ISSN 1932-2321

GNEDENKO FORUM
INTERNATIONAL GROUP ON RELIABILITY

Reliability: Theory and Applications

ELECTRONIC JOURNAL OF INTERNATIONAL GROUP ON RELIABILITY
JOURNAL IS REGISTERED IN THE LIBRARY OF THE U.S. CONGRESS

Special Issue 8 (85) July 2025

Reliability and Performance in Stochastic Models

SELECTED ARTICLES

Edited by

Oleg Lukashenko
Evsey Morozov
Ruslana Nekrasova
Alexander Rumyantsev

Journal of International Group on Reliability

Reliability: Theory and Applications

Vol.20 Special Issue No.8 (85),
July 2025

Reliability and Performance in Stochastic Models

SELECTED ARTICLES

Edited by

Oleg Lukashenko,
Evsey Morozov,
Ruslana Nekrasova,
Alexander Rumyantsev

San Diego
2025

Preface

This Special Issue of the Reliability Theory and Applications journal is a collection of original articles dedicated to various aspects of reliability and performance in stochastic models. The topics of the papers cover a relatively wide area of subjects, from theoretic analysis to simulation, and the range of models include queues, queueing-inventory, production-inventory, reliability and retrial systems as well as networks.

The volume was edited by the organizers of SMARTY event series, and we thank the anonymous referees for their valuable time, as well as the respected authors for their contribution. Our special thanks to the members of Gnedenko Forum and Editors of the Electronic Journal Reliability: Theory & Applications (ISSN 1932-2321), for their valuable support in making this Special Issue a reality.

Special Issue editors,

Oleg Lukashenko,

Evsey Morozov,

Ruslana Nekrasova,

Alexander Rumyantsev

Table of Contents

A PRODUCTION INVENTORY MODEL WITH PROTECTION FOR FEW STAGES OF PRODUCTION

Nisha Mathew[1], V.C Joshua[2], A Krishnamoorthy[3], Ambily P. Mathew[2,*]

•

[1]Department of Mathematics, B.K College Amalagiri,Kottayam, India
[2]Department of Mathematics, CMS College Kottayam, India
[3] Centre for Research in Mathematics, CMS College Kottayam, India
nishatmathew@gmail.com vcjoshua@cmscollege.ac.in
achyuthacusat@gmail.com
* corresponding author: ambilypm@cmscollege.ac.in

Abstract

We consider a two server production inventory model with positive service time. Customers arrive to the system according to a Markovian Arrival Process. Service time of customers follow identical but independent phase type distribution. The production of inventory follows (s, S) policy. Production of inventory is by one unit at a time and the production time follows Erlang distribution. While in production shocks occur and consequently breakdown of the production machinery takes place. The shock/damage process occurs according to a Poisson process. After repair, the production process restarts, discarding the item in production. The repair time follows phase type distribution. In order to minimize the product loss due to shock, protection is given to the last k stages of production. Protection of the production process involves additional cost. As a result of this protection, the item, while in the last k stages of production, will not be affected by shocks. Steady state analysis of the model is performed. Some performance measures and distributions of certain important performance characteristics are evaluated. We formulate an optimization problem related to the number of stages of the production process to be protected.

Keywords: production inventory, protection, phase type distribution, Erlang distribution, Markovian Arrival Process

1. Introduction

Inventory with positive service time was first introduced by Melikov and Molchanov[1] and Sigman and Simchi Levi[2], independently of each other. After that a lot of works have been carried out in this area. A survey of inventory with positive service time is given in Krishnamoorthy et al.[3]. Attached to production inventory, positive service time was introduced in Krishnamoorthy et al. [4]. They considered a production inventory system with a single server and server vacation. In that the customer arrival process was assumed to follow a Markovian arrival process(MAP), and the time for producing each item was assumed to have Markovian production scheme. The service process and vacation durations follow independent phase type distributions. Krishnamoorthy et al. in [5] considered two (s, S) production inventory systems with positive service time. Here arrival of customers was according to a Poisson process and service and production time follow independent exponential distributions. In both models, the steady state distributions were obtained in product form. In [6], Baek et al. studied an $M/M/1$ queue with an attached production inventory system. Along with an internal production process

following a Poisson process, they considered the (r, Q) inventory control policy. In [6], the customers arriving during stock out period are considered as lost. In [7], Krishnamoorthy et al. introduced the idea of protection in a queueing system where the service process is subject to interruptions. Here the last $m - n$ phases of the Erlang service process were given protection from interruption. Krishnamoorthy et al. in [8] considered (s, S) production inventory system with positive service time and interruptions. Here the customers arrive according to a Poisson process, time for producing each item and service time to each customer follow independent Erlang distributions. Both service process and the production process were subjected to interruptions and certain number of phases in both these processes were given protection from interruption. Anoop and Jacob in [9] considered a queueing system where arrival of customers follows Poisson process. They considered the (s, S) production inventory as servers of the queueing system and the service time follows an exponential distribution. In [10] Yue and Qin considered a production inventory system with positive service times and vacations to the production facility. Customers arrive to the system according to a Poisson process, the service times are exponentially distributed, and has a single production facility that produces one type of product, whose production times are exponentially distributed. Baek et al. in [11] considered an (s, S) production inventory system with an attached Markovian service queue with c servers. In [11], customers leave the system with exactly one item at the service completion epochs and the customers arriving during stock out period are considered as lost. Yue and Qin in [12] considered a production inventory system with service time and product returns, dependent on the characteristics of online shopping behaviors. Here the customers arrive according to a Poisson process, and the service time is exponentially distributed. Jose et al. in [13] considered a single server perishable inventory system in which customers arrive according to a Poisson process. In [13], when a customer arrives if the server is available with a positive level of inventory, that customer enters service. Otherwise, the customer goes to a orbit of infinite capacity with pre determined probability or exits the system with complementary probability. Each customer in the orbit tries to access the server in an exponentially distributed time interval and after every unsuccessful retrial, the customer returns to the orbit with a pre allotted probability or is lost forever with complementary probability.

The highlights of our model are:

- In this model, the production process is subject to shocks which interrupts the production. When the production process is interrupted due to shock, the item in production is lost.
- In order to reduce the effect of loss due to shock, protection is given to the last k stages of production
- The item which is lost due to interruption is sold at scrap value, which is much lower than the selling price of finished goods.
- The cost function is dependent on the number of phases of protection.

In real life, there are many situations in which the production process is subjected to some shock or damage. In such cases, in order to reduce the loss, protection may be given to the production process. For example, suppose the production requires uninterrupted power supply. In case of a power failure, the item in production will be lost. So in order to reduce such loss due to power supply, a protection is provided in the form of a backup power like generator, battery etc. But providing protection requires extra cost. Since economic viability is not feasible for a full time protection, it is provided only in last few stages where major loss is incurred. Our model gains its motivation from such production processes. We encounter many situations in manufacturing which can be modeled as multi server models. Here we consider a two server queueing inventory model, as all the works related considered so far in literature are related to single server models.

The remaining sections of the paper are organized as follows: in Section 2, the model is described. Mathematical formulation of the model is done in Section 3. Steady-state analysis of the model is done in section 4. In section 5 some performance measures are evaluated. In section 6, we derive the expected length of a production period. In section 7, some numerical

and graphical illustrations are done. Cost analysis is done in section 8. We conclude the paper in section 9.

2. Model Description

We consider a two server queueing inventory system with production of items to be given to customers at the end of service. Customers arrive to the system according to a Markovian Arrival Process(MAP) with representation (D_0, D_1) of order m. When a customer arrives, if both the servers are idle and there is at least one item in inventory, that customer is served by the first server. When a customer arrives, if only one server is idle and there is at least one item in inventory, that customer is served by the idle server. Otherwise they wait in queue. Backlog is allowed regardless of the inventory/production state. The customers are served one by one on a first come first served(FCFS) basis. The service time of customers by both the servers follow phase type distribution with representation PH$(\boldsymbol{\alpha}, T)$ with m_1 phases, where T^0 is such that $T\boldsymbol{e} + T^0 = \mathbf{0}$. When a customer leaves the system after service completion, the inventory level drops by one unit.

The production of inventory follows (s, S) policy and production is by one unit at a time. The production process is turned on when the inventory level falls to s and it is turned off only when the inventory level reaches S. Production time of each item follows Erlang (β, W) distribution of order m_2 and parameter θ. Here β is the initial probability vector corresponding to this Erlang distribution. Production process is subject to shocks which interrupts the process. Shock/damage process strikes the production machinery only while the production process is on. The arrival of shock process is Poisson with parameter ν. The production machinery fails with the arrival of shock/damage process. When shock/damage occurs, the item being produced is lost. After repair, the production restarts. The repair time of the production machine follows phase type distribution with representation PH$(\boldsymbol{\gamma}, U)$ with m_3 phases,where U^0 is such that $U\boldsymbol{e} + U^0 = \mathbf{0}$.

When the production process is interrupted due to shock, the item in production is lost and so sold at scrap value, which is lower than the selling price of finished goods. So to reduce the effect of loss due to shock, protection is given to the last k stages of production. This means that if the production process is in the last k stages, protection is provided to the machine against shocks. Thus in the last k stages, it is protected from being interrupted. The first $m_2 - k$ phases of production are unprotected. The protection is given at extra cost. The optimal value of k is investigated for an appropriate cost function.

3. Mathematical Formulation

We introduce the following notations:

- $N(t)$: the number of customers in the system at time t.
- $I(t)$: the number of items in the inventory at time t.
- $P(t)$: the status of production at time t.

$$P(t) = \begin{cases} 0 \text{ production process is off} \\ 1 \text{ production process is on} \\ 2 \text{ production machine is under repair} \end{cases}$$

- $C(t)$: the protection status at time t.

$$C(t) = \begin{cases} 0, \text{ production process is not protected} \\ 1, \text{ production process is protected} \end{cases}$$

- $J_1(t)$: the phase of the service process of server-1 at time t.

- $J_2(t)$: the phase of the service process of server-2 at time t.
- $J_3(t)$: the phase of the production process at time t.
- $R(t)$: the phase of the repair process at time t.
- $A(t)$: the phase of the arrival process at time t.

Then $\{(N(t), I(t), P(t), C(t), J_1(t), J_2(t), J_3(t), R(t), A(t)); t \geq 0\}$ is a continuous time Markov chain on the state space $\Omega = \bigcup_{n=0}^{\infty} l(n)$, where $l(n)$ denotes the collection of states in level n and are defined by considering the first three states as $l(0) = (0,i), l(1) = (1,i)$ and for $n \geq 2$, $l(n) = (n,i)$. Let the ordering of the elements of Ω be lexiographical.

$(0,i) = \{(0,i,1) \bigcup (0,i,2)/0 \leq i \leq s\} \bigcup \{(0,i,0) \bigcup (0,i,1)) \bigcup (0,i,2)/s+1 \leq i \leq S-1\} \bigcup (0,S,0)$.
$(1,i) = (1,0,1) \bigcup (1,0,2) \bigcup \{(1,i,1) \bigcup (1,i,2)/1 \leq i \leq s\}$
$\bigcup \{(1,i,0) \bigcup (1,i,1) \bigcup (1,i,2)/s+1 \leq i \leq S-1\} \bigcup (1,S,0)$.
$(n,i) = (n,0,1) \bigcup (n,0,2) \bigcup (n,1,1) \bigcup (n,1,2) \bigcup \{(n,i,1) \bigcup (n,i,2)/2 \leq i \leq s\} \bigcup$
$\{(n,i,0) \bigcup (n,i,1) \bigcup (n,i,2)/s+1 \leq i \leq S-1\} \bigcup (n,S,0)$.

The transitions, by considering the first three states, are given in the tables 1,2,3,4,5,6 and 7.

The infinitesimal generator Q of the LIQBD (Level Independent Quasi Birth Death) process describing the above two server queueing inventory system is of the form

$$Q = \begin{pmatrix} B_{00} & B_{01} & O & \dots & \dots & \dots & \dots & \dots \\ B_{10} & B_{11} & B_{12} & O & \dots & \dots & \dots & \dots \\ O & B_{21} & A_1 & A_0 & O & \dots & \dots & \dots \\ O & O & A_2 & A_1 & A_0 & O & \dots & \dots \\ O & O & O & A_2 & A_1 & A_0 & O & \dots \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \end{pmatrix}.$$

As there are two servers, the behaviour of the system is different for levels 0 and 1. Level 0 corresponds to no customer in the system and level 1 corresponds to 1 customer in the system and hence in service. At level 0, both servers are idle and at level 1, one of the servers is idle. But from level 2 onwards, there are always two or more customers in the system, the two of them are in service. So from level 2 onwards, both servers will be busy, provided there are at least two inventory in the system. So the first three levels are boundary levels and after that the structure gets repeated resulting in a quasi-Toeplitz matrix. The structure of the matrices $B_{00}, B_{01}, B_{10}, B_{11}, B_{12}, B_{21}, A_0, A_1$ and A_2 are given in Appendix A.

4. Steady-State Analysis

For the steady state analysis, we use the Matrix Geometric method by Neuts [14].

4.1. Stability Condition

Theorem-1: The Markov chain with generator Q is stable if and only if

$$\boldsymbol{\pi}[I_{K_4} \otimes D_1]\boldsymbol{e} < \boldsymbol{\pi}_1 H_1 \boldsymbol{e} + \boldsymbol{\pi}_2 H_5 \boldsymbol{e} + \sum_{j=3}^{s} \boldsymbol{\pi}_j H_9 \boldsymbol{e} + \boldsymbol{\pi}_{s+1} H_{10} \boldsymbol{e} + \sum_{j=s+2}^{S-1} \boldsymbol{\pi}_j H_{11} \boldsymbol{e} + \boldsymbol{\pi}_S H_{12} \boldsymbol{e}. \quad (1)$$

Proof. Let $A = A_0 + A_1 + A_2$. Then A is an irreducible matrix and the stationary vector $\boldsymbol{\pi}$ of A is obtained by solving

$$\boldsymbol{\pi} A = 0; \boldsymbol{\pi} \boldsymbol{e} = 1.$$

The Markov chain with generator Q is stable if and only if

$$\boldsymbol{\pi} A_0 e < \boldsymbol{\pi} A_2 \boldsymbol{e},$$

i.e, the system is stable if and only if

$$\boldsymbol{\pi}[I_{K_4} \otimes D_1]\boldsymbol{e} < \boldsymbol{\pi}_1 H_1 \boldsymbol{e} + \boldsymbol{\pi}_2 H_5 \boldsymbol{e} + \sum_{j=3}^{s} \boldsymbol{\pi}_j H_9 \boldsymbol{e} + \boldsymbol{\pi}_{s+1} H_{10} \boldsymbol{e} + \sum_{j=s+2}^{S-1} \boldsymbol{\pi}_j H_{11} \boldsymbol{e} + \boldsymbol{\pi}_S H_{12} \boldsymbol{e}.$$

■

4.2. Stationary Distribution

The stationary distribution of the Markov process under consideration is obtained by solving the set of equations

$$\mathbf{x} Q = \mathbf{0}; \mathbf{x}\mathbf{e} = \mathbf{1}. \tag{2}$$

Let $\mathbf{x}$ be decomposed in conformity with Q. Then
$\mathbf{x} = (\mathbf{x_0}, \mathbf{x_1}, \mathbf{x_2}, \dots)$, where $\mathbf{x_i} = (\mathbf{x_{i0}}, \mathbf{x_{i1}}, \dots \dots \mathbf{x_{iS}})$, for $i = 0, 1, 2, \dots$.
$\mathbf{x_{ij}} = (\mathbf{x_{ij1}}, \mathbf{x_{ij2}})$, for $j = 0, 1, 2, \dots, s$,
$\mathbf{x_{ij}} = (\mathbf{x_{ij0}}, \mathbf{x_{ij1}}, \mathbf{x_{ij2}})$, for $j = s+1, 2, \dots, S-1$ and
$\mathbf{x_{iS}} = (\mathbf{x_{iS0}})$.
For $j = 0, 1, 2, \dots, S-1$,

$$\mathbf{x_{ij1}} = (\mathbf{x_{ij10}}, \mathbf{x_{ij11}}).$$

From $\mathbf{x}Q = 0$, we get the following equations:

$$\mathbf{x_0} B_{00} + \mathbf{x_1} B_{10} = 0, \tag{3}$$

$$\mathbf{x_0} B_{01} + \mathbf{x_1} B_{11} + \mathbf{x_2} B_{21} = 0, \tag{4}$$

$$\mathbf{x_1} B_{12} + \mathbf{x_2} A_1 + \mathbf{x_3} A_2 = 0, \tag{5}$$

$$\mathbf{x_{i-1}} A_0 + \mathbf{x_i} A_1 + \mathbf{x_{i+1}} A_2 = 0, i = 3, 4, \dots. \tag{6}$$

It may be shown that there exists a constant matrix R such that

$$\mathbf{x_i} = \mathbf{x_{i-1}} R, i = 3, 4, \dots. \tag{7}$$

The sub vectors $\mathbf{x_i}$ are geometrically related by the equation

$$\mathbf{x_i} = \mathbf{x_2} R^{i-1}, i = 3, 4, \dots. \tag{8}$$

R is the minimal non negative solution to the matrix quadratic equation

$$R^2 A_2 + R A_1 + A_0 = O. \tag{9}$$

5. Performance Measures

In this section we evaluate a few performance characteristics of the system that are essential for computation of optimal value of k.

1. Expected number of customers in the system:

$$E[N] = \sum_{i=1}^{\infty} i \mathbf{x_i} \mathbf{e}. \tag{10}$$

2. Expected inventory level:

$$E[I] = \sum_{i=0}^{\infty} \sum_{j=1}^{S} j\mathbf{x_{ij}e}. \tag{11}$$

3. Expected number of customers waiting in the system due to lack of inventory:

$$E[W] = \sum_{i=2}^{\infty} (i-1)\mathbf{x_{i1}e} + \sum_{i=1}^{\infty} i\mathbf{x_{i0}e}. \tag{12}$$

4. Probability that both servers are idle:

$$b_0 = \sum_{i=1}^{\infty} \mathbf{x_{i0}e} + \sum_{\mathbf{j=0}}^{\mathbf{S}} \mathbf{x_{0j}e}. \tag{13}$$

5. Probability that one of the server is busy and the other idle:

$$b_1 = \sum_{i=1}^{\infty} \mathbf{x_{i1}e} + \sum_{\mathbf{j=2}}^{\mathbf{S}} \mathbf{x_{1j}e}. \tag{14}$$

6. Probability that both servers are busy:

$$b_2 = \sum_{i=2}^{\infty} \sum_{j=2}^{S} \mathbf{x_{ij}e}. \tag{15}$$

7. Probability that production process is off:

$$p_0 = \sum_{i=0}^{\infty} \sum_{j=s+1}^{S} \mathbf{x_{ij0}e}. \tag{16}$$

8. Probability that production process is on:

$$p_1 = \sum_{i=0}^{\infty} \sum_{j=0}^{S-1} \mathbf{x_{ij1}e}. \tag{17}$$

9. Probability that production machine is under repair:

$$p_2 = \sum_{i=0}^{\infty} \sum_{j=0}^{S-1} \mathbf{x_{ij2}e}. \tag{18}$$

10. Probability that production process is on and is protected:

$$p_p = \sum_{i=0}^{\infty} \sum_{j=0}^{S-1} \mathbf{x_{ij11}e}. \tag{19}$$

11. Probability that production process is on and is not protected:

$$p_u = \sum_{i=0}^{\infty} \sum_{j=0}^{S-1} \mathbf{x_{ij10}e}. \tag{20}$$

12. Probability that production process is on, is not protected and is in the r^{th} phase of production:

$$p_r = \sum_{i=0}^{\infty} \sum_{j=0}^{S-1} \sum_{s_1=0}^{m_1} \sum_{s_2=0}^{m_1} \mathbf{x_{ij10s_1s_2r}e}. \tag{21}$$

13. Expected rate of interruptions of the production process in unit time:

$$E[RI] = \nu \sum_{i=0}^{\infty} \sum_{j=0}^{S-1} \mathbf{x_{ij10}} \mathbf{e}. \tag{22}$$

14. Probability that inventory level is zero:

$$I_0 = \sum_{i=0}^{\infty} \mathbf{x_{i0}} \mathbf{e}. \tag{23}$$

15. The probability that the inventory level is less than or equal to s:

$$p_s = \sum_{i=0}^{\infty} \sum_{j=0}^{s} \mathbf{x_{ij}} \mathbf{e}. \tag{24}$$

16. The probability that the maximum storage capacity S is utilized:

$$p_S = \sum_{i=0}^{\infty} \mathbf{x_{iS}} \mathbf{e}. \tag{25}$$

6. Expected length of a production period

Here, we derive the expected length of a production period. The expected length of a production period is the time starting from the production switch on time when inventory falls to s, to the time production is switched off for the first time when inventory reaches S. For this, we consider the Markov process
$\{I(t), N(t), P(t), C(t), J_1(t), J_2(t), J_3(t), R(t), A(t); t \geq 0\}$. All variables have the same definition as in section 3, except that $P(t)$ takes the values 1 and 2 only and does not take the value 0. Here we consider the number of inventories as levels and the number of customers in the system is truncated to N. The truncation level N is taken such that, the probability of number of customers in the system going above the truncated size is of the order less than ϵ. Here the inventory level varies from 0 to S and level S is an absorbing state, denoting that production is switched off. The infinitesimal generator of this process is of the form

$$Q_1 = \begin{pmatrix} U_1 & U_1^0 \\ O & O \end{pmatrix}, \tag{26}$$

where U_1 is a sub matrix of Q_1 corresponding to transition among the transient states and U_1^0 corresponding to transition rates to the absorbing state. The structure of U_1 and U_1^0 are as follows:

$$U_1 = \begin{pmatrix} E_{00} & E_{01} & & & & & & \\ E_{10} & E_{11} & E_{12} & & & & & \\ & E_{21} & A_1^{'} & A_0^{'} & & & & \\ & & A_2^{'} & A_1^{'} & A_0^{'} & & & \\ & & & A_2^{'} & A_1^{'} & A_0^{'} & & \\ & & & & \ddots & \ddots & \ddots & \\ & & & & & A_2^{'} & A_1^{'} & A_0^{'} \\ & & & & & & A_2^{'} & A_1^{'} \end{pmatrix},$$

and

$$U_1^0 = \begin{pmatrix} O \\ O \\ \vdots \\ A_0^{''} \end{pmatrix}.$$

The structure of the matrices $E_{00}, E_{01}, E_{10}, E_{11}, E_{12}, E_{21}, A_0^{'}, A_1^{'}, A_2^{'}$ and $A_0^{''}$ are given in Appendix B.

The expected period length is the time until absorption of the Markov chain $U_1(t)$. The period length follows PH distributions with representation $\mathrm{PH}(\boldsymbol{\alpha_1}, U_1)$, where $\boldsymbol{\alpha_1} = \left(O_{1\times s}, 1, O_{1\times S-s-1}\right)$. Thus we arrive at the following theorem.

Theorem-2: The expected value of the length of the period is approximately given by

$$E[U_1(t)] = -\boldsymbol{\alpha_1} U_1^{-1} \boldsymbol{e}. \tag{27}$$

7. Numerical Examples

In this section, we give some numerical illustrations of variation in performance measures with regard to variation in values of the parameters. Here the MAP describing the arrival is represented by (D_0, D_1). The following values are kept fixed:

$m = 2, m_1 = 3, m_2 = 5, m_3 = 4, k = 3, s = 4, S = 9, \theta = 20,$

$$D_0 = \begin{pmatrix} -3.2 & 1 \\ 1.5 & -4.6 \end{pmatrix}; D_1 = \begin{pmatrix} 1.3 & 0.9 \\ 1.4 & 1.7 \end{pmatrix};$$

$$T^0 = \begin{pmatrix} 2 \\ 3 \\ 4 \end{pmatrix}; W^0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 20 \end{pmatrix}; U^0 = \begin{pmatrix} 1.8 \\ 2 \\ 2.5 \\ 3 \end{pmatrix};$$

$$\alpha = \begin{pmatrix} 0.2 & 0.4 & 0.4 \end{pmatrix}; T = \begin{pmatrix} -7.5 & 3. & 2.5 \\ 4.8 & -12 & 4.2 \\ 2.6 & 4.5 & -11.1 \end{pmatrix};$$

$$\beta = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \end{pmatrix}; W = \begin{pmatrix} -20 & 20 & 0 & 0 & 0 \\ 0 & -20 & 20 & 0 & 0 \\ 0 & 0 & -20 & 20 & 0 \\ 0 & 0 & 0 & -20 & 20 \\ 0 & 0 & 0 & 0 & -20 \end{pmatrix};$$

$$\gamma = \begin{pmatrix} 0.3 & 0.2 & 0.2 & 0.3 \end{pmatrix}; U = \begin{pmatrix} -7.3 & 2 & 1 & 2.5 \\ 2 & -7.3 & 2 & 1.3 \\ 1.6 & 2 & -9.6 & 3.5 \\ 1.2 & 3 & 1.5 & -8.7 \end{pmatrix}.$$

Effect of production interruption rate ν on various performance measures

Tables 8 and 9 shows the variation in various measures of performance for different values of ν

- From table 8 and figures 1 and 2 it is clear that the expected number of customers in the system and the expected number of customers waiting in the system due to lack of inventory increases as ν increases. But the expected inventory level decreases as ν increases. This is because as the production interruption rate ν increases, the production gets interrupted more frequently. This reduces the production rate. As a result, the inventory level decreases and so the expected number of customers in the system and the expected number of customers waiting in the system due to lack of inventory increases.
- As the production interruption rate ν increases, we see from table 8 and figure 3 that probability that both servers are idle decreases slightly, probability that one of the server is busy and the other idle increases and probability that both servers are busy decreases.

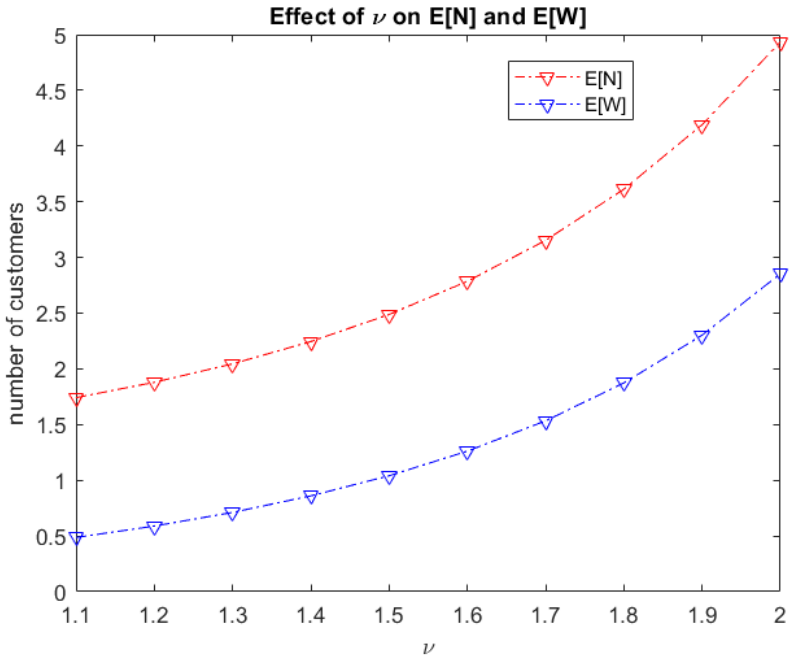


Figure 1

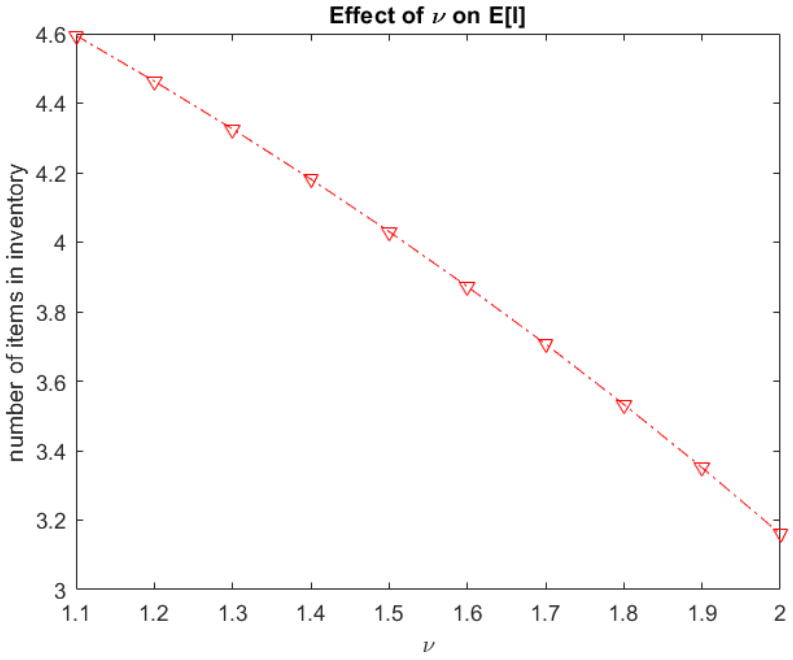


Figure 2

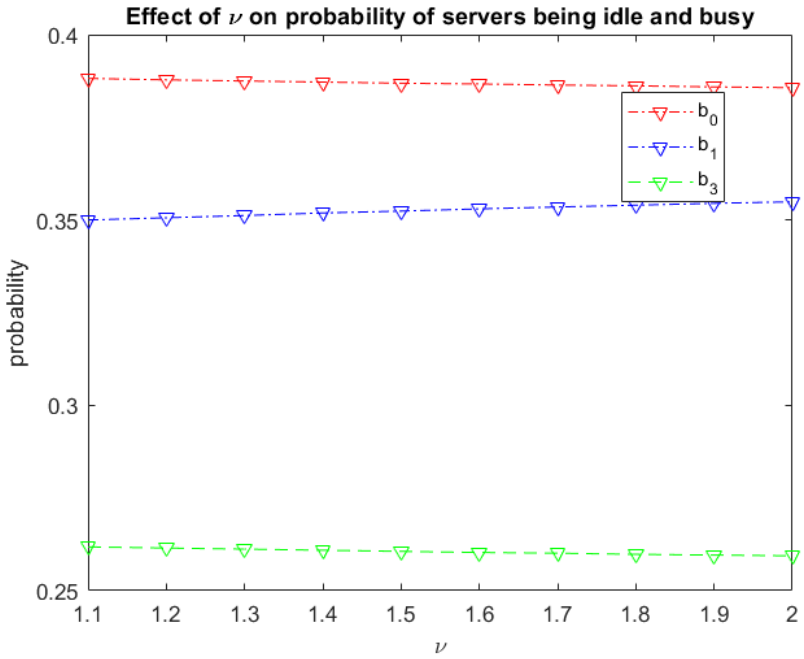


Figure 3

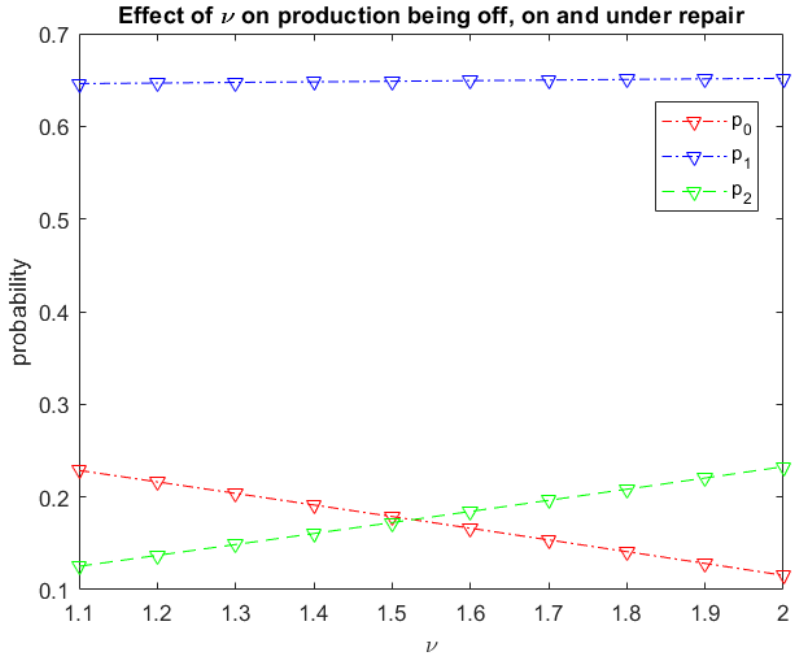


Figure 4

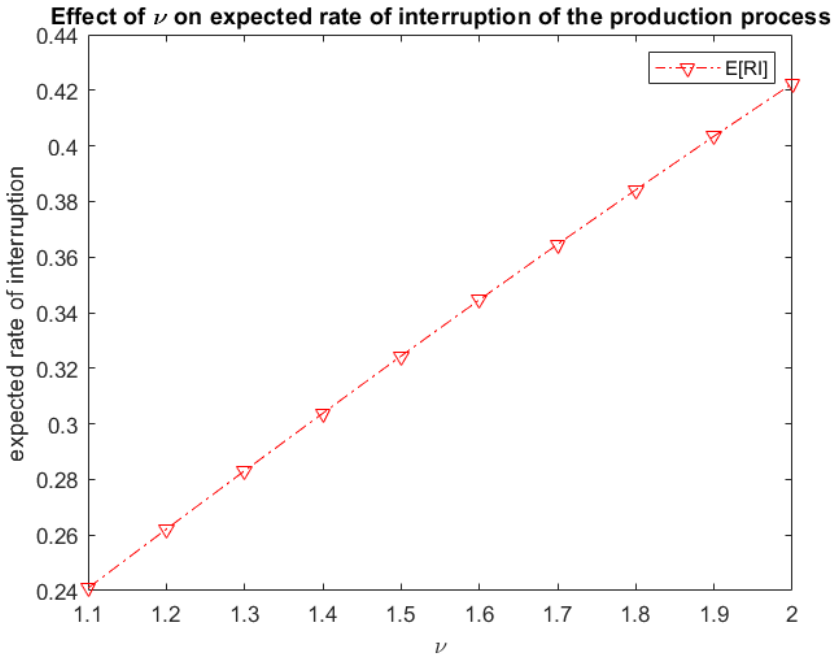


Figure 5

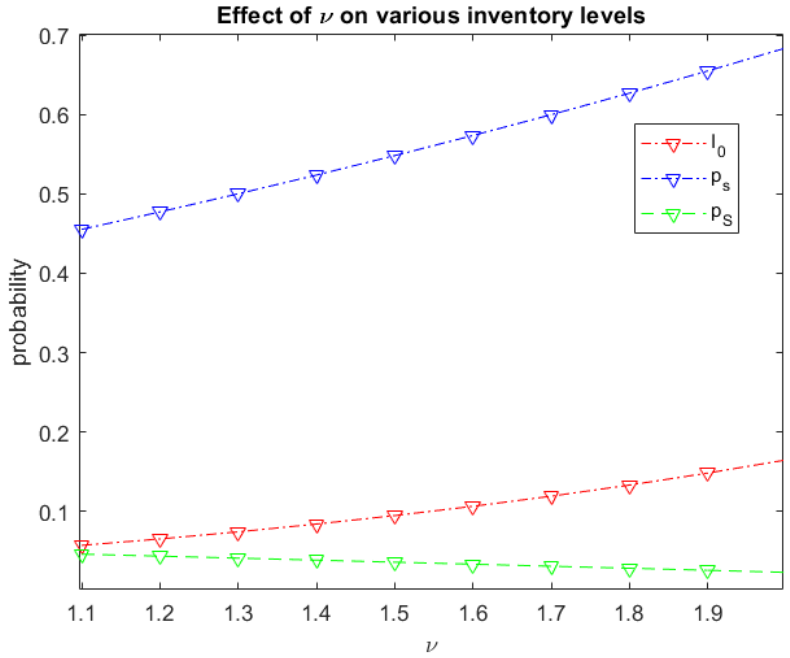


Figure 6

This is because as ν increases, the inventory level decreases and both servers are busy only if there is at least two inventory. So the probability that both servers are busy decreases slightly as ν increases. Thus probability that one of the server is busy and the other idle increases slightly as it happens if there is at least one inventory.

- It is clear from table 8 and figure 4 that as ν increases, the probability that production is off decreases and the probability that production is on and under repair increases. This is due to the fact that as the production interruption rate ν increases, the production gets interrupted more frequently and the probability that the production process is under repair increases. This reduces the inventory level and so the probability that production is on increases and the probability that production is off decreases.
- Also it is clear from table 9 and figure 5 that as the production interruption rate ν increases, the expected rate of interruption also increases.
- From table 9 and figure 6, we see that as ν increases, the probability that the inventory level is 0 increases, the probability that the inventory level is less than or equal to s increases and the probability that the maximum storage capacity S is used decreases. This is because as ν increases, the inventory level decreases.

8. Cost Analysis

For determining the optimal number of production phases to be protected, we construct a cost function. For the cost analysis we define the following costs:

- C_1: Holding cost per customer per unit time.
- C_2: Holding cost per item per unit time.
- C_3: Unit time cost for running the production machinery.
- C_4: Unit time cost incurred due to protection of the production machinery.
- C_5: Unit time cost incurred due to repair of the production machinery.
- C_6: Fixed cost for a production period.
- a: Cost incurred if the item is lost at the first phase of production.
- b: Cost incurred if the item is lost after the first phase of production.

The expected total cost is

$$ETC = C_1 * E[N] + C_2 * E[I] + C_3 * p_1 + C_4 * p_p + C_5 * p_2 + C_6/E[X(t)] + \sum_{l=1}^{m_2-k} \nu * a * b^{l-1} * p_l.$$

We fix the following values:
$\nu = 1, C_1 = 100, C_2 = 110, C_3 = 240, C_4 = 300, C_5 = 250, C_6 = 450, a = 50, b = 2$.
The values of $s, S, m, m_1, m_2, m_3, D_0, D_1, T, T^0, W, W^0, U, U^0, \alpha, \beta, \gamma$ are same as in section 7.

- From table 10 and figure 7, we see that the optimal value of the cost function is at $k = 3$. Thus it is optimal to give protection to the last 3 phases of production. The total cost is maximum when $k = 1$. This means that if we give protection only to the last phase, the total cost is high.

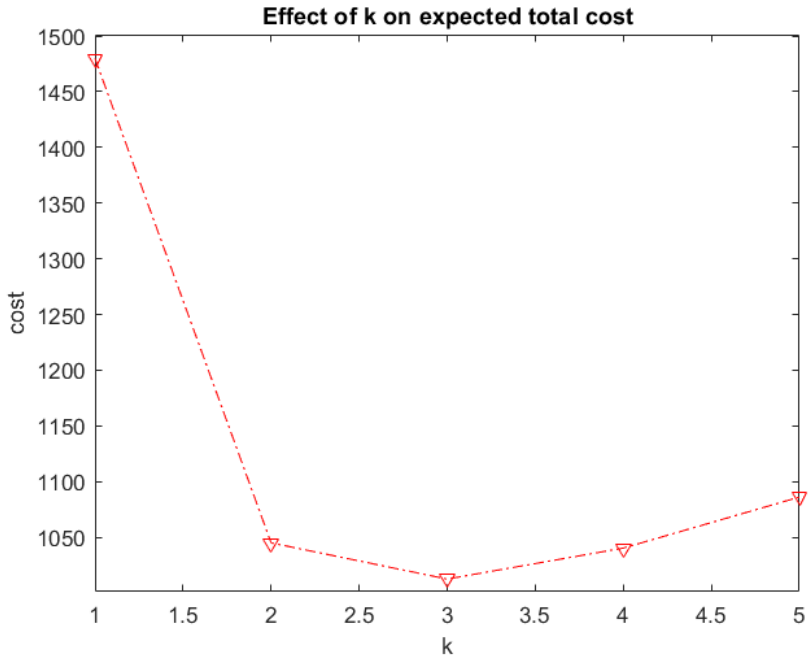


Figure 7

9. Conclusion

In this paper, we considered a production inventory model with two servers and positive service time. Production process is subject to shocks which interrupts the process and so to reduce the effect of loss due to shock, protection is given to the last k stages of production. Steady state analysis of the model is performed and some performance measures are evaluated. The expected length of a production period is calculated. We investigated numerically the variation in performance measures with regard to variation in values of the parameters. We formulated an optimization problem related to the number of stages of the production process to be protected.

Appendix A

Let $\beta_1 = \beta \otimes I_m$, $V_0 = \begin{pmatrix} \nu \otimes I_{m_2-k} & O \\ O & O \end{pmatrix}$, β is a $1 \times m_2$ matrix with one in the first column and zero elsewhere.

$$W^0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \theta \end{pmatrix}; W = \begin{pmatrix} -\theta & \theta & 0 & 0 & 0 \\ 0 & -\theta & \theta & 0 & 0 \\ 0 & 0 & -\theta & \theta & 0 \\ 0 & 0 & 0 & -\theta & \theta \\ 0 & 0 & 0 & 0 & -\theta \end{pmatrix};$$

$V = \begin{pmatrix} \nu \\ \vdots \\ \nu \\ 0 \\ \vdots \\ 0 \end{pmatrix}$, V is a column matrix with first $m_2 - k$ rows entries as ν and zero for the last k rows.

$$B_{01} = \begin{pmatrix} I_{(m_1+m_2)} \otimes D_1 & O & O & O \\ O & I_s \otimes E_3 & O & O \\ O & O & I_{(S-s-1)} \otimes E_4 & O \\ O & O & O & F_3 \end{pmatrix}, \tag{28}$$

$$B_{00} = E_1 + E_2, \tag{29}$$

$$B_{11} = E_8 + E_9, \tag{30}$$

$$B_{10} = \begin{pmatrix} O & O & O & O & O \\ I_s \otimes H_1 & O & O & O & O \\ O & H_2 & O & O & O \\ O & O & I_{(S-s-2)} \otimes H_3 & O & O \\ O & O & O & H_4 & O \end{pmatrix}, \tag{31}$$

$$B_{12} = \begin{pmatrix} I_{(2m_1+1)(m_2+m_3)} \otimes D_1 & O & O & O \\ O & I_{(s-1)} \otimes E_6 & O & O \\ O & O & I_{(S-s-1)} \otimes E_7 & O \\ O & O & O & F_6 \end{pmatrix}, \tag{32}$$

$$B_{21} = \begin{pmatrix} O & O & O & O & O & O \\ H_1 & O & O & O & O & O \\ O & I_{s-1} \otimes H_5 & O & O & O & O \\ O & O & H_6 & O & O & O \\ O & O & O & I_{(S-s-2)} \otimes H_7 & O & O \\ O & O & O & O & H_8 & O \end{pmatrix}, \tag{33}$$

$$A_2 = \begin{pmatrix} O & O & O & O & O & O & O \\ H_1 & O & O & O & O & O & O \\ O & H_5 & O & O & O & O & O \\ O & O & I_{s-2} \otimes H_9 & O & O & O & O \\ O & O & O & H_{10} & O & O & O \\ O & O & O & O & I_{S-s-2} \otimes H_{11} & O & O \\ O & O & O & O & O & H_{12} & O \end{pmatrix}, \tag{34}$$

$$A_0 = I_{K_4} \otimes D_1, \tag{35}$$

$$A_1 = A_{11} + A_{12}, \tag{36}$$

$$E_1 = \begin{pmatrix} I_{s+1} \otimes C_1 & O & O \\ O & I_{S-s-1} \otimes C_4 & O \\ O & O & D_0 \end{pmatrix},$$

$$E_2 = \begin{pmatrix} O & I_s \otimes C_2 & O & O & O \\ O & O & C_3 & O & O \\ O & O & O & I_{S-s-2} \otimes C_5 & O \\ O & O & O & O & C_6 \\ O & O & O & O & O \end{pmatrix},$$

$$E_3 = \begin{pmatrix} F_1 & O \\ O & F_2 \end{pmatrix}, E_4 = \begin{pmatrix} F_3 & O & O \\ O & F_1 & O \\ O & O & F_2 \end{pmatrix},$$

$$H_1 = \begin{pmatrix} h_{11} & O \\ O & h_{12} \end{pmatrix}, H_2 = \begin{pmatrix} h_{21} & O \\ h_{11} & O \\ O & h_{12} \end{pmatrix},$$

$$H_3 = \begin{pmatrix} h_{31} & O & O \\ O & h_{11} & O \\ O & O & h_{12} \end{pmatrix}, H_4 = \begin{pmatrix} h_{31} & O \end{pmatrix},$$

$$E_8 = \begin{pmatrix} C_1 & O & O & O \\ O & I_s \otimes C_8 & O & O \\ O & O & I_{S-s-1} \otimes C_{11} & O \\ O & O & O & C_{14} \end{pmatrix},$$

$$E_6 = \begin{pmatrix} F_4 & O \\ O & F_5 \end{pmatrix}, E_7 = \begin{pmatrix} F_6 & O & O \\ O & F_4 & O \\ O & O & F_5 \end{pmatrix},$$

$$E_9 = \begin{pmatrix} O & C_7 & O & O & O & O \\ O & O & I_{s-1} \otimes C_9 & O & O & O \\ O & O & O & C_{10} & O & O \\ O & O & O & O & I_{S-s-2} \otimes C_{12} & O \\ O & O & O & O & O & C_{13} \\ O & O & O & O & O & O \end{pmatrix},$$

$$H_5 = \begin{pmatrix} h_{51} & O \\ O & h_{52} \end{pmatrix}, H_6 = \begin{pmatrix} h_{61} & O \\ h_{51} & O \\ O & h_{52} \end{pmatrix},$$

$$H_7 = \begin{pmatrix} h_{71} & O & O \\ O & h_{51} & O \\ O & O & h_{52} \end{pmatrix}, H_8 = \begin{pmatrix} h_{71} & O \end{pmatrix},$$

$$K_4 = (m_2 + m_3)[1 + 2m_1 + (S-2)m_1^2] + (S-s)m_1^2,$$

$$A_{11} = \begin{pmatrix} C_1 & O & O & O & O \\ O & C_8 & O & O & O \\ O & O & I_{s-1} \otimes G_2 & O & O \\ O & O & O & I_{S-s-1} \otimes G_3 & O \\ O & O & O & O & G_4 \end{pmatrix},$$

$$A_{12} = \begin{pmatrix} O & C_7 & O & O & O & O & O \\ O & O & G_1 & O & O & O & O \\ O & O & O & I_{s-2} \otimes G_5 & O & O & O \\ O & O & O & O & G_6 & O & O \\ O & O & O & O & O & I_{S-s-2} \otimes G_7 & O \\ O & O & O & O & O & O & G_8 \\ O & O & O & O & O & O & O \end{pmatrix},$$

$$H_9 = \begin{pmatrix} h_{91} & O \\ O & h_{92} \end{pmatrix}, H_{10} = \begin{pmatrix} h_{10} & O \\ h_{91} & O \\ O & h_{92} \end{pmatrix},$$

$$H_{11} = \begin{pmatrix} h & O & O \\ O & h_{91} & O \\ O & O & h_{92} \end{pmatrix}, H_{12} = \begin{pmatrix} h & O & O \end{pmatrix},$$

$$F_1 = \begin{pmatrix} O & \alpha \otimes (I_{m_2} \otimes D_1) \end{pmatrix}, F_2 = \begin{pmatrix} O & \alpha \otimes (I_{m_3} \otimes D_1) \end{pmatrix},$$

$$F_3 = \begin{pmatrix} O & \alpha \otimes D_1 \end{pmatrix}, F_4 = \begin{pmatrix} \alpha \otimes (I_{m_1 m_2} \otimes D_1) \\ I_{m_1} \otimes (\alpha \otimes I_{m_2}) \otimes D_1 \end{pmatrix},$$

$$F_5 = \begin{pmatrix} \alpha \otimes (I_{m_1 m_3} \otimes D_1) \\ I_{m_1} \otimes (\alpha \otimes I_{m_2}) \otimes D_1 \end{pmatrix}, F_6 = \begin{pmatrix} \alpha \otimes (I_{m_1} \otimes D_1) \\ I_{m_1} \otimes (\alpha \otimes D_1) \end{pmatrix},$$

$$C_1 = \begin{pmatrix} W \oplus D_0 - V_0 & (\gamma \otimes V) \otimes I_m \\ U^0 \otimes \beta_1 & U \oplus D_0 \end{pmatrix},$$

$$C_2 = \begin{pmatrix} W^0 \otimes \beta_1 & O \\ O & O \end{pmatrix}, C_3 = \begin{pmatrix} O & W^0 \otimes \beta_1 & O \\ O & O & O \end{pmatrix},$$
$$C_4 = \begin{pmatrix} D_0 & O \\ O & C_1 \end{pmatrix}, C_5 = \begin{pmatrix} O \\ C_3 \end{pmatrix},$$
$$C_6 = \begin{pmatrix} O \\ W^0 \otimes I_m \\ C_3 \end{pmatrix}, C_7 = \begin{pmatrix} 0 & (\alpha \otimes W^0) \otimes \beta_1) & O \\ & O & O \end{pmatrix},$$
$$C_8 = \begin{pmatrix} I_2 \otimes [T \oplus (W \oplus D_0)] - I_{2m_1} \otimes V_0 & I_{2m_1} \otimes [(\gamma \otimes V) \otimes I_m] \\ I_{2m_1} \otimes (U^0 \otimes \beta_1) & I_2 \otimes [T \oplus (U \oplus D_0)] \end{pmatrix},$$
$$C_9 = \begin{pmatrix} I_{2m_1} \otimes (W^0 \otimes \beta_1) & O \\ O & O \end{pmatrix},$$
$$C_{10} = \begin{pmatrix} O & I_{2m_1} \otimes (W^0 \otimes \beta_1) & O \\ O & O & O \end{pmatrix},$$
$$C_{11} = \begin{pmatrix} c_{11} & c_{12} \end{pmatrix},$$
$$c_{11} = \begin{pmatrix} I_2 \otimes [T \oplus D_0] & O \\ O & I_2 \otimes [T \oplus (W \oplus D_0)] - I_{2m_1} \otimes V_0 \\ O & I_{2m_1} \otimes (U^0 \otimes \beta_1) \end{pmatrix},$$
$$c_{12} = \begin{pmatrix} O \\ I_{2m_1} \otimes (\gamma \otimes V) \otimes I_m \\ I_2 \otimes [T \oplus (U \oplus D_0)] \end{pmatrix},$$
$$C_{12} = \begin{pmatrix} O \\ C_{10} \end{pmatrix}, C_{13} = \begin{pmatrix} O \\ I_{2m_1} \otimes W^0 \otimes I_m \\ O \end{pmatrix},$$
$$C_{14} = I_2 \otimes [T \oplus D_0], G_1 = \begin{pmatrix} g_1 & O \\ O & O \end{pmatrix},$$
$$G_2 = \begin{pmatrix} g_{21} & I_{m_1^2} \otimes ((\gamma \otimes V) \otimes I_m) \\ I_{m_1^2} \otimes (U^0 \otimes \beta_1) & g_{22} \end{pmatrix},$$
$$G_3 = \begin{pmatrix} g_{31} & O & O \\ O & g_{21} & I_{m_1^2} \otimes ((\gamma \otimes V) \otimes I_m) \\ O & I_{m_1^2} \otimes (U^0 \otimes \beta_1) & g_{22} \end{pmatrix},$$
$$G_4 = T \oplus [T \oplus D_0], G_5 = \begin{pmatrix} I_{m_1^2} \otimes (W^0 \otimes \beta_1) & O \\ O & O \end{pmatrix},$$
$$G_6 = \begin{pmatrix} O & I_{m_1^2} \otimes (W^0 \otimes \beta_1) & O \\ O & O & O \end{pmatrix},$$
$$G_7 = \begin{pmatrix} O & O & O \\ O & I_{m_1^2} \otimes (W^0 \otimes \beta_1) & O \\ O & O & O \end{pmatrix},$$
$$G_8 = \begin{pmatrix} O \\ I_{m_1^2} \otimes (W^0 \otimes I_m) \\ O \end{pmatrix},$$
$$g_1 = \begin{pmatrix} (\alpha \otimes (I_{m_1} \otimes W^0)) \otimes \beta_1 \\ I_{m_1} \otimes ((\alpha \otimes W^0) \otimes \beta_1) \end{pmatrix},$$
$$g_{21} = T \oplus [T \oplus (W \oplus D_0)] - I_{m_1^2 m} \otimes V_0,$$
$$g_{22} = T \oplus [T \oplus (U \oplus D_0)],$$
$$g_{31} = T \oplus [T \oplus D_0],$$
$$h_{11} = \begin{pmatrix} T^0 \otimes I_{m_2 m} \\ T^0 \otimes I_{m_2 m} \end{pmatrix}, h_{12} = \begin{pmatrix} T^0 \otimes I_{m_3 m} \\ T^0 \otimes I_{m_3 m} \end{pmatrix},$$
$$h_{21} = \begin{pmatrix} T^0 \otimes \beta I_m \\ T^0 \otimes \beta I_m \end{pmatrix}, h_{31} = \begin{pmatrix} T^0 \otimes I_m \\ T^0 \otimes I_m \end{pmatrix},$$

$h_{51} = \begin{pmatrix} T^0 \otimes I_{m_1 m_2 m} & I_{m_1} \otimes (T^0 \otimes I_{m_2 m}) \end{pmatrix}$,
$h_{52} = \begin{pmatrix} T^0 \otimes I_{m_1 m_3 m} & I_{m_1} \otimes (T^0 \otimes I_{m_3 m}) \end{pmatrix}$,
$h_{61} = \begin{pmatrix} (T^0 \otimes I_{m_1}) \otimes \beta_1 & I_{m_1} \otimes (T^0 \otimes \beta_1) \end{pmatrix}$,
$h_{71} = \begin{pmatrix} T^0 \otimes I_{m m_1} & I_{m_1} \otimes (T^0 \otimes I_m) \end{pmatrix}$,
$h_{91} = (T^0 \otimes \alpha) \otimes I_{m_1 m_2 m} + I_{m_1} \otimes ((T^0 \otimes \alpha) \otimes I_{m_2 m})$,
$h_{92} = (T^0 \otimes \alpha) \otimes I_{m_1 m_3 m} + I_{m_1} \otimes ((T^0 \otimes \alpha) \otimes I_{m_3 m})$,
$h_{10} = ((T^0 \otimes \alpha) \otimes I_{m_1}) \otimes \beta_1 + I_{m_1} \otimes ((T^0 \otimes \alpha) \otimes \beta_1)$,
$h = (T^0 \otimes \alpha) \otimes I_{m_1 m} + I_{m_1} \otimes ((T^0 \otimes \alpha) \otimes I_m)$.

