Journal of Applied Mathematics and Stochastic Analysis, **12**:2 (1999), 151-160.

EXACT SOLUTION OF THE BELLMAN EQUATION FOR A β-DISCOUNTED REWARD IN A TWO-ARMED BANDIT WITH SWITCHING ARMS

DONCHO S. DONCHEV
Higher Institute of Food and Flavor Industries
26, Maritza str.
4002 Plovdiv, Bulgaria

(Received May, 1998; Revised October, 1998)

We consider the symmetric Poissonian two-armed bandit problem. For the case of switching arms, only one of which creates reward, we solve explicitly the Bellman equation for a β-discounted reward and prove that a myopic policy is optimal.

Key words: Two-Armed Bandit, Continuous Time, Switching Arms, β-Discounted Reward.

AMS subject classifications: 49L20, 60J99.

1. Introduction

We discuss the Two-Armed Bandit (TAB) problem which is a continuous analogue of the discrete-time TAB studied by Feldman [6]. A self-contained statement of the discrete-time TAB can be found also in DeGroot [1] and Dynkin and Yushkevich [5]. The Poissonian version of this problem was stated and studied in detail by Sonin [8] and Presman and Sonin [7]. The symmetric Poissonian TAB is determined by a matrix

$$\Lambda = \begin{pmatrix} \lambda & \mu \\ \mu & \lambda \end{pmatrix}, \quad 0 \leq \mu < \lambda.$$

The columns of Λ represent the left and the right arms of the TAB; the rows correspond to hypotheses H_1 and H_2. Both arms generate Poisson flows of particles. According to H_1, the intensities of the flows corresponding to the left and the right arms are equal to λ and μ, respectively. According to H_2, these intensities are μ and λ. An input flow with intensity 1 can be directed either to the left or to the right arms. A controller, at each time $t \geq 0$, selects probabilities u_t and $1 - u_t$, with which an arriving particle (if any) is directed to the input of the left or the right arm. If a particle gets into an arm with a higher intensity, then it is immediately observed at the output of this arm. If a particle gets into an arm with lower intensity, then the same happens with probability μ/λ; with probability $1 - \mu/\lambda$, the particle remains

Printed in the U.S.A. ©1999 by North Atlantic Science Publishing Company

unobserved. The aim of the controller is to maximize the expected number of observed particles if he knows *a priori* probability x of hypothesis H_1 at time $t = 0$. In this setting, it is shown by Presman and Sonin [7] that the optimal policy is given by the rule $u_t = \varphi(x_t)$, where x_t is a posteriori probability of H_1 at time t, and

$$\varphi(x) = \begin{cases} 0, & \text{if } 0 \le x < \frac{1}{2}, \\ \frac{1}{2}, & \text{if } x = \frac{1}{2}, \\ 1, & \text{if } \frac{1}{2} < x \le 1. \end{cases} \tag{1}$$

A new feature studied in this paper is the random switching of arms (or equivalently, hypotheses) in the TAB. This switching occurs at jump-times of an unobserved Poisson process with known rate $a > 0$. In the case of an infinite horizon and discounted rewards, Donchev [2] has proved the optimality of policy (1) for all sufficiently small values of μ/a. However, this is not done by using the explicit solution of the optimality equation. In that paper (Donchev [2]), we made use of a comparison of a special functional-differential equation with delay, and two ordinary differential equations utilizing some recent results for the solution of the equation with delay.

Here, we consider the particular case of a matrix

$$\Lambda = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}.$$

Since for this matrix $\mu/a = 0$, the above result implies that policy φ from Equation (1) is optimal with respect to the β-discounted reward criterion for any $a > 0$. In this paper, we suggest a direct proof of this statement by solving explicitly the Bellman equation for the corresponding controlled process $\{x_t\}$ of posteriori probabilities. We also provide an explicit formula for the β-discounted reward corresponding to policy φ. A direct solution of the average optimality equation in the case $\mu = 0$ is given in Donchev and Yushkevich [4].

In Section 2, we characterize the process $\{x_t\}$, which is a piecewise-deterministic process and describe sufficient optimality conditions. In Section 3, we solve the equation which the value function, provided that policy φ is optimal, should satisfy. In Section 4, we show that the solution of this equation also solves the Bellman equation.

