A. V. Kolnogorov

POISSONIAN TWO-ARMED BANDIT: BAYESIAN APPROACH[1]

We consider a Bayesian approach to a continuous time two-armed bandit problem, in which incomes are described by poissonian processes. The problem is studied in a discrete approximation. To do this, a control horizon is divided into equal consequtive half-intervals, at which the strategy remains piece-wise constant and incomes arise in batches corresponding to these half-intervals. For finding the piece-wise constant Bayesian strategy and corresponding Bayesian risk, a recursive difference equation is obtained. In the limiting case as the number of half-intervals grows infinitely, the existence of a limiting value of the Bayesian risk is established and a partial differential equation for its determining is derived.

We consider a Bayesian approach to poissonian two-armed bandit, which is different from presented in [1]. Poissonian two-armed bandit is a right-continuous jump-like controlled random process $\{X(t), 0 \le t \le T\}$, which values are interpreted as incomes increasing by one at the time points of jumps. A control is carried out using two actions. Let's use a notation $y((t,t+\varepsilon]) = \ell$ if on the half-interval $t' \in (t, t+\varepsilon]$, $\varepsilon > 0$ the action $y(t') = \ell$ was permanently used ($\ell = 1, 2$). If this permanent control is used then increments of the process $X(t)$ depend on chosen actions as follows

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

$i = 0, 1, 2, \dots;\ \ell = 1, 2$. The value $X(t+\varepsilon) - X(t)$ is interpreted as a batch of incomes obtained on the half-interval $(t, t+\varepsilon]$. So, a vector parameter $\theta = (\lambda_1, \lambda_2)$, where λ_1, λ_2 are intensities of incomes' generation, completely describes poissonian two-armed bandit. The set Θ of admissible values of parameter is assumed to be known.

For a control, piece-wise constant strategies are used. At the start of the control both actions are used on the half-intervals of the length t_0. Then a control horizon is divided into equal half-intervals of the length ε, on which the actions remain constant. A control strategy σ at the point of time t, corresponding to the start of the current half-interval, determines a choice (generally speaking, a random) of the action $y((t, t+\varepsilon])$ depending on the known history (X_1, t_1, X_2, t_2).

[1]Supported by RFBR, project number 20-01-00062.

Here t_1, t_2 are current cumulative times of both actions applications ($t_1 + t_2 = t$) and X_1, X_2 are corresponding cumulative incomes.

Let's denote by $X_1(t), X_2(t)$ the current values of incomes at the point of time t. If the values of intensities λ_1, λ_2 were known, one should always choose the action corresponding to the largest of them, the total expected income on the control horizon T is thus $T\max(\lambda_1, \lambda_2)$. The actual expected income is less then the maximal one by the value $L_{\varepsilon,T}(\sigma, \theta) = T\max(\lambda_1, \lambda_2) - \mathbf{E}_{\sigma,\theta}\left(X_1(T) + X_2(T)\right)$, which is called the regret. By $\mathbf{E}_{\sigma,\theta}$ we denote the mathematical expectation computed over the measure generated by the strategy σ and the parameter θ. Here and below the index ε highlights the usage of piece-wise constant strategies. Let's assign a prior distribution density $\mu(\theta) = \mu(\lambda_1, \lambda_2)$ on the set Θ. Bayesian risk computed with respect to a prior distribution density $\mu(\theta)$ is

$$R^B_{\varepsilon,T}(\mu) = \inf_{\{\sigma\}} \int_{\Theta} L_T(\sigma, \mu)\mu(\theta)d\theta, \tag{1}$$

corresponding optimal strategy is called a Bayesian strategy.

Theorem 1. *Consider a recursive difference equation*

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

where

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

if $t_1 + t_2 = T$ *and then*

$$\begin{aligned} R^{(1)}_\varepsilon(X_1, t_1, X_2, t_2) &= \varepsilon g^{(1)}(X_1, t_1, X_2, t_2) + \mathbf{T}^{(1)}_\varepsilon R_\varepsilon(X_1, t_1 + \varepsilon, X_2, t_2), \\ R^{(2)}_\varepsilon(X_1, t_1, X_2, t_2) &= \varepsilon g^{(2)}(X_1, t_1, X_2, t_2) + \mathbf{T}^{(2)}_\varepsilon R_\varepsilon(X_1, t_1, X_2, t_2 + \varepsilon) \end{aligned} \tag{4}$$