Appendix B

$$E_{00} = \begin{pmatrix} C_1 & C_1' & & & \\ & C_1 & C_1' & & \\ & & \ddots & \ddots & \\ & & & C_1 & C_1' \\ & & & & C_1'' \end{pmatrix},$$

$$E_{01} = \begin{pmatrix} C_2 & O \\ O & I_N \otimes C_7 \end{pmatrix}, E_{10} = \begin{pmatrix} O & O \\ I_N \otimes H_1 & O \end{pmatrix},$$

$$E_{11} = \begin{pmatrix} C_1 & E_3 & & & & \\ & C_8 & E_3' & & & \\ & & C_8 & E_3' & & \\ & & & \ddots & \ddots & \\ & & & & C_8 & E_3' \\ & & & & & C_8' \end{pmatrix},$$

$$E_{12} = \begin{pmatrix} C_2 & O & O \\ O & C_9 & O \\ O & O & I_{N-1} \otimes G_1 \end{pmatrix}, E_{21} = \begin{pmatrix} O & O & O \\ H_1 & O & O \\ O & I_{N-1} \otimes H_5 & O \end{pmatrix},$$

$$A_1' = \begin{pmatrix} C_1 & E_3 & & & & \\ & C_8 & E_6 & & & \\ & & G_2 & E_6' & & \\ & & & \ddots & \ddots & \\ & & & & G_2 & E_6' \\ & & & & & G_2' \end{pmatrix},$$

$$A_2' = \begin{pmatrix} O & O & O & O \\ H_1 & O & O & O \\ O & H_5 & O & O \\ O & O & I_{N-2} \otimes H_9 & O \end{pmatrix},$$

$$A_0' = \begin{pmatrix} C_2 & O & O \\ O & C_9 & O \\ O & O & I_{N-1} \otimes G_5 \end{pmatrix}, A_0'' = \begin{pmatrix} C_6' & O & O \\ O & C_{13}' & O \\ O & O & I_{N-1} \otimes G_8' \end{pmatrix},$$

$C_1' = I_{m_2+m_3} \otimes D_1$,

$$C_1'' = \begin{pmatrix} W \oplus (D_0 + D_1) - V_0 & (\gamma \otimes V) \otimes I_m \\ U^0 \otimes \beta_1 & U \oplus (D_0 + D_1) \end{pmatrix},$$

$$C_8' = \begin{pmatrix} I_2 \otimes [T \oplus (W \oplus (D_0 + D_1))] - I_{2m_1} \otimes V_0 & I_{2m_1} \otimes [(\gamma \otimes V) \otimes I_m] \\ I_{2m_1} \otimes (U^0 \otimes \beta_1) & I_2 \otimes [T \oplus (U \oplus (D_0 + D_1))] \end{pmatrix},$$

$$E_3 = \begin{pmatrix} e_{31} & O \\ O & e_{32} \end{pmatrix}, E_6 = \begin{pmatrix} e_{61} & O \\ O & e_{62} \end{pmatrix},$$

$$E_3^{'} = I_{2m_1(m_2+m_3)} \otimes D_1,\ E_6^{'} = I_{m_1^2(m_2+m_3)} \otimes D_1,$$

$$G_2^{'} = \begin{pmatrix} g'_{21} & I_{m_1^2} \otimes [(\gamma \otimes V) \otimes I_m] \\ I_{m_1^2} \otimes (U^0 \otimes \beta_1) & g'_{22} \end{pmatrix},$$

$$C_6^{'} = \begin{pmatrix} W^0 \otimes I_m \\ O \end{pmatrix},\ C_{13}^{'} = \begin{pmatrix} I_{2m_1} \otimes (W^0 \otimes I_m) \\ O \end{pmatrix},$$

$$G_8^{'} = \begin{pmatrix} I_{m_1^2} \otimes (W^0 \otimes I_m) \\ O \end{pmatrix},$$

$$e_{31} = \begin{pmatrix} O & \alpha \otimes (I_{m_2} \otimes D_1) \end{pmatrix},\ e_{32} = \begin{pmatrix} O & \alpha \otimes (I_{m_3} \otimes D_1) \end{pmatrix},$$

$$e_{61} = \begin{pmatrix} \alpha \otimes (I_{m_1 m_2} \otimes D_1) \\ I_{m_1} \otimes (\alpha \otimes (I_{m_2} \otimes D_1)) \end{pmatrix},\ e_{62} = \begin{pmatrix} \alpha \otimes (I_{m_1 m_3} \otimes D_1) \\ I_{m_1} \otimes (\alpha \otimes (I_{m_3} \otimes D_1)) \end{pmatrix},$$

$$g_{21}^{'} = T \oplus [T \oplus (W \oplus (D_0 + D_1))] - I_{m_1^2} \otimes V_0,$$

$$g_{22}^{'} = T \oplus [T \oplus (U \oplus (D_0 + D_1))].$$

Funding

Ambily P Mathew received support from DST-RSF research project number 22-49-02023(RSF) and research project number 64800(DST) for the preparation of this publication.

References

[1] A. Z. Melikov and A. A. Molchanov. "Stock optimization in transportation/storage systems". In: *Cybernetics and Systems Analysis* 28.3 (1992), pp. 484–487. ISSN: 1573-8337. DOI: 10.1007/bf01125431.

[2] Karl Sigman and David Simchi-Levi. "Light traffic heuristic for anM/G/1 queue with limited inventory". In: *Annals of Operations Research* 40.1 (Dec. 1992), pp. 371–380. ISSN: 1572-9338. DOI: 10.1007/bf02060488.

[3] Achyutha Krishnamoorthy, Dhanya Shajin, and Viswanath C. Narayanan. "Inventory with Positive Service Time: a Survey". In: *Queueing Theory 2*. Wiley, Apr. 2021, pp. 201–237. ISBN: 9781119755234. DOI: 10.1002/9781119755234.ch6.

[4] A. Krishnamoorthy and Viswanath C. Narayanan. "Production inventory with service time and vacation to the server". In: *IMA Journal of Management Mathematics* 22.1 (2011), pp. 33–45. DOI: 10.1093/imaman/dpp025.

[5] A. Krishnamoorthy and Narayanan C. Viswanath. "Stochastic decomposition in production inventory with service time". In: *European Journal of Operational Research* 228.2 (2013), pp. 358–366. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2013.01.041.

[6] Jung Woo Baek and Seung Ki Moon. "The M/M/1 queue with a production-inventory system and lost sales". In: *Applied Mathematics and Computation* 233 (May 2014), pp. 534–544. ISSN: 0096-3003. DOI: 10.1016/j.amc.2014.02.033.

[7] A. Krishnamoorthy, B. Gopakumar, and C.N. Viswanath. "An M/E^m/1 Queue With Protected and Unprotected Phases From Interruption". 5th International Conference on Queueing Theory and Network Applications. 2010.

[8] A. Krishnamoorthy, Sajeev S. Nair, and Viswanath C. Narayanan. "Production inventory with service time and interruptions". In: *International Journal of Systems Science* 46.10 (Oct. 2013), pp. 1800–1816. ISSN: 1464-5319. DOI: 10.1080/00207721.2013.837538.

[9] Anoop N. Nair and M. J. Jacob. "An (s,S) Production Inventory Controlled Self-Service Queuing System". In: *Journal of Probability and Statistics* 2015 (2015), pp. 1–8. ISSN: 1687-9538. DOI: 10.1155/2015/505082.

[10] Dequan Yue and Yaling Qin. "A Production Inventory System with Service Time and Production Vacations". In: *Journal of Systems Science and Systems Engineering* 28.2 (Feb. 2019), pp. 168–180. ISSN: 1861-9576. DOI: 10.1007/s11518-018-5402-8.

[11] Jung Woo Baek and Seung Ki Moon. "A production–inventory system with a Markovian service queue and lost sales". In: *Journal of the Korean Statistical Society* 45.1 (Mar. 2016), pp. 14–24. ISSN: 1226-3192. DOI: 10.1016/j.jkss.2015.05.002.

[12] Yaling Qin and Dequan Yue. "Performance Analyses of Production Inventory Systems Considering Service Time and Product Returns of Online Shopping". In: *Journal of Systems Science and Complexity* 32.3 (2018), pp. 888–906. ISSN: 1559-7067. DOI: 10.1007/s11424-018-7230-9.

[13] K. P. Jose and P. S. Reshmi. "A production inventory model with deteriorating items and retrial demands". In: *OPSEARCH* 58.1 (Aug. 2020), pp. 71–82. ISSN: 0975-0320. DOI: 10.1007/s12597-020-00471-8.

[14] M. F. Neuts. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Baltimore: Johns Hopkins University Press, 1981. ISBN: 978-0-486-68342-3.

TABLES

Table 1: *Transition table corresponding to arrivals*

From	To	Rate	Description
$(n,0,1)$	$(n+1,0,1)$	$I_{m_2} \otimes D_1$	arrival when both servers are idle and no inventory in the system, production is on, $n \geq 0$
$(n,0,2)$	$(n+1,0,2)$	$I_{m_3} \otimes D_1$	arrival when both servers are idle and no inventory in the system, production process is under repair, $n \geq 0$
$(0,i,1)$	$(1,i,1)$	$\begin{pmatrix} 0 & \alpha \otimes (I_{m_2} \otimes D_1) \end{pmatrix}$	$1 \leq i \leq S-1$, arrival when both servers are idle, at least one inventory in the system, and production is on
$(0,i,2)$	$(1,i,2)$	$\begin{pmatrix} 0 & \alpha \otimes (I_{m_3} \otimes D_1) \end{pmatrix}$	$1 \leq i \leq S-1$, arrival when both servers are idle, at least one inventory in the system, and production process is under repair
$(0,i,0)$	$(1,i,0)$	$\begin{pmatrix} 0 & \alpha \otimes D_1 \end{pmatrix}$	$s+1 \leq i \leq S$, arrival when both servers are idle and production is off
$(n,1,1)$	$(n+1,1,1)$	$I_{2m_1m_2} \otimes D_1$	arrival when one server is busy, the other server is idle, only one inventory in system, production is on, $n \geq 1$
$(n,1,2)$	$(n+1,1,2)$	$I_{2m_1m_3} \otimes D_1$	arrival when one server is busy, the other server is idle, only one inventory in system, production process is under repair, $n \geq 1$

Table 2: *Transition table corresponding to arrivals*

From	To	Rate	Description
$(1,i,1)$	$(2,i,1)$	$\begin{pmatrix} \alpha \otimes (I_{m_1m_2} \otimes D_1) \\ I_{m_1} \otimes (\alpha \otimes (I_{m_2} \otimes D_1)) \end{pmatrix}$	$2 \leq i \leq S-1$, arrival when one server is busy, the other server is idle, at least two inventory in the system, and production is on
$(1,i,2)$	$(2,i,2)$	$\begin{pmatrix} \alpha \otimes (I_{m_1m_3} \otimes D_1) \\ I_{m_1} \otimes (\alpha \otimes (I_{m_3} \otimes D_1)) \end{pmatrix}$	$2 \leq i \leq S-1$, arrival when one server is busy, the other server is idle, at least two inventory in the system, and production process is under repair
$(1,i,0)$	$(2,i,0)$	$\begin{pmatrix} \alpha \otimes (I_{m_1} \otimes D_1) \\ I_{m_1} \otimes (\alpha \otimes D_1) \end{pmatrix}$	$s+1 \leq i \leq S$, arrival when one server is busy, the other server is idle, and production is off
$(n,i,1)$	$(n+1,i,1)$	$I_{m_1^2m_2} \otimes D_1$	$n \geq 2, 2 \leq i \leq S-1$, arrival when both servers are busy and production is on
$(n,i,2)$	$(n+1,i,2)$	$I_{m_1^2m_3} \otimes D_1$	$n \geq 2, 2 \leq i \leq S-1$, arrival when both servers are busy and production process is under repair
$(n,i,0)$	$(n+1,i,0)$	$I_{m_1^2} \otimes D_1$	$n \geq 2, s+1 \leq i \leq S$, arrival when both servers are busy and production is off

Table 3: *Transition table corresponding to no change*

$(0,i,0)$	$(0,i,0)$	D_0	$s+1 \leq i \leq S$
$(0,i,1)$	$(0,i,1)$	$W \oplus D_0 - V_0$	$0 \leq i \leq S-1$
$(0,i,2)$	$(0,i,2)$	$U \oplus D_0$	$0 \leq i \leq S-1$
$(n,0,1)$	$(n,0,1)$	$W \oplus D_0 - V_0$	$n \geq 1$
$(n,0,2)$	$(n,0,2)$	$U \oplus D_0$	$n \geq 1$
$(1,i,1)$	$(1,i,1)$	$I_2 \otimes [T \oplus (W \oplus D_0)] - I_{2m_1} \otimes V_0$	$1 \leq i \leq S-1$
$(1,i,2)$	$(1,i,2)$	$I_2 \otimes [T \oplus (U \oplus D_0)]$	$1 \leq i \leq S-1$
$(1,i,0)$	$(1,i,0)$	$I_2 \otimes [T \oplus D_0]$	$s+1 \leq i \leq S$
$(n,i,0)$	$(n,i,0)$	$T \oplus [T \oplus D_0]$	$n \geq 2, s+1 \leq i \leq S$
$(n,1,1)$	$(n,1,1)$	$I_2 \otimes [T \oplus (W \oplus D_0)] - I_{2m_1} \otimes V_0$	$n \geq 2, 1 \leq i \leq S-1$
$(n,1,2)$	$(n,1,2)$	$I_2 \otimes [T \oplus (U \oplus D_0)]$	$n \geq 2, 1 \leq i \leq S-1$
$(n,i,1)$	$(n,i,1)$	$T \oplus [T \oplus (W \oplus D_0)] - I_{m_1^2} \otimes V_0$	$n \geq 2, 1 \leq i \leq S-1$
$(n,i,2)$	$(n,i,2)$	$T \oplus [T \oplus (U \oplus D_0)]$	$n \geq 2, 1 \leq i \leq S-1$

Table 4: *Transition table corresponding to service completions*

From	To	Rate	Description
$(1,i,1)$	$(0,i-1,1)$	$\begin{pmatrix} T^0 \otimes I_{m_2 m} \\ T^0 \otimes I_{m_2 m} \end{pmatrix}$	$1 \leq i \leq S-1$
$(1,i,2)$	$(0,i-1,2)$	$\begin{pmatrix} T^0 \otimes I_{m_3 m} \\ T^0 \otimes I_{m_3 m} \end{pmatrix}$	$1 \leq i \leq S-1$
$(1,s+1,0)$	$(0,s,1)$	$\begin{pmatrix} T^0 \otimes \beta_1 \\ T^0 \otimes \beta_1 \end{pmatrix}$	
$(1,i,0)$	$(0,i-1,0)$	$\begin{pmatrix} T^0 \otimes I_m \\ T^0 \otimes I_m \end{pmatrix}$	$s+2 \leq i \leq S$
$(n,1,1)$	$(n-1,0,1)$	$\begin{pmatrix} T^0 \otimes I_{m_2 m} \\ T^0 \otimes I_{m_2 m} \end{pmatrix}$	$n \geq 2$
$(n,1,2)$	$(n-1,0,2)$	$\begin{pmatrix} T^0 \otimes I_{m_3 m} \\ T^0 \otimes I_{m_3 m} \end{pmatrix}$	$n \geq 2$
$(2,i,1)$	$(1,i-1,1)$	$\begin{pmatrix} T^0 \otimes I_{m_1 m_2 m} & I_{m_1} \otimes (T^0 \otimes I_{m_2 m}) \end{pmatrix}$	$2 \leq i \leq S-1$
$(2,i,2)$	$(1,i-1,2)$	$\begin{pmatrix} T^0 \otimes I_{m_1 m_3 m} & I_{m_1} \otimes (T^0 \otimes I_{m_3 m}) \end{pmatrix}$	$2 \leq i \leq S-1$
$(2,s+1,0)$	$(1,s,1)$	$\begin{pmatrix} (T^0 \otimes I_{m_1}) \otimes \beta_1 & I_{m_1} \otimes (T^0 \otimes \beta_1) \end{pmatrix}$	
$(2,i,0)$	$(1,i-1,0)$	$\begin{pmatrix} T^0 \otimes I_{m_1 m} & I_{m_1} \otimes (T^0 \otimes I_m) \end{pmatrix}$	$s+2 \leq i \leq S$
$(n,2,1)$	$(n-1,1,1)$	$\begin{pmatrix} T^0 \otimes I_{m_1 m_2 m} & I_{m_1} \otimes (T^0 \otimes I_{m_2 m}) \end{pmatrix}$	$n \geq 2$
$(n,2,2)$	$(n-1,1,2)$	$\begin{pmatrix} T^0 \otimes I_{m_1 m_3 m} & I_{m_1} \otimes (T^0 \otimes I_{m_3 m}) \end{pmatrix}$	$n \geq 2$
$(n,i,1)$	$(n-1,i-1,1)$	$(T^0 \otimes \alpha) \otimes I_{m_1 m_2 m} +$ $I_{m_1} \otimes ((T^0 \otimes \alpha) \otimes I_{m_2 m})$	$n \geq 3$, $3 \leq i \leq S-1$
$(n,i,2)$	$(n-1,i-1,2)$	$(T^0 \otimes \alpha) \otimes I_{m_1 m_3 m} +$ $I_{m_1} \otimes ((T^0 \otimes \alpha) \otimes I_{m_3 m})$	$n \geq 3$, $3 \leq i \leq S-1$
$(n,s+1,0)$	$(n-1,s,1)$	$((T^0 \otimes \alpha) \otimes I_{m_1}) \otimes \beta_1 +$ $I_{m_1} \otimes ((T^0 \otimes \alpha) \otimes \beta_1)$	$n \geq 3$
$(n,i,0)$	$(n-1,i-1,0)$	$(T^0 \otimes \alpha) \otimes I_{m_1 m} +$ $I_{m_1} \otimes ((T^0 \otimes \alpha) \otimes I_m)$	$n \geq 3$, $s+2 \leq i \leq S$

Table 5: *Transition table corresponding to interruptions*

$(0,i,1)$	$(0,i,2)$	$(\gamma \otimes V) \otimes I_m$	$0 \leq i \leq S-1$
$(n,0,1)$	$(n,0,2)$	$(\gamma \otimes V) \otimes I_m$	$n \geq 1$
$(1,i,1)$	$(1,i,2)$	$I_{2m_1} \otimes ((\gamma \otimes V) \otimes I_m)$	$1 \leq i \leq S-1$
$(n,1,1)$	$(n,1,2)$	$I_{2m_1} \otimes ((\gamma \otimes V) \otimes I_m)$	$n \geq 2$
$(n,i,1)$	$(n,i,2)$	$I_{m_1{}^2} \otimes ((\gamma \otimes V) \otimes I_m)$	$n \geq 2, 2 \leq i \leq S-1$

Table 6: *Transition table corresponding to production completions*

$(0,i,1)$	$(0,i+1,1)$	$W^0 \otimes \beta_1$	$0 \leq i \leq S-2$
$(0,S-1,1)$	$(0,S,0)$	$W^0 \otimes I_m$	
$(n,0,1)$	$(n,1,1)$	$\begin{pmatrix} 0 & (\alpha \otimes W^0) \otimes \beta_1 \end{pmatrix}$	$n \geq 1$
$(1,i,1)$	$(1,i+1,1)$	$I_{2m_1} \otimes (W^0 \otimes \beta_1)$	$1 \leq i \leq S-2$
$(1,S-1,1)$	$(1,S,0)$	$I_{2m_1} \otimes (W^0 \otimes I_m)$	
$(n,1,1)$	$(n,2,1)$	$\begin{pmatrix} (\alpha \otimes (I_{m_1} \otimes W^0)) \otimes \beta_1 \\ I_{m_1} \otimes ((\alpha \otimes W^0) \otimes \beta_1) \end{pmatrix}$	$n \geq 2$
$(n,i,1)$	$(n,i+1,1)$	$I_{m_1{}^2} \otimes (W^0 \otimes \beta_1)$	$n \geq 2, 2 \leq i \leq S-2$
$(n,S-1,1)$	$(n,S,0)$	$I_{m_1{}^2} \otimes (W^0 \otimes I_m)$	$n \geq 2$

Table 7: *Transition table corresponding to repair completion*

$(0,i,2)$	$(0,i,1)$	$U^0 \otimes \beta_1$	$0 \leq i \leq S-1$
$(n,0,2)$	$(n,0,1)$	$U^0 \otimes \beta_1$	$n \geq 1$
$(1,i,2)$	$(1,i,1)$	$I_{2m_1} \otimes (U^0 \otimes \beta_1)$	$1 \leq i \leq S-1$
$(n,1,2)$	$(n,1,1)$	$I_{2m_1} \otimes (U^0 \otimes \beta_1)$	$n \geq 2$
$(n,i,2)$	$(n,i,1)$	$I_{m_1{}^2} \otimes (U^0 \otimes \beta_1)$	$n \geq 2, 2 \leq i \leq S-1$

Table 8: *Effect of ν on some measures of performance*

ν	$E[N]$	$E[I]$	$E[W]$	b_0	b_1	b_2	p_0	p_1	p_2
1.1	1.7417	4.5946	0.4866	0.3882	0.3500	0.2618	0.2287	0.6461	0.1252
1.2	1.8794	4.4632	0.5888	0.3878	0.3506	0.2615	0.2163	0.6467	0.1370
1.3	2.0449	4.3256	0.7117	0.3875	0.3512	0.2612	0.2039	0.6474	0.1487
1.4	2.2451	4.1814	0.8602	0.3872	0.3519	0.2609	0.1914	0.648	0.1606
1.5	2.4883	4.0305	1.0406	0.3869	0.3524	0.2606	0.1789	0.6486	0.1724
1.6	2.7861	3.8723	1.2613	0.3867	0.3530	0.2603	0.1663	0.6493	0.1844
1.7	3.1537	3.7066	1.5338	0.3864	0.3535	0.2601	0.1537	0.6499	0.1964
1.8	3.6123	3.5330	1.8736	0.3862	0.3540	0.2598	0.1410	0.6506	0.2084
1.9	4.1915	3.3510	2.3028	0.3859	0.3545	0.2596	0.1283	0.6512	0.2205
2.0	4.9349	3.1602	2.8535	0.3857	0.3549	0.2594	0.1155	0.6518	0.2327

Table 9: *Effect of ν on some measures of performance*

ν	p_p	p_u	$E[RI]$	I_0	p_s	p_S
1.1	0.4271	0.2190	0.2409	0.0574	0.4552	0.0463
1.2	0.4283	0.2184	0.2621	0.0655	0.4772	0.0438
1.3	0.4296	0.2178	0.2831	0.0744	0.5000	0.0413
1.4	0.4310	0.2170	0.3038	0.0842	0.5236	0.0388
1.5	0.4324	0.2162	0.3244	0.0949	0.5481	0.0362
1.6	0.4339	0.2154	0.3446	0.1066	0.5734	0.0337
1.7	0.4355	0.2145	0.3646	0.1193	0.5996	0.0311
1.8	0.4371	0.2135	0.3842	0.1333	0.6267	0.0286
1.9	0.4388	0.2124	0.4035	0.1484	0.6547	0.0260
2.0	0.4406	0.2112	0.4224	0.1649	0.6837	0.0234

Table 10: *Effect of k on expected total cost*

k	ETC
1	1479.7
2	1045.0
3	1012.5
4	1040.3
5	1086.1

SPREADING OF A LIMITED LIFETIME INFORMATION IN NETWORKS EVOLVING BY PREFERENTIAL ATTACHMENT

Natalia Markovich

•

V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences,
Profsoyuznaya Str. 65, 117997 Moscow Russia
nat.markovich@gmail.com

Abstract

The paper is devoted to the information spreading (propagation) on random graphs evolving by a linear preferential attachment (PA) model. The PA is proposed to play a double role, namely, as the evolution model, i.e. the tool to add new edges and nodes to the network and (or) to remove existing nodes and edges, and as the spreading tool. We assume that a single message is to be propagated within a fixed time interval. In practice, a message may become old and not relevant. A node having a message instantaneously passes on information to one of its neighbour nodes which does not have the message yet. This neighbour may be either a node newly appended to the graph or an existing node. By probabilities of $\alpha-$, $\beta-$ and $\gamma-$schemes of the used PA model a new directed edge is drawn between a new node appended to the graph and an existing node or a new edge is drawn between a pair of existing nodes. By convention the propagation is provided if the new node (or one of the existing nodes) without the message has an incoming edge to an existing node having the information. Distributions of the number of nodes that received the message and the total number of nodes as well as the ratio of the latter random numbers in a fixed time interval with regard to parameters of the PA are obtained.

Keywords: information spreading, directed random graphs, evolution, linear preferential attachment

1. Introduction

Spreading information attracts interest due to many applications like multi-agent systems, internet traffic, parallel computation [1], [2], social networks and spreading of infections [3], [4], percolation [5] and gossip algorithms [6]. We consider the problem of spreading a single message through the directed network evolving by the preferential attachment (PA) model within a fixed time interval. A somewhat similar idea is propagated in [7] where the contact process that is evolving on graphs that are themselves evolving is studied. The PA plays a double role, namely, as the evolution model, i.e. the tool to add new edges and nodes to the network and (or) to remove existing nodes and edges, and as the spreading tool. In [8]-[9] the spreading of a unique message among a fixed number of nodes given beforehand has been considered. In [10] and [9] the PA evolution respectively without and with node and edge deletion was considered. However, a reasonable spreading time may be limited because the information may be outdated. Our objective is to obtain distributions of the number of nodes that received a single message and the total number of nodes in the graph as well as the ratio of the latter random variables (r.v.s) in the fixed time interval $[0, T^*]$ with regard to the parameters of the PA.

Let us denote the graph at evolution step k as $G(k) = (V(k), E(k))$, where $V(k)$ and $E(k)$ are sets of vertices (nodes) and edges, respectively. Let $N(k) = \|V(k)\|$ be the number of nodes in the network at the evolution step k, and let $\|A\|$ denote a cardinality of the set A. The evolution begins with an arbitrary initial directed graph $G(0)$ with at least one node and $\|E(0)\|$ edges. $N(0)$ and $\|E(0)\|$ are assumed to be fixed. We use the linear PA evolution model described in Section 2. Then $N(k)$ is a r.v. since a new node is appended to the graph with some probability.

Let $S(k)$ denote a set of nodes which have the message at the evolution step k. Indeed, $S(k) \subseteq V(k)$ holds. The spreading starts from an initial set of nodes $S(0)$ in which at least one node has the message. The spreading is unsuccesfull at step $k+1$ if a new node $j \in N(k+1)$ joins to node $i \notin S(k)$ by a new edge $(j \to i)$. Hence, $\|S(k)\|$ is a r.v. and $\|S(k)\| \leq N(k)$ holds.

Clocks of nodes are assumed to be asynchronized. As in [1], [2], we assume that new nodes arrive by Poisson time ticks of a global clock, see Fig. 1 (top).

Let $\{\tau_i\}$, $i = 0, 1, 2, ...$, $\tau_0 = 0$, be independent identically distributed (i.i.d.) exponential r.v.s with parameter λ. The sequence $\{\tau_i\}$ means the inter-arrival times of appended new nodes. The message may be propagated not at each tick of the global clock.

Since clock ticks build a Poissonian sequence, then the probability that the number of ticks $\nu(t)$ in time t is equal to $k = 0, 1, 2, ...$ is the following

$$P_k(t) \quad = \quad P\{\nu(t) = k\} = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \tag{1}$$

and the mean number of ticks in time t (or the renewal function) is $E(\nu(t)) = \lambda t$. By (1) we get $P_k(T^*)$ and $E(\nu(T^*))$ for a fixed time $T^* \geq 0$.

We aim to find the distributions of the number of nodes which have obtained the information by $K^* = \nu(T^*)$ evolution steps, i.e. $\|S(K^*)\|$, as well as of $N(K^*)$ and of the proportion of such nodes $\|S(K^*)\|/N(K^*)$. The information may be delivered to all nodes at some step k, i.e. $\|S(k)\| = N(k)$, but the propagation will be continued until the time T^* which cannot be exceeded. We suppose the evolution is without deletion of nodes and edges. Therefore, the propagated message cannot be lost since the number of nodes with the message does not decrease.

The paper is organized as follows. In Section 2, the PA model for the evolution of directed graphs is recalled. Section 3 contains our main results, namely, probability mass functions (pmfs) of $\|S(K^*)\|$ and $N(K^*)$ as well as the distribution of $\|S(K^*)\|/N(K^*)$. We finalize with conclusions. Proofs are presented in the Appendix.

2. Preferential attachment for directed graphs

By the PA model, networks are built recursively by adding nodes and edges in such a way that new nodes prefer to be connected to existing nodes if they have high node degrees. The PA networks are called scale-free [11]-[12]. It means that the node degree distribution is a power law. A discrete r.v. X exhibits a power-law distribution if

$$\mathrm{P}\left(X = i\right) \sim C i^{-(1+\iota)}, \quad i \to \infty,$$

holds for some positive constants C and ι.

We will use the so-called $\alpha-$, $\beta-$ and $\gamma-$schemes of the linear PA proposed in [11], [13] to model the evolution and information spreading. The latter PA model allows to build evolving random graphs with self-loops and multiple edges generated by the β-scheme. Examples of such evolution and spreading are shown in Fig. 1 (bottom).

Let us recall $\alpha-$, $\beta-$ and $\gamma-$schemes of the PA given in [13]. A type of the new edge is selected by flipping a 3-sided coin with probabilities α, β and γ such that $\alpha + \beta + \gamma = 1$. To this end, the iid trinomial r.v.s with values 1, 2 and 3 and the corresponding probabilities α, β and γ are generated. The parameters $\delta_{in}, \delta_{out} > 0$ allow us to determine the probabilities to select existing nodes when their in- or out-degrees are zero-valued. Let $I_k(w)$ and $O_k(w)$ denote the in- and out-degrees of node w at evolution step k. We assume $I_0(w)$ and $O_0(w)$, the in- and out-degrees in the initial graph $G(0)$, to be fixed.

Let us denote a complement as $A \setminus B$. The graph G_k is obtained from the existing graph G_{k-1} by the following $\alpha-$, $\beta-$ and $\gamma-$schemes.

- By the α-scheme, one appends a new node $v \in V(k) \setminus V(k-1)$ to $G(k-1)$, $k \geq 1$, and a new edge $(v \to w)$ to an existing node $w \in V(k-1)$ with probability α. The node w is chosen with a probability depending on its in-degree in $G(k-1)$

$$P_\alpha(k,w) \quad = \quad \frac{I_{k-1}(w)+\delta_{in}}{k-1+\delta_{in}N(k-1)}.$$

- By the β-scheme, one adds a new edge $(v \to w)$ to $E(k-1)$, $k \geq 1$, with probability β, where the existing nodes $v, w \in V(k-1) = V(k)$ are chosen independently from the nodes of $G(k-1)$ with probabilities

$$P_\beta(k,w,v) \quad = \quad \frac{O_{k-1}(v)+\delta_{out}}{k-1+\delta_{out}N(k-1)} \cdot \frac{I_{k-1}(w)+\delta_{in}}{k-1+\delta_{in}N(k-1)}.$$

- By the γ-scheme, one adds a new node $v \in V(k) \setminus V(k-1)$ to $G(k-1)$, $k \geq 1$, and an edge $(w \to v)$ with probability γ. The existing node $w \in V(k-1)$ is chosen with probability

$$P_\gamma(k,w) \quad = \quad \frac{O_{k-1}(w)+\delta_{out}}{k-1+\delta_{out}N(k-1)}.$$

Here, α, β, γ, δ_{in} and δ_{out} are parameters of the PA model. By convention, the information can be spread at step $k+1$ to a new node $v \notin S(k)$ having no information if a new edge is created by the $\alpha-$scheme or if a new edge is created by the $\beta-$scheme between two existing nodes v, w and one of them has the information (i.e. $v \in S(k)$ or $w \in S(k)$). In both cases, the new edge is to be directed to an existing node having the information.

3. Probability mass functions of $\|S(K^*)\|$ and $N(K^*)$

We assume that the nodes and edges are not removed from the graph. Then both $\|S(k)\|$ and $N(k)$ are non-decreasing functions in time. We prove that the success probability depends on the PA parameters α, β, δ_{in}, δ_{out}, and $N(K^*)$ has a Poisson distribution given a graph G_{K^*-1}.

Let the initial graph $G(0)$ contain a unique isolated node that has a message to be spread, i.e. $\|S(0)\| = \|V(0)\| = 1$, $\|E(0)\| = 0$ hold. We denote for brevity $\Omega_k = \{G(0), \cdots, G(k)\} = \{G_0, \cdots, G_k\}$ and the success probabilities for a given G_{k-1} as p_k.

Let $\sum_{c,k,j}$ denote the sum of all $\binom{k}{j} = k!/(j!(k-j)!)$ distinct index combinations among $\{j_1, j_2, ..., j_k\}$ of length j and $\mathbf{1}(A)$ denote the indicator of the event A. For example, for $k=3$ evolution steps, the combinations with exactly two successes are the following: $(101), (110), (011)$.

Lemma 1. The conditional pmf of $\|S(k)\|$ for a maximum number of evolution steps K^* in the fixed time T^* is the following. For $1 \leq i < K^*$ it holds

$$P\{\|S(K^*)\| \quad = \quad i|G_{K^*-1}\} = e^{-\lambda T^*} \sum_{k=i}^{\infty} \frac{(\lambda T^*)^k}{k!} P\{\|S(k)\| = i|G_{k-1}\}, \tag{2}$$

where

$$P\{\|S(k)\| = i|G_{k-1}\} = \sum_{c,k,i-1} \prod_{n=1}^{i-1} p_{j_n} \prod_{m=i}^{k} (1-p_{j_m}) \mathbf{1}\{k \geq i \geq 2\}$$

$$+ \quad \prod_{m=1}^{k} (1-p_m) \mathbf{1}\{k \geq i = 1\} + \prod_{n=1}^{k} p_n \mathbf{1}\{i = k+1\} = \psi(i,j,k), \tag{3}$$

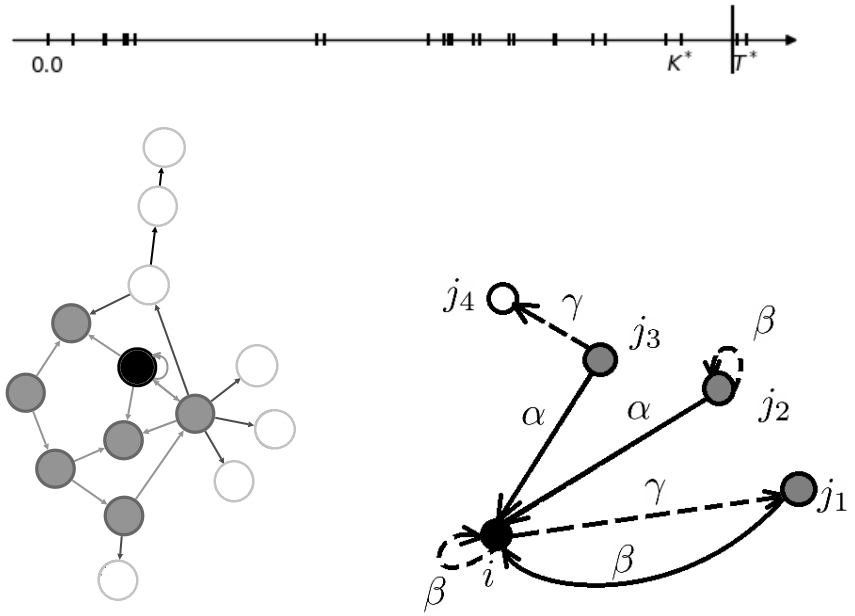


Figure 1: *Poissonian clock ticks with the fixed time T^* and the corresponding maximum number of ticks K^* are shown on top; spreading a message from an initial black node by the PA schemes with parameters $(\alpha, \beta, \gamma) = (0.35, 0.25, 0.4)$ at evolution step $k = 20$, where nodes that obtained the message are shown by grey color and nodes without the message by white color, arrows corresponding to successful spreading are shown by grey and to unsuccessful spreading by black (bottom, left); the scheme of the message spreading from node i to nodes $j_1 - j_4$ (bottom, right) with the same node coloring.*

$$P\{\|S(k)\| = i | G_{k-1}\} = 0, \qquad otherwise,$$

and success probabilities are

$$p_k = \alpha \sum_{w \in S(k-1)} P_\alpha(k, w) \quad + \quad \beta \sum_{w \in S(k-1)} \sum_{v \in V(k-1) \setminus S(k-1)} P_\beta(k, w, v), \quad k = 1, 2, ... \tag{4}$$

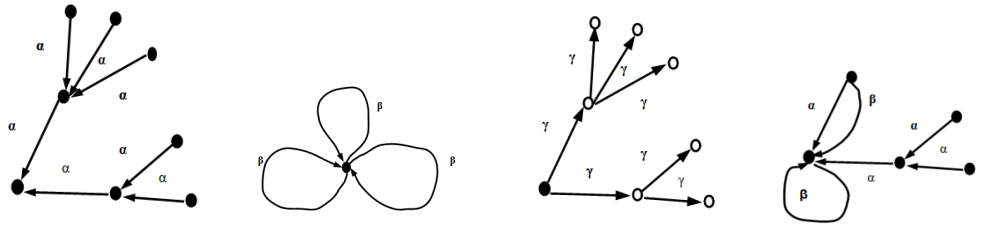
Remark 1. Note that the success probabilities are different at each evolution step and the first term in (3) corresponds to a Poisson binomial distribution.[1]

Remark 2. By (4), $p_1 = \alpha$ holds. It implies the probability that the first new node v is connected to a single node $w \in G(0)$ with the message and w shares the message with v. The second term in (4) equals to zero since the self-loop in the initial node built by the $\beta-$scheme does not lead to the success (sharing the message) or formally, since the set $V(k-1) \setminus S(k-1)$ is empty.

Corollary 1. The collection of conditional probabilities (2) forms a conditional probability distribution.

Example 1. Let us give examples of a full spreading and a full non-spreading of the message to all nodes in the graph. To this end, we consider the PA with parameters $(\alpha, \beta, \gamma) = \{(1,0,0), (0,1,0), (0,0,1)\}$. By (4), if $\alpha = 1$ holds at each step $k \geq 1$, then it implies $p_k = 1$, $\forall k \geq 1$ and the spreading to each new node, see Fig. 2(a); $\beta = 1$, $\forall k \geq 1$ leads to further non-spreading due to self-loops at the initial node, Fig. 2(b); $\gamma = 1$, $\forall k \geq 1$ means $p_k = 0$, $\forall k \geq 1$, i.e. no further spreading, Fig.2(c). The PA model with $\alpha + \beta = 1$, $\alpha \neq 0$, $\beta \neq 0$, $\gamma = 0$ implies $p_k = \alpha$, $\forall k \geq 1$ that means the full spreading. Despite the second term in (4) is equal to zero, $\beta-$scheme with $\beta \neq 0$ contributes to node degrees and thus, to $\sum_{w \in S(k-1)} P_\alpha(k, w)$, Fig. 2(d).

[1] An integer-valued r.v. X is called Poisson binomial and denoted as $X \sim PB(p_1, p_2, ..., p_k)$, if $X =^d \xi_1 + ... + \xi_k$, where $\xi_1, ..., \xi_k$ are independent Bernoulli r.v.s with parameters $p_1, p_2, \ldots, p_k$. The probability distribution of X is $P\{X = k\} = \sum_{A \in [k], \|A\| = j} \left(\prod_{i \in A} p_i \prod_{i \neq A} (1 - p_i)\right)$, where the sum ranges over all subsets of $[k] = \{1, ..., k\}$ of size j [14].



(a) (b) (c) (d)

Figure 2: *Examples of the PA-schemes with $(\alpha,\beta,\gamma)=(1,0,0)$ corresponding to a full spread tree (Fig. 2(a)); self-loops with $(\alpha,\beta,\gamma)=(0,1,0)$ (Fig. 2(b)) and the tree with $(\alpha,\beta,\gamma)=(0,0,1)$ (Fig.2(c)) both as full non-spreading; with (α,β,γ) such that $\alpha+\beta=1$, $\alpha\neq 0$, $\beta\neq 0$ corresponding to the full spreading (Fig. 2(d)).*

Lemma 2. The pmf of $N(k)$, $k\geq 1$ for a maximum number of evolution steps K^* in a fixed time T^* is the Poissonian:

$$P\{N(K^*)=j|G_{K^*-1}\} \quad = \quad \frac{a^{j-1}}{(j-1)!e^a}, \qquad a=(\alpha+\gamma)\lambda T^*, \quad j=1,2,...,K^*+1. \tag{5}$$

Corollary 2. If $(\alpha,\beta,\gamma)=(0,0,1)$ holds at each evolution step $k\geq 1$, then

$$P\{\|S(K^*)\|=1|G_{K^*-1}\}=1, \quad P\{N(K^*)=K^*+1|G_{K^*-1}\}=1. \tag{6}$$

If $(\alpha,\beta,\gamma)=(0,1,0)$ holds at each evolution step $k\geq 1$, then

$$P\{\|S(K^*)\|=N(K^*)=1|G_{K^*-1}\}=1. \tag{7}$$

If $(\alpha,\beta,\gamma)=(1,0,0)$ holds at each evolution step $k\geq 1$, then

$$P\{\|S(K^*)\|=N(K^*)=K^*+1|G_{K^*-1}\}=1. \tag{8}$$

Let us consider now the proportion of nodes which obtained the message to the total number of nodes $\|S(K^*)\|/N(K^*)$ in time T^*.

Lemma 3. Let K^* be a Poisson r.v. with the pmf $P\{K^*=k\}=(\lambda^*)^k e^{-\lambda^*}/k!$, $\lambda^*=\lambda T^*$, $k=0,1,2,....$ Then it holds

$$\begin{aligned} 1 \quad &\leq \quad e^{\lambda^*}P\{\|S(K^*)\|/N(K^*)\leq x|G_{K^*-1}\} \\ &\leq \quad e^{-(\alpha+\gamma)\lambda^*}\sum_{k=1}^{\infty}\frac{(\lambda^*)^k}{k!}\sum_{j=1}^{k+1}\frac{(\alpha+\gamma)^{j-1}(\lambda^*)^{j-1}}{(j-1)!}\sum_{i=1}^{\lfloor xj\rfloor}\psi(i,j,k)+1, \end{aligned} \tag{9}$$

where $\psi(i,j,k)$ is determined by (3).

Lemma 4. $P\{\|S(K^*)\|/N(K^*)\leq x|G_{K^*-1}\}$, $0<x\leq 1$, is linear in each $0\leq p_j\leq 1$, $j=1,...,k$ for each fixed x.

The maximum of $P\{\|S(K^*)\|/N(K^*)\leq x|G_{K^*-1}\}$ is achieved if at least one of the $\{p_j\}$ is equal to 1. The linearity does not depend on the way to select K^*.

4. Conclusions

We study the propagation of one message among nodes in an evolving network within a fixed time interval T^*. The propagation starts from the initial graph containing a single node with the

message to be spread. Considering the network evolution we assume that it can be modeled by the $\alpha-$, $\beta-$ and $\gamma-$ schemes of the linear PA model proposed in [13]. It implies that a new node appended to the network at some evolution step will likely link to an existing node with a large degree. The message may be spread not at each step of the evolution but only when a new edge is directed from a new node (or an existing node) without the message to the existing node with the message. These cases correspond to the $\alpha-$ and $\beta-$ PA schemes, respectively.

The pmfs of the number of nodes obtaining the message and the total number of nodes as well as the distribution function of their ratio in the PA evolved network at time T^* are derived.

5. Acknowledgements

The author was supported by the Russian Science Foundation RSF, project number 24-21-00183.

References

[1] Damon Mosk-Aoyama and Devavrat Shah. "Computing separable functions via gossip". In: *Proceedings of the twenty-fifth annual ACM symposium on Principles of distributed computing*. PODC06. ACM, July 2006. DOI: 10.1145/1146381.1146401.

[2] Keren Censor Hillel and Hadas Shachnai. "Partial information spreading with application to distributed maximum coverage". In: *Proceedings of the 29th ACM SIGACT-SIGOPS symposium on Principles of distributed computing*. PODC '10. ACM, July 2010. DOI: 10.1145/1835698.1835739.

[3] Siddharth Patwardhan et al. "Epidemic Spreading in Group-Structured Populations". In: *Physical Review X* 13.4 (Dec. 2023). ISSN: 2160-3308. DOI: 10.1103/physrevx.13.041054.

[4] Lang Zeng et al. "Contagion dynamics in time-varying metapopulation networks with node's activity and attractiveness". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 34.5 (May 2024). ISSN: 1089-7682. DOI: 10.1063/5.0204497.

[5] Peter Gracar and Arne Grauer. *Geometric scale-free random graphs on mobile vertices: broadcast and percolation times*. 2024. DOI: 10.48550/ARXIV.2404.15124.

[6] Devavrat Shah. "Gossip Algorithms". In: *Foundations and Trends® in Networking* 3.1 (2007), pp. 1–125. ISSN: 1554-0588. DOI: 10.1561/1300000014.

[7] Emmanuel Jacob, Amitai Linker, and Peter Mörters. *The contact process on dynamical scale-free networks*. 2022. DOI: 10.48550/ARXIV.2206.01073.

[8] Natalia M. Markovich and Maksim S. Ryzhov. "Information Spreading with Application to Non-homogeneous Evolving Networks". In: *Distributed Computer and Communication Networks*. Springer International Publishing, 2022, pp. 284–292. ISBN: 9783030971106. DOI: 10.1007/978-3-030-97110-6_22.

[9] Natalia M. Markovich and Maksim S. Ryzhov. "Information Spreading in Non-homogeneous Evolving Networks with Node and Edge Deletion". In: *Distributed Computer and Communication Networks: Control, Computation, Communications*. Springer Nature Switzerland, 2024, pp. 119–128. ISBN: 9783031504822. DOI: 10.1007/978-3-031-50482-2_10.

[10] Natalia M. Markovich and Maksim S. Ryzhov. "Information Spreading and Evolution of Non-Homogeneous Networks". In: *Advances in Systems Science and Applications* 22.2 (June 2022), pp. 21–33. DOI: 10.25728/assa.2022.22.2.1186.

[11] Béla Bollobás et al. "Directed scale-free graphs". In: Jan. 2003, pp. 132–139. DOI: 10.1145/644108.644133.

[12] Natalia Markovich and Marijus Vaičiulis. "Extreme Value Statistics for Evolving Random Networks". In: *Mathematics* 11.9 (May 2023), p. 2171. ISSN: 2227-7390. DOI: 10.3390/math11092171.

[13] Phyllis Wan et al. "Are extreme value estimation methods useful for network data?" In: *Extremes* 23.1 (Aug. 2019), pp. 171–195. ISSN: 1572-915X. DOI: 10.1007/s10687-019-00359-x.

[14] Wenpin Tang and Fengmin Tang. "The Poisson Binomial Distribution – Old & New". In: *Statistical Science* 38.1 (Feb. 2023). ISSN: 0883-4237. DOI: 10.1214/22-sts852.

Appendix

5.1. Proof of Lemma 1

Proof.

After the clock tick k, we have either $\|S(k)\| = \|S(k-1)\|$ or $\|S(k)\| = \|S(k-1)\| + 1$, $k = 1, 2, ...$, where $\|S(0)\| = 1$. The size increases, i.e. $\|S(k)\| = \|S(k-1)\| + 1$ holds, if a node $v \notin S(k-1)$ contacts a node $w \in S(k-1)$ and if the edge $(v \to w)$ is directed from v to w. The new edge $(v \to w)$ leading to the success (the increase of $\|S(k)\|$) may be created between two existing nodes $v, w \in V(k-1) = V(k)$ with probability $\beta P_\beta(k, w, v)$ if $v \in V(k-1) \setminus S(k-1) = V(k) \setminus S(k-1)$, $w \in S(k-1)$, or between a newly appended node $v \in V(k) \setminus V(k-1)$ and an existed node $w \in S(k-1)$ with probability $\alpha P_\alpha(k, w)$. Otherwise, the size does not increase, i.e. $\|S(k)\| = \|S(k-1)\|$. The $\gamma-$scheme does not lead to the information spreading from node w to node v by convention. Here, α, β and γ are defined as in Section 2.

The conditional success probability p_k to increase the number of nodes with the message at step k is the following

$$\begin{aligned} & p_k = E\{S(k) - S(k-1) | \Omega_{k-1}\} \qquad (10) \\ = \; & \alpha \sum_{w \in S(k-1)} P_\alpha(k, w) + \beta \sum_{w \in S(k-1)} \sum_{v \in V(k-1) \setminus S(k-1)} P_\beta(k, w, v), \quad \text{for } k = 2, 3, ..., \\ p_1 \; = \; & \alpha P_\alpha(1, w \in S(0)) = \alpha, \qquad \text{for} \qquad k = 1. \qquad (11) \end{aligned}$$

Since $G(0)$ contains a unique node, $V(0) \setminus S(0) = \emptyset$, the $\beta-$scheme can be applied at step $k = 1$, but it does not lead to the "success". The $\beta-$scheme may provide self-loops.

Let us consider the case $k \geq i \geq 2$. The creation of a direction of a new edge may be taken as the experiment. The experiments are independent since they are generated by the iid trinomial r.v.s.. We obtain the Poisson binomial pmf

$$P\{\|S(k)\| = i | \Omega_{k-1}\} = \sum_{c,k,i-1} \prod_{n=1}^{i-1} p_{j_n} \prod_{m=i}^{k} (1 - p_{j_m}), \qquad (12)$$

where the sequence $\{p_{j_n}\}$ is determined by (10), (11). By the Markov property one can substitute Ω_{k-1} by G_{k-1} in the latter equality and further.
For $k = i = 1$ it holds

$$P\{\|S(1)\| = 1 | G_0\} = \gamma P_\gamma(1, w \in S(0)) + \beta P_\beta(1, w \in S(0)) = \gamma + \beta = 1 - p_1 \qquad (13)$$

due to $P_\gamma(1, w \in S(0)) = P_\beta(1, w \in S(0) = 1$.
For $k > i = 1$ it follows

$$\begin{aligned} P\{\|S(k)\| = 1 | G_{k-1}\} \; &= \; \big(\gamma P_\gamma(1, w \in S(0)) + \beta P_\beta(1, w \in S(0))\big) \prod_{m=2}^{k} (1 - p_m) \\ &= \; (\gamma + \beta) \prod_{m=2}^{k} (1 - p_m) = \prod_{m=1}^{k} (1 - p_m), \qquad (14) \end{aligned}$$

The event $\{\|S(k)\| = 1 | G_{k-1}\}$ means that the set $S(k)$ contains only the initial node with the message.

Let us consider the case $0 \leq k < i$. For $k = 0$ it trivially follows

$$P\{\|S(0)\| = 1\} \; = \; 1, \qquad P\{\|S(0)\| = j\} = 0, \qquad j > 1. \qquad (15)$$

For $i > k \geq 1$ we have

$$P\{\|S(k)\| = i|G_{k-1}\} = \begin{cases} \alpha^k \prod_{j=1}^{k} \sum_{\omega \in S(j-1)} P_\alpha(j,\omega), & i = k+1, \\ 0, & i > k+1. \end{cases} \tag{16}$$

Note that it holds

$$\alpha^k \prod_{j=1}^{k} \sum_{\omega \in S(j-1)} P_\alpha(j,\omega) = \prod_{n=1}^{k} p_n. \tag{17}$$

Summarizing (12)-(17) we get (3). Then it holds

$$P\{\|S(K^*)\| = i|G_{K^*-1}\} = \sum_{k=i}^{\infty} P\{\|S(k)\| = i|K^* = k, G_{k-1}\}P\{K^* = k\}$$
$$= \sum_{k=i}^{\infty} P\{\|S(k)\| = i|G_{k-1}\}P_k(T^*)$$

since $S(k)$ and K^* are independent. Using $P_k(T^*)$ in (1) we get (2).

■

5.2. Proof of Corollary 1

Proof.

Since $\|S(k)\|$ has a Poisson binomial distribution and $P\{\|S(k)\| = 0\} = 0$, the collection of probabilities $P\{\|S(K^*)\| = i|\Omega_{K^*-1}\} = P\{\|S(K^*)\| = i|G_{K^*-1}\}$ forms a probability distribution:

$$\sum_{i=0}^{\infty} P\{\|S(K^*)\| = i|G_{K^*-1}\} = 1.$$

■

5.3. Proof of Lemma 2

Proof. It holds $N(k+1) = N(k) + 1$ with probability $\alpha + \gamma$ and $N(k+1) = N(k)$ with probability β. Hence, for a fixed $k \geq 1$ and $N(0) = 1$ we get

$$P\{N(k) = i+1|G_{k-1}\} = \binom{k}{i}(\alpha+\gamma)^i \beta^{k-i}, \qquad i = 0, 1, 2, ..., k.$$

Since $\sum_{i=0}^{\infty} a^i/i! = e^a$, $a > 0$ holds and due to independence of the r.v.s K^* and $N(k)$, $N(k) \leq k+1$, it follows

$$P\{N(K^*) = i+1|G_{K^*-1}\} = \sum_{k=i}^{\infty} P\{N(k) = i+1|K^* = k, G_{K^*-1}\}P\{K^* = k\}$$
$$= \sum_{k=i}^{\infty} \binom{k}{i}(\alpha+\gamma)^i \beta^{k-i} P_k(T^*) = \sum_{k=i}^{\infty} \frac{(\alpha+\gamma)^i(1-\alpha-\gamma)^{k-i}}{i!(k-i)!}(\lambda T^*)^k e^{-\lambda T^*}$$
$$= \frac{a^i}{i!e^a}, \qquad \text{where} \quad a = (\alpha+\gamma)\lambda T^*.$$

■

5.4. Proof of Corollary 2

Proof. Let $(\alpha,\beta,\gamma)=(0,0,1)$ hold. By (4), we have $p_k=0$, $k=1,2,....$ By (3) it follows

$$P\{\|S(k)\|=i|G_{k-1}\}=\prod_{m=1}^{k}(1-p_m)\mathbf{1}\{k\geq i=1\}=1$$

and hence, the first statement in (6) holds. Furthermore, we get

$$\begin{aligned}
& P\{N(K^*)=K^*+1|G_{K^*-1}\}\\
= & \sum_{k=1}^{\infty}P\{N(k)=k+1|K^*=k,G_{K^*-1}\}P\{K^*=k\}+P\{N(0)=1\}P\{K^*=0\}\\
= & \sum_{k=0}^{\infty}P\{K^*=k\}=1.
\end{aligned}$$

For $(\alpha,\beta,\gamma)=(0,1,0)$ (7) follows by (3) since the $\beta-$scheme forms only self-loops in the initial node and $p_k=0$, $k=1,2,....$ For $(\alpha,\beta,\gamma)=(1,0,0)$ (8) follows since each new node gets the message and $p_k=1$, $k=1,2,....$ ■

5.5. Proof of Lemma 3

Proof.