2. The Process $\{x_t\}$. Optimality Conditions

Let x_t be a posterior probability of H_1 at time t. It depends on the given probability $x_0 = x$, on the known controls u_s, $0 \le s < t$, and on the observations over the time interval $[0, t]$. It is well known that the process $\{x_t\}$ forms a sufficient statistic for the TAB problem, which is a problem with incomplete information. In the case where there is an absence of switching, hence in our problem for $a = 0$, a rigorous proof of this assertion can be found in Presman and Sonin [7]. On an intuitive level, this can be done by means of the Bayes rule. Utilizing it to re-evaluate x_{t+dt} given x_t, u_t, and observations on the interval $(t, t + dt)$, one may see the following:

(*i*) In the absence of observed particles, the process x_t satisfies the ordinary differential equation

$$x'_t = \lambda(1-2u_t)x_t(1-x_t) + a(1-2x_t); \tag{2}$$

(*ii*) With probability $\lambda u_t x_t dt$, a particle is observed at the output of the left arm and in this case, $x_{t+dt} = 1$;

(*iii*) With probability $\lambda(1-u_t)(1-x_t)dt$, a particle is observed at the output of the right arm, and in this case $x_{t+dt} = 0$; and

(*iv*) Since we get income when observing particles, the expected increment of rewards is equal to the sum of these two probabilities (for more details, see Donchev [2]).

This characterization of the process $\{x_t\}$ allows to put our problem into the framework of controlled, piecewise-deterministic processes. Following Yushkevich [9] (for the case of no impulse actions, no restrictions on the continuous actions, and stationary in time components of the model), we define a model Z which is determined as follows:

(*a*) A state space $X = [0,1]$;

(*b*) A space of admissible controls $U = [0,1]$;

(*c*) A drift coefficient

$$b(x,u) = \lambda(1-2u)x(1-x) + a(1-2x); \tag{3}$$

(*d*) A jump rate-measure $q(x,u,dy)$ on X defined by

$$q(x,u,0) = \lambda(1-u)(1-x), q(x,u,1) = \lambda ux, q(x,u,(0,1)) = 0; \tag{4}$$

(*e*) A reward rate

$$r(x,u) = \lambda ux + \lambda(1-u)(1-x), \quad x \in X, \quad u \in U. \tag{5}$$

A policy π is a random process $\{u_t\}$, $t \geq 0$, which is progressively measurable with respect to the past observations, and such that some usual conditions for solvability of the equation $dy_t = b(y_t, u_t)dt$ are satisfied. For more details, we refer the read to Yushkevich [9]. Here we discuss only the concept of a Markov policy, which we need in the sequel. A stationary policy is a Borel mapping ψ from X to U such that the equation $dy = b(y, \psi(y))dt$ has a unique solution y_t, $t \geq 0$, for every initial condition $y_0 = x \in X$, with values $y_t \in X$. Let Π and Φ denote, respectively, the sets of all policies and all Markov policies in Z. To every policy π and *a priori* probability x, there corresponds a measure P^π_x on the space of all sample paths of the process $\{x_t\}$. Denoting the corresponding expectation by E^π_x, we consider a problem with a criterion

$$v^\pi(x) = E^\pi_x \int_0^\infty e^{-\beta t} r(x_t, u_t)dt, \; x \in X, \quad \pi = \{u_t\} \in \Pi, \tag{6}$$

with $r(x,u)$ being defined by Equation (5). In Equation (6), $\beta > 0$ is a discounting factor which ensures finiteness of the expectation in its right-hand side. The problem is to maximize the expected total β-discounted reward over $\pi \in \Pi$. Let us denote by

$$v(x) = \sup_{\pi \in \Pi} v^\pi(x) \tag{7}$$

the value function in this problem

To formulate sufficient optimality conditions, let $C^1[X]$ be the space of all continuously differentiable functions defined on $X = [0, 1]$. Utilizing Equations (2) through (5), and taking into account the definitions of the model, we define the operators

$$L_1 f(x) = [-\lambda x(1-x) + a(1-2x)]f'(x) + \lambda x[f(1) - f(x) + 1], \tag{8}$$

$$L_2 f(x) = [\lambda x(1-x) + a(1-2x)]f'(x) + \lambda(1-x)[f(0) - f(x) + 1], \tag{9}$$

acting on functions $f \in C^1[X]$. Let us set

$$Lf(x,u) = uL_1 f(x) + (1-u)L_2 f(x), \quad u \in U, \ x \in X. \tag{10}$$

Proposition 2.1: *If a Markov policy $\varphi(x)$ and a function $V(x) \in C^1[X]$ satisfy the relations*

$$LV(x, \varphi(x)) = \beta V(x) = \max_{u \in U} LV(x,u), \quad x \in X, \tag{11}$$

then v is a value function and φ is an optimal policy for the problem presented in Equations (6) *and* (7).

