Poissonian two-armed bandit problem

A. V. Kolnogorov [1]

[1]Yaroslav-the-Wise Novgorod State University, Velikiy Novgorod, Russia, Alexander.Kolnogorov@novsu.ru

Bayesian setting for poissonian two-armed bandit is considered. Recursive equation for piece-wise constant strategies and a partial differential equation in the limiting case are obtained.

Introduction. We consider the two-armed bandit problem (see, e.g. Berry and Fristedt [1], Presman and Sonin [2]) which has numerous applications in medicine, economics, data processing, internet technologies, etc. The essential feature of considered setting is a continuous time. Currently relatively few continuous time settings of the two-armed bandit problem are known. For example, continuous time one-armed bandit problem was considered for Wiener processes in Berry and Fristedt [1], Chernoff and Ray [4]. Two settings were considered in Mandelbaum [5] for deteriorating and diffusion multi-armed bandits with geometric discounting on the infinite control horizon. For poisson processes, the problem was considered in Presman and Sonin [2], Presman [3], where a number of important results were obtained, e.g., a thresholding nature of control strategy and a method of the optimal control synthesis. However, an important disadvantage of the approach in Presman and Sonin [2], Presman [3] is that it is restricted to the finite number sets of admissible values of parameters. This is because a control strategy in Presman and Sonin [2], Presman [3] depends on the evolution of the posterior distribution on the set of parameters. This evolution is described by the system of ODE which dimension is precisely equal to the number of parameters. In what follows, we propose an approach which is free of the requirement of the finiteness of the set of parameters.

Poissonian two-armed bandit. Formally, poissonian two-armed bandit is a right-continuous jump-like random process $\{X(t), 0 \le t \le T\}$ which values are interpreted as cumulative incomes increasing by one at the moments of jumps. A control is carried out using two actions. Let's use a notation $y((t, t+\varepsilon]) = \ell$ if at the half-interval $t' \in (t, t+\varepsilon]$, $\varepsilon > 0$ the action $y(t') = \ell$ was permanently chosen ($\ell = 1, 2$). By using such control the increments of the process $X(t)$ depend on chosen actions as follows

$$\Pr\left(X(t+\varepsilon) - X(t) = i | y((t, t+\varepsilon]) = \ell\right) = p(i, \varepsilon; \lambda_\ell) = \frac{(\lambda_\ell \varepsilon)^i}{i!} e^{-\lambda_\ell \varepsilon},$$

$i = 0, 1, 2, \ldots$; $\ell = 1, 2$. Note that for small ε the following approximate formulas hold: $p(0, \varepsilon; \lambda_\ell) = 1 - \lambda_\ell \varepsilon + o(\varepsilon)$, $p(1, \varepsilon; \lambda_\ell) = \lambda_\ell \varepsilon + o(\varepsilon)$, $p(i, \varepsilon; \lambda_\ell) = o(\varepsilon)$, $i = 2, 3, \ldots$; $\ell = 1, 2$. Hence, a vector parameter $\theta = (\lambda_1, \lambda_2)$, where λ_1, λ_2 are

the rates of the flow, completely describes poissonian two-armed bandit. The set Θ of admissible values of parameter θ is known.

A control strategy σ at the point of time t determines the choice of action $y((t,t+\varepsilon])$ on the half-interval $(t,t+\varepsilon]$ depending on the current history, i.e., cumulative times of both actions applications t_1, t_2 $(t_1+t_2=t)$ and corresponding cumulative incomes X_1, X_2. Denote current values of incomes X_1, X_2 at the point of time t by $X_1(t), X_2(t)$. The loss function is defined as

$$L_T(\sigma,\theta) = T\max(\lambda_1,\lambda_2) - \mathbf{E}_{\sigma,\theta}\left(X_1(T)+X_2(T)\right)$$

and describes expected losses of cumulative income with respect to its maximum possible value due to incomplete information. Here $\mathbf{E}_{\sigma,\theta}$ denotes mathematical expectation over the measure generated by strategy σ and parameter θ. Let's assign the prior distribution density $\mu(\theta)=\mu(\lambda_1,\lambda_2)$ on Θ. Bayesian risk computed with respect to the prior distribution density $\mu(\theta)$ is equal to

$$R_T^B(\mu) = \inf_{\{\sigma\}} L_T(\sigma,\mu), \tag{1}$$

corresponding optimal strategy σ^B is called Bayesian strategy.