We have by (1)

$$\begin{aligned}
& P\{\|S(K^*)\|/N(K^*)\leq x|\Omega_{K^*-1}\}\\
= & \sum_{j=1}^{\infty}P\{\|S(K^*)\|\leq xj|N(K^*)=j,\Omega_{K^*-1}\}P\{N(K^*)=j|\Omega_{K^*-1}\}\\
= & \sum_{j=1}^{\infty}\sum_{i=1}^{\lfloor xj\rfloor}P\{\|S(K^*)\|=i,N(K^*)=j|\Omega_{K^*-1}\}\\
= & \sum_{k=1}^{\infty}\sum_{j=1}^{k+1}\sum_{i=1}^{\lfloor xj\rfloor}P\{\|S(k)\|=i,N(k)=j|K^*=k,\Omega_{k-1}\}P\{K^*=k\}\\
+ & \mathbf{1}\{i=j=1,k=0\}P\{K^*=0\}\\
\geq & \mathbf{1}\{\|S(0)\|=N(0)=1\}P\{K^*=0\}=e^{-\lambda^*},
\end{aligned}\tag{18}$$

where $x\in(0,1]$. Using the Markov property, we can substitute Ω_k by G_k in expressions above. Note that the r.v.s $N(k)$ and $\|S(k)\|$ are dependent, $1\leq\|S(k)\|\leq N(k)\leq k+1$ for any $k\geq 0$ holds. Then by (1), (3) and (5) we get

$$\begin{aligned}
& P\{\|S(K^*)\|/N(K^*)\leq x|G_{K^*-1}\}\\
\leq & \sum_{k=1}^{\infty}\sum_{j=1}^{k+1}\sum_{i=1}^{\lfloor xj\rfloor}P\{\|S(k)\|=i|K^*=k,G_{k-1}\}P\{N(k)=j|K^*=k,G_{k-1}\}P\{K^*=k\}+e^{-\lambda^*}\\
= & \sum_{k=1}^{\infty}\sum_{j=1}^{k+1}\sum_{i=1}^{\lfloor xj\rfloor}\psi(i,j,k)\frac{a^{j-1}}{(j-1)!e^a}\frac{(\lambda^*)^ke^{-\lambda^*}}{k!}+e^{-\lambda^*}\\
= & e^{-(1+\alpha+\gamma)\lambda^*}\sum_{k=1}^{\infty}\sum_{j=1}^{k+1}\sum_{i=1}^{\lfloor xj\rfloor}\psi(i,j,k)\frac{(\alpha+\gamma)^{j-1}(\lambda^*)^{j-1+k}}{(j-1)!k!}+e^{-\lambda^*}.
\end{aligned}\tag{19}$$

By (18) and (19) the statement (9) follows. ■

5.6. Proof of Lemma 4

Proof.

Let us substitute the Poisson binomial r.v. $\|S(k)\|$, $k \geq 1$ in (18) by the sum of independent Bernoulli r.v.s $X_1, ..., X_k$ with success probabilities $p_1, ..., p_k$, respectively. Denoting

$$b_{i,j,k} = P\{X_2 + ... + X_k = i, N(k) = j | K^* = k, G_{k-1}\},$$
$$a_{i,j,k} = P\{X_2 + ... + X_k = i-1, N(k) = j | K^* = k, G_{k-1}\} - b_{i,j,k},$$

we get that the next probability is linear in p_1:

$$\begin{aligned}
& P\{\|S(K^*)\|/N(K^*) \leq x | G_{K^*-1}\} \\
= & \sum_{k=1}^{\infty} \sum_{j=1}^{k+1} \sum_{i=1}^{\lceil xj \rceil} P\{X_1 + ... + X_k = i, N(k) = j | K^* = k, G_{k-1}\} P_k(T^*) \\
+ & \mathbf{1}\{i = j = 1, k = 0\} P_0(T^*) \\
= & \sum_{k=1}^{\infty} \sum_{j=1}^{k+1} \sum_{i=1}^{\lceil xj \rceil} \Big(P\{X_2 + ... + X_k = i, N(k) = j | K^* = k, G_{k-1}\}(1-p_1) + \mathbf{1}\{i = j = 1, k = 0\} e^{-\lambda T^*} \\
+ & P\{X_2 + ... + X_k = i-1, N(k) = j | K^* = k, G_{k-1}\} p_1 \Big) P_k(T^*) + \mathbf{1}\{i = j = 1, k = 0\} e^{-\lambda T^*} \\
= & \sum_{k=1}^{\infty} \sum_{j=1}^{k+1} \sum_{i=1}^{\lceil xj \rceil} \Big(b_{i,j,k} + p_1 a_{i,j,k} \Big) P_k(T^*) + \mathbf{1}\{i = j = 1, k = 0\} e^{-\lambda T^*}.
\end{aligned}$$

One can rewrite the latter expression in the linear form regarding p_1:

$$\phi(x, p_1) \quad = \quad f_1(x) p_1 + f_2(x),$$

where

$$\phi(x, p_1) = P\{\|S(K^*)\|/N(K^*) \leq x | G_{K^*-1}\},$$
$$f_1(x) = \sum_{k=1}^{\infty} \sum_{j=1}^{k=1} \sum_{i=1}^{\lceil xj \rceil} a_{i,j,k} P_k(T^*), \qquad f_2(x) = \sum_{k=1}^{\infty} \sum_{j=1}^{k+1} \sum_{i=1}^{\lceil xj \rceil} b_{i,j,k} P_k(T^*) + e^{-\lambda T^*}.$$

Due to symmetry, the linearity can be shown in any p_j, $1 \leq j \leq k+1$. ■

INFINITE-SERVER QUEUEING SYSTEM WITH WAITING NEGATIVE CUSTOMERS

Danil Korolev
Svetlana Moiseeva*
Alexander Moiseev
Sardor Saidov

•

National Research Tomsk State University, Tomsk, Russia
Karshi State University, Department of Mathematics Computer Sciences, Karshi, Uzbekistan

Abstract

The paper considers a queueing system with waiting negative customers. The system has two arrival processes: one for positive customers, another for negative ones. In this model, arrived negative customers do not contact with present positive ones but immediately destroy new positive arrivals. To find the joint probability distribution of the number of positive and negative customers, we use the method of asymptotic analysis under the condition of high rate of arrivals. As the result, we derive the approximation of characteristic function of the distribution. Using it, we obtain that one-dimensional stationary probability of the number of positive customers can be approximated by Gaussian distribution. Using numerical evaluations and simulation experiments, we estimate an error and an applicability area of the approximation.

Keywords: queueing system, negative customers, asymptotic analysis

1. Introduction

In today's world, service industries such as customer services, transportation systems, medical facilities and many others are faced with increasing flows of customers, where both positive and negative calls require effective management. From e-commerce to social media, our digital infrastructure is subjected to a constant stream of requests, which can be thought of as customers in a queueing system. In real life, the information technology domain faces various challenges in handling a variety of customers and identifying optimal management strategies. For example, malicious attacks or outages can negatively affect the performance and security of information systems.

Consider the process of processing network requests in a data center. In this system, positive customers are requests for access to specific data or resources that are processed by the data center. Such requests are received continuously in the system for a random time. On the other hand, negative customers can be represented by malicious attacks or DDoS (Distributed Denial of Service) attacks. When such a negative customer arrives, it does not interact with

*The research is supported by Russian Science Foundation according to the research project No.24-21-00454, https://rscf.ru/project/24-21-00454

the previous processed positive customer, but begins its destructive activity of overloading the network infrastructure. When new data access requests arrive, the interference caused by DDoS attacks can block or destroy those requests.

Systems and networks with negative customers were proposed by E. Gelenbe in [1], [2], where he introduced a new class of queueing networks with two types of customers. The first type of customers are regular customers and the server serves them in the usual way. Such customers are called positive or persistent customers. A positive customers obeys the established service and routing rules that determine the dynamics of the network under consideration. On the other hand, the second type of negative customers acts as a signal to induce a positive customers at a node, if any, to leave the node immediately. Such networks were originally used to model biophysical neural networks. In this context, a node represents a neuron. Positive and negative requests routed in the network represent excitation and inhibition signals that increase or decrease in unit potential of the neuron to which they arrive. Extensions of the original Gelenbe network lead to a universal class called G-networks in the literature because it provides a unifying framework for queuing systems and neural networks. This analogy has been discussed in detail in survey articles [3], [4].

The first papers on G-networks and single-node queues with negative customers (G-queues) were written at the end of the 20th century. Thus, the main results were developed in a short period of time. Perhaps because of this novelty and special interest in G-networks, many authors simultaneously worked on similar research topics. The most comprehensive review including more than 300 references was presented in [5].

Queues with negative customers can be used to model failures and packet loss, task completion under speculative parallelism, faulty components in production systems and server breakdowns, and the reaction network of interacting molecules. Negative queries with appropriate destruction discipline extend the modeling capabilities of these queueing models, since real-world phenomena such as failures, packet loss in radio interfaces, load balancing, and disasters can be easily captured. For example, in [6], a retrial queue with positive and negative arrival was considered. Negative arrivals lead to server failures, after which the server should perform a repair. The authors of [7, 8] considered retrial queues in which the negative customers clear the pool of retrial calls. The authors used the method of asymptotic analysis for the study. Queueing systems with catastrophes, server failures, and repairs are considered in [9, 10, 11, 12]. The authors analyzed steady-state behavior of the systems under different conditions.

In [13], an infinite-server queue with negative customers with waiting was considered. Situation when an incoming negative customer instantly destroys a serving positive customer were considered. In contrast to the mentioned work, the presented paper considers the case when an incoming negative customer does not interact with the available positive ones, but waits for the arrival of a new positive customer, destroys it and leaves the system. Let us imagine a system of automatic deployment and rollback of changes in the IT infrastructure. Positive customers are requests to deploy a new version of an application, and negative customers are requests to roll back to the previous version. The deployment server acts as a service appliance. When a deployment request arrives, it occupies the server, the deployment takes place, and the server is released. If a rollback request arrives during the deployment, it 'hangs', waiting for the deployment to complete. Once the new deployment is complete, and a positive customer (to deploy the next change or update) comes in, a negative customer (rollback) "intercepts" this new positive customer, cancels it (preventing the new deployment from starting) and starts the rollback process to the previous version, at which point both customers (rollback and canceled deployment) leave the system.

The rest text is organized as follows. In Sec. 2, mathematical model in the form of a queueing system is proposed, system of Kolmogorov equations for it is constructed. In Sec. 3, we perform asymptotic analysis to find the solution of the system. In Sec. 4, results of numerical experiments are provided. Basing on them, we estimate the accuracy of the obtained approximations and area of their applicability.

2. Mathematical model

2.1. Problem statement

Consider a queueing system with an unlimited number of servers. Let two Poisson flows of customers arrive at the system. The first one with intensity λ delivers normal (positive) customers, the other one with intensity α delivers negative customers. A positive customer arrived at the system when there are no negative customers in it, occupies any free server and immediately starts service within it during exponentially distributed period with parameter μ. We consider a problem with waiting negative customers. This means that when a negative customer arrives, it does not interact with positive customers present in the system but waits for the arrival of a new positive customer. The capacity of the waiting queue for negative customers is unbounded. As soon as a new positive customer arrives, if there are negative customers in the system, one of them destroys this new customer and they both leave the system. Figure 1 shows the scheme of the described system. During the system evolution we may have any combinations of the positive and negative customers presence. For example, some positive customers can arrive when there are no any negative ones in the system. They go to the servers. After that, new negative customers can arrive and they will stay in the system until new positive ones arrive. So, we may have both positive and negative customers being in the system at the same time.

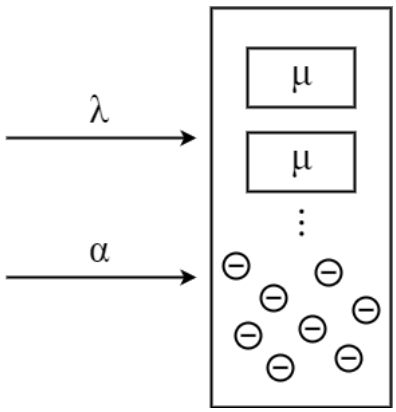


Figure 1: *Queueing system with waiting negative customers*

The paper studies Markov random process $\{i(t), l(t)\}$, where $i(t)$ is the number of positive and $l(t)$ is the number of negative customers in the system at time moment t.

2.2. System of Kolmogorov equations

For probability distribution $P(i,l,t) = \mathrm{P}\{i(t) = i, l(t) = l\}$ of the considered Markov process, we can construct the following system of Kolmogorov differential equations for steady-state regime. For this purpose, let us take time t tending to infinity in probabilities $P(i,l,t)$ and denote $P(i,l) = P(i,l,\infty)$:

$$\begin{aligned}
-(\lambda+\alpha)P(0,0) + \mu P(1,0) + \lambda P(0,1) &= 0,\\
-(\lambda+\alpha+i\mu)P(i,0) + \lambda P(i-1,0) + (i+1)\mu P(i+1,0) + \lambda P(i,1) &= 0,\\
-(\lambda+\alpha)P(0,l) + \mu P(1,l) + \alpha P(0,l-1) + \lambda P(0,l+1) &= 0,\\
-(\lambda+\alpha+i\mu)P(i,l) + \alpha P(i,l-1) + (i+1)\mu P(i+1,l) + \lambda P(i,l+1) &= 0.
\end{aligned} \tag{1}$$

Question about the stability condition will be discussed later (see Sec. 3.1).

Let us introduce partial characteristic functions

$$H(u,l) = \sum_{i=0}^{\infty} e^{jui} P(i,l),\ j = \sqrt{-1},\ l = 0,1,2,\ldots,\ u \in \mathbb{R}.$$

Using the method of partial characteristic functions, system (1) can be written in the form

$$j\mu(1-e^{-ju})\frac{dH(u,0)}{du}+(\lambda e^{ju}-(\lambda+\alpha))H(u,0)+\lambda H(u,1)=0,$$

$$j\mu(1-e^{-ju})\frac{dH(u,l)}{du}-(\lambda+\alpha)H(u,l)+\alpha H(u,l-1)+\lambda H(u,l+1)=0. \quad (2)$$

So, we have obtained system of infinite number of differential equations with variable parameters for infinite number of unknown functions $H(u,l)$, $l=0,1,2,\dots$ Unfortunately, we do not see ways of its direct solution. Because this reason, we apply a special technique "the asymptotic analysis method" [14] which allows to obtain approximation of the system solution under some conditions.

3. Asymptotic analysis

3.1. First-order asymptotic analysis

Since the direct solution of the obtained system of differential equations is not possible, we apply the method of asymptotic analysis [14] under the condition of the equivalent growing intensities of arrival processes. This condition can be defined as

$$\lambda=\lambda N,\quad \alpha=\alpha N, \quad (3)$$

where we suppose $N\to\infty$. Also, we name this condition as a condition of high intensity of arrivals. The method of asymptotic analysis allows to obtain limit solution of system (2) and use the derived limit expressions for functions $H(u,l)$ as approximations of solution of system (2) when parameter N has big values. So, for numerical evaluations, we take enough big values of N in the obtained expressions (see section 4).

Substituting (3) into (2), we obtain the following system of equations:

$$\frac{1}{N}j\mu(1-e^{-ju})\frac{dH(u,0)}{du}+(\lambda e^{ju}-(\lambda+\alpha))H(u,0)+\lambda H(u,1)=0,$$

$$\frac{1}{N}j\mu(1-e^{-ju})\frac{dH(u,l)}{du}-(\lambda+\alpha)H(u,l)+\alpha H(u,l-1)+\lambda H(u,l+1)=0,\ l\geq 1. \quad (4)$$

Let us denote $\frac{1}{N}=\varepsilon$ and perform substitutions

$$u=\varepsilon w,\qquad H(u,l)=F_1(w,l,\varepsilon).$$

Then system (4) can be rewritten as follows:

$$\varepsilon j\mu(1-e^{-jw\varepsilon})\frac{dF_1(w,0,\varepsilon)}{\varepsilon dw}+(\lambda e^{jw\varepsilon}-(\lambda+\alpha))F_1(w,0,\varepsilon)+\lambda F_1(w,1,\varepsilon)=0, \quad (5)$$

$$\varepsilon j\mu(1-e^{-jw\varepsilon})\frac{dF_1(w,l,\varepsilon)}{\varepsilon dw}-(\lambda+\alpha)F_1(w,l,\varepsilon)+\alpha F_1(w,l-1,\varepsilon)+\lambda F_1(w,l+1,\varepsilon)=0,\ l\geq 1.$$

Consider function $F_1(w,\varepsilon)=\sum_{l=0}^{\infty}F_1(w,l,\varepsilon)$ and sum equations of the system (5). Then we obtain the expression

$$e^{-jw\varepsilon}j\mu\frac{dF_1(w,\varepsilon)}{dw}+\lambda F_1(w,0,\varepsilon)=0.$$

Performing limit transition $\varepsilon\to 0$, we obtain

$$j\mu F_1'(w)+\lambda F_1(w,0)=0, \quad (6)$$

where $F_1(w)=\lim_{\varepsilon\to 0}F_1(w,\varepsilon)$.

Let $\varepsilon \to 0$ in (5). Using notation $F_1(w,l) = \lim_{\varepsilon\to 0} F_1(w,l,\varepsilon)$, we obtain the system of equations

$$\alpha F_1(w,0) + \lambda F_1(w,1) = 0, \tag{7}$$

$$-(\lambda+\alpha)F_1(w,l) + \alpha F_1(w,l-1) + \lambda F_1(w,l+1) = 0,\ l \geq 1.$$

We will seek the solution of this system in the form of $F_1(w,l) = \Phi_1(w)r(l)$, where $\{r(l)\}$ is the stationary probability distribution of the number of negative customers in the system. Substituting this expression into (7), we obtain:

$$\alpha r(0) + \lambda r(1) = 0, \tag{8}$$

$$-(\lambda+\alpha)r(l) + \alpha r(l-1) + \lambda r(l+1) = 0,\ l \geq 1.$$

Solving this system, we obtain

$$r(0) = 1 - \frac{\alpha}{\lambda},$$

$$r(l) = r(0)\Big(\frac{\alpha}{\lambda}\Big)^l,\ l = 1,2,\dots$$

We see that system (8) can be solved in terms of probabilities only if $\alpha < \lambda$. This is the stability condition. It also obviously follows from the model: under the opposite inequality, the number of negative customers is increasing unlimitedly.

Using $F_1(w,l) = \Phi_1(w)r(l)$ and taking into account that

$$F_1(w) = \sum_{l=0}^{\infty} F_1(w,l) = \Phi(w)\sum_{l=0}^{\infty} r(l) = \Phi(w),$$

we find $\Phi_1(w)$ from equation (6):

$$j\mu\Phi_1'(w) + \lambda\Phi_1(w)r(0) = 0,$$

$$\frac{d\Phi_1(w)}{dw} = -\frac{\lambda}{j\mu}\Phi_1(w)r(0),$$

$$\frac{d\Phi_1(w)}{\Phi_1(w)} = -\frac{\lambda}{j\mu}r(0)dw,$$

$$\Phi_1(w) = \exp\{j\kappa_1 w\}, \tag{9}$$

where

$$\kappa_1 = \frac{\lambda}{\mu}r(0) = \frac{\lambda-\alpha}{\mu}.$$

3.2. Second-order asymptotic analysis

Taking into account result (9), let us represent functions $H(u,l)$ in the form

$$H(u,l) = H_2(u,l)\exp\{j\kappa_1 uN\}, \tag{10}$$

where $H_2(u,l)$ are some functions ($l = 0,1,2,\dots$). Substituting (10) into (4), we obtain the following system of equations:

$$\frac{j\mu}{N}(1-e^{-ju})\frac{dH_2(u,0)}{du} - \kappa_1\mu(1-e^{-ju})H_2(u,0) +$$

$$+(\lambda e^{ju} - (\lambda+\alpha))H_2(u,0) + \lambda H_2(u,1) = 0, \tag{11}$$

$$\frac{j\mu}{N}(1-e^{-ju})\frac{dH_2(u,l)}{du} - \kappa_1\mu(1-e^{-ju})H_2(u,l) -$$

$$-(\lambda+\alpha)H_2(u,l) + \alpha H_2(u,l-1) + \lambda H_2(u,l+1) = 0.$$

Let us denote $\frac{1}{N} = \varepsilon^2$ and perform substitutions

$$u = \varepsilon w, \qquad H_2(u,l) = F_2(w,l,\varepsilon).$$

Then we derive

$$\begin{gathered}
\varepsilon j\mu(1-e^{-j\varepsilon w})\frac{dF_2(w,0,\varepsilon)}{dw} - \kappa_1\mu(1-e^{-j\varepsilon w})F_2(w,0,\varepsilon)+ \\
+(\lambda e^{j\varepsilon w} - (\lambda+\alpha))F_2(w,0,\varepsilon) + \lambda F_2(w,1,\varepsilon) = 0, \qquad (12) \\
\varepsilon j\mu(1-e^{-j\varepsilon w})\frac{dF_2(w,l,\varepsilon)}{dw} - \kappa_1\mu(1-e^{-j\varepsilon w})F_2(w,l,\varepsilon)- \\
-(\lambda+\alpha)F_2(w,l,\varepsilon) + \alpha F_2(w,l-1,\varepsilon) + \lambda F_2(w,l+1,\varepsilon) = 0.
\end{gathered}$$

Similarly to [14] and taking into account features of the model under consideration, we will look for solution $F_2(w,l,\varepsilon)$ in the form of expansion

$$F_2(w,l,\varepsilon) = \Phi_2(w)(r(l) + jw\varepsilon g(l)) + O(\varepsilon^2),$$

where $g(l)$ are some unknown function of discrete argument ($l = 0,1,2,...$). Now we substitute this expression into system (12):

$$\varepsilon j\mu(1-e^{-j\varepsilon w})\frac{d\Phi_2(w)(r(0)+jw\varepsilon g(0)+O(\varepsilon^2))}{dw} - +\kappa_1\mu(1-e^{-j\varepsilon w})\Phi_2(w)(r(0)+jw\varepsilon g(0)+O(\varepsilon^2))+$$

$$+(\lambda e^{j\varepsilon w} - (\lambda+\alpha))\Phi_2(w)(r(0)+jw\varepsilon g(0)+O(\varepsilon^2)) + \lambda\Phi_2(w)(r(1)+jw\varepsilon g(1)+O(\varepsilon^2)) = 0, \quad (13)$$

$$\begin{gathered}
\varepsilon j\mu(1-e^{-j\varepsilon w})\frac{d\Phi_2(w)(r(l)+jw\varepsilon g(l)+O(\varepsilon^2))}{dw} - +\kappa_1\mu(1-e^{-j\varepsilon w})\Phi_2(w)(r(l)+jw\varepsilon g(l)+O(\varepsilon^2))- \\
-(\lambda+\alpha)\Phi_2(w)(r(l)+jw\varepsilon g(l)+O(\varepsilon^2)) + \alpha\Phi_2(w)(r(l-1)+jw\varepsilon g(l-1)+O(\varepsilon^2))+ \\
+\lambda\Phi_2(w)(r(l+1)+jw\varepsilon g(l+1)+O(\varepsilon^2)) = 0.
\end{gathered}$$

In this system, we expand exponents into the Taylor series until power ε^2:

$$\begin{gathered}
\varepsilon j\mu(1-(1-j\varepsilon w))(\Phi_2'(w)(r(0)+jw\varepsilon g(0)) + \Phi_2(w)j\varepsilon g(0))- \\
-\kappa_1\mu\left(j\varepsilon w + \frac{(j\varepsilon w)^2}{2}\right)\Phi_2(w)(r(0)+jw\varepsilon g(0))+ \\
+\left(\lambda\left(1+j\varepsilon w+\frac{(j\varepsilon w)^2}{2}\right) - (\lambda+\alpha)\right)\Phi_2(w)(r(0)+jw\varepsilon g(0))+ \\
+\lambda\Phi_2(w)(r(1)+jw\varepsilon g(1)) + O(\varepsilon^3) = 0, \qquad (14) \\
\varepsilon j\mu(1-(1-j\varepsilon w))(\Phi_2'(w)(r(l)+jw\varepsilon g(l)) + \Phi_2(w)j\varepsilon g(l))- \\
-\kappa_1\mu\left(j\varepsilon w + \frac{(j\varepsilon w)^2}{2}\right)\Phi_2(w)(r(l)+jw\varepsilon g(l))- \\
-(\lambda+\alpha)\Phi_2(w)(r(l)+jw\varepsilon g(l)) + \alpha\Phi_2(w)(r(l-1)+jw\varepsilon g(l-1))+ \\
+\lambda\Phi_2(w)(r(l+1)+jw\varepsilon g(l+1)) + O(\varepsilon^3) = 0.
\end{gathered}$$

Let us consider in (14) terms of different powers of ε. Taking into account only terms without ε, we obtain

$$\begin{gathered}
-\alpha r(0) + \lambda r(1) = 0, \qquad (15) \\
-(\lambda+\alpha)r(l) + \alpha r(l-1) + \lambda r(l+1) = 0,
\end{gathered}$$

which coincides with system (8).

Taking into account only terms with ε in system (14), we obtain the following system of the inhomogeneous finite-difference equations with respect to the discrete-argument function $g(l)$:

$$-\alpha g(0) + \lambda g(1) = r(0)(\lambda - \kappa_1 \mu), \tag{16}$$

$$-(\lambda + \alpha)g(l) + \alpha g(l-1) + \lambda g(l+1) = \kappa_1 \mu r(l).$$

Consider the corresponding homogeneous system of equations

$$-\alpha g(0) + \lambda g(1) = 0,$$

$$-(\lambda + \alpha)g(l) + \alpha g(l-1) + \lambda g(l+1) = 0.$$

The general solution of this system has the form

$$g(l) = C\Big(\frac{\alpha}{\lambda}\Big)^l,$$

where C is an arbitrary constant, which we find from the boundary condition

$$\sum_{l=0}^{\infty} g(l) = C \sum_{l=0}^{\infty} \Big(\frac{\alpha}{\lambda}\Big)^l = C \cdot \frac{1}{1 - \frac{\alpha}{\lambda}} = 1,$$

implying

$$C = 1 - \frac{\alpha}{\lambda}.$$

Then the solution of the inhomogeneous system is the sum of the general solution of the homogeneous system and the partial solution of the inhomogeneous system (16). So, we can write the solution of the inhomogeneous system (16) in the form

$$g(l) = \Big(1 - \frac{\alpha}{\lambda}\Big)\Big(\frac{\alpha}{\lambda}\Big)^l + \frac{\kappa_1 \mu}{\alpha}\Big(1 - \frac{\alpha}{\lambda}\Big)\Big(\frac{\alpha}{\lambda}\Big)^l. \tag{17}$$

Let us consider terms of system (14) which has power ε^2:

$$\Phi_2'(w)r(0) - \kappa_1 w \Phi_2(w) g(0) - \kappa_1 \frac{w}{2}\Phi_2(w) r(0) +$$

$$+\frac{\lambda}{\mu}\Phi_2(w)\frac{w}{2}r(0) + \frac{\lambda}{\mu} w \Phi_2(w) g(0) = 0,$$

$$\Phi_2'(w)r(l) - \kappa_1 w \Phi_2(w) g(l) - \kappa_1 \frac{w}{2}\Phi_2(w) r(l) = 0,\ l \geq 1.$$

Let us summarize the equations of this system. Denoting

$$G = \sum_{l=0}^{\infty} g(l) = \sum_{l=0}^{\infty} \Big(1 - \frac{\alpha}{\lambda}\Big)\Big(\frac{\alpha}{\lambda}\Big)^l + \frac{\kappa_1 \mu}{\alpha}\Big(1 - \frac{\alpha}{\lambda}\Big)\Big(\frac{\alpha}{\lambda}\Big)^l =$$

$$= 1 + \frac{\lambda - \alpha}{\alpha}\frac{1}{1 - \frac{\alpha}{\lambda}} = 1 + \frac{\lambda}{\alpha},$$

$$\sum_{l=0}^{\infty} r(l) = \sum_{l=0}^{\infty} \Big(1 - \frac{\alpha}{\lambda}\Big)\Big(\frac{\alpha}{\lambda}\Big)^l = \Big(1 - \frac{\alpha}{\lambda}\Big)\left(\frac{1}{1 - \frac{\alpha}{\lambda}}\right) = 1,$$

we obtain:

$$\Phi_2'(w)\sum_l r(l) - \kappa_1 w \Phi_2(w)\sum_l g(l) - \kappa_1 \frac{w}{2}\Phi_2(w)\sum_l r(l) +$$

$$+\frac{\lambda}{\mu}\Phi_2(w)\frac{w}{2}r(0) + \frac{\lambda}{\mu} w \Phi_2(w) g(0) = 0,$$

$$\Phi_2'(w) - \kappa_1 w\Phi_2(w)G - \kappa_1\frac{w}{2}\Phi_2(w) + \frac{\lambda}{\mu}\Phi_2(w)\frac{w}{2}r(0) + \frac{\lambda}{\mu}w\Phi_2(w)g(0) = 0.$$

After grouping the terms, we obtain a first-order differential equation with separating variables:

$$\frac{d\Phi_2(w)}{\Phi_2(w)} = w\left[\kappa_1\left(\frac{1}{2} + \frac{\lambda}{\alpha}\right) + \frac{\lambda}{\mu}\left(\frac{3}{2}\left(1 - \frac{\alpha}{\lambda}\right)\right)\right],$$

which has solution

$$\Phi_2(w) = \exp\left\{-\frac{\kappa_2 w^2}{2}\right\},$$

where

$$\kappa_1 = \frac{\lambda - \alpha}{\mu}, \qquad \kappa_2 = \kappa_1\left(\frac{1}{2} + \frac{\lambda}{\alpha}\right) + \frac{\lambda}{\mu}\left(\frac{3}{2}\left(1 - \frac{\alpha}{\lambda}\right)\right). \tag{18}$$

Turning back to functions $H_2(u,l)$, we obtain:

$$H_2(u,l) = F_2(w,l,\varepsilon) = \Phi_2(w)(r(l) + jw\varepsilon g(l)) + O(\varepsilon^2) =$$

$$= \exp\left\{-\frac{\kappa_2 w^2}{2}\right\}\left(1 - \frac{\alpha}{\lambda}\right)\left(\frac{\alpha}{\lambda}\right)^l = \exp\left\{-\frac{\kappa_2\frac{u^2}{\varepsilon^2}}{2}\right\}\left(1 - \frac{\alpha}{\lambda}\right)\left(\frac{\alpha}{\lambda}\right)^l \approx$$

$$\approx \exp\left\{-\frac{\kappa_2 u^2 N}{2}\right\}\left(1 - \frac{\alpha}{\lambda}\right)\left(\frac{\alpha}{\lambda}\right)^l.$$

Sign $\approx$ is used because ε tends to 0 (this is equivalent to $N \to \infty$) but we take its pre-limit non-zero value to obtain pre-limit expressions for functions $H_2(u,l)$ and $H(u,l)$.

Given the expression (8), we obtain the partial characteristic function of the joint distribution of the number of positive and negative customers:

$$H(u,l) = \exp\left\{juN\kappa_1 + \frac{(ju)^2 N\kappa_2}{2}\right\}\left(1 - \frac{\alpha}{\lambda}\right)\left(\frac{\alpha}{\lambda}\right)^l,$$

where κ_1 and κ_2 are determined by expressions (18).

If we sum up this expressions over $l = 0, 1, 2, \ldots$, we can obtain the characteristic function of the one-dimensional random process of the number of positive customers in the system in the steady-state regime:

$$H(u) = \exp\left\{juN\kappa_1 + \frac{(ju)^2 N\kappa_2}{2}\right\}.$$

Thus, the asymptotic stationary probability distribution of the number of positive customers in the system under consideration is Gaussian with mean $N\kappa_1$ and variance $N\kappa_2$.

If we put $u = 0$, we obtain the probability distribution of the number of negative customers in the system:

$$P(l) = \left(1 - \frac{\alpha}{\lambda}\right)\left(\frac{\alpha}{\lambda}\right)^l,\ l \geq 0.$$

That is, the one-dimensional distribution of the number of negative customers in the system in the stationary regime in the specified asymptotic condition is geometric.

4. Numerical examples and estimation of approximation accuracy

To evaluate the accuracy of the obtained approximation and to establish the limits of its applicability, a series of numerical experiments were conducted, in which the asymptotic distributions were numerically compared with the empirical ones obtained as results of simulation modeling. A program for simulation of the model with waiting negative customers was developed in Python.

The following parameter values were chosen for the experiments: $\lambda = 2$, $\alpha = 1$, $\mu = 1$. In this case, $\kappa_1 = 1$ and the asymptotic average number of positive customers in the system equals to N, which is convenient for understanding the applicability area of the results. Figure 2 shows plots of asymptotic probability distributions of the number of positive customers and probability distributions obtained by the simulation model for different values of parameter N.

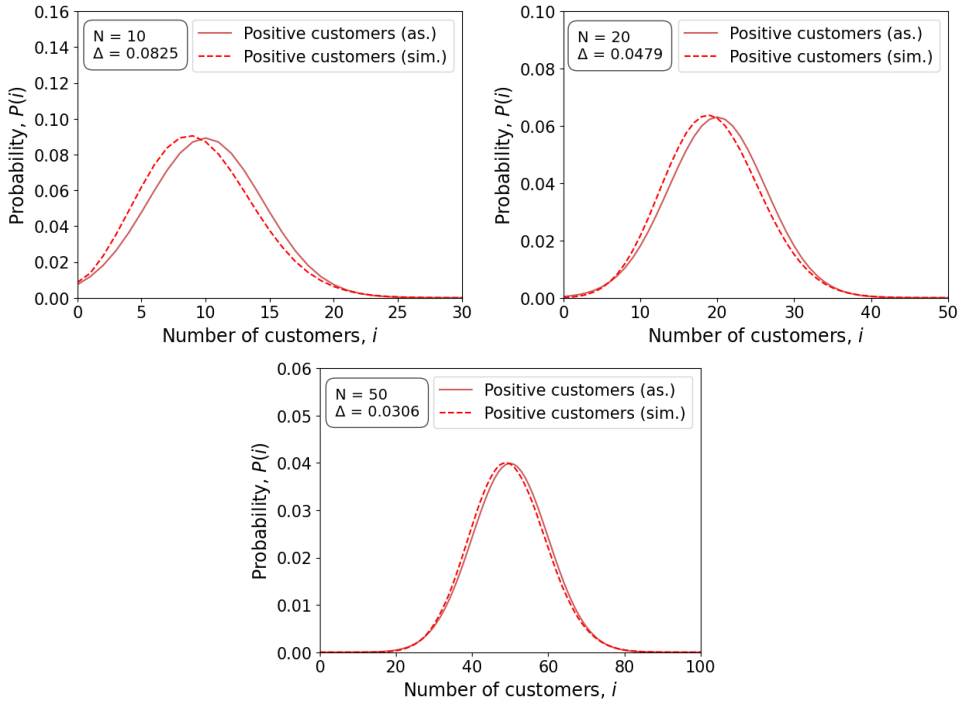


Figure 2: *Comparison of empirical and asymptotic probability distributions of the number of positive customers for various values of parameter N*

To estimate the error of the obtained asymptotic probability distribution of the number of positive and negative customers, let us use the Kolmogorov distance, which is calculated by the formula

$$\Delta = \max_{0 \le i < \infty} \left| \sum_{n=0}^{i} \Big(P_{sim}(n) - P_{as}(n) \Big) \right|,$$

where P_{sim} is the probability distribution obtained from the simulation experiment and P_{as} is the asymptotic probability distribution.

Table 1 presents values of the Kolmogorov distances between the asymptotic distributions of positive customers and the corresponding distributions obtained from the simulations. As we can see from the table and graphs, the Kolmogorov distance decreases while the high-rate parameter N increases. If we choose error $\Delta \le 0.05$ as an acceptable value, then we can conclude that the obtained Gaussian approximation for distribution of positive customers can be applied for $N \ge 20$.

Table 1: *Kolmogorov distance for the distribution of the number of positive customers*

N	10	20	50	100	200	500	1000
Δ	0.0825	0.0479	0.0306	0.0231	0.0172	0.0124	0.0041

Figure 3 shows the plot of the asymptotic probability distribution of the number of negative customers and the corresponding probability distribution obtained from the simulation model for

$N = 100$. The curves almost match, and the Kolmogorov distance is small enough even for small values of asymptotic parameter N (e.g. for $N = 10$, $\Delta = 0.0023$).

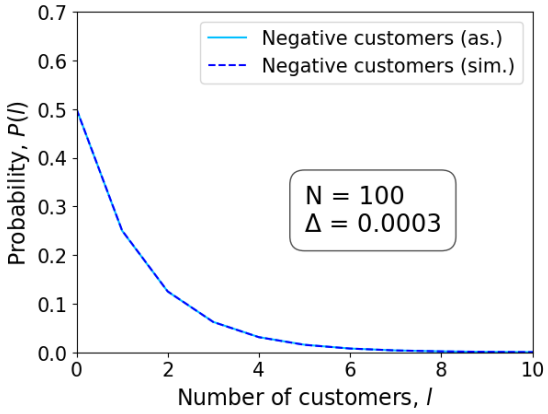


Figure 3: *Comparison of empirical and asymptotic probability distributions of the number of negative customers*

5. Conclusion

In the paper, we consider a mathematical model of a queueing system with waiting negative customers. The asymptotic analysis of this system is performed, the partial characteristic functions of the joint distribution of the number of positive and negative customers are obtained. One-dimensional asymptotic probability distributions of the number of positive and negative customers under the condition of a high rate of arrivals are derived. A series of numerical experiments and comparison with the results of simulation modeling are carried out. Based on these results, we have estimated the accuracy and applicability area of the obtained approximations.

The study may be extended to the models with MAP arrivals or some other non-Poisson processes as well as with an arbitrary distribution of service times.

References

[1] Erol Gelenbe, Peter Glynn, and Karl Sigman. "Queues with negative arrivals". In: *Journal of Applied Probability* 28.1 (Mar. 1991), pp. 245–250. ISSN: 1475-6072. DOI: 10.2307/3214756.

[2] Jean-Michel Fourneau, Erol Gelenbe, and Rina Suros. "G-networks with multiple classes of negative and positive customers". In: *Theoretical Computer Science* 155.1 (Feb. 1996), pp. 141–156. ISSN: 0304-3975. DOI: 10.1016/0304-3975(95)00018-6.

[3] J.R. Artalejo. "G-networks: A versatile approach for work removal in queueing networks". In: *European Journal of Operational Research* 126.2 (Oct. 2000), pp. 233–249. ISSN: 0377-2217. DOI: 10.1016/s0377-2217(99)00476-2.

[4] Erol Gelenbe and Ali Labed. "G-networks with multiple classes of signals and positive customers". In: *European Journal of Operational Research* 108.2 (July 1998), pp. 293–305. ISSN: 0377-2217. DOI: 10.1016/s0377-2217(97)00371-8.

[5] Tien Van Do. "An initiative for a classified bibliography on G-networks". In: *Performance Evaluation* 68.4 (Apr. 2011), pp. 385–394. ISSN: 0166-5316. DOI: 10.1016/j.peva.2010.10.001.

[6] Ioannis Dimitriou. "A mixed priority retrial queue with negative arrivals, unreliable server and multiple vacations". In: *Applied Mathematical Modelling* 37.3 (Feb. 2013), pp. 1295–1309. ISSN: 0307-904X. DOI: 10.1016/j.apm.2012.04.011.

[7] Natalya P. Meloshnikova, Anatoly A. Nazarov, and Ekaterina A. Fedorova. "Study of the M/M/1 retrial queueing system with disasters". In: *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika* 68 (2024), pp. 28–37. ISSN: 2311-2085. DOI: 10.17223/19988605/68/3.

[8] Natalya Meloshnikova, Ekaterina Fedorova, and Danil Plaksin. "Asymptotic Analysis of a Multiserver Retrial Queue with Disasters in the Service Block". In: *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*. Springer Nature Switzerland, 2023, pp. 55–67. ISBN: 9783031329906. DOI: 10.1007/978-3-031-32990-6_5.

[9] Sherif I. Ammar. "Transient behavior of a two-processor heterogeneous system with catastrophes, server failures and repairs". In: *Applied Mathematical Modelling* 38.7–8 (Apr. 2014), pp. 2224–2234. ISSN: 0307-904X. DOI: 10.1016/j.apm.2013.10.033.

[10] Yunna Han et al. "On a retrial queue with negative customers, passive breakdown, and delayed repairs". In: *Probability in the Engineering and Informational Sciences* 38.2 (Oct. 2023), pp. 428–447. ISSN: 1469-8951. DOI: 10.1017/s0269964823000219.

[11] B. Krishna Kumar et al. "Transient analysis of a single server queue with catastrophes, failures and repairs". In: *Queueing Systems* 56.3–4 (Mar. 2007), pp. 133–141. ISSN: 1572-9443. DOI: 10.1007/s11134-007-9014-0.

[12] Ahmed M. K. Tarabia. "Transient and steady state analysis of an M/M/1 queue with balking, catastrophes, server failures and repairs". In: *Journal of Industrial & Management Optimization* 7.4 (2011), pp. 811–823. ISSN: 1553-166X. DOI: 10.3934/jimo.2011.7.811.

[13] A. A. Nazarov and N. M. Feropontova. "Study of the Interaction of Fluxes of Annihilating Particles". In: *Russian Physics Journal* 58.8 (Dec. 2015), pp. 1118–1127. ISSN: 1573-9228. DOI: 10.1007/s11182-015-0621-7.

[14] Alexander Moiseev and Anatoly Nazarov. "Queueing network MAP-(GI/∞)K with high-rate arrivals". In: *European Journal of Operational Research* 254.1 (Oct. 2016), pp. 161–168. ISSN: 0377-2217. DOI: 10.1016/j.ejor.2016.04.011.

PARTIAL ASYMPTOTIC ANALYSIS METHOD FOR TWO-CLASS RETRIAL QUEUE WITH CONSTANT RETRIAL RATE*

Ekaterina Fedorova, Anatoly Nazarov, Elena Bulgakova

•

Tomsk State University, Tomsk, Russian Federation, Russia
ekat_fedorova@mail.ru nazarov.tsu@gmail.com Len190703@gmail.com

Abstract

In the paper, a single-server retrial queueing system with two types of arrivals and a constant retrial policy is considered as a mathematical model of a multimodal telecommunication network. Service, inter-arrival and inter-retrial times have exponential distributions. The constant retrial policy means that only the first customer from an orbit performs an attempt to get a service. The method of partial asymptotic analysis under a condition of a heavy load of one class of customers is proposed. The formula for the asymptotic characteristic function of the stationary marginal probability distribution of the number of customers of one class is derived. In addition, the system stability conditions are discussed. Some numerical examples are presented.

Keywords: two-class retrial queueing system, constant retrial policy, partial asymptotic analysis, heavy load

1. Introduction

Modern telecommunication networks have several types of transmitted information (text, sound, image, service information, etc.). The transmission and processing of heterogeneous data need more complex preliminary analysis. An example of heterogeneous networks is the multimodal system [1, 2]. Most studies of multimodal systems are based on simulation. Although various mathematical models of such systems have been proposed, their analytical study is almost not carried out because of the need to study multidimensional random processes. For networks optimizing and planing, assessing their reliability and efficiency, queuing theory is usually used [3].

Retrial queueing systems (RQ or queueing system with repeated calls) [4, 5] are new class of queuing models widely applied in various communication systems [6, 7]. In spite of the large number of studies of retrial queueing systems in various configurations, heterogeneous models are weak investigated. Retrial queues with several types of customers (and several orbits too) are called as multiclass retrial queues and considered in [8, 9, 10, 11, 12, 13, 14]. Most of the cited papers are devoted only the stability analysis, while probability distributions or even means of processes under study are hardly investigated.

In the paper, we propose an partial asymptotic analysis method for two-dimensional Markov process analyzing. The asymptotic analysis method [15, 16] is applied for study of various classes

*SUPPORTED BY RUSSIAN SCIENCE FOUNDATION ACCORDING TO THE RESEARCH PROJECT NO.24-21-00454, HTTPS://RSCF.RU/PROJECT/24-21-00454/

of queueing systems. It allows us to solve Kolmogorov's equations under some asymptotic (i.e. limit) condition when they cannot be solved directly. But usually this method is used for one-dimensional processes under study. Here we try to apply the asymptotic analysis approach for multiclass queueing models. Also we demonstrate the method applying for retrial queues with constant retrial policy for the first time.

The rest of the paper is organized as follows. In Section 2, the mathematical model and the process under study are described. In Section 2.1, Kolmogorov equations are written and expressions for the stationary probabilities of the server states are obtained. In Section 2.2, there are discussions about stability and partial stability areas of the model. Section 3 is devoted to applying the original partial asymptotic analysis method under a heavy load of one class of customers. In Section 4, some numerical results are presented. Section 5 consists of some conclusions.

2. Mathematical Model

We consider a retrial queueing system with two classes of customers. The n-th class customer comes to the system according to Poisson arrival process with parameter λ_n, where $n = 1, 2$. There is one server. If the server is idle, an arrival (the n-th class customer) starts its servicing during the exponentially distributed random time with rate μ_n. If the arrival finds the server busy, it goes to the corresponding orbit and waits for a random time exponentially distributed with rate σ_n there. There are two orbits for each class customers. We consider the constant retrial policy, which means that only the first customer from an orbit has access to the server. The model structure is presented in Figure 1.

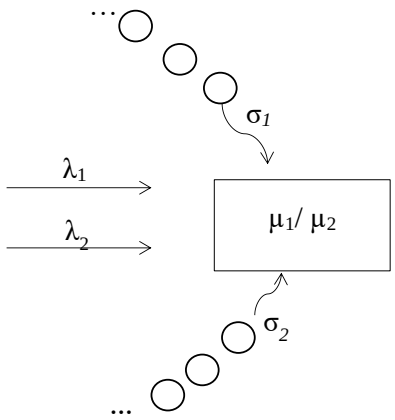


Figure 1: *Two-class retrial queueing system with constant retrial rate*

Let us denote the random processes of the number of customers in the n-th orbit by $i_n(t)$, where $n = 1, 2$. Process $k(t)$ determines the state of the server in the following way: $k(t) = 0$ if the server is free and $k(t) = n$ if the n-th class customer is servicing. Process $\{k(t), i_1(t), i_2(t)\}$ is a three-dimensional continuous-time Markov chain.

2.1. Kolmogorov equations

Denote the probability that the server has state k and there are i_1 customers in the first orbit and i_2 customers in the second orbit at time t by $P\{k(t) = k, i_1(t) = i_1, i_2(t) = i_2\} = P(k, i_1, i_2, t)$. Let us write the following system of Kolmogorov equations for $P(k, i_1, i_2, t)$:

$$\begin{cases} \dfrac{\partial P(0,i_1,i_2,t)}{\partial t} = -(\lambda_1+\lambda_2+\delta_{i_1}\sigma_1+\delta_{i_2}\sigma_2)P(0,i_1,i_2,t)+ \\ +\mu_1 P(1,i_1,i_2,t)+\mu_2 P(2,i_1,i_2,t), \\ \dfrac{\partial P(1,i_1,i_2,t)}{\partial t} = -(\lambda_1+\lambda_2+\mu_1)P(1,i_1,i_2,t)+\lambda_1 P(0,i_1,i_2,t)+ \\ +\sigma_1 P(0,i_1+1,i_2,t)+\lambda_2 P(1,i_1,i_2-1,t)+\lambda_1 P(1,i_1-1,i_2,t), \\ \dfrac{\partial P(2,i_1,i_2,t)}{\partial t} = -(\lambda_1+\lambda_2+\mu_2)P(2,i_1,i_2,t)+\lambda_2 P(0,i_1,i_2,t)+ \\ +\sigma_2 P(0,i_1,i_2+1,t)+\lambda_1 P(2,i_1-1,i_2,t)+\lambda_2 P(2,i_1,i_2-1,t), \end{cases} \tag{1}$$

where $\delta_i = \{0 \text{ if } i=0; 1 \text{ if } i \neq 0\}$ is Kronecker delta.

Then we obtain the following balance equations in steady state

$$\begin{cases} -(\lambda_1+\lambda_2+\delta(i_1)\sigma_1+\delta(i_2)\sigma_2)P(0,i_1,i_2)+ \\ +\mu_1 P(1,i_1,i_2)+\mu_2 P(2,i_1,i_2)=0, \\ -(\lambda_1+\lambda_2+\mu_1)P(1,i_1,i_2)+\lambda_1 P(0,i_1,i_2)+ \\ +\sigma_1 P(0,i_1+1,i_2)+\lambda_2 P(1,i_1,i_2-1)+\lambda_1 P(1,i_1-1,i_2)=0, \\ -(\lambda_1+\lambda_2+\mu_2)P(2,i_1,i_2)+\lambda_2 P(0,i_1,i_2)+ \\ +\sigma_2 P(0,i_1,i_2+1)+\lambda_1 P(2,i_1-1,i_2)+\lambda_2 P(2,i_1,i_2-1)=0. \end{cases} \tag{2}$$

Let us introduce the partial characteristic functions as follows

$$H(k,u_1,u_2) = \sum_{i_1=0}^{\infty}\sum_{i_2=0}^{\infty} e^{ju_1 i_1}\cdot e^{ju_2 i_2} P(k,i_1,i_2),$$

$$h_2(u_1) = \sum_{i_1=0}^{\infty} e^{ju_1 i_1} P(0,i_1,0), \quad h_1(u_2) = \sum_{i_2=0}^{\infty} e^{ju_2 i_2} P(0,0,i_2).$$

Then Equations (2) are rewritten as

$$\begin{cases} -(\lambda_1+\lambda_2+\sigma_1+\sigma_2)H(0,u_1,u_2)+ \\ +\mu_1 H(1,u_1,u_2)+\mu_2 H(2,u_1,u_2)+\sigma_1 h_1(u_2)+\sigma_2 h_2(u_1)=0, \\ (\lambda_1(e^{ju_1}-1)+\lambda_2(e^{ju_2}-1)-\mu_1)H(1,u_1,u_2)+ \\ +(\lambda_1+\sigma_1 e^{-ju_1})H(0,u_1,u_2)-\sigma_1 e^{-ju_1}h_1(u_2)=0, \\ (\lambda_1(e^{ju_1}-1)+\lambda_2(e^{ju_2}-1)-\mu_2)H(2,u_1,u_2)+ \\ +(\lambda_2+\sigma_2 e^{-ju_2})H(0,u_1,u_2)-\sigma_2 e^{-ju_2}h_2(u_1)=0. \end{cases} \tag{3}$$

By summing up all equations of System (3), we obtain an additional equation

$$\begin{gathered} (\sigma_1(e^{-ju_1}-1)+\sigma_2(e^{-ju_2}-1))H(0,u_1,u_2)+ \\ +(\lambda_1(e^{ju_1}-1)+\lambda_2(e^{ju_2}-1))(H(1,u_1,u_2)+H(2,u_1,u_2))+ \\ +\sigma_1(1-e^{-ju_1})h_1(u_2)+\sigma_2(1-e^{-ju_2})h_2(u_1)=0. \end{gathered} \tag{4}$$

Let us denote the stationary probability of the server states by $R_k = H(k,0,0)$ for $k=0,1,2$ and the stationary probability that the server is idle and the corresponding orbit is empty by $r_1 = h_1(0)$ and $r_2 = h_2(0)$.

From Equations (3)–(4), the following equations for probabilities R_k, r_n can be obtained

$$\begin{cases} -(\lambda_1+\lambda_2+\sigma_1+\sigma_2)R_0+\mu_1 R_1+\mu_2 R_2\sigma_1 r_1+\sigma_2 r_2=0, \\ -\mu_1 R_1+(\lambda_1+\sigma_1)R_0-\sigma_1 r_1=0, \\ -\mu_2 R_2+(\lambda_2+\sigma_2)R_0-\sigma_2 r_2=0, \\ -\sigma_2 R_0+\lambda_2(R_1+R_2)\sigma_2 r_2=0, \\ -\sigma_1 R_0+\lambda_1(R_1+R_2)\sigma_1 r_1=0. \end{cases} \tag{5}$$

Taking into account the normalization condition $R_0 + R_1 + R_2 = 1$ in Equations (5), it is easy to derive expressions for probabilities R_k, r_n.

$$\begin{aligned} R_1 &= \frac{\lambda_1}{\mu_1}, \quad R_2 = \frac{\lambda_2}{\mu_2}, \quad R_0 = 1 - \frac{\lambda_1}{\mu_1} - \frac{\lambda_2}{\mu_2}, \\ r_1 &= \frac{1}{\sigma_1}\left[\sigma_1 - (\lambda_1 + \sigma_1)\left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}\right)\right], \\ r_2 &= \frac{1}{\sigma_2}\left[\sigma_2 - (\lambda_2 + \sigma_2)\left(\frac{\lambda_1}{\mu_1} + \frac{\lambda_2}{\mu_2}\right)\right]. \end{aligned} \tag{6}$$

2.2. Stability

From Expressions (6), due to non-negative values of probabilities, it can be obtained the necessary conditions of the system stability:

$$\begin{gathered} \frac{\lambda_1}{\mu_1} - \frac{\lambda_2}{\mu_2} < 1, \\ \sigma_1 > \lambda_1 \frac{R_1 + R_2}{R_0}, \quad \sigma_2 > \lambda_2 \frac{R_1 + R_2}{R_0} \end{gathered} \tag{7}$$

that matches with the results [9], where it is proved that inequalities (7) are necessary and sufficient conditions of system stability.

Also, we can rewrite inequalities (7) as

$$R_1 + R_2 < \frac{\sigma_1}{\sigma_1 + \lambda_1}, \quad R_1 + R_2 < \frac{\sigma_2}{\sigma_2 + \lambda_2}. \tag{8}$$

where $R_1 = \lambda_1/\mu_1$ and $R_2 = \lambda_2/\mu_2$ have the meaning of the corresponding class load parameters.

In other words, failure to satisfy condition (8) entails negative values of stationary probabilities r_n, so there is no steady state for the customers of the corresponding class (i.e., the number of customers in the orbit tends to infinity).

Note that the existence of steady-state distributions in the considered queueing system with constant retrial policy depends on values of all model parameters, while a classical retrial queue stability does not depend on values of retrial rates.

In addition, we demonstrate the stability and instability areas of the model through numerical examples. First, we denote the maximum possible (unachievable) values of each class arrival and retrial parameters when the system is in steady state by S_n and G_n, respectively. Sometimes, parameters S_1 and S_2 are called throughput. In other words, parameters S_n and G_n are a constraint on system rates λ_n and σ_n, if they are not met, inequalities (8) are not satisfied and the system is not in steady state. From expressions (6), we can obtain that

$$G_1 = \frac{\lambda_1(\lambda_1\mu_2 + \lambda_2\mu_1)}{\mu_1\mu_2 - \lambda_1\mu_2 - \lambda_2\mu_1}, \quad G_2 = \frac{\lambda_2(\lambda_1\mu_2 + \lambda_2\mu_1)}{\mu_1\mu_2 - \lambda_1\mu_2 - \lambda_2\mu_1} \tag{9}$$

and

$$\begin{aligned} S_1 &= \left(\frac{\lambda_2}{\mu_2} + \frac{\sigma_1}{\mu_1}\right)\frac{\mu_1}{2}\left\{\sqrt{1 + 4\frac{\frac{\sigma_1}{\mu_1}\left(1 - \frac{\lambda_2}{\mu_2}\right)}{\left(\frac{\lambda_2}{\mu_2} + \frac{\sigma_1}{\mu_1}\right)^2}} - 1\right\}, \\ S_2 &= \left(\frac{\lambda_1}{\mu_1} + \frac{\sigma_2}{\mu_2}\right)\frac{\mu_2}{2}\left\{\sqrt{1 + 4\frac{\frac{\sigma_2}{\mu_2}\left(1 - \frac{\lambda_1}{\mu_1}\right)}{\left(\frac{\lambda_1}{\mu_1} + \frac{\sigma_2}{\mu_2}\right)^2}} - 1\right\}. \end{aligned} \tag{10}$$

So, the system stability conditions can be written as $\sigma_n > G_n$ or $\lambda_n < S_n$.