This proposition follows from well known results in continuous-time dynamic programming. In the framework used here, a formal reference should be made to Theorem 7.1 in Yushkevich [9] (with some obvious simplifications due to the absence of impulses and constraints on continuous actions).

3. Solution of the Equation for the Value Function

In this section, we find a solution of the equation

$$LV(x, \varphi(x)) = \beta V(x), \tag{12}$$

which is the first relation in Equation (11), assuming that the policy φ defined by Equation (1) is optimal. In view of the symmetry of φ and elements (c), (d) and (e) of the model Z with respect to $x = 1/2$, we seek a solution of Equation (12) with a symmetric function V, so that

$$V(x) = V(1-x), \quad 0 \le x \le 1. \tag{13}$$

Obviously, such a function satisfies the relations

$$V(0) = V(1), \tag{14}$$

$$V'\left(\frac{1}{2}\right) = 0, \tag{15}$$

$$L_1 V(x) = L_2 V(1-x). \tag{16}$$

To avoid misunderstanding, we denote the restriction of $V(x)$ on the interval $1/2 \le x \le 1$ by $v(x)$. On this interval, the equation $LV(x, \varphi(x)) = \beta V(x)$ becomes

$L_1 v(x) = \beta v(x)$, or equivalently

$$F(x)v'(x) - (\lambda x + \beta)v(x) = -\lambda x[1 + v(1)], \tag{17}$$

where

$$F(x) = -\lambda x(1-x) + a(1-2x) = -\lambda(x_1 - x)(x - x_2), \tag{18}$$

$$x_{1,2} = \frac{1}{2} + \frac{a}{\lambda} \pm \sqrt{\frac{1}{4} + \left(\frac{a}{\lambda}\right)^2}. \tag{19}$$

Equation (17) is a linear, first-order ordinary differential equation for the function v. Its general solution depends on $v(1)$, the unknown value $v(x)$ for $x = 1$, as well as on the integration constant C. To determine these two quantities, we utilize Equation (15) and the identity $v(1) = v(1)$. By applying to Equation (16) the standard solution formula, we get that its general solution is

$$v(x) = e^{-I(x)}\left[C + (1 + v(1))\int e^{I(x)}\frac{x}{(x_1 - x)(x - x_2)}dx\right], \frac{1}{2} \le x \le 1, \tag{20}$$

where

$$I(x) = \int \frac{x + \beta/\lambda}{(x_1 - x)(x - x_2)}dx = c\ \ln(x - x_2) - (1 + c)\ln(x_1 - x) \tag{21}$$

and

$$c = \frac{x_2 + \beta/\lambda}{x_1 - x_2}. \tag{22}$$

Let us mention the following useful relations

$$cx_1 - (1 + c)x_2 = \frac{\beta}{\lambda}, \tag{23}$$

$$c(1 + c) = \frac{R}{(x_1 - x_2)^2}, \tag{24}$$

and

$$R = \frac{a}{\lambda} + \frac{\beta}{\lambda} + 2\frac{a\beta}{\lambda^2} + \left(\frac{\beta}{\lambda}\right)^2, \tag{25}$$

which follow from Equations (19) and (22).

To compute the integral in the right-hand side of Equation (20), we use a decomposition of the fraction under the integral as a sum of elementary fractions:

$$\frac{x}{(x_1 - x)(x - x_2)} = \frac{1}{x_1 - x_2}\left(\frac{x_1}{x_1 - x} + \frac{x_2}{x - x_2}\right). \tag{26}$$

By Equation (21), we have

$$e^{I(x)} = (x - x_2)^c (x_1 - x)^{-1-c}. \tag{27}$$

Utilizing Equations (26) and (27), we get

$$\int e^{I(x)}\frac{x}{(x_1 - x)(x - x_2)}dx = \frac{x_1}{x_1 - x_2}\int (x - x_2)^c (x_1 - x)^{-c-2}dx \tag{28}$$

$$+\frac{x_2}{x_1-x_2}\int(x-x_2)^{c-1}(x_1-x)^{-c-1}dx$$

$$:=I_1+I_2.$$

Both integrals I_1 and I_2 can be solved as follows

$$I_1=\frac{x_1}{x_1-x_2}\int\left(\frac{x-x_2}{x_1-x}\right)^c\cdot\frac{dx}{(x_1-x)^2}$$

