Ann Oper Res (2006) 141: 193–210
DOI 10.1007/s10479-006-5299-3

Optimal multi-threshold control by the BMAP/SM/1 retrial system

Che Soong Kim · Valentina Klimenok · Alexander Birukov · Alexander Dudin

Abstract A single server retrial system having several operation modes is considered. The modes are distinguished by the transition rate of the batch Markovian arrival process (BMAP), kernel of the semi-Markovian (SM) service process and the intensity of retrials. Stationary state distribution is calculated under the fixed value of the multi-threshold control strategy. Dependence of the cost criterion, which includes holding and operation cost, on the thresholds is derived. Numerical results illustrating the work of the computer procedure for calculation of the optimal values of thresholds are presented.

Keywords Batch Markovian arrival process · Controlled operation modes · Cost criterion · Optimal control

1. Introduction

Queueing theory plays a significant role in improving system operation in practice by choosing appropriately the system configuration and parameters. While a number of papers address this aspect using static optimization, in reality one has to rely on dynamic optimization. This is a very challenging but interesting problem and this paper addresses this aspect by looking at a specific retrial queueing model with batch arrivals and services that are offered in various

C. S. Kim
Department of Industrial Engineering, Sangji University, Wonju, Kangwon, Korea 220–702
e-mail: dowoo@sangji.ac.kr

V. Klimenok · A. Birukov
Department of Applied Mathematics and Computer Science, Belarusian State University, 4 F.Skorina Ave, 220050, Minsk, Belarus
e-mail: klimenok@bsu.by; al_birukov@mail333.com

A. Dudin
Department of Applied Mathematics and Computer Science. Belarusian State University, Minsk, Belarus
e-mail: dudin@bsu.by

modes of operation. We assume that the arrival process, the retrial process and the service process depend on the mode of operation.

Note, that often the state of the orbit is not observable in real-life systems and there is no chance to make any control basing on the knowledge of the system state. However, in some computer systems, e.g., local area communication networks, Internet access points, etc., the attempts of the users are registered by the system and information about the state of the orbit (virtual place where the repeated customers wait for another attempt) is available. Our queueing model is addressed to such systems.

We assume that the system operation mode can be changed at any service completion epoch depending on the number of customers in the orbit according to the so-called multi-threshold strategy. This strategy has proven its optimality in the class of all Markovian strategies in some queueing systems, see, e.g., paper by Tijms (1976). In the present model, we do not try to prove the optimality but investigate a way for the selection of an optimal set of thresholds. This way, we find the conditions for the system stability under fixed values of the thresholds and fixed distributions characterizing the arrival, service and retrial processes. Then we calculate the stationary state distribution of the system and performance characteristics. This reduces the problem of finding the optimal multi-threshold strategy just to minimization of a known function of several integer variables.

It looks like there are no recent surveys of controlled queues in literature. For some list of papers devoted to the problem of optimal control by queues with the *BMAP* input see, e.g., (Dudin and Chakravarthy, 2002). Note that the controlled retrial queues are less investigated compared to the controlled queues with a buffer. To the best of our knowledge, the set of publications devoted to the controlled retrial queues having such a general input process as the *BMAP* consists of only the paper (Choi, Chung, Dudin, 2001). The correspondence between that paper and the present one is the following. In (Choi, Chung, Dudin, 2001), a bit more general hysteresis strategy of control is considered, but only the case of two available modes is dealt with. In addition, the total retrial rates are assumed to be independent of the number of customers in the orbit under the fixed mode of operation. Here we consider the case of R, $R \geq 2$, available operation modes switched according to the multi-threshold strategy. The dependence of the total retrial rate on the number of customers is allowed.

2. The model

We consider a single-server retrial system with a *BMAP* input flow and *SM* service process which can operate in R, $R \geq 2$, different modes.

The standard *BMAP* is described (see, e.g., paper by Lucantoni (1991) and survey Chakravarthy (2001)) as follows. Arrivals are governed by the so-called directing process which is the irreducible continuous time Markov chain ν_t, $t \geq 0$, with the finite state space $\{0, 1, \ldots, W\}$. The *BMAP* is described by the directing process ν_t, $t \geq 0$, and the matrix transition generating function $D(z) = \sum_{k=0}^{\infty} D_k z^k$, $|z| < 1$. The matrices D_k characterize transitions of the chain ν_t, $t \geq 0$, which are accompanied by generating a batch of k customers, $k \geq 0$. The matrix $D(1)$ is the infinitesimal generator of the process ν_t, $t \geq 0$. So, the stationary probability row vector $\vec{\theta}$ of this process is the unique solution to the following system of linear algebraic equations:

$$\vec{\theta} D(1) = \vec{0}, \vec{\theta}\mathbf{e} = 1.$$

Here $\vec{0}$ is zero row vector and e is a column vector of 1's. The dimension of vectors here and in the sequel is defined from the context.

The average intensity λ of arrivals (fundamental rate) in the *BMAP* is calculated as $\lambda = \vec{\theta} D'(1)\mathbf{e}$. The average intensity $\lambda^{(b)}$ of group's arrival is defined as $\lambda^{(b)} = \vec{\theta}(-D_0)\mathbf{e}$. The variation coefficient c_{var} of intervals between successive group arrivals is defined by the formula

$$c_{var}^2 = 2\lambda^{(b)}\vec{\theta}(-D_0)^{-1}\mathbf{e} - 1.$$

The correlation coefficient C_{cor} of successive intervals between group's arrival is given by

$$C_{cor} = (\lambda^{(b)}\vec{\theta}(-D_0)^{-1}(D(1) - D_0)(-D_0)^{-1}\mathbf{e} - 1)/c_{var}^2.$$

This coefficient can differ from zero. So, the *BMAP* is a popular descriptor of the correlated information flows in modern telecommunication networks. The *BMAP* includes many previously studied input flows, the *PH* (phase type) input flow in particular.

In the present queueing model we assume that the *BMAP* has the same directing process $\nu_t, t \geq 0$, for all operation modes. But under the rth operation mode, the transitions of the process $\nu_t, t \geq 0$ and the batches generation are defined by the matrix generating function $D^{(r)}(z) = \sum_{k=0}^{\infty} D_k^{(r)} z^k$, $|z| < 1$. The fundamental rate λ_r of the input under the rth operation mode is calculated as $\lambda_r = \vec{\theta}_r (D^{(r)}(z))'|_{z=1}\mathbf{e}$ where the vector $\vec{\theta}_r$ is the unique solution to the system $\vec{\theta}_r D^{(r)}(1) = \vec{0}, \vec{\theta}_r \mathbf{e} = 1, r = \overline{1, R}$.

A standard semi-Markovian service process is described in the following way. Semi-Markovian process $m_t, t \geq 0$, which is characterized by the state space $\{1, \ldots, M\}$ and the semi-Markovian kernel $B(t) = \|B_{m,m'}(t)\|_{m,m'=\overline{1,M}}$, is fixed. The successive service times are defined as the successive sojourn times of the process $m_t, t \geq 0$ in its states. The transition probability matrix $B(\infty)$ of the embedded Markov chain is assumed to be irreducible. The average service time is calculated as $b_1 = \vec{\delta} \int_0^{\infty} t dB(t)\mathbf{e}$ where $\vec{\delta}$ is the unique solution to the system $\vec{\delta} B(\infty) = \vec{\delta} \cdot \vec{\delta}\mathbf{e} = 1$. Semi-Markovian service process allows to take into account correlation of successive service times. The recurrent service is just a partial case of the *SM* service process when $M = 1$.

In the present controlled model, the service time is defined in the following way. At the given service completion epoch, the operation mode, say r, is defined according to the fixed control strategy. After this epoch, sojourn time and the transition of the process m_t are determined according to the kernel $B_r(t), r = \overline{1, R}$. The average service time in the rth mode is calculated as $b_1^{(r)} = \vec{\delta}_r \int_0^{\infty} t dB_r(t)\mathbf{e}$, where $\vec{\delta}_r$ is the unique solution to the system $\vec{\delta}_r B_r(\infty) = \vec{\delta}_r, \vec{\delta}_r \mathbf{e} = 1, r = \overline{1, R}$.

If the batch of arriving customers enters the system when the server is idle, then the server begins the processing of one customer immediately. The rest of the batch becomes the repeated customers. If the batch of primary customers enters the system when the server is busy then all customers of this batch become the repeated customers. The repeated customers are said to be in "orbit". These customers stay in the orbit until they get service. The attempts to get a service are made in exponentially distributed times in such a way that the total intensity of retrials from the orbit is equal to $\alpha_i^{(r)}, i \geq 1 (\alpha_0^{(r)} = 0)$ when the system operates in the rth mode and the number of customers in the orbit is equal to i. The dependence of parameters $\alpha_i^{(r)}$ on the value i can be arbitrary for $r = \overline{1, R-1}$. But, aiming to investigate the stationary behavior of the model, we have to assume that for $r = R$ the limit $\lim_{i\to\infty} \alpha_i^{(R)}$ exists. We will distinguish the cases when this limit is infinite (this case includes, e.g., the

classic strategy of retrials when $\alpha_i^{(R)} = i\alpha^{(R)}$ and the linear repeated attempts, which were introduced by Artalejo and Gomez-Corral (1997), when $\alpha_i^{(R)} = i\alpha^{(R)} + \gamma R$) and finite (it includes the case of a constant retrial rate when $\alpha_i^{(R)} = \gamma_R$, $i \geq 1$).

The mode of operation can be changed at any service completion epoch depending on the current number of customers in the orbit. The aim of the control by operation modes is minimization of a cost criterion. Following to existing traditions in controlled queues, we fix the following form of the expected total cost criterion:

$$E = a\tau^{-1}L + \sum_{r=1}^{R} c_r P_r \ , \tag{1}$$

which is the average charge per unit time. Here L is the average number of customers in the orbit at the service completion epochs, τ is the average inter- departure time, P_r is the average fraction of time when the rth mode is exploited, $r = \overline{1, R}$, a is a charge, which is paid for one customer staying in the orbit at a service completion epoch (holding cost), c_r is the cost of the rth mode utilization (service cost) per unit time, $r = \overline{1, R}$. Physically, the cost c_r can include the paying for faster service, lower input rate and retrial process control.

Our aim is to find the strategy of control which minimizes the cost criterion (1).

We do not impose any restrictions on the values of the traffic intensities $P_r = \lambda_r b_1^{(r)}$ and costs C_r, $r = \overline{1, R}$ except the stability condition which is presented in the next section. In case of systems with buffers, a popular assumption is that the operation modes are enumerated in such a way as

$$\rho_1 \geq \rho_2 \geq \cdots \geq \rho_R, \quad c_1 \leq c_2 \leq \cdots \leq c_R \ , \tag{2}$$

i.e., the modes with smaller indices are characterized by larger traffic and lower cost of using. Sometimes (see, e.g., Tijms 1976) it is possible to prove that the optimal Markovian control strategy belongs to the class of multi-threshold strategies when condition (2) holds true. Such a kind of strategies is intuitively reasonable and suitable for technical realization in the real life systems.

In case of our retrial model where the arrangement of retrial intensities may be even more important than arrangement of the traffic intensities and costs, we do not try to establish conditions for optimality of the multi-threshold strategies in the class of the Markovian strategies. But we restrict ourselves in advance and solve the problem of finding the optimal (or suboptimal) strategy in class of the multi-threshold strategies. These strategies are defined in the following way.

Let some integers $j_1, \ldots, j_{R-1}$ such as $-1 = j_0 < j_1 \leq j_2 \leq \cdots \leq j_{R-1} < J_R = +\infty$ be fixed. If the number i of customers in the orbit at a given service completion epoch belongs to the interval $j_{r-1} < i \leq j_r$ then the system will operate in the rth mode, $r = \overline{1, R}$.

So, we intend to develop a procedure for calculation of the optimal set $(j_1^*, \ldots, j_{R-1}^*)$ of thresholds which provides a minimal value to criterion (1).

To this end, we exploit so-called direct approach that consists of the following. We fix a set of thresholds $(j_1, \ldots, j_{R-1})$ and calculate the stationary state distribution of the system at the service completion and at arbitrary epochs. Then, we calculate the value of the cost criterion. As a result we have a numerical procedure to calculate the cost criterion for any fixed value of the thresholds. The optimal set of the thresholds is calculated just by enumeration. In principle, some more effective procedures can be elaborated, but the enumeration in a reasonable region works sufficiently fast.

3. The embedded Markov chain

Let some set of the thresholds $(j_1, \ldots, j_{R-1})$, $j_{R-1} < +\infty$ be fixed. Let t_n be the nth service completion epoch. Consider the 3-dimensional process $\{i_n, \nu_n, m_n\}$, $n \geq 1$, where i_n is the number of customers in the orbit at the epoch $t_n + 0$, $i_n \geq 0$; ν_n is the state of the directing process ν_t, $t \geq 0$ at the epoch t_n, $\nu_n = \overline{0, W}$; m_n is the state of the process m_t, $t \geq 0$ at the epoch $t_n + 0$, $m_n = \overline{1, M}$.

It is easy-to see that under the fixed set of the thresholds the process $\xi_n = \{i_n, \nu_n, m_n\}$, $n \geq 1$ is the Markov chain.

Let $P_{i,l}$, $i, l \geq 0$, be the square matrices formed from the one-step transition probabilities

$$P\{i_{n+1} = l, \nu_{n+1} = \nu', m_{n+1} = m' | i_n = i, \nu_n = \nu, m_n = m\}$$

of the Markov chain ξ_n, $n \geq 1$ listed in the lexicographic order.

Lemma 3.1. *Non-zero matrices $P_{i,l}$ are defined as follows:*

$$P_{i,l} = \alpha_i^{(r)}(\alpha^{(r)} I - \tilde{D}_0^{(r)})^{-1} \hat{Y}_{l-i+1}^{(r)} + (\alpha_i^{(r)} I - \tilde{D}_0^{(r)})^{-1} \sum_{k=1}^{l-i+1} \tilde{D}_k^{(r)} \hat{Y}_{l-i+1-k}^{(r)},$$

$$j_{r-1} < i \leq j_r, l \geq \max\{i-1, 0\}, r = \overline{1, R}\,, \tag{3}$$

where $\tilde{D}_k^{(r)} = D_k^{(r)} \otimes I_M$, $k \geq 0$, the matrices $\hat{Y}_n^{(r)}$, $n \geq 0$, $r = \overline{1, R}$ are defined as the coefficients in the following matrix expansion:

$$\sum_{n=0}^{\infty} \hat{Y}_n^{(r)} z^n = \hat{\beta}_r(z) = \int_0^{\infty} e^{D^{(r)}(z)t} \otimes dB_r(t), r = \overline{1, R}\,. \tag{4}$$

Here and in the sequel I is the identity matrix (possible low index indicates the dimension of this matrix in case it is not clear from context) and $\otimes$ is the symbol of Kronecker product of matrices.

The proof of the lemma consists of the analysis of the chain ξ_n, $n \geq 1$ transitions accounting the probabilistic meaning of involved matrices. The matrices $\tilde{Y}_l^{(r)}$ define the probability of l, $l \geq 0$, customers arrival and corresponding transition of the components $\{\nu_n, m_n\}$ during one customer service in the rth mode. The matrix $\alpha_i^{(r)}(\alpha_i^{(r)} I - \tilde{D}_0^{(r)})^{-1}$ defines the transition probabilities of the components $\{\nu_n, m_n\}$ during the idle period since the service completion epoch until the successful retrial when the system operates in the rth mode. The matrix $(\alpha_i^{(r)} I - \tilde{D}_0^{(r)})^{-1} \tilde{D}_k)$ defines the corresponding probabilities in case where the idle period is finished by arrival of a primary batch having size k, $k \geq 1$.

In the sequel, we exploit the notion of the multi-dimensional asymptotically quasi-toeplitz Markov chain which was introduced in (Dudin and Klimenok, 2000). These Markov chains are defined as follows.

Definition 3.1. Discrete time Markov chain $\eta_n, n \geq 1$, is called as the multi-dimensional asymptotically quasi-toeplitz Markov chain (AQTMC) if its nonzero

one-step transition probability matrices $P_{i,l}, l \geq max\{0, i-1\}, i \geq 0$, can be represented in the form:

$$P_{i,l} = \sum_{j=1}^{J} Q_j^{(i)} Y_{l-i+1}^j \ , \tag{5}$$

where the matrices $Q_j^{(i)}$ satisfy the conditions:

$$\lim_{t\to\infty} Q_j^{(i)} = Q_j < +\infty, j = \overline{1, J} \ , \tag{6}$$

the matrices $\sum_{j=1}^{J} Q_j$ and $\sum_{l=0}^{\infty} Y_l^{(j)}$, $j = \overline{1, J}$ are stochastic.

In (Dudin and Klimenok, 2000), sufficient conditions for the AQTMC stationary distribution existence were presented and the algorithm for calculating the stationary distribution was elaborated. A more stable algorithm was presented later in (Breuer, Dudin and Klimenok, 2002).

To have an opportunity to exploit the known results for the AQTMC we should establish the correspondence between the considered Markov chain $\xi_n, n \geq 1$ and the class of the AQTMC.

Lemma 3.2. 3-dimensional Markov chain $\xi_n, n \geq 1$, defined by the blocks (3) of the one-step transition probability matrix belongs to the class of the AQTMC.

To prove lemma, we just match denotations in definition (5), (6) and formula (3). Thus, it is easy to check that we get (3) from (5) if we put:

$$J = 2R,$$

$$Q_{2m-1}^{(i)} = \begin{cases} \{\alpha_i^{(m)}(\alpha_i^{(m)} I - \tilde{D}_0^{(m)})^{-1}, \ i \in (j_{m-1,j_m}], \\ 0, \qquad\qquad \text{otherwise}, \end{cases}$$

$$Q_{2m}^{(i)} = \begin{cases} \{(\alpha_i^{(m)} I - \tilde{D}_0^{(m)})^{-1}, (-\tilde{D}_0^{(m)}), \ i \in (j_{m-1,j_m}], \\ 0, \qquad\qquad \text{otherwise}, \end{cases}$$

$$Y_l^{(2m-1)} = \hat{Y}_l^{(m)}, l \geq 0 \ ,$$

$$Y_l^{(2m)} = (-\tilde{D}_0^{(m)})^{-1} \sum_{k=1}^{l} \tilde{D}_k^{(m)} \hat{Y}_{l-k}^{(m)}, l \geq 0, m = \overline{1, R} \ .$$

So, we can use results (Breuer, Dudin and Klimenok, 2002; Dudin and Klimenok, 2000) to deal with the chain $\xi_n, n \geq 1$.

As follows from (Dudin and Klimenok, 2000), the sufficient condition for the stationary distribution of the AQTMC $\eta_n, n \geq 1$, existence is expressed in terms of the transition probabilities of the corresponding limiting Markov chain.

Definition 3.2. Multi-dimensional Markov chain $\tilde{\eta}_n, n \geq 1$, is said to be limiting with respect to the AQTMC $\eta_n, n \geq 1$, if its nonzero transition probability blocks $\tilde{P}_{i,l}$ are defined as follows:

$$\tilde{P}_{i,l} = \sum_{j=0}^{J} Q_j Y_{l-i+1}^{(j)}, l \geq i-1, i \geq 1. \tag{7}$$

Denote $\Psi(z) = \sum_{l=i-1}^{\infty} \tilde{P}_{i,l} z^{l-i+1}, i \geq 1$. Because the transition blocks $\tilde{P}_{i,l}$ depend on i and l only via $l-i$, denotation $\Psi(z)$ is correct although formally some arbitrary i is involved into definition of the function $\Psi(z)$.

In (Dudin and Klimenok, 2000), it was proven that the stationary distribution of the AQTMC exists if the following inequality holds:

$$[det(zI - \Psi(z))]'|_{z=1} > 0 \, . \tag{8}$$

It was proven in (Klimenok 2000) that inequality (Dudin and Klimenok, 2000) is equivalent to the inequality

$$\vec{X}\Psi'(1)\mathbf{e} < 1 \, , \tag{9}$$

where the vector $\vec{X}$ is the unique solution to the following system of equations:

$$\vec{X}\Psi(1) = \vec{X}, \vec{X}\mathbf{e} = 1 \, . \tag{10}$$

Taking into account the explicit expressions for blocks $\tilde{P}_{i,l}$ defined by formula (7), we can calculate the matrix generating function $\Psi(z)$.

In case of the infinitely increasing retrial rate in the Rth operation mode ($\lim_{i\to\infty} \alpha_i^{(R)} = \infty$), the function $\Psi(z)$ has the form:

$$\Psi(z) = \hat{\beta}_R(z) \, . \tag{11}$$

In case of a finite limiting retrial rate in the Rth operation mode ($\lim_{i\to\infty} \alpha_i^{(R)} = \gamma_R < +\infty$), the function $\Psi(z)$ is defined by

$$\Psi(z) = (I + (\gamma_R I - \tilde{D}_0^{(R)})^{-1} \tilde{D}_0^{(R)}(z))\hat{\beta}_R(z) \, . \tag{12}$$

Taking into account (9)-(10) and the explicit form (11), (12) of the matrix generating function $\Psi(z)$, the following statement is proven.

Theorem 3.1. *The sufficient condition for existence of the AQTMC $\xi_n, n \geq 1$. stationary distribution has the form:*

$$\rho_R = \lambda_R b_1^{(R)} < 1 \tag{13}$$

in the case of infinite limiting retrial rate and the form:

$$\vec{X}[\hat{\beta}_R'(1) + (\gamma_R I - \tilde{D}_0^{(R)})^{-1} \hat{\beta}_R(1) \tilde{D}^{(R)}(z))'|_{z=1}]\mathbf{e} < 1, \tag{14}$$

in the case of the finite limiting retrial rate.

Here the vector $\vec{X}$ is the unique solution to the system

$$\vec{X}(I - \hat{\beta}_R(1) - (\gamma_R I - \tilde{D}_0^{(R)})^{-1}\hat{\beta}_R(1)\tilde{D}^{(R)}(1)) = \vec{0}, \tag{15}$$

$$\vec{X}\mathbf{e} = 1.$$

In the sequel, we assume condition (13) or (14) being fulfilled.

Introduce the stationary state probabilities

$$\pi(i, \nu, m) = \lim_{n\to\infty} P\{i_n = i, \nu_n = \nu, m_n = m\}, i \geq 0, \nu = \overline{0, W}, m = \overline{1, M},$$

and the row vectors $\vec{\pi}_i, i \geq 0$ of these probabilities listed in the lexicographic order.

Because the Markov chain $\xi_n, n \geq 1$ belongs to the class of the AQTMC, the following algorithm for calculation of the vectors $\vec{\pi}_i, i \geq 0$ follows directly from (Breuer, Dudin and Klimenok, 2002; Klimenok and Dudin, 2003).

Step 3.0. Calculate the matrix G as a minimal non-negative solution to the following matrix equation:

$$G = \Psi(G)\,. \tag{16}$$

The entries of the stochastic matrix G are the transition probabilities of the components $\{\nu_n, m_n\}$ of the Markov chain $\tilde{\eta}_n, n \geq 1$ during the time interval when the component i_n of the chain $\tilde{\eta}_n, n \leq 1$ reaches value i first time starting from the value $i + 1$. Sometimes it is referred to as Neuts' G matrix. There exists a variety of known numerical algorithms for solving equation (16) starting from the book by Neuts (1989).

Step 3.0. Calculate the matrices $G_0, \ldots, G_{i_0-1}$ using the backward recursion

$$G_i = (I - \sum_{l=i+1}^{\infty} P_{i+1,l}G_{l-1}G_{l-2}\ldots G_{i+1})^{-1}P_{i+1,i}, i = \overline{0, i_0 - 1} \tag{17}$$

with the boundary condition $G_i = G, i \geq i_0$.

The value i_0 is found empirically in such a way that the matrices G_{i_0-1} and G_{i0} practically do not distinguish in norm from each other and from the matrix G defined as solution to (16).

The inverted matrix in (17) always exists because the matrix. which is subtracted from the identity matrix I, is an irreducible sub-stochastic matrix.

Step 3.0. Calculate the matrices $\bar{P}_{i,l}$ as follows:

$$\bar{P}_{i,l} = P_{i,l} + \sum_{n=l+1}^{\infty} P_{i,n}G^{max\{0,n-\min\{i_0.l\}\}}G_{min\{i_0,n\}-1\ldots}G_l, l \geq i, i \geq 0, \tag{18}$$

where the matrices $P_{i,l}$ are defined by formulas (3).

Step 3.0. Calculate the matrices $\Phi_l, l \geq 1$ using the recursion:

$$\Phi_l = (\bar{P}_{0,l} + \sum_{i=1}^{l-1} \Phi_i \bar{P}_{i,l})(I - \bar{P}_{l,l})^{-1}, l \geq 1. \tag{19}$$

The matrix $\bar{P}_{l,l}$ is the irreducible sub-stochastic matrix, so the matrix $(I - \bar{P}_{l,l})^{-1}$ exists and is non-negative. Thus, recursion (19) includes only non-negative matrices and is numerically stable.

Step 3.0. Calculate the vector $\vec{\pi}_0$ as the unique solution to the system

$$\vec{\pi}_0(I - \bar{P}_{0,0}) = \bar{0}, \tag{20}$$

$$\vec{\pi}_0(I + \sum_{l=1}^{\infty} \Phi_l)\mathbf{e} = 1. \tag{21}$$

Step 3.0. Calculate the vectors $\vec{\pi}_l$ by

$$\vec{\pi}_l = \vec{\pi}_0 \Phi_l, l \geq 1. \tag{22}$$

Derivation of this algorithm based on the idea of the censored Markov chains (see, e.g. Kemeni, Shell and Knapp, 1966) is presented in (Klimenok and Dudin, 2003). Computer realization confirms the high stability of the described algorithm.

Remark 3.1. Rigorously saying, the presented algorithm should be treated as the approximative one. It is due to necessity to find the value i_0 on step 2 in empiric way. However, in some sense this algorithm can be considered as the exact one. The reason is that we can require to choose i_0 such as the matrix $G_{i_0-1} - G$ has the norm equal to zero with available accuracy of computer calculations. Thus the algorithm can be referred to as exact in the same extent as any algorithm requiring computer calculations.

4. The stationary state distribution of the system at arbitrary time

Denote $r_t, r_t = \overline{1, R}$, the number of the operation mode which is exploited at the moment $t, t \geq 0$. Consider the process $(i_t, \nu_t, m_t, r_t), t \geq 0$.

Introduce the stationary state distribution of this process as

$$p(i, \nu, m, r) = \lim_{t \to \infty} P\{i_t = i, \nu_t = \nu, m_t = m, r_t = r\}$$

and form the vectors $\vec{p}(i, r)$ of such probabilities, $i \geq 0, r = \overline{1, R}$. Let also $\vec{P}_r(z) = \sum_{i=j_{r-1}+1}^{\infty} \vec{p}(i, r) z^i, |z| < 1, r = \overline{1, R}$ be the partial generating functions of these vectors.

The process $(i_t, \nu_t, m_t, r_t), t \geq 0$ is a semi-regenerative one. Using the results by Cinlar (1975), the following statement can be proven.

Theorem 4.1. *The stationary state probability vectors $\bar{p}(i,r)$ are calculated as follows:*

$$\vec{p}(0,1) = \tau^{-1}\vec{\pi}_0(-\tilde{D}_0^{(1)})^{-1},$$

$$\vec{p}(i,r) = \tau^{-1}\left\{\sum_{l=j_{r-1}+1}^{j_r}\delta_{i,l}\vec{\pi}_i(\alpha_i^{(r)}I - \tilde{D}_0^{(r)})^{-1} + \sum_{l=j_{r-1}+1}^{\min\{i,j_r\}}\vec{\pi}_l\alpha_l^{(r)}\times\right.$$

$$\left.\times(\alpha_l^{(r)}I - \tilde{D}_0^{(r)})^{-1}\Gamma_{i-l}^{(r)} + \sum_{l=j_{r-1}+1}^{\min\{i,j_r\}}\vec{\pi}_l(\alpha_l^{(r)}I - \tilde{D}_0^{(r)})^{-1}\sum_{k=1}^{i-1}\tilde{D}_k^{(r)}\Gamma_{i-l-k}^{(r)}\right\},$$

$$i \geq \max\{j_{r-1},0\}+1, r = \overline{1,R} \tag{23}$$

where $\delta_{i,l}$ is the Kronecker symbol, the matrices $\Gamma_l^{(r)}, l \geq 0, r = \overline{1,R}$ are calculated as:

$$\Gamma_l^{(r)} = \int_0^\infty P^{(r)}(l,t)\otimes(I - \nabla_B^{(r)}(t))dt, l \geq 0, r = \overline{1,R},$$

the matrices $P^{(r)}(l,t)$ are defined as the coefficients of the expansion

$$\mathbf{e}^{D^{(r)}(z)t} = \sum_{l=0}^{\infty}P^{(r)}(l,t)z^l,$$

the matrix function $\nabla_B^{(r)}(t)$ is calculated as:

$$\nabla_B^{(r)}(t) = \operatorname{diag}\left\{\sum_{m'=1}^{M}(B_r(t))_{m,m'}, m = \overline{1,m}\right\},$$

diag stands for the diagonal matrix with the diagonal entries listed in the brackets, $(A)_{m,m'}$ denotes the (m,m')th entry of the matrix A, τ is the average inter-departure time and is calculated as

$$\tau = \sum_{r=1}^{R}\sum_{l=j_{r-1}+1}^{jr}\vec{\pi}_l\left[(\alpha_l^{(r)}I - \tilde{D}_0^{(r)})^{-1} + b_1^{(r)}I\right]\mathbf{e}\,. \tag{24}$$

Corollary 4.1. *The partial vector generating functions $\vec{P}_r(z)$ are calculated as follows:*

$$\vec{P}_r(z) = \tau^{-1}\sum_{i=j_{r-1}+1}^{jr}\vec{\pi}_i z^i\left\{(\alpha_i^{(r)}I - \tilde{D}_0^{(r)})^{-1} + \right.$$

$$\left. + \left[I + (\alpha_i^{(r)}I - \tilde{D}_0^{(r)})^{-1}\tilde{D}^{(r)}(z)\right]\Gamma^{(r)}(z)\right\},\ r = \overline{1,R}, \tag{25}$$

where

$$\Gamma^{(r)}(z) = \int_0^\infty e^{D^{(r)}(z)t}\otimes(I - \nabla_B^{(r)}(t))dt,\ r = \overline{1,R}.$$

5. Optimization problem

As soon as the stationary state distribution have been computed, we are able to get the dependence of the cost criterion on the thresholds.

Theorem 5.1. *Dependence of the cost criterion (1) on the thresholds* $(j_1, \ldots, j_{R-1})$ *has the following form:*

$$E(j_1, \ldots, j_{R-1}) = \left\{ a \sum_{r=1}^{R} \sum_{l=j_{r-1}+1}^{jr} i\vec{\pi}_i + \sum_{r=1}^{R} c_r \sum_{i=j_{r-1}+1}^{jr} \vec{\pi} i \left[(\alpha_i^{(r)} I - \tilde{D}_0^{(r)})^{-1} + \right. \right.$$

$$\left. \left. + b_1^{(r)} I \right] \right\} \boldsymbol{e} \left\{ \sum_{r=1}^{R} \sum_{i=j_{r-1}+1}^{jr} \vec{\pi}_i \left[(\alpha_i^{(r)} I - \tilde{D}_0^{(r)})^{-1} + b_1^{(r)} I \right] \boldsymbol{e} \right\}^{-1} . \qquad (26)$$

Proof: Calculate the characteristics $L, \tau, P_\tau, r = \overline{1, R}$, which appear in criterion (1).

The average number L of customers in the orbit at departure epochs is calculated trivially:

$$L = \sum_{i=1}^{\infty} i\vec{\pi}_i \mathbf{e} . \qquad (27)$$

The average shares P_r of using the rth operation mode are calculated as

$$P_r = \vec{P}_r(1)\mathbf{e}, r = \overline{1, R}.$$

Taking into account (25), we get

$$P_r = \tau^{-1} \sum_{i=j_{r-1}+1}^{j_r} \vec{\pi}_i \left[(\alpha_i^{(r)} I - \tilde{D}_0^{(r)})^{-1} + b_1^{(r)} I \right] \mathbf{e}, r = \overline{1, R} . \qquad (28)$$

The value of inter-departure time τ is calculated from (24). Substituting (24), (27), (28) into (1) we get (26).

The theorem is proven. □

6. Numerical examples

The algorithm for calculating the stationary distribution described above is realized as a computer program on C++ using tools of software "SIRIUS++" (see, e.g., (Dudin, Klimenok, Klimenok, 2000)). Extensive numerical experience has demonstrated the high stability of this algorithm.

The procedure for calculating the optimal set $(j_1^*, .., j_{R-1}^*)$ of the thresholds is implemented on computer as well. The search of this set is performed by means of calculation of the cost criterion (1) value for all possible combinations of the thresholds $(j_1, .., j_{R-1})$ in some predefined region $0 \le j_1 \le .. \le j_{R-1} \le J$. The value J is taken arbitrarily. If the optimal set of the thresholds in this region is such as $j_{R-1}^* = J$, then the new, larger, value of J

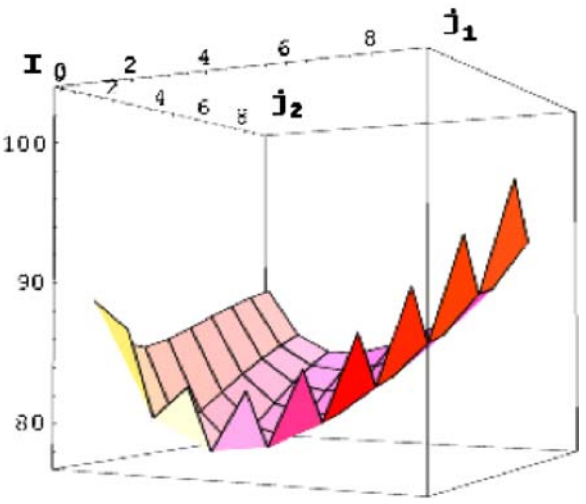

Fig. 1 Dependence of cost function value on j_1, j_2

is selected and procedure is continued to calculate the value of the cost criterion for the thresholds belonging to the extended region of a search. The extension of the region by means of increasing J is repeated until we get $j^*_{R-1} < J$. For visualization of the search procedure, the curve or surface representing the dependence of the cost criterion on the thresholds are built up in cases of two and three available modes.

We can guarantee theoretically that the procedure gives exactly the optimal set of the thresholds only under some rather restrictive assumptions about the system parameters. In general, this procedure gives at least the optimal set of the thresholds in the search region.

To illustrate the work of the algorithm, consider the following model. The system has three available operation modes.

When the system operates in the first mode, the parameters of the system are the following:

$$D_0^{(1)} = \begin{pmatrix} -1.45 & 0.45 \\ 0.6 & -4.6 \end{pmatrix}, \; D_1^{(1)} = D_2^{(1)} = \begin{pmatrix} 0.5 & 0 \\ 0 & 2 \end{pmatrix}, \; D_k^{(1)} = 0, \; k > 2,$$

$$B_1(t) = \operatorname{diag}\{B_1^{(1)}(t), B_2^{(1)}(t)\}P, \tag{29}$$

the (transition matrix P has the form: $P = \left(\begin{smallmatrix} 0.6 & 0.4 \\ 0.35 & 0.65 \end{smallmatrix}\right)$, $B_l^{(1)}(t)$ is the distribution function of a degenerate random having value 0.5 for $l = 1$ and value 0.4 for $l = 2$.

The described *BMAP* has the fundamental rate $\lambda_1 = 3.42857$, group rate $\lambda_1^{(b)} = 2.28571$, squared variation coefficient 1.68878 and coefficient of correlation 0.127455. The average service time in the first mode is equal to $b_1^{(1)} = 0.44667$

Table 1 Stationary state probabilities of embedded Markov chain for the optimal set of thresholds

$\pi_0 e$	$\pi_1 e$	$\pi_2 e$	$\pi_3 e$	$\pi_4 e$	$\pi_5 e$	$\pi_6 e$
0.06407	0.11493	0.21798	0.24859	0.16382	0.09078	0.04995
$\pi_7 e$	$\pi_8 e$	$\pi_9 e$	$\pi_{10} e$	$\pi_{11} e$	$\pi_{12} e$	$\pi_{13} e$
0.02556	0.01287	0.00622	0.00291	0.00132	0.00058	0.00025
$\pi_{14} e$	$\pi_{15} e$	$\pi_{16} e$	$\pi_{17} e$	$\pi_{18} e$	$\pi_{19} e$	$\pi_{20} e$
0.0001	0.00004	0.00002	0.000005	0.000001	0.0000003	0.00000005

Under the second operation mode, the matrices defining the *BMAP* are defined as follows:

$$D_0^{(2)} = \begin{pmatrix} -1.45 & 0.45 \\ 0.6 & -2.6 \end{pmatrix}, D_1^{(2)} = D_2^{(2)} = \begin{pmatrix} 0.5 & 0 \\ 0 & 2 \end{pmatrix}, D_k^{(2)} = 0, k > 2$$

The semi-Markovian kernel has a structure analogous to (29) with the matrix P of a form $P = \begin{pmatrix} 0.3 & 0.7 \\ 0.75 & 0.25 \end{pmatrix}$, the distribution function $B_1^{(2)}(t)$ corresponding to the exponential distribution with the parameter 4 and $B_2^{(2)}(t) = B_2^{(1)}(t)$.

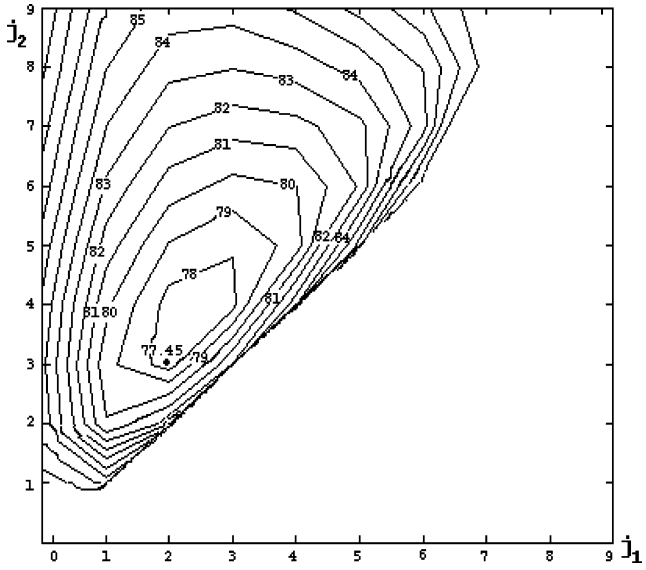

Fig. 2 Level lines of the cost criterion value

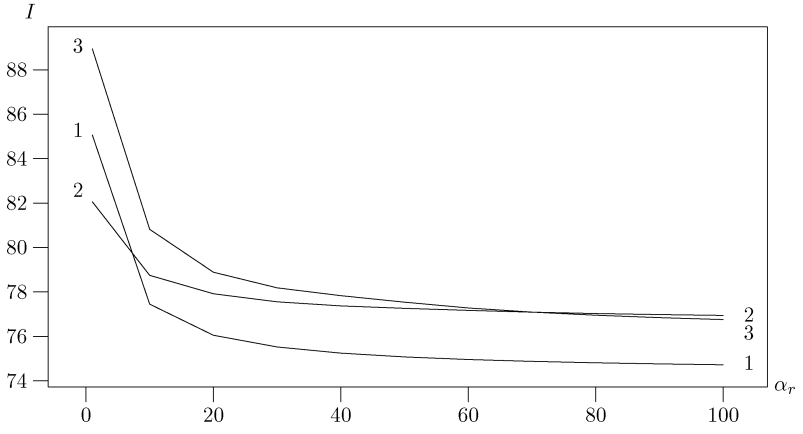

Fig. 3 Dependence of optimal cost function value on $\alpha^{(r)}$

The second *BMAP* has the fundamental rate $\lambda_2 = 2.14286$, group rate $\lambda_2^{(b)} = 1.42857$, squared variation coefficient 1.13944 and coefficient of correlation 0.0350749. The average service time in the second mode is $b_1^{(2)} = 0.322414$.

In the third mode, the *BMAP* is defined by the matrices

$$D_0^{(3)} = \begin{pmatrix} -2.45 & 1.45 \\ 0.06 & -0.6 \end{pmatrix}, D_1^{(3)} = \begin{pmatrix} 0.5 & 0 \\ 0.2 & 0.1 \end{pmatrix},$$

$$D_2^{(3)} = \begin{pmatrix} 0.5 & 0 \\ 0.04 & 0.2 \end{pmatrix}, D_k^{(3)} = 0, k > 2.$$

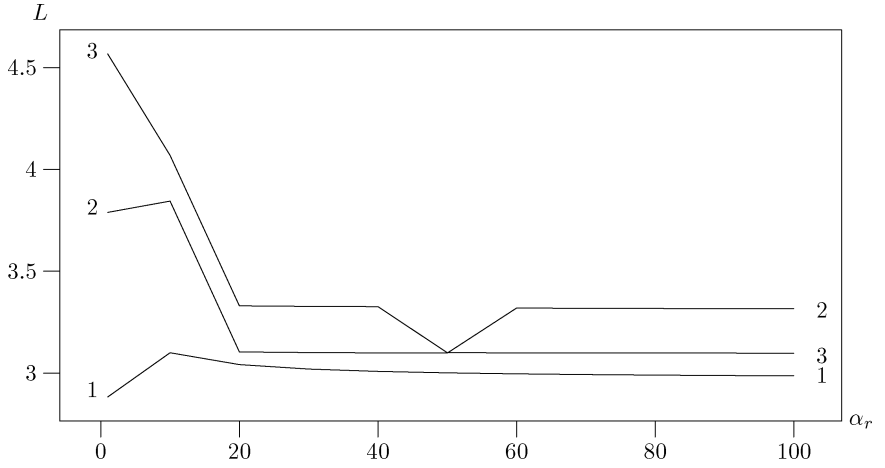

Fig. 4 Dependence of mean queue length on $\alpha^{(r)}$

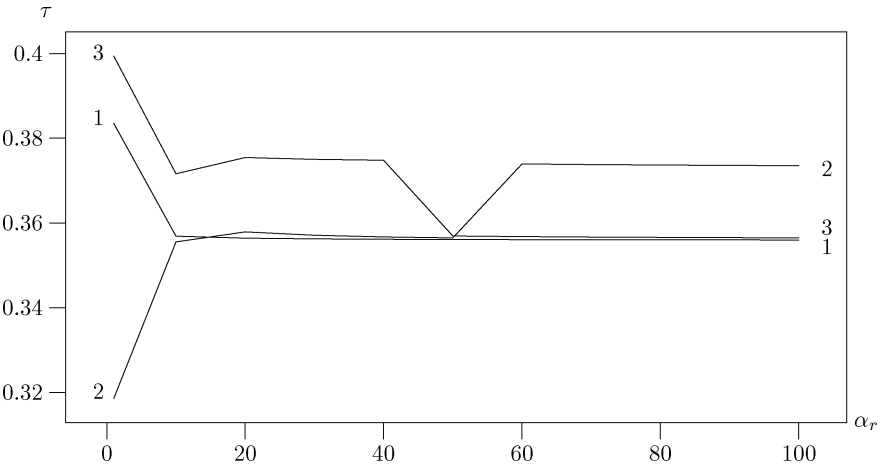

Fig. 5 Dependence of mean interdeparture time on $\alpha^{(r)}$

The matrix P in description of the semi-Markovian kernel $B_3(t)$ is defined as

$$P = \begin{pmatrix} 0.5 & 0.5 \\ 0.55 & 0.45 \end{pmatrix}.$$

Both distributions in the form like (29) correspond to the degenerate random with values 0.1 and 0.07 correspondingly.

So, in the third mode, we have the fundamental rate of the *BMAP* $\lambda_3 = 0.903429$, group rate $\lambda_3^{(b)} = 0.618857$, squared variation coefficient 1.17575, correlation coefficient 0.0162124. The average service time is equal to $b_1^{(3)} = 0.0857143$.

The values of the traffic intensity $\rho_r = \lambda_r b_1^{(r)}$ for the described data are

$$\rho_1 = 1.531429,\ \rho_2 = 0.690887,\ \rho_3 = 0.077437.$$

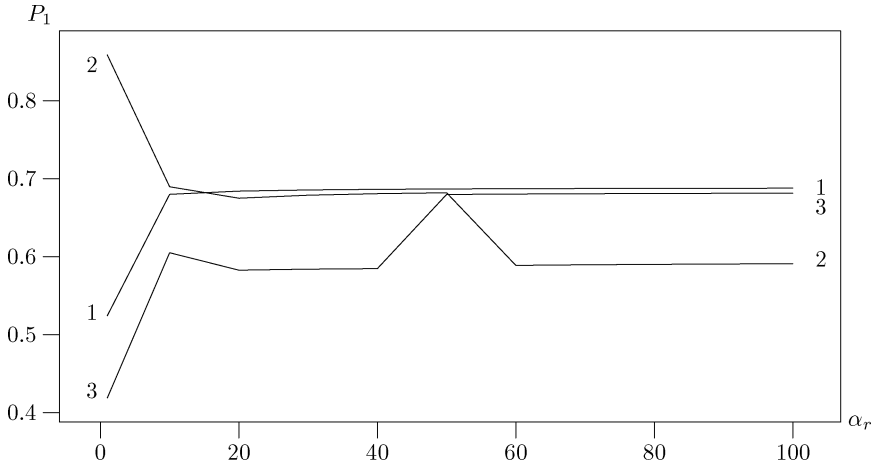

Fig. 6 Dependence of P_1 on $\alpha^{(r)}$

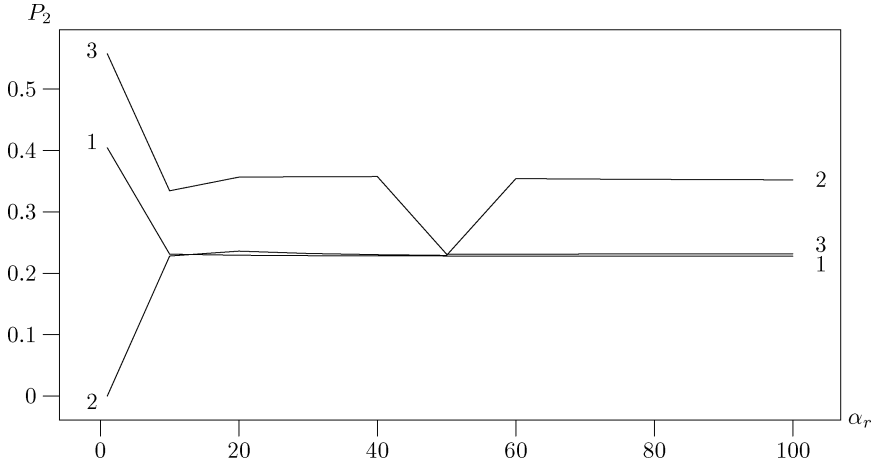

Fig. 7 Dependence of P_2 on $\alpha^{(r)}$

The dependence $\alpha_i^{(r)}$ is given as $\alpha_i^{(r)} = i\alpha^{(r)}$, $r = \overline{1,3}$, where

$$\alpha^{(1)} = 10, \alpha^{(2)} = 35, \alpha^{(3)} = 53. \tag{30}$$

Let the cost coefficients in the cost criterion be the following: $a = 2, c_1 = 2, c_2 = 100, c_3 = 400$.

Now the system and the criterion are defined completely. We can see that the first mode is very cheap, but the traffic intensity is greater than 1 and the use of this only mode does not provide the stationary regime of the system operation. So, here $L = \infty$ and the value E_1 of the cost criterion (1) is infinite.

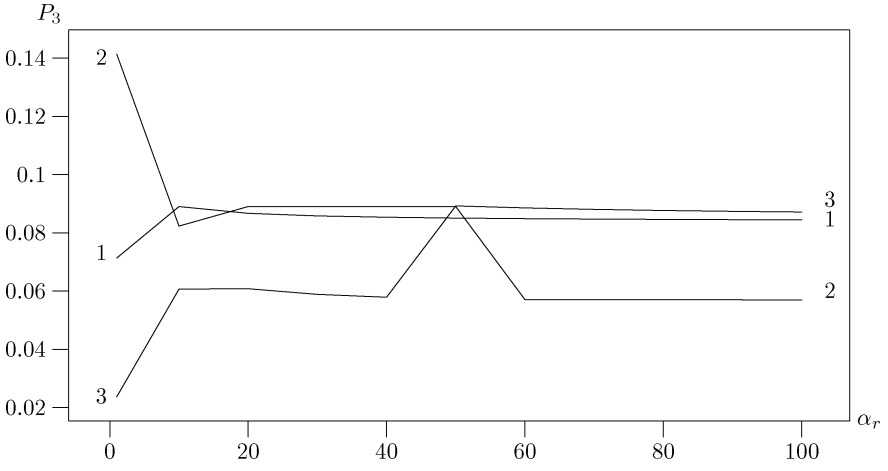

Fig. 8 Dependence of P_3 on $\alpha^{(r)}$

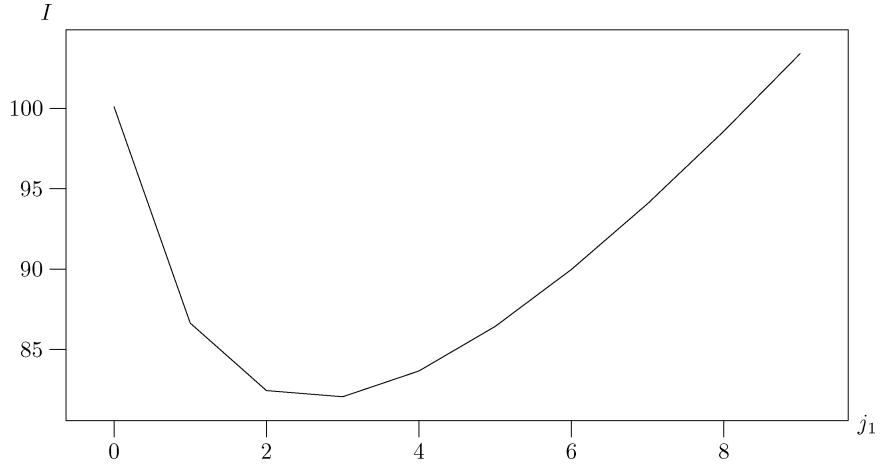

Fig. 9 Dependence of cost function value on j_1

The second mode is much more expensive, but can provide the stationary regime of the system operation. The value E_2 of the cost criterion is equal to_114.9238 if the system operates only in this mode.

The third mode is very expensive, but very productive. The value E_3 of the cost criterion (1) is equal to 400.8305.

The results of the cost criterion (1) calculation for the different sets of the thresholds (j_1, j_2) are presented on Figure 1.

The represented surface has a minimum at the point (2,3) and the optimal value of the criterion (1) is equal to $E^* = 77.4499$. Comparing this value with the value of $E_2 = \min\{E_2, E_3\}$, we see that $E_2/E^* = 1.4838$. So, in comparison with the single mode, the control by the operation modes is profitable. The ratio E_2/E^* is essentially changed when the cost c_3 varies. E.g., if $c_3 = 500$, the optimal thresholds are (2,5) and the ratio is equal to 1.3876. if $c_3 = 150$, this ratio is equal to 2.6835, the optimal thresholds are (1,1).

Figure 2 illustrates level lines for the surface represented on figure 1.

The values $\vec{\pi}_i\mathbf{e}$, of a probability to see i, $i \geq 0$, customers in the orbit upon the customer departure epoch under the optimal set (2,3) of the thresholds are given in the Table 1.

Figures 3–8 illustrate the dependence of the optimal value of criterion (1) and its components: L, τ, P_1, P_2, P_3 on the value of the coefficient $(\alpha^{(r)})$ (the individual retrial rate under the fixed, rth, mode of operation). Three curves in each picture correspond to variation of the value of $\alpha^{(r)}$, $r = \overline{1,3}$ under the value of others intensities $\alpha^{(r')} r' \neq r$ fixed by formula (30). The curve marked as r, $r = \overline{1,3}$, corresponds to the case of the intensity $\alpha^{(r)}$ variation.

Figure 9 represents the dependence of the cost criterion (1) on the threshold if only two modes (first and third) are used.

More detailed analysis of the system (e.g., dependence of the optimal strategy on the input intensity, variation and correlation coefficients; mean value and variation coefficient of the service process; holding and service costs; mechanism of retrials, etc.) can be easily performed basing on the created PC program. This program is built into the software "SIRIUS++" and is available on request from the authors.

Acknowledgments This research was supported by the KOSEF Overseas Lab Program 2003. The authors are thankful to referees for careful reading and valuable remarks.

References

Artalejo J.R. and Gomez-Corral A. (1997). "Steady State Solution of a Single-Server Queue with Linear Repeated Requests." *Journal of Applied Probability* 34, 223–233.

Breuer L., Dudin A.N. and Klimenok V.I. (2002). "A retrial $BMAP|PH|N$ system." *Queueing Systems* 40, 433–457.

Chakravarthy S.R. (2001). "The batch Markovian arrival process: A review and future work." in: *Advances in Probability Theory and Stochastic Processes*, eds. A. Krishnamoorthy, et al.(Notable Publications) pp. 21–39.

Choi B.D., Chung Y.H., Dudin A.N. (2001). "The $BMAP|SM|1$ retrial queue with controllable operation modes." *European Journal of Operational Research* 131, 16–30.

Cinlar E. (1975). *Introduction to Stochastic Processes* (Prentice-Hall).

Dudin A.N. and Chakravarthy S.R. (2002). "Optimal hysteretic control for the $BMAP|G|1$ system with single and group service modes." *Annals of Operations Research* 112, 153–169.

Dudin A.N. and Klimenok V.I. (1999). "Multi-dimensional quasitoeplitz Markov chains." *Journal of Applied Mathematics and Stochastic Analysis* 12, 393–415.

Dudin A.N. and Klimenok V.I. (2000). "A retrial $BMAP/SM/1$ system with linear repeated requests." *Queueing Systems* 34, 47–66.

Dudin A.N., Klimenok V.I., Klimenok I.A., et al. (2000). "Software "SIRIUS+" for evaluation and optimization of queues with the BMAP-input." in: *Advances in Matrix Analytic Methods for Stochastic Models*, eds. G. Latouche and P. Taylor (Notable Publications, Inc., New Jersey) pp. 115–133.

Kemeni J., Shell J. and Knapp A. (1966). "Van Nostrand, New York." *Denumerable Markov chains*.

Klimenok V.I. and Dudin A.N. (2003). "Application of censored Markov chains for calculating the stationary distribution of the multi-dimensional left-skip-free Markov chains." *Queues: flows, systems, networks* 17, 121–128.

Klimenok V.I. (2000). "About stationary distribution existence conditions in queueing sytems with the MAP and retrials." *Reports of Belarusian Academy of Science* 39, 128–132 (in Russian).

Lucantoni D.M. (1991). "New results on the single server queue with a batch markovian arrival process." *Communications in Statistics-Stochastic Models* 7, 1–46.

Neuts M.F. (1989). *Structured Stochastic Matrices of M/G/1 type and their applications* (Marcel Dekker).

Tijms H.C. (1976). "On the optimality of a switch-over policy for controlling the queue size in an M/G/1 queue with variable service rate." *Lecture Notes in Computer Sciences 40*, 736–742.