Then we demonstrate the areas of the system stability and instability in numerical examples. Let the system parameters be $\mu_1 = 1$, $\mu_2 = 2$. Expressions (9) are illustrated in Figure 2 depending on values of σ_n for arrival rates $\lambda_1 = 0.1$, $\lambda_2 = 0.3$. Expressions (10) are shown in Figure 3 depending on values of λ_n for $\sigma_1 = 0.1$, $\sigma_2 = 0.2$.

As we can see in Figures 3,2, there are four zones:

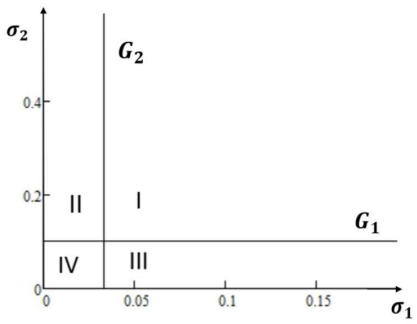


Figure 2: *Stability areas vs.* σ_1, σ_2

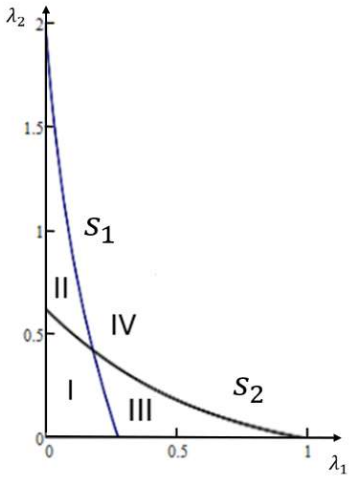


Figure 3: *Stability areas vs.* λ_1, λ_2

- zone I – stability of both orbits (both $\lambda_n < S_n$ and $\sigma_n > G_n$);
- zone II – stability of the first orbit, while the number of customers on the second orbit tends to infinity ($\lambda_1 < S_1$ and $\sigma_1 > G_1$, but $\lambda_2 > S_2$ and $\sigma_2 < G_2$);
- zone III – stability of the second orbit, while the number of customers on the first orbit tends to infinity ($\lambda_2 < S_2$ and $\sigma_2 > G_2$, but $\lambda_1 > S_1$ and $\sigma_1 < G_1$);
- zone IV – instability of both orbits (both $\lambda_n > S_n$ and $\sigma_n < G_n$).

3. Method

Equations (3)–(4) contain five unknown functions of two variables, so the direct solution is impossible. We propose applying an asymptotic analysis approach, namely, the method of partial asymptotic analysis under the condition of a heavy load of one class of customers to Equations (3)–(4) solving. Let us choose a condition of a heavy load of the second class of customers, so $\lambda_2 \to S_2$. Graphically, we can demonstrate the asymptotic condition as the red curve in Figure 4. Further we will derive the steady state asymptotic probability distribution of the number of customers in the first orbit. Obviously, the asymptotic results must be closer to the exact probability distribution with λ_2 closer to S_2.

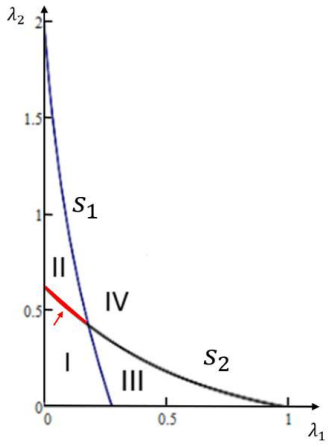


Figure 4: *Asymptotic condition (red curve)*

3.1. Asymptotic equations

For System (3)-(4) solving, we introduce an infinitesimal parameter ε. Then the asymptotic condition can be rewritten as $\lambda_2 = S_2(1-\varepsilon)$, where $\varepsilon \to 0$. Also we introduce the following notations

$$\begin{array}{l} u_2 = \varepsilon w, \quad H(k,u_1,u_2) = F(k,u_1,w,\varepsilon), \\ h_1(u_2) = f_1(w,\varepsilon), \quad h_2(u_1) = \varepsilon f_2(u_1). \end{array} \tag{11}$$

Substituting (11) into Equations (3)–(4), we obtain

$$\left\{\begin{array}{l} -(\lambda_1 + S_2(1-\varepsilon) + \sigma_1 + \sigma_2)F(0,u_1,w,\varepsilon) + \\ +\mu_1 F(1,u_1,w,\varepsilon) + \mu_2 F(2,u_1,w,\varepsilon) + \sigma_1 f_1(w,\varepsilon) + \sigma_2 \varepsilon f_2(u_1) = O(\varepsilon^2), \\ (\lambda_1(e^{ju_1}-1) + S_2(1-\varepsilon)(e^{jw\varepsilon}-1) - \mu_1)F(1,u_1,w,\varepsilon) + \\ +(\lambda_1 + \sigma_1 e^{-ju_1})F(0,u_1,w,\varepsilon) - \sigma_1 e^{-ju_1} f_1(w,\varepsilon) = O(\varepsilon^2), \\ (\lambda_1(e^{ju_1}-1) + S_2(1-\varepsilon)(e^{jw\varepsilon}-1) - \mu_2)F(2,u_1,w,\varepsilon) + \\ +(S_2(1-\varepsilon) + \sigma_2 e^{-jw\varepsilon})F(0,u_1,w,\varepsilon) - \sigma_2 e^{-jw\varepsilon}\varepsilon f_2(u_1) = O(\varepsilon^2), \\ (\sigma_1(e^{-ju_1}-1) + \sigma_2(e^{-jw\varepsilon}-1))F(0,u_1,w,\varepsilon) + \\ +(\lambda_1(e^{ju_1}-1) + S_2(1-\varepsilon)(e^{jw\varepsilon}-1))(F(1,u_1,w,\varepsilon) + F(2,u_1,w,\varepsilon)) + \\ +\sigma_1(1-e^{-ju_1})f_1(w,\varepsilon) + \sigma_2(1-e^{-jw\varepsilon})\varepsilon f_2(u_1) = O(\varepsilon^2). \end{array}\right. \tag{12}$$

Taking into account $\lim\limits_{\varepsilon\to 0} f_1(w,\varepsilon) = r_1$, Equations (12) are written under $\varepsilon \to 0$ as

$$\left\{\begin{array}{l} -(\lambda_1 + S_2 + \sigma_1 + \sigma_2)F(0,u_1,w) + \mu_1 F(1,u_1,w) + \mu_2 F(2,u_1,w) + \sigma_1 r_1 = 0, \\ (\lambda_1(e^{ju_1}-1) - \mu_1)F(1,u_1,w) + (\lambda_1 + \sigma_1 e^{-ju_1})F(0,u_1,w) - \sigma_1 e^{-ju_1} r_1 = 0, \\ (\lambda_1(e^{ju_1}-1) - \mu_2)F(2,u_1,w) + (S_2 + \sigma_2)F(0,u_1,w) = 0, \\ \sigma_1(e^{-ju_1}-1)F(0,u_1,w) + \lambda_1(e^{ju_1}-1)(F(1,u_1,w) + F(2,u_1,w)) + \\ +\sigma_1(1-e^{-ju_1})r_1 = 0. \end{array}\right. \tag{13}$$

To obtain the marginal probability distribution, we will set $w = 0$ in System (13) and denote $H^{(1)}(k,u) = F(k,u_1,0)$.

$$\left\{\begin{array}{l} -(\lambda_1 + S_2 + \sigma_1 + \sigma_2)H^{(1)}(0,u) + \mu_1 H^{(1)}(1,u) + \mu_2 H^{(1)}(2,u) + \sigma_1 r_1 = 0, \\ (\lambda_1(e^{ju}-1) - \mu_1)H^{(1)}(1,u) + (\lambda_1 + \sigma_1 e^{-ju})H^{(1)}(0,u) - \sigma_1 e^{-ju} r_1 = 0, \\ (\lambda_1(e^{ju}-1) - \mu_2)H^{(1)}(2,u) + (S_2 + \sigma_2)H^{(1)}(0,u) = 0, \\ \sigma_1 H^{(1)}(0,u) + \lambda_1 e^{ju}(H^{(1)}(1,u) + H^{(1)}(2,u)) + \sigma_1 r_1 = 0. \end{array}\right. \tag{14}$$

The solution of System (14) is following

$$H^{(1)}(0,u) = \frac{\sigma_1 r_1(\lambda_1(1-e^{ju})+\mu_2)(\mu_1-\lambda_1 e^{ju})}{(\sigma_1\mu_1-\lambda_1 e^{ju}(\sigma_1+\lambda_1)(\mu_2+\lambda_1(1-e^{ju}))-\lambda_1 e^{ju}(S_2+\sigma_2)(\mu_1+\lambda_1(1-e^{ju}))},$$
$$H^{(1)}(1,u) = \frac{(\lambda_1+\sigma_1 e^{-ju})H^{(1)}(0,u)-\sigma_1 r_1 e^{-ju}}{\lambda_1(1-e^{ju})+\mu_1}, \quad H^{(1)}(2,u) = \frac{(S_2+\sigma_2)}{\lambda_1(1-e^{ju})+\mu_2}H^{(1)}(0,u). \tag{15}$$

In this way, we can obtain the partial characteristic function of the number of customers in the first orbit under a heavy load of the second class customers

$$H^{(1)}(u) = H^{(1)}(0,u) + H^{(1)}(1,u) + H^{(1)}(2,u),$$

then the probability distribution of the number of customers in the first orbit can be approximate as

$$P_1(i_1) \approx \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-jui_1}H^{(1)}(u)du. \tag{16}$$

Note that we can derive the marginal asymptotic probability distribution of the number of customers in the second orbit in the same way.

4. Results

Using Expressions (15), we can be obtained main characteristics of the model under study such as marginal probability distribution of each class customers (16), probabilities of server states (6) or means and high order moments as

$$\mathrm{E}[(i_1(t))^n] = j^{-n}\frac{d^n H^{(1)}(u)}{du}\bigg|_{u=0}.$$

For a demonstration of the results of the asymptotic method, we consider a numerical example. Consider the system parameters have the following values:

$$\lambda_1 = 0.1, \quad \sigma_1 = 5, \quad \sigma_2 = 10, \quad \mu_1 = 1, \quad \mu_2 = 2$$

and the second class of customers makes a heavy load to the system. We introduce a load parameter ρ such as $\lambda_2 = \rho S_2$; under asymptotic condition $\rho \to 1$.

First, we analyze S_1 and S_2 dependence on the parameters of the second class (λ_2 and σ_2) (Figure 5).

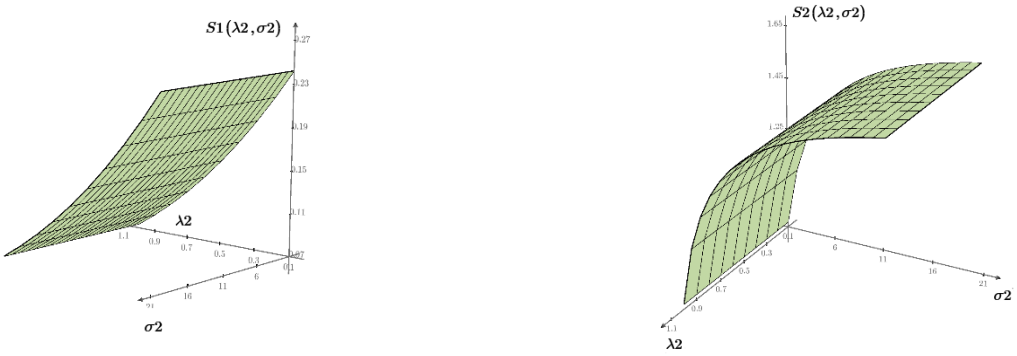


Figure 5: *Values of S_1 and S_2 vs. λ_2 and σ_2*

We can conclude that the first class throughput parameter S_1 decreases with arrival rate λ_2 increases and does not depend on retrial rate σ_2. In addition, the second class throughput parameter S_2 increases with retrial rate increasing.

In Figure 6, the probability distribution of the number of customers of the first class is presented for $\rho = 0.9$. In Table 1, the values of the mean of the number of customers in the first

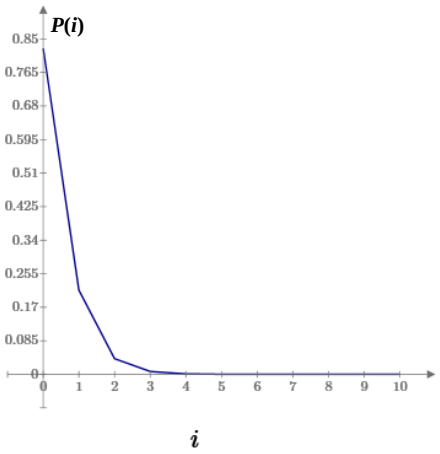


Figure 6: *Probability distribution of the number of customers in the first orbit for* $\lambda_2 = 0.9S_2$

orbit are given for different values of ρ.

Table 1: *Mean of the number of customers in the first orbit*

ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
$\mathrm{E}\{i_1(t)\}$	1.509	1.365	1.221	1.077	0.933	0.789	0.645	0.501	0.357	0.213

In Figure 7, the stationary probabilities of the server states depending on ρ are presented. As

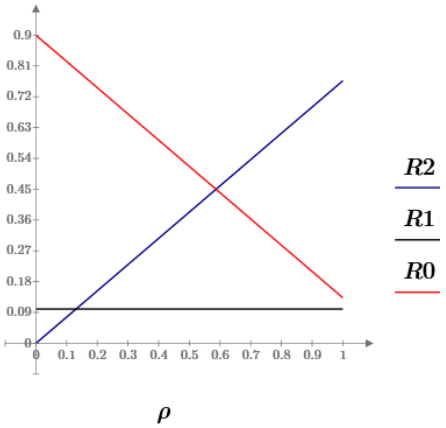


Figure 7: *Probabilities of the server states* R_k *vs.* ρ

we can see, probability R_1 does not depend on parameter ρ (as on $\lambda_2 = \rho S_2$), it is also proved by formulas (6). Probabilities R_0 and R_2 have linear dependence on ρ.

In Figure 8, the stationary probabilities of empty orbits r_n are presented. We can conclude that the probabilities that corresponding orbit and the server are empty decrease ($r_2 \to 0$) with $\rho \to 1$. That can be explained by the fact that the stability boundary is approaching, so the probability of an empty orbit becomes smaller and smaller.

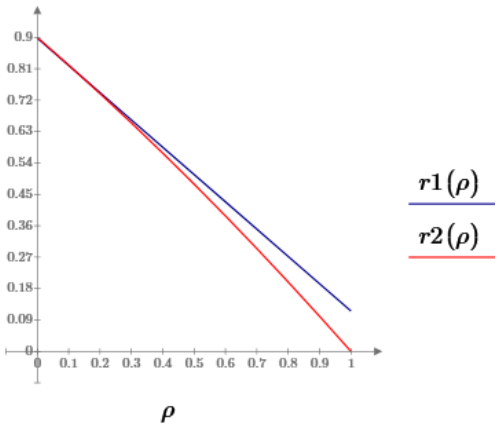


Figure 8: *Probabilities of empty orbits r_n vs. ρ*

5. Conclusions

In this way, we have proposed the partial asymptotic analysis method for the two-class retrial queuing systems with constant retrial rate. During the research, we have derived the asymptotic characteristic function of the number of customers in one orbit under a heavy load of the system by other class customers, from which all marginal characteristics of the model can be found. Obviously, asymptotic formulas will give us more precise results if an arrival process rate tends to its throughput: $\lambda_n \to S_n$. Numerical examples are also presented.

The asymptotic method allows us to derive analytical formulas in case of impossibility of explicit formulas deriving. In this way, in future studies, the proposed methods will be applied for more complex models such as multiclass retrial queue, systems with MMPP arrivals, priority customers, etc. Also, we plan to develop the approach for other asymptotic conditions, for example, an overload of one orbit.

References

[1] Artem Ryndin et al. "Modelling of multi-path transmission system of various priority multimodal information". In: *2020 IEEE 14th International Conference on Application of Information and Communication Technologies (AICT)*. IEEE, Oct. 2020, pp. 1–5. DOI: 10.1109/aict50176.2020.9368802.

[2] Mohammad Al Jaafreh et al. "Multimodal Systems, Experiences, and Communications: A Review Toward the Tactile Internet Vision". In: *Recent Trends in Computer Applications*. Springer International Publishing, 2018, pp. 191–220. ISBN: 9783319899145. DOI: 10.1007/978-3-319-89914-5_12.

[3] Valeriy Naumov et al. *Matrix and Analytical Methods for Performance Analysis of Telecommunication Systems*. Springer International Publishing, 2021. ISBN: 9783030831325. DOI: 10.1007/978-3-030-83132-5.

[4] Jesús R. Artalejo and Antonio Gómez-Corral. *Retrial Queueing Systems*. Springer Berlin Heidelberg, 2008. ISBN: 9783540787259. DOI: 10.1007/978-3-540-78725-9.

[5] J. Falin and James G. C. Templeton. *Retrial Queues*. en. London: Chapman and Hall/CRC, 1997.

[6] Tuan Phung-Duc. *Retrial Queueing Models: A Survey on Theory and Applications*. 2019. DOI: 10.48550/ARXIV.1906.09560.

[7] Elena Makeeva, Irina Kochetkova, and Reem Alkanhel. "Retrial Queueing System for Analyzing Impact of Priority Ultra-Reliable Low-Latency Communication Transmission on Enhanced Mobile Broadband Quality of Service Degradation in 5G Networks". In: *Mathematics* 11.18 (Sept. 2023), p. 3925. ISSN: 2227-7390. DOI: 10.3390/math11183925.

[8] Evsey Morozov et al. "Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking". In: *Performance Evaluation* 134 (Oct. 2019), p. 102005. ISSN: 0166-5316. DOI: 10.1016/j.peva.2019.102005.

[9] Konstantin Avrachenkov, Evsey Morozov, and Ruslana Nekrasova. "Stability analysis of two-class retrial systems with constant retrial rates and general service times". In: *Performance Evaluation* 159 (Jan. 2023), p. 102330. ISSN: 0166-5316. DOI: 10.1016/j.peva.2022.102330.

[10] A. Krishnamoorthy, V. C. Joshua, and Ambily P. Mathew. "A Retrial Queueing System with Multiple Hierarchial Orbits and Orbital Search". In: *Developments in Language Theory*. Springer International Publishing, 2018, pp. 224–233. ISBN: 9783319986548. DOI: 10.1007/978-3-319-99447-5_19.

[11] Bara Kim and Jeongsim Kim. "Proof of the conjecture on the stability of a multi-class retrial queue with constant retrial rates". In: *Queueing Systems* 104.3–4 (June 2023), pp. 175–185. ISSN: 1572-9443. DOI: 10.1007/s11134-023-09881-z.

[12] Bara Kim and Jeongsim Kim. "Stability of a multi-class multi-server retrial queueing system with service times depending on classes and servers". In: *Queueing Systems* 94.1–2 (Oct. 2019), pp. 129–146. ISSN: 1572-9443. DOI: 10.1007/s11134-019-09634-x.

[13] Konstantin Avrachenkov. "Stability and partial instability of multi-class retrial queues". In: *Queueing Systems* 100.3–4 (Apr. 2022), pp. 177–179. ISSN: 1572-9443. DOI: 10.1007/s11134-022-09814-2.

[14] Yang Woo Shin and Dug Hee Moon. "M/M/c Retrial Queue with Multiclass of Customers". In: *Methodology and Computing in Applied Probability* 16.4 (Apr. 2013), pp. 931–949. ISSN: 1573-7713. DOI: 10.1007/s11009-013-9340-0.

[15] A Nazarov, S Paul, and O Lizyura. "Asymptotic analysis of Markovian retrial queue with unreliable server and multiple types of outgoing calls". en. In: *Global and Stochastic Analysis* 8.3 (2021), pp. 143–149.

[16] Elena Danilyuk, Daria Kuznetsova, and Svetlana Moiseeva. "Asymptotic Analysis of Retrial Queueing System MMPP/M/1 with Impatient Customers, Collisions and Unreliable Server". In: *2023 5th International Conference on Problems of Cybernetics and Informatics (PCI)*. IEEE, Aug. 2023, pp. 1–4. DOI: 10.1109/pci60110.2023.10325990.

PERFORMANCE AND NUMERICAL ANALYSIS OF $(GI|GI|N,M)$ QUEUES USING MARKED MARKOV PROCESS

Vladimir Rykov[1,2], Nika Ivanova[3], Evsey Morozov[4,5]

•

[1]Peoples' Friendship University of Russia (RUDN University), Moscow, Russia
[2]Gubkin Russian State Oil and Gas University, Moscow, Russia
vladimir_rykov@mail.ru
[3] V.A.Trapeznikov Institute of Control Sciences of RAS, Moscow, Russia
nm_ivanova@bk.ru
[4]Institute of Applied Mathematical Research,
Karelian Research Centre of RAS, Petrozavodsk, Russia
[5]Petrozavodsk State University, Petrozavodsk, Russia
emorozov@karelia.ru

Abstract

We study the key performance characteristics of a finite-buffer multi-server queuing system denoted as $(GI|GI|n,m)$, with general inter-arrival and service times distributions. The concept called Marked Markov Processes is employed to analyze such a system. Its mathematical model is constructed, and marks' transformations are introduced, which are further applied to calculate the performance characteristics of the system using a special simulation algorithm. Numerical study validates the proposed method employing the comparison of the obtained results with well-known results for $(M|M|1)$, $(M|GI|1)$, and $(M|M|n,m)$ models.

Keywords: $(GI|GI|n,m)$ queuing system, Marked Markov Process, general inter-arrival and service times distributions, steady-state probabilities, stationary performance metrics, numerical analysis.

1. Introduction

Many real stochastic phenomena can be modeled as processes with discrete random intervention (PDRI), first proposed by Kolmogorov. This class of processes is discussed in [1] and some other works. Jump-wise Markov, semi-Markov, Generalized Semi-Markov Processes (GSMP) are notable examples. In the paper, we consider an alternative construction of such processes and demonstrate its application to a widely studied queuing system (QS) with renewal input, buffer capacity m, and n servers with general independent identically distributed (iid) service times. Using Kendall's notation [2], slightly adapted for modern needs in [3], we denote such a system as $(GI|GI|n,\ m)$.

The initial application of Markov processes in stochastic system analysis focused on processes with discrete (finite or countable) state spaces before being extended in various directions, including, in particular, the method of supplementary variables [4], linear Markov processes [5], and piecewise linear Markov processes [6]. Furthermore, since the publication of Cinlar's paper [7], semi-Markov processes have gained widespread acceptance in the field.

Since Smith's classic paper [8], the regenerative approach has played a pivotal role in studying stochastic systems. Also, it has been established that a wide class of stochastic processes possess a regenerative structure (see, for instance, [9, 10, 11]). This approach proves particularly powerful for analyzing stochastic processes describing complex queuing models.

A critical challenge in studying stochastic systems is that initial information is rarely known with high precision. Thus, analyzing the sensitivity of model output characteristics to the shape and parameters of input distributions is essential. In this context, we highlight:

(i) classic study by Sevastyanov [12] on the insensitivity of the Erlang loss system steady state distribution to the shape of service time distribution, as well as

(ii) the works by Gnedenko [13, 14] and Soloviev [15] on the convergence of the reliability function of the double cold-standby system to the exponential one, regardless of the life/repair time distributions of elements. These findings can be interpreted as the asymptotic insensitivity of the system reliability function to component characteristics. Later on these studies have been extended in a series of our works [16, 17, 18, 19] to a broader class of the reliability systems.

In recent publications, a concept of Discrete Event System (DES) modeling [20, 21, 22, 23] and its mathematical formulation as GSMP [24, 25] has been proposed. These studies demonstrate that GSMPs effectively describe models in queuing theory, reliability theory, and stochastic networks. Moreover, it was shown that GSMPs exhibit a regenerative structure, enabling proofs of the Law of Large Numbers (LLN) and Central Limit Theorem (CLT) in [26, 27]. This regenerative framework offers significant potential for the DES modeling and simulation [20, 21, 22, 23, 24]. While the classical results have been applied mainly to the finite-state processes, the recent research identify the regenerative structures in models with more general state spaces.

Some articles [28, 29, 30] present research results, including numerical, of specific models of stochastic systems based on DES using GSMP. In this paper we propose another mathematical model of PDRI.

In contrast, inspired by the Marked Point Processes, in [31] the concept of the Marked Markov Process (MMP) has been proposed to analyze a hot-standby double system. The subsequent works [32, 33] use this approach to analyze a repairable k-out-of-n systems with general life and repair times distributions of the components. This approach enables both computation of the key reliability metrics and also a sensitivity analysis with respect to the input data.

This paper employs MMPs to study the $(GI|GI|n,\ m)$ system with focus on the development of the methods and procedures to calculate the key metrics of the system. We mention a few monographs [10, 11, 34, 35, 36] where different queueing systems are considered in various aspects at a fairly advanced mathematical level. In this regard we note that the monograph [22] which also focuses on the algorithms to calculate the performance metrics of the stochastic systems is being the most close to the current paper.

The paper is organized as follows. A general description of the MMP is given in Section 2. In Section 3 the description of the system $(GI|GI|n,\ m)$, the key notations are given. Moreover, the basic assumptions and the problem setting are presented in this section. A detailed description of the transformations of marks and the calculation of the corresponding transition probabilities are given in Section 4. In Section 5, the analytical expressions for the marks distributions are deduced. The stability conditions are given in Section 6. By the complexity of the computation procedures, Section 7 presents the main formulae expressing the key metrics in terms of the marks. Section 8 contains various numerical examples, which include validation of the proposed methodology. Appendices A and B contain auxiliary information and a general simulation-based algorithm to calculate the performance metrics.

2. Marked Markov Process

Most of all mentioned in the Introduction processes are PDRI. A fairly general mathematical model of PDRI is the Marked Markov Process. By MMP we understand a discrete-time sequence

$$Z := Z(t) = \{(J(t), \mathbf{V}(t)),\ t = 0, 1, \dots\},$$

at the times $S(t)$ of random intervention which are calculated with the help of marks and allow investigating of all needed characteristics of the system in continuous time. Here

- $J(t)$ is the main component, which describes the system states with a set of states $\mathcal{J}$ with $|\mathcal{J}| \leq \infty$,
- $\mathbf{V}(t)$ is a set of random marks $\mathbf{V}(t) = \{\mathbf{V}_j(t) : j \in \mathcal{J}\}$, which makes the process Z Markov, and, for each j, the mark $\mathbf{V}_j$ takes values in a measurable space $(E_j, \mathcal{E}_j)$.

Such a process is determined by:

- The transition probabilities $p_{ij}(\mathbf{V}_i) = \mathbf{P}\{J(t+1) = j | J(t) = i, \mathbf{V}_i(t)\}$ of the component $J(t)$, which depend on the mark $\mathbf{V}_i(t)$ in state i (at step t);
- The marks transformation operators $\Phi_{ij}(\mathbf{V}_i)$ for the transition from state i to state j, based on the content of the mark $\mathbf{V}_i$ in state i, and
- The sequence of the iid random variables (rv's) ξ_t, $t \in \mathbb{N}_0 := [0, 1, \ldots]$ describing the input data and discrete random interventions. These rv's determine the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ on which all other components of the process are also determined. A detailed description of these data depends on a concrete model and, for the considered case, will be given below.

A complete description of such a process includes also
(i) the initial distribution $\alpha = (\alpha_j : \ j \in \mathcal{J})$ of the main component, where $\alpha_j = \mathbf{P}\{J(0) = j\}$,
(ii) the conditional initial distribution $\mu_j(0, \cdot) = \mathbf{P}\{\mathbf{V}(0) \in \cdot \, | J(0) = j\}$ of the marks $\mathbf{V}(t)$ and
(iii) the distribution $F(\cdot)$ of the rv's ξ_t.

This research does not cover a full theory of the MMP but only illustrates some possibilities to calculate the performance of a general queueing model.

3. Model description

3.1. Assumptions and notations

Consider an n-server QS, denoted by $(GI|GI|n, m)$, with buffer capacity m, the renewal input process, and the iid service times. Denote by

– A_i : $(i = 1, 2, \ldots)$ iid inter-arrival times of customers with absolutely continuous cumulative distribution function (cdf) $A(t) = \mathbf{P}\{A_i \leq t\}$, probability density function (pdf) $a(t) = A'(t)$, finite mean $\mu_A = \mathbf{E}[A_i] < \infty$ and finite coefficient of variation (CoV) defined as the ratio of the standard deviation σ_A to the mean μ_A, $v_A = \sigma_A/\mu_A < \infty$;

– B_i $(i = 1, 2, \ldots)$ the iid service times with absolutely continuous cdf $B(t) = \mathbf{P}\{B_i \leq t\}$, pdf $b(t) = B'(t)$, finite mean $\mu_B = \mathbf{E}[B_i] < \infty$ and finite CoV $v_B = \sigma_B/\mu_B < \infty$.

In what follows, the rv's are denoted by capital Latin letters, $X, Y, \ldots$, and their values are denoted by the corresponding lowercase Latin letters, $x, y, \ldots$. Multidimensional rv's and their values are denoted by the bold letters. Furthermore, for an iid sequence $\{X_i\}$, the rv X denotes the generic element.

3.2. The MMP-based System Modeling

The dynamic of the system $(GI|GI|n, m)$ is described by the following MMP

$$Z := \{Z(t) = (J(t), \mathbf{V}(t)),\ t \in \mathbb{N}_0\},$$

where t is the number (step) of intervention (change of the state of system). The (main) component $J(t) \in \mathcal{J} = \{0, 1, 2, \ldots, n+m\}$ denotes the number of customers in the system (at step t) and

$$\mathbf{V} := \{\mathbf{V}(t) = \{\mathbf{V}_j(t), j \in \mathcal{J}\},\ t \in \mathbb{N}_0\},\ \mathbf{V}_j(t) = (X_j(t),\ \mathbf{Y}_j(t)) \tag{1}$$

is the sequence of marks where, in state j in step t, $X_j(t)$ denotes the residual arrival time and $\mathbf{Y}_j(t) = (Y_j^{(1)}(t), \ldots, Y_j^{(j\wedge n)}(t))$ are the residual service times (arranged in an ascending order) and $j \wedge n = \min(j, n)$. (By assumption, $Y_0^{(1)}(t) = \infty$ if $J(t) = 0$.) The changes of the process Z in continuous time occur at the instants

$$S(t) = S(t-1) + T_{J(t)}(t), \tag{2}$$

where the (continuous) time interval $T_{J(t)}(t)$ between the switching of states is

$$T_{J(t)}(t) = \min\left[X_{J(t)}(t),\, Y_{J(t)}^{(1)}(t)\right],\ t \in \mathbb{N}_0,\ S(0) = S(-1) := 0. \tag{3}$$

We note that the dimension of the vector $\mathbf{Y}_j(t)$ depends on $J(t) = j$ but is fixed for fixed j.

The 1st customer arrives in the empty system and starts a new service, $J(0) = 1,\ X_1(0) = A_0,\ Y_1(0) = B_0$. The transition probabilities of $J(t)$ do not depend on t (time-homogeneous process) but depend on the associated marks as follows:

$$\begin{aligned} p_j(\mathbf{V}_j) &= \mathbf{P}\{J(t+1) = j+1 \,|\, J(t) = j\} = \mathbf{P}\{X_j(t) \le Y_j^{(1)}(t)\}, \\ q_j(\mathbf{V}_j) &= \mathbf{P}\{J(t+1) = j-1 \,|\, J(t) = j\} = \mathbf{P}\{X_j(t) > Y_j^{(1)}(t)\},\ j \in \mathcal{J}. \end{aligned}$$

These transitions are illustrated in Figure 1 and resemble transitions of a typical birth and death process. Note that relation (2) allows to calculate the required metrics in continuous time. To this

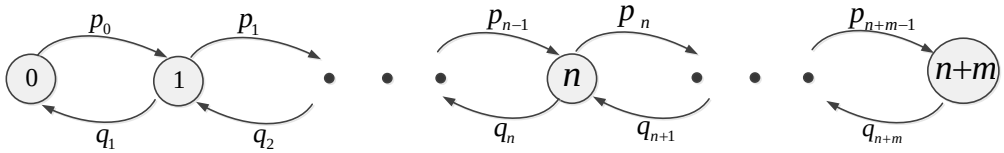


Figure 1: *Transition graph of the process $J(t)$, where $p_j = p_j(\mathbf{V}_j)$ and $q_j = q_j(\mathbf{V}_j)$*

end, we introduce the process

$$N(s) = \min\{t:\ S(t) \le s\},\ s \ge 0,$$

counting (in continuous time) the intervention times in interval $[0, s]$. Then, the number of customers in the system, $L(s)$, in (continuous) time s is defined as

$$L(s) = J(N(s)),\ s \ge 0. \tag{4}$$

If $J(t-1) = 0,\ J(t) = 1$, then t is a regeneration step, and the discrete-time regenerations $\{\tau_i\}$ can be defined by the following recursion,

$$\tau_i = \min\{t:\ t > \tau_{i-1}:\ J(t-1) = 0,\ J(t) = 1\},\ i \ge 1,\ \tau_0 := 0.$$

It is now seen that the instant $S(\tau_i)$ (see (2)) is the beginning of the ith regeneration cycle, that is the ith regeneration point in continuous time. Hence,

$$R_i = S(\tau_i) - S(\tau_{i-1}) = \sum_{\tau_{i-1} < t \le \tau_i} T_{J(t)}(t), \tag{5}$$

is the ith (continuous-time) regeneration cycle length, $i \ge 1$. Figure 2 shows the (right-continuous) trajectory of the process $L(s)$, with the 1st regeneration period $R_1 = S(\tau_1)$. Note that the duration of the ith busy period Π_i is

$$\Pi_i = \sum_{\tau_{i-1} < t \le \tau_i - 1} T_{J(t)}(t) = R_i - T_{J(\tau_i - 1)}(\tau_i - 1),\ i \ge 1. \tag{6}$$

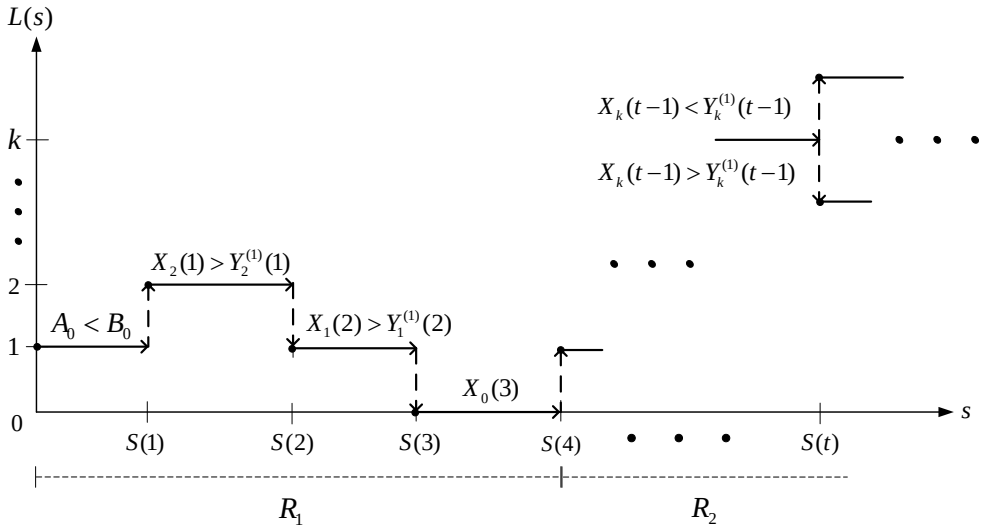


Figure 2: *Trajectory of continuous time process $L(s)$*

Define the number of regenerations in $[0,s]$,

$$K(s) := \min\{k : S(\tau_k) \le s\},\ s \ge 0,$$

and assume that the mean regeneration period length $\mu_R := \mathbf{E}[R] < \infty$. Then, by (4), the stationary distribution $\pi_j = \lim_{s\to\infty} \mathbf{P}\{L(s) = j\}$ is calculated, in terms of marks, as follows:

$$\begin{aligned}\pi_j &= \lim_{s\to\infty}\frac{1}{s}\int_0^s 1_{\{L(u)=j\}}du = \lim_{s\to\infty}\frac{1}{s}\left[\sum_{k=1}^{K(s)}\int_{\tau_{k-1}}^{\tau_k} 1_{\{J(N(u))=j\}}du + \int_{\tau_{K(s)}}^{s} 1_{\{J(N(u))=j\}}du\right] = \\ &= \lim_{s\to\infty}\frac{K(s)}{s}\frac{1}{K(s)}\sum_{k=1}^{K(s)}\int_{\tau_{k-1}}^{\tau_k} 1_{\{J(N(u))=j\}}du = \frac{1}{\mu_R}\mathbf{E}\left[\sum_{t=1}^{\tau_1} T_j(t)1_{\{J(t)=j\}}\right],\ j\in\mathcal{J}. \qquad (7)\end{aligned}$$

It is worth mentioning that the regenerative simulation method based on the regenerative version of the CLT allows constructing correct interval estimates, not only the point ones of the performance metrics [23]. However in the numerical analysis in this research we demonstrate the proposed algorithm for the sample-mean estimation while the interval estimation is postponed for a future research.

3.3. The problem setting

The main purpose of the paper is to study the basic stationary performance metrics of the system $(GI|GI|n,m)$ using the MMP framework. Relations (5)-(7) as well as the regenerative structure of the process Z allow to use one (say the 1st) regeneration cycle to calculate the stationary metrics including:

- cdf $F_\Pi(s) = \mathbf{P}\{\Pi \le s\}$ and the rth moments $m_r(\Pi) = \mathbf{E}[\Pi^r]$, $r > 0$;
- cdf $F_R(s) = \mathbf{P}\{R \le s\}$ and the rth moments $m_r(R) = \mathbf{E}[R^r]$, $r > 0$;
- stationary distribution π_j, $j \in \mathcal{J}$;
- the loss probability $\pi_{loss} = \pi_{n+m}$ (if $m < \infty$);
- the rth moments of the stationary number of customers $m_r(L) = \sum_{0\le j\le n+m} j^r\pi_j$;

- the rth moments of the queue length $m_r(Q) = \sum_{n\le j\le n+m}(j-n)^r\pi_j$;
- the moments of the stationary waiting time W and sojourn time V.

4. Marks Transformations and Distribution

4.1. Transformations of Marks

Consider the sequence of the marks $\{\mathbf{V}_n\}$ and a rv V. Denote by $Sh[\mathbf{V}_n, V]$ the operator subtracting the rv V from the variation series $\mathbf{V}_n$ provided $V \le V^{(1)}$. (If $V = V^{(1)}$, then the vector $\mathbf{V}$ is shifted by one component to the left.) Also denote by $Ad[\mathbf{V}_n, V]$ the operator adding an independent rv V to the variation series $\mathbf{V}_n$. More exactly, these operators are defined as

$$Sh[\mathbf{V}_n, V] = V^{(i)} - V, \text{ for } i = \overline{1,n}; \tag{8}$$

$$Ad[\mathbf{V}_n, V] = \begin{cases} V^{(i)} & \text{for } i < l, \\ V & \text{for } i = l, \\ V^{(i+1)} & \text{for } i > l, \end{cases} \tag{9}$$

where $l = \max\{i:\ V^{(i)} \le V\}$. Introduce the operators that transform the process J from state j to $j+1$ and $j-1$, respectively,

$$\mathbf{V}_{j+1} = \Phi_{j,j+1}[\mathbf{V}_j] =: \Phi_j[\mathbf{V}_j], \quad \mathbf{V}_{j-1} = \Phi_{j,j-1}[\mathbf{V}_j] =: \Psi_j[\mathbf{V}_j].$$

Now we prove the following statement.

Statement 1. Under transition $j \to j+1$, i.e. if $X_j(t) \le Y_j^{(1)}(t)$, the marks are transformed as follows:

$$X_{j+1}(t+1) = \Phi_j[X_j(t)] = A_{t+1}. \tag{10}$$

For $j < n$, $i = \overline{1, n-j}$:

$$Y_{j+1}^{(i)}(t+1) = \Phi_j[\mathbf{Y}_j(t)] = Ad[Sh[\mathbf{Y}_j(t), X_j(t)],\, B_{t+1}] =$$
$$= \begin{cases} Y_j^{(i)}(t) - X_j(t) & \text{for } i < l, \\ B_{t+1} & \text{for } i = l, \\ Y_j^{(i+1)}(t) - X_j(t) & \text{for } i > l, \end{cases} \tag{11}$$

where $l = \max\{i:\ Y_j^{(i)}(t) - X_j(t) \le B_{t+1}\}$;
for $j \ge n$

$$Y_{j+1}^{(i)}(t+1) = \Phi_j[\mathbf{Y}_j(t)] = Sh[\mathbf{Y}_j(t), X_j(t)] = Y_j^{(i+1)}(t) - X_j(t) \ \ (i = \overline{1, n-j}),\ j \in \mathcal{J}. \tag{12}$$

Analogously, under transition $j \to j-1$, i.e. if $X_j(t) > Y_j^{(1)}(t)$, then mark X_{j-1} is transformed as

$$X_{j-1}(t+1) = \Psi_j[X_j(t)] = Sh[X_j(t), Y_j^{(1)}(t)] = X_j(t) - Y_j^{(1)}(t),\ j \in \mathcal{J}. \tag{13}$$

However, the transformation of the mark $\mathbf{Y}_j$ depends on state j and has the following form:
for $j \le n$

$$Y_{j-1}^{(i)}(t+1) = \Psi_j[\mathbf{Y}_j(t)] = Sh[\mathbf{Y}_j(t),\, Y_j^{(1)}(t)] = Y_j^{(i+1)}(t) - Y_j^{(1)}(t) \ \ (i = \overline{1, j}); \tag{14}$$

and for $j > n$,

$$\begin{aligned} Y_{j-1}^{(i)}(t+1) &= \Psi_j[\mathbf{Y}_j(t)] = Ad[Sh[\mathbf{Y}_j(t),\, Y_j^{(1)}(t)],\, B_{t+1}] \\ &= \begin{cases} Y_j^{(i)}(t) - Y_j^{(1)}(t) & \text{for } i < l, \\ B_{t+1} & \text{for } i = l, \\ Y_j^{(i+1)}(t) - Y_j^{(1)}(t) & \text{for } i > l, \end{cases} \end{aligned} \tag{15}$$

where $l = \max\{i:\, Y_j^{(i)}(t) - Y_j^{(1)}(t) \le B_{t+1}\};\ j \in \mathcal{J}$.

Comments. Indeed, the transition $j \to j+1$ at instant t occurs when a new customer arrives, then the mark is $X_{j+1}(t+1) = A_{t+1}$, implying (10). To find the change of the mark $\mathbf{Y}_j$, we note that the residual service times $Y_j^{(i)}(t)$ are decreased by $X_j(t)$. Further, if there are available servers ($j < n$), a new customer (with the service time B_{t+1}) takes the ith position in the mark $\mathbf{Y}_{j+1}(t)$. At that, the marks $Y_{j+1}^{(i)}(t)$ preceding the $(i-1)$th position remain unchanged and the marks in the subsequent positions are shifted by one position to the right, it implies (11). If $j \ge n$, then the new customer is queued, if $j < n+m$, and lost, otherwise. At that, the value $X_j(t)$ is subtracted from all components $Y_{j+1}^{(i)}(t)$ in the mark $\mathbf{Y}_j(t)$, and formula (12) follows.

Similarly, the transition $j \to j-1$ is only possible when a service is finished. Then the residual time $X_j(t)$ decreases by $Y_j^{(1)}(t)$, implying (13). Finally, the transformation of $\mathbf{Y}_j(t)$ caused by the transition $j \to j-1$ depends on the state of the system. Namely, for $j \le n$, the components of $\mathbf{Y}_j(t)$ are shifted by $Y_j^{(1)}(t)$, and the value $Y_j^{(1)}(t)$ is removed from the (vector) mark $\mathbf{Y}_j(t)$, proving (14). Otherwise, the 1st awaiting customer joins service, and then $\mathbf{Y}_j(t)$ is transformed according to (15).

It is easy to see that, because the rv's A_t and B_t are independent of the pre-history of the process Z, then the process Z turns out to be Markovian.

Remark 1. When a new customer arrives in the idle system in step t, then $J(t) = 1$ and the new marks take (independent) values $X_1(t) = A$, $Y_1^{(1)}(t) = B$. In other words, step t and the instant $S(t)$ are the regeneration step and the regeneration instant, respectively.

4.2. Transition probabilities of marks transformation

Based on transformations of the marks $\mathbf{V}_j = (X_j, \mathbf{Y}_j)$, we extend the action of operators Sh and Ad to the (non-random) values $\mathbf{v}_j = (x_j,\, \mathbf{y}_j)$. Namely, we put

$$Sh[\mathbf{v}_n,\, v] = v^{(i)} - v, \ \text{when } i = \overline{1,n}; \tag{16}$$

$$Ad[\mathbf{v}_n,\, v] = \begin{cases} v^{(i)} & \text{for } i < l, \\ v & \text{for } i = l, \\ v^{(i+1)} & \text{for } i > l, \end{cases} \tag{17}$$

where $l = \max\{i:\, v^{(i)} < v\}$. Denote transition probabilities by

$$\begin{aligned} P_j(\mathbf{v}_j;\, C_{j+1}) &= \mathbf{P}\{X_j \le Y_j^{(1)},\ \mathbf{V}_{j+1} \in C_{j+1} | \mathbf{V}_j = \mathbf{v}_j\}; \\ Q_j(\mathbf{v}_j;\, C_{j-1}) &= \mathbf{P}\{X_j > Y_j^{(1)},\ \mathbf{V}_{j-1} \in C_{j-1} | \mathbf{V}_j = \mathbf{v}_j\}, \end{aligned}$$

where C is a subset of the state space of marks (to be specified below). Under transitions of the main component, the values $\mathbf{v}_j = (x_j, \mathbf{y}_j)$ are transformed by Theorem 1, and we need the following notation

$$\Delta_j^i Sh[\mathbf{y}_j,\, x_j] = (y_j^{i-1} - x_j,\, y_j^i - x_j],\ \Delta_j^i B(Sh[\mathbf{y}_j,\, x_j]) = B(y_j^i - x_j) - B(y_j^{i-1} - x_j),\ i = \overline{2, n-j},\ j \in \mathcal{J}$$

$$\Delta_j^i Sh[\mathbf{y}_j,\, y_j^1] = (y_j^{i-1} - y_j^1,\, y_j^i - y_j^1],\ \Delta_j^i B(Sh[\mathbf{y}_j,\, y_j^1]) = B(y_j^i - y_j^1) - B(y_j^{i-1} - y_j^1),\ i = \overline{2,j},\ j \in \mathcal{J}.$$

Now we define the set $C_{j+1}^{(i)}(\mathbf{v}_j, x)$ corresponding new arrival implying transition $j \to j+1$, (provided $j < n$). This set consists of the values $\mathbf{v}_{j+1}$ (of mark $\mathbf{V}_{j+1}$) for which the component $x_{j+1} \in [0, x]$, while the value $\mathbf{y}_j$ (of mark $\mathbf{Y}_j$), after being shifted by x_j, is complemented by a new component at position i. More exactly,

$$C_{j+1}^{(i)}(\mathbf{v}_j,\, x) = \{\mathbf{v}_{j+1} :\ Ad\big[Sh[\mathbf{v}_j,\, x_j],\ A_{t+1} \in [0, x],\ B_{t+1} \in \Delta_j^i(Sh[\mathbf{y}_j,\, x_j])\big]\}.$$

Similarly, for $j \geq n$, the set $C_{j+1}(\mathbf{v}_j, x)$ contains the values $\mathbf{v}_{j+1}$ with the component $x_{j+1} \in [0, x]$. The new customer is either queued or lost, resulting in shift $\mathbf{Y}_j$ by x_j. Namely,

$$C_{j+1}(\mathbf{v}_j, x) = \{\mathbf{v}_{j+1} :\ A_{t+1} \in [0, x],\ Sh[\mathbf{y}_j,\, x_j]\}.$$

Assume $j \leq n$ and a customer finishes service implying transition $j \to j-1$. Then we define the set $C_{j-1}(\mathbf{v}_j)$ which contains the values of the mark $\mathbf{v}_{j-1}$ shifted by y_j^1, i.e.,

$$C_{j-1}(\mathbf{v}_j) = \{\mathbf{v}_{j-1} :\ Sh[\mathbf{v}_j,\, y_j^1]\}.$$

Similarly, for $j > n$, the set $C_{j-1}^{(i)}(\mathbf{v}_j)$ contains values $\mathbf{v}_j$ shifted by y_j^1. In this case $\mathbf{Y}_j$, is also complemented by the service time B_{t+1} of the customer taken from the ith position in the queue. Formally,

$$C_{j-1}^{(i)}(\mathbf{v}_j) = \{\mathbf{v}_{j-1} :\ Ad\big[Sh[\mathbf{v}_j,\, y_j^1],\ B_{t+1} \in \Delta_j^i(Sh[\mathbf{y}_j,\, y_j^1])\}.$$

The analysis above is summarized in the following statement.

Lemma 1. The kernels of the mark transformations, under transition $j \to j+1$, are: for $j < n$, $i = \overline{2, n-j}$,

$$P_j(\mathbf{v}_j;\, C_{j+1}) \quad = \quad \begin{cases} A(x) \cdot \Delta B_j^i(Sh[\mathbf{y}_j,\, x_j]) & \text{for } C_{j+1} = C_{j+1}^{(i)}(\mathbf{v}_j,\, x), \\ 0, & \text{otherwise;} \end{cases}$$

for $j \geq n$, $i = \overline{2, n-j}$,

$$P_j(\mathbf{v}_j;\, C_{j+1}) \quad = \quad \begin{cases} A(x) & \text{for } C_{j+1} = C_{j+1}(\mathbf{v}_j,\, x), \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Analogously, when it goes from state j to state $j-1$ the kernels have the form: for $j \leq n$, $i = \overline{2, j}$,

$$Q_j(\mathbf{v}_j;\, C_{j-1}) \quad = \quad \begin{cases} 1_{\{Ad[\mathbf{v}_j, y_j^1] \in C_{j-1}\}} & \text{for } C_{j-1} = C_{j-1}(\mathbf{v}_j), \\ 0, & \text{otherwise;} \end{cases} \tag{19}$$

and for $j > n$, $i = \overline{2, j}$

$$Q_j(\mathbf{v}_j;\, C_{j-1}) \quad = \quad \begin{cases} \Delta B_j^i(Sh[\mathbf{y}_j, y_j^1], e_j(1)]) & \text{for } C_{j-1} = C_{j-1}(\mathbf{v}_j), \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Using these expressions, in the next section, we will consider the distribution of the marks.

5. Distributions of marks

Denote by $\mu_j(t,\cdot) = \mathbf{P}\{\mathbf{V}_j(t) \in \cdot\}$ the measure (distribution) of the mark $\mathbf{V}_j(t) = (X_j(t),\, \mathbf{Y}_j(t))$ in step $t \in \mathbb{N}_0$, $j \in \mathcal{J}$. Based on the kernels P_j and Q_j, we may write down the equations for the measures $\mu_j(t,\cdot)$

$$\mu_j(t+1,\cdot) = \int \mu_{j-1}(t, \mathbf{d}\mathbf{v}_{j-1})P_{j-1}(\mathbf{v}_{j-1}\,\cdot) + \int \mu_{j+1}(t, \mathbf{d}\mathbf{v}_{j+1})Q_{j+1}(\mathbf{v}_{j+1},\cdot),$$

which, in the terms of operators, can be expressed as follows:

$$\mu_j(t+1,\cdot) = \mu_{j-1}P_{j-1}(t,\cdot) + \mu_{j+1}Q_{j+1}(t,\cdot).$$

The following statement holds.

Corollary 1. By assumption A and B follow absolutely continuous distributions, then the measures of marks are also absolutely continuous and

$$\mu_j(t;C) = \int_C \cdots \int f_j(t;x,\, y^1,\ldots,y^{j\wedge n})dx \prod_{1\le i\le j\wedge n} dy^i,$$

where $f_j(\cdot)$ is the pdf of marks.

Proof. Using the proof by induction, assume that the initial state $J(0) = 0$ (at time $S(0) = 0$) with the mark $X(0) = A_0$. At step $t = 1$ (at time $S(1) = A_0$), $J(1) = 1$ and $X(1) = A_1,\ Y(1) = B_1$. It is shown in lemma 2 (Appendix A) that in this case subtracting and adding operators keep absolute continuity. Therefore, the measures of the marks are also absolutely continuous. ■

Introduce the following notations,

$$\mathbf{y}_j^i = (y_j^1,\ldots y_j^{i-1}, y_j^{i+1},\ldots y_j^{j\wedge n}) \text{ and } \mathbf{y}_j^i + u = (y_j^1+u,\ldots y_j^{i-1}+u, y_j^{i+1}+u,\ldots y_j^{j\wedge n}+u).$$

Theorem 1. The pdf f_j satisfies the following recursion:

$$f_1(t+1;x,y) = \left[\int_{u\ge 0} f_0(t;u,\infty)du\right] a(x)b(y) + \int_{u\ge 0} f_2(t;x+u,u,y+u)du; \tag{21}$$

for $2 \le j \le n$

$$\begin{aligned} f_j(t+1;x_j,\mathbf{y}_j) &= \sum_{1\le i\le j} a(x_j)b(y_j^i)\left(B(y_j^{i+1}) - B(y_j^{i-1})\right) \int_{u\ge 0} f_{j-1}(t;u,\mathbf{y}_j^i+u)\,du + \\ &+ \int_{u\ge 0} f_{j+1}(t;x_j+u,u,\mathbf{y}_j+u)du; \end{aligned} \tag{22}$$

for $j > n$

$$\begin{aligned} f_j(t+1;x_j,\mathbf{y}_j) &= a(x_j) \int_{u\ge 0} f_{j-1}(t;u,\mathbf{y}_j^i+u)\,du + \\ &+ \sum_{1\le i\le n} b(y_j^i)\left(B(y_j^{i+1}) - B(y_j^{i-1})\right) \int_{u\ge 0} f_{j+1}(t;x_j+u,u,\mathbf{y}_j+u)du. \end{aligned} \tag{23}$$

Also the following additional condition holds:

$$f_0(t+1;x,\infty) = \int_{u\ge 0} f_1(t;x+u,u)du. \tag{24}$$

Proof. We outline the straightforward proof. The state of the process Z such that $J(t+1)=1$ with the residual time x and the remaining arrival time y may occur if and only if the following points a) and b) hold:

(a) $J(t)=0$ (with any remaining time u to new arrival) and, by the agreement, the residual service time is infinite; during this step a new customer arrives and his service time is y with the probability $b(y)dy$, while the remaining arrival time is x with the probability $a(x)dx$; this explains the 1st summand in (21),

(b) $J(t)=2$ with the (arbitrary) minimal residual service time u and the appropriate residual times; during the step the time u is expired, and it explains the 2nd summand in (21).