$$=\frac{x_1}{(x_1-x_2)^2}\int\left(\frac{x-x_2}{x_1-x}\right)^c d\frac{x-x_2}{x_1-x}=\frac{x_1}{(c+1)(x_1-x_2)^2}\left(\frac{x-x_2}{x_1-x}\right)^{c+1},$$

$$I_2=\frac{x_2}{x_1-x_2}\int\left(\frac{x_1-x}{x-x_2}\right)^{-c-1}\cdot\frac{dx}{(x-x_2)^2}$$

$$=-\frac{x_2}{(x_1-x_2)^2}\int\left(\frac{x_1-x}{x-x_2}\right)^{-c-1}d\frac{x_1-x}{x-x_2}=\frac{x_2}{c(x_1-x_2)^2}\left(\frac{x-x_2}{x_1-x}\right)^c.$$

We add the expressions for I_1 and I_2 from the last two formulas and multiply I_1+I_2 by $e^{-I(x)}=(x_1-x)^{c+1}(x-x_2)^{-c}$. After some elementary algebra in which Equations (23) and (24) are used, we get that the second term in the right-hand side of Equation (20) is equal to

$$\frac{1+v(1)}{R}\left(\frac{a}{\lambda}+\frac{\beta}{\lambda}x\right).$$

Thus we obtain the following formula for $v(x)$:

$$v(x)=C(x_1-x)^{c+1}(x-x_2)^{-c}+\frac{1+v(1)}{R}\left(\frac{a}{\lambda}+\frac{\beta}{\lambda}x\right). \tag{29}$$

Next, we set $x=1$ in Equation (29) in order to express the constant C by means of $v(1)$. Easy calculations show that

$$C=\frac{R-\frac{a}{\lambda}-\frac{\beta}{\lambda}}{R}(x_1-1)^{-c-1}(1-x_2)^cv(1)-\frac{\frac{a}{\lambda}+\frac{\beta}{\lambda}}{R}(x_1-1)^{-c-1}(1-x_2)^c. \tag{30}$$

Substituting Equation (30) in Equation (29), we get

$$Rv(x)=[1+v(1)]\left[\frac{a}{\lambda}+\frac{\beta}{\lambda}x-\left(\frac{a}{\lambda}+\frac{\beta}{\lambda}\right)X(x)\right]+Rv(1)X(x), \tag{31}$$

where

$$X(x)=\left(\frac{x_1-x}{x_1-1}\right)^{c+1}\left(\frac{x-x_2}{1-x_2}\right)^{-c},\quad x\in\left[\frac{1}{2},1\right]. \tag{32}$$

The function $X(x)$ plays an important role in the sequel. We summarize some of its properties in the next lemma.

Lemma 3.1: *The function $X(x)$ satisfies relations*

$$X(1)=1, \tag{33}$$

$$X\left(\frac{1}{2}\right) = \sqrt{\frac{\lambda}{4a}}\left(\frac{x_1}{x_2}\right)^{c+1/2}, \tag{34}$$

$$X'(x) = \frac{x + \frac{\beta}{\lambda}}{-x(1-x) + \frac{a}{\lambda}(1-2x)}X(x), \tag{35}$$

$$X''(x) = \frac{R}{\left[-x(1-x) + \frac{a}{\lambda}(1-2x)\right]^2}X(x). \tag{36}$$

The first of these formulas is trivial; the third follows from the fact that $X(x)$ solves the homogeneous part of Equation (17). To check Equation (34), we represent the powers in Equation (32) as $c+1 = (c+\frac{1}{2}) + \frac{1}{2}$, $-c = -(c+\frac{1}{2}) + \frac{1}{2}$, and use identities

$$\left(x_1 - \frac{1}{2}\right)(1-x_2) = \frac{x_1}{2},\ \left(\frac{1}{2} - x_2\right)(x_1 - 1) = \frac{x_2}{2},$$
$$\left(\frac{1}{2} - x_2\right)\left(x_1 = \frac{1}{2}\right) = \frac{1}{4},\ (1-x_2)(x_1-1) = \frac{a}{\lambda}.$$

Finally, we get Equation (36) by a direct computation.

Now, let us return to the function $v(x)$. It remains to determine only $v(1)$, the value of $v(x)$ at $x = 1$. Making use of Equations (15) and (34), we obtain an easy equation for $v(1)$. Solving it, we get

$$Tv(1) = \left(\frac{a}{\lambda} + \frac{\beta}{\lambda}\right)\left(1 + 2\frac{\beta}{\lambda}\right)\sqrt{\frac{\lambda}{a}}\left(\frac{x_1}{x_2}\right)^{c+1/2} + \frac{\beta}{\lambda}, \tag{37}$$

$$T[1 + v(1)] = R\left(1 + 2\frac{\beta}{\lambda}\right)\sqrt{\frac{\lambda}{a}}\left(\frac{x_1}{x_2}\right)^{c+1/2}, \tag{38}$$

$$T = \frac{\beta}{\lambda}\left(2\frac{a}{\lambda} + \frac{\beta}{\lambda}\right)\left(1 + 2\frac{\beta}{\lambda}\right)\sqrt{\frac{\lambda}{a}}\left(\frac{x_1}{x_2}\right)^{c+1/2} - \frac{\beta}{\lambda}, \tag{39}$$

with R being given by Equation (25).

Substituting Equations (37) and (38) into Equation (31), we obtain an explicit formula for $v(x)$:

$$Tv(x) = \left(1 + 2\frac{\beta}{\lambda}\right)\sqrt{\frac{\lambda}{a}}\left(\frac{x_1}{x_2}\right)^{c+1/2}\left(\frac{a}{\lambda} + \frac{\beta}{\lambda}x\right) + \frac{\beta}{\lambda}X(x),\ \frac{1}{2} \le x \le 1. \tag{40}$$

Finally, we set

$$V(x) = \begin{cases} v(x), & \text{if } \frac{1}{2} \le x \le 1 \\ v(1-x), & \text{if } 0 \le x < \frac{1}{2}. \end{cases} \tag{41}$$

Now, we are in a position to formulate our main result. Its proof will be given in Section 4.

Theorem 3.1: *The policy φ and the function V defined by Equations* (1) *and* (41) *are, respectively, an optimal policy and the value function for the problem presented in Equations* (6) *and* (7).

4. Main Results. (Proof of Theorem 3.1)

By construction, the function $V(x)$ satisfies the first equation in Equation (11) on $\left[\frac{1}{2},1\right]$. In view of Equation (16), it satisfies the same equation on the whole segment $[0,1]$. Thus, it remains to prove only the second relation in Equation (11). On the other hand, since the operator L is linear with respect to u, it suffices to show only that

$$L_2V(x) \leq \beta V(x), \text{ if } \tfrac{1}{2} \leq x \leq 1,$$

and

$$L_1V(x) \leq \beta V(x), \text{ if } 0 \leq x \leq \tfrac{1}{2}.$$

By a symmetry, we need to prove only the first of these inequalities, which is equivalent to

$$L_1v(x) + L_2v(x) \leq 2\beta v(x), \quad \tfrac{1}{2} \leq x \leq 1. \tag{42}$$

According to Equations (8) and (9), the inequality in Equation (42) reduces to

$$2a(1-2x)v'(x) + \lambda[1+v(1)] \leq (2\beta+\lambda)v(x), \tag{43}$$

where we have used the equality $V(0) = V(1) = v(1)$, which follows from Equation (14). The same equality, along with Equation (15), implies the relation

$$LV\left(\tfrac{1}{2},\tfrac{1}{2}\right) = \lambda\left[v(1) + 1 - v\left(\tfrac{1}{2}\right)\right] = 2\beta v\left(\tfrac{1}{2}\right). \tag{44}$$

Comparing Equations (43) and (44), we see that inequality (43) is equivalent to

$$2a(1-2x)v'(x) \leq (\lambda+2\beta)\left[v(x) - v\left(\tfrac{1}{2}\right)\right]. \tag{45}$$

To prove inequality (45), we need the next lemma.

Lemma 4.1: *The following inequalities hold:*

$$v(1) > v\left(\tfrac{1}{2}\right) > 0. \tag{46}$$

The probabilistic meaning of these inequalities is quite apparent. The first one means that, if at time $t=0$, we know exactly that the left arm is the better arm, then the expected total number of observed particles is greater than the corresponding number in the case of a full uncertainty about arms. According to the second inequality in Equation (46), even in the worst case, we get a strictly positive income.