if $2t_0 \le t < T$. *Here functions* $\{g^{(\ell)}(X_1, t_1, X_2, t_2)\}$ *and operators* $\{\mathbf{T}^{(\ell)}_\varepsilon\}$ *are as follows*

$$\begin{aligned} g^{(1)}(X_1, t_1, X_2, t_2) &= \iint_{\Theta} (\lambda_2 - \lambda_1)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} \mu(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2, \\ g^{(2)}(X_1, t_1, X_2, t_2) &= \iint_{\Theta} (\lambda_1 - \lambda_2)^+ \lambda_1^{X_1} e^{-\lambda_1 t_1} \lambda_2^{X_2} e^{-\lambda_2 t_2} \mu(\lambda_1, \lambda_2) d\lambda_1 d\lambda_2, \\ \mathbf{T}^{(1)}_\varepsilon F(X_1, t_1, X_2, t_2) &= \sum_{j=0}^{\infty} F(X_1 + j, t_1, X_2, t_2) \times \frac{\varepsilon^j}{j!}, \\ \mathbf{T}^{(2)}_\varepsilon F(X_1, t_1, X_2, t_2) &= \sum_{j=0}^{\infty} F(X_1, t_1, X_2 + j, t_2) \times \frac{\varepsilon^j}{j!}. \end{aligned} \tag{5}$$

Bayesian strategy prescribes to choose the ℓth action (i.e., $\sigma_\ell(X_1,t_1,X_2,t_2) = 1$) if $R_\varepsilon^{(\ell)}(X_1,t_1,X_2,t_2)$ has the smaller value ($\ell = 1,2$). In the 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 of the action is arbitrary. Bayesian risk (1) *is*

$$R_{\varepsilon,T}(\mu) = t_0 \iint\limits_{\Theta} |\lambda_1 - \lambda_2| \mu(\lambda_1,\lambda_2) d\lambda_1 d\lambda_2 + \sum_{X_1=0}^{\infty} \sum_{X_2=0}^{\infty} R_\varepsilon(X_1,t_0,X_2,t_0) \frac{t_0^{X_1} t_0^{X_2}}{X_1! X_2!}, \tag{6}$$

and, in particular, $R_{\varepsilon,T}(\mu) = R_\varepsilon(0,0,0,0)$ if $t_0 = 0$.

In what follows, let's assume that $\varepsilon \to 0$. From (2)–(6) the theorem follows.

Theorem 2. *A limit $R(X_1,t_1,X_2,t_2) = \lim_{\varepsilon\to+0} R_\varepsilon(X_1,t_1,X_2,t_2)$ exists if $t_1 \ge t_0$, $t_2 \ge t_0$. This limit is bounded and satisfies Lipschitz conditions for t_1, t_2. A limiting Bayesian risk* (1) *is*

$$R_T(\mu) = \lim_{t_0\to+0,\,\varepsilon\to+0} R_{\varepsilon,T}(\mu) = \lim_{t_0\to+0} R(0,t_0,0,t_0). \tag{7}$$

A limit $R(X_1,t_1,X_2,t_2)$ satisfies partial differential equation

$$\begin{aligned} &\min\left(\frac{\partial R}{\partial t_1} + R(X_1+1,t_1,X_2,t_2) + g^{(1)}(X_1,t_1,X_2,t_2),\right. \\ &\left.\frac{\partial R}{\partial t_2} + R(X_1,t_1,X_2+1,t_2) + g^{(2)}(X_1,t_1,X_2,t_2)\right) = 0 \end{aligned} \tag{8}$$

with initial condition $R(X_1,t_1,X_2,t_2) = 0$ at $t_1 + t_2 = T$. A limiting Bayesian risk (1) is computed according to (7). Differential equation (8) describes at the same time the evolution of the Bayesian risk $R(X_1,t_1,X_2,t_2)$ and also the Bayesian strategy, which prescribes to choose the ℓth action if the ℓth term on the left-hand side of (8) has the smaller value; in the case of a draw the choice of the action can be arbitrary.

[1] Presman, E.L. and Sonin, I.M. *Sequential Control with Incomplete Information*, New York: Academic, 1990.