For $2 \le J(t+1)=j \le n$, the process Z will be in a 'neighborhood' of the point $\mathbf{v}_j=(x_j,\ \mathbf{y}_j)$ with the probability $f_j(t+1;x_j,\ \mathbf{y}_j)\,dx_j\mathbf{dy}_j$, if and only if in step t, either:

(a) $J(t)=j-1$ with the probability $f_{j-1}(t;u,\mathbf{y}_j^i+u)\,du\mathbf{dy}_j^i$ in a 'neighborhood' of the point $(u,\mathbf{y}_j^i+u)$, the residual arrival time u is expired and the components of vector $\mathbf{v}$ are shifted by u; the new arrival time is x_j with the probability $a(x_j)dx_j$, the new customer finds free server and takes the service time y_j^i which is added to the shifted vector $\mathbf{y}_j$ (at the position keeping order) with the probability $(B(y_j^{i+1})-B(y_j^{i-1}))b(y_j^i)dy_j^i$. It explains the 1st term on the right-hand side of (22); or

(b) $J(t)=j+1$ in a 'neighborhood' of the point $(x_j+u,u,\mathbf{y}_j+u)$, with the probability $f_{j+1}(t;x_j+u,u,\mathbf{y}_j+u)dx_jdu\mathbf{dy}_j$, and after completion of service after time u, the process switches to state j in the 'neighborhood' of the point $(x_j,\mathbf{y}_j)$. This explains the 2nd term on the right side of (22).

By a similar way, with the evident changes, can be explained formula (23) when all servers are busy, $n<j\le n+m$, in which case new customer joins the queue or lost.

Finally, to explain (24) we note that $J(t+1)=0$, $X_0(t+1)=x$ and $Y_0(t+1)=\infty$ if and only if $J(t)=1$ with an arbitrary $Y_0(t)=u$ and an appropriate residual arrival time, and if the service ends before the new arrival. ■

6. Steady-state analysis

In this section we give the positive recurrence (stability) conditions of the Markov (regenerative) process Z.

6.1. Positive recurrence conditions of the process Z

According to the regeneration theory (see, for example, [10]), the stationary distribution of the discrete-time regenerative process $\{Z(t),\ t\ge 0\}$, as $t\to\infty$, exists if $\mathbf{E}[\tau]<\infty$ and the regeneration cycle length τ is aperiodic rv. Because the system under consideration is finite (the number of customers $\le n+m<\infty$) then the key condition $\mathbf{E}[\tau]<\infty$ is satisfied if the following easily verified condition (connecting the predefined input data) holds true:

$$\mathbf{P}\{A>B\}=\int_0^\infty B(x)A(dx)>0. \tag{25}$$

It is worth noting, that condition (25) is automatically satisfied, for example, for a Poisson input process, as well as for any renewal process where the inter-arrival interval A has an unbounded support. Note that all components of Z regenerate simultaneously when a new customer arrives in an empty system.

In the context of this study, it is useful to mention the regenerative stability analysis of a repairable finite system containing n unreliable elements with non-exponential lifetimes studied in the recent paper [37].

Remark 2. The regenerative stability analysis remains basically unchanged if the buffer size is unlimited, $m = \infty$. However, in this scenario, we need the well-known negative drift condition [10, 36]

$$\rho := \frac{\mu_B}{n\mu_A} < 1, \tag{26}$$

in addition to the 'regeneration' condition (25). We note that if $n = 1$, then condition (26) implies (25) and that, because the main component process $J(t)$ is itself regenerative, then (under conditions (25), (26) the stationary distribution $\{\pi_j\}$ exists as well.

If the number of servers $n = \infty$, then condition (25) implies stationarity, provided conditions $\mu_A < \infty$, $\mu_B < \infty$ are fulfilled (see [36]). Note that condition (25) is not redundant, since otherwise a basic process describing the dynamics of this system has no classical regenerations.

Remark 3. It is important to note that, within the framework of the approach, we can also consider the unfinished work process in state $J(t)$,

$$W_{J(t)}(t) = \sum_{1\le i\le J(t)\wedge n} Y_j^{(i)} + \sum_{1\le i\le [J(t)-n]^+} B_i,$$

which also regenerates at the moment when a new customer arrives in an empty system.

Remark 4. It follows from the Wald's identity that the mean regeneration cycle length in continuous time is finite $\mathbf{E}[R] = \mathbf{E}[\tau]\mathbf{E}[A] < \infty$ provided $\mathbf{E}[\tau] < \infty$. Because the cdf $A(\cdot)$, being absolutely continuous is non-lattice, then the stationary distribution of the process in continuous time also exists. Moreover in this case the cycle length R is spread-out and the convergence to stationary distribution in total variation holds as well [10].

6.2. Steady-state Equations

If the stationary distributions (measures) of marks exist then they satisfy the following equations,

$$\mu_j(\cdot) = \int \mu_{j-1}(\mathbf{dv}_{j-1})P_{j-1}(\mathbf{v}_{j-1},\ \cdot) + \int \mu_{j+1}(\mathbf{dv}_{j+1})Q_{j+1}(\mathbf{v}_{j+1},\ \cdot)\ \ (j \in \mathcal{J}),$$

with the evident boundary conditions for $j = 0$ and $j = m + n$. These equations can be rewritten in the operator's form as

$$\mu_j(\cdot) = \mu_{j-1}P_{j-1}(\cdot) + \mu_{j+1}Q_{j+1}(\cdot)\ \ (j \in \mathcal{J}). \tag{27}$$

It is worth mentioning that the kernel $P_j + Q_j$ is the identity transformation expressed as $P_j + Q_j = 1_{\{\mathbf{v}_j\in\cdot\}}$. Furthermore, we can utilize operator notation for transition kernels as

$$\mu_j(P_j(\cdot) + Q_j(\cdot)) = \mu_{j-1}P_{j-1}(\cdot) + \mu_{j+1}Q_{j+1}(\cdot),$$

and obtain the recurrence relation

$$\mu_{j-1}P_{j-1}(\cdot) - \mu_jQ_j(\cdot) = \mu_jP_j(\cdot) - \mu_{j+1}Q_{j+1}(\cdot), \tag{28}$$

with the equation

$$\mu_0(C_0) = \mu_1Q_1(C_0) \tag{29}$$

for the boundary state $j = 0$. Because $y_0^1 = \infty$ in state $j = 0$, then $P_0(\mathbf{v}_0,\ C_0) = 1_{\{\mathbf{v}_0\in\cdot\}}$, and (29) can be rewritten as

$$\mu_0P_0(C_0) = \mu_1Q_1(C_0) \quad \text{or} \quad \mu_1Q_1(C_0) - \mu_0P_0(C_0) = 0.$$

In turn, the latter result allows us to rewrite relation (28) in the form

$$\mu_{j+1}Q_{j+1}(\cdot) = \mu_jP_j(\cdot)\ \ \forall j \in \mathcal{J}. \tag{30}$$

The results above have rather theoretical interest since analytical solutions of (28), (29) seem inaccessible while the numerical solutions are labor-intensive. For this reason, in the next section, we outline how to calculate the main performance metrics directly using the marks.

7. Calculation of the main performance metrics in terms of the marks

In this section, we show that the main performance metrics can be directly obtained in terms of marks of the process Z. For this we use representation of the process $L(s)$ in continuous time (see Figure 2) and its regenerative structure. The regenerative structure allows constructing confidence interval using a single (large enough) trajectory based on the regenerative version of the Central Limit Theorem [23]. Thus, to estimate the main Quality of Service (QoS) metrics we simulate a large enough number K of the paths of the process Z up to the 1st regeneration time $S(\tau_1)$. Thus (supplying the index k to variables from the k-th cycle) based on the formulas (5 – 7), we obtain the sample-mean estimates of

- the cdf of the busy period and its first moment,

$$\hat{F}_{\Pi}(x) = \frac{1}{K}\sum_{1\le k\le K} 1_{\{\Pi^{(k)}\le x\}} \text{ and } \hat{\mu}_{\Pi} = \frac{1}{K}\sum_{1\le k\le K}\Pi^{(k)};$$

- the cdf of the regeneration period and its first moment,

$$\hat{F}_{R}(x) = \frac{1}{K}\sum_{1\le k\le K} 1_{\{R^{(k)}\le x\}} \text{ and } \hat{\mu}_{R} = \frac{1}{K}\sum_{1\le k\le K} R^{(k)};$$

- the steady state probabilities

$$\hat{\pi}_j = \frac{1}{\hat{\mu}_R}\hat{S}_j^{(k)}(\tau) = \frac{\sum_{1\le k\le K} S_j^{(k)}(\tau)}{\sum_{1\le k\le K} R^{(k)}},$$

where

$$\hat{S}_j^{(k)}(\tau) = \sum_{1\le t\le \hat{S}^{(k)}(\tau)} \hat{T}_j(t) 1_{\{J(t)=j\}}$$

is the estimation of the time spent by the process in state j during the 1st regeneration period along the kth trajectory and $\hat{S}^{(k)}(\tau)$ is the estimate of the kth regeneration period length;

- all other mean estimates are calculated by replacing in formulas of Section 3.3 the probabilities π_j by their estimates $\hat{\pi}_j$.

Appendix B presents the Algorithm for calculating the stationary performance metrics of the model. The algorithm constructs the process Z trajectories based on the transformations of marks and regenerative simulation method. For numerical analysis, the programming code in Python has been developed.

8. Numerical examples

In this section, we consider a few numerical examples to compare the performance metrics estimated by the Algorithm and the available analytical expressions for systems $(M|GI|1)$ and $(M|M|n,m)$ (or $M/GI/1$ and $M/M/n/m$ in classical notation [2]). Some additional examples of systems $(GI|GI|n,m)$ are presented as well. In all experiments, $K = 10^5$ realizations (paths) of the Algorithm are used. Each path is time time duration of the 1st regeneration cycle.

We use the standard metrics, μ_A, μ_B and CoV's v_A, v_B, for the inter-arrival and service time, respectively. Table 1 demonstrates the characteristics of the used distributions Gnedenko-Weibull, $GW := GW(\alpha,\beta)$, and Gamma, $\Gamma := \Gamma(\alpha,\beta)$, where α is the shape parameter, β is the scale parameter.

Table 1: *Distributions and their characteristics*

Distribution	$GW(\alpha,\beta)$	$\Gamma(\alpha,\beta)$
pdf	$f(s)=\dfrac{\alpha e^{-(s/\beta)^\alpha}(s/\beta)^{\alpha-1}}{\beta}$, $s>0$	$f(s)=\dfrac{\beta^\alpha e^{-s/\beta}s^{\alpha-1}}{\Gamma(\alpha)}$, $s>0$
mean	$\mu=\beta\Gamma\left(1+\dfrac{1}{\alpha}\right)$	$\mu=\dfrac{\alpha}{\beta}$
variance	$\sigma^2=\beta^2\Gamma\left(1+\dfrac{2}{\alpha}\right)-\mu^2$	$\sigma^2=\dfrac{\alpha}{\beta^2}$
CoV	$v=\dfrac{\sigma}{\mu}$	$v=\dfrac{\sqrt{\alpha}}{\alpha}$
distribution parameters	$\alpha,\beta=\dfrac{\mu}{\Gamma(1+1/\alpha)}$	$\alpha=v^{-2},\quad \beta=\mu v^2$

8.1. Comparison with analytical results and numerical analysis

The analytical expressions for the mean performance metrics of a $(M|GI|1)$ system can be found, for instance, in [10, 34, 35, 38]. It is known that these expressions depend on the 1st and 2nd moments of service time only, while the cdf F_Π of the busy period depends on the service time cdf [34],

$$F_\Pi(t)=\int_0^t\sum_{k=1}^{\infty}e^{-\mu_A^{-1}x}\frac{(\mu_A^{-1}x)^{k-1}}{k!}b_{(k)}(x)dx, \tag{31}$$

where $b_{(k)}(x)$ is the k-fold convolution of the service time pdf $b(x)$.

To study the infinite-buffer $(M|GI|1)$ system using the Algorithm (Appendix B) we take buffer size $m=10^4$. First, consider the service time distribution $B(\cdot)\sim\Gamma$. (We denote this system by $(M|\Gamma|1)$.) The mean inter-arrival time of the Poisson input is $\mu_A=2$. The mean service time is $\mu_B=\mu_A\rho$, where traffic intensity $\rho=0.5,\,0.8,\,0.95$, while $v_B=0.5,\,1,\,10$.

Fig. 3 shows the cdf F_Π[1] of the busy period where we use: solid lines for $\rho=0.5$, dashed lines

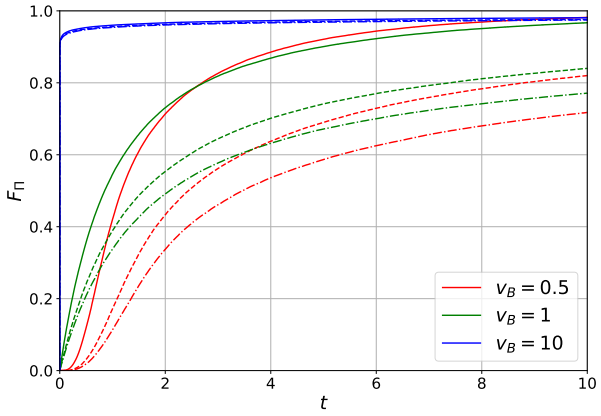


Figure 3: *F_Π of the system $(M|\Gamma|1)$ for different values of v_B and ρ*

for $\rho=0.8$, and dash-dotted lines for $\rho=0.95$. It is seen that v_B significantly affects F_Π. The biggest value of v_B has the minimal impact on F_Π, resulting in all curves (black lines) being close to 1. In contrast, for small values ($v_B<10$), F_Π considerably depends on ρ.

Tables 2 and 4 show $\hat{\mu}_\Pi$ and $\hat{v}_\Pi$ obtained by the algorithm for $B(\cdot)\sim\Gamma$ and GW. Using known analytical formulas (for instance, [10]), $\mu_\Pi=\mu_B/(1-\rho)$, then, for $\rho=0.5,\,0.8,\,0.95$, it follows

[1]We will use the notation F_Π instead of $F_\Pi(\cdot)$.

that $\mu_\Pi = 2$, 8, 37.99, respectively. The calculated by the algorithm values of $\hat{\mu}_\Pi$ are close to the analytical ones.

Table 2: *$\hat{\mu}_\Pi$ of the system $(M|GI|1)$ for $B(\cdot) \sim \Gamma$ / GW (by the algorithm)*

$\hat{\mu}_\Pi$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
$v_B = 0.5$	2.0062 / 2.0034	8.0093 / 8.0229	38.2819 / 38.0612
$v_B = 1$	1.9947 / 1.9976	7.9947 / 7.9713	38.5953 / 38.1866
$v_B = 10$	2.0202 / 1.9721	7.9357 / 8.4624	38.8837 / 31.2089

Table 3 presents v_Π calculated, by [35], as

$$v_\Pi = +\sqrt{(\rho + v_B^2)/(1-\rho)}. \tag{32}$$

Table 4 contains values of $\hat{v}_\Pi$ obtained by the algorithm for $B(\cdot) \sim \Gamma/$ GW which are consistent

Table 3: *v_Π calculated by (32)*

v_Π	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
$v_B = 0.5$	1.2247	2.2913	4.8989
$v_B = 1$	1.7321	3.0000	6.2449
$v_B = 10$	14.1774	22.4499	44.9333

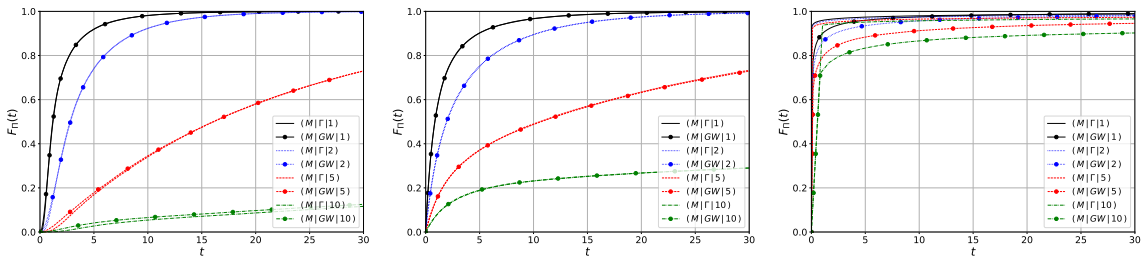
with the results in Table 3. In particular, for fixed ρ, the increase of v_B increases $\hat{v}_\Pi$ as well, while the value of $\hat{v}_\Pi$ is the same, when $B(\cdot) \sim \Gamma$, GW.

Table 4: *$\hat{v}_\Pi$ for $B(\cdot) \sim \Gamma$ / GW (by the algorithm)*

v_Π	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.95$
$v_B = 0.5$	1.2271 / 1.2218	2.2682 / 2.2856	4.9411 / 4.9091
$v_B = 1$	1.7394 / 1.7374	2.9561 / 2.9794	6.2628 / 6.5084
$v_B = 10$	13.6669 / 14.0985	22.9887 /21.8648	38.8837 / 35.3533

8.2. Numerical analysis of the system $(M|GI|n)$

In this section, we present the numerical study of n-server system $(M|GI|n)$ with $n = 1$, 2, 5, 10, $\mu_A = 2$, and the fixed traffic intensity $\rho = \dfrac{\mu_B}{n \cdot \mu_A} = 0.5$. It then follows that $\mu_B = 1$, 2, 5, 10. (The case $n = 1$ is included to compare with the earlier obtained results.) Fig. 4 demonstrates F_Π for $B(\cdot) \sim GW$ and Γ.



a) $v_B = 0.5$ b) $v_B = 1$ c) $v_B = 10$

Figure 4: *F_Π for $B(\cdot) \sim \Gamma$ and GW for $n = 1, 2, 5, 10$*

It is seen that F_Π curves are quite close for the given $B(\cdot)$ and v_B for all n, except $v_B = 10$ (fig. 4c). These results demonstrate that the numerical analysis is informative, since illustrates and complements the analytical solutions when they exist. (In this regard compare the results from (31) and the black curves demonstrated F_Π for the $(M|GI|1)$ system.)

As $v_B = 10$, the shape of $B(\cdot)$ plays a significant role and the curves of F_Π for $B(\cdot) \sim GW$ are below than that for $B(\cdot) \sim \Gamma$. For each v_B, F_Π increases as the number of servers n increases. Note that, as $v_B > 1$, the shape parameter $0 < \alpha < 1$ and thus GW service time distribution is heavy-tailed. This observation may be important for a further study of the insensitivity problem.

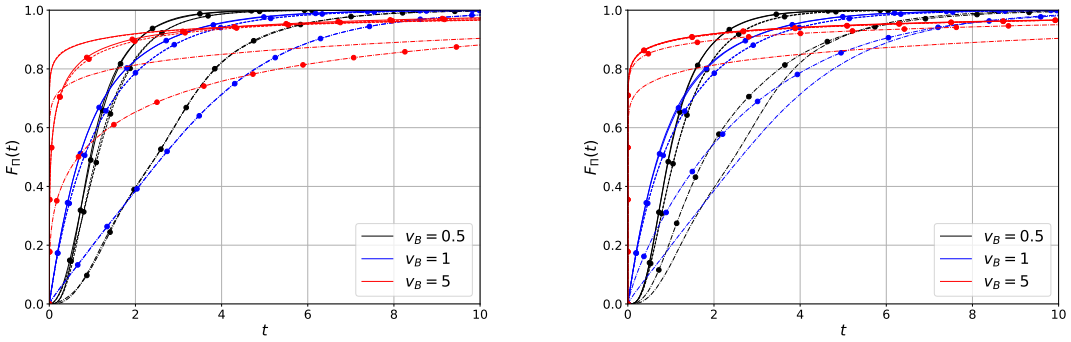
Table 5 shows close values of $\hat{\mu}_\Pi$ for the given cdf's $B(\cdot)$ and v_B if $v_B \leq 1$. For a fixed v_B, increasing n increases $\hat{\mu}_\Pi$ because μ_B increases as well. This interesting observation shows that an increasing μ_B makes performance worse in spite of the increasing the number of servers provided $\rho = 0.5$ is fixed.

Table 5: $\hat{\mu}_\Pi$ *of the system* $(M|GI|n)$ *for* $B(\cdot) \sim \Gamma$ / GW

$\hat{\mu}_\Pi$	$n = 2$	$n = 5$	$n = 10$
$v_B = 0.5$	4.1047 / 4.1043	23.2388 / 23.1257	297.1719 / 296.7376
$v_B = 1$	4.0125 / 4.0245	22.8843 / 23.0876	295.5491 / 296.3262
$v_B = 10$	4.0356 / 3.8161	23.0863 / 22.1556	305.3809 / 301.4704

8.3. Numerical analysis of the system $(GI|GI|3,5)$

In this section, we investigate the dependence of the system's metrics $(GI|GI|3,5)$ on the shape of the inter-arrival time distribution $A(\cdot)$ and its CoV v_A, assuming $\mu_A = 2$, $\mu_B = 2$ and $v_A = v_B = 0.5, 1, 5$. Fig. 5a) demonstrates F_Π of the system $(\Gamma|GI|3,5)$ for $B(\cdot) \sim \Gamma, GW$ where the curves with markers (with no markers) relate to $B(\cdot) \sim GW$ $(B(\cdot) \sim \Gamma)$. Moreover, we use solid lines for $v_A = 0.5$, dashed lines for $v_A = 1$, and dash-dotted lines for $v_A = 5$.



a) $B(\cdot) \sim \Gamma$ and GW varying v_A, v_B b) $A(\cdot) \sim \Gamma$ and GW varying v_A, v_B

Figure 5: F_Π *for* $(GI|GI|3,5)$ *system*

It is seen that an increase of v_A results in a stochastically decrease of F_Π. At that, for the given cdf $B(\cdot)$, a proximity of the curves with and without markers shows that F_Π weakly depends on the shape of $B(\cdot)$ when $v_B \leq 1$, and strongly depends on both values v_A and v_B (Fig. 5a).

Fig. 5b) shows F_Π for $B(\cdot) \sim \Gamma$ and $A(\cdot) \sim GW, \Gamma$. Here the curves with markers (with no markers) relate to $A(\cdot) \sim GW$ $(A(\cdot) \sim \Gamma)$. It shows a weak dependence of F_Π on the shape of $A(\cdot)$ when $v_A \leq 1$ and $v_B \leq 1$ and if $A(\cdot) \sim \Gamma$ and GW.

Table 6 demonstrates that $\hat{\mu}_\Pi$ strongly depends on both v_A and v_B and changes slightly if $B(\cdot) \sim \Gamma$ and GW. Conversely, v_A and the shape of $B(\cdot)$ have weak impact on $\hat{v}_\Pi$ when $v_A \leq 1$ and $v_B \leq 1$. An increase of v_B implies increasing $\hat{\mu}_\Pi$ and $\hat{v}_\Pi$, while an increase of v_A results in increasing of $\hat{\mu}_\Pi$ and decreasing of $\hat{v}_\Pi$ for fixed v_B.

Table 6: $\hat{\mu}_\Pi$, $\hat{v}_\Pi$ *of the system* $(\Gamma|GI|3,5)$ *for* $B(\cdot) \sim \Gamma/GW$

	$\hat{\mu}_\Pi$			$\hat{v}_\Pi$		
	$v_B = 0.5$	$v_B = 1$	$v_B = 5$	$v_B = 0.5$	$v_B = 1$	$v_B = 5$
$v_A = 0.5$	1.1340 / 1.1323	1.1407 / 1.1449	1.2421 / 1.1956	0.6444 / 0.6439	1.1528 / 1.1519	5.3147 / 5.3594
$v_A = 1$	1.2955 / 1.3024	1.3032 / 1.3001	1.2925 / 1.3228	0.6537 / 0.6605	1.1176 / 1.1184	5.1876 / 4.9622
$v_A = 5$	2.7021 / 2.6981	3.0914 / 3.0906	3.4152 / 4.6756	0.6236 / 0.6184	0.7811 / 0.7784	3.2407 / 2.8206

Comparing Tables 6 and 7, we may conject that both $\hat{\mu}_\Pi$ and $\hat{v}_\Pi$ slightly depend on the shape of $A(\cdot)$ when $v_A \le 1$ and $v_B \le 1$ and for given cdf $A(\cdot)$.

Table 7: $\hat{\mu}_\Pi$, $\hat{v}_\Pi$ *for the system* $(GW|\Gamma|3,5)$

	$\hat{\mu}_\Pi$			$\hat{v}_\Pi$		
	$v_B = 0.5$	$v_B = 1$	$v_B = 5$	$v_B = 0.5$	$v_B = 1$	$v_B = 5$
$v_A = 0.5$	1.1378	1.1535	1.2973	0.6352	1.1369	5.2320
$v_A = 1$	1.2995	1.3050	1.3009	0.6600	1.1193	5.2295
$v_A = 5$	2.3516	2.5265	1.8555	0.7777	1.0115	4.6257

Finally, Table 8 shows that for the given $A(\cdot)$, $B(\cdot)$ the metrics $\hat{\mu}_I$, $\hat{\mu}_Q$, $\hat{\mu}_V$ and $\hat{\mu}_W$ of the system $(GI|GI|3,5)$ change slightly with changing v_B, the shape of $B(\cdot)$, when $v_B \le 1$, and the shape of $A(\cdot)$, when $v_A \le 1$. However they strongly depend on v_A, as $v_A \le 1$.

Table 8: *The performance metrics of the system* $(GI|GI|3,5)$

System performance		$A(\cdot) \sim \Gamma$, $B(\cdot) \sim \Gamma$ / $A(\cdot) \sim \Gamma$, $B(\cdot) \sim GW$			$A(\cdot) \sim GW$, $B(\cdot) \sim \Gamma$		
		$v_B = 0.5$	$v_B = 1$	$v_B = 5$	$v_B = 0.5$	$v_B = 1$	$v_B = 5$
$v_A = 0.5$	$\hat{\mu}_I$	0.5018 / 0.4997	0.4979 / 0.5002	0.5077 / 0.5026	0.5015	0.5001	0.5240
	$\hat{\mu}_Q$	$2 \cdot 10^{-6}$ / $2 \cdot 10^{-6}$	$5 \cdot 10^{-5}$ / $7 \cdot 10^{-5}$	0.0127 / 0.0078	$2 \cdot 10^{-5}$	10^{-3}	0.0143
	$\hat{\mu}_V$	$4 \cdot 10^{-6}$ / $4 \cdot 10^{-6}$	10^{-3} / 10^{-3}	0.0253 / 0.0155	$4 \cdot 10^{-5}$	0.0002	0.0286
	$\hat{\mu}_W$	1.0037 / 0.9993	0.9958 / 1.0003	1.0153 / 1.0053	1.0029	1.0003	1.0481
$v_A = 1$	$\hat{\mu}_I$	0.5019 / 0.5014	0.5047 / 0.5051	0.5148 / 0.5195	0.5000	0.5055	0.5127
	$\hat{\mu}_Q$	0.0021 / 0.0024	0.0031 / 0.0032	0.0177 / 0.0128	0.0021	0.0031	0.0168
	$\hat{\mu}_V$	0.0042 / 0.0047	0.0063 / 0.0063	0.0354 / 0.0257	0.0043	0.0062	0.0337
	$\hat{\mu}_W$	1.0037 / 1.0027	1.093 / 1.0102	1.0296 / 1.0389	1.0000	1.0109	1.0254
$v_A = 5$	$\hat{\mu}_I$	0.5209 / 0.5207	0.5165 / 0.5152	0.5164 / 0.4832	0.6287	0.6166	0.5286
	$\hat{\mu}_Q$	0.1872 / 0.1874	0.1732 / 0.1729	0.0903 / 0.0940	0.1644	0.1502	0.0653
	$\hat{\mu}_V$	0.3743 / 0.3748	0.3465 / 0.3459	0.1805 / 0.1880	0.3288	0.3005	0.1306
	$\hat{\mu}_W$	1.0418 / 1.0415	1.0330 / 1.0304	1.0329 / 0.9664	1.2574	1.2332	1.0573

Note that the case $v_A = v_B = 1$ corresponds to the system $(M|M|3,5)$, and these numerical results are close to that obtained in Table 9 by analytical formulas [38].

Table 9: *The performance metrics of the system* $(M|M|3,5)$, *analytical results*

m_I	m_Q	m_V	m_W
0.5044	0.0030	0.0060	1.0060

9. Conclusion

The purpose of the current research is to analyze the system $(GI|GI|n,m)$ using the Marked Markov Process and transformations of the marks to compute the main performance indicators

of the system. The approach provides the analytical expressions for marks' distributions, as well as conditions for the existence of the stationary distribution of the main process.

Because of the high complexity of the system $(GI|GI|n,m)$, a simulation algorithm is also proposed to evaluate the system's performance metrics. At that, the numerical examples demonstrate consistency with the known analytical expressions. Moreover, the proposed approach allows the simulation-based numerical analysis of the stationary performance indicators, in particular, to study their dependency concerning to given inter-arrival and service time distributions and the coefficients of variation. The obtained in this paper numerical results do not allow making unambiguous conclusions on the sensitivity/insensitivity of the basic performance indicators. However, they show that conclusions based solely on the two first moments of the inter-arrival and service times may lead to inaccurate conclusions of the system performance indicators. Furthermore, the proposed approach applies the simulation directly to construct the trajectories of the process based on the marks transformations, and by our opinion, it opens some new opportunities in stochastic modeling and simulation. In future research, the authors are going to extend the proposed concept of the MMP to consider more complex stochastic systems.

Acknowledgments

This work was partially supported by the Moscow Center for Fundamental and Applied Mathematics (recipient E. Morozov).

References

[1] A.N. Shiryaev. *Fundamentals of Stochastic Financial Mathematics. Volume 1. Facts. Models*. in Russ. Moscow: FAZIS, 1998.

[2] D. G. Kendall. "Stochastic Processes Occurring in the Theory of Queues and Their Analysis by the Method of Embedded Markov Chains". In: *Annals of Math. Stat.* 24 (1953), pp. 338–354.

[3] V. V. Rykov and D. V. Kozyrev. *Fundamentals of Queuing Theory*. in Russ. NIC INFRA-M, 2016.

[4] D. R. Cox. "The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 51 (1955), pp. 433–441.

[5] Y. K. Belyaev. "Linear Markov processes and their applications to reliability problems". In: Proceedings of the VI Alunuon Workshop on Probability Theory and Mathematical Statistics. Tomsk State University. Vilnius, Lithuania, 1962.

[6] I. N. Kovalenko. *Investigations on Analysis of Complex Systems Reliability*. Naukova Dumka, 1976.

[7] E. Cinlar. "On semi-Markov processes on arbitrary spaces". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 66.2 (1969), pp. 381–392. DOI: 10.1017/S0305004100045096.

[8] W. L. Smith. "Regenerative stochastic processes". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 232 (1955), pp. 6–31.

[9] K. Sigman and R Wolff. "A review of regenerative processes". In: *SIAM Rev.* 35 (1993), pp. 269–288.

[10] S. Asmussen. *Applied Probability and Queues, 2nd ed.* Springer, 2003.

[11] R. Serfozo. *Basics of Applied Stochastic Processes*. Springer, 2009.

[12] B. A. Sevast'yanov. "An Ergodic Theorem for Markov Processes and Its Application to Telephone Systems with Refusals". In: *Theory Probab. Appl.* 2.1 (1957), pp. 104–112.

[13] B. V. Gnedenko. "On an unloaded duplication". In: *Eng. Cybern.* 4 (1964). in Russian, pp. 3–12.

[14] B. V. Gnedenko. "On a repairable duplication". In: *Eng. Cybern.* 5 (1964). in Russian, pp. 111–118.

[15] A.D. Soloviev. "Asymptotic distribution of the life time of a doubled element". In: *Izv. Akad. Nauk SSSR. Tehn. Kibernet* 5 (1964). in Russian, pp. 119–121.

[16] V. V. Rykov and D. V. Kozyrev. "Analysis of Renewable Reliability Systems by Markovization Method". In: *Analytical and Computational Methods in Probability Theory*. Ed. by Vladimir V. Rykov, Nozer D. Singpurwalla, and Andrey M. Zubkov. Cham: Springer International Publishing, 2017, pp. 210–220.

[17] V. Rykov et al. "On Sensitivity Analysis of Steady State Probabilities of Double Redundant Renewable System with Marshall-Olkin Failure Model". In: *Distributed Computer and Communication Networks*. Ed. by Vladimir M. Vishnevskiy and Dmitry V. Kozyrev. Cham: Springer International Publishing, 2018, pp. 234–245.

[18] V. Rykov. "On Reliability of Renewable Systems". In: *Reliability Engineering*. CRC Press, 2018, pp. 173–196. DOI: 10.1201/9781351130363-9.

[19] V. V. Rykov, N. M. Ivanova, and D. V. Kozyrev. "Sensitivity Analysis of a k-out-of-n:F System Characteristics to Shapes of Input Distribution". In: *Distributed Computer and Communication Networks*. Ed. by Vladimir M. Vishnevskiy, Konstantin E. Samouylov, and Dmitry V. Kozyrev. Cham: Springer International Publishing, 2020, pp. 485–496.

[20] P. W. Glynn and D. L. Iglehart. "Simulation methods for queues: An overview". In: *Queueing Systems* 3 (1988), pp. 221–255.

[21] S. G. Henderson and P. W. Glynn. "Regenerative steady-state simulation of discrete-event systems". In: *ACM Trans. Model. Comput. Simul.* 11.4 (Oct. 2001), pp. 313–345. ISSN: 1049-3301. DOI: 10.1145/508366.508367.

[22] Peter J. Haas. *Stochastic Petri Nets: Modelling, Stability, Simulation*. Springer, 2002.

[23] Søren Asmussen and Peter W. Glynn. *Stochastic Simulation: Algorithms and Analysis*. Springer, 2007.

[24] M. Miyazawa. "Palm Calculus, Reallocatable GSMP and Insensitivity Structure". In: *Queueing Networks: A Fundamental Approach*. Ed. by R. J. Boucherie and N. M. van Dijk. Boston, MA: Springer US, 2011, pp. 141–215. DOI: 10.1007/978-1-4419-6472-4_4.

[25] P. W. Glynn. "A GSMP Formalism for Discrete Event Systems". In: *Proceedings of the IEEE* 77.1 (1989), pp. 14–23. DOI: 10.1109/5.21067.

[26] P. W. Glynn and P. J. Haas. "Laws of Large Numbers and Functional Central Limit Theorems for Generalized Semi-Markov Processes". In: *Stochastic Models* 22.2 (2006), pp. 201–231. DOI: 10.1080/15326340600648997.

[27] P. W. Glynn and P. J. Haas. "On Transience and Recurrence in Irreducible Finite-State Stochastic Systems". In: *ACM Transactions on Modeling and Computer Simulation* 25.4 (2015), pp. 1–19. DOI: 10.1145/2699721.

[28] M. Miyazawa and G. Yamazaki. "The basic equations for a supplemented GSMP and its applications to queues". In: *Journal of Applied Probability* 25.3 (1988), pp. 565–578. DOI: 10.2307/3213985.

[29] A. Coyle. "Sensitivity Bounds on a GI/M/n/n Queueing System". In: *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics* (1989).

[30] X. Yin, X. Ma, and K. S. Trivedi. "MAC and application level performance evaluation of beacon message dissemination in DSRC safety communication". In: *Perform. Evaluation* 71 (2014), pp. 1–24.

[31] V. Rykov and N. Ivanova. "On the reliability of double redundant system with arbitrary distributions of life and repair times of its elements". In: Materials of the XXII International Conference named after A.F. Terpugov. in Russ. Tomsk State University. Tomsk, 2023.

[32] V. Rykov and N. Ivanova. "On the dependability function of a $< GI_{k\le n}|GI|l >$ system. Part I. Analytical results". In: *Dependability* 3 (2024), pp. 34–43. DOI: 10.21683/1729-2646-2024-24-3-34-43.

[33] V. Rykov and N. Ivanova. "On the dependability function of a $< GI_{k\le n}|GI|l >$ system. Part II. Numerical study and sensitivity analysis". In: *Dependability* 4 (2024), pp. 3–11. DOI: 10.21683/1729-2646-2024-24-4-3-11.

[34] J. W. Cohen. *The Single Server Queue (2nd ed.)* North-Holland, 1982.

[35] L. Kleinrock. *Queueing Systems - Vol. 1: Theory*. Wiley, 1975.

[36] E. Morozov and B. Steyaert. *Stability analysis of regenerative queueing models*. Springer, 2021.

[37] E. Morozov and V. Rykov. "On the Positive Recurrence of Finite Regenerative Stochastic Models". In: *Mathematics* 11.4754 (2023). DOI: 10.3390/math11234754.

[38] P. P. Bocharov, C. D'Apice, and A. V. Pechinkin. *Queueing Theory*. Berlin, Boston: De Gruyter, 2003. DOI: doi:10.1515/9783110936025.

[39] H. A. David and H. N. Nagaraja. *Order Statistics*. John Wiley & Sons, 2003. DOI: 10.1002/0471722162.

Appendix A. Auxiliary information

Denote by

$$X^{(1)} \le X^{(2)} \le \cdots \le X^{(i)} \le \cdots \le X^{(n)}$$

the variation series for independent sample of rv's X_i $(i = \overline{1,n})$. In order to maintain consistency in the notation for the members of the variation series, we use upper indices in brackets instead of the traditional lower indices. The corresponding rv's values are also denoted with superscripts but without parentheses.

It is well-known (see [39]) that the joint pdf $f(\mathbf{x}) = f(x^1,\dots,x^n)$ of the variation series $X^{(i)}$ of a sample X_i of independent rv's with pdf $f(x)$ has the form

$$f(x^1,\dots,x^n) = n!\prod f(x^i) \text{ where } x^1 \le \cdots \le x^i \le \cdots \le x^n.$$

Lemma 2. Suppose that the pdf $f_{\mathbf{X}_n}(\mathbf{x}_n)$ of the original vector $\mathbf{X}_n$ is continuous in all variables. Then the pdf $f_{Sh[\mathbf{X}_n,X]}(\mathbf{x}_n)$ of the variation series $Sh[\mathbf{X}_n,X]$ obtained by subtracting an independent rv X with the cdf $F_X(x)$ and the pdf $f_X(x)$ satisfying the condition $\mathbf{P}\{X \le X^{(1)}\} = 1$, is defined by the relation

$$f_{Sh[\mathbf{X}_n,X]}(\mathbf{x}_n) = \int_0^{x^1} f_{\mathbf{X}_n}(\mathbf{x}_n + u) f_X(u)du. \tag{33}$$

In case $X = X^{(1)}$ it holds

$$f_{Sh[\mathbf{X}_n,X]}(\mathbf{x}_{n-1}) = \int_0^{x^1} f_{\mathbf{X}_n}(u, \mathbf{x}_{n-1} + u)du. \tag{34}$$

The pdf $f_{Ad[\mathbf{X}_n,X]}(\mathbf{x}_{n+1})$ of a variation series $Ad[\mathbf{X}_n,X]$ obtained by adding an independent rv X, with the cdf $F_X(x)$ and the pdf $f_X(x)$, has the form

$$f_{Ad[\mathbf{X}_n,X]}(\mathbf{x}_{n+1}) = \sum_{1\le i\le n+1} (F_X(x^i) - F_X(x^{i-1}))f_X(x^i) f_{\mathbf{X}_n}(\mathbf{x}_{n+1} \setminus \{x^i\}), \tag{35}$$

where $x^0 = 0,\ x^{n+2} = \infty$.

Proof. Indeed, the subtracting can be viewed as a non-degenerate linear transformation with a determinant equal to one. Formulas (33, 34) follow by the total probability law taking into account the independence of the rv X. Denote by $d\mathbf{x}_n = dx^1 dx^2 \dots dx^n$. Then, if $X \le X^{(1)}$, we get formula (33):

$$\begin{aligned} f_{Sh[\mathbf{X}_n,X]}(\mathbf{x}_n)d\mathbf{x}_n &= \mathbf{P}\{\mathbf{X}_n - X \in [\mathbf{x}_n, \mathbf{x}_n + d\mathbf{x}_n]\} \\ &= \int_0^{x^1} \mathbf{P}\{\mathbf{X}_n - X \in [\mathbf{x}_n, \mathbf{x}_n + d\mathbf{x}_n] | X \in [u, u+du]\}\mathbf{P}\{X \in [u, u+du]\} \\ &= \int_0^{x^1} \mathbf{P}\{\mathbf{X}_n \in [\mathbf{x}_n + u, \mathbf{x}_n + u + d\mathbf{x}_n]\}\mathbf{P}\{X \in [u, u+du]\} \\ &= \int_0^{x^1} f_{\mathbf{X}_n}(\mathbf{x}_n + u) f_X(u) du\, d\mathbf{x}_n. \end{aligned}$$

If $X = X^{(1)}$, then the variation series $Sh[\mathbf{X}_n, X]$ transforms to vector $(X^{(2)} - X^{(1)}, \dots, X^{(n)} - X^{(1)})$ of dimension $n-1$. The joint pdf of this vector satisfies the following relation,

$$\begin{aligned} &f_{Sh[\mathbf{X}_n,X]}(\mathbf{x}_{n-1})d\mathbf{x}_{n-1} = \mathbf{P}\{X^{(2)} - X^{(1)} \in [x^1, x^1 + dx^1], \dots, X^{(n)} - X^{(1)} \in [x^{n-1}, x^{n-1} + dx^{n-1}]\} \\ &= \int_0^{x^1} \mathbf{P}\{X^{(1)} \in [u, u+du], X^{(2)} - X^{(1)} \in [x^1, x^1 + dx^1], \dots, X^{(n)} - X^{(1)} \in [x^{n-1}, x^{n-1} + dx^{n-1}]\} \\ &= \int_0^{x^1} f_{\mathbf{X}_n}(u, x^1 + u, \dots, x^{n-1} + u) du\, d\mathbf{x}_{n-1} = \int_0^{x^1} f_{\mathbf{X}_n}(u, \mathbf{x}_{n-1} + u) du\, d\mathbf{x}_{n-1}, \end{aligned}$$

which implies relation (34).

Further, when a rv X by is added to the series $\mathbf{X}_n$, this rv takes any position $X^{(i)}$ between the values x^{i-1} and x^{i+1} with probability $F_X(x^{i+1}) - F_X(x^{i-1})$. At that it can also takes the value in a small neighborhood of a point x^i with probability $f_X(x^i)dx^i$. Therefore, representing its joint pdf in terms of probabilities, we have

$$\begin{aligned} &f_{Ad[\mathbf{X}_n,X]}(\mathbf{x}_{n+1})d\mathbf{x}_{n+1} = \mathbf{P}\{X^{(1)} \in [x^1, x^1 + dx^1], \dots, X^{(n+1)} \in [x^{n+1}, x^{n+1} + dx^{n+1}]\} = \\ &= \sum_{1 \le i \le n+1} \mathbf{P}\{x^{i-1} < X \le x^{i+1}\}\mathbf{P}\{X \in [x^i, x^i + dx^i]\}\mathbf{P}\{X^{(j)} \in [x^j, x^j + dx^j],\ j \ne i\} = \\ &= \sum_{1 \le i \le n+1} (F_X(x^{i+1}) - F_X(x^{i-1})) f_X(x^i) f_{\mathbf{X}_n}(\mathbf{x}_{n+1} \setminus \{x^i\}) d\mathbf{x}_{n+1}, \end{aligned}$$

that proves the relation (35) and completes the proof of this lemma. ■

Appendix B. Algorithm

We use simulation to calculate the empirical estimates of the required steady-state indicators of the model. To describe the algorithm we use operators introduced in (16) and apply them to the values of the appropriate arrays,

$$\begin{aligned} Sh[\mathbf{v}_j, v] &\equiv \{v^i_{sh} = v^i_j - v,\ \ (i = \overline{1,j},\ \ v \le v^i_j)\} \\ Ad[\mathbf{v}_j, v] &\equiv \begin{cases} v^i_{ad}(v) = v^i_j, \text{ as } i < l, \\ v^l_{ad}(v) = v, \\ v^{i+1}_{ad}(v) = v^i_j \text{ as } \ i > l \ \ (i = \overline{1, j-1}), \end{cases} \end{aligned}$$

where $l = \max\{i : v^i_j < v\}$. Remind that the dimension of the array $\mathbf{y}_j$ is $n + m$.

Remark 5. Remind that due to the regenerative structure of the model it is enough to investigate the process behavior only along one (the first) regenerative period. Thus, the algorithm simulates K trajectories up to the first regeneration point. Note, that in the Algorithm letters t, τ have

another sense than in the main text, but their new sense is additionally explained in the Algorithm to avoid misunderstandings. Especially, the letter t_j is used as an estimator for the time spent by the process in state j along the trajectory of the process up to the first regeneration, $t_j = \hat{S}_j^{(k)}(R)$.

Algorithm.

Preparation: Initialize the following initial data: integers n, m; K is the number of model trajectories. Set the distributions $A(\cdot)$, $B(\cdot)$ of rv's A_i, B_i, along with the corresponding mean (μ_A, μ_B) and CoV (v_A, v_B).

Prepare the counters: $(\nu_0, \dots, \nu_{n+m})$ is the number of visits to system states; $(t_0, \dots t_{n+m})$ are dwell times in system states; k is the current number of trajectory; τ is the number of loss customers; and an array of length K Π is the time until the system returns to state 0 from state 1 for each kth trajectory.

Beginning. Put $j = 0$, $k = 1$, $\Pi^{(k)} = 0$. Calculate marks in the initial system state $x_0 = A_0$, $\mathbf{y}_0 = \{y_0^1 = \infty\}$.

Step 1. If $k < K$, go to Step 2, if no, go to Step 6.

Step 2. If $j = 0$, calculate $t_j := t_j + x_j$, $\nu_j := \nu_j + 1$, $j := j + 1$. Find $x_j = A_j$, $\mathbf{y}_j = Ad[B]$.
Go to Step 5.

Step 3. While $0 < j < n + m$, repeat:
it $x_j \le y_j^1$:
put $t_j := t_j + x_j$, $\nu_j := \nu_j + 1$, $\Pi^{(k)} := \Pi^{(k)} + x_j$, $j := j + 1$;
calculate $x_j = A_j$;
if $j \le n$, then $\mathbf{y}_j = Ad[Sh[\mathbf{y}_{j-1}, x_{j-1}], B]$
in another case $j > n$, then $\mathbf{y}_j = Sh[\mathbf{y}_{j-1}, x_{j-1}]$
if $x_j > y_j^1$:
put $t_j := t_j + y_j^1$, $\nu_j := \nu_j + 1$, $\Pi^{(k)} := \Pi^{(k)} + y_j^1$, $j := j - 1$;
calculate $x_j = Sh[x_{j-1}, y_{j-1}^1]$;
if $j \le n$, then $\mathbf{y}_j = Sh[\mathbf{y}_{j-1}, x_{j-1}]$
in another case $j > n$, then $\mathbf{y}_j = Ad[Sh[\mathbf{y}_{j-1}, x_{j-1}], B]$

If $j = 0$, then put $y_0^{(1)} = \infty$ and go to Step 2. If $j = n + m$, then go to Step 4. Otherwise, repeat Step 3, while the condition $0 < j < n + m$ is true.

Step 4. If $j = n + m$,
If $x_j \le y_j^1$:
put $\tau := \tau + 1$
calculate $\mathbf{y}_j = Sh[\mathbf{y}_{j-1}, x_{j-1}]$; $x_j = A_j$;
Repeat Step 4 from the beginning.
if $x_j > y_j^1$:
put $t_j := t_j + y_j^1$, $\nu_j := \nu_j + 1$, $\Pi^{(k)} := \Pi^{(k)} + y_j^1$, $j := j - 1$;
calculate $x_j = Sh[x_{j-1}, y_{j-1}^1]$, $\mathbf{y}_j = Ad[Sh[\mathbf{y}_{j-1}, y_{j-1}^1], B]$.
Go to Step 3.

Step 5. Collect statistics:

- Filling counters $\nu_0, \dots, \nu_{n+m}$,
- Filling the array Π by values $\Pi^{(k)}$,
- Filling counters $t_0, \dots t_{n+m}$.

Put $k := k + 1$ and go to Step 3.

Step 6. Processing statistics:

- Calculating the distribution of the number ν_j of visits to the states, $\hat{v}_j = \dfrac{\nu_j}{\sum_{j \le n+m} \nu_j}$,

- Calculating steady-state probabilities $\hat{\pi}_j = \dfrac{t_j}{\sum\limits_{0\le j\le n+m} t_j}$,
- Calculating cdf of the busy period $\hat{F}_{\Pi}(x) = \frac{1}{K}\sum\limits_{1\le k\le K} 1_{\{\Pi^{(k)}\le x\}}$;
- Calculating the mean busy period $\hat{\mu}_{\Pi} = \frac{1}{K}\sum\limits_{1\le k\le K} \Pi^{(k)}$;
- Calculating the loss probability $\hat{\pi}_{loss} = \hat{\pi}_{n+m}$ and loss rate $\hat{\lambda}_{loss} = \mu_A^{-1}\hat{\pi}_{n+m}$,
- Calculating mean queue length $\hat{\mu}_Q = \sum_{n<j\le n+m}(j-n)\hat{\pi}_j$,
- Calculating the mean number of customers in the system $\hat{\mu}_J = \sum_{1\le j\le n+m} j\hat{\pi}_j$,
- Calculating the mean waiting time $\hat{\mu}_W = \mu_A\hat{\mu}_J$ and the mean sojourn time $\hat{\mu}_V = \mu_A\hat{\mu}_Q$.

Stop.

ON PARTIAL STABILITY OF PREEMPTIVE PRIORITY RETRIAL MODEL

Ruslana Nekrasova

•

Institute of Applied Mathematical Research Karelian Research Centre, RAS,
Petrozavodsk State University, Russia
ruslana.nekrasova@mail.ru

Abstract

We consider a retrial model under constant retrial rate policy with two classes of customers characterized by different priorities. Preemptive priority arrivals, who meet the server busy by the other class customer, immediately start the processing, while interrupted customers lose the residual service times and join the end of the corresponding orbit queue. The system is fed by a superposition of two Poisson inputs, retrial times are exponential, service times are generally distributed and independent and iid in each class. We study the model in a partial stable state, when one orbit queue (independently of its class priority) is stochastically bounded, and other orbit infinitely grows in probability. We rely in preliminary results for a convenient two-class retrial model with no interruptions, where partial stability is equivalent to the transience of an associated Markov Chain (MC). Based on MC approach, we obtain transience conditions for embedded two-dimensional orbit size process and then verify partial stability behavior in transient zones by simulation.

Keywords: retrial model, constant retrial rate, preemptive priority, partial stability

1. Introduction

The paper deals with a single server two-class retrial model under constant retrial rate policy. Multi-class retrial systems have a wide sphere of applicability. For instance, such models can be successfully used in description of call centers [1, 2] or modern computer networks and protocols [3, 4]. Various types of retrial models are well presented in a corresponding literature. In this regard it is worth mention [5, 6, 7]. In particular [8, 9, 10] are dedicated to priority retrial queues.

We consider retrial system with two non-equivalent classes, characterized by different priorities. If the high (preemptive) priority arrival meets the server busy by the low priority customer, the service interruption occurs. In such a moment the low priority customer joins the corresponding orbit queue, while the high priority one starts processing. The system is fed by a superposition of two Poisson inputs, service times are generally distributed and independent and identically distributed (iid) in each class. The system contains two orbits under constant retrial rate policy, orbit attempts are exponential and retrial rates are defined by the corresponding class. Stability conditions for such a model have been obtained in a recent paper [11], where authors used an embedded MC approach. Previously such a method was successfully applied to the stability analysis of a two-class retrial model [12] and also to a retrial model with unreliable server [13]. The main purpose of the present paper is to study the so-called partially stable state, which arises when one orbit is stochastically bounded and other orbit infinitely grows in probability. In such a case the model is not totally stable, but we are able to use steady-state techniques to analyze

the stable orbit. Note that in [14] regenerative method of confidence estimation was applied to a partially stable retrial model with two equivalent priority classes.

Partial stability conditions have been obtained in [12] for a conventional two-class retrial model. Based on the MC approach, the authors in [12] had shown that partial stability regimes are equivalent to the transient states of a two-dimensional MC associated with the orbit queue size components. Thus our goal is to obtain in an explicit form the transience conditions for the embedded MC in the model under consideration, then to verify partial stability behavior in transient mode by simulation.

The paper is organized as follows. Section 2 is devoted to the detailed description of the model. In Section 3 we present MC approach to stability analysis of two-class retrial model. Such a method allows to obtain the ergodicity/transience conditions of a MC associated with the orbit size components. Section 4 contains the new basic analytical result for transience conditions. In Section 5 we present simulation results for the model in transient and non-ergodic modes. The obtained results show that transient regimes correspond to the partial stability states. Section 6 concludes the paper.

2. Model Description

We construct two-class retrial model under constant retrial rate policy. The system is fed by a superposition of two Poisson inputs with corresponding rates λ_k, $k = 1, 2$. Define by τ_k generic interarrival time for the class-k customers. Note $\mathsf{E}\tau_k = 1/\lambda_k$, $k = 1, 2$.

If class-k new arrival is unable to receive service because of busy server, it joins to the corresponding infinite capacity orbit. Constant retrial rate policy implies that the orbit customers obey to FIFO discipline: the first in the orbit queue customer makes retrial attempts to capture the server. The retrial times are exponentially distributed and class-dependent. Define class-k retrial rate by σ_k.

Note that the model has non-equivalent classes. The first class of customers has so-called preemptive or high priority. Namely if the high priority arrival meets the other class customer on service, it captures the server immediately, while the low priority costumer joins the corresponding (class-2) orbit queue and an interruption occurs. Note that in case of successful retrial attempt an interrupted customer gets new independent service time, while its previous residual service time is lost. From this point of view the model under consideration has no work-conserving property. The low priority arrivals and the high priority arrivals who meet the server busy by the same class customer obey to the constant retrial rate policy. Note that class-1 retrials (orbit customers) loss its priority and are unable to interrupt the second class service.