We only sketch the details of the proof which is quite technical. First, we show that $T > 0$, where T is given by Equation (39). To do this, we cancel the factor $\frac{\beta}{\lambda}$ in the expression for T and get the inequality

$$\left(2\frac{a}{\lambda}+\frac{\beta}{\lambda}\right)\left(1+2\frac{\beta}{\lambda}\right)\sqrt{\frac{\lambda}{a}}\left(\frac{x_1}{x_2}\right)^{c+1/2} > 1,$$

which, in view of Equations (19) and (22), is to be proved only for $\beta = 0$. In this case, the last inequality reduces to

$$\left(\frac{\frac{1}{2}+z+\sqrt{\frac{1}{4}+z^2}}{\frac{1}{2}+z-\sqrt{\frac{1}{4}+z^2}}\right)^{\frac{1/2+z}{\sqrt{1+4z^2}}} > \frac{1}{2\sqrt{z}}, \tag{47}$$

with $z = \frac{a}{\lambda}$. Since the fraction and the power in the left-hand side of the inequality in Equation (47) are greater than 1 and $\frac{1}{2}$, respectively, it would hold provided that

$$\frac{\frac{1}{2}+z+\sqrt{\frac{1}{4}+z^2}}{\frac{1}{2}+z-\sqrt{\frac{1}{4}+z^2}} = \frac{\left(\frac{1}{2}+z+\sqrt{\frac{1}{4}+z^2}\right)^2}{z} > \frac{1}{4z}.$$

Obviously, this inequality holds for all positive values of z. Now, both inequalities in Equation (46) follow easily from Equations (34), (37) and (40).

Lemma 4.1 allows us to complete the proof of Theorem 3.1, making use of the same idea as in Donchev and Yushkevich [4]. Returning to Equation (45), we notice that it certainly would hold provided that

$$v'(x) \geq 0, \frac{1}{2} \leq x \leq 1,$$

and

$$v(x) - v\left(\frac{1}{2}\right) \geq 0, \ \frac{1}{2} \leq x \leq 1.$$

The second of these two inequalities is a corollary of the first of them, hence it remains to prove only that $v'(x) \geq 0$ for $\frac{1}{2} \leq x \leq 1$. Assume the contrary. By Equation (46), we have that $v'(x)$ is positive somewhere in the interval $\left(\frac{1}{2}, 1\right]$. Since $v'(x)$ is continuous in $\left[\frac{1}{2}, 1\right]$, the assumption implies that $v'(\xi) = 0$ at some point $\xi \neq \frac{1}{2}$. Then, in view of Equation (15), we deduce that $v''(x)$ must attain the value 0 at some point of the interval $\left(\frac{1}{2}, 1\right]$, which obviously contradicts Equations (32), (36) and (40). This completes the proof of Theorem 3.1.

Acknowledgements

My thanks are addressed to both the anonymous referees and the associate editor for their very careful reading of the manuscript and constructive suggestions which improved both the content and the style of the present article.

References

[1] DeGroot, M., *Optimal Statistical Decisions*, McGraw-Hill, New York 1970.

[2] Donchev, D.S., On the two-armed bandit problem with non-observed Poissonian switching arms, *Math. Methods Oper. Res.* **47**:3 (1988), 401-422.

[3] Donchev, D.S., A system of functional-differential equations associated with the optimal detection problem for jump-times of a Poisson process, *J. of Appl. Math and Stoch. Anal.* **11**:2 (1998), 179-192.

[4] Donchev, D.S. and Yushkevich, A.A., Average optimality in a Poissonian bandit with switching arms, *Math. Methods Oper. Res.* **45**:2 (1997), 265-280.

[5] Dynkin, E.B. and Yushkevich, A.A., *Controlled Markov Processes*, Springer-Verlag, Berlin 1979.

[6] Feldman, D., Contributions on the "two-armed bandit" problem, *Ann. Math. Stat.* **33** (1962), 847-856.

[7] Presman, E.L. and Sonin, I.M., *Sequential Control with Incomplete Data: Bayesian Approach*, Academic Press, New York 1990.

[8] Sonin, I.M., A model of resource distribution with incomplete information, In: *Modeling Scientific-Technological Progress and Control of Economic Processes under Incomplete Information*, CEMI, USSR Academy of Sciences Press, Moscow (1976), 161-201 (in Russian).

[9] Yushkevich, A.A., Verification theorems for Markov decision processes with a controlled deterministic drift and gradual and impulsive controls, *Theory Probab. Appl.* **34**:3 (1989), 474-496.