots, r$ have maximal possible values on Θ and the optimal durations of learning stages and magnitudes of thresholds are obtained by balancing all $L_T(\pi, \theta_i)$, $i = 1, \dots, r$. It can be shown that $\tau_i/\tau_{i+1} \to 0$ as $T \to \infty$ for $i = 1, \dots, r-1$.

Note, that asymptotically unimprovable value of the minimax risk for all possible strategies is presented in (Vogel, 1960). Since this value is of the order $T^{1/2}$, the formula (4) gives result close to unimprovable one for moderate r.

3. SCHEMES OF PROVES

3.1 Some auxiliary estimates

Let's introduce the notations

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right),\Phi(x)=\int\limits_{-\infty}^{x}\varphi(t)dt.$$

Lemma 3.1. The limiting equalities hold

$$\lim_{x\to+\infty}\Phi(-x)\left(x^{-1}\varphi(x)\right)^{-1}=1, \tag{8}$$

$$\lim_{b\to+\infty}eb^2\varphi^{-1}(b)\left(\max_{x\geq 0}x\Phi(-b-x)\right)=1, \tag{9}$$

$$\lim_{b\to+\infty}b^{-1}\left(\max_{x\in U(b)}x\Phi(b-x)\right)=1, \tag{10}$$

$$\lim_{b\to+\infty}b^{-1}\left(\max_{x\notin U(b)}x\Phi(b-x)\right)=0, \tag{11}$$

where $U(b)=\{x: Cb^{1/2}\leq x\leq 2b, C>0\}$.

Proof. The estimate (8) is represented in the reference book (Prokhorov and Rozanov, 1987).

To prove (9) we use (8), so as $\Phi(-b-x)\sim(2\pi)^{-1/2}(b+x)^{-1}\exp\left(-b^2/2-bx-x^2/2\right)$. Then noting that $\max_{0\leq x\leq 2b^{-1}}x\exp(-bx)=(be)^{-1}$ at $x=b^{-1}$ and $\max_{x\geq 2b^{-1}}x\exp(-bx)=2e^{-1}(be)^{-1}<(be)^{-1}$ at $x=2b^{-1}$ we see that $\max_{x\geq 0}x\Phi(-b-x)=\max_{0\leq x\leq 2b^{-1}}x\Phi(-b-x)\sim(eb^2)^{-1}\varphi(b)$ as $b\to+\infty$.

To prove (10), applying $x=b(1-b^{-1/2})$ we get the lower bound estimate $\max_{x\geq 0}b^{-1}x\Phi(b-x)\geq(1-b^{-1/2})\Phi(b^{1/2})\to 1$ as $b\to+\infty$. The upper bound estimates for $x\leq b$ and $x>b$ are the following: $\max_{x\leq b}b^{-1}x\Phi(b-x)\leq 1$ and $\max_{x>b}b^{-1}x\Phi(b-x)=\max_{y>0}b^{-1}(b+y)\Phi(-y)\leq 0.5+b^{-1}\max_{y>0}y\Phi(-y)\leq 1$ for sufficiently large b. It is easy to see that (11) is correct as well. ■

For $i=1,\dots,r-2$, consider $X_i(\ell)$, $\ell=1,k$, the total numbers of successes on the i-th stage corresponding to probabilities of success $p_\ell=p-m_\ell$, where $m_\ell=x_\ell(2V/\tau_i)^{1/2}$; $\ell=1,k$; $x_k\geq x_1=0$.

Lemma 3.2. Let $p\in[d,1-d]$, $0<d<0.5$, $b_i\to+\infty$, $\tau_i\to+\infty$ as $T\to+\infty$ and $i=1,\dots,r-2$. Then the following estimates hold as $T\to\infty$

$$\sup_{x_k\in U(b_i)}\frac{m_k\mathbf{P}(X_i(k)>X_i(1)-\Delta_i)}{b_i(2V/\tau_i)^{1/2}}\sim 1, \tag{12}$$

$$\sup_{x_k\notin U(b_i)}\frac{m_k\mathbf{P}(X_i(k)>X_i(1)-\Delta_i)}{b_i(2V/\tau_i)^{1/2}}\to 0, \tag{13}$$

$$\sup_{x_k\geq 0}\frac{m_k\mathbf{P}(X_i(1)\leq X_i(k)-\Delta_i)}{(2V/\tau_i)^{1/2}b_i^{-2}e^{-1}\varphi(b_i)}\lesssim 1. \tag{14}$$

Proof. Considered magnitudes of the difference $X_i(k)-X_i(1)$, including large deviations, obey the central limit theorem (see reference book (Prokhorov and Rozanov, 1987)). Denote $V_\ell=p_\ell(1-p_\ell)$, $V(p)=p(1-p)$. Then

$$\mathbf{E}(X_i(k)-X_i(1))=-x_k(2V\tau_i)^{1/2},$$

$$\mathbf{Var}(X_i(k)-X_i(1))=(V_1+V_k)\tau_i\sim 2V(p)\tau_i$$

as $\tau_i\to\infty$. To prove (12) we write

$$\mathbf{P}(X_i(k)>X_i(1)-\Delta_i))\sim\Phi((b_i-x_k)(V/V(p))^{1/2})$$

and then

$$\sup_{x_k\in U(b_i)}m_k\mathbf{P}(X_i(k)>X_i(1)-\Delta_i)\sim(2V/\tau_i)^{1/2}\times$$

$$\times\sup_{x_k\in U(b_i)}x_k\Phi((b_i-x_k)(V/V(p))^{1/2})\sim b_i(2V/\tau_i)^{1/2}$$

according to the estimate (10). The estimates (13) and (14) can be proved in the similar way using the estimates (11) and (9). ■

For $X_{r-1}(\ell)$, $\ell=1,\dots,k$ consider the probabilities of success $p_\ell=p-m_\ell$; $m_\ell=z_\ell(V/\tau_{r-1})^{1/2}$; $\ell=1,\dots,k$; $z_k\geq\dots\geq z_1=0$. Then define $C_1=0$ and for $k\geq 2$

$$C_k=\max_U\sum_{i=1}^{k}z_i\int\limits_{-\infty}^{\infty}\varphi(t+z_i)\prod_{\substack{j=1\\(j\neq i)}}^{k}\Phi(t+z_j)dt,$$

where $U=\{(z_1,\dots,z_k): z_k\geq\dots\geq z_1=0\}$. These constants can be calculated by numerical methods. Some particular values are given below

$$C_2\approx 0.240, C_3\approx 0.372, C_4\approx 0.463.$$

Lemma 3.3. Let $p\in[d,1-d]$, $0<d<0.5$, $b_{r-2}\to+\infty$, $\tau_{r-1}\to+\infty$ as $T\to+\infty$ and $z_k\geq\dots\geq z_1=0$. Then following estimates hold as $T\to\infty$

$$\sup_{z_k\leq b_{r-2}}\left(\sum_{l=1}^{k}m_l\mathbf{P}\left(X_{r-1}(l)=\max_{\ell=1,\dots,k}X_{r-1}(\ell)\right)\right)\times$$

$$\times(V/\tau_{r-1})^{-1/2}\sim(V(p)/V)^{1/2}C_k, \tag{15}$$

$$\sup_{z_2\geq b_{r-2}}\left(\sum_{l=1}^{k}m_l\mathbf{P}\left(X_{r-1}(l)=\max_{\ell=1,\dots,k}X_{r-1}(\ell)\right)\right)\times$$

$$\times(V/\tau_{r-1})^{-1/2}\to 0, \tag{16}$$

and for $k>2$

$$C_k\leq C_2(k-1). \tag{17}$$

Proof. The estimates (15) and (16) can be proved in the similar way as (12). To prove (17) note, that

$$C_2=\max_{z_2\geq 0}z_2\int\limits_{-\infty}^{\infty}\varphi(t+z_2)\Phi(t)dt,$$

$$C_k\leq\max_U\sum_{i=2}^{k}z_i\int\limits_{-\infty}^{\infty}\varphi(t+z_i)\Phi(t)dt$$

for $k>2$ and, hence, (17) holds. ■

3.2 Scheme of estimation of the minimax risk

Let's make some notations. Without loss of generality we assume that $w_1(\theta) \geq w_2(\theta) \cdots \geq w_K(\theta)$ and denote $m_\ell = w_1(\theta) - w_\ell(\theta)$, so that $m_K \geq \cdots \geq m_2 \geq m_1 = 0$. Denote by $\mathcal{D}_i = (\mathcal{I}_i, \ell_{ij_0}, \mathcal{J}_i)$ the statistics $\{X_i(\ell); \ell \in \mathcal{I}_i\}$ collected at the i-th learning stage with the set of tested alternatives $\mathcal{I}_i$ which results to the better alternative ℓ_{ij_0} and the set of rejected alternatives $\mathcal{J}_i$, by $\mathcal{C}_i = \mathcal{D}_1 \dots \mathcal{D}_i$ the total statistics up to the i-th learning stage describing the changes of $\mathcal{D}_1, \dots, \mathcal{D}_i$. Obviously, $\mathcal{C}_i = \mathcal{C}_{i-1}\mathcal{D}_i$ if $i \geq 2$. Denote by $\#\mathcal{I}_i$ the number of elements in the set of indices $\mathcal{I}_i$. We denote by $\mathcal{C}_i^+ = \mathcal{C}_i \cap \{\#\mathcal{I}_{i+1} > 1\}$ the subsets of $\mathcal{C}_i$ which prescribe to continue learning stages, by $\mathcal{D}_i^- = \mathcal{D}_i \cap \{\#\mathcal{I}_{i+1} = 1\}$ the subsets of $\mathcal{D}_i$ which prescribe immediately to go to the final stage. If $r = 2$ then

$$L_T(\pi, \theta) = \sum_{\ell=1}^{K} m_\ell \tau_1 + \sum_{\{\mathcal{D}_1^-\}} \mathbf{P}(\mathcal{D}_1^-) m_{\ell_{1j_0}} (T - t_1).$$

If $r \geq 3$ then

$$L_T(\pi, \theta) = \sum_{\ell=1}^{K} m_\ell \tau_1 + \sum_{\{\mathcal{D}_1^-\}} \mathbf{P}(\mathcal{D}_1^-) m_{\ell_{1j_0}} (T - t_1) +$$

$$+ \sum_{i=2}^{r-1} \left\{ \sum_{\{\mathcal{C}_{i-1}^+\}} \mathbf{P}(\mathcal{C}_{i-1}^+) \sum_{\ell \in \mathcal{I}_i} m_\ell \tau_i + \right.$$

$$\left. + \sum_{\{\mathcal{C}_{i-1}^+ \mathcal{D}_i^-\}} \mathbf{P}(\mathcal{C}_{i-1}^+ \mathcal{D}_i^-) m_{\ell_{ij_0}} (T - t_i) \right\}.$$

To simplify notations, dependence of $\mathbf{P}(\mathcal{D}_1^-)$, $\mathbf{P}(\mathcal{C}_{i-1}^+)$, $\mathbf{P}(\mathcal{C}_{i-1}^+ \mathcal{D}_i^-)$ and m_ℓ on θ is omitted.

In the sequel we focus on the case $r \geq 3$. Let's denote by $\hat{\mathcal{C}}_i^+ = \mathcal{C}_i^+ \cap \{\ell_{i,1} = 1, \ell_{i+1,1} = 1\}$, $\hat{\mathcal{D}}_i^- = \mathcal{D}_i^- \cap \{\ell_{i,1} = 1\}$, i.e. the alternative $\ell = 1$ was not rejected up to the i-th stage. Obviously, estimation holds

$$L_T(\pi, \theta) \geq \hat{L}_T(\pi, \theta) = \sum_{\ell=1}^{K} m_\ell \tau_1 +$$

$$+ \sum_{i=2}^{r-1} \sum_{\{\hat{\mathcal{C}}_{i-1}^+\}} \mathbf{P}(\hat{\mathcal{C}}_{i-1}^+) \sum_{\ell \in \mathcal{I}_i} m_\ell \tau_i +$$

$$+ \sum_{\{\hat{\mathcal{C}}_{r-2}^+ \hat{\mathcal{D}}_{r-1}^-\}} \mathbf{P}(\hat{\mathcal{C}}_{r-2}^+ \hat{\mathcal{D}}_{r-1}^-) m_{\ell_{r-1,j_0}} (T - t_{r-1}).$$

The estimate (4) can be obtained for $\hat{L}_T(\pi, \theta)$ and then generalized to $L_T(\pi, \theta)$. Let $\tau_r = T - t_{r-1}$ and suppose at first that the requirements hold

$$\tau_1 \gtrsim T^{\varepsilon_0}, \ \tau_{i+1}/\tau_i \gtrsim T^{\varepsilon_0}, \ i = 1, \dots, r-1 \quad (18)$$

for some $\varepsilon_0 > 0$ and

$$b_i \to +\infty, \ b_i T^{-\varepsilon} \to 0, \ i = 1, \dots, r-2 \quad (19)$$

for all $\varepsilon > 0$ as $T \to \infty$. Let $b_{r-1} = 0$.

For $w \in [d, 1-d]$ with sufficiently small d define intervals $P_i(w)$ $(i = 1, \dots, r)$

$$P_1(w) = \{p : p = w - x, b_1^{-1} \leq x \leq \Delta W\} \cap [0, w],$$

$$P_i(w) = \{p : p = w - x(2V/\tau_{i-1})^{1/2}, x \in U(b_{i-1})\},$$

$$i = 2, \dots, r-1,$$

$$P_r(w) = \{p : p = w - z(V/\tau_{r-1})^{1/2}, 0 \leq z \leq b_{r-2}\}$$

and

$$P_0(w) = [0, w] \setminus \bigcup_{i=1}^{r} P_i(w).$$

$P_i(w) \cap P_j(w) = \emptyset$ for sufficiently large T and all $i \neq j$, $w \in [d, 1-d]$. Obviously, $\bigcup_{i=0}^{r} P_i(w) = [0, w]$.

Then define the set of indices $\mathcal{N} = \{n_2, \dots, n_K\}$ with $0 \leq n_\ell \leq r$, $\ell = 2, \dots, K$ and $n_2 \geq \cdots \geq n_K$. For every $w \in [d, 1-d]$ let

$$\Theta(w, \mathcal{N}) = \{\theta : p_1 = w, p_\ell \in P_{n_\ell}(w), \ell = 2, \dots, K\}$$

and denote by $\#_i\mathcal{N}$ the number of elements $n_\ell \in \mathcal{N}$ such that $n_\ell = i$. One can say of ℓ-th component of $\Theta(w, \mathcal{N})$ that it can contribute to total loss on the n_ℓ-th stage only ($n_\ell > 0$) and its contribution on other stages is negligible because m_ℓ is either too small or sufficiently large but ℓ-th alternative has been already rejected on previous stages. If $n_\ell = 0$ then contribution of this component is negligible on all stages. Denote

$$f_1(\pi, w) = \min(w, \Delta W)\tau_1,$$

$$f_i(\pi) = b_{i-1}(2V/\tau_{i-1})^{1/2}\tau_i,$$

$$f_r(\pi, n) = C_n(V/\tau_{r-1})^{1/2}T,$$

$$i = 2, \dots, r-1, n = 1, \dots, K.$$

It can be shown that

$$\sup_{\Theta(w,\mathcal{N})} \hat{L}_T(\pi, \theta) = \mathcal{L}_T(\pi, w, \mathcal{N}),$$

where

$$\mathcal{L}_T(\pi, w, \mathcal{N}) = (\#_1\mathcal{N} + \varepsilon_{1\mathcal{N}}) f_1(\pi, w) +$$

$$+ \sum_{i=2}^{r-1} (\#_i\mathcal{N} + \varepsilon_{i\mathcal{N}}) f_i(\pi) + (1 + \varepsilon_{r\mathcal{N}}) f_r(\pi, \#_r\mathcal{N} + 1),$$

and $\varepsilon_{i\mathcal{N}} \to 0$ as $T \to \infty$ for all $\mathcal{N}$, i. The estimates of $\sup\left(\sum_{\{\hat{\mathcal{C}}_{i-1}^+\}} \sum_{\ell \in \mathcal{I}_i} m_\ell \mathbf{P}(\hat{\mathcal{C}}_{i-1}^+)\right)$, $i = 2, \dots, r-1$ use (12), (13). The estimate of $\sup\left(\sum_{\{\hat{\mathcal{C}}_{r-2}^+ \hat{\mathcal{D}}_{r-1}^-\}} m_{\ell_{r-1,j_0}} \mathbf{P}(\hat{\mathcal{C}}_{r-2}^+ \hat{\mathcal{D}}_{r-1}^-)\right)$ uses (15), (16). If $w \notin [d, 1-d]$, so that the central limit theorem cannot be used, the estimates can be based on the Chebishev's inequality, the corresponding value of the loss function is negligible in comparison with presented above. Hence

$$R_T(\Theta) \geq \inf_\pi \sup_\Theta \hat{L}_T(\pi, \theta) =$$

$$= \inf_\pi \max_{\mathcal{N}} \sup_{w \in [d, 1-d]} \sup_{\Theta(w,\mathcal{N})} \hat{L}_T(\pi, \theta) =$$

$$= \inf_\pi \max_{\{\mathcal{N}\}} \mathcal{L}_T(\pi, \Delta W, \mathcal{N}),$$

and, therefore,

$$R_T(\Theta) \geq \inf_{\pi} \max_{i=1,\ldots,r} \mathcal{L}_T(\pi, \Delta W, \mathcal{N}_i), \qquad (20)$$

where each $\mathcal{N}_i = (n_2, \ldots, n_K)$ is defined as $n_2 = i$, $n_3 = \cdots = n_K = 1$. Since each function $f_i(\cdot)$ is increasing in τ_i and decreasing in τ_{i-1} (excepts $f_r(\cdot)$ and $f_1(\cdot)$ which do not depend on τ_i and τ_{i-1} respectively) the inf in (20) is achieved for the strategy π^* balancing functions

$$\mathcal{L}_T(\pi^*, \Delta W, \mathcal{N}_1) = \cdots = \mathcal{L}_T(\pi^*, \Delta W, \mathcal{N}_r)$$

or equivalently

$$f_1(\pi^*, \Delta W) = f_i(\pi^*) = f_r(\pi^*, 2), \qquad (21)$$
$$i = 2, \ldots, r-1.$$

Since $f_r(\pi, n+1) \leq n f_r(\pi, 2)$, $n \geq 2$, due to (17), it follows from (21) that

$$\mathcal{L}_T(\pi^*, \Delta W, \mathcal{N}) \leq \mathcal{L}_T(\pi^*, \Delta W, \mathcal{N}_1)$$

holds for strategy π^* for all $\{\mathcal{N}\}$. Assuming that $\tau_i = (a_i T^{\alpha_i})^2$, $i = 1, \ldots, r-1$, where $\{a_i\}$ are slow functions of T (i.e. $a_i^{-1} T^{-\varepsilon} \to 0$, $a_i T^{-\varepsilon} \to 0$ as $T \to \infty$, $\forall \varepsilon > 0$), we find from (21) that $\{\alpha_i\}$ satisfy conditions (1). The least possible $\{b_i\}$ such that

$$\mathcal{L}_T^{-1}(\pi^*, \Delta W, \mathcal{N}_1) \times \left(\sum_{\{\hat{\mathcal{D}}_1^-\}} \mathbf{P}(\hat{\mathcal{D}}_1^-) m_{\ell_{1 j_0}}(T - t_1) + \right.$$

$$\left. + \sum_{i=2}^{r-2} \sum_{\{\hat{\mathcal{C}}_{i-1}^+ \hat{\mathcal{D}}_i^-\}} \mathbf{P}(\hat{\mathcal{C}}_{i-1}^+ \hat{\mathcal{D}}_i^-) m_{\ell_{i j_0}}(T - t_i) \right) \to 0$$

as $T \to +\infty$ should be chosen from (2) (here estimate (14) is used). Then $\{a_i\}$ are chosen from (3). Then it can be shown that π^* asymptotically does not increase the value of risk for all probability events $\{\mathcal{C}_i^+, \mathcal{D}_i^-\}$. Finally, restrictions (18), (19) can be removed because for any other strategy π there exists such parameter θ that corresponding loss exceeds $\mathcal{L}_T(\pi^*, \Delta W, \mathcal{N}_1)$.

4. CONCLUSION

As mentioned above, considered approach allows parallel processing of handled units on all stages. In this case total duration of the process is often mainly influenced by the number of stages r rather than by the number of processed units T. Since the minimax risk is close to unimprovable one for moderate r, considered approach can be especially recommended for usage in such systems.

As a more vivid example of this property the medical treatment of a large group of T patients by K alternative medicines with different and unknown effectiveness can be considered. The goal is to maximize the expected total number of successful treatments.

The application of the r-stage approach assumes that at the beginning of each stage the medicines are given to the patients of appropriate groups simultaneously. Then some time is necessary for passage of the course of treatment. At the end of the stage the results of treatment are analyzed.

In this example the duration of course of treatment is much larger than the duration of medicines assignment and analyzing the results. Hence, the total duration of the process is determined by the duration of r successive courses of treatment and not by the number of patients T.

Since all estimates have asymptotic character, it is interesting to know how they behave for particular finite values T for different r. The general rule is that conditions $\tau_i \ll \tau_{i+1}$, $i = 1, \ldots, r-2$ and $\tau_1 + \cdots + \tau_{r-1} \ll T$ should hold, where τ_i, $i = 1, \ldots, r-1$ are determined by theorems 2.1 and 2.2. If these conditions do not hold, the approach considered in (Kolnogorov, 2000) for close distributions may be useful.

REFERENCES

Berry, D.A. and B. Fristedt (1985). *Bandit Problems: Sequential Allocation of Experiments.* Chapman and Hall. London, New York.

Cheng, Y. (1994). Multistage decision problems. *Sequential Analysis* **13**, 329–350.

Kolnogorov, A.V. (1999). A simple strategy of behavior in a stationary medium with exponential guaranteed convergence rate. *Automation and Remote Control* **60**, 1136–1140. (Translated from Russian).

Kolnogorov, A.V. (2000). On optimal prior learning time in the two-armed bandit problem. *Probl. Inf. Transm.* **36**, 387–396. (Translated from Russian).

Kolnogorov, A.V. and S.V. Melnikova (2005). On optimal duration of initial stage in the two-stage model of expedient behaviour in random medium. *Vestnik Novgorodskogo Universiteta* **34**, 73–75. (In Russian).

Prokhorov, Yu.V. and Yu.A. Rozanov (1987). *Theory of Probability.* Nauka. Moscow. (In Russian).

Sragovich, V.G. (1981). *Adaptive Control.* Nauka. Moscow. (In Russian).

Tsetlin, M.L. (1973). *Automation Theory and Modeling of Biological Systems.* Academic Press. New York. (Translated from Russian).

Vogel, W. (1960). An asymptotic minimax theorem for the two-armed bandit problem. *Ann. Math. Stat.* **31**, 444–451.

Witmer, J.A. (1986). Bayesian multistage decision problems. *Ann. Stat.* **14**, 283–297.