Recursive equation. Given a prior distribution density $\mu(\lambda_1,\lambda_2)$, denote by $R_\varepsilon^B(X_1,t_1,X_2,t_2)$ Bayesian risk computed with respect to the posterior distribution density $\mu(\lambda_1,\lambda_2|X_1,t_1,X_2,t_2)$. The subscript ε is due to the usage of piece-wise constant strategy σ. Denote $R_\varepsilon(X_1,t_1,X_2,t_2) = R_\varepsilon^B(X_1,t_1,X_2,t_2)\mu(X_1,t_1,X_2,t_2)$ with

$$\mu(X_1,t_1,X_2,t_2) = \iint_\Theta p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2.$$

Then the following recursive equation holds

$$R_\varepsilon(X_1,t_1,X_2,t_2) = \min(R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2), R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)), \tag{2}$$

where

$$R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2) = R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2) = 0 \tag{3}$$

if $t_1+t_2=T$ and

$$\begin{aligned} R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2) &= \varepsilon g^{(1)}(X_1,t_1,X_2,t_2) \\ +\sum_{j=0}^{\infty} R_\varepsilon(X_1+j,t_1+\varepsilon,X_2,t_2) &\times \frac{t_1^{X_1}\varepsilon^j(X_1+j)!}{(t_1+\varepsilon)^{X_1+j}X_1!j!}, \\ R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2) &= \varepsilon g^{(2)}(X_1,t_1,X_2,t_2) \\ +\sum_{j=0}^{\infty} R_\varepsilon(X_1,t_1,X_2+j,t_2+\varepsilon) &\times \frac{t_2^{X_2}\varepsilon^j(X_2+j)!}{(t_2+\varepsilon)^{X_2+j}X_2!j!}, \end{aligned} \tag{4}$$

if $0 \le t_1 + t_2 < T$, where

$$g^{(1)}(X_1,t_1,X_2,t_2) = \iint_\Theta (\lambda_2-\lambda_1)^+ p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2,$$
$$g^{(2)}(X_1,t_1,X_2,t_2) = \iint_\Theta (\lambda_1-\lambda_2)^+ p(X_1,t_1;\lambda_1)p(X_2,t_2;\lambda_2)\mu(\lambda_1,\lambda_2)d\lambda_1 d\lambda_2.$$

Bayesian strategy σ^B prescribes to choose ℓ-th action if $R_\varepsilon^{(\ell)}(X_1,t_1,X_2,t_2)$ has smaller value. In case of a draw $R_\varepsilon^{(1)}(X_1,t_1,X_2,t_2) = R_\varepsilon^{(2)}(X_1,t_1,X_2,t_2)$ the choice is arbitrary. Bayesian risk (1) is computed by the formula $R_{\varepsilon,T}(\mu) = R_\varepsilon(0,0,0,0)$.

Limiting description. It can be proved that there exists a limit $R(X_1,t_1,X_2,t_2) = \lim_{\varepsilon\to 0} R_\varepsilon(X_1,t_1,X_2,t_2)$ which is continuous in t_1,t_2 and satisfies the partial differential equation

$$\min_{\ell=1,2}\left(\frac{\partial R}{\partial t_\ell} + D^{(\ell)}R(X_1,t_1,X_2,t_2) + g^{(\ell)}(X_1,t_1,X_2,t_2)\right) = 0 \qquad (5)$$

with initial condition $R(X_1,t_1,X_2,t_2) = 0$ if $t_1+t_2 = 1$, where

$$D^{(1)}R(X_1,t_1,X_2,t_2) = \frac{R(X_1+1,t_1,X_2,t_2)(X_1+1) - R(X_1,t_1,X_2,t_2)X_1}{t_1},$$
$$D^{(2)}R(X_1,t_1,X_2,t_2) = \frac{R(X_1,t_1,X_2+1,t_2)(X_2+1) - R(X_1,t_1,X_2,t_2)X_2}{t_2}.$$

Bayesian risk (1) is computed by the formula $R_T(\mu) = \lim_{t_0\to 0} R(0,t_0,0,t_0)$. Partial differential equation (5) follows from (2)–(4).

We also consider asymptotic behavior of $R_T(\mu)$ as $T\to\infty$ computed with respect to the worst-case prior distribution densities and show that it is the same as the behavior of the minimax risk of the Gaussian two-armed bandit (see, Kolnogorov [6], Kolnogorov [7]).

Acknowledgements. The reported study was funded by RFBR, project number 20-01-00062.

References

1. D. A. Berry, B. Fristedt, *Bandit Problems: Sequential Allocation of Experiments*, London, Chapman & Hall, 1985.

2. E. L. Presman, I. M. Sonin, *Sequential Control with Incomplete Information*, New York, Academic, 1990.

3. E. L. Presman, Poisson version of the two-armed bandit problem with discounting, *Theory of Probability and its Applications* **35**:2 (1990) 307–317.

4. H. Chernoff, S. N. Ray, A Bayes sequential sampling inspection plan, *Ann. Math. Statist.* **36** (1965) 1387–1407.

5. A. Mandelbaum, Continuous multi-armed bandits and multiparameter processes, *Ann. Probab.* **15**:4 (1987) 1527–1556.

6. A. V. Kolnogorov, On a limiting description of robust parallel control in a random environment, *Autom. Remote Control* **76**:7 (2015) 1229–1241.

7. A. V. Kolnogorov, Gaussian two-armed bandit: limiting description, *Problems of Information Transmission* **56**:3 (2020) 278–301.