UDC 519.244.3

An Asymptotic Minimax Theorem for Gaussian Two-Armed Bandit

A. V. Kolnogorov*

* *Department of Applied Mathematics and Computer Science, Yaroslav-the-Wise Novgorod State University, B-St-Petersburgskaya str. 41, Velikiy Novgorod, 173003, Russia*

Abstract. The asymptotic minimax theorem for Bernoulli two-armed bandit problem states that minimax risk has the order $N^{1/2}$ as $N \to \infty$, where N is the control horizon, and provides the estimates of the factor. For Gaussian two-armed bandit with unit variances of one-step incomes and close expectations, we improve the asymptotic minimax theorem as follows: the minimax risk is approximately equal to $0.637N^{1/2}$ as $N \to \infty$.

Keywords: two-armed bandit problem, minimax and Bayesian approaches, an asymptotic minimax theorem, parallel processing.

1. Introduction

We consider the two-armed bandit problem which is also well-known as the problem of expedient behavior in a random environment and the problem of adaptive control (see, e.g. [1], [2]). Let ξ_n, $n = 1, \ldots, N$ be a controlled random process which values are interpreted as incomes, depend only on currently chosen actions y_n ($y_n \in \{1, 2\}$) and are distributed as follows

$$\Pr(\xi_n = 1 | y_n = \ell) = p_\ell, \quad \Pr(\xi_n = 0 | y_n = \ell) = q_\ell,$$

$p_\ell + q_\ell = 1$, $\ell = 1, 2$. This is Bernoulli two-armed bandit. It can be described by a parameter $\theta = (p_1, p_2)$ with the set of values $\Theta = \{\theta : 0 \le p_\ell \le 1; \ell = 1, 2\}$. The results can be applied to optimizing of data processing. In this case $\xi_n = 1$ and $\xi_n = 0$ correspond to successively and unsuccessively processed data. The goal is to maximize (in some sense) the total expected number of successively processed data.

Control strategy σ at the point of time n assigns a random choice of the action y_n depending on the current history of the process, i.e. responses $x^{n-1} = x_1, \ldots, x_{n-1}$ to applied actions $y^{n-1} = y_1, \ldots, y_{n-1}$. Denote by $m_\ell = p_\ell$ one-step expected income if the ℓ-th action is applied. If the parameter θ is known then the optimal strategy prescribes always to choose the action corresponding to the largest of m_1, m_2, the total expected income is thus equal to $N(m_1 \vee m_2)$. If the parameter is unknown then the regret

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

describes the expected losses of income due to incomplete information. Minimax risk is equal to

$$R_N^M(\Theta) = \inf_{\sigma} \sup_{\theta} L_N(\sigma, \theta), \tag{1}$$

corresponding optimal strategy is called the minimax strategy. The minimax approach to the problem was proposed in [3] and caused a considerable interest to it. It was shown in [4] that explicit determination of the minimax strategy and minimax risk is virtually impossible already for $N \geq 5$. However, an asymptotic minimax theorem was proved in [5] using some indirect techniques. This theorem states that the following estimates hold as $N \to \infty$:

$$0.612 \leq (DN)^{-1/2} R_N^M(\Theta) \leq 0.752, \tag{2}$$

where $D = 0.25$ is the maximal variance of one-step income. Presented here the lower estimate was obtained in [6]. The maximal values of the regret calculated for proposed in [5] strategy correspond to $|p_1 - p_2| \approx 3.78(D/N)^{1/2}$ with additional requirement that p_1, p_2 are close to 0.5.

The goal of our paper is to improve the estimates (2). Our approach is based on the main theorem of the theory of games which sets the relation between minimax and Bayesian approaches. Let $\lambda(\theta)$ be a prior probability density. The value

$$R_N^B(\lambda) = \inf_{\sigma} \int_{\Theta} L_N(\sigma, \theta) \lambda(\theta) d\theta \tag{3}$$

is called the Bayesian risk and corresponding optimal strategy is called the Bayesian strategy. Minimax risk (1) and Bayesian risk (3) are related as follows

$$R_N^M(\Theta) = R_N^B(\lambda_0) = \sup_{\lambda} R_N^B(\lambda), \tag{4}$$

where λ_0 is called the worst-case prior distribution. The advantage of the Bayesian approach is that it allows to find explicitly Bayesian strategy and Bayesian risk by solving a recursive Bellman-type equation. However, a direct usage of the main theorem of the theory of games is virtually impossible due to the high computational complexity. Therefore, at first we describe the properties of the worst-case prior and this allows to simplify Bellman-type equation significantly. This is easily made for Gaussian two-armed bandit which occurs if the parallel control in the random environment is implemented, i.e. the same actions are applied to groups of incoming data and then cumulative incomes are used for the control (see [7, 8]). For Gaussian two-armed bandit we obtain the invariant Bellman-type recursive equation with the unit control horizon and second-order partial differential equation in the limiting case. Finally, the asymptotic estimate of the minimax risk is obtained by numerical methods. For Gaussian two-armed bandit with unit variances of one-step incomes

and close expectations, we estimate the minimax risk as approximately equal to $0.637N^{1/2}$ as $N \to \infty$.

2. Main section

Incomes of the Gaussian two-armed bandit ξ_n, $n = 1, \dots, N$ have probability densities $f(x|m_\ell)$ if $y_n = \ell$, $\ell = 1, 2$ where

$$f(x|m) = (2\pi)^{-1/2} \exp\left\{-(x-m)^2/2\right\}.$$

So, Gaussian two-armed bandit is described by a vector parameter $\theta = (m_1, m_2)$. The set of parameters is $\Theta = \{\theta : |m_1 - m_2| \le 2cN^{-1/2}\}$ with $c > 0$ large enough. It is a set of "close expectations", maximal expected regret is attained just on this set of parameters.

According to the approach based on the main theorem of the theory of games (see equality (4)) we characterize the worst-case prior distribution. It is convenient to modify parametrization. Let's put $m_1 = m + v$, $m_2 = m - v$, then $\theta = (m+v, m-v)$ and $\Theta = \{\theta : |v| \le cN^{-1/2}\}$. Asymptotically the worst-case prior distribution density can be chosen as

$$\nu_a(m, v) = \kappa_a(m)\rho(v), \tag{5}$$

where $\kappa_a(m)$ is the uniform density if $|m| \le a$, $\rho(v)$ is a symmetric density (i.e. $\rho(-v) = \rho(v)$) if $|v| \le cN^{-1/2}$ and $a \to \infty$.

Let's present the dynamic programming equation for calculation Bayesian risk and Bayesian strategy with respect to (5). Denote by n_1, n_2, X_1, X_2 current applications of both actions and corresponding total incomes up to the point of time $n = n_1 + n_2$. We assume that once being chosen action is applied M times (so, the group control is considered). It turned out for the prior (5) that at the time point n control is completely determined by a triple (U, n_1, n_2) with $U = (X_1 n_2 - X_2 n_1)n^{-1}$.

We present the equation in invariant form with unity control horizon. Let's put $\varepsilon = MN^{-1}$, $t_1 = n_1 N^{-1}$, $t_2 = n_2 N^{-1}$, $t = nN^{-1}$, $u = UN^{-1/2}$, $w = vN^{1/2}$, $c = CN^{1/2}$, $\varrho(w) = N^{1/2}\rho(v)$. Denote $f_D(u) = (2\pi D)^{-1/2} \exp(-u^2/(2D))$.

Bayesian strategy at the initial stage $t \le 2\varepsilon$ ($n \le 2M$) applies actions turn-by-turn. Then it can be determined by solving the following recursive equation:

$$r_\varepsilon(u, t_1, t_2) = \min_{\ell=1,2} r_\varepsilon^{(\ell)}(u, t_1, t_2), \tag{6}$$

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

$$r_\varepsilon^{(1)}(u, t_1, t_2) = \varepsilon g^{(1)}(u, t_1, t_2) + r_\varepsilon(u, t_1 + \varepsilon, t_2) * f_{\varepsilon t_2^2 t^{-1}(t+\varepsilon)^{-1}}(u), \tag{7}$$

$$r_\varepsilon^{(2)}(u,t_1,t_2) = \varepsilon g^{(2)}(u,t_1,t_2) + r_\varepsilon(u,t_1,t_2+\varepsilon) * f_{\varepsilon t_1^2 t^{-1}(t+\varepsilon)^{-1}}(u) \quad (8)$$

if $t_1 + t_2 < 1$. Here '$*$' denotes convolution,

$$g^{(\ell)}(u,t_1,t_2) = \int_0^c 2w \exp\left((-1)^\ell 2uw - 2w^2 t_1 t_2 t^{-1}\right) \varrho(w) dw,$$

$\ell = 1, 2$. If $t > 2\varepsilon$ ($n > 2M$) then the ℓ-th action is currently optimal iff $r_\varepsilon^{(\ell)}(u,t_1,t_2)$ has smaller value ($\ell = 1, 2$). Bayesian risk corresponding to the prior (5) is calculated according to the formula

$$N^{-1/2} R_N^B(\rho(v)) = 4\varepsilon \int_0^c w\varrho(w)dw + \int_{-\infty}^{\infty} r_\varepsilon(u,\varepsilon,\varepsilon) f_{0.5\varepsilon}(u) du, \quad (9)$$

where

$$N^{-1/2} R_N^B(\rho(v)) = \lim_{a\to\infty} N^{-1/2} R_N^B(\nu_a(m,v)).$$

It is proved in [7,8] that there exists a limit $r(u,t_1,t_2) = \lim r_\varepsilon(u,t_1,t_2)$ as $\varepsilon \to +0$. The limiting description of $r(u,t_1,t_2)$ is given by the second-order partial differential equation which follows from (6), (7), (8). Equation (6) must be written as $\min(r_\varepsilon^{(\ell)}(u,t_1,t_2) - r_\varepsilon(u,t_1,t_2)) = 0$, $\ell = 1, 2$. The differential equation is as follows

$$\min_{\ell=1,\,2} \left(\frac{\partial r}{\partial t_\ell} + \frac{t_{\overline{\ell}}^2}{2t^2} \times \frac{\partial^2 r}{\partial u^2} + g^{(\ell)}(u,t_1,t_2) \right) = 0 \quad (10)$$

with $\overline{\ell} = 3 - \ell$. Initial and boundary conditions take the form

$$r(u,t_1,t_2)|_{t_1+t_2=1} = 0,\ r(\infty,t_1,t_2) = r(-\infty,t_1,t_2) = 0. \quad (11)$$

The optimal strategy prescribes to chose the ℓ-th action if the ℓ-th member in the left-hand side of (10) has minimal value. Note that $\varepsilon \to 0$ implies $N \to \infty$. So, the limiting value of Bayesian risk according to (9) is given by

$$\lim_{N\to\infty} N^{-1/2} R_N^B(\rho(v)) = r(0,0,0). \quad (12)$$

To estimate the limiting value of the minimax risk according to (4) and (12) we assumed that the worst-case prior $\varrho(w)$ is a degenerate one and is concentrated at two points $w = \pm d$ with equal probabilities. Calculations of $r(u,t_1,t_2)$ as a function of d were implemented according to (10), (11) for $t_1 + t_2 \geq \varepsilon$ with $\varepsilon = 0.001$. Partial derivatives were replaced by

partial differences with $\Delta u = 0.023$, $\Delta t = 2000^{-1}$. For $0.5 \le d \le 2.5$ maximum of $2d\varepsilon + r(0, \varepsilon, \varepsilon)$ was approximately equal to 0.637 at $d \approx 1.57$. Finally, the determined optimal strategy was applied to calculate regrets for $0.5 \le d \le 2.5$. As the regret does not exceed the value 0.637 at $d \approx 1.57$, this confirms the assumption of the worst-case prior.

3. Conclusions

The asymptotic minimax theorem for Gaussian two-armed bandit was obtained for the case of close mathematical expectations. To generalize it to the case $\Theta = \{\theta : |m_1 - m_2| \le 2C\}$, $0 < C < \infty$ one should provide the separation of close and distant expectations at the initial stage of control. Some ideas are presented in [7]. Applications of the results to Bernoulli two-armed bandit may be done by usage of group control.

References

1. *Berry D. A., Fristedt B.* Bandit Problems: Sequential Allocation of Experiments. — Chapman and Hall, London, New York, 1985.
2. *Sragovich V. G.* Mathematical Theory of Adaptive Control. — World Scientific. Interdisciplinary Mathematical Sciences, New Jersey, London, 2006. — Vol. 4.
3. *Robbins H.* Some aspects of the sequential design of experiments // Bulletin AMS. — 1952. — Vol. 58, no. 5. — P. 527–535.
4. *Fabius J., van Zwet W. R.* Some remarks on the two-armed bandit // Ann. Math. Statist. — 1970. — Vol. 41. — P. 1906–1916.
5. *Vogel W.* An asymptotic minimax theorem for the two-armed bandit problem // Ann. Math. Stat. — 1960. — Vol. 31. — P. 444–451.
6. *Bather J. A.* The minimax risk for the two-armed bandit problem // Lecture Notes in Statistics. Springer-Verlag, New York. — 1983. — Vol. 20. — P. 1–11.
7. *Kolnogorov A. V.* Parallel design of robust control in the stochastic environment (the two-armed bandit problem) // Automation and Remote Control. — 2012. — Vol. 73, no. 4. — P. 689–701.
8. *Kolnogorov A. V.* On a limiting description of robust parallel control in a random environment // Automation and Remote Control. — 2015. — Vol. 76, no. 7. — P. 1229–1241.