Moreover we consider general and iid class-dependent service times. As the model is not Markovian. Stability analysis becomes much more complicated. Let S_k define class-k generic service time with the corresponding distribution function (d.f.) F_k. Thus

$$\mathsf{p}_0 := (S_2 \le \tau_1) = \int_0^\infty e^{-\lambda_1 x} dF_2(x) \tag{1}$$

is the probability that the second class customer completed its service with no interruption. Note that p_0 is the Laplace-Stieltjes transform of the service time S_2. Let define class-k load coefficient $\rho_k = \lambda_k \mathsf{E}S_k$, $k = 1, 2$.

Next we define by $Y^{(k)}(t)$, $k = 1, 2$, and by $N(t) \in \{0, 1\}$, class-k orbit queue size and the number of customers on the server at instant $t \ge 0$, respectively. Now we construct a basic three-dimensional process

$$\mathbf{X} = \Big(Y^{(1)}(t),\, Y^{(2)}(t),\, N(t),\, t \ge 0\Big), \tag{2}$$

describing the dynamics of the system.

In this paper we are interested in so-called partial stability, which actually means that one orbit is tight, while other orbit infinitely grows in probability. Note the process $Y^{(k)}(t)$ is called

tight [15], if for any finite constant $C \geq 0$ exists $\delta > 0$

$$\inf_t \mathsf{P}\big(Y^{(k)}(t) \leq C\big) \geq 1 - \delta.$$

3. Markov chain approach

Stability analysis based on the properties of the process $\mathbf{X}$ seems rather intuitive while the search of partial stability conditions is a challenging problem. To solve this problem, we use a method associated with the embedded MC. In this regard we consider two-dimension discrete time process

$$\mathbf{Y} = \Big(Y_n^{(1)}, Y_n^{(2)}, \quad n \geq 1\Big), \tag{3}$$

where $Y_n^{(k)}$ defines the number of customers at class-k orbit just after the n-th departure instant (after service competition). As input stream is Poisson and retrial times are exponentially distributed, it is easy to show that the process $\mathbf{Y}$ is a homogeneous irreducible aperiodic MC.

It has been shown in paper [12] that, for a convenient two-class retrial system, stability of a basic process $\mathbf{X}$ is equivalent to the ergodicity of embedded MC $\mathbf{Y}$. Our goal is to extend the results from [12] for preemptive priority model. Note that in further analysis we define stability as ergodicity of MC $\mathbf{Y}$.

The method developed in [16] allows to obtain ergodicity and transience conditions for two-dimensional MC. Such an approach actually represents two-dimensional analogue of the negative drift condition. Namely, from Theorem 3.3.1 , [16] ergodicity and transience conditions for two-dimensional MC $\mathbf{Y}_n$ are expressed via appropriate combinations of the following four conditions or their opposites.

1. The first orbit negative drift condition:

$$\mathsf{E}\Big[Y_{n+1}^{(1)} - Y_n^{(1)}\Big|Y_n^{(1)} > 0, Y_n^{(2)} > 0\Big] < 0. \tag{4}$$

2. The second orbit negative drift condition:

$$\mathsf{E}\Big[Y_{n+1}^{(2)} - Y_n^{(2)}\Big|Y_n^{(1)} > 0, Y_n^{(2)} > 0\Big] < 0. \tag{5}$$

3. The first "joint" condition:

$$\begin{aligned}\mathsf{E}\Big[Y_{n+1}^{(1)} - Y_n^{(1)}\Big|Y_n^{(1)} = 0, Y_n^{(2)} > 0\Big]\mathsf{E}\Big[Y_{n+1}^{(2)} - Y_n^{(2)}\Big|Y_n^{(1)} > 0, Y_n^{(2)} > 0\Big] - \\ \mathsf{E}\Big[Y_{n+1}^{(1)} - Y_n^{(1)}\Big|Y_n^{(1)} > 0, Y_n^{(2)} > 0\Big]\mathsf{E}\Big[Y_{n+1}^{(2)} - Y_n^{(2)}\Big|Y_n^{(1)} = 0, Y_n^{(2)} > 0\Big] \quad \geq \quad 0.\end{aligned} \tag{6}$$

4. The second "joint" condition:

$$\begin{aligned}\mathsf{E}\Big[Y_{n+1}^{(1)} - Y_n^{(1)}\Big|Y_n^{(1)} > 0, Y_n^{(2)} > 0\Big]\mathsf{E}\Big[Y_{n+1}^{(2)} - Y_n^{(2)}\Big|Y_n^{(1)} > 0, Y_n^{(2)} = 0\Big] - \\ \mathsf{E}\Big[Y_{n+1}^{(1)} - Y_n^{(1)}\Big|Y_n^{(1)} > 0, Y_n^{(2)} = 0\Big]\mathsf{E}\Big[Y_{n+1}^{(2)} - Y_n^{(2)}\Big|Y_n^{(1)} > 0, Y_n^{(2)} = 0\Big] \quad \geq \quad 0.\end{aligned} \tag{7}$$

Note that in [11] by MC method were obtained stability conditions for a model under consideration. Authors in [12] applied MC approach for analysis of a retrial model with no interruptions and showed that transience state of two-dimensional MC associated with orbit queue components corresponds to partial stability regime. It is worth mention that results from [16] are applicable under some extra conditions, which automatically hold for the models with Poisson input, see Appendix A in [12] for details.

4. Transience conditions

One of our main goals in this paper is to obtain transience conditions for MC **Y** and then to verify partial stability behavior by simulation in transient mode. Theorem 3.3.1 from [16] defines two transience cases. Relying on [16] we consider transience - I, if

- the first orbit negative drift condition (4) holds;
- the second orbit negative drift condition (5) is violated;
- the first joint condition (6) holds true.

Symmetrically we define transience - II, if

- the first orbit negative drift condition (4) is violated;
- the second orbit negative drift condition (5) holds;
- the second joint condition (7) holds true.

Taking into account the results from [12], we can expect that transience - I mode corresponds to the first class partial stability (the first orbit is tight, the second orbit infinitely grows), while transience - II mode corresponds to the second class partial stability.

To obtain transience conditions in an explicit form we first deduce expressions for the mentioned above conditional mean drifts.

4.1. Mean drifts

We start from analysis of the expression

$$\mathsf{E}\Big[Y_{n+1}^{(1)} - Y_n^{(1)}\Big| Y_n^{(1)} > 0, Y_n^{(2)} > 0\Big],$$

which represents the first orbit mean drift on one step of MC under the condition that in the previous step both orbits were not empty.

Recall that input stream is Poisson and consider the probability that $j \geq 0$ customers join the class-k orbit on a random time interval distributed as S_i as follows:

$$p_k^{S_i}(j) := \int_{x=0}^{\infty} e^{-\lambda_k x} \frac{(\lambda_k x)^j}{j!} dF_i(x), \quad k = 1, 2;\ i = 1, 2;\ j \geq 0. \tag{8}$$

Next we analyze the first orbit mean increment for the one step of MC **Y** under the condition that on previous step both orbits were not empty. In case the first class customer is on service, j customers join the orbit with probability (w.p.) $p_1^{S_1}(j)$. From the other hand, if the second class customer is on service, the interruption may occur w. p. $(1 - \mathsf{p}_0)$. In this case the arrival immediately starts its service and does not affect the orbit, while the further first class arrivals saturate the orbit on time interval distributed as S_1. In case of no interruption which arise w. p. $\mathsf{p}_0 = \mathsf{P}(S_2 \leq \tau_1)$ the first orbit drift is equal to zero (the low priority customer finishes its service before the high priority arrival joins the system). Thus

$$\mathsf{E}\Big[Y_{n+1}^{(1)} - Y_n^{(1)}\Big| Y_n^{(1)} > 0, Y_n^{(2)} > 0\Big] = \frac{\lambda_1}{\lambda_1 + \sigma_1 + \lambda_2 + \sigma_2} \sum_{j=0}^{\infty} j p_1^{S_1}(j)$$
$$+ \frac{\sigma_1}{\lambda_1 + \sigma_1 + \lambda_2 + \sigma_2} \sum_{j=0}^{\infty} (j-1) p_1^{S_1}(j) + (1 - \mathsf{p}_0) \frac{\lambda_2 + \sigma_2}{\lambda_1 + \sigma_1 + \lambda_2 + \sigma_2} \sum_{j=0}^{\infty} j p_1^{S_1}(j).$$

After some computation efforts (see [11] for details) we obtain

$$\mathsf{E}\Big[Y_{n+1}^{(1)} - Y_n^{(1)}\Big| Y_n^{(1)} > 0, Y_n^{(2)} > 0\Big] = \frac{\lambda_1\rho_1 - \sigma_1(1-\rho_1) + (\sigma_2 + \lambda_2)(1 - \mathsf{p}_0)\rho_1}{\lambda_1 + \sigma_1 + \lambda_2 + \sigma_2}. \tag{9}$$

Next we discuss the second orbit mean drift under condition that both orbits are not empty. If the first class customer is on service (or if no interruption occurs), class-2 customers join the corresponding orbit on interval distributed as S_1 (or S_2). In case of interruption, arrivals join the orbit on interval distributed as $\tau_1 + S_1$. Namely if the second class customer occupies the server and interruption occurs, w.p. $(1 - \mathsf{p}_0)$ the second class mean orbit drift is defined by

$$\frac{\lambda_2}{\lambda_1+\sigma_1+\lambda_2+\sigma_2}\sum_{j=0}^{\infty}(j+1)\sum_{k=0}^{j}p_2^{\tau_1}(k)p_2^{S_1}(j-k)+\frac{\sigma_2}{\lambda_1+\sigma_1+\lambda_2+\sigma_2}\sum_{j=0}^{\infty}j\sum_{k=0}^{j}p_2^{\tau_1}(k)p_2^{S_1}(j-k)\Big).$$

The total expression has the following view

$$\mathsf{E}\Big[Y_{n+1}^{(2)}-Y_n^{(2)}\Big|Y_n^{(1)}>0,Y_n^{(2)}>0\Big]=\Bigg[(\lambda_1+\sigma_1)\lambda_2\mathsf{E}S_1+\mathsf{p}_0\Big(\lambda_2\rho_2-\sigma_2(1-\rho_2)\Big)$$

$$+(1-\mathsf{p}_0)\lambda_2\Big(\lambda_2\mathsf{E}S_1+\frac{\lambda_2}{\lambda_1}+\lambda_2+\sigma_2\mathsf{E}S_1+\frac{\sigma_2}{\lambda_1}\Big)\Bigg]\Big/(\lambda_1+\sigma_1+\lambda_2+\sigma_2). \tag{10}$$

Joint conditions (6) and (7) contain mean orbit drifts for the cases when one of the orbits is empty on previous step. We can easily obtain such expressions from (9) and (10), setting $\sigma_1 = 0$ ($\sigma_2 = 0$), if $Y_k^{(1)} = 0$, $Y_k^{(2)} > 0$ ($Y_k^{(1)} > 0$, $Y_k^{(2)} = 0$).

4.2. Transience zones

To simplify further calculations we define some auxiliary values

$$\frac{1}{A_1} = (1-\mathsf{p}_0)\frac{\rho_1}{1-\rho_1}, \tag{11}$$

$$B_1 = \big(\lambda_1+\lambda_2(1-\mathsf{p}_0)\big)\frac{\rho_1}{1-\rho_1}, \tag{12}$$

$$\frac{1}{A_2} = \big(-\mathsf{p}_0(1-\rho_2)+(1-\mathsf{p}_0)(\lambda_2\mathsf{E}S_1+\frac{\lambda_2}{\lambda_1})\big)/\lambda_2\mathsf{E}S_1, \tag{13}$$

$$B_2 = \big(\rho_1+\mathsf{p}_0\rho_2+(1-\mathsf{p}_0)(\lambda_2\mathsf{E}S_1+\frac{\lambda_2}{\lambda_1}+1)\big)/\mathsf{E}S_1. \tag{14}$$

Note that $B_2 > 0$, the sign of A_2 may vary and for $\rho_1 < 1$ we have A_1, $B_1 > 0$. Thus we obtain

$$\mathsf{E}\Big[Y_{n+1}^{(1)}-Y_n^{(1)}\Big|Y_n^{(1)}>0,Y_n^{(2)}>0\Big] = \frac{1-\rho_1}{\lambda_1+\sigma_1+\lambda_2+\sigma_2}\big(-\sigma_1+\frac{1}{A_1}\sigma_2+B_1\big), \tag{15}$$

$$\mathsf{E}\Big[Y_{n+1}^{(2)}-Y_n^{(2)}\Big|Y_n^{(1)}>0,Y_n^{(2)}>0\Big] = \frac{\lambda_2\mathsf{E}S_1}{\lambda_1+\sigma_1+\lambda_2+\sigma_2}\big(\sigma_1+\frac{1}{A_2}\sigma_2+B_2\big) \tag{16}$$

and then formulate joint conditions (6) and (7) as follows.
The first joint condition:

$$(1-\rho_1)\Bigg(\Big(\frac{1}{A_1}\sigma_2+B_1\Big)\Big(\sigma_1+\frac{1}{A_2}\sigma_2+B_2\Big)-\Big(-\sigma_1+\frac{1}{A_1}\sigma_2+B_1\Big)\Big(\frac{1}{A_2}\sigma_2+B_2\Big)\Bigg) \geq 0$$

$$(1-\rho_1)\Bigg(\sigma_2\Big(\frac{1}{A_1}+\frac{1}{A_2}\Big)+B_1+B_2\Bigg) \geq 0. \tag{17}$$

The second joint condition:

$$\frac{(1-\rho_1)}{A_1A_2}\Bigg(\sigma_1\Big(A_1+A_2\Big)+B_2A_2-B_1A_1\Bigg) \geq 0. \tag{18}$$

Next we present the basic analytical result.

Theorem 1. Consider two-class retrial model with constant retrial rate and preemptive priority of the first class customers.

1. If $\rho_1 < 1$, then the first orbit negative drift condition

$$\lambda_1\rho_1 - \sigma_1(1-\rho_1) + (\sigma_2+\lambda_2)\big(1-\mathsf{p}_0\big)\rho_1 < 0$$

defines transience-I regime, if one of the following alternative cases holds true

a).

$$\begin{cases} \lambda_2(1-\mathsf{p}_0) < \lambda_1\mathsf{p}_0(1-\rho_1)(1-\rho_2) \\ \sigma_2 \le \lambda_2\frac{\mathsf{p}_0\lambda_1\big(1-(1-\rho_1)(1-\rho_2)\big)+(\lambda_1+\lambda_2)(1-\mathsf{p}_0)}{\lambda_1\mathsf{p}_0\big(1-(1-\rho_1)(1-\rho_2)\big)-\lambda_2(1-\mathsf{p}_0)} \end{cases}; \tag{19}$$

b).

$$\lambda_1\mathsf{p}_0(1-\rho_1)(1-\rho_2) \le \lambda_2(1-\mathsf{p}_0) < \lambda_1\mathsf{p}_0(1-\rho_2) - \lambda_2(1-\mathsf{p}_0)\rho_1; \tag{20}$$

c).

$$\lambda_2(1-\mathsf{p}_0)(\rho_1+1) \ge \lambda_1\mathsf{p}_0(1-\rho_2). \tag{21}$$

2. If $\lambda_2(1-\mathsf{p}_0)(\rho_1+1) < \lambda_1\mathsf{p}_0(1-\rho_2)$, then the second orbit negative drift condition

$$(\lambda_1+\sigma_1)\lambda_2\mathsf{E}S_1 + \mathsf{p}_0\Big(\lambda_2\rho_2 - \sigma_2(1-\rho_2)\Big) + (1-\mathsf{p}_0)\lambda_2\Big(\lambda_2\mathsf{E}S_1 + \frac{\lambda_2}{\lambda_1} + \lambda_2 + \sigma_2\mathsf{E}S_1 + \frac{\sigma_2}{\lambda_1}\Big) < 0$$

defines transience-II regime, if one of the following alternative cases holds true

a).

$$\begin{cases} \lambda_2(1-\mathsf{p}_0) < \lambda_1\mathsf{p}_0(1-\rho_1)(1-\rho_2) \\ \sigma_1 \le \lambda_1\frac{\lambda_1\mathsf{p}_0\rho_1(1-\rho_2)}{\lambda_1\mathsf{p}_0(1-\rho_1)(1-\rho_2)-\lambda_2(1-\mathsf{p}_0)} \end{cases}; \tag{22}$$

b).

$$\begin{cases} \rho_1 < 1 \\ \lambda_2(1-\mathsf{p}_0) \ge \lambda_1\mathsf{p}_0(1-\rho_1)(1-\rho_2) \end{cases}; \tag{23}$$

c).

$$\rho_1 \ge 1. \tag{24}$$

Proof. First our goal is to analyze the sets of transience conditions from [16] for various signs of parameters A_k, B_k. Let discuss all the possible cases separately.

Case 1: $\rho_1 < 1$.

The current case automatically implies A_1, $B_1 > 0$, see (11), (12). The first orbit negative drift condition is obtained as follows

$$\sigma_2 < A_1\sigma_1 - A_1B_1 =: f_1(\sigma_1), \tag{25}$$

where f_1 is a linear increasing function with an argument σ_1 and coefficients defined via λ_k, $\mathsf{E}S_k$, $k = 1, 2$ and p_0. Next we analyze various signs of A_2.

Case 1.1: $\rho_1 < 1$, $A_2 < 0$.

Now we express the second orbit negative drift condition as

$$\sigma_2 > -A_2\sigma_1 - A_2B_2 =: f_2(\sigma_1), \tag{26}$$

where f_2 is a linear increasing function.

Note that condition $A_2 < 0$ or equivalently $1/A_2 < 0$ is presented as

$$\lambda_2(1-\mathsf{p}_0)(\rho_1+1) < \lambda_1\mathsf{p}_0(1-\rho_2). \tag{27}$$

Case 1.1.1: $\rho_1 < 1$, $A_2 < 0$, $A_1 + A_2 > 0$.
We define the first and the second joint conditions respectively as

$$\sigma_2 \leq -A_1A_2\frac{(B_1+B_2)}{A_1+A_2} := \sigma_2^*, \tag{28}$$

$$\sigma_1 \leq \frac{A_1B_1 - A_2B_2}{A_1+A_2} := \sigma_1^*. \tag{29}$$

Note that in this case σ_1^*, $\sigma_2^* > 0$. Moreover for any arbitrary signs of A_k, B_k we can show that $f_1(\sigma_1^*) = f_2(\sigma_1^*) = \sigma_2^*$, see [11] for detailed calculations. Transience-I zone in the current case corresponds to the following set of conditions:

$$\sigma_2 < f_1(\sigma_1),\ \sigma_2 \leq f_2(\sigma_1),\ \sigma_2 \leq \sigma_2^*,$$

while transience-II is defined by:

$$\sigma_2 \geq f_1(\sigma_1),\ \sigma_2 > f_2(\sigma_1),\ \sigma_1 \leq \sigma_1^*.$$

The relation $A_1 > -A_2$ implies $f_1(\sigma_1) > f_2(\sigma_1)$ for $\sigma_1 > \sigma_1^*$, $\sigma_2 > \sigma_2^*$. Transience zones for the particular case of uniform service times distributed on corresponding intervals $[a_k, b_k]$, $k = 1, 2$ (define $S_k \sim \mathsf{U}[a_k, b_k]$) are presented on Fig 1, the left picture. Namely we set

$$\lambda_1 = 1.5,\ \lambda_2 = 0.8,\ S_1 \sim U[0.1, 0.7],\ S_2 \sim U[0.05, 0.45] \tag{30}$$

and obtain

$$\mathsf{p}_0 \approx 0.6976,\ A_1 = 2.2,\ A_2 = -1.1,\ B_1 = 2.6,\ B_2 = 3.2,\ \sigma_1^* = 8.1,\ \sigma_2^* = 12.1. \tag{31}$$

Note that in [11] was obtained that under conditions $\rho_1 < 1$, $A_2 < 0$, $A_1 + A_2 > 0$ the zone $\sigma_1 > \sigma_1^*$, $\sigma_2 > \sigma_2^*$ defines stability region for the model under consideration.

Case 1.1.1: $\rho_1 < 1$, $A_2 < 0$, $A_1 + A_2 > 0$. Case 1.1.2: $\rho_1 < 1$, $A_2 < 0$, $A_1 + A_2 < 0$.

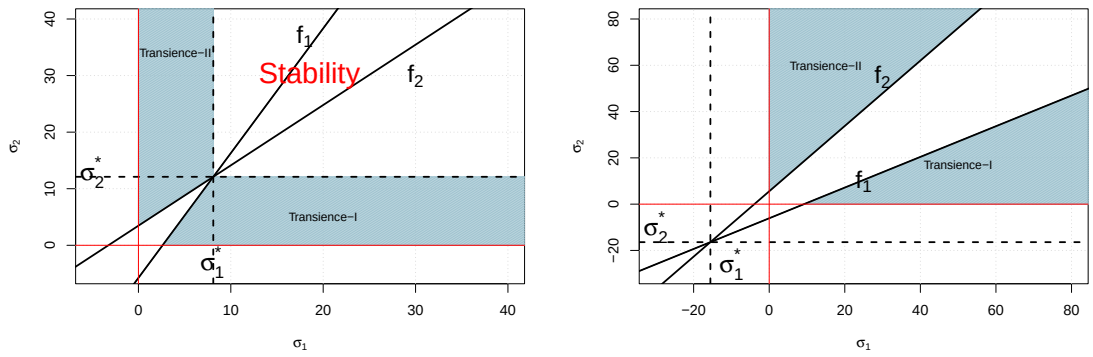


Figure 1: *Transience zones.*

Moreover the left picture from Figure 1 illustrates that in both transience modes non-negative drift conditions are redundant.

The condition $A_1 + A_2 > 0$ is equivalent to

$$\lambda_2\mathsf{E}S_1(1-\mathsf{p}_0)\rho_1 < (1-\rho_1)\Big(\mathsf{p}_0(1-\rho_2) - (1-\mathsf{p}_0)(\lambda_2\mathsf{E}S_1 + \lambda_2/\lambda_1\Big)$$
$$\lambda_2(1-\mathsf{p}_0) < \lambda_1\mathsf{p}_0(1-\rho_1)(1-\rho_2). \tag{32}$$

From (27) and (32) we have

$$\lambda_2(1-\mathsf{p}_0) < \lambda_1\mathsf{p}_0(1-\rho_2) - \max\Big(\lambda_2(1-\mathsf{p}_0),\ \lambda_1\mathsf{p}_0(1-\rho_2)\Big)\rho_1. \tag{33}$$

The relation $\lambda_2(1-\mathsf{p}_0) \geq \lambda_1\mathsf{p}_0(1-\rho_2)$ violates the condition (27). Thus (33) turns to

$$\lambda_2(1-\mathsf{p}_0) < \lambda_1\mathsf{p}_0(1-\rho_1)(1-\rho_2). \tag{34}$$

Hence the pair of conditions $A_2 < 0$ and $A_1 + A_2 > 0$ is implied by (34).

Case 1.1.2: $\rho_1 < 1$, $A_2 < 0$, $A_1 + A_2 < 0$.
For this case both joint conditions change the signs and turn to $\sigma_k \geq \sigma_k^*$, while $f_2(\sigma_1)$ dominates $f_1(\sigma_1)$ for $\sigma_1 > \sigma_1^*$, $\sigma_2 > \sigma_2^*$. Moreover $\sigma_k < 0$, $k = 1, 2$, see expressions in (28), (29).

Transience-I corresponds to the set of conditions

$$\sigma_2 < f_1(\sigma_1),\ \sigma_2 \leq f_2(\sigma_1),\ \sigma_2 \geq \sigma_2^*,$$

and transience-II is defined by :

$$\sigma_2 \geq f_1(\sigma_1),\ \sigma_2 < f_2(\sigma_1),\ \sigma_1 \geq \sigma_1^*.$$

Transience zones presented on Figure 1 (the right picture) illustrate that k-type transience is totally defined by the corresponding negative drift condition. Note that for this case we set

$$\lambda_1 = 2,\ \lambda_2 = 0.8,\ S_1 \sim U[0.1,\ 0.7],\ S_2 \sim U[0.05,\ 0.45] \tag{35}$$

and obtain

$$\mathsf{p}_0 \approx 0.6228,\ A_1 = 0.7,\ A_2 = -1.4,\ B_1 = 8.7,\ B_2 = 4.0,\ \sigma_1^* = -15.6,\ \sigma_2^* = -16.4. \tag{36}$$

Case 1.1.3: $\rho_1 < 1$, $A_2 < 0$, $A_1 + A_2 = 0$.
In this case both joint conditions (17) and (18) hold true because the left hand sides do not depend on retrial rates and are strictly positive, moreover note that f_1 and f_2 are parallel linear functions. The relation

$$-A_1B_1 < 0 < -A_2B_2$$

implies $f_1(\sigma_1) < f_2(\sigma_1)$ for any σ_1. Thus similar to the the case 1.1.2. we can show that k-type transience is defined by the corresponding negative drift condition.

Next we analyze the condition $A_1 + A_2 \leq 0$ which corresponds to cases 1.1.2 and 1.1.3. Relying on (32) we obtain

$$\lambda_2(1-\mathsf{p}_0)\rho_1 \geq \lambda_1\mathsf{p}_0(1-\rho_2) - \lambda_1\mathsf{p}_0(1-\rho_2). \tag{37}$$

Combining with (27) we have

$$\lambda_1\mathsf{p}_0(1-\rho_2) - \lambda_1\mathsf{p}_0(1-\rho_2)\rho_1 \leq \lambda_2(1-\mathsf{p}_0) < \lambda_1\mathsf{p}_0(1-\rho_2) - \lambda_2(1-\mathsf{p}_0)\rho_1. \tag{38}$$

Note that by (27) $\lambda_1\mathsf{p}_0(1-\rho_2)\rho_1 > \lambda_2(1-\mathsf{p}_0)\rho_1$, thus the solution set for (38) is not empty. Recall that (38) is equivalent to $A_2 < 0$, $A_1 + A_2 \leq 0$.

Case 1.2: $\rho_1 < 1$, $A_2 > 0$.
The second orbit negative drift condition transforms to $\sigma_2 < f_2(\sigma_1)$, where f_2 is a linear decreasing function. The first joint condition transforms to $\sigma_2 \geq \sigma_2^*$, and always holds true, as $\sigma_2^* < 0$. The second joint condition transforms to $\sigma_1 \geq \sigma_1^*$, the sign of σ_1^* may vary. Transience-I is described by the conditions

$$\sigma_2 < f_1(\sigma_1),\ \sigma_2 \geq f_2(\sigma_1),\ \sigma_2 \geq \sigma_2^*,$$

while transience-II is defined by the following set

$$\sigma_2 \geq f_1(\sigma_1),\ \sigma_2 < f_2(\sigma_1),\ \sigma_1 \geq \sigma_1^*.$$

In the current case the condition $f_2(\sigma_1) > 0$ holds only for negative values of σ_1. Hence the second orbit negative drift condition is violated and transience-II zone is empty. From the other hand, transience-I is defined by the first orbit negative drift condition, see Figure 2, where

$$\lambda_1 = 2,\ \lambda_2 = 0.8,\ S_1 \sim U[0.1,\ 0.7],\ S_2 \sim U[0.05,\ 0.95] \tag{39}$$

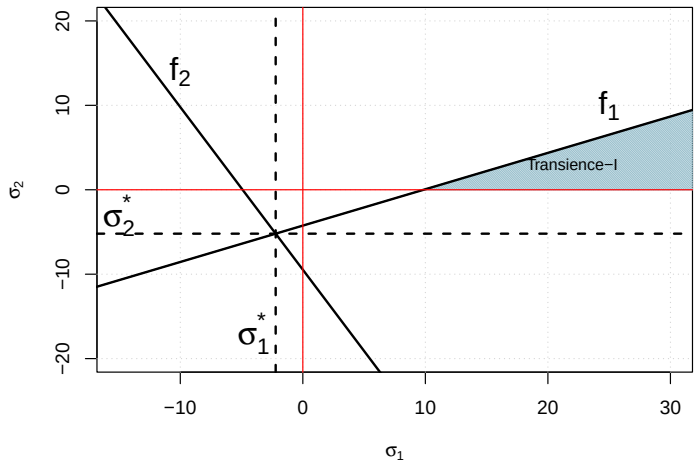


Figure 2: *Transience zones,* $\rho_1 < 1$, $A_2 > 0$.

and

$$\mathsf{p}_0 \approx 0.4196,\ A_1 = 0.4,\ A_2 = 1.9,\ B_1 = 10.8,\ B_2 = 5.0,\ \sigma_1^* = -2.2,\ \sigma_2^* = -5.2. \tag{40}$$

Case 1.3: $\rho_1 < 1$, $1/A_2 = 0$.
The second orbit negative drift condition turns to $\sigma_1 + B_2 < 0$ and is violated. Thus transience-II zone is empty. Hence the second orbit non negative drift condition always holds true, moreover the first joint condition always holds true, see (17). Thus transience-I zone is defined by the first orbit negative drift condition.

Note that $1/A_2 \geq 0$ is equivalent to

$$\lambda_2(1 - \mathsf{p}_0)(\rho_1 + 1) \geq \lambda_1 \mathsf{p}_0(1 - \rho_2). \tag{41}$$

Case 2: $\rho_1 > 1$.
This case automatically implies $A_1 < 0$, $B_1 < 0$. Thus after multiplying (15) by $A_1/(1 - \rho_1) > 0$ we have the first orbit negative drift condition as $\sigma_2 < f_1(\sigma_1)$, where $f_1(\sigma_1) = A_1\sigma_1 - A_1B_1$ is a **decreasing** function. Moreover by definition we have $f_1 < 0$ for $\sigma_1 \geq 0$. Thus the first orbit negative drift condition is violated and transience-I zone is empty.

Case 2.1: $\rho_1 > 1$, $A_2 < 0$.
The second orbit negative drift condition is defined as $\sigma_2 > f_2(\sigma_1)$, where f_2 increases. The second joint condition evaluates to $\sigma_1 \geq \sigma_1^*$, while $\sigma_1^* < 0$. Thus negative drift condition defines transience mode for the second class orbit.

Case 2.2: $\rho_1 > 1$, $A_2 > 0$.
The second orbit negative drift condition is defined as $\sigma_2 < f_2(\sigma_1)$ and f_2 decreases. We have

$$f_2(\sigma_1) = -A_2\sigma_1 - A_2B_2.$$

Hence $f_2 < 0$ for $\sigma_1 \geq 0$, the second orbit negative drift condition is violated and transience-II zone is empty.

Case 2.3: $\rho_1 > 1$, $1/A_2 = 0$.
The second orbit negative drift condition turns to $\sigma_1 < -B_2$, thus transience-II zone is empty.

Case 3: $\rho_1 = 1$.
In this case $A_1 = 0$, while $B_1A_1 > 0$, see (11), (12). The second joint condition evaluates to

$$\sigma_1 \geq \frac{A_1B_1}{A_2} - B_2 \equiv \sigma_1^*. \tag{42}$$

If $A_2 < 0$ holds true, transience-II zone is defined by the second orbit negative drift condition. Next similar to the cases 2.2 and 2.3 we can show that for the case $1/A_2 \geq 0$ transience-II zone is empty.

Summing up, we group all the cases in Table 1.

Table 1: *Transience zones.*

	$\rho_1 < 1$			$\rho_1 \geq 1$	
	$1/A_2 < 0$		$1/A_2 \geq 0$	$1/A_2 < 0$	$1/A_2 \geq 0$
	$A_1 + A_2 > 0$	$A_1 + A_2 \leq 0$			
Transience-I	Neg. drift, $\sigma_2 \leq \sigma_2^*$.	Neg. drift.		-	
Transience-II	Neg. drift, $\sigma_1 \leq \sigma_1^*$.	Neg. drift.	-	Neg. drift.	-

Finally, relying on analysis from [11] we can show that

$$\sigma_1^* = \lambda_1 \frac{\lambda_1 \mathsf{p}_0 \rho_1 (1-\rho_2)}{\lambda_1 \mathsf{p}_0 (1-\rho_1)(1-\rho_2) - \lambda_2 (1-\mathsf{p}_0)}, \tag{43}$$

$$\sigma_2^* = \lambda_2 \frac{\mathsf{p}_0 \lambda_1 \big(1-(1-\rho_1)(1-\rho_2)\big) + (\lambda_1+\lambda_2)(1-\mathsf{p}_0)}{\lambda_1 \mathsf{p}_0 \big(1-(1-\rho_1)(1-\rho_2)\big) - \lambda_2 (1-\mathsf{p}_0)}. \tag{44}$$

Hence basing on the results in table 1 and (43), (44), we prove the theorem. ■

Remark. Now we formulate a few comments to the presented Theorem.

1. Condition $\rho_1 < 1$ is a transience-I necessary condition. This demand is rather natural, as in this case we can expect the first orbit partial stability.
2. Condition $\lambda_1(1-\mathsf{p}_0)(\rho_1+1) < \lambda_1 \mathsf{p}_0 (1-\rho_2)$ or equivalently
$$(1-\mathsf{p}_0)\lambda_2 \mathsf{E}\big(S_1+\tau_1\big) < \mathsf{p}_0(1-\rho_2) \tag{45}$$
is a transience-II necessary condition. Note that $\mathsf{E}\big(S_1+\tau_1\big)$ defines a mean time from the moment when the second class "interrupted to be" customer captures the server up to the moment when the server becomes idle. Thus $\lambda_2 \mathsf{E}\big(S_1+\tau_1\big)$ defines some kind of the probability that the server is busy in case of interruption: the second class customer is on service and then it is replaced by the preemptive priority arrival. From this point of view the left hand side of (45) defines the probability that interruption actually occurs. While the right hand side of (45) defines the probability that the interruption was not detected (the server is not occupied by the second class customer w.p. $(1-\rho_2)$). Note that (45) automatically implies a weaker necessary stability condition $\rho_2 < 1$.
3. From [11] we have that model is stable only if $\rho_1 < 1$, $A_2 < 0$, $A_1 + A_2 > 0$. In this case the pair of conditions $\sigma_1 < \sigma_1^*$, $\sigma_2 < \sigma_2^*$ is a stability criterion. Thus the second conditions in systems (19) and (22) provide instability for the the second and the first orbit, respectively.

5. Simulations

In this section we present simulation results in various cases from Theorem 1. Namely our goal is to show that the k-th type of transience mode defines the partial stability of the corresponding orbit.

We estimate mean orbit dynamics as follows. Consider $y_i^{(k)}(j)$ – the number of customers at k-class orbit just before the j-th arrival instant based on the i-th trajectory, $i = 1, \ldots, m$, $m = 100$. Namely we construct the following two-component sequence

$$\Big(\frac{1}{m}\sum_{i=1}^{m} y_i^{(1)}(j),\ \frac{1}{m}\sum_{i=1}^{m} y_i^{(2)}(j)\Big), \quad j = 1, \ldots n,\ n = 20\,000$$

and calculate the maximal values for the corresponding components

$$Max.Orb_k := \max_j \Big(\frac{1}{m} \sum_{i=1}^{m} y_i^{(k)}(j) \Big), \quad k = 1, 2.$$

In k-type transience mode we expect stable behavior of k-class orbit process and increasing of the other orbit process. Moreover a significant deviation of $Max.Orb_1$ and $Max.Orb_2$ also may indicate the partial stability of the orbit with the smaller maximal value.

Next we consider the model with uniform service time and star from the case 1.1.1 where $\rho_1 < 1$, $A_2 < 0$, $A_1 + A_2 > 0$. Input rates and mean service times coincide with the corresponding values (30). See table 2 for the explicit values for retrial rates used in simulation tests, test points (σ_1, σ_2) in transience and non-ergodic zones are presented on Figure 3, the left picture.

Case 1.1.1 Case 1.1.2

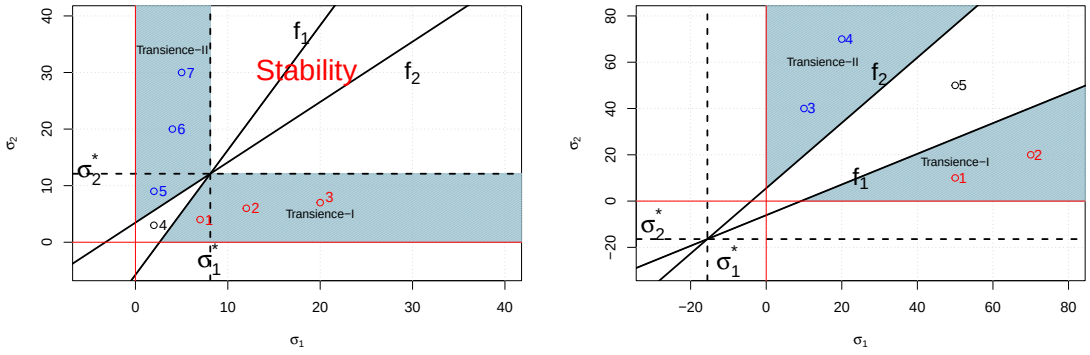


Figure 3: *Test points.*

Recall that in case 1.1.1 we have $\rho_1 = 0.6$, $A_1 = 2.1$, $A_2 = -1.1$. Tests 1–3 belong to transience-I zone, moreover moving from 1 to 3 we go closer to the stability border, and as a consequence the mean values for maximal orbit size decrease, see table 2. From the other hand, in these tests the maximal value of the first class orbit (Max. Orb_1) is not greater that 6.1, which may define the stable behavior, while the maximal value of the second class orbit (Max. Orb_2) strongly dominates Max. Orb_1. Such a phenomenon occurs in case of the first orbit partial stability. The results for transience-II zone (tests 5–7) are symmetrical, see table 2.

Orbit dynamics in transience modes (tests 2 and 6) are presented on Figure 4, the left picture. The grey lines correspond to the model in transience-I zone. The solid line illustrates the behavior of the first class orbit process. The dynamics is stable. From the other hand the dash line increases, thus we can expect that the mean value of the second class orbit size goes to infinity. Hence we observe the partly stable mode. The orbits behavior in transience-II zone also correspond to the assumption of partial stability. The black lines from the Figure 4 (the left picture) illustrate the test 6. The second class orbit (solid line) is stable, the first class orbit (dash line) infinitely grows.

The other interest is a model behavior in non-ergodic regime (test 4). See the mean orbits dynamics on Figure 4, the right picture, grey lines. Both orbits grow. Such a phenomenon corresponds to total instability and is predictable in non-ergodic zone. Note that the first class orbit process (solid line) strongly dominated the second class orbit process (dash line). Thus the first orbit is much more saturated. Such a result from the first site looks a bit confusing, as the first class customers have the high priority. The explanation is the following. Note that in the current example $\rho_1 = 0.6$, $\rho_2 = 0.2$, $\lambda_1 = 1.5$, $\lambda_2 = 0.8$ and $\sigma_1 = 2$, $\sigma_2 = 3$. Thus $\rho_1 > \rho_2$, hence the "load" from the first class is greater, the arrivals are more intensive. Moreover $\sigma_1 < \sigma_2$, consequently the second orbit is "faster". As a result we obtain the less number of the second class arrivals and more intensive attempts from the corresponding orbit, which lead to the smaller (in comparison with the other class) number of orbit customers.

Next we discuss the case 1.1.2 where $\rho_1 < 1$, $A_2 < 0$, $A_1 < -A_2$ and the basic parameters are defined in (35). Retrial rates are presented in table 2 and on Figure 3, the right picture. Tests 1–4 illustrate the orbits behavior in transience modes. The obtained results correspond to the assumption of partial stability and are additionally confirmed by the maximal values of mean orbit sizes, see table 2. The test 5 corresponds to non-ergodic zone, and we can expect the total instability. Note that obtained values $Max.Orb_1 = 1204$ and $Max.Orb_2 = 85$ are rather small in comparison with the other unstable cases. Thus to study such a test in details, we illustrate the mean orbit dynamics on Figure 4, the right picture, black lines. The first class orbit (dash line) strongly dominates the second class orbit (solid line). Note that $\rho_1 = 0.8$, $\rho_2 = 0.2$, $\lambda_1 = 2$, $\lambda_2 = 0.8$ and $\sigma_1 = 50$, $\sigma_2 = 50$. The first class arrivals are more intensive, which leads to the greater values of orbit size process. Note that the second class orbit process also grows. Thus the model is totally unstable.

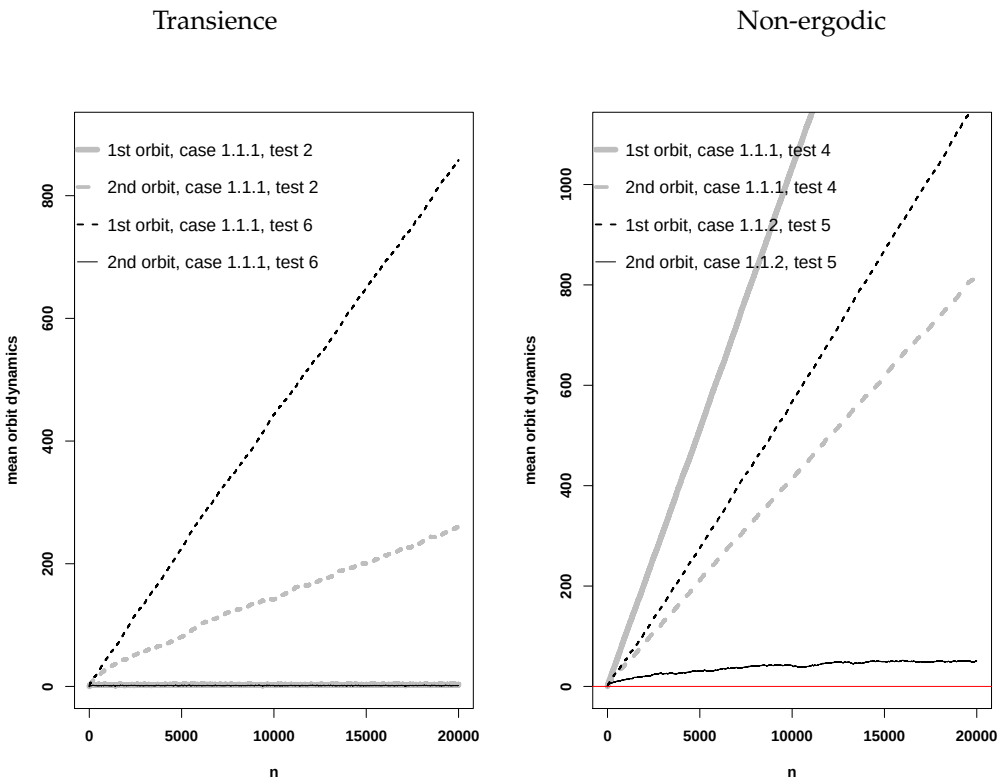


Figure 4: *Orbit dynamics.*

The results for the case 1.2, where $\rho_1 < 1, A_2 > 0$ and parameters are defined in (39), are presented in table 2. The assumption of partial stability is confirmed. Finally for the case 2.1, where $\rho_1 > 1, A_2 < 0$, we define

$$\lambda_1 = 1.5,\ \lambda_2 = 0.8,\ S_1 \sim U[0.5,\ 1.5],\ S_2 \sim U[0.05,\ 0.45]$$

and obtain

$$\mathsf{p}_0 \approx 0.6977,\ A_1 = -1.1,\ A_2 = -5.2,\ B_1 = -5.2,\ B_2 = 2.4, \sigma_1^* = -2.9,\ \sigma_2^* = -2.6.$$

The simulation results correspond with the assumption of patial stability. The model is totally unstable in non-ergodic zone.

Table 2: *Simulation results.*

Case	Test	σ_1	σ_2	Max. Orb_1	Max. Orb_2	Mode
1.1.1	1	7	4	6.1	319.9	Tr.-I
	2	12	6	3.8	260.8	Tr.-I
	3	20	7	2.6	65.1	Tr.-I
	4	2	3	2059.4	809.9	non-ergodic
	5	22	9	2284.2	2.8	Tr.-II
	6	4	20	857.9	2.1	Tr.-II
	7	5	30	369.9	1.9	Tr.-II
1.1.2	1	50	10	8.07	2591.2	Tr.-I
	2	70	20	9.54	1965.2	Tr.-I
	3	10	40	2596.3	2.7	Tr.-II
	4	20	70	1817.7	3.1	Tr.-II
	5	50	50	1204	89	non-ergodic
1.2	1	27	3	10.87	4652.5	Tr.-I
	2	12	6	9.54	1965.2	Tr.-I
	3	10	10	2705	2991	non-ergodic
2.1	1	7	60	6669	38	Tr.-II
	2	5	45	6921	20	Tr.-II
	3	20	45	6363	754	non-ergodic
	4	40	60	6106	2037	non-ergodic

6. Conclusion

The research is related to the phenomenon of partial stability in a two-class retrial model with preemptive priority of the first class customers and interruptions. Partially stable mode arises when queue size process in one of the orbits is stable, while in the other infinitely increases. Relying on MC approach we obtained transient conditions for MC, associated with orbit components at departure instants. Then basing on preliminary results for a convenient two-class retrial model with no interruptions, where transience defines partial stability, we verified by simulation partially stable state in transient zones. The simulation results had shown that under obtained transient conditions the model illustrates partially stable behavior.

Summing up, we can expect that transient zone implies the stability for one of the classes. Moreover relying on the assumption of partial stability in transient state, we can manage the model parameters to provide the service completion at finite time for the customers from the selected (even low priority) class, while the other (high priority) class customers sojourn time at corresponding orbit infinitely grows, and the whole model is unstable.

References

[1] Tuan Phung-Duc et al. “Retrial queues with balanced call blending: analysis of single-server and multiserver model”. In: *Annals of Operations Research* 239.2 (Apr. 2014), pp. 429–449. ISSN: 1572-9338. DOI: 10.1007/s10479-014-1598-2.

[2] Evsey Morozov et al. “Performance analysis and stability of multiclass orbit queue with constant retrial rates and balking”. In: *Performance Evaluation* 134 (Oct. 2019), p. 102005. ISSN: 0166-5316. DOI: 10.1016/j.peva.2019.102005.

[3] Konstantin Avrachenkov, Philippe Nain, and Uri Yechiali. “A retrial system with two input streams and two orbit queues”. In: *Queueing Systems* 77.1 (Aug. 2013), pp. 1–31. ISSN: 1572-9443. DOI: 10.1007/s11134-013-9372-8.

[4] Evsey Morozov and Tuan Phung-Duc. "Regenerative Analysis of Two-Way Communication Orbit-Queue with General Service Time". In: *Queueing Theory and Network Applications*. Springer International Publishing, 2018, pp. 22–32. ISBN: 9783319937366. DOI: 10.1007/978-3-319-93736-6_2.

[5] J. Falin and James G. C. Templeton. *Retrial Queues*. en. London: Chapman and Hall/CRC, 1997.

[6] Jesús R. Artalejo and Antonio Gómez-Corral. *Retrial Queueing Systems*. Springer Berlin Heidelberg, 2008. ISBN: 9783540787259. DOI: 10.1007/978-3-540-78725-9.

[7] Tuan Phung-Duc. *Retrial Queueing Models: A Survey on Theory and Applications*. 2019. DOI: 10.48550/ARXIV.1906.09560.

[8] Yi Peng. "On the discrete-time Geo/G/1 retrial queueing system with preemptive resume and Bernoulli feedback". In: *OPSEARCH* 53.1 (July 2015), pp. 116–130. ISSN: 0975-0320. DOI: 10.1007/s12597-015-0218-5.

[9] Shan Gao. "A preemptive priority retrial queue with two classes of customers and general retrial times". In: *Operational Research* 15.2 (Apr. 2015), pp. 233–251. ISSN: 1866-1505. DOI: 10.1007/s12351-015-0175-z.

[10] Sherif I. Ammar and Pakkirisamy Rajadurai. "Performance Analysis of Preemptive Priority Retrial Queueing System with Disaster under Working Breakdown Services". In: *Symmetry* 11.3 (Mar. 2019), p. 419. ISSN: 2073-8994. DOI: 10.3390/sym11030419.

[11] Ruslana Nekrasova. "Stability Conditions of Two-Class Preemptive Priority Retrial System with Constant Retrial Rate". In: *Distributed Computer and Communication Networks*. Springer Nature Switzerland, 2025, pp. 69–80. ISBN: 9783031808531. DOI: 10.1007/978-3-031-80853-1_6.

[12] Konstantin Avrachenkov, Evsey Morozov, and Ruslana Nekrasova. "Stability analysis of two-class retrial systems with constant retrial rates and general service times". In: *Performance Evaluation* 159 (Jan. 2023), p. 102330. ISSN: 0166-5316. DOI: 10.1016/j.peva.2022.102330.

[13] Ruslana Nekrasova et al. "Stability analysis of a two-class system with constant retrial rate and unreliable server". In: *Annals of Operations Research* 331.2 (Feb. 2023), pp. 1029–1051. ISSN: 1572-9338. DOI: 10.1007/s10479-023-05216-6.

[14] Ruslana Nekrasova. "Regeneration estimation in partially stable two class retrial queue". In: *Annales Mathematicae et Informaticae* 56 (2023), pp. 84–94. ISSN: 1787-6117. DOI: 10.33039/ami.2022.12.008.

[15] Evsey Morozov and Bart Steyaert. *Stability Analysis of Regenerative Queueing Models: Mathematical Methods and Applications*. Springer International Publishing, 2021. ISBN: 9783030824389. DOI: 10.1007/978-3-030-82438-9.

[16] G. Fayolle, V. A. Malyshev, and M. V. Menshikov. *Topics in the Constructive Theory of Countable Markov Chains*. Cambridge University Press, May 1995. ISBN: 9780511984020. DOI: 10.1017/cbo9780511984020.

ON THE RELIABILITY ESTIMATION OF THE GAUSSIAN DEGRADATION SYSTEM WITH A PATH-DEPENDENT MEAN DEGRADATION RATE

Oleg Lukashenko

•

Institute of Applied Mathematical Research,
Karelian Research Centre of RAS, Petrozavodsk, Russia;
Petrozavodsk State University, Petrozavodsk, Russia
lukashenko@krc.karelia.ru

Abstract

We consider a system whose degradation dynamic is described by an underlying stochastic process that consists of two components: a centered Gaussian process and a drift term with a so-called path-dependent intensity rate, which means its dependence on the degradation history. The main goal is to estimate the reliability of the system via simulation methods, as its analytical expression is generally not available. The cross-entropy method has been applied to estimate the required quantity with acceptable accuracy. A few numerical experiments have been conducted to study the properties of the proposed estimator.

Keywords: Reliability, Degradation process, Gaussian process, Variance reduction methods, Importance sampling, Cross-entropy method

1. Introduction

The development and evaluation of the models describing the degradation process is an actual research area in reliability analysis since it allows simulation of the damage accumulation process giving an opportunity to estimate the failure probability of the system when its degradation level reaches some critical threshold value.

The degradation arises randomly for many technical systems due to various sources of uncertainties. Thus, it seems quite natural to model the degradation evolution of the system as a stochastic process. There are a few typical classes of stochastic processes that are often used to model the degradation evolution. If the degradation process is monotonic, the Gamma processes [1, 2, 3, 4] and Inverse Gaussian processes [5, 6, 7] (i.e. stochastic processes with independent and stationary increments having Gamma and inverse Gaussian distribution respectively) are often used to model the degradation evolution. Nevertheless, the degradation dynamic of many systems exhibits non-monotonic behavior, that is why the models based on the Wiener process (a well-recognized example of the Gaussian process with stationary and independent increments) are widely used instead [8, 9, 10, 11, 12, 13, 14, 15].

The correlation structure of the degradation process can be rather complicated, thus independence of the increments is not a realistic assumption. Hence, general Gaussian processes whose distributions are completely defined by the mean and covariance function seems a good choice for the degradation modeling [16, 17, 18]. Degradation models based on the Gaussian processes have been successfully applied for different practical issues, such as modeling of the

The standard models assume a fixed mean degradation rate, which is not realistic in practice. The multi-phase Wiener degradation system with a deterministic sequence of change points has been proposed in [15]. The more general case of general Gaussian process with stationary and possibly dependent increments was considered in [21], where the required performance measures were estimated via the Monte Carlo simulation technique using the special variant of the conditional Monte Carlo method to reduce the variance of the estimator. In the next work [22] the case of random change points has been studied and a few variance reduction methods including importance sampling and control variates have been applied.

In this paper, we consider a more general setting when the degradation intensity at each time instant is a random variable whose distribution could depend on the path of the underlying stochastic process, i.e. on the degradation history up to the current time instant. Such an assumption can model possible acceleration or deceleration of the degradation process. In order to estimate the reliability of the considered degradation system we apply the cross-entropy method aiming at the approximation of the so-called zero-variance proposal distribution followed by the standard importance sampling step.

The rest of this paper is organized as follows. Section 2 describes the proposed Gaussian degradation model with a path-dependent degradation intensity. Section 3 is devoted to estimating performance measures of the considered reliability model via Monte Carlo simulation focusing on the variance reduction techniques. The general idea of the cross-entropy method is discussed as well as a few implementation details related to the considered rare-event simulation problem. The results of the numerical experiments are presented in Section 4. Finally, a few concluding remarks are given in Section 5.

2. Model description

The degradation dynamic of the considered reliability system is governed by the stochastic process $\{A(t),\, t \in \mathcal{T}\}$ defined as

$$A(t) = \Lambda(t) + X(t), \tag{1}$$

where the terms on the right-hand side are defined as follows:

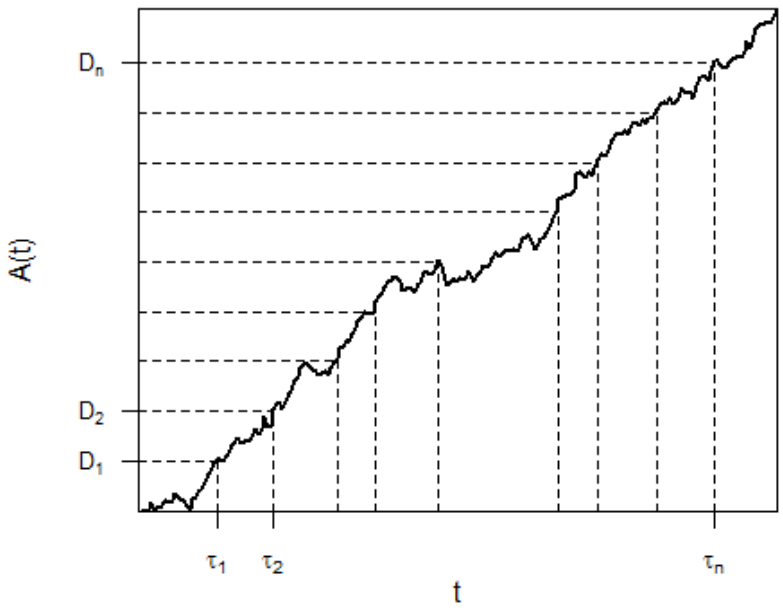


Figure 1: *Path-dependent change-points of degradation intensity.*

- $\{X(t),\ t \in \mathcal{T}\}$ is the centered Gaussian process with a covariance function

$$\Gamma(t,s) := \mathbb{E}\left[X(t)X(s)\right].$$

  The process X can be seen as an additive correlated noise reflecting real-world uncertainties and variations in degradation evolution due to external latent factors.

