Journal of Computational and Applied Mathematics 366 (2020) 112425

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Effective algorithm for computation of the stationary distribution of multi-dimensional level-dependent Markov chains with upper block-Hessenberg structure of the generator

Sergei Dudin [a,b], Alexander Dudin [a,b,*], Olga Kostyukova [c], Olga Dudina [a,b]

[a] Belarusian State University, 4, Nezavisimosti Ave., Minsk, 220030, Belarus
[b] Peoples' Friendship University of Russia (RUDN University), 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation
[c] Institute of Mathematics, National Academy of Sciences of Belarus, Minsk, Belarus

ARTICLE INFO

Article history:
Received 15 February 2019
Received in revised form 26 June 2019

Keywords:
Level-dependent multi-dimensional Markov chains
Effective algorithm
Batch Markovian arrival process
Unreliable service
Retrial queue

ABSTRACT

Multi-dimensional level-dependent Markov chains with the upper block-Hessenberg structure of the generator have found extensive applications in applied probability for solving the problems of queueing, reliability, inventory, etc. However, the problem of computing the stationary distribution of such chains is not completely solved. There is a known algorithm for multi-dimensional Asymptotically Quasi-Toeplitz Markov Chains, but, it is required a large amount of computer resources and time-consuming. In this paper, we propose a new effective algorithm that is much less time- and memory-consuming. The new algorithm can be used for analyzing any multi-dimensional Markov chain with the considered structure of the generator. To numerically demonstrate the advantages of this algorithm over the known one, we use it for analysis of a novel single-server retrial queueing system with the batch Markovian arrival process (*BMAP*), a finite buffer, non-persistent customers and an unreliable server. We derive a transparent ergodicity condition for this queueing system. Then, assuming that this condition is fulfilled, we apply the new algorithm and demonstrate its advantages over the known one.

1. Introduction

Queueing theory is very useful for solving numerous problems of capacity planning, performance evaluation and optimization of various real-world systems. The majority of the research in the borders of this theory assumes the existence of buffers in which arriving customers are stored in the case they cannot be accepted for service immediately upon arrival. However, it is typical in many practically important systems, see, e.g., [1–3] that the customers that cannot be immediately accepted or stored in some buffer temporarily leave the service area and make the repeated attempts (retrials) to obtain service.

An analysis of queueing systems with retrials is more complicated than the analysis of their counterparts with an infinite buffer due to the following two reasons.

* Corresponding author.
E-mail addresses: dudins@bsu.by (S. Dudin), dudin@bsu.by (A. Dudin), kostyukova@im.bas-net.by (O. Kostyukova), dudina@bsu.by (O. Dudina).

https://doi.org/10.1016/j.cam.2019.112425
0377-0427/

(a) One of the main random processes describing any queueing system is the number of customers in the system. In the simplest settings, when all distributions characterizing the dynamics of the system are exponential, this process is a partial case of *one-dimensional* Markov chain (the so-called birth-and-death process), which represents a very well studied in the probability theory subject, see, e.g. [4]. The generator of the birth-and-death process has the following tridiagonal structure:

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \dots \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & \dots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & \dots \\ 0 & 0 & \mu_3 & -(\lambda_3+\mu_3) & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{1}$$

where the parameters λ_i and μ_i are the rates (intensities) of the birth and the death at the state $i,\ i \geq 0$.

For the corresponding systems with retrials, to construct the Markov process, it is necessary to split the number of customers in the system into two dependent processes: the number of customers making retrials (these customers are assumed to be staying at some virtual place called *orbit*) and the number of customers in the service area. Therefore, the dynamics of the retrial system even in the simplest settings is described by a *two-dimensional* birth-and-death process or quasi-birth-and-death (*QBD*) process having one denumerable component sometimes called as the *level* and one finite component called as *phase*, see, e.g., [5,6] and [7].

The generator of this process has the following block-tridiagonal structure:

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & O & O & \dots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & O & \dots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \dots \\ O & O & Q_{3,2} & Q_{3,3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{2}$$

where the entries of matrix (2) are not the scalars (as in matrix (1)), but the matrices.

If the blocks of matrix (2) are defined as $Q_{i,i-1} = Q_0,\ Q_{i,i} = Q_1,\ Q_{i,i+1} = Q_2$ for all $i,\ i \geq 1$, the *QBD* process is called as a **level-independent** *QBD* process. Otherwise, the *QBD* process is called as a **level-dependent** *QBD* process.

(b) In the majority of retrial queueing models (except a bit artificial case of the so-called constant retrial rate), the process describing the behavior of the system is a level-dependent *QBD* process, while for the similar queueing models with buffers this process is a level-independent *QBD* process. The investigation of level-dependent birth-and-death and *QBD* processes is more complicated than the analysis of their level-independent analogs. For birth-and-death processes, the analysis in the level-independent case is possible even for time-dependent probabilities of the states of the process. While in the level-dependent version, only formulas for the stationary probabilities are known.

A situation in the analysis of *QBD* processes is worse. Explicit formulas for the stationary probabilities are not known. In the case of the level-independent *QBD* process, there exist the elegant *algorithmic* results for the computation of the stationary probabilities, see [5,6] and [7]. In the case of level-dependent *QBD* processes, only some algorithmic results for the computation of the stationary distribution are presented in [8,9] and [10].

Often, in real-world systems customers can arrive in batches. The multi-dimensional Markov chains describing behavior of such systems have a block upper-Hessenbergian structure of the generator

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & \dots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & Q_{1,3} & \dots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \dots \\ O & O & Q_{3,2} & Q_{3,3} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{3}$$

This structure is more general than (2) and assumes that the denumerable component i_t of the Markov chain can make jumps up from the state i not only to the state $i+1$, but to the states $i+2, i+3, \dots$.

The important particular case of structure (3) is as follows:

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & \dots \\ Q_0 & Q_1 & Q_2 & Q_3 & \dots \\ O & Q_0 & Q_1 & Q_2 & \dots \\ O & O & Q_0 & Q_1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{4}$$

The matrix of form (4) is quasi-Toeplitz upper-Hessenbergian. The multi-dimensional Markov chains with structure (4) of the generator are called as $M/G/1$-type (or Quasi-Toeplitz) Markov chains. Such Markov chains are comprehensively studied in the book [11] by M. Neuts.

Analysis of the queueing models with the generator of form (3) is a complicated problem. Except a series of papers where the stationary distribution of queueing models with such a generator is found by means of direct truncation of the system of equations for the stationary probabilities, or only the conditional probabilities of the states given that the state of the denumerable component of the chain does not exceed a certain level, only the paper [10] is known in the literature. In that paper, the Markov chains with the generator of form (3) are analyzed under an additional assumption that the following limits exist:

$$Y^{(n)} = \lim_{i\to\infty} R_i^{-1} Q_{i,i+n-1} + \delta_{n,1} I,\ n \geq 0, \tag{5}$$

where R_i is a diagonal matrix with the diagonal entries defined as the moduli of the corresponding diagonal entries of the matrix $Q_{i,i},\ i \geq 0$, I is an identity matrix and $\delta_{n,1}$ is the symbol of Kronecker. Such Markov chains are called in [10] as Asymptotically Quasi-Toeplitz Markov Chains (*AQTMC*).

This additional assumption holds good for many retrial queues, queues with impatient customers, tandem queues, etc. This makes the range of application of results for *AQTMC* quite wide. In [10], the *AQTMC* with the discrete and continuous-time are investigated. Sufficient conditions for ergodicity and non-ergodicity are derived. An algorithm for the computation of the stationary distribution is proposed. The main idea of this algorithm is not to solve the equilibrium equations for the stationary probabilities, but to derive and solve an alternative system of equations for the stationary probabilities. This system is derived by means of the construction of series of so-called censored Markov chains (see [12]) with various levels of censoring.

An important step of this algorithm is the computation of matrices G_i the entries of which define the probabilities of transitions of finite components of the chain in the time interval during which the denumerable component first time transits from the state $i+1$ to the state i. These matrices are computed from a backward recursion and the problem of the choice of the initial state of this recursion arises. A certain solution to this problem is proposed in [10]. This solution leads to the successful computation of the stationary probability vectors. However, this solution often implies the necessity of the computation of a huge set of matrices $G_i,\ i \geq 0$, which is both time- and memory-consuming.

Recently, an alternative way for the computation of the matrices $G_i,\ i \geq 0$, and stationary distribution of the Markov chain is offered and its advantages are illustrated in [9] for the level-dependent *QBD* process with block tridiagonal structure (2) of the generator and property (5).

In many cases, the blocks $Q_{i,j}$ of the generator (3) are equal to zero matrix for sufficiently large values of j, e.g., when $j > i + L$ where L is a certain constant. In application to analysis of real-world systems, the parameter L corresponds to the maximum of the size of arriving batch of customers. Therefore, it is important to analyze Markov chains with the generator of form (6):

$$Q = \begin{pmatrix} Q_{0,0} & Q_{0,1} & Q_{0,2} & Q_{0,3} & \cdots & Q_{0,L} & O & O & O & O & \cdots \\ Q_{1,0} & Q_{1,1} & Q_{1,2} & Q_{1,3} & \cdots & Q_{1,L} & Q_{1,L+1} & O & O & O & \cdots \\ O & Q_{2,1} & Q_{2,2} & Q_{2,3} & \cdots & Q_{2,L} & Q_{2,L+1} & Q_{2,L+2} & O & O & \cdots \\ O & O & Q_{3,2} & Q_{3,3} & \cdots & Q_{3,L} & Q_{3,L+1} & Q_{3,L+2} & Q_{3,L+3} & O & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}. \tag{6}$$

In this paper, we extend the results from [9] obtained for the level-dependent *QBD* process with block tridiagonal structure (2) of the generator to a more general case of the Markov chain with the generator of form (6). It is important to mention that our algorithm does not exploit, in contrast to [10], the asymptotic properties of the Markov chain and can be applied for the analysis of an arbitrary ergodic level-dependent Markov chain with the upper block-Hessenberg structure of the generator of form (6).

The remainder of the paper is organized as follows. In the next section, the known algorithm for the computation of the stationary probability vectors is given in brief. New algorithms for the computation of the stationary distribution of the Markov chains with the generator of form (6) are presented in Section 3. The rest part of the paper is devoted to the illustration of the application of the proposed algorithm. In Section 4, we describe the mathematical model of the *BMAP*/*PHF*/1/*N* retrial system with a finite buffer and non-persistent customers. The dynamics of the system is described by a multi-dimensional continuous-time Markov chain with the generator of type (6). The explicit form of the generator is presented. The sufficient condition for the ergodicity of the Markov chain is derived. Formulas for computing the key performance indicators of the system are presented. The results of numerical experiments are presented. Finally, some conclusions can be found in Section 5.

2. Known algorithm for computation of the vectors of stationary probabilities of the Markov chain

We consider a multi-dimensional Markov chain with one countable component i_t and the generator of form (6). We call as level i the set of the states of this Markov chain having the value i of the denumerable component, $i \geq 0$. We assume that the Markov chain is ergodic. Then, the row vectors $\boldsymbol{\pi}_i$ consisting of the stationary probabilities of the states from the level i, $i \geq 0$, exist.

It is well known that the probability vectors $\boldsymbol{\pi}_i,\ i \geq 0$, can be found from the system of Chapman–Kolmogorov equations

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)\mathbf{Q} = \mathbf{0}, \tag{7}$$

$$(\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \dots)\mathbf{e} = 1. \tag{8}$$

Here, $\mathbf{e}$ is a column vector of 1's and $\mathbf{0}$ is a row vector of 0's.

System (7), (8) has the infinite number of equations and the problem of solving this system is very difficult except the case when the matrix Q has the level-independent structure. The more detailed form of Eq. (7) is

$$\sum_{i=0}^{l+1} \boldsymbol{\pi}_i Q_{i,l} = \mathbf{0},\ \ l \geq 0. \tag{9}$$

Theoretically, it could be possible to organize a recursive procedure for the computation of the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, like

$$\boldsymbol{\pi}_{i+1} = \sum_{k=0}^{i} \boldsymbol{\pi}_k Q_{k,i}(-Q_{i+1,i})^{-1},\ \ i \geq 0. \tag{10}$$

However, such a procedure does not work due to two reasons:

(i) a realization of such a procedure requires the knowledge of the vector $\boldsymbol{\pi}_0$ as the initial condition of the recursion. Indeed, the problem of computing this vector is very difficult and is solved only in the case of level-independent Markov chains with the generator in form (4) (using considerations of analyticity of the generating function of the vectors $\boldsymbol{\pi}_i,\ i \geq 0$, in the unit disk of the complex plane or an interpretation of the stationary probability of a state of the Markov chain in terms of the expectation of the number of jumps between two consecutive visits to this state).

(ii) very often, the matrix $Q_{i+1,i}$ is singular and its inversion in (10) is illegal.

Therefore, to compute unknown vectors $\boldsymbol{\pi}_i,\ i \geq 0$, it was proposed in [10] not to solve system (7) but to derive another infinite system of equations for these vectors. This was done in [10] by means of the construction of a series of the so-called censored Markov chains (see [12]) with various levels of censoring. As the result, the following system of equations was derived:

$$\sum_{i=0}^{l} \boldsymbol{\pi}_i A_{i,l} = \mathbf{0},\ \ l \geq 0. \tag{11}$$

Details of derivation of system of equations (11) can be found in [10]. Having system (11), we do not have principal difficulties in its solving. If we set in (11) $l = 0$, we obtain the equation

$$\boldsymbol{\pi}_0 A_{0,0} = \mathbf{0} \tag{12}$$

for the unknown vector $\boldsymbol{\pi}_0$. When the ergodicity condition is fulfilled, the matrix $A_{0,0}$ is singular, but its rank is equal to the dimension of the vector $\boldsymbol{\pi}_0$ minus one. So, by replacing one equation in (12) with the inhomogeneous equation derived from normalization condition (8), we can compute the unique solution of this equation.

The rest of the vectors $\boldsymbol{\pi}_i,\ i \geq 1$, are easily computed from the recursion that evidently follows from (11):

$$\boldsymbol{\pi}_l = \sum_{i=0}^{l-1} \boldsymbol{\pi}_i A_{i,l}(-A_{l,l})^{-1},\ \ l \geq 1.$$

Here, the matrix $A_{l,l}$ for all $l,\ l \geq 1$, is non-singular as the irreducible sub-generator.

The key roles in the computation of the matrices $A_{i,l}$ for system (11) are played by the matrices $G_i,\ i \geq 0$, the entries of which define the probabilities of transition of finite components of the chain in the time interval during which the chain first time transits from the level $i+1$ to the level i. The brief outline of the algorithm for the computation of the matrices $A_{i,l}$, which was elaborated in [10], adapted to the Markov chain with the generator of form (6) is as follows.

- Compute the matrices G_i by recursion

$$G_i = (-\sum_{n=i+1}^{i+1+L} Q_{i+1,n} G_{n-1} G_{n-2} \dots G_{i+1})^{-1} Q_{i+1,i},\ \ i \geq 0. \tag{13}$$

- Compute the matrices $A_{i,l}$ as

$$A_{i,l} = Q_{i,l} + \sum_{n=l+1}^{i+L} Q_{i,n} G_{n-1} G_{n-2} \dots G_l,\ \ l = \overline{i, i+L},\ i \geq 0.$$

- Compute the matrices $F_l,\ l \geq 0$, as

$$F_0 = I, F_l = \sum_{i=\max\{0,l-L\}}^{l-1} F_i A_{i,l}(-A_{l,l})^{-1},\ l \geq 1. \tag{14}$$

- Compute the vector $\boldsymbol{\pi}_0$ as the unique solution to the system

$$\begin{cases} \boldsymbol{\pi}_0(-A_{0,0}) = \mathbf{0}, \\ \boldsymbol{\pi}_0 \sum_{l=0}^{\infty} F_l \mathbf{e} = 1. \end{cases} \tag{15}$$

- Compute the vectors $\boldsymbol{\pi}_l,\ l \geq 1$, as

$$\boldsymbol{\pi}_l = \boldsymbol{\pi}_0 F_l,\ \ l \geq 1. \tag{16}$$

The presented scheme looks quite simple. However, some technical problems arise at the stage of its computer implementation. Two main problems are the following ones.

(a) Recursion (13) is the backward one. To compute the matrix $G_l,\ \ l \geq 0$, we have to know all matrices $G_k,\ \ k > l$. Therefore, the problem of fixing the initial (terminal) condition for this recursion exists.

(b) The second equation in (15) contains the sum of an infinite number of summands and it is necessary to decide how to truncate this sum.

To solve the problem (a), the following reasonings can be used. It follows from the definition of *AQTMC*, that for large values of the denumerable component i_t of Markov chain $\xi_t,\ \ t \geq 0$, the behavior of *AQTMC* becomes very close to the behavior of Markov chain of $M/G/1$ type with the transition probability matrix of form (4) where the blocks Q_k are replaced with the matrices $Y^{(k)}$ from the definition of *AQTMC*. Thus, the sequence of the matrices G_i tends, when i approaches infinity, to the matrix G that defines the probabilities of transitions of finite components of the $M/G/1$ type Markov chain in the time interval during which the denumerable component first time transits from the state $i+1$ to the state i. Because the $M/G/1$ type Markov chain has a level-independent transition probability matrix, this matrix indeed does not depend on the value of i.

It is well known, see, e.g., the book [11], that the matrix G is the minimal nonnegative solution to the matrix equation

$$G = \sum_{k=0}^{\infty} Q_k G^k.$$

This matrix equation can be solved by means of the various variants of iteration method, see, e.g., [13]. One of the simplest versions of iterations organization is as follows:

$$G^{(0)} = O,\ G^{(m+1)} = \sum_{k=0}^{L+1} Q_k (G^{(m)})^k,\ \ m > 0.$$

Iterations are stopped when the norm of matrix $G^{(m+1)} - G^{(m)}$ becomes less than the pre-assigned small value. After that the matrix G is set to be equal $G^{(m+1)}$.

Because we mentioned above that the sequence of the matrices G_i tends when i approaches infinity, to the matrix G (formally this is stated in [10]), according to the definition of the limit this implies that for any predefined small number $\varepsilon_G > 0$, there exists such a value i^* that the norm of the matrix $G_i - G$ is less than ε_G for all $i,\ i \geq i^*$.

Therefore, to choose the value i_0 such as the matrices G_i will be set equal to the matrix G for all $i,\ i \geq i_0$, we have to implement the following steps:

Step 1. Fix an arbitrary small number $\varepsilon_G > 0$, and i^* as an arbitrary positive integer, say, $i^* = 1000$.

Step 2. Set $G_{i^*+k} = G$ for all $k,\ k \geq 1$, and compute the matrix G_{i^*} from relation (13).

Step 3. Compare the norm of the matrix $G_{i^*+1} - G_{i^*}$ with the value ε_G. If this norm is less than ε_G, we set $i_0 = i^*$ and stop computations. The required value i_0 is found. Otherwise, we increase the value i^*, e.g., multiply it by the factor 2, and go to Step 2.

Problem (b) is solved more easily. All the matrices $(-A_{l,l})^{-1}$ in direct recursion (14) exist and are non-negative because the matrices $A_{l,l}$ for all $l,\ \ l \geq 1$, are the irreducible sub-generators. Thus, it is clear from (14) that all matrices $F_i,\ \ i \geq 0$, have non-negative entries and, because we assume that the considered Markov chain is ergodic, the norm of the matrix F_i tends to zero when i approaches infinity. So, recursive calculation of the matrices F_i from (14) can be stopped when the norm of the matrix F_i becomes less than some preassigned accuracy level ε_F.

Therefore, in principle, problems (a) and (b) are solved. However, the use of this solution in computer implementation meets significant difficulties such as a very long computational time and the excessive use of computer resources. Although the number, say i^*, of matrices G_i directly used for computation of the stationary probabilities may be relatively small, to compute the matrices $G_l,\ l = \overline{0, i^*}$, it is necessary to implement a huge number of steps of the backward recursion. E.g., to compute the stationary probabilities of the retrial queueing system, which parameters are presented in Section 4 (for $N = 10$), using the algorithm from [10], it is required to compute $i_0 = 342\,200$ matrices G_i. While only 397 of them (about 0.116%) are indeed required for computation of the required number of stationary probability vectors.

3. New algorithm for the computation of stationary distribution of *AQTMC*

3.1. Key idea

For developing a new algorithm we should overcome the disadvantage of the known algorithm, i.e., we need to avoid the necessity to compute a large number of the matrices G_i that is not directly used for computation of the stationary probabilities. Namely, we need to elaborate a procedure for computing the matrices $G_i, \ldots, G_0$ for some arbitrary i without computing all the matrices $G_k,\ k > i$. To this end, let us formulate the following assertion.

Lemma 1. *For any value $i,\ \ i \geq 1$, of the countable component i_t of the ergodic Markov chain $\xi_t,\ t \geq 0$, and for any fixed arbitrarily small value $\epsilon,\ \epsilon > 0$, there exists $s_0,\ s_0 < \infty$, such that, for all $s \geq s_0$, the probability of the transition of the chain $\xi_t,\ t \geq 0$, from the level $i+1$ to the level $i+s$ without visiting the level i is less than ϵ.*

The validity of the lemma follows from the fact that if there exists such a value ϵ that for all s_0 there exists $s \geq s_0$, such that the probability of the transition of the chain $\xi_t,\ t \geq 0$, from the level $i+1$ to the level $i+s$ without visiting the level i is greater than ϵ, then the chain is evidently not ergodic.

We already mentioned that the entries of the matrix G_i define the transition probabilities of the finite components of the Markov chain ξ_t during the time when the chain first time transits from the level $i+1$ to the level i. It can be shown by repeating recursion (13) s times that the matrix G_i can be expressed via the generator blocks and the matrices $G_{i+s+l},\ l \geq 0$. Based on Lemma 1, we can conclude that for sufficiently large s, the probability of reaching the level $i+s$ starting from the level $i+1$ without visiting the level i is negligible. Thus, entries of the matrix G_i do not depend on transitions that the finite components of the chain make starting from the level $i+s$ because the chain has no chance to reach the level $i+s$ starting from the level $i+1$ without falling to the level i. Therefore, the matrix G_i does not depend on the matrices $G_{i+s+l},\ l \geq 0$, where the value s is sufficiently large.

So, to compute the matrices $G_m,\ m = \overline{0,i}$, we can fix some sufficiently large s, set $G_{i+s+l} = C,\ l \geq 0$, where C is some arbitrary matrix of the corresponding dimension which does not entail the degeneracy of the matrices $\sum_{n=i+1}^{i+1+L} Q_{i+1,n}C^{n-i-1}$ and use recursion (13). Now, we should solve the problem of choosing the sufficiently large value s. This problem as well as the problem of computing the matrix $G_l,\ l = \overline{0,i}$, can be solved by means of Procedure 1.

Procedure 1.

1. We fix s as an arbitrary large integer value, e.g., $s = 100$, and some small ε_G, e.g., $\varepsilon_G = 10^{-12}$.
2. We set $G^{(1)}_{i+s+l} = O,\ l \geq 0$, and $G^{(2)}_{i+s+l} = I,\ l \geq 0$.
3. Sequentially calculate the matrices $G^{(1)}_i$ and $G^{(2)}_i$ based on recursion (13).
4. Compute the norm of the matrix $G^{(1)}_i - G^{(2)}_i$. If the norm is less than ε_G, the chosen value s is sufficiently large, and G_i does not depend on $G_{i+s+l},\ l \geq 0$, and we go to step 5. Otherwise, we increase s, e.g., multiply s by some factor, and go to step 2.
5. We set $G_{i+l} = G^{(1)}_{i+l},\ l \geq 0$, and compute the matrices $G_k,\ k < i$, by recursion (13).

Explanation. The main idea of Procedure 1 is the following. We fix some large level s and compute the matrix G_i from recursion (13) assuming absolutely different initial conditions $G_{i+s+l} = C_1,\ l \geq 0$, and $G_{i+s+l} = C_2,\ l \geq 0$. If the matrices G_i obtained with the use of different initial conditions coincide, we assume that the value s sufficiently large. Otherwise, we increase s. As initial conditions we choose $G_{i+s+l} = C_1 = O,\ l \geq 0$, and $G_{i+s+l} = C_2 = I,\ l \geq 0$. Note, that the initial conditions are not relevant to the actual values of the matrices $G_{i+s+l},\ l \geq 0$. Moreover, all matrices $G_i,\ i \geq 0$ are stochastic matrices. We formally set $G_{i+s+l} = O,\ l \geq 0$, and $G_{i+s+l} = I,\ l \geq 0$, to choose the value of s such that the value of G_i does not depend on $G_{i+s+l},\ l \geq 0$. Such a choice of the initial conditions provide the non-degeneracy of all inverted matrices. The entries of the inverse matrices are non-negative. If we found that the matrix G_i does not depend on $G_{i+s+l},\ l \geq 0$, then all matrices $G_k,\ k < i$, also do not depend on $G_{i+s+l},\ l \geq 0$. Thus, assuming $G_{i+l} = G^{(1)}_{i+l},\ l \geq 0$, we can compute all matrices $G_k,\ k < i$.

After we succeed in developing Procedure 1, the scheme of computing the stationary distribution can be represented as follows.

We fix some value i_0 as the anticipated number of the matrices F_l, required for computing the stationary distribution, and compute all the matrices $G_i,\ i = \overline{0,i_0}$, based on Procedure 1. Then, we compute the matrices F_i for $i = \overline{0,i_0}$, from (14). If, during the computation we obtain that $\|F_{l^*}\| < \varepsilon_F,\ l^* \leq i_0$, we can stop the computation of the matrices F_i and find the stationary probabilities $\boldsymbol{\pi}_i,\ \ 0 \leq i \leq l^*$, of the system from (15)–(16) using the level l^* as a truncation level of the infinite sum. If the inequality $\|F_i\| < \varepsilon_F$ is not fulfilled for all $i = \overline{0,i_0}$, then the value of i_0 is not large enough. We increase this value and compute additional amount of the matrices G_i and F_i. We repeat these manipulations until the norm of the matrix F_i becomes less than ε_F for some i.

Based on all the presented above reasonings, in the next section, we formulate a new algorithm for computing the stationary distribution of level-dependent multidimensional Markov chain with the generator of form (6).

3.2. New algorithm

Algorithm 1.

Step 1. (Initialization of parameters) Let us fix the small values $\epsilon_G > 0$ and $\epsilon_F > 0$, e.g., $\epsilon_G = \epsilon_F = 10^{-12}$, set i_0 as a large integer value, e.g. $i_0 = 100$, and set the number n_F of already computed matrices F_i equal to 0.

Step 2. (Computation of matrices G_{k^*+l}, $l = \overline{0, L-1}$, for some $k^* \geq i_0$)

This step consists of the following sub-steps:

Step 2.1. Set the parameter s as an arbitrary large integer number, e.g., $s = 100$.

Step 2.2. Set $k = i_0 + s + L$,

$$C_l = O, \quad B_l = I, \; l = \overline{1, L}.$$

Step 2.3. Compute the matrices

$$C = -(Q_{k+1,k+1} + \sum_{n=k+2}^{k+1+L} Q_{k+1,n} C_{n-k-1} C_{n-k-2} \dots C_1)^{-1} Q_{i+1,i}, \tag{17}$$

$$B = -(Q_{k+1,k+1} + \sum_{n=k+2}^{k+1+L} Q_{k+1,n} B_{n-k-1} B_{n-k-2} \dots B_1)^{-1} Q_{i+1,i}.$$

Then, redefine the matrices $C_l, B_l, \; l = \overline{2, L}$, as

$$C_{l+1} = C_l, \;\; B_{l+1} = B_l, \;\; l = L-1, \dots, 1.$$

Then, set

$$C_1 = C, \;\; B_1 = B.$$

Step 2.4. Compute the norm of the matrix $C_1 - B_1$. If $\|C_1 - B_1\| < \epsilon_G$, then set $k^* = k - L + 1$ and go to Step 2.5. Otherwise, set $k := k - 1$. In the case when $k = i_0 + L - 1$, increase the value s, e.g., $s := 2s$ and go to Step 2.2. If $k \geq i_0 + L$, return to Step 2.3.

Step 2.5. Set $G_{k^*+L-1} = C_1$.

Set $m = 2$.

While $m \leq L$ do the following steps:

- Set $k := k - 1$;
- Compute the matrix C by formula (17);
- Redefine the matrices $C_l, \; l = \overline{2, L}$, as

$$C_{l+1} = C_l, \;\; l = L-1, \dots, 1;$$

- Set $C_1 = C$;
- Set $G_{k^*+L-m} = C_1$;
- Set $m := m + 1$.

Step 3. (Computation of the matrices G_i for $i < k^*$)

Further, we compute all matrices G_i, via the backward recursion

$$G_i = -(Q_{i+1,i+1} + \sum_{n=i+2}^{i+1+L} Q_{i+1,n} G_{n-1} G_{n-2} \dots G_{i+1})^{-1} Q_{i+1,i}$$

where $i = k^* - 1, k^* - 2, \dots, 0$ if $n_F = 0$ and $i = k^* - 1, k^* - 2, \dots, n_F + L$ if $n_F > 0$.

Explanation. Steps 2–3 correspond to the implementation of Procedure 1.

Step 4. (Computation of matrices $A_{l,m}$ and F_l)

Step 4.1. Set $l = n_F + 1$. If $l = 1$, we set $F_0 = I$ and calculate the matrices $A_{0,m}$ by the formulae

$$A_{0,m} = Q_{0,m} + \sum_{n=m+1}^{L} Q_{0,n} G_{n-1} G_{n-2} \dots G_m, \quad m = \overline{0, L}.$$

Step 4.2. Calculate the matrices $A_{l,m}$ by the formulae

$$A_{l,m} = Q_{l,m} + \sum_{n=m+1}^{l+L} Q_{l,n} G_{n-1} G_{n-2} \dots G_m, \quad m = \overline{l, l+L}.$$

Step 4.3. Compute the matrix F_l as

$$F_l = \sum_{i=\max\{0,l-L\}}^{l-1} F_i A_{i,l}(-A_{l,l})^{-1}.$$

Step 4.4. Calculate the norm of the matrix F_l. If $\|F_l\| < \epsilon_F$, then set $l^* = l$ and go to Step 5. Otherwise, if $l < k^*$, increase l by 1 and go to Step 4.2. If $l = k^*$, set $i_0 = k^* + s_1$ where s_1 is some large integer value, e.g., $s_1 = 100$, and set $n_F = k^*$ and go to Step 2.1.

Explanation. If the termination condition $\|F_{l^*}\| < \epsilon_F$ is fulfilled, we can compute the stationary distribution of the system. If $\|F_{k^*}\| \geq \epsilon_F$, the number k^* is not large enough and we need to increase it and go to Step 2.1. Note, that we do not need recalculate the computed matrices $F_l,\ l = \overline{1, k^*}$, thus we set $n_F = k^*$.

Step 5. (Computation of the vectors $\boldsymbol{\pi}_l,\ l = \overline{0, l^*}$)

Compute the vector $\boldsymbol{\pi}_0$ as the unique solution to the system of the linear algebraic equations

$$\boldsymbol{\pi}_0(-A_{0,0}) = \mathbf{0},\ \boldsymbol{\pi}_0 \sum_{l=0}^{l^*} F_l \mathbf{e} = 1,$$

and calculate the stationary distribution of the Markov chain as

$$\boldsymbol{\pi}_l = \boldsymbol{\pi}_0 F_l,\ \ l = \overline{1, l^*}.$$

It is worth noting that it is clear that, due to the ergodicity of the Markov chain, $\|\boldsymbol{\pi}_i\| \to 0$ and $\|F_i\| \to 0$ when $i \to \infty$. Therefore, the algorithm stops its work (the termination condition $\|F_{l^*}\| < \epsilon_F$ is fulfilled) after a finite number of steps. But, we do not consider here the rate of its convergence due to the possible description of the chain by lots of parameters. Results of the numerical experiments for several queueing systems show that this rate essentially depends on too many factors like the load of the system, rate of convergence of retrial intensities, correlation in the arrival process, variance of inter-arrival and service times, higher moments of their distributions, probabilities of the service without the error, etc, and various combinations of these factors.

3.3. Modification of the new algorithm

The new algorithm can be optimized by means of calculation of the vectors $\boldsymbol{\phi}_i$ instead of the matrices F_i as follows.

Algorithm 2.

Step 1. Let us fix the small values $\epsilon_G > 0$ and $\epsilon_{\boldsymbol{\phi}} > 0$, e.g., $\epsilon_G = 10^{-12}$, $\epsilon_{\boldsymbol{\phi}} = 10^{-14}$, and set i_0 as an arbitrary positive integer large value, e.g. $i_0 = 100$, and set $n_{\boldsymbol{\phi}} = 0$.

Step 2. Implement Steps 2–3 from Algorithm 1, only replace the parameter n_F by the parameter $n_{\boldsymbol{\phi}}$.

Step 3.1. Set $l = n_{\boldsymbol{\phi}} + 1$. If $l = 1$, calculate the matrices $A_{0,m}$ by the formulae

$$A_{0,m} = Q_{0,m} + \sum_{n=m+1}^{L} Q_{0,n} G_{n-1} G_{n-2} \dots G_m,\ \ m = \overline{0, L},$$

and find the vector $\boldsymbol{\phi}_0$ as the unique solution to the system

$$\boldsymbol{\phi}_0(-A_{0,0}) = \mathbf{0},\quad \boldsymbol{\phi}_0 \mathbf{e} = 1.$$

Step 3.2. Calculate the matrices $A_{l,m}$ by the formulae

$$A_{l,m} = Q_{l,m} + \sum_{n=m+1}^{l+L} Q_{l,n} G_{n-1} G_{n-2} \dots G_m,\ \ m = \overline{l, l+L}.$$

Step 3.3. Calculate the vectors $\boldsymbol{\phi}_l$ using the formulae

$$\boldsymbol{\phi}_l = \sum_{i=\max\{0,l-L\}}^{l-1} \boldsymbol{\phi}_i A_{i,l}(-A_{l,l})^{-1}.$$

Step 3.4. Compute the norm of the vector $\boldsymbol{\phi}_l$. If $\|\boldsymbol{\phi}_l\| < \epsilon_{\boldsymbol{\phi}}$, then set $l^* = l$ and go to Step 4. Otherwise, if $l < k^*$, increase l by one and go to Step 3.2. If $l = k^*$, set $i_0 = k^* + s_1$, where s_1 is some arbitrary positive integer large value, set $n_{\boldsymbol{\phi}} = k^*$ and go to Step 2.

Step 4. Compute a constant $a = \left(\sum_{l=0}^{l^*} \boldsymbol{\phi}_l \mathbf{e}\right)^{-1}$.

Then calculate the vectors $\boldsymbol{\pi}_l,\ l = \overline{0, l^*}$, as

$$\boldsymbol{\pi}_l = a\boldsymbol{\phi}_l,\ l = \overline{0, l^*}.$$

Remark 1. It is worth to note that the results elaborated in [10] essentially exploit asymptotic assumption (5) for construction of the algorithm, namely, for choosing the terminal condition for the backward recursion (13). This assumption is quite natural for Markov chains describing the behavior of a variety of practically important queueing systems, including retrial systems with the classical and linear strategy of retrials, systems with impatient customers, systems with the infinite number of servers, tandem queueing systems, etc. This property holds true, in particular, for the queueing system under study in the next section. However, definitely, there are situations when the asymptotic assumption (5) is not fulfilled, e.g., when the total retrial rate of customers α_i when i customers are in the orbit cyclically depends on i, e.g., it is equal to a_1 for odd values of i and is equal to a_2, otherwise. The algorithms developed in our paper, exploit only the ergodicity of the Markov chain and do not use asymptotic properties (5). Therefore, these algorithms can be applied to compute the stationary distribution of an arbitrary ergodic level-dependent multidimensional Markov chain with the generator of form (6).

4. Application of algorithms to analysis of a single-server retrial queueing system with the batch Markovian arrival process, a finite buffer, non-persistent customers and an unreliable server

4.1. Motivation of the model and its novelty

The elaborated algorithms can be applied for computation of the stationary distribution of the states and optimization of a variety of queueing and inventory models, especially models with customer retrials and impatience. To illustrate their application to the analysis of a concrete retrial queueing model, we consider in brief a novel single-server retrial queueing system with the *BMAP*, a finite buffer, non-persistent customers and an unreliable server. The *BMAP* was proposed by the research team of M. Neuts as a significant generalization of the stationary Poisson arrival process. The stationary Poisson process is still popular in the engineering literature but badly describes real information flows in modern telecommunication networks and contact centers. These flows exhibit fluctuation of instantaneous arrival rate and correlation of inter-arrival times. The *BMAP* ideally fits to take into account these features of real-world processes. The single-server queueing system with the *BMAP*, general service time distribution, and infinite buffer was first comprehensively analyzed in [14] and [15]. The first paper where a single-server retrial queueing system with the *BMAP* was analyzed is [16]. Here, we consider the model that differs from [16] in the following aspects.

(i) A pure retrial model was considered in [16]. There is no buffer in the system and an arriving customer, which meets busy server, moves the orbit for making repeated attempts. However, in some real-world systems, an additional small buffer exists. E.g., in modeling cellular networks, the customer that requires handover procedure during moving to another cell has a chance to wait during his/her sojourn in the overlapping area of current and target cells. Therefore, the account of the existence of a buffer is important and practically motivated.

(ii) Customers were supposed to be absolutely persistent in [16] while in the real-world systems they can terminate retrials and leave the system permanently. The technique of embedded Markov chains applied in [16] cannot be effectively extended to the case of non-persistent customers.

(iii) The server was supposed to be absolutely reliable in [16]. However, in reality, the servers can be unreliable. Unreliable retrial queues were considered, e.g., in [17,18]. The model considered in [17] has many servers and phase-type service time distribution. The model considered in [18] assumes a general distribution of service and repair times. It is assumed in both these papers that the server may fail. Service is terminated and the server becomes broken and has to be repaired during a certain random time. In our model, we assume that the failure occurrence does not imply the breakdown of the server and the necessity of its recovering.

The model considered in this section can be applied for performance evaluation and capacity planning of telecommunication networks, in particular, wireless networks, including sensor networks designed for security provisioning of some objects under the impact of external influences. In the application of retrial queueing models for modeling wireless communications, the unreliability of the server is related mainly to the noise in the transmitting thread. Therefore, a failure in transmission does not require the repair of the server. Only the repeated service of a customer, during service of which an error occurs, is required. Thus, the results from [17] and [18] cannot be directly applied for modeling systems with unreliable wireless transmission of information, while the mechanism of unreliable service considered in our paper ideally fits such systems.

4.2. Description of the mathematical model

The structure of the considered queueing system is given in Fig. 1.

Customers arrive at the system in batches according to the *BMAP*, see, e.g., [19] and [14]. In the *BMAP*, the potential moments of batches arrival are defined as the moments of jumps of the irreducible continuous-time Markov chain ν_t, $t \geq 0$, with the finite state space $\{0, 1, \ldots, W\}$ that is called as the underlying process of the *BMAP*. The *BMAP* is completely characterized by the matrices D_k which consist (except the diagonal entries of the matrix D_0) of the intensities of jumps of the process ν_t that are accompanied by the arrival of a batch of k customers. In this paper, we suppose that the *BMAP* has finite support, i.e., the maximum size of the batch is L where $L < \infty$. The matrix $D(1) = \sum_{k=0}^{L} D_k$ represents the generator of the process ν_t. The average arrival rate (fundamental rate) λ of the *BMAP* is defined by $\lambda = \boldsymbol{\theta} \sum_{k=1}^{L} kD_k\mathbf{e}$

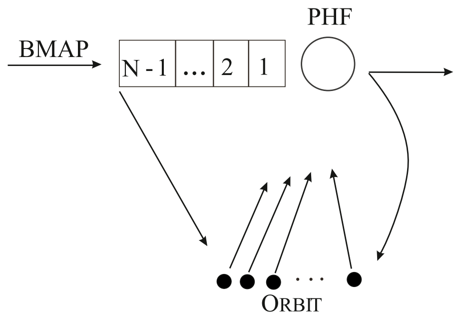

Fig. 1. Structure of queueing system under study.

where $\boldsymbol{\theta}$ is the invariant vector of the stationary distribution of the process ν_t. This vector is the unique solution to the system $\boldsymbol{\theta}D(1) = \mathbf{0}$, $\boldsymbol{\theta}\mathbf{e} = 1$.

The system has a single server and a finite buffer of capacity $N-1$, $N > 1$. The presence of a finite buffer in the retrial model is assumed, e.g., in [20]. If the server is idle at the moment of a batch arrival, one customer of a batch immediately starts service. The rest of the batch occupies the corresponding number of places in the buffer. If there is no place in the buffer for all customers, the redundant customers join the orbit of infinite capacity. If the server is busy at the moment of a batch arrival, the customers of this batch occupy the corresponding number of places in the buffer. Analogously, if there is no place in the buffer for all customers, the redundant customers join the orbit.

The customers staying in the orbit try to obtain access, independently of each other. The total retrial rate when i customers are in the orbit is equal to α_i, $i \geq 1$, such as $\alpha_i \to \infty$ when $i \to \infty$. If the buffer is full at the moment of a retrial attempt, the customer leaves the system permanently with probability p, $0 \leq p \leq 1$. With the complementary probability, the customer returns to the orbit.

We assume that the server is not absolutely reliable and errors can occur during service. The account of the unreliability of the server is very important for the correct prediction of the values of performance measures of any queueing system, for references see, e.g., [17,21]. In our model, we assume that error occurrence does not mean the server breakdown and the necessity of its repairing. Error occurrence just implies the termination of a customer's service. The customer, during service of which the error occurs, joins the orbit or repeats service from the early beginning or from the phase at which the error occurs. To take these features into account, we suppose that the service time of the customer has the so-called *PHF* (phase type with failures) distribution, that is the natural extension of the classical *PH* (phase-type) distribution, see [5] and [7], to the case when the server is unreliable.

The *PHF* distribution was recently introduced in [22]. This distribution is defined by the underlying irreducible continuous-time Markov chain η_t, $t \geq 0$, with a finite state space $\{1, 2, \dots, M, M+1, M+2\}$ and the set of parameters $(\boldsymbol{\beta}, S, \mathbf{S}_1, \mathbf{S}_2, q_1, q_2)$. The row vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_M)$ defines the choice of the state of the process η_t from the set $\{1, 2, \dots, M\}$ at the service beginning instant. The sub-generator S defines the transition rates of the chain within this set. The column vector $\mathbf{S}_1$ defines the rates of transition to the absorbing state $M+1$ which corresponds to the successful service completion. The column vector $\mathbf{S}_2$, $\mathbf{S}_2 = -S\mathbf{e} - \mathbf{S}_1$, defines the rates of transition to the absorbing state $M+2$ which corresponds to a failure occurrence. After the failure occurrence the customer, service of which is terminated, joins the orbit with probability q_1, restarts service from the beginning with probability q_2, and restarts service from the phase, at which the failure occurred, with probability $1 - q_1 - q_2$. For more information about the *PHF* distribution, see [22].

4.3. The process of system states

The dynamics of the system under study is described by the process

$$\xi_t = \{i_t, n_t, \nu_t, \eta_t\},\ t \geq 0,$$

where, at the moment t, $t \geq 0$,

i_t is the number of customers in the orbit, $i_t \geq 0$;

n_t is the number of customers in the system (in service and in the buffer), $n_t = \overline{0, N}$;

ν_t is the state of the underlying process of the *BMAP*, $\nu_t = \overline{0, W}$;

η_t is the state of the *PHF* underlying process, $\eta_t = \overline{1, M}$, if $n_t > 0$. When $n_t = 0$, the component η_t is not defined and is not included into the process ξ_t.

It is easy to see that the process ξ_t is the regular irreducible continuous-time Markov chain.

Let us enumerate the states of the Markov chain ξ_t in the lexicographic order and call the set of the states having the value i of the first component as the level i. Each level consists of $(W+1)(NM+1)$ states. Let Q be the generator of the Markov chain.

Lemma 2. *The infinitesimal generator Q of the Markov chain ξ_t, $t \geq 0$, has structure (6) with the non-zero blocks $Q_{i,j}$ having the following form:*

$$Q_{i,i-1} = \alpha_i H,$$

where

$$H = \begin{pmatrix} O & I_{\bar{W}} \otimes \boldsymbol{\beta} & O & \dots & O \\ O & O & I_{\bar{W}M} & \dots & O \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ O & O & O & \dots & I_{\bar{W}M} \\ O & O & O & \dots & pI_{\bar{W}M} \end{pmatrix},$$

$$Q_{i,i+1} = \begin{pmatrix} O & O & O & \dots & O & \delta_{N+1\leq L} D_{N+1} \otimes \boldsymbol{\beta} \\ q_1 I_{\bar{W}} \otimes \mathbf{S}_2 & O & O & \dots & O & \tilde{D}_N \\ O & q_1 I_{\bar{W}} \otimes \mathbf{S}_2 \boldsymbol{\beta} & O & \dots & O & \tilde{D}_{N-1} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ O & O & O & \dots & O & \tilde{D}_2 \\ O & O & O & \dots & q_1 I_{\bar{W}} \otimes \mathbf{S}_2 \boldsymbol{\beta} & \tilde{D}_1 \end{pmatrix},$$

$$Q_{i,i+k} = \begin{pmatrix} O & \dots & O & \delta_{k+N\leq L} D_{k+N} \otimes \boldsymbol{\beta} \\ O & \dots & O & \tilde{D}_{k+N-1} \\ O & \dots & O & \tilde{D}_{k+N-2} \\ \vdots & \ddots & \vdots & \vdots \\ O & \dots & O & \tilde{D}_k \end{pmatrix}, k = \overline{2, L},$$

$$(Q_{i,i})_{n,n'} = \begin{cases} O & n' < n-1,\ n = \overline{2,N}, \\ I_{\bar{W}} \otimes \mathbf{S}_1, & n' = 0,\ n = 1, \\ I_{\bar{W}} \otimes \mathbf{S}_1 \boldsymbol{\beta}, & n' = n-1,\ n = \overline{2,N}, \\ D_0 - \alpha_i I_{\bar{W}}, & n' = n = 0, \\ A - \alpha_i I_{\bar{W}M}, & n' = n,\ n = \overline{1, N-1}, \\ A - p\alpha_i I_{\bar{W}M}, & n' = n = N, \\ \delta_{l\leq L} D_l \otimes \boldsymbol{\beta}, & n' = l,\ l = \overline{1,N},\ n = 0, \\ \tilde{D}_l, & n' = n + l,\ l = \overline{1, N-n},\ n = \overline{1,N}, \end{cases} \qquad i \geq 0.$$

Here, $\delta_{condition}$ is an indicator, which is equal to 1 if the condition is fulfilled, and is equal to 0 otherwise,

$$\tilde{D}_k = \delta_{k\leq L} D_k \otimes I_M,\ \ k \geq 1,$$
$$A = D_0 \oplus S + (1 - q_1 - q_2) I_{\bar{W}} \otimes \mathrm{diag}\{\mathbf{S}_2\} + q_2 I_{\bar{W}} \otimes \mathbf{S}_2 \boldsymbol{\beta},$$

$\oplus$ and $\otimes$ are symbols of the Kronecker sum and product of matrices, respectively, see, e.g., [23], $\bar{W} = W + 1$.

Proof of the lemma is implemented via the analysis of the intensities of transitions of the process ξ_t, $t \geq 0$.

Lemma 3. *The Markov chain ξ_t, $t \geq 0$, with the generator defined by Lemma 2 is AQTMC, see [10].*

Proof. To prove this lemma, it is required to verify that limits (5) for the Markov chain ξ_t with the generator defined by Lemma 2 exist.

It can be shown that if $p > 0$, the matrices $Y^{(n)}$, $n \geq 0$, have the following form:

$$Y^{(0)} = H,$$

$$Y^{(1)} = \begin{pmatrix} O & O & O & O & \dots & O & O \\ O & O & O & O & \dots & O & O \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & O & \dots & O & O \\ O & O & O & O & \dots & O & (1-p) I_{\bar{W}M} \end{pmatrix},\ Y^{(n)} = O,\ n = \overline{2, L+1}.$$

If $p = 0$, these matrices have the following form:

$$Y^{(0)} = \begin{pmatrix} O & I_{\bar{W}} \otimes \boldsymbol{\beta} & O & O & \dots & O & O \\ O & O & I_{\bar{W}M} & O & \dots & O & O \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & O & \dots & O & I_{\bar{W}M} \\ O & O & O & O & \dots & O & O \end{pmatrix}, \; Y^{(1)} = \begin{pmatrix} O & O & \dots & O & O \\ O & O & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \dots & O & O \\ O & O & \dots & Z^{-1}(I_{\bar{W}} \otimes \mathbf{S}_1\boldsymbol{\beta}) & Z^{-1}A + I \end{pmatrix},$$

$$Y^{(2)} = \begin{pmatrix} O & O & \dots & O & O \\ O & O & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \dots & O & O \\ O & O & \dots & Z^{-1}(q_1 I_{\bar{W}} \otimes \mathbf{S}_2\boldsymbol{\beta}) & Z^{-1}\tilde{D}_1 \end{pmatrix}, \; Y^{(n)} = \begin{pmatrix} O & O & \dots & O & O \\ O & O & \dots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \dots & O & O \\ O & O & \dots & O & Z^{-1}\tilde{D}_{n-1} \end{pmatrix}, \; n = \overline{3, L+1},$$

where Z is the diagonal matrix the diagonal entries of which are defined as the modulus of the corresponding diagonal entries of the matrix A. Therefore, for the Markov chain under study all the limits in the definition of *AQTMC* exist and, hence, this chain belongs to the class of *AQTMC*.

Theorem 1. *In the case $p = 0$, the Markov chain ξ_t, $t \geq 0$, is ergodic if the inequality*

$$\boldsymbol{\mu}\mathbf{S}_1 > \lambda$$

holds true where the vector $\boldsymbol{\mu}$ is the unique solution to the system

$$\boldsymbol{\mu}(S + \mathbf{S}_1\boldsymbol{\beta} + (1 - q_1 - q_2)\text{diag}\{\mathbf{S}_2\} + (q_1 + q_2) \otimes \mathbf{S}_2\boldsymbol{\beta}) = \mathbf{0}, \; \boldsymbol{\mu}\mathbf{e} = 1.$$

In the case $p > 0$, the Markov chain ξ_t, $t \geq 0$, is ergodic for all values of the system parameters.

Proof. To obtain the sufficient condition for the ergodicity of the Markov chain ξ_t we apply the theory of *AQTMC* from [10]. This condition is formulated in terms of the matrix generating function $Y(z) = \sum_{k=0}^{L} Y^{(k)} z^k$, $|z| \leq 1$. The condition has different form depending on whether or not the matrix $Y(z)$ is irreducible. In our case, we have more complicated case of the reducible matrix. In this case, the sufficient condition for the ergodicity has the following form. *AQTMC* is ergodic if, for all irreducible blocks $Y_l(z)$ of the canonical normal form of the matrix $Y(z)$ (for definition see [24]), the inequalities

$$\mathbf{y}_l \frac{dY_l(z)}{dz}|_{z=1}\mathbf{e} < 1 \tag{18}$$

are fulfilled where the vectors $\mathbf{y}_l$ are the unique solutions to the equations

$$\mathbf{y}_l Y_l(1) = \mathbf{y}_l, \; \mathbf{y}_l\mathbf{e} = 1. \tag{19}$$

The explicit form of the matrix generating function $Y(z)$ in the case $p = 0$ is as follows:

$$Y(z) = \begin{pmatrix} O & I_{\bar{W}} \otimes \boldsymbol{\beta} & O & O & \dots & O & O \\ O & O & I_{\bar{W}M} & O & \dots & O & O \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & O & O & \dots & O & I_{\bar{W}M} \\ O & O & O & O & \dots & Z^{-1}((I_{\bar{W}} \otimes \mathbf{S}_1\boldsymbol{\beta})z + q_1(I_{\bar{W}} \otimes \mathbf{S}_2\boldsymbol{\beta})z^2) & zI_{\bar{W}M} + Z^{-1}(Az + \sum_{k=1}^{L} \tilde{D}_k z^{k+1}) \end{pmatrix}.$$

By means of matched permutations of block rows and columns of the matrix $Y(z)$, it is possible to check that the single irreducible block $Y_1(z)$ of the matrix $Y(z)$, which participates in formulation of the ergodicity condition, has the following form:

$$Y_1(z) = \begin{pmatrix} O & I_{\bar{W}M} \\ Z^{-1}((I_{\bar{W}} \otimes \mathbf{S}_1\boldsymbol{\beta})z + q_1(I_{\bar{W}} \otimes \mathbf{S}_2\boldsymbol{\beta})z^2) & zI_{\bar{W}M} + Z^{-1}(Az + \sum_{k=1}^{L} \tilde{D}_k z^{k+1}) \end{pmatrix}.$$

For $z = 1$, this matrix has the form:

$$Y_1(1) = \begin{pmatrix} O & I_{\bar{W}M} \\ Z^{-1}X_1 & Z^{-1}X_2 + I \end{pmatrix}$$

where

$$X_1 = I_{\bar{W}} \otimes \mathbf{S}_1\boldsymbol{\beta} + q_1 I_{\bar{W}} \otimes \mathbf{S}_2\boldsymbol{\beta},$$
$$X_2 = D(1) \otimes I_M + I_{\bar{W}} \otimes S + (1 - q_1 - q_2)I_{\bar{W}} \otimes \text{diag}\{\mathbf{S}_2\} + q_2 I_{\bar{W}} \otimes \mathbf{S}_2\boldsymbol{\beta}.$$

Let us represent the vector $\mathbf{y}_1$, which is the solution to system (19), in the form

$$\mathbf{y}_1 = (\mathbf{y}^{(N-1)}, \mathbf{y}^{(N)}). \tag{20}$$

By substituting the vector $\mathbf{y}_1$ in form (20) into system (19), one can verify that the vector $\mathbf{y}^{(N)}Z^{-1}$ is the solution of the equation

$$\mathbf{y}^{(N)}Z^{-1}\Big(I_{\bar W} \otimes \mathbf{S}_1\boldsymbol{\beta} + q_1 I_{\bar W} \otimes \mathbf{S}_2\boldsymbol{\beta} + D(1) \otimes I_M + I_{\bar W} \otimes S + (1 - q_1 - q_2)I_{\bar W} \otimes \mathrm{diag}\{\mathbf{S}_2\} + q_2 I_{\bar W} \otimes \mathbf{S}_2\boldsymbol{\beta}\Big) = \mathbf{0},$$

or

$$\mathbf{y}^{(N)}Z^{-1}\Big(I_{\bar W} \otimes (\mathbf{S}_1\boldsymbol{\beta} + S + (1 - q_1 - q_2)\mathrm{diag}\{\mathbf{S}_2\} + (q_1 + q_2) \otimes \mathbf{S}_2\boldsymbol{\beta}) + D(1) \otimes I_M\Big) = \mathbf{0} \tag{21}$$

while the vector $\mathbf{y}^{(N-1)}$ is computed by

$$\mathbf{y}^{(N-1)} = \mathbf{y}^{(N)}Z^{-1}(I_{\bar W} \otimes (\mathbf{S}_1 + q_1\mathbf{S}_2)\boldsymbol{\beta}). \tag{22}$$

By the direct substitution into Eq. (21), it can be verified that the vector $\mathbf{y}^{(N)}Z^{-1}$ has the form

$$\mathbf{y}^{(N)}Z^{-1} = a(\boldsymbol{\theta} \otimes \boldsymbol{\mu}) \tag{23}$$

where a, $a > 0$, is a normalizing constant, $\boldsymbol{\theta}$ is the stationary distribution vector of the underlying process of the *BMAP*, and $\boldsymbol{\mu}$ is the unique solution to the system

$$\boldsymbol{\mu}(S + \mathbf{S}_1\boldsymbol{\beta} + (1 - q_1 - q_2)\mathrm{diag}\{\mathbf{S}_2\} + (q_1 + q_2) \otimes \mathbf{S}_2\boldsymbol{\beta}) = \mathbf{0},\ \boldsymbol{\mu}\mathbf{e} = 1.$$

From (22), we obtain

$$\mathbf{y}^{(N-1)} = a(\boldsymbol{\theta} \otimes \boldsymbol{\mu})(I_{\bar W} \otimes (\mathbf{S}_1 + q_1\mathbf{S}_2)\boldsymbol{\beta}). \tag{24}$$

Substituting the vector $\mathbf{y}_1$ in form (20) into inequality (18), it is possible to verify that this inequality can be transformed to the form

$$a(\boldsymbol{\theta} \otimes \boldsymbol{\mu})(I_{\bar W} \otimes (\mathbf{S}_1 + q_1\mathbf{S}_2)\boldsymbol{\beta})\mathbf{e} > a(\boldsymbol{\theta} \otimes \boldsymbol{\mu})(\sum_{k=1}^{L} kD_k \otimes I_M + q_1 I_{\bar W} \otimes \mathbf{S}_2\boldsymbol{\beta})\mathbf{e}.$$

After some algebra, using the mixed product rule for the Kronecker product of matrices (see [23]) we can rewrite this inequality into the form

$$(\boldsymbol{\theta} I_{\bar W} \otimes \boldsymbol{\mu}(\mathbf{S}_1 + q_1\mathbf{S}_2)\boldsymbol{\beta})(\mathbf{e}_{\bar W} \otimes \mathbf{e}_M) > (\boldsymbol{\theta}\sum_{n=1}^{L} nD_n \otimes \boldsymbol{\mu} I_M + q_1\boldsymbol{\theta} I_{\bar W} \otimes \boldsymbol{\mu}\mathbf{S}_2\boldsymbol{\beta})(\mathbf{e}_{\bar W} \otimes \mathbf{e}_M),$$

and, then,

$$\boldsymbol{\mu}(\mathbf{S}_1 + q_1\mathbf{S}_2\boldsymbol{\beta})\mathbf{e}_M > \lambda + \boldsymbol{\mu} q_1\mathbf{S}_2\boldsymbol{\beta}\mathbf{e}_M,$$

or

$$\boldsymbol{\mu}\mathbf{S}_1 > \lambda.$$

Let us now consider the case $p > 0$. In this case, the irreducible stochastic blocks of the matrix $Y(1)$ all are equal to scalar 1 and the inequalities of type (18) are reduced to the inequality $1 - p < 1$ that is always true in the considered case. The theorem is proved.

Remark 2. Condition $\boldsymbol{\mu}\mathbf{S}_1 > \lambda$ is easily tractable. The vector $\boldsymbol{\mu}$ defines the stationary distribution of the underlying process η_t of service when the system is overloaded. The number $\boldsymbol{\mu}\mathbf{S}_1$ defines the rate of customers departure from the system when it is overloaded. It is intuitively clear that the system is stable (the Markov chain describing its behavior is ergodic) if the rate of customers departure from the system when it is overloaded exceeds the arrival rate of customers λ.

Remark 3. As follows from [10], the condition

$$\boldsymbol{\mu}\mathbf{S}_1 < \lambda$$

is sufficient for the non-ergodicity of the Markov chain $\xi_t,\ t \geq 0$.

Let us assume that conditions of Theorem 1 are fulfilled. Then the following limits exist

$$\pi(i, 0, \nu) = \lim_{t\to\infty} P\{i_t = i,\ n_t = 0,\ \nu_t = \nu\},$$

$$\pi(i,n,\nu,m) = \lim_{t\to\infty} P\{i_t = i,\ n_t = n,\ \nu_t = \nu,\ \eta_t = m\},$$
$$i \geq 0,\ n = \overline{1,N},\ \nu = \overline{0,W}, m = \overline{1,M},$$

and are called as the stationary probabilities of the states of the Markov chain $\xi_t,\ t \geq 0$.

Let us form the row vectors $\boldsymbol{\pi}(i,n)$ $i \geq 0$, $n = \overline{1,N}$, of the probabilities $\pi(i,n,\nu,m)$ enumerated in the direct lexicographic order of the components (ν,m). Then, let us form the row vectors

$$\boldsymbol{\pi}_i = (\boldsymbol{\pi}(i,0), \boldsymbol{\pi}(i,1),\dots,\boldsymbol{\pi}(i,N)),\ i \geq 0.$$

To compute these vectors, we can apply one of the algorithms elaborated in Section 3. Then, we can compute the key performance measures of the system.

4.4. Performance measures

The average number N_{system} of customers in the system (in the server and in the buffer) is computed by

$$N_{system} = \sum_{i=0}^{\infty}\sum_{n=1}^{N} n\boldsymbol{\pi}(i,n)\mathbf{e}.$$

The average number L_{orbit} of customers in the orbit is computed by

$$L_{orbit} = \sum_{i=1}^{\infty} i\boldsymbol{\pi}_i\mathbf{e}.$$

The average intensity λ_{out} of the flow of customers, who successfully obtain service, is computed by

$$\lambda_{out} = \sum_{i=0}^{\infty}\sum_{n=1}^{N} \boldsymbol{\pi}(i,n)(I_{\bar{W}} \otimes \mathbf{S}_1\boldsymbol{\beta})\mathbf{e}.$$

The loss probability P_{loss} of an arbitrary customer is computed by

$$P_{loss} = 1 - \frac{\lambda_{out}}{\lambda}.$$

4.5. Numerical examples

For computations, we use a PC with an Intel Core i7-8700 CPU and 16 GB RAM.

Example 1. Effectiveness of old and new algorithms

To construct the *BMAP*, we use the *MAP* defined by the matrices

$$D_0 = \begin{pmatrix} -25.53984 & 0.393329 & 0.361199 \\ 0.14515 & -2.2322 & 0.200007 \\ 0.295961 & 0.387445 & -1.752618 \end{pmatrix},\ D_1 = D = \begin{pmatrix} 24.24212 & 0.466868 & 0.076324 \\ 0.034097 & 1.666864 & 0.186082 \\ 0.009046 & 0.255481 & 0.804685 \end{pmatrix}. \tag{25}$$

This arrival process has the average arrival rate $\lambda = 5$, the coefficient of correlation of two successive intervals between arrivals $c_{cor} = 0.3$, and the squared coefficient of variation of the intervals between customer arrivals $c_{var} = 2$.

We assume that the number of customers in a batch has a truncated geometric distribution, i.e., the probability that an arbitrary batch consists of k customers is equal to $q(k) = q^{k-1}\frac{1-q}{1-q^L}$, $k = \overline{1,L}$, where q is the parameter of the distribution and $L = 5$. Based on the described *MAP*, we construct the *BMAP* as follows. We set the matrix D_0 the same as above. The matrices $D_k,\ k = \overline{1,5}$, are initially defined by $D_k = Dq(k),\ k = \overline{1,5}$, where the matrix D is given in (25) and $q = 0.8$. After that, we multiply all the matrices $D_k,\ k = \overline{0,5}$, by the factor $\frac{5}{12.8153}$ to obtain the *BMAP* having the fundamental intensity $\lambda = 5$.

The service time of a customer has a *PHF* distribution with the following parameters:

$$\boldsymbol{\beta} = (0.5, 0.3, 0.2),$$
$$S = \begin{pmatrix} -13 & 2 & 1 \\ 0 & -12 & 2 \\ 1 & 1 & -15 \end{pmatrix},$$
$$\mathbf{S}_1 = (9.5, 9.8, 12.9)^T,\ \mathbf{S}_2 = (0.5, 0.2, 0.1)^T,\ q_1 = 0.5,\ q_2 = 0.3.$$

Under these parameters, the probability of successful completion of an arbitrary service is equal to 0.971262. Correspondingly, the probability of interruption of an arbitrary service is equal to 0.028738.

The individual intensity of customers' retrial is $\alpha = 0.5$, the total retrial intensity when i customers stay in the orbit is $\alpha_i = i\alpha,\ i \geq 0$. The probability that the customer leaves the system after unsuccessful retrial is $p = 0.1$.

Table 1
Information related to the implementation of the known algorithm.

N	*Size*	n_G	n_π	Runtime
2	21	341 600	402	242 s
3	30	342 000	399	420 s
4	39	342 200	398	670 s
5	48	342 200	397	981 s
6	57	342 200	397	1318 s
7	66	342 200	397	1815 s
8	75	342 200	397	2290 s
9	84	342 200	397	2842 s
10	93	342 200	397	3453 s

Table 2
Information related to the implementation of Algorithm 2.

N	*Size*	n_G	n_π	Runtime
2	21	641	409	1.6 s
3	30	641	407	2.4 s
4	39	641	406	3.5 s
5	48	641	406	5.4 s
6	57	641	406	7.3 s
7	66	641	406	9.5 s
8	75	641	405	11.7 s
9	84	641	405	14.7 s
10	93	641	405	18 s
15	138	641	403	37 s
20	183	641	402	60 s
30	273	641	399	81 s
40	363	641	397	154 s
50	453	641	395	267 s

Let us fix $\epsilon_G = 10^{-3}$ and $\epsilon_F = 10^{-11}$ and use the known algorithm for *AQTMC* from [10]. We compute the stationary distribution of the system with capacity N varying in the interval $[2, 10]$. The information related to the implementation of the known algorithm (the size of generator's blocks $Size = \bar{W}(1 + NM)$, the number n_G of computed matrices G_i, the number n_π of computed stationary probabilities vectors, and the runtime) is presented in Table 1.

Now, we vary the system capacity N over the interval $[1, 50]$ and compute the performance measures of the system based on Algorithm 2. All the constants in the algorithm are assumed the same as in Section 3.

The information related to the implementation of Algorithm 2 for each number N is presented in Table 2.

Comparing Tables 1 and 2, we observe a huge advantage of the new algorithm with respect to the runtime. E.g., for the capacity $N = 2$ of the system the new algorithm is 151 times faster and requires the computation of 532 times smaller number of matrices G_i. For the capacity $N = 10$ of the system the new algorithm is 191 times faster. The runtime is only 18 s, while the known algorithm requires about 58 min. The new algorithm requires a bit more than 4 min for $N = 50$. While the known algorithm has essential difficulties with available computer memory even for $N = 11$.

Note, that all the data presented in Table 2 are obtained for values $i_0 = s = 100$ that are default fixed in the algorithm as the initial values. Taking into account the features of the queueing model, to which the algorithm is applied, one could be more flexible in choosing the initial values. It is seen from Table 2 that for $N = 2$ the number of non-negligible probability vectors $\boldsymbol{\pi}_i$ is about 410. It is clear that the increase of the buffer size $N - 1$ causes the decrease in the number of customers in the orbit. Thus, for $N > 2$ the number of non-negligible probability vectors $\boldsymbol{\pi}_i$ is not larger than 410. Therefore, it is reasonable to increase the initial values of i_0 and s in the computation of the stationary distribution of the system states for $N > 2$. E.g., if one chooses $i_0 = 400$ and $s = 150$, the runtime for $N = 50$ will be about 141, i.e., almost twice less than in Table 2.

Example 2. Illustration of the effect of batch arrivals

The necessity of the account of correlation and variation of inter-arrival times in the *BMAP* is quite well illustrated in the literature. Therefore, we will briefly illustrate the effect of the batch arrivals. To build *BMAP*s with various batch size distributions, we again use the *MAP* defined by formula (25). The number of customers in a batch has a truncated geometric distribution. We consider the following three different values of the parameter q of this distribution: 0.2, 0.5 and 0.8. The maximal batch size is 5. For each fixed q, we scale the matrices $D_k,\ k = \overline{0, 5}$, to get the same fundamental rate $\lambda = 5$ for all these *BMAP*s. Note, that the larger is the parameter q the larger is the average batch size (it is equal to 1.2484, 1.8387 and 2.5630 for q equal to 0.2, 0.5 and 0.8, correspondingly). Because we set the fundamental rates for these *BMAP*s equal, a smaller number of q implies that the customers arrive more frequently, but in batches of smaller size.

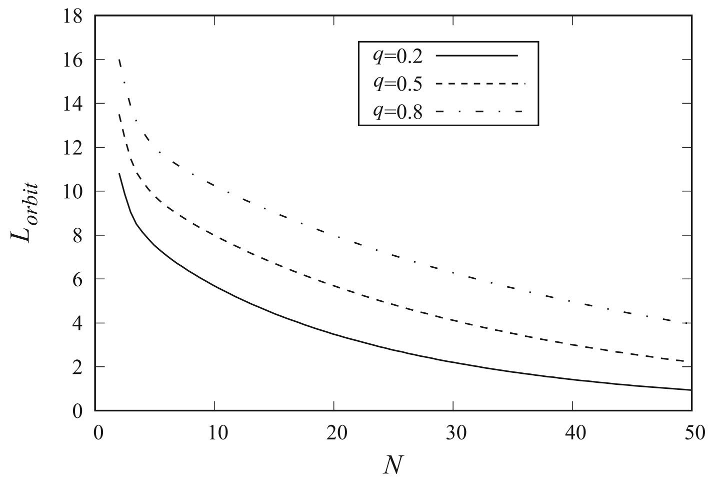

Fig. 2. Dependence of the average number L_{orbit} of customers in the orbit on the value N for various values of q.

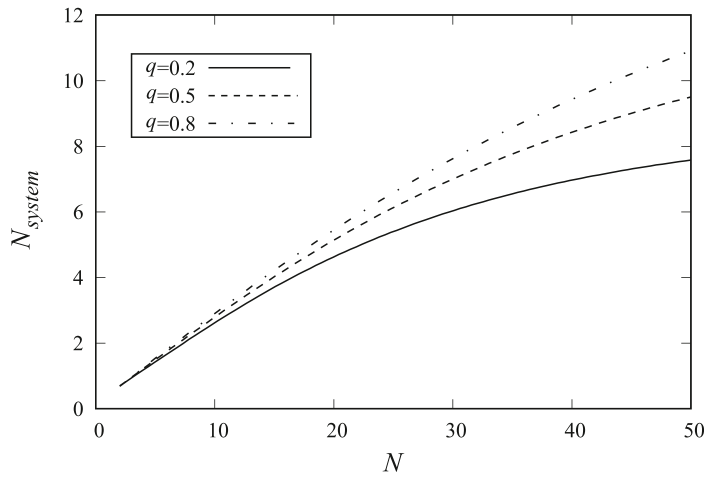

Fig. 3. Dependence of the average number N_{system} of customers in the system on the value N for various values of q.

Figs. 2–4 illustrate the dependence of the average number L_{orbit} of customers in the orbit, the average number of customers in the system N_{system} (excluding the number of customers in the orbit) and the probability of an arbitrary customer loss P_{loss} on the values of N for various values of q.

It is easy to see from these figures, that the performance measures of the system are better for smaller values of q. This is explained by the made above conclusion that smaller value of q implies more regular arrival of customers (more frequent and in smaller batches) and, therefore, higher chances of an arbitrary customer to avoid visiting the orbit. The last fact is evident from the following observations. When $q = 0.2$, the average batch size is 1.2484 and, if the arriving batch sees the server idle, with probability $\frac{1}{1.2484} = 0.8$ an arbitrary customer succeeds to start service immediately. In the case $q = 0.8$, the average batch size is 2.5630 and the corresponding probability is $\frac{1}{2.5630} = 0.39$.

5. Conclusion

In this paper, new numerically stable algorithms for computation of the stationary distribution of ergodic multi-dimensional Markov chains with the level-dependent structure of the generator are developed. To demonstrate the application of these algorithms, we analyze a novel single-server retrial queue with the batch Markov arrival process and a finite buffer. The total retrial intensity arbitrarily depends on the number of customers in the orbit. Customers in the orbit may be non-persistent and depart from the orbit without service after each unsuccessful attempt. The server is unreliable. Service consists of a random number of phases. Service can be interrupted what implies the transition of the customer to the orbit or service repetition from the beginning or from the phase at which the interruption occurred. For this system, the ergodicity condition is derived. Under fulfillment of this condition, the stationary distribution of a

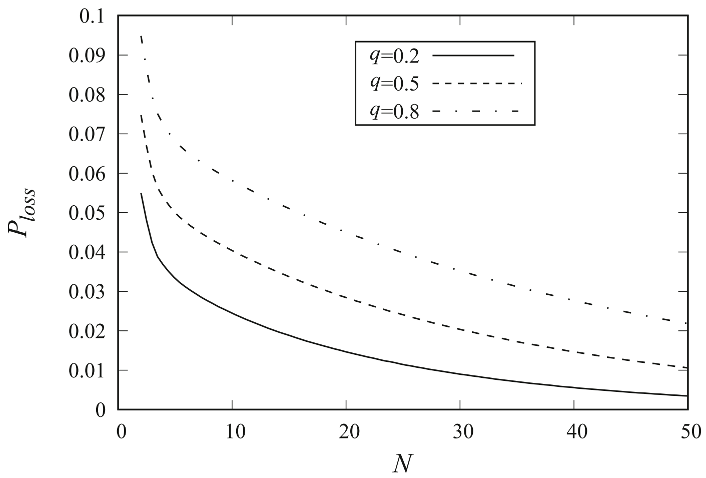

Fig. 4. Dependence of the probability of an arbitrary customer loss P_{loss} on the value N for various values of q.

multi-dimensional Markov chain that describes the behavior of the system is computed via the known algorithm and the algorithm developed in this paper. The essential advantages of the new algorithm in terms of runtime and requirements to computer memory over the known algorithm are numerically demonstrated.

Acknowledgments

The publication has been prepared with the support of the "RUDN University Program, Russia 5-100" (participants Dudin S.A. and Dudin A.N., new algorithm development and queueing model analysis). The publication was supported by grant F18MV-003 of Belarusian Republican Foundation for Fundamental Research (participant Dudina O.S., modified new algorithm development and numerical experiments).

References

[1] J.R. Artalejo, A. Gomez-Corral, Retrial Queueing Systems: A Computational Approach, Springer, Berlin, Heidelberg, 2008.
[2] I.M.K. Atencia, J.L. Galán-García, G. Aguilera-Venegas, P. Rodríguez-Cielos, M. Galán-García, A $Geo^{[X]}/G^{[X]}/1$ retrial queueing system with removal work and total renewal discipline, Appl. Math. Comput. 319 (2018) 245–253.
[3] G.I. Falin, J.G.C. Templeton, Retrial Queues, Chapman & Hall, London, 1997.
[4] W. Feller, An Introduction to Probability Theory and its Applications, Vol. 1, John Wiley & Sons Inc., 1957.
[5] D.A. Bini, G. Latouche, B. Meini, Numerical Methods for Structured Markov Chains, Oxford University, 2005.
[6] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modelling, first ed., ASA SIAM, 1999.
[7] M. Neuts, Matrix-Geometric Solutions in Stochastic Models — An Algorithmic Approach, John Hopkins University Press, 1981.
[8] L. Bright, P.G. Taylor, Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes, Comm. Stat. Stoch. Models 11 (1995) 497–525.
[9] S.A. Dudin, O.S. Dudina, Retrial multi-server queueing system with PHF service time distribution as a model of a channel with unreliable transmission of information, Appl. Math. Model. 65 (2019) 676–695.
[10] V.I. Klimenok, A.N. Dudin, Multi-dimensional asymptotically quasi-Toeplitz Markov chains and their application in queueing theory, Queueing Syst. 54 (2006) 245–259.
[11] M.F. Neuts, Structured Stochastic Matrices of M/G/1 Type and their Applications, Marcel Dekker, New York, 1989.
[12] J.G. Kemeni, J.L. Snell, A.W. Knapp, Denumerable Markov Chains, Van Nostrand, New York, 1966.
[13] Q. Ye, High accuracy algorithms for solving nonlinear matrix equations in queueing models, in: Advances in Algorithmic Methods for Stochastic Models, Notable Publications, NJ, 2000.
[14] D.M. Lucantoni, New results on the single server queue with a batch Markovian arrival process, Comm. Stat. Stoch. Models 7 (1991) 1–46.
[15] V. Ramaswami, The $N/G/1$ queue and its detailed analysis, Adv. Appl. Probab. 12 (1980) 222–261.
[16] A. Dudin, V. Klimenok, Queueing system $BMAP/G/1$ with repeated calls, Math. Comput. Modelling 30 (1999) 115–128.
[17] V.I. Klimenok, D.S. Orlovsky, C.S. Kim, The $BMAP/PH/N$ retrial queue with Markovian flow of breakdowns, European J. Oper. Res. 189 (2008) 1057–1072.
[18] Q.-L. Li, Y. Ying, Y.Q. Zhao, A $BMAP/G/1$ retrial queue with a server subject to breakdowns and repairs, Ann. Oper. Res. 141 (2006) 233–270.
[19] S.R. Chakravarthy, The batch Markovian arrival process: a review and future work, in: A. Krishnamoorthy, N. Raju, V. Ramaswami (Eds.), Advances in Probability Theory and Stochastic Processes, Notable Publications Inc., New Jersey, 2001, pp. 21–29.
[20] Q.M. He, H. Li, Y.Q. Zhao, Ergodicity of the $BMAP/PH/S/S+K$ retrial queue with PH-retrial times, Queueing Syst. 35 (2000) 323–347.
[21] A. Krishnamoorthy, P. Pramod, S. Chakravarthy, A survey on queues with interruptions, TOP 22 (2014) 290–320.
[22] A. Dudin, S. Dudin, Analysis of a priority queue with phase-type service and failures, Int. J. Stoch. Anal. (2016) 1–11, Article ID 9152701.
[23] A. Graham, Kronecker Products and Matrix Calculus with Applications, Ellis Horwood, Cichester, 1981.
[24] F.R. Gantmakher, The Matrix Theory, Science, Moscow, 1967 (in Russian).