- The drift term $\Lambda(t) = m(t)t$ has a time-dependent degradation rate $m(t)$, which generally could be a random variable. In this research, it is additionally assumed that $m(t)$ depends on the path of the degradation process (i.e. degradation history) $(A(s),\ s < t)$ up to the current time instant t. Thus, we call it a path-dependent degradation intensity rate.

Now we give the following concrete examples of the path-dependent degradation rates:

1. The path-dependent change-points of the degradation intensity (see. fig. 1):

$$m(t) = \sum_{i\geq 1} m_i \cdot I(\tau_{i-1} < t \leq \tau_i),$$

   where I denotes the indicator function, $0 = \tau_0 < \tau_1 < \tau_2 < \dots$ is a sequence of random variables defined as the successive hitting times

$$\tau_i = \min\{t :\ A(t) \geq D_i\}$$

   with $D_1 < D_2 <$ ... being increasing sequence of the given intermediate thresholds; $\{m_i\}$ are given values. Note that if, at some time instant t, $m(t) = m_i$, then, after instant t, degradation rate can not take any value m_k with $k < i$.

2. The current degradation rate depends on the previous degradation level (see. fig. 2)

$$m(t) = \begin{cases} \sum\limits_{i\geq 1} m_i \cdot I(D_{i-1} < A(t-1) \leq D_i), & t \geq 1, \\ m_0, & t < 1, \end{cases}$$

   where $\{m_i\}$ are given values, $0 = D_0 < D_1 < D_2 <$ ... is increasing sequence of the given intermediate thresholds. This example is quite similar to the previous one but allows $m(t)$ return to the previous levels (such an event potentially occur when the degradation process can locally decrease).

Note that the two examples given above illustrate the general idea of the path-dependent intensity rate. We believe that such an assumption reflects different scenarios of the acceleration or deceleration of degradation dynamic depending on the prehistory of degradation evolution. At the same time, more complex models can be considered.

The lifetime of the considered system is defined as follows

$$T_D := \min\{t \in \mathcal{T} :\ A(t) \geq D\}, \tag{2}$$

where D is the given last threshold. We are interested in the estimating the reliability of the system defined as the tail distribution of the lifetime:

$$R(u) := \mathbb{P}(T_D \geq u), \quad u > 0. \tag{3}$$

The closed-form expression of the target performance measure (3) is not available in general except a few simple particular cases of the Wiener degradation model with a deterministic sequence of the degradation intensity change-points (see [15] for more details). Thus, evaluate the tail distribution (3), one has to rely on the Monte Carlo methods which are discussed in the next section.

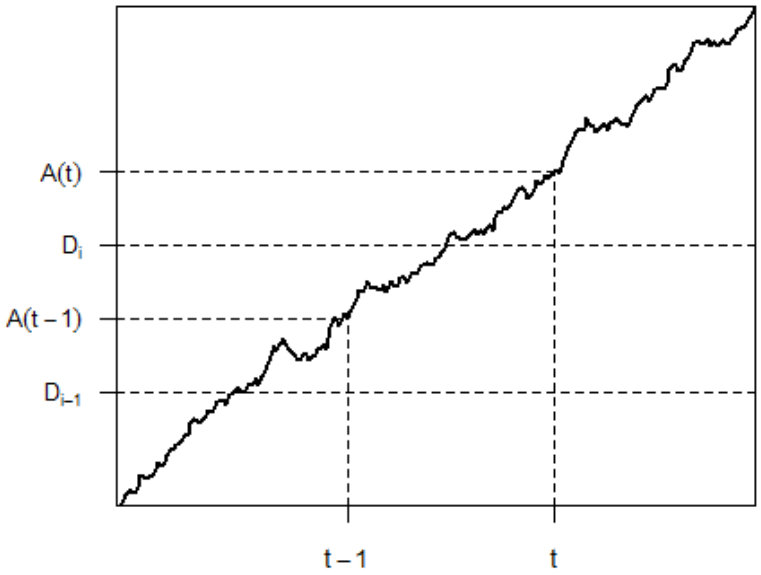


Figure 2: *Dependence of the current degradation rate on the previous degradation level: when $D_{i-1} < A(t-1) \leq D_i$ the current degradation rate is m_i.*

3. Monte Carlo estimation

Denote by Z_u an unbiased estimator of $R(u)$, that is $\mathbb{E}Z_u = R(u)$. Obviously, $R(u) \to 0$ as $u \to \infty$, thus u is called the rarity parameter. To estimate $R(u)$ by the Monte Carlo (MC) simulation, one has to sample from the distribution of the random variable Z_u and calculate the sample mean

$$\widehat{R}_u := \frac{1}{N} \sum_{n=1}^{N} Z_u^{(n)}. \tag{4}$$

The measure of the quality of the estimator is expressed by the *relative error* (RE):

$$\mathrm{RE}\left[\widehat{R}_u\right] := \frac{\sqrt{\mathbb{V}\mathrm{ar}\left[\widehat{R}_u\right]}}{\mathbb{E}\left[\widehat{R}_u\right]}. \tag{5}$$

The standard MC approach is based on the indicator of the target event, i.e.

$$Z_u^{\mathrm{MC}} = I(T_D \geq u).$$

It is straightforward to show that

$$\mathrm{RE}\left[\widehat{R}_u^{\mathrm{MC}}\right] \sim \frac{1}{\sqrt{N \cdot R(u)}}, \quad \text{as} \quad u \to \infty,$$

where $a \sim b$ means $a/b \to 1$. Thus, the RE of the standard MC estimator tends to infinity when the target probability tends to zero, hence a large sample size is required to get a suitable RE. Moreover, in order to have bounded RE the sample size N must grow at least at the same rate as $1/R(u)$ when $u \to \infty$.

There are a few rare event simulation techniques [23, 24] aiming at modifying the estimator (4) to reduce its variance, hence requiring less sample size for the desired accuracy. One class of these methods, namely importance sampling, is briefly discussed below.

In what follows, we restrict ourselves to the finite-dimensional case (enough for the simulation needs) when $\mathcal{T} = \{t_1, ..., t_L\}$, where L is the required simulation length (then t_L is the simulation horizon).

3.1. Importance Sampling

Importance sampling is a widely used method for variance reduction. Its main idea is selecting the proposal distribution so that the target rare event becomes more likely to occur.

Let $f(\mathbf{x})$ be the probability density function (pdf) of the Gaussian random vector $(X(t_1), ..., X(t_L))$ and

$$h_u\mathbf{x}) = I(T_D(\mathbf{x}) \geq u), \quad \mathbf{x} \in \mathbb{R}^L.$$

Note that

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{L/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right), \quad \mathbf{x} \in \mathbb{R}^L, \tag{6}$$

where the covariance matrix

$$\mathbf{\Sigma} = ||\Gamma(t_i, t_j)||_{i,j=1,...,L}. \tag{7}$$

Having some proposal pdf $g(\mathbf{x})$, the target probability is

$$R(u) = \int h_u(\mathbf{x})\frac{f(\mathbf{x})}{g(\mathbf{x})}g(\mathbf{x})d\mathbf{x} = \mathbb{E}_g\left[h_u(\mathbf{X})\frac{f(\mathbf{X})}{g(\mathbf{X})}\right], \tag{8}$$

where $\mathbb{E}_g$ denotes the expectation with respect to the new pdf g. Thus,

$$Z_u^{\text{IS}} = h_u(\mathbf{X})\frac{f(\mathbf{X})}{g(\mathbf{X})}, \quad \mathbf{X} \sim g, \tag{9}$$

is the unbiased estimator of $R(u)$.

The main problem arising here is how to choose the proposal distribution g in order to reduce the variance of the estimator Z_u^{IS}. It is quite straightforward to show (see for example [24]) that the optimal density g_* which provides the zero variance of the estimator has the following form:

$$g_*(\mathbf{x}) = \frac{h_u(\mathbf{x})f(\mathbf{x})}{R(u)}. \tag{10}$$

However, it is not implementable in practice because it requires knowledge of the target quantity $R(u)$. Nevertheless, sometimes it is possible to find a precise approximation of the optimal density g_*.

3.1.1 Cross-entropy method

The aim of this method is to find the proposal distribution g close to the desired zero-variance distribution g_* in the sense of the Kullback-Leibler divergence defined as [25, 26]

$$\begin{aligned}\mathcal{D}(g_*, g) &= \mathbb{E}_{g_*}\left[\log\frac{g_*(\mathbf{X})}{g(\mathbf{X})}\right] \\ &= \int g_*(\mathbf{x})\log g_*(\mathbf{x})d\mathbf{x} - \int g_*(\mathbf{x})\log g(\mathbf{x})d\mathbf{x}.\end{aligned}$$

When both proposal and nominal distributions are selected from some parametric set of distributions, this leads to the finite dimensional optimization problem. For this reason let's consider the parametric class of the multivariate normal distributions $f(\cdot;\boldsymbol{\theta})$ with parameters $\boldsymbol{\theta} = \{\mathbf{v}, \sigma\}$, i. e., with the mean vector $\mathbf{v} \in \mathbb{R}^L$ and covariance matrix $\mathbf{\Sigma}' = \sigma\mathbf{\Sigma}$, where $\mathbf{\Sigma} \in \mathbb{R}^{L\times L}$ is defined by (7) and $\sigma > 0$ is a scaling parameter.

The nominal pdf f defined by (6) belongs to this parametric class with $\boldsymbol{\theta}_0 = \{\mathbf{0}, 1\}$. Let's further choose the proposal density g from the same family with another parameter vector $\boldsymbol{\theta}$ further referred as a reference parameter. The cross-entropy (CE) method is based on finding an optimal reference parameter:

$$\begin{aligned}\boldsymbol{\theta}^* := \{\mathbf{v}^*, \sigma^*\} &= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, \mathcal{D}(g_*, f(\cdot;\boldsymbol{\theta})) \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{\eta}} \left[h_u(\mathbf{X}) W(\mathbf{X};\boldsymbol{\theta}_0,\boldsymbol{\eta}) \log f(\mathbf{X};\boldsymbol{\theta})\right],\end{aligned}$$

where $\mathbb{E}_{\boldsymbol{\eta}}$ denotes the expectation with respect to the distribution $f(\cdot;\boldsymbol{\eta})$ and

$$W(\mathbf{X};\boldsymbol{\theta}_0,\boldsymbol{\eta}) = \frac{f(\mathbf{X};\boldsymbol{\theta}_0)}{f(\mathbf{X};\boldsymbol{\eta})}.$$

The given above stochastic optimization problem is replaced by its stochastic counterpart:

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \frac{1}{M} \sum_{k=1}^{M} h_u(\mathbf{X}^k) W(\mathbf{X}^k,\boldsymbol{\theta}_0,\boldsymbol{\eta}) \log f(\mathbf{X}^k;\boldsymbol{\theta}), \tag{11}$$

where $\mathbf{X}^1,..,\mathbf{X}^M$ is the sample from the distribution $f(\cdot;\boldsymbol{\eta})$. Setting the gradient of (11) with respect to $\mathbf{v}$ and σ to zero it is straightforward to obtain that

$$\mathbf{v}^* = \frac{\sum\limits_{i=1}^{M} h_u(\mathbf{X}^i) W(\mathbf{X}^i;\boldsymbol{\theta}_0,\boldsymbol{\eta}) \mathbf{X}^i}{\sum\limits_{i=1}^{M} h_u(\mathbf{X}^i) W(\mathbf{X}^i;\boldsymbol{\theta}_0,\boldsymbol{\eta})}, \tag{12}$$

$$\sigma^* = \frac{\sum\limits_{i=1}^{M} h_u(\mathbf{X}^i) W(\mathbf{X}^i;\boldsymbol{\theta}_0,\boldsymbol{\eta}) (\mathbf{X}^i - \mathbf{v}^*)^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}^i - \mathbf{v}^*)}{L \sum\limits_{i=1}^{M} h_u(\mathbf{X}^i) W(\mathbf{X}^i;\boldsymbol{\theta}_0,\boldsymbol{\eta})}. \tag{13}$$

The main problem arising here is that most values of the $h_u(\mathbf{X}^i)$ are zero. In this case the so-called multi-level procedure [25, 26] can be applied. According to this approach the sequence $(u_t, \boldsymbol{\theta}_t)$ of both rarity and reference parameters is constructed, such that $\boldsymbol{\theta}_t$ is a solution of the problem (11) with the previous value of the reference parameter, i. e. $\boldsymbol{\eta} = \boldsymbol{\theta}_{t-1}$.

To be more precise, we start from the nominal vector of the parameters $\boldsymbol{\theta}_0$. Let further $(u_{t-1}, \boldsymbol{\theta}_{t-1})$ be the current values of the rarity and reference parameters respectively. The subsequent value of u_t is obtained by drawing a sample $\mathbf{X}^1,..,\mathbf{X}^M$ from the distribution $f(\cdot;\boldsymbol{\theta}_{t-1})$ as follows [27]:

$$u_t = T_D^{((1-\alpha)M)}, \tag{14}$$

where $T_D^{(j)}$ is the j-th order-statistics of the sequence $T_D(\mathbf{X}^1),...,T_D(\mathbf{X}^M)$; α is a free parameter chosen not very small (the typical value in practice is $\alpha = 0.05$).

Then, the next value of the reference parameter $\boldsymbol{\theta}_t$ is derived as a solution of the following CE program:

$$\boldsymbol{\theta}_t = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{\theta}_{t-1}} \left[h_{u_t}(\mathbf{X}) W(\mathbf{X};\boldsymbol{\theta}_0,\boldsymbol{\theta}_{t-1}) \log f(\mathbf{X};\boldsymbol{\theta})\right].$$

Thus, the solution of the corresponding stochastic counterpart has the form (12)-(13) with $\boldsymbol{\eta} = \boldsymbol{\theta}_{t-1}$ and $u = u_t$.

The described above iterative algorithm terminates when $u_t \geq u$. After that, the target quantity is estimated as

$$\widehat{R}_u^{\mathrm{CE}} := \frac{1}{N} \sum_{i=1}^{N} h_u(\mathbf{X}^i) W(\mathbf{X}^i;\boldsymbol{\theta}_0,\boldsymbol{\theta}_S), \quad \mathbf{X}^1,...,\mathbf{X}^N \sim f(\cdot;\boldsymbol{\theta}_S), \tag{15}$$

where S denotes the last iteration number. The resulting CE procedure is summarized in the Algorithm 1 below.

Algorithm 1 The main CE algorithm

$\boldsymbol{\theta}_0 \leftarrow \{\mathbf{0}, 1\}$
$u_0 \leftarrow 0$
$t \leftarrow 0$
while $u_t < u$ **do**
  $t \leftarrow t+1$
  Generate a sample $\mathbf{X}^1, ..., \mathbf{X}^M \sim f(\cdot, \boldsymbol{\theta}_{t-1})$
  Compute u_t as (14)
  Compute $\boldsymbol{\theta}_t$ as a solution of CE program (12)-(13) with $\boldsymbol{\eta} = \boldsymbol{\theta}_{t-1}$ and $u = u_t$
end while
Compute the reliability according to (15)

Remark 1. Instead of estimating the mean vector $\mathbf{v}$ and the scaling factor σ of the covariance matrix one can try to estimate the full covariance matrix, namely choosing the parametrization as $\boldsymbol{\theta} = \{\mathbf{v}, \boldsymbol{\Sigma}\}$, where both mean vector $\mathbf{v}$ and covariance matrix $\boldsymbol{\Sigma}$ are estimated from the optimization problem (11). But the dimension of the corresponding optimization problem significantly increases, thus the number of samples M in (11) should be increased appropriately, otherwise, the overfitting problem can occur.

4. Simulation Results

In this section, we provide a simulation analysis of the accuracy of the proposed estimator.

All experiments were conducted for the case when the process X is the fractional Brownian motion (FBM) with the covariance function

$$\Gamma(t,s) := \frac{1}{2}\left(t^{2H} + s^{2H} - |t-s|^{2H}\right).$$

Now we describe in brief the simulation procedure. First note that it is enough to simulate the FBM over the interval $[0, u]$. Sample paths of the FBM are drawn as realizations of the random vector:

$$(X(t_1), ...X(t_L)),$$

where $t_1, ..., t_L$ is a uniform partition of the interval $[0, u]$.

In all experiments performed below $N = 10000$ trajectories of the FBM with Hurst parameter $H = 0.7$ were generated.

The first numerical experiment deals with the case of a single change-point:

$$m(t) = \begin{cases} m_1, & t < \tau, \\ m_2, & t \geq \tau, \end{cases} \tag{16}$$

where

$$\tau = \min\{t : \ A(t) \geq D_1\},$$

and $D_1 < D$ is a given intermediate threshold.

The following values of the other parameters were used: $m_1 = 1, m_2 = 3$; $D_1 = 10$, $D = 20$. The number of samples M in (11) is 10^4. To verify the accuracy of the proposed estimators, we considered the dependence of the relative error on the rarity parameter u for both CE and standard MC estimators. The numerical results are presented in Table 1. The obtained results demonstrate that the CE estimator significantly outperforms the standard MC one. Note that the relative error is also estimated. To study the variability of the RE we performed 100 simulation runs of the described above experiment for the fixed value of the rarity parameter $u = 150$. We repeat this procedure for both $M = 10^4$ and $M = 10^5$ samples in the CE optimization problem (11) and calculate the empirical distribution of the RE for both cases. The obtained results presented

Table 1: *Performance of the estimators in case of the single change-point defined as* (16)*:* $m_1 = 1$, $m_2 = 3$.

u	$\widehat{R}^{\mathrm{MC}}$	$\widehat{R}^{\mathrm{CE}}$	$\mathrm{RE}(\widehat{R}^{\mathrm{MC}})$	$\mathrm{RE}(\widehat{R}^{\mathrm{CE}})$
40	0.0112	0.0096	0.0939	0.0239
50	0.0028	0.0035	0.1887	0.0280
60	0.0021	0.0014	0.2179	0.0416
70	9e-04	6.9e-04	0.3332	0.0336
80	7e-04	3.3e-04	0.3778	0.0385
90	3e-04	1.7e-04	0.5772	0.0488
100	–	8.7e-05	–	0.0581
110	–	4.9e-05	–	0.0675
120	–	2.6e-05	–	0.0999
130	–	1.4e-05	–	0.1269
140	–	8.3e-06	–	0.1541
150	–	5.9e-06	–	0.1861
160	–	1.9e-06	–	0.1827
170	–	1.4e-06	–	0.1979
180	–	8.8e-07	–	0.2371
190	–	1.6e-07	–	0.3514

in Fig. 3 indicate that the behavior of the CE estimator is much more robust in the sense of the variance of the RE when $M = 10^5$ since the larger value of M leads to the more precise approximation of the zero-variance distribution and consequently to the smaller RE which has the order 10^{-2} in case of $M = 10^5$.

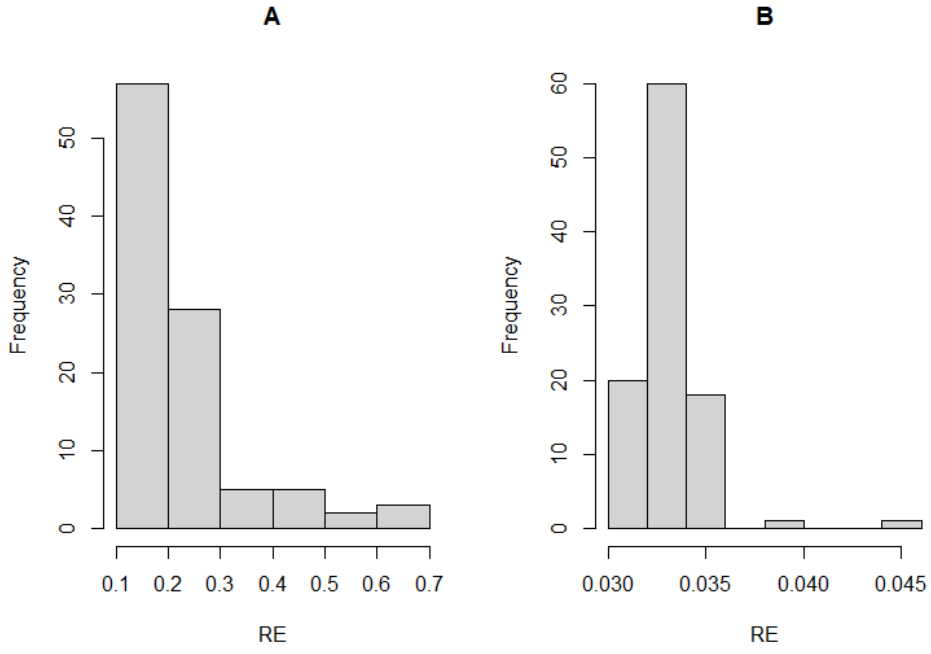


Figure 3: *Histogram of the RE in case of the degradation density defined as* (16)*: the number of samples used in CE optimization problem* (11) $M = 10^4$ (**A**), $M = 10^5$ (**B**).

For the second experiment, we consider the degradation intensity being dependent on the

Table 2: *Performance of the estimators for the degradation intensity defined as* (17)*:* $m_1 = 1$, $m_2 = 3$.

u	$\widehat{R}^{\mathrm{MC}}$	$\widehat{R}^{\mathrm{CE}}$	$\mathrm{RE}(\widehat{R}^{\mathrm{MC}})$	$\mathrm{RE}(\widehat{R}^{\mathrm{CE}})$
40	0.0092	0.0091	0.1037	0.0285
50	0.0055	0.0034	0.1344	0.0449
60	0.0021	0.0016	0.2179	0.0734
70	7e-04	7e-04	0.3778	0.0303
80	4e-04	3.5e-04	0.4999	0.0395
90	1e-04	1.7e-04	1	0.0412
100	–	8.7e-05	–	0.0606
110	–	5.7e-05	–	0.0881
120	–	2.4e-05	–	0.1102
130	–	1.4e-05	–	0.1466
140	–	9.8e-06	–	0.1387
150	–	3.7e-06	–	0.1247
160	–	3.2e-06	–	0.1535
170	–	1.6e-06	–	0.2080
180	–	6.8e-07	–	0.1847
190	–	4.9e-07	–	0.2229

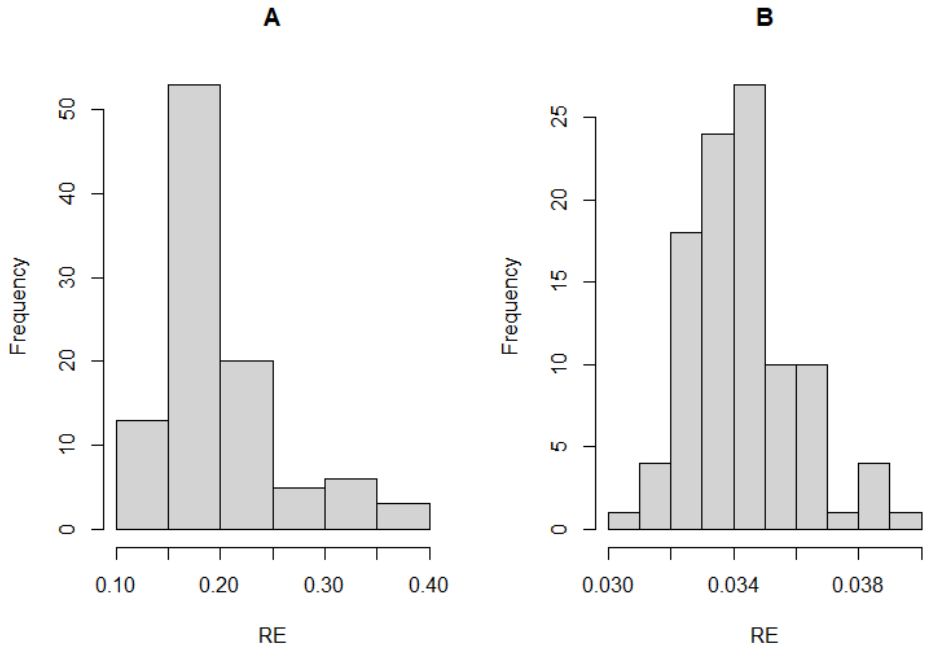


Figure 4: *Histogram of the RE in case of the degradation density defined as* (17)*: the number of samples used in CE optimization problem* (11) $M = 10^4$ *(***A***),* $M = 10^5$ *(***B***).*

degradation level at the previous time instant:

$$m(t) = \begin{cases} m_1, & A(t-1) < D_1, \\ m_2, & A(t-1) \geq D_1, \end{cases} \quad t \geq 1. \tag{17}$$

We conducted the same experiments with the same values of the parameters as a "sanity check". The results are summarized in Table 2 and Fig. 4. The results are quite similar to the ones obtained from the previous experiment since according to the definition (17) the switching

time instant to the intensity m_2 will be close to the same value in the change-point model (16). Moreover, since $0 < m_1 < m_2$ the probability that the intensity will switch back to the level m_1 is rather small for FBM with $H = 0.7$.

5. Conclusion

In this paper, we have considered the Gaussian degradation model with the degradation intensity being dependent on the degradation history. Such a model has a very complicated dependence structure that makes it difficult to obtain the analytical expressions for the required performance measures. Thus, simulation remains the only available tool for analyzing such systems. To estimate the reliability of the considered degradation system the cross-entropy method has been applied which allows to calculate probabilities of rare events having a limited sample size. The numerical results indicate that the proposed estimator provides a significant reduction of the relative error in comparison with the standard Monte Carlo approach. The problem of accurate approximation of the zero-variance distribution seems to be an interesting topic of a further research.

References

[1] Zhengqiang Pan and Narayanaswamy Balakrishnan. "Reliability modeling of degradation of products with multiple performance characteristics based on gamma processes". In: *Reliability Engineering amp; System Safety* 96.8 (Aug. 2011), pp. 949–957. DOI: 10.1016/j.ress.2011.03.014.

[2] Xiaolin Wang et al. "Real-time Reliability Evaluation for an Individual Product Based on Change-point Gamma and Wiener Process: Real-time Reliability Evaluation". In: *Quality and Reliability Engineering International* 30.4 (Feb. 2013), pp. 513–525. DOI: 10.1002/qre.1504.

[3] Xiaofei Wang et al. "Degradation data analysis based on gamma process with random effects". In: *European Journal of Operational Research* 292.3 (Aug. 2021), pp. 1200–1208. DOI: 10.1016/j.ejor.2020.11.036.

[4] J.M. van Noortwijk. "A survey of the application of gamma processes in maintenance". In: *Reliability Engineering amp; System Safety* 94.1 (Jan. 2009), pp. 2–21. DOI: 10.1016/j.ress.2007.03.019.

[5] Zhi-Sheng Ye and Nan Chen. "The Inverse Gaussian Process as a Degradation Model". In: *Technometrics* 56.3 (July 2014), pp. 302–311. DOI: 10.1080/00401706.2013.830074.

[6] Chien-Yu Peng. "Inverse Gaussian Processes With Random Effects and Explanatory Variables for Degradation Data". In: *Technometrics* 57.1 (Jan. 2015), pp. 100–111. DOI: 10.1080/00401706.2013.879077.

[7] Weiwen Peng et al. "Inverse Gaussian process models for degradation analysis: A Bayesian perspective". In: *Reliability Engineering amp; System Safety* 130 (Oct. 2014), pp. 175–189. DOI: 10.1016/j.ress.2014.06.005.

[8] Waltraud Kahle and Axel Lehmann. "The Wiener Process as a Degradation Model: Modeling and Parameter Estimation". In: *Advances in Degradation Modeling*. Birkhäuser Boston, Oct. 2009, pp. 127–146. ISBN: 9780817649241. DOI: 10.1007/978-0-8176-4924-1_9.

[9] Xiao-Sheng Si et al. "A Wiener-process-based degradation model with a recursive filter algorithm for remaining useful life estimation". In: *Mechanical Systems and Signal Processing* 35.1–2 (Feb. 2013), pp. 219–237. DOI: 10.1016/j.ymssp.2012.08.016.

[10] Meng Xiao et al. "Degradation Modeling Based on Wiener Process Considering Multi-Source Heterogeneity". In: *IEEE Access* 8 (2020), pp. 160982–160994. DOI: 10.1109/access.2020.3020723.

[11] Guru Prakash and Anshul Kaushik. "A change-point-based Wiener process degradation model for remaining useful life estimation". In: *Safety and Reliability* 39.3–4 (Aug. 2020), pp. 253–279. DOI: 10.1080/09617353.2020.1801165.

[12] G. A. Whitmore. "Estimating degradation by a wiener diffusion process subject to measurement error". In: *Lifetime Data Analysis* 1.3 (1995), pp. 307–319. DOI: 10.1007/bf00985762.

[13] Wei-an Yan et al. "Real-time reliability evaluation of two-phase Wiener degradation process". In: *Communications in Statistics - Theory and Methods* 46.1 (Sept. 2016), pp. 176–188. DOI: 10.1080/03610926.2014.988262.

[14] Donghui Pan et al. "Degradation Data Analysis Using a Wiener Degradation Model With Three-Source Uncertainties". In: *IEEE Access* 7 (2019), pp. 37896–37907. DOI: 10.1109/access.2019.2906325.

[15] Hongda Gao, Lirong Cui, and Dejing Kong. "Reliability analysis for a Wiener degradation process model under changing failure thresholds". In: *Reliability Engineering amp; System Safety* 171 (Mar. 2018), pp. 1–8. DOI: 10.1016/j.ress.2017.11.006.

[16] Zhihua Wang et al. "A generalized degradation model based on Gaussian process". In: *Microelectronics Reliability* 85 (June 2018), pp. 207–214. DOI: 10.1016/j.microrel.2018.05.001.

[17] W. J. Padgett and Meredith A. Tomlinson. "Inference from Accelerated Degradation and Failure Data Based on Gaussian Process Models". In: *Lifetime Data Analysis* 10.2 (June 2004), pp. 191–206. DOI: 10.1023/b:lida.0000030203.49001.b6.

[18] Zhen Chen et al. "Two-phase degradation data analysis with change-point detection based on Gaussian process degradation model". In: *Reliability Engineering amp; System Safety* 216 (Dec. 2021), p. 107916. DOI: 10.1016/j.ress.2021.107916.

[19] C. Park and W.J. Padgett. "New Cumulative Damage Models for Failure Using Stochastic Processes as Initial Damage". In: *IEEE Transactions on Reliability* 54.3 (Sept. 2005), pp. 530–540. DOI: 10.1109/tr.2005.853278.

[20] Yu Wang, Zhi-Sheng Ye, and Kwok-Leung Tsui. "Stochastic Evaluation of Magnetic Head Wears in Hard Disk Drives". In: *IEEE Transactions on Magnetics* 50.5 (May 2014), pp. 1–7. DOI: 10.1109/tmag.2013.2293636.

[21] Oleg Lukashenko. "On the Reliability Estimation of the Gaussian Multi-phase Degradation System". In: *Distributed Computer and Communication Networks: Control, Computation, Communications*. Springer Nature Switzerland, 2022, pp. 410–421. DOI: 10.1007/978-3-031-23207-7_32.

[22] Oleg Lukashenko. "On the Variance Reduction Methods for Estimating the Reliability of the Multi-phase Gaussian Degradation System". In: *Distributed Computer and Communication Networks: Control, Computation, Communications*. Springer Nature Switzerland, 2024, pp. 197–208. DOI: 10.1007/978-3-031-50482-2_16.

[23] Sheldon M. Ross. *Simulation*. 4th ed. OCLC: ocm69672100. Amsterdam ; Boston: Elsevier Academic Press, 2006. ISBN: 978-0-12-598063-0.

[24] Dirk P. Kroese, Thomas Taimre, and Zdravko I. Botev. *Handbook of Monte Carlo Methods*. Wiley, Feb. 2011. ISBN: 9781118014967. DOI: 10.1002/9781118014967.

[25] Dirk P. Kroese, Reuven Y. Rubinstein, and Peter W. Glynn. "The Cross-Entropy Method for Estimation". In: *Handbook of Statistics - Machine Learning: Theory and Applications*. Elsevier, 2013, pp. 19–34. DOI: 10.1016/b978-0-444-53859-8.00002-3.

[26] R.Y. Rubinstein and D.P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Information Science and Statistics. Springer, 2004. ISBN: 978-0-387-21240-1.

[27] Pieter-Tjerk de Boer et al. "A Tutorial on the Cross-Entropy Method". In: *Annals of Operations Research* 134.1 (Feb. 2005), pp. 19–67. DOI: 10.1007/s10479-005-5724-z.

REGENERATION AND APPROXIMATION OF A QUEUEING SYSTEM FED BY SUPERPOSED INPUT WITH WEIBULL COMPONENTS

Irina Peshkova, Michele Pagano, Evsey Morozov

•

Petrozavodsk State university, Russia
Institute of Applied Mathematical Research of the Karelian Research Centre of RAS,
Department of Information Engineering, University of Pisa, Italy
iaminova@petrsu.ru

Abstract

We study a single-server queueing system with a superposed input process formed by independent stationary renewal processes with Weibull interarrival time distributions. An approximating system with renewal input process based on Palm construction is considered. Moreover, the accuracy of the approximation in the terms of Kolmogorov distance is discussed. Finally, we demonstrate how to construct, in the initially non-regenerative queueing system, the artificial regenerations based on the exponential splitting technique.

Keywords: exponential splitting, regenerative estimation, Weibull distribution, superposition of renewal inputs

1. Introduction

In this research, we consider a single-server queueing system, denoted by $\sum_{i=1}^{m} GI_i/G/1$, in which the input process is a superposition of m independent stationary renewal processes (components). Such systems play an important role in modeling the modern communication systems, characterized by the superposition of independent heterogeneous traffic flows.

It is well-known that systems with the superposed input process are not (classically) regenerative unless the input components are Poisson processes. In particular, it is shown in [1] that such a superposed process obeys a weaker property, namely, the so-called *one-dependent regeneration*.

On the other hand, the presence of regenerative structure in the processes describing the dynamics of the system makes its analysis much more effective and accurate.

The study of systems with superposed input process has a long story, see for instance, the papers [2]- [3], in which various aspects of such systems are considered, including the covariance between interarrival times in the superposed process [2] and conditions for its convergence to a Poisson process [4, 5]. Moreover, a considerable attention has been paid to the analysis of the approximation of such a system by a classic $GI/G/1$ queueing system with renewal input, which, in turn, is constructed as a Palm stationary process composed from the (properly weighted) distributions of the interarrival times in the component processes [6]. In a companion paper [7], we studied such a system provided the interarrival times in the component processes have either exponential or Pareto distributions. The presence of the heavy-tailed distributions permits to apply the so-called exponential splitting [8] to construct classical regenerations in a modified system, which is equivalent, in the terms of marginal (one-dimensional) distributions, to the

original one. This approach uses a possibility to split the corresponding 'heavy-tailed' density in such a way that it is represented as a two-component mixture containing (weighted) exponential density.

The main contribution of the present research is that we now consider, in the superposed input process, independent (stationary) heavy-tailed Weibull components, instead of the Pareto distribution used in the related paper [7]. We apply the exponential splitting to construct classical regenerations and, using regenerative method, obtain then the confidence interval for the mean stationary workload (unfinished work) in the system. One more contribution of this work is that, for such a system, we construct an approximating system based on the Palm theory and investigate the accuracy of this approximation using various metrics. Numerical results illustrating theoretical findings are included as well. In summary, in this paper we realize, in the main features, the research program which has been developed in our previous paper [7] for the superposed process containing Weibull distributed components.

The paper is organized as follows. In Section 2, we describe the basic model and an approximating renewal process, based on the Palm approach, with the same (one-dimensional) interevent distribution as in the original superposed process. Then we focus on the input process containing two independent renewal processes with Weibull interevent distributions. As we show, this approximating distribution is a two-component mixture, again containing Weibull distributions. Section 3 deals with the splitting procedure of the heavy-tailed Weibullian density, which then is used to construct regenerations of both the (superposed) input process and the processes describing the dynamics of the whole system. Section 4 contains simulation results related to the accuracy (in terms of Kolmogorov distance) of the approximation of the original system by a $GI/G/1$ queueing system (Section 4.1) and to the efficiency of the exponential splitting (Section 4.2), measured in the term of the frequency of the regenerations in a limited simulation time.

2. Basic model

We remind that the basic system is fed by a superposition of m independent renewal processes. We assume that the ith input component is a stationary renewal process, which is defined by the independent identically distributed (iid) interarrival times between customers (class-i ones) $\{\tau_k^{(i)}, k \geq 1\}$ with distribution function A_i and mean $\mathsf{E}\tau^{(i)} = 1/\lambda_i < \infty$, ($\tau^{(i)}$ is the generic interarrival time) and the time up to the 1st arrival has the following *integrated-tail distribution*

$$\lambda_i \int_0^x (1 - A_i(u))du, \;\; i = 1, \ldots, m.$$

It is well-known that such a choice guarantees stationarity of each component (renewal) process and as a result the stationarity of the superposed point process composed from the (ordered) points of the components [9] (Sect.3, Ch. 5.) The stationary interarrival time in the superposed (stationary) process follows the Palm distribution [3, 10]:

$$A(x) = 1 - \sum_{i=1}^{m} \frac{\lambda_i}{\lambda}(1 - A_i(x)) \prod_{j, j \neq i} \lambda_j \int_x^{\infty} (1 - A_j(x))dx, \tag{1}$$

where $\lambda = \lambda_1 + \cdots + \lambda_m$. We note that the (tail) distribution $1 - A(x)$ is a m-component mixture (with the weights $p_i := \lambda_i/\lambda$) of the original distributions describing interarrival times in the component renewal processes. Also denote by $\{S_k^{(i)}, k \geq 1\}$ the iid service times of the class-i customers (from the ith component process) with distribution function B_i and finite mean service rate $\mu_i := 1/\mathsf{E}S^{(i)}$, $i = 1, \ldots, m$. Then the service time distribution (of an arbitrary customer) in the system is also represented as the mixture

$$B(x) := \sum_{i=1}^{m} p_i B_i(x), \;\; x \geq 0. \tag{2}$$

The stability criterion of this system is well-known [6, 11]:

$$\rho = \sum_{i=1}^{m} \lambda_i \mathsf{E}S^{(i)} < 1, \tag{3}$$

which also is the stability criterion of the original $\sum_{i=1}^{m} GI/G/1$ system fed by a stationary input process [10]. Note that in the latter case, in the absence of regenerative structure, the proof of stability is based on construction proposed in [12]. We remark that, as it is easy to check, condition (3) can also be written as $\lambda ES < 1$, where S has distribution $B(x)$ from (2).

In what follows we will consider, as the main example, a superposed input composed by two independent stationary components with iid interarrival times $\{\tau_k^{(i)}, k \geq 1\}$ having Weibull distribution (denoted further by $We(\alpha_i, \beta_i)$) with parameters α_i, β_i:

$$A_i(x) = 1 - e^{-(x/\alpha_i)^{\beta_i}}, \;\; x \geq 0, \; \alpha_i > 0, \; \beta_i > 0, \;\; i = 1, 2. \tag{4}$$

Remark. The Weibull distribution was introduced and studied by several prominent researchers in the 1930s, besides W. Weibull, these include R. Fisher, L. Tippett, and L. von Mises. B. Gnedenko later has established the conditions under which a suitably normalized sequence of extreme values converges to one of three limiting distributions, including Weibull distribution, [13]. Because of this prominent research, the generalized two-parameter distribution (4) sometimes also referred to as the Weibull-Gnedenko distribution, see for instance, [14].

Substituting distributions (4) in formula (1), we obtain the distribution of the stationary *renewal input process* in the form:

$$A(x) = 1 - \frac{p}{\Gamma(1/\beta_2)} e^{-(x/\alpha_1)^{\beta_1}} \Gamma(1/\beta_2, (x/\alpha_2)^{\beta_2}) - \frac{1-p}{\Gamma(1/\beta_1)} e^{-(x/\alpha_2)^{\beta_2}} \Gamma(1/\beta_1, (x/\alpha_1)^{\beta_1}),$$

where

$$\Gamma(\xi, x) = \int_x^{\infty} e^{-t} t^{\xi - 1} dt$$

is the upper incomplete Gamma function,

$$p = \frac{\beta_1 \alpha_2 \Gamma(1/\beta_2)}{\beta_1 \alpha_2 \Gamma(1/\beta_2) + \beta_2 \alpha_1 \Gamma(1/\beta_1)}, \tag{5}$$

is the 'mixing proportion' and $\Gamma(\xi)$ is the Gamma function. Finally, applying the relation

$$\Gamma(\xi, x) = \Gamma(\xi) - \gamma(\xi, x),$$

where $\gamma(\xi, x)$ is the lower incomplete gamma function, we obtain the following (tail) Palm distribution of the renewal intervals in the $GI/G/1$ approximating system:

$$\begin{aligned} \overline{A}(x) \;\; &:= \;\; 1 - A(x) = p e^{-(x/\alpha_1)^{\beta_1}} \left(1 - \frac{1}{\Gamma(1/\beta_2)} \gamma(1/\beta_2, (x/\alpha_2)^{\beta_2})\right) + \\ &+ \;\; (1-p) e^{-(x/\alpha_2)^{\beta_2}} \left(1 - \frac{1}{\Gamma(1/\beta_1)} \gamma(1/\beta_1, (x/\alpha_1)^{\beta_1})\right). \end{aligned} \tag{6}$$

Notice that the tail distribution (6) is the mixture,

$$\overline{A}(x) = p\overline{F}_1(x) + (1-p)\overline{F}_2(x), \tag{7}$$

containing tail distributions

$$\begin{aligned} \overline{F}_1(x) \;\; &= \;\; e^{-(x/\alpha_1)^{\beta_1}} \left(1 - \frac{1}{\Gamma(1/\beta_2)} \gamma(1/\beta_2, (x/\alpha_2)^{\beta_2})\right), \\ \overline{F}_2(x) \;\; &= \;\; e^{(-x/\alpha_2)^{\beta_2}} \left(1 - \frac{1}{\Gamma(1/\beta_1)} \gamma(1/\beta_1, (x/\alpha_1)^{\beta_1})\right), \end{aligned} \tag{8}$$

with the mixing proportion p defined by (5).

Denote by τ the (generic) interarrival time with distribution (6). Let distribution F_1 define a random variable (r.v.) Y, and distribution F_2 define a r.v. Z. Then τ can be expressed by the two-component mixture as

$$\tau = IY + (1 - I)Z, \tag{9}$$

where I is the indicator function with $\mathsf{P}(I = 1) = p$ and

$$\begin{aligned} Y &= \min(Y_1, Y_2) \text{ with } Y_1 \sim We(\alpha_1, \beta_1), \;\; Y_2 \sim SGGa(\alpha_2, 1/\beta_2, \beta_2), \\ Z &= \min(Z_1, Z_2) \text{ with } \;\; Z_1 \sim We(\alpha_2, \beta_2), \;\; Z_2 \sim SGGa(\alpha_1, 1/\beta_1, \beta_1), \end{aligned}$$

where the symbol $\sim$ connects a r.v. and its distribution, and $SGGa(\alpha, a, c)$ denotes Stacy's generalized Gamma distribution

$$F(x) = \frac{1}{\Gamma(a)}\gamma(a, (x/\alpha)^c), \quad x \geq 0, \tag{10}$$

(where $a = 1/\beta_i, c = \beta_i, i = 1, 2$), with density function

$$f(x) = \frac{c}{\alpha\Gamma(a)}(x/\alpha)^{ca-1}e^{-(x/\alpha)^c}, \quad x \geq 0,$$

and moments

$$EX = \frac{\alpha\Gamma(a + 1/c)}{\Gamma(a)}, \;\; EX^2 = \frac{\alpha^2\Gamma(a + 2/c)}{\Gamma(a)}.$$

In this setting, the stability condition (3) of approximating system $GI/G/1$ becomes

$$\rho = \frac{\beta_1}{\alpha_1\Gamma(1/\beta_1)}\mathsf{E}S^{(1)} + \frac{\beta_2}{\alpha_2\Gamma(1/\beta_2)}\mathsf{E}S^{(2)} < 1.$$

Below we illustrate the analysis by numerical results based on Monte-Carlo simulation and regenerative approach.

3. Construction of classical regeneration by exponential splitting

In this section we discuss the application of exponential splitting to regenerative simulation of queues with superposed heavy-tailed Weibull inputs. This approach is based on the memoryless property of exponential distribution and the synchronization of a regeneration point of the input process and an empty state of the system. When the component processes have Pareto interevent distributions, this approach has been developed in a recent paper [7].

As far as the authors know, the idea of exponential splitting has been firstly described in [8]. It involves replacing the original r.v. T by a stochastically equivalent r.v. T' (defined on an enlarged probability space). More exactly, an absolutely continuous positive r.v. T with the density g is called *exponentially split* if there exist constants $\eta > 0$ and $\delta \in (0, 1)$ such that

$$g(x) \geq \delta\eta e^{-\eta x}, \quad x \geq 0. \tag{11}$$

Let us define a new r.v. T' as follows:

$$T' = I_T T_0 + (1 - I_T)T_1, \tag{12}$$

where the r.v. T_0 has density $g_0(x) = \eta e^{-\eta x}$, the r.v. T_1 has density

$$g_1(x) = \frac{g(x) - \delta g_0(x)}{1 - \delta}, \tag{13}$$

and I_T is the Bernoulli r.v. (called *splitting indicator*) such that $\mathsf{P}(I_T = 1) = \delta$. If the event $\{I_T = 1\}$ happens, we say that the exponential phase takes place. For Weibull distribution (4), the inequality (11) transforms into

$$\frac{\beta}{\alpha}(x/\alpha)^{\beta-1}e^{-(x/\alpha)^\beta} \geq \delta\eta e^{-\eta x}, \tag{14}$$

which indeed holds under the following conditions connecting the parameters:

$$0 < \delta \le \beta/\alpha, \quad \eta = \alpha^{-\beta},\ 0.5 \le \beta < 1,\ \alpha \ge 1. \tag{15}$$

In our example the r.v. T_1 with the density (13) has distribution function

$$G_1(x) = 1 - \frac{1}{1-\delta} e^{-(x/\alpha)^\beta} + \frac{\delta}{1-\delta} e^{-\eta x},\ \ x \ge 0.$$

From the simulation viewpoint, splitting means that, instead of generating an r.v. T with Weibull distribution, a triple (T_0, T_1, I_T) is generated, where T_0 is exponential with parameter η, T_1 has distribution function G_1 and I_T is the splitting indicator. Moreover, the inequalities (15) ensure that the basic inequality (11) holds. We note that (11) is a particular case of the so-called *minorization condition*, which plays an important role in the theory of general Markov chains, see for instance, [15].

To construct the regeneration points for the waiting time process $W = \{W_n, n \ge 1\}$ in the $\sum_{i=1}^m GI/G/1$ system, we denote by $\{t_k^{(i)}\}$ the arrival instances of the i-th input and by $\{t_k\}$ the arrival points in the superposed input process.

Now we select and fix an arbitrary component input process, denoted further by i_0, with the arrival instances $\{t_k^{(i_0)} \equiv t_k^{(0)},\ k \ge 1\}$. Also define index $n(k)$ as $t_k = t_{n(k)}^{(0)}$. In other words, the k-th arrival in the superposed input (at instant t_k) is indeed the $n(k)$-th class-i_0 arrival. Let indicator function $I_j(t) = 1$ if, at instant t, the j-th component process is in the exponential phase (see decomposition (12)), and $I_j(t) = 0$, otherwise. Now define the following events:

$$\mathcal{E}_k^{(0)} = \{I_j(t_k^{(0)}) = 1,\ j = 1, \dots, m;\ j \ne i_0)\},\ k \ge 1. \tag{16}$$

In other words, the event $\mathcal{E}_k^{(0)}$ means that, at the arrival instant $t_k^{(0)}$ of a class-i_0 customer, the interarrival times of all inputs $j \ne i_0$ have exponential phase. It is easy to check that, on the event $\mathcal{E}_k^{(0)}$, the superposed input process classically regenerates. To illustrate it, we return to the basis model (see Section 2) in which the interarrival time in the i-th renewal input has distribution A_i, and let its exponential phase have parameter λ_i. Then, on the event $\mathcal{E}_k^{(0)}$, the remaining time up to the next event in the superposed process at the arrival epoch $t_k^{(0)}$ of a class-i_0 customer, denoted by $\tau(t_k^{(0)})$, has (tail) distribution

$$\mathsf{P}(\tau(t_k^{(0)}) > x) = (1 - A_{i_0}(x)) \exp\{-x \sum_{j \ne i_0} \lambda_j\},\ \ x \ge 0,$$

which is independent of instant $t_k^{(0)}$. Thus, the regeneration instances $\{\gamma_n\}$ of the superposed input process can be defined as follows:

$$\gamma_0 = 0,\ \ \gamma_{k+1} = \min\{i : 1(t_{n(i)}^{(0)} > \gamma_k) \cdot 1(\mathcal{E}_{n(i)}^{(0)}) = 1\},\ k \ge 0. \tag{17}$$

To construct regenerations of the entire model (that is, the basic processes describing the dynamics of the system) we need one more step. Take the waiting time process $W = \{W_n, n \ge 1\}$ as a basic process. Then it *classically* regenerates at arrival instant of a class-i_0 customer, if i) it is a regeneration instant of the superposed input process and ii) the system is idle at this instant. Formally, these instants can be defined recursively as

$$\beta_0 = 0,\ \beta_{k+1} = \min\{\gamma_i > \beta_k : W_{\gamma_i} = 0\},\ k \ge 0. \tag{18}$$

Then the random distances $\widehat{\Delta}_k := \beta_{k+1} - \beta_k$, $k \ge 1$, are the iid regeneration cycle lengths.

In what follows we simulate and estimate by the regenerative method [16] the stationary performance of the waiting time process W when the superposed input process is composed by two components with heavy-tailed Weibull distributions.

4. Simulation results

The aim of this Section is to investigate through discrete event simulation the two main theoretical contributions of this paper. At first, Section 4.1 deals with the renewal approximation of the original system fed by two independent Weibull flows with different values of β_i (including the case $\beta_i = 1$ that corresponds to Poisson arrivals). Then, in Section 4.2 we focused on the heavy-tailed case (i.e., $\beta_i < 1$) and evaluated the efficiency of the exponential splitting in terms of the frequency of the regenerations for different values of the splitting parameters.

4.1. Approximation by $GI/G/1$

We consider a queueing systems fed by two independent input processes with Weibull interarrival times. In all the simulations described in this subsection 10^{10} arrivals are generated and common (class-independent) exponential service times with rate μ are assumed.

We investigate the accuracy of the approximation through the Kolmogorov distance $d(W, W_A)$ between the empirical distributions of the stationary waiting time process W in the basic system $\sum_{i=1}^{2} We(\alpha_i, \beta_i)/M/1$ and the waiting time process W_A in the approximating system $GI/M/1$ for different values of traffic intensity ρ.

The first set of simulations is carried out with the following values of the parameters: $\alpha_1 = 1$, $\alpha_2 = 4$, $\beta_1 = \beta_2 = 1$, which corresponds to the superposition of two Poisson processes. It is easy to check that in this case the approximating system is an $M/M/1$ system with Poisson input with parameter $1/\alpha_1 + 1/\alpha_2$. We can compare simulation results of the actual waiting time trajectories for the original system W and the waiting time W_A in the corresponding $M/M/1$ queueing system. The absolute error does not exceed 8×10^{-5} for all values of traffic intensity $0.1 < \rho < 0.9$.

The next two sets of simulations investigates the goodness of the approximation in case of light-tailed ($\beta_i > 1$)and heavy-tailed ($\beta_i < 1$) Weibull distributions, respectively.

The comparison between W and W_A (see Tables 1 and 2, where the simulation settings are also reported) includes not only the Kolmogorov distance, but also the values of the average workload process and its variance in the two cases. It is worth noticing that in the first case (see Table 1), the Kolmogorov distance $d(W, W_A)$ does not exceed 5% only for values $\rho < 0.5$, while in presence of heavy-tailed Weibull distributions (see Table 2) the Kolmogorov distance $d(W, W_T)$ is very small for any value of traffic intensity ρ.

$\beta_1 = 5, \beta_2 = 10,\ \alpha_1 = \alpha_2 = 1$					
ρ	$d(W, W_A)$	$Mean(W)$	$Mean(W_A)$	$Var(W)$	$Var(W_A)$
0.1	0.00277	0.00234	0.00260	0.00021	0.00025
0.2	0.01181	0.00948	0.01189	0.00169	0.00237
0.3	0.02499	0.02295	0.03152	0.006107	0.00983
0.4	0.03849	0.04763	0.06805	0.01745	0.03007
0.5	0.05085	0.09289	0.13311	0.04568	0.07991
0.6	0.06189	0.17661	0.24946	0.11651	0.20209
0.7	0.07168	0.33954	0.46978	0.30896	0.52809
0.8	0.08027	0.70258	0.95107	0.95681	1.61544
0.9	0.08808	1.87013	2.47942	4.89730	8.23029

Table 1: *Simulation results* $\beta_1 = 5, \beta_2 = 10,\ \alpha_1 = \alpha_2 = 1$.

$\beta_1 = 0.1, \beta_2 = 0,2, \alpha_1 = \alpha_2 = 1$					
ρ	$d(W, W_A)$	$Mean(W)$	$Mean(W_A)$	$Var(W)$	$Var(W_A)$
0.1	1.909×10^{-5}	95.0969	95.0917	11325.2	11322.9
0.2	8.228×10^{-5}	355.945	355.870	143765	143709
0.3	9.561×10^{-5}	874.066	873.498	827005	825552
0.4	2.952×10^{-5}	1827.71	1827.88	$3.52\times10^{+6}$	$3.52\times10^{+6}$
0.5	11.72×10^{-5}	3572.09	3571.48	$1.32\times10^{+7}$	$1.32\times10^{+7}$
0.6	20.18×10^{-5}	6869.13	6865.69	$4.82\times10^{+7}$	$4.81\times10^{+7}$
0.7	13.88×10^{-5}	13652.6	13652.4	$1.89\times10^{+8}$	$1.89\times10^{+8}$
0.8	69.37×10^{-5}	30214.2	30144.3	$9.19\times10^{+8}$	$9.13\times10^{+8}$
0.9	14.63×10^{-4}	90455.3	90122.8	$8.19\times10^{+9}$	$8.11\times10^{+9}$

Table 2: *Simulation results* $\beta_1 = 0.1, \beta_2 = 0,2,\ \alpha_1 = \alpha_2 = 1$.

4.2. Artificial regeneration by exponential splitting

In this subsection we use several sets of simulation (with 10^9 arrivals) to evaluate the influence of the selected process i_0, splitting parameter δ, exponential parameter η and the traffic intensity ρ, see minorization inequality (11). We focus on the heavy-tailed case, namely, with $\alpha_1 = 1.5$, $\beta_1 = 0.5$, $\alpha_2 = 2$ and $\beta_2 = 2/3$.

As regenerative simulation allows us to calculate confidence intervals, in the following tables we report (for different values of ρ) not only the mean value of the waiting time process W, Mean(W), but also the half-width of the 99% confidence interval, 99%CI, and the number of regeneration cycles, RegW, of the queueing system. It is worth noticing that the number of regenerations in the superposed input (i.e., the number of events (16)), RegIn, does not depend on ρ, while it is proportional to δ.

Exponential splitting, $i_0 = 2$

ρ	$\delta = 0.33$			$\delta = 0.1$		
	RegW	Mean(W)	99%CI	RegW	Mean(W)	99%CI
0.1	5.2×10^7	0.048	1.8×10^{-5}	1.5×10^7	0.048	1.8×10^{-5}
0.2	3.8×10^7	0.184	7.1×10^{-5}	1.1×10^7	0.184	7.1×10^{-5}
0.3	2.7×10^7	0.439	18.5×10^{-5}	8.3×10^6	0.439	18.5×10^{-5}
0.4	1.9×10^7	0.875	41.7×10^{-5}	5.8×10^6	0.875	41.6×10^{-5}
0.5	1.3×10^7	1.604	89.3×10^{-5}	4.0×10^6	1.604	89.2×10^{-5}
0.6	8.9×10^6	2.852	19.4×10^{-4}	2.7×10^6	2.854	19.4×10^{-4}
0.7	5.6×10^6	5.162	45.6×10^{-4}	1.7×10^6	5.162	4.6×10^{-3}
0.8	3.2×10^6	10.139	13.1×10^{-3}	9.6×10^5	10.142	13.2×10^{-3}
0.9	1.4×10^7	25.905	65.6×10^{-3}	4.1×10^5	25.885	64.7×10^{-3}
0.92	1.0×10^6	33.977	10.7×10^{-2}	3.1×10^5	33.927	10.6×10^{-2}
0.95	6.2×10^5	58.202	28.9×10^{-2}	1.8×10^5	58.190	25.3×10^{-2}

Table 3: *Simulation results for* $\eta = 0.8164, \beta_1 = 0.5, \beta_2 = 0.67, \alpha_1 = 1.5, \alpha_2 = 2$.

At first we select the second component of the superposed process as i_0-process, so the

Exponential splitting, $i_0 = 1$

ρ	$\delta = 0.33$			$\delta = 0.1$		
	RegW	Mean(W)	99%CI	RegW	Mean(W)	99%CI
0.1	6.2×10^7	0.048	1.8×10^{-5}	1.85×10^7	0.048	1.8×10^{-5}
0.2	4.6×10^7	0.184	7.1×10^{-5}	1.38×10^7	0.184	7.1×10^{-5}
0.3	3.4×10^7	0.439	18.5×10^{-5}	1.03×10^7	0.439	18.5×10^{-5}
0.4	2.5×10^7	0.875	41.7×10^{-5}	7.63×10^6	0.875	41.7×10^{-5}
0.5	1.8×10^7	1.604	89.2×10^{-5}	5.50×10^6	1.604	89.3×10^{-5}
0.6	1.3×10^7	2.853	19.3×10^{-4}	3.818×10^6	2.855	19.4×10^{-4}
0.7	8.3×10^6	5.161	4.5×10^{-3}	2.49×10^6	5.162	45.6×10^{-4}
0.8	4.8×10^6	10.147	1.3×10^{-2}	1.45×10^6	10.148	13.1×10^{-3}
0.9	2.1×10^6	25.855	6.5×10^{-2}	6.33×10^5	25.904	65.5×10^{-2}
0.92	1.6×10^6	33.956	0.107	4.93×10^5	33.903	10.7×10^{-2}
0.95	9.8×10^5	58.278	0.291	2.96×10^5	58.275	29.4×10^{-2}

Table 4: *Simulation results for* $\eta = 0.63,\ \beta_1 = 0.5, \beta_2 = 0.67, \alpha_1 = 1.5, \alpha_2 = 2.$

exponential splitting is applied to the first flow. From the inequalities (15) it is easy to see that, choosing $\eta = 1/\sqrt{1.5}$, the exponential splitting is possible for $0 < \delta \leq 1/3$. For sake of brevity, the simulation results are reported only for two values of δ.

As highlighted in Table 3, when the traffic intensity ρ increases, RegW decreases, while Mean(W) and 99%CI increase. The value of δ does not significantly affect the reported parameters, apart from RegW. In more detail, for $\delta = 0.33$ RegIn $\approx 7.2 \times 10^7$, while for $\delta = 0.1$ the value of RegIn $\approx 2.16 \times 10^7$ is around ≈ 3.3 times less. Similar consideration may be applied to the values of RegW.

Similar considerations can be drawn when the exponential splitting is applied to the second component (see Table 4). In this case, according to the inequalities (15), we take $\eta = 2^{-2/3}$ and the exponential splitting is still possible for $0 < \delta \leq 1/3$. Note that in this case the input process regenerates more frequently, but the ratio between the number of regenerations and the value of δ remains almost the same (namely, RegIn $\approx 9.3 \times 10^7$ and RegIn $\approx 2.8 \times 10^7$ for $\delta = .33$ and $\delta = 0.1$, respectively).

5. Conclusion

In this paper we analyzed a queueing system fed by the superposition of independent stationary renewal processes with Weibull interarrival time distribution, focusing on two relevant issues: the renewal approximation of the input process and the regenerative simulation of the original queueing system.

In more detail, the accuracy of the renewal approximation was evaluated in terms of the Kolmogorov distance for the corresponding distributions of the stationary workload (the remaining work to be processed). Simulation experiments pointed out that the approximation works perfectly for any value of traffic intensity ρ in the case of heavy-tailed Weibull components and only for $\rho < 0.6$ in the light-tail case. This unexpected result needs further investigation and explanation.

Moreover, the exponential splitting was applied to construct artificial regenerations and estimate the mean stationary workload (with the confidence interval) in the basic (non-regenerative) system. Sufficient conditions for the applicability of the exponential splitting have been derived in the case of heavy-tailed Weibull interarrival time distributions and the theoretical results have

been verified by simulation.

We demonstrated the efficiency of the approach based on the construction of regenerations using exponential splitting technique. It seems to be highly effective for the queueing systems containing heavy-tailed distributions. In this regard the authors hope that this research (as well as the related previous paper [7]) creates a methodological basis for reliable regeneration-based estimation of a broad class of the systems described by heavy-tailed distributions.

Acknowledgments

This work was partially supported by the Moscow Center for Fundamental and Applied Mathematics (recipient E. Morozov).

References

[1] Karl Sigman. "One-Dependent Regenerative Processes and Queues in Continuous Time". In: *Mathematics of Operations Research* 15.1 (1990), pp. 175–189. URL: http://www.jstor.org/stable/3689937 (visited on 06/06/2025).

[2] S. L. Albin. "On Poisson Approximations for Superposition Arrival Processes in Queues". In: *Management Science* 28.2 (Feb. 1982), pp. 126–137. DOI: 10.1287/mnsc.28.2.126.

[3] Ward Whitt. "Approximating a Point Process by a Renewal Process, I: Two Basic Methods". In: *Operations Research* 30.1 (Feb. 1982), pp. 125–147. DOI: 10.1287/opre.30.1.125.

[4] A.Ia. Khinchin. *Mathematical Methods in the Theory of Queueing*. Griffin's Statistical Monographs. London, Griffin, 1960. ISBN: 978-0-02-847880-7.

[5] G. F. Newell. "Approximations for Superposition Arrival Processes in Queues." In: *Management Science* 30.5 (1984), pp. 623–632. (Visited on 06/06/2025).

[6] Søren Asmussen, Hanspeter Schmidli, and Volker Schmidt. "Tail probabilities for non-standard risk and queueing processes with subexponential jumps". In: *Advances in Applied Probability* 31.2 (June 1999), pp. 422–447. DOI: 10.1239/aap/1029955142.

[7] Irina Peshkova, Evsey Morozov, and Michele Pagano. "Regenerative Analysis and Approximation of Queueing Systems with Superposed Input Processes". In: *Mathematics* 12.14 (July 2024), p. 2202. DOI: 10.3390/math12142202.

[8] Alexander Andronov. "Artificial regeneration points for stochastic simulation of complex systems". In: *Simulation Technology: Science and Art. 10th European Simulation Symposium ESS'98*. SCS, Delft, The Netherlands, 1998, pp. 34–40.

[9] Søren Asmussen. *Applied probability and queues*. English. New York: Springer, 2003. ISBN: 978-0-387-00211-8.

[10] François Baccelli and Pierre Brémaud. *Elements of Queueing Theory*. Springer Berlin Heidelberg, 2003. ISBN: 9783662116579. DOI: 10.1007/978-3-662-11657-9.

[11] Evsey Morozov and Bart Steyaert. *Stability Analysis of Regenerative Queueing Models: Mathematical Methods and Applications*. Springer International Publishing, 2021. ISBN: 9783030824389. DOI: 10.1007/978-3-030-82438-9.

[12] R. M. Loynes. "The stability of a queue with non-independent inter-arrival and service times". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 58.3 (July 1962), pp. 497–520. DOI: 10.1017/s0305004100036781.

[13] B. Gnedenko. "Sur La Distribution Limite Du Terme Maximum D'Une Série Aléatoire". In: *Annals of Mathematics* 44.3 (1943), pp. 423–453. URL: http://www.jstor.org/stable/1968974 (visited on 06/06/2025).

[14] Frank Beichelt and Peter Franken. *Zuverlässigkeit und Instandhaltung: mathematische Methoden Instandhaltung, Verfügbarkeit, Zuverlässigkeit*. Vol. 1. Veb Verlag Technik, 1983.

[15] Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Springer London, 1993. ISBN: 9781447132677. DOI: 10.1007/978-1-4471-3267-7.

[16] Søren Asmussen and Peter W. Glynn. *Stochastic simulation: algorithms and analysis*. en. Stochastic modelling and applied probability 57. OCLC: ocn123113652. New York: Springer, 2007. ISBN: 978-0-387-69033-9.

QUEUEING-INVENTORY *K*-OUT-OF-*N* SYSTEM WITH HEAVY TAILS

Arya P S[1,a], Manikandan Rangaswamy[1,b], Alexander Rumyantsev[2,3,c]

•

[1] Department of Mathematics, Central University of Kerala,
Periye, Kasaragod, 671320, Kerala, India
[2] Institute of Applied Mathematical Research, Karelian Research Center of RAS,
Petrozavodsk, Russia
[3]Petrozavodsk State University, Petrozavodsk, Russia

arya.p.s@cukerala.ac.in[a], mani@cukerala.ac.in[b], ar0@krc.karelia.ru[c]

Abstract

In this paper, we study the so-called k-out-of-n queueing-inventory system with a single repair unit, identical elements that are subject to failure, stock of spare elements, and state-dependent replenishment policy. The finite state space Markov chain model is described, and key stationary performance measures are defined. The key focus of this research is on the non-Markov case, in which the random repair and replenishment times may have infinite means, which may affect the positive recurrence of the states of the model. This case is investigated numerically.

Keywords: *k*-out-of-*n* : *G* system, degradation, queueing-inventory model, heavy tails, positive recurrence, generalized semi-Markov processes

1. Introduction

The reliability analysis of *k*-out-of-*n* systems is a topic of considerable interest due to its wide application in various real-world systems, including telecommunications, manufacturing, and power generation. These systems ensure that a system remains operational as long as at least *k* components out of *n* are functioning. The complexity of *k*-out-of-*n* systems arises from the intricate nature of component failures and the repair mechanisms employed to maintain system functionality.

Over time, various models have been developed to analyze and optimize the reliability of such systems, taking into account component failure rates, repair policies, and replenishment strategies. These aspects were considered in a wide range of applications. Failure analysis and statistical modeling were used in enhancing the system resilience of a large-scale hybrid supercomputer with repair and replenishment of components [1]. Reliability evaluation of power systems using multi-state warm standby components and performance-sharing mechanisms was done in [2]. The repair and replenishment are also applied in industrial applications like solar power management at Cochin International Airport [3], where prompt repair or replacement of malfunctioning equipment is critical in maintaining efficiency, as detailed in [4].

In systems with repair, replenishment orders are crucial for maintaining operational availability, especially in COLD, WARM, and HOT systems. A system is classified as COLD when operational components stop deteriorating after failure until the system is restored, as WARM

when operational components continue to deteriorate at a reduced rate after failure, and as HOT when operational components deteriorate at the same rate regardless of system failure. In the preliminary study conducted by [5], the system performance is optimized through the hiring of a repair server upon the failure of N components. This study is further extended in [6], where they investigate a k-out-of-n : G system with an N-policy replenishment, where components are replaced upon failure, leading to optimized long-run state probabilities, and in [7], which explores k-out-of-n systems with the retrial of failed units, focusing on reliability across cold, warm, and hot standby systems. Incorporating an unreliable server and phase-type repair times, [8] extends this analysis under the (N, T) policy. Recently, [9] analyzes a k-out-of-n repairable system with a server offering single or bulk service to failed components. Through a Continuous-time Markov Chain (CTMC) model, the study evaluates system reliability, server performance, and cost functions, demonstrating the advantages of the single/bulk service model over the traditional single-unit service policy, whereas [10] introduces an innovative repair policy for circular consecutive-k-out-of-n : F systems, prioritizing emergency repairs to prevent system failure while managing ordinary repairs in a queue, and provides valuable insights into minimizing repair costs while enhancing system performance.

Additionally, [11] examines a k-out-of-n system with repair, extending service to external customers through a retrial queue, providing insights into system size probabilities and performance metrics. Emphasizing the significance of general repair time distributions on system reliability, [12] provides a comprehensive probabilistic analysis of a k-out-of-n system, where a mathematical model, along with an algorithm to calculate the reliability function and validate their findings through numerical investigations, which pave the way for future sensitivity analyses of reliability characteristics in such systems.

Traditional models, such as Markov processes and fault tree analysis, have been extensively used to assess the reliability of the k-out-of-n systems. However, the models often rely on the assumption of exponential distributions and the memoryless property, which may only sometimes reflect the true complexity of real-world systems. As highlighted in [13], these assumptions can lead to inaccurate reliability estimates if they involve more complex service time distributions or non-exponential failure and repair rates; there is a need for more sophisticated models in real-world systems. To address this limitation, Generalized Semi-Markov Processes (GSMPs) have been introduced. GSMPs extend classical Markov models by allowing arbitrary, non-exponential distributions for the timing of events. These processes provide a flexible framework where states transition based on event clocks, accommodating the modeling of systems with complex, variable timing behaviors [14, 15, 16]. Simulation-based methods offer a robust tool for analyzing the dynamic behavior of k-out-of-n systems, where complex interactions between failure, repair, and replenishment processes are modeled. For example, [17] utilize matrix-based reliability methods for k-out-of-n systems, revealing challenges in analyzing system reliability due to the statistical dependence between component failures. By extending such approaches, it is possible to capture various performance metrics, such as system downtime, repair time, and cost optimization.

Using simulation, [18] explores the asymptotic insensitivity of k-out-of-n systems, while [19] analyzes their steady-state reliability under full repair. Studies such as [20] examine repairable k-out-of-n systems where some components are suspended during repair, and [21] further considers systems with repairmen taking vacations and a shut-off rule for non-failed components. [22] investigates the optimization of inspection and maintenance schedules for k-out-of-n : G warm standby systems, emphasizing the need to activate standby components effectively.

As systems become more intricate, they require advanced models accounting for non-exponential distributions, time-dependent transitions, and interdependency between components. These demands have led to the increasing use of simulation-based approaches, particularly Discrete-Event Simulation (DES), which is highly flexible and capable of accurately modeling complex systems. The DES approach has been effectively applied in fields such as operations research, healthcare, and manufacturing to evaluate processes involving events like failures, repairs, and maintenance activities over time [23, 24]. Additionally, research has explored the implications of external customer services on system reliability. Studies such as [25] and [26]

examine systems where servers handle component repairs and external customer services. These models introduce new layers of complexity, with performance measures such as server idle time utilization and queue management playing a crucial role in overall system reliability.

Building upon these foundational studies, our work focuses on a k-out-of-n : G, WARM system with repair and replenishment orders. We develop a simulation model within the GSMP framework to analyze the system with general distributions for repair times, replenishment orders, and system reliability. The key question in this direction is the positive recurrence of the system states, which, in general, is not guaranteed in advance. We address this question by a numerical study of the simulation model under the heavy-tailed distribution assumptions, which is the main contribution of the present study.

The rest of the paper is arranged as follows: Section 2 provides a detailed description and key performance metrics of the model in the Markovian case, while Section 3 studies the recurrence of the system states (in general case) by constructing a simulation model. The paper ends with a brief conclusion.

2. Model Description

The k-out-of-n : G queueing-inventory system has $n > 0$ identical elements, a single repair server, and a finite inventory for storing *spare* elements. The lifetimes of the *working* elements are independent and identically distributed (iid). Depending on the number of (remaining) working elements, the system can be in *failure state* (if less than $k > 0$ elements are working) or in *active state* (otherwise). When the system is in an active state, the lifetimes of the elements follow an exponential distribution with rate $\lambda > 0$, and in the failure state they follow an exponential distribution with rate $\theta > 0$. The defective elements form a queue and the repair process starts when the number of working elements drops to $L < n$. The system assigns iid repair times following an exponential distribution with rate γ and restores repaired elements to a "good as new" condition. The repair stops when the number of working elements increases to H, $L < H < n$. Apart from repair, restoration of the elements in the system also happens due to replenishment. That is, if the number of working elements reduces to N, $k < N < L$, an order for replenishment is placed. The size of the replenishment order depends on the state of the system in which replenishment occurs:

1. In the active system state, we place a fixed order size of $(n - k + 1)$ elements.
2. In the failure system state, the total number of working elements is restored to n (without spare).

The lead time follows an exponential distribution with rate β in both cases. At the time of replenishment, we discard the failed elements. The excess elements, after materialization, are kept as spares and used to replace defective elements. Thus, the maximal number of elements (including both working and spare) in the system is

$$B_{max} = n + H - k. \tag{1}$$

We define $X(t)$ as the number of elements (working and spare) at time t and let $Y(t)$ represent the restoration state at time t, which takes one of three values,

$$Y(t) = \begin{cases} 0, & \text{when repair and replenishment are OFF.} \\ 1, & \text{when repair is ON, replenishment is OFF,} \\ 2, & \text{when repair is ON, replenishment is ON.} \end{cases}$$

The CTMC $\{X(t), Y(t)\}_{t\geq 0}$ has finite state space

$$\Omega = \{(i,0)|\ L+1 \leq i \leq B_{max}\} \cup \{(i,1)|\ L \leq i \leq H-1\} \cup \{(i,2)|\ 0 \leq i \leq H-1\}.$$

The generator matrix $\boldsymbol{Q}$ is constructed using the following CTMC transition rates:

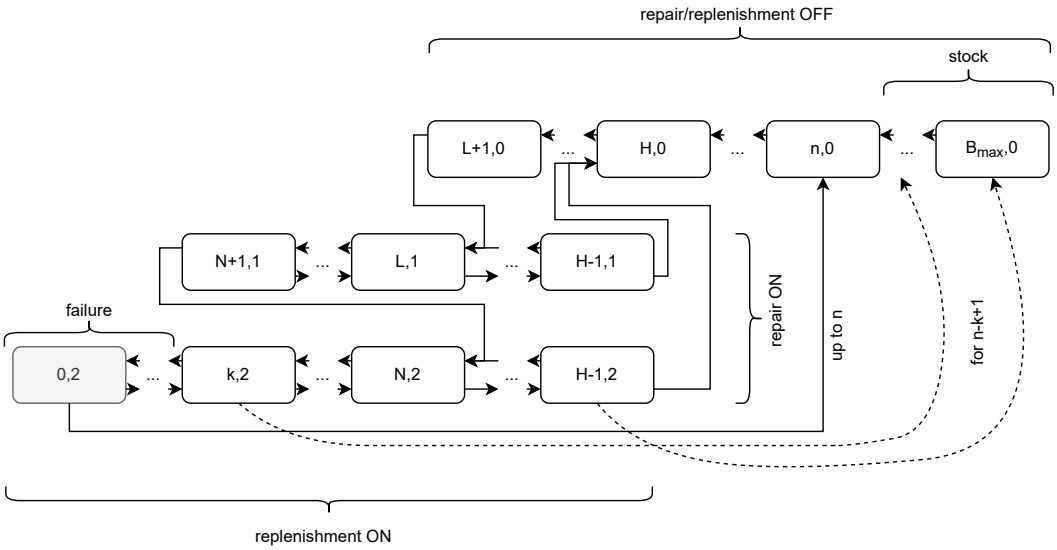


Figure 1: *State transition diagram for the k-out-of-n : G model*

(i) Degradation:

- in active state: $(i,0) \to (i-1,0)$ with rate $\min(i,n)\lambda$, for $i = L+2,\ldots,B_{max}$;
- starting repair: $(L+1,0) \to (L,1)$ with rate $(L+1)\lambda$;
- in repair ON state: $(i,1) \to (i-1,1)$ with rate $i\lambda$, for $i = N+2,\ldots,H-1$;
- placing replenishment order: $(N+1,1) \to (N,2)$ with rate $(N+1)\lambda$;
- in repair ON and replenishment ON: $(i,2) \to (i-1,2)$ with rate $i\lambda$, for $i = k,\ldots,H-1$;
- in failure state: $(i,2) \to (i-1,2)$ with rate $i\theta$, for $i = 0,\ldots,k-1$.

(ii) Repair:

- if replenishment ON: $(i,2) \to (i+1,2)$ with rate γ, for $i = 0,\ldots,H-2$;
- if only repair is ON: $(i,1) \to (i+1,1)$ with rate γ, for $i = N+1,\ldots,H-2$;
- when switching repair OFF: $(H-1,j) \to (H,0)$ with rate γ, for $j = 1,2$.

(iii) Replenishment:

- in active state: $(i,2) \to (i+n-k+1,0)$ with rate β, for $i = k,\ldots,H-1$;
- in failure state: $(i,2) \to (n,0)$ with rate β, for $i = 0,\ldots,k-1$.

All other transition rates (and corresponding off-diagonal elements of the generator matrix) are zero. The diagonal elements are negative to guarantee $\boldsymbol{Q}\mathbf{1} = \mathbf{0}$. The CTMC is irreducible and aperiodic and has unique steady-state stochastic vector $\boldsymbol{\pi} = ||\pi_{i,j}||_{(i,j)\in\Omega}$ solving the linear system

$$\boldsymbol{\pi}\boldsymbol{Q} = \mathbf{0}, \quad \boldsymbol{\pi}\mathbf{1} = 1. \tag{2}$$

Using this steady-state distribution, we can illustrate how to derive some critical performance and reliability measures of the system in the Markov case.

1. Expected number of working elements when the repair is ON, E_{RON}.

$$E_{RON} = \sum_{i=N+1}^{H-1} i\pi_{i,1} + \sum_{i=0}^{H-1} i\pi_{i,2}. \tag{3}$$

2. Probability that the server is busy, P_B.

$$P_B = \sum_{i=N+1}^{H-1} \pi_{i,1} + \sum_{i=0}^{H-1} \pi_{i,2}. \tag{4}$$

We can find the probability that the server is idle as $P_I = 1 - P_B$.

3. Probability that the system is in the failed state, P_F.

$$P_F = \sum_{i=0}^{k-1} \pi_{i,2}. \tag{5}$$

4. Effective replenishment rate, E_{RR}.

$$E_{RR} = \beta \sum_{i=0}^{H-1} \pi_{i,2}. \tag{6}$$

Other performance/reliability metrics can be defined in a similar way.

Remark 1. Note that the steady-state distribution allows one to obtain the expected cycle length of various cycles that appear as the times at which the CTMC revisits some fixed state. In particular, the average cycle length $ET_{i,j}$, where $T_{i,j}$ is the (generic) time between the consecutive visits to the state $(i,j) \in \Omega$ of the CTMC is calculated as

$$ET_{i,j} = \left[-Q_{(i,j),(i,j)}\pi_{i,j}\right]^{-1}, \tag{7}$$

where the diagonal element $-Q_{(i,j),(i,j)}$ is the outgoing rate from state (i,j) and $\pi_{i,j}$ is the steady-state probability of that state.

3. Simulation Model

In order to study the sensitivity of the k-out-of-n : G model to the repair/replenishment distributions, we construct a simulation model using the DES approach within the framework of the so-called GSMPs [16]. This framework allows one to construct confidence intervals [27] for the desired performance estimates using regenerative simulation [28]. We note that historically, GSMP construction dates back to the 1970s and has several variations known as *service scheme* [29] or reallocatable GSMP [30].

To describe a GSMP, one needs to define a multi-dimensional stochastic process

$$\Theta = \{\boldsymbol{X}(t), \boldsymbol{T}(t)\}_{t\geq 0}, \tag{8}$$

which describes the system *state* vector $\boldsymbol{X}(t) \in \mathcal{X}$ evolving in time. The simulation time advances according to the *timers* (clocks) vector $\boldsymbol{T}(t) \geq \boldsymbol{0}$ among which the so-called *active timers* enumerated from the set $A(\boldsymbol{x}) \neq \varnothing$ (for any state $\boldsymbol{x} \in \mathcal{X}$) decrease linearly with rates given by a (state-dependent) vector $\boldsymbol{r}(\boldsymbol{x}) \geq \boldsymbol{0}$, positive for active timers:

$$r_i(\boldsymbol{x}) > 0 \text{ if and only if } i \in A(\boldsymbol{x}).$$

More formally,

$$\boldsymbol{T}(t+h) = \boldsymbol{T}(t) - h\boldsymbol{r}(\boldsymbol{X}(t)), \quad h \leq \min_{i\in A(\boldsymbol{X}(t))} T_i(t)/r_i(\boldsymbol{X}(t)).$$

At *an event* of type i, that is., a time epoch $t \geq 0$ such that $T_i(t-) = 0$, $i \in A(\boldsymbol{X}(t))$ (assuming events appearing singly), the state vector $\boldsymbol{X}$ transitions from some state $\boldsymbol{x} \in \mathcal{X}$ into $\boldsymbol{x}' \in \mathcal{X}$ in accordance with transition probability matrix $\boldsymbol{P}^{(i)} = ||P^{(i)}_{\boldsymbol{x},\boldsymbol{x}'}||_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}}$, where

$$P^{(i)}_{\boldsymbol{x},\boldsymbol{x}'} = \mathrm{P}\{\boldsymbol{X}(t) = \boldsymbol{x}' | \boldsymbol{X}(t-) = \boldsymbol{x}, T_i(t-) = 0\}. \tag{9}$$

After such a transition, the set of active timers may also change from $A(\boldsymbol{x})$ into $A(\boldsymbol{x}')$. Thus, the *new active* timers $i' \in A(\boldsymbol{x}') \setminus (A(\boldsymbol{x}) \setminus \{i\})$ are initialized from some given density

$$f_{i'}(u, \boldsymbol{x}, \boldsymbol{x}', i) = \mathrm{P}\{T_{i'}(t) \in du | \boldsymbol{X}(t-) = \boldsymbol{x}, \boldsymbol{X}(t) = \boldsymbol{x}', T_i(t-) = 0\}, \quad u > 0. \tag{10}$$

Since the state does not change between the events, tracking the system only at event epochs (standard for the DES) is necessary.

In this paper, we construct a GSMP with *simple* events [31], which means that the timer initialization density function depends only on the timer itself,

$$f_{i'}(u, \boldsymbol{x}, \boldsymbol{x}', i) \equiv f_{i'}(u).$$

This approach adds some complexity to the state space description, although it still keeps the model within the class of finite state space models. Additionally, note that the timer initialization densities are general, no longer restricting the model to exponential distributions.

The components of the state vector have the following meaning (according to the number of the component):

$1, \ldots, n$: an indicator of the state (working/broken) of the elements in active state (since there are at most n working elements);

$n+1, \ldots, n+k-1$: an indicator of the state (working/broken) of the elements in a failure state (since there can be at most $k-1$ working elements);

$n+k$: inventory state (number of spare elements);

$n+k+1$: indicator of the restoration state (0 – repair/replenishment OFF; 1 – repair ON, replenishment OFF; 2 – repair/replenishment ON).

Accordingly, the timers have the following functions:

$1, \ldots, n$: residual lifetimes of the elements in active state;

$n+1, \ldots, n+k-1$: residual lifetimes of the elements in failure state;

$n+k$: residual repair time;

$n+k+1$: residual replenishment time.

That is, the system state vector $\boldsymbol{X}$ and timers vector $\boldsymbol{T}$ have the same dimension $n+k+1$, where n and k are the model parameters (maximum number of working elements and failure threshold, respectively).

The timer initialization densities $f_{i'}(u)$ are then chosen according to the given p.d.f. for the corresponding event, that is, the life/repair/replenishment times. We assign them for the specific experiments separately.

A technical *restriction* of the model is that at the system state change (active/failure), the timers are initialized anew in a memoryless way. This assumption is due to the fact that the lifetime distribution is affected by the system state (elements' lifetimes have different distributions in active and failure states). Although it is possible to use a conditional distribution for the residual lifetime based on the already attained lifetime, we omit this possibility hereafter.

An important result for the finite state space GSMP models is the recurrence condition of the system states. As demonstrated in [31], if two or more clocks have clock setting distributions with an infinite mean, the recurrence of all states is not guaranteed. In contrast, all states are positive recurrent (that is., the hitting times of states have finite expectations regardless of the initial state) if the initialized clocks have finite means [31, Proposition 3.1]. To study this effect in our model, we perform a numerical experiment.

We define a GSMP model for the k-out-of-n : G system so that one-timer has a distribution with an infinite mean. To do so, we fix $n = 10$, $H = 8$, $L = 6$, $N = 4$, $k = 2$ and take all the timers

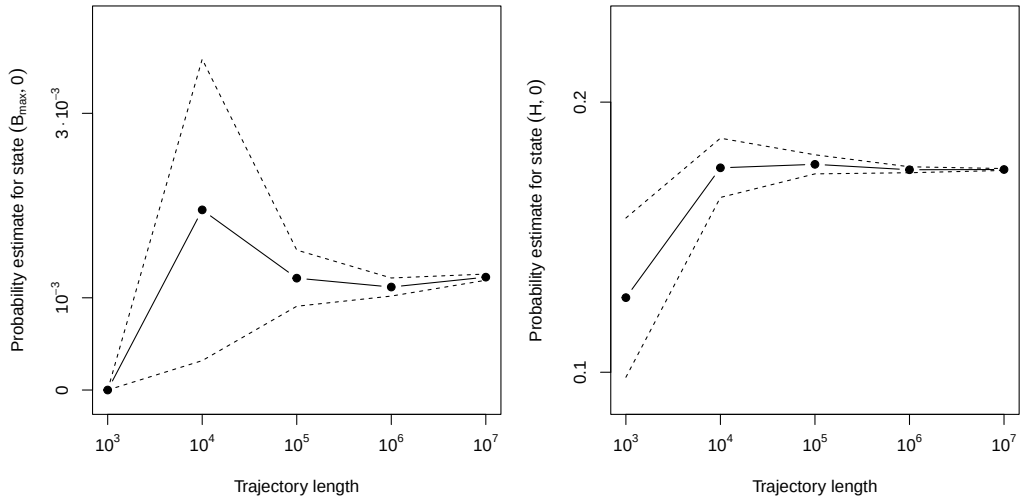


Figure 2: *Regenerative estimates for the stationary probabilities of states* $(B_{max}, 0)$ *(left) and* $(H, 0)$ *(right) in a GSMP k-out-of-n : G model, at the confidence level* 0.05*, for given trajectory length* 10^i *for* $i = 3, \ldots, 7$*. The repair times have heavy-tailed (Pareto) distribution* (11) *with* $\alpha = 0.8$ *and, consequently, infinite mean. The estimates obtained by* `simulato` *package.*

to be exponentially distributed, apart from one being (type-II) Pareto distributed with shape parameter $\alpha = 0.8$, that is, having tail distribution function $\overline{F}(x) = 1 - F(x)$ of the following form,

$$\overline{F}(x) = x_0^{\alpha}(x + x_0)^{-\alpha}. \tag{11}$$

Namely, the system assigns exponentially distributed lifetimes to elements in the working and failure states, with rates $\lambda = 1$ and $\theta = 0.1$, respectively. Repair or replenishment times (for the corresponding cases) are also exponentially distributed with rates $\gamma = \beta = 10$. We consider two cases: one where the residual repair timer T_{n+k} has a heavy-tailed distribution with an infinite mean, and another where the residual replenishment timer T_{n+k+1} has such a distribution. Using the built-in capabilities of the `simulato` package[32] for `R` language, we look at the regenerative estimates of the stationary probabilities of the state $(H, 0)$ (in the original system) that corresponds to H working elements and no ongoing repair/replenishment, and the state $(B_{max}, 0)$. Note that the latter is only accessible through replenishment from the state $(H - 1, 2)$, and thus, one would expect this state to have infinite recurrence time if the replenishment timer has an infinite mean. We check this observation by plotting the estimate of such a probability with a growing number of transitions within the GSMP model.

Before doing so, we note that for validation purposes, we compared the accuracy of the GSMP model to the CTMC in case all timers have exponential distributions. The relative accuracy of the estimates did not exceed 0.4% for the GSMP trajectory length 10^7. For the aforementioned model configuration, in the CTMC model, the states $(H, 0)$ and $(B_{max}, 0)$ had probabilities 0.1838 and 0.0013, respectively.

Now we analyze the GSMP model with the *replenishment* timer having heavy-tailed distribution and infinite mean. Figure 2 shows that the estimate for the state $(B_{max}, 0)$ is relatively small but positive. The state $(H, 0)$ has a positive steady-state probability. In both cases, the confidence interval decreases with increasing trajectory length. Note also that the values of the corresponding estimates are rather close to the purely exponential CTMC case mentioned above.

Figure 3 depicts the results for the system with a heavy-tailed *repair* timer distribution (having an infinite mean). As expected, the state $(H, 0)$ has a non-zero steady-state probability estimate with a decreasing confidence interval. The state $(B_{max}, 0)$ estimate is also positive but remains relatively small, falling below the simulation accuracy. This observation is somewhat counterintuitive. However, a closer look at the diagram of the possible model transitions given in Figure 1 can invoke a possible explanation for this phenomenon. In fact, the only possibility of

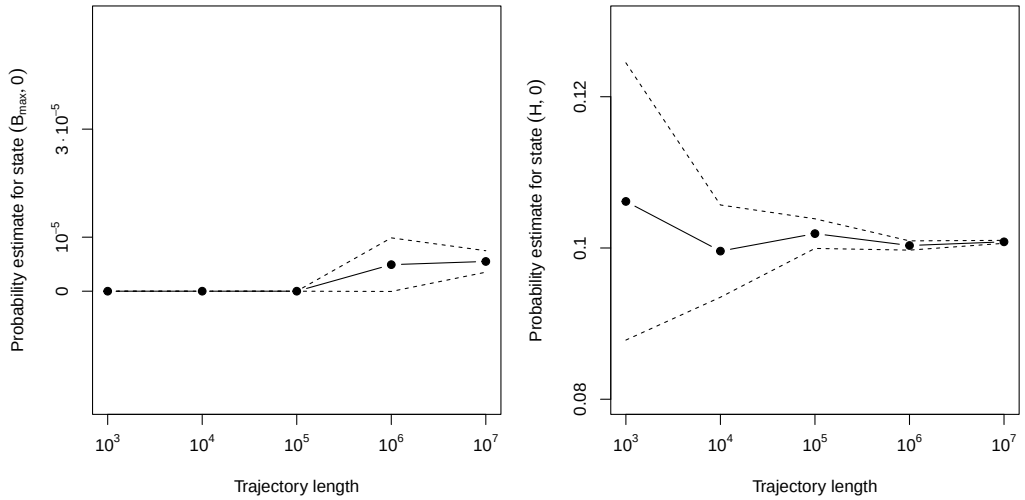


Figure 3: *Regenerative estimates for the stationary probabilities of states* $(B_{max}, 0)$ *(left) and* $(H, 0)$ *(right) in a GSMP k-out-of-n* : *G model, at the confidence level* 0.05*, for given trajectory length* 10^i *for* $i = 3, \ldots, 7$*. The repair times have heavy-tailed (Pareto) distribution* (11) *with* $\alpha = 0.8$ *and, consequently, infinite mean.*

entering the state $(B_{max}, 0)$ is through the state $(H-1, 2)$. The latter is only reachable from states with fewer working elements due to repair. Thus, if the repair time has infinite mean, the state $(B_{max}, 0)$ may be suspected to be null recurrent.

Now we consider a system in which *both* the repair and replenishment timers are Pareto distributed. In this case, the model can enter the state $(B_{max}, 0)$ only through at least two sequential events: repair to enter $(H-1, 2)$ and replenish to enter $(B_{max}, 0)$. If both timers have an infinite mean, it is more likely that the degradation event interrupts this chain of two heavy-tailed events, and following the arguments in [31], one may suspect the state $(B_{max}, 0)$ to be transient. To observe this phenomenon, we plot the stationary probability estimate of the state $(B_{max}, 0)$ for the cases when both timers have Pareto distribution (11) with $\alpha = 0.8$ vs. $\alpha = 1.2$. While in the latter case, this state has a small positive probability (having an order of the simulation accuracy), the probability estimate is negligible in the former case. We extend this investigation by plotting the probability estimates of the number of working elements for two trajectory lengths, 10^4 and 10^7. Figure 5 shows that most states appear transient in the model with $\alpha = 0.8$, except for the zero states where no elements are working. We also see that the estimate of the probability of the zero state increases with increasing trajectory length, that is., 0.6795 for 10^4 and 0.9505 for 10^7 transitions.

Conclusion

In this paper, we consider the so-called *k*-out-of-*n* : *G* model with a single repair element, stock of spare elements, state-dependent replenishment policy, and (threshold-based) state-dependent element degradation. While the Markov case does not display any difficulties, since the model is a finite state space Markov chain, in the general case positive recurrence of the states is questionable. We investigated the general model by simulation and demonstrated the effects of heavy tails (for repair and replenishment time distributions) on the probabilities of the system states. As we observed from the results of numerical experiments, the model seems to possess null-recurrence and transience for some states, which is similar to the effects reported in [31]. A rigorous proof of those effects is a promising direction for further research.

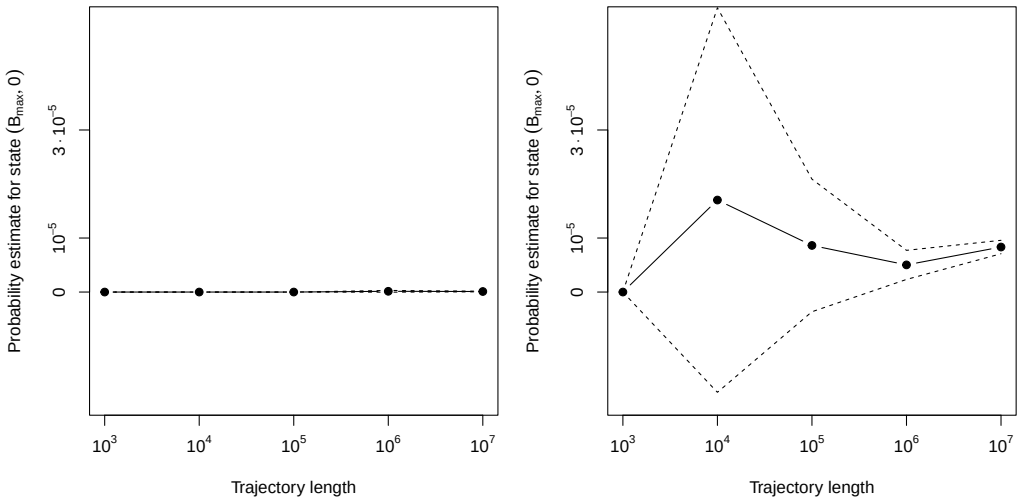


Figure 4: *Regenerative estimates for the stationary probabilities of states* $(B_{max}, 0)$ *in a GSMP k-out-of-n* : *G model, at the confidence level* 0.05*, for given trajectory length* 10^i *for* $i = 3, \ldots, 7$*. Both the repair and replenishment times have heavy-tailed (Pareto) distribution* (11) *with infinite mean for* $\alpha = 0.8$ *(left) and finite mean for* $\alpha = 1.2$ *(right).*

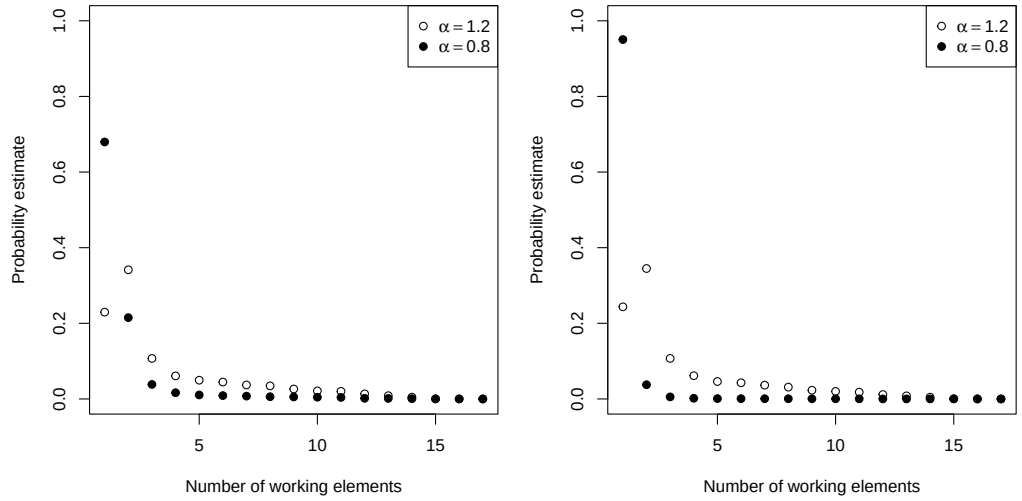


Figure 5: *Relative frequencies vs. number of working elements in the GSMP k-out-of-n* : *G model for trajectory length* 10^7*. Both the repair and replenishment times have heavy-tailed (Pareto) distribution* (11) *with infinite mean for* $\alpha = 0.8$ *(black dots) and finite mean for* $\alpha = 1.2$ *(white dots) for simulation trajectory of length* 10^4 *(left) and* 10^7 *(right).*

Acknowledgments

This work was partially supported by the Moscow Center for Fundamental and Applied Mathematics (recipient A. Rumyantsev).

References

[1] Elvis Rojas et al. "Towards a Model to Estimate the Reliability of Large-Scale Hybrid Supercomputers". In: *Euro-Par 2020: Parallel Processing*. Springer International Publishing, 2020, pp. 37–51. ISBN: 9783030576752. DOI: 10.1007/978-3-030-57675-2_3.

[2] Heping Jia et al. "Reliability evaluation of power systems with multi-state warm standby and multi-state performance sharing mechanism". In: *Reliability Engineering & System Safety* 204 (Dec. 2020), p. 107139. ISSN: 0951-8320. DOI: 10.1016/j.ress.2020.107139.

[3] *Cochin International Airport Ltd. CIAL's green energy generation touches 25 Cr. Units*. URL: https://www.cial.aero/news-Updates/CIAL-s-green-energy.

[4] Sreenath Sukumaran and K. Sudhakar. "Fully solar powered airport: A case study of Cochin International airport". In: *Journal of Air Transport Management* 62 (July 2017), pp. 176–188. ISSN: 0969-6997. DOI: 10.1016/j.jairtraman.2017.04.004.

[5] A. Krishnamoorthy, P. Ushakumari, and B. Lakshmy. "k-out-of-n-system with repair: The N-policy". In: *Asia-Pacific Journal of Operational Research* 19 (May 2002), pp. 47–61.

[6] Achyutha Krishnamoorthy. "A Reliability-Inventory Problem Under N-policy of replenishment of component". In: (2021). DOI: 10.24412/1932-2321-2021-465-73-87.

[7] A. Krishnamoorthy and P. V. Ushakumari. "Reliability of ak-out-of-n system with repair and retrial of failed units". In: *Top* 7.2 (Dec. 1999), pp. 293–304. ISSN: 1863-8279. DOI: 10.1007/bf02564728.

[8] Srinivas R. Chakravarthy, A. Krishnamoorthy, and P. V. Ushakumari. "A k-out-of-n reliability system with an unreliable server and phase type repairs and services: the (N, T) policy". In: *International Journal of Stochastic Analysis* 14.4 (Jan. 2001), pp. 361–380. ISSN: 2090-3340. DOI: 10.1155/s1048953301000326.

[9] P V Ushakumari. "k-out-of-n System with Repair under Single/Bulk Service". en. In: *Degres Journal* 9.7 (2024), pp. 27–36.

[10] Shan Gao, Jinting Wang, and Qin Chen. "Reliability Evaluation for a Circular Con/k/n:F System With a Novel Differential Repair Policy". In: *IEEE Transactions on Reliability* (2025), pp. 1–14. ISSN: 1558-1721. DOI: 10.1109/tr.2024.3524329.

[11] A. Krishnamoorthy, Vishwanath C. Narayanan, and T. G. Deepak. "Optimal utilization of service facility for ak-out-of-n system with repair by extending service to external customers in a retrial queue". In: *Journal of Applied Mathematics and Computing* 25.1–2 (Sept. 2007), pp. 389–405. ISSN: 1865-2085. DOI: 10.1007/bf02832364.

[12] Vladimir Rykov et al. "On reliability function of a k-out-of-n system with general repair time distribution". In: *Probability in the Engineering and Informational Sciences* 35.4 (May 2020), pp. 885–902. ISSN: 1469-8951. DOI: 10.1017/s0269964820000285.

[13] Ameneh Farahani, Ahmad Shoja, and Hamid Tohidi. "Markov and semi-Markov models in system reliability". In: *Engineering Reliability and Risk Assessment*. Elsevier, 2023, pp. 91–130. ISBN: 9780323919432. DOI: 10.1016/b978-0-323-91943-2.00010-1.

[14] Oliver C. Ibe. *Markov processes for stochastic modeling*. eng. 2nd ed. Elsevier insights. London: Elsevier, 2013. ISBN: 978-0-12-407839-0.

[15] Peter J Haas. *Stochastic Petri nets modelling, stability, simulation*. English. New York: Springer, 2002. ISBN: 978-0-387-21552-5. (Visited on 08/17/2014).

[16] P. W. Glynn. "A GSMP formalism for discrete event systems". In: *Proceedings of the IEEE* 77.1 (1989), pp. 14–23. DOI: 10.1109/5.21067.

[17] Ji-Eun Byun, Hee-Min Noh, and Junho Song. "Reliability growth analysis of k-out-of-N systems using matrix-based system reliability method". In: *Reliability Engineering & System Safety* 165 (Sept. 2017), pp. 410–421. ISSN: 0951-8320. DOI: 10.1016/j.ress.2017.05.001.

[18] Nika Ivanova. "Modeling and Simulation of Reliability Function of a k-out-of-n:F System". In: *Distributed Computer and Communication Networks: Control, Computation, Communications*. Springer International Publishing, 2020, pp. 271–285. ISBN: 9783030662424. DOI: 10.1007/978-3-030-66242-4_22.

[19] N. M. Ivanova. "On Steady State Reliability and Sensitivity Analysis of a k-out-of-n System Under Full Repair Scenario". In: *Distributed Computer and Communication Networks: Control, Computation, Communications*. Springer Nature Switzerland, 2022, pp. 422–434. ISBN: 9783031232077. DOI: 10.1007/978-3-031-23207-7_33.

[20] Xiaohu Li, Ming J. Zuo, and Richard C.M. Yam. "Reliability analysis of a repairable k-out-of-n system with some components being suspended when the system is down". In: *Reliability Engineering & System Safety* 91.3 (Mar. 2006), pp. 305–310. ISSN: 0951-8320. DOI: 10.1016/j.ress.2005.01.010.

[21] Yuanyuan Zhang, Wenqing Wu, and Yinghui Tang. "Analysis of an k-out-of-n:G system with repairman's single vacation and shut off rule". In: *Operations Research Perspectives* 4 (2017), pp. 29–38. ISSN: 2214-7160. DOI: 10.1016/j.orp.2017.02.002.

[22] Hui Wu, Yan-Fu Li, and Christophe Bérenguer. "Optimal inspection and maintenance for a repairable k-out-of-n: G warm standby system". In: *Reliability Engineering & System Safety* 193 (Jan. 2020), p. 106669. ISSN: 0951-8320. DOI: 10.1016/j.ress.2019.106669.

[23] J. Banks. *Discrete-event System Simulation*. Prentice Hall, 2010. ISBN: 9780136062127. URL: https://books.google.ru/books?id=cqSNnmrqqbQC.

[24] M. Averill Law and W.D. Kelton. *Simulation Modeling and Analysis*. 4th ed. New York: McGraw Hill, 2007.

[25] A. N. Dudin, A. Krishnamoorthy, and V. C. Narayanan. "Idle time utilization through service to customers in a retrial queue maintaining high system reliability*". In: *Journal of Mathematical Sciences* 191.4 (May 2013), pp. 506–517. ISSN: 1573-8795. DOI: 10.1007/s10958-013-1336-3.

[26] A. Krishnamoorthy, V. C. Narayanan, and T. G. Deepak. "Reliability of a k-out-of-n system with repair by a service station attending a queue with postponed work". In: *International Journal of Reliability, Quality and Safety Engineering* 14.04 (Aug. 2007), pp. 379–398. ISSN: 1793-6446. DOI: 10.1142/s0218539307002702.

[27] Shane G. Henderson and Peter W. Glynn. "Regenerative steady-state simulation of discrete-event systems". en. In: *ACM Transactions on Modeling and Computer Simulation* 11.4 (Oct. 2001), pp. 313–345. ISSN: 10493301. DOI: 10.1145/508366.508367. (Visited on 04/19/2018).

[28] Søren Asmussen and Peter W. Glynn. *Stochastic simulation: algorithms and analysis*. en. Stochastic modelling and applied probability 57. OCLC: ocn123113652. New York: Springer, 2007. ISBN: 978-0-387-69033-9.

[29] D. König, V. V. Rykov, and V. Schmidt. "Stationary queuing systems with dependencies". en. In: *Journal of Soviet Mathematics* 21.6 (1983), pp. 938–994. ISSN: 0090-4104, 1573-8795. DOI: 10.1007/BF01089194. (Visited on 11/15/2018).

[30] Masakiyo Miyazawa. "Insensitivity and Product-Form Decomposability of Reallocatable GSMP". In: *Advances in Applied Probability* 25.2 (1993). Publisher: Applied Probability Trust, pp. 415–437. ISSN: 00018678. DOI: 10/d5fqss. (Visited on 03/26/2020).

[31] Peter W. Glynn and Peter J. Haas. "On Transience and Recurrence in Irreducible Finite-State Stochastic Systems". en. In: *ACM Transactions on Modeling and Computer Simulation* 25.4 (May 2015), pp. 1–19. ISSN: 10493301. DOI: 10/f767qh. (Visited on 03/26/2020).

[32] A. Rumyantsev and S. Astafiev. *simulato: Simulation framework for GSM processes*. URL: `https://R-Forge.R-project.org/projects/simulato/`.

EXACT SAMPLING FOR HETEROGENEOUS MULTISERVER JOB MODEL

Alexander S. Golovin

•

Institute of Applied Mathematical Research, KarRC RAS, Russia,
Petrozavodsk State University, Russia,
golovin@krc.karelia.ru

Abstract

The paper presents an approach to the simulation of a multi-server queueing system, known as the multi-server job model, where the (random number of) servers are seized/released by a customer simultaneously. This model is widely used in scenarios like cloud computing and parallel processing. Due to the inherent difficulty in obtaining analytical solutions for such systems, we adopt the exact sampling technique to generate the steady-state workload samples accurately. This technique allows one to obtain unbiased estimates of the steady-state performance, such as the per-class waiting times of customers.

Keywords: Exact sampling, dominated coupling from the past, multiserver model with simultaneous service, modified Kiefer-Wolfowitz workload vector.

1. Introduction

Since 1965, the multiserver job model (MJM) has been investigated, in which a customer may need simultaneous service on one or more servers [1]. Such models naturally arise in many applications including telecommunication, computer and emergency systems. In particular, a communication system with multiple-address messages, where a message is transmitted simultaneously to many receivers, was treated in [2]. Computer systems with competition for storage space and programs having diverse storage size and time requirements were considered in the paper [3] using a loss MJM, where jobs which cannot immediately receive service are dropped. Analytical results were provided for M/M/2-type model in papers [4, 5]. In [6], using a matrix-analytical approach, a criterion was obtained for the stability of M/M/c-type model. Much of the recent research on MJM is motivated by the need to understand and optimize the performance of data centers and cloud computing infrastructure, see, e.g. [7, 8, 9]. The research directions in MJM in recent years are described in sufficient detail in papers [10, 11]. While extensive literature exists on the MJM, the general case characterized by an arbitrary service time distribution and more than two servers remains largely unexplored, necessitating further investigation.

When dealing with complex queueing systems, often with non-standard arrival or service processes, analytical derivations of the performance characteristics can become intractable or impossible. *Exact sampling* (perfect simulation) techniques [12] provide powerful simulation-based alternative. Those allow one to obtain samples directly from the steady-state distribution, avoiding the need to wait for sufficiently long "warm-up" period during simulation.

This paper is dedicated to adopting the Dominated Coupling From the Past (DCFTP) exact sampling technique to the MJM. In [13] exact sampling technique was applied to MJM, to the best of our knowledge, for the first time, where the cases considered included heavy-tailed (Pareto), phase-type and exponential service time distributions. However, the regenerative exact simulation

approach suggested in [13] has serious technical issues with (on average) infinite number of iterations while there is no such drawback in the approach proposed in the present paper.

We develop our approach on the basis of [14]. In the paper [14] the DCFTP algorithm [15] was adopted to sampling from the stationary customer delay in First-Come-First-Served (FCFS) M/G/c queue in the "super-stable" case, when the system load $\rho < 1$ (we call this scheme *the original algorithm* below, for short). Note that this condition on the load is heavily restrictive for the classical multiserver queue, since in general the stability requires $\rho < c$ only. At the same time, as shown in the present paper, the condition $\rho < 1$ is reasonable for the MJM case, which advocates the applicability of the proposed approach. So, our key contribution is a novel adaptation of the DCFTP algorithm to *heterogeneous* MJM which, in contrast to [13], guarantees finite average runtime.

The structure of the paper is as follows. In Section 2, we discuss the original approach to exact simulation [14] of the stationary customer delay in the M/G/c-FCFS queue based on the general DCFTP method where the single-server M/G/1 queue under the processor sharing (PS) discipline serves as an upper bound. Description of the MJM and modifications (for class-independent and class-dependent service times) of the original algorithm are given in Section 3. In Section 4, we present the results of numerical experiments and validation, which, in particular, include the comparison with an exact result obtained in [4] for the M/M/2-type model. We end up with a conclusion discussing possible further research directions.

2. Original Algorithm

This section is devoted to a detailed description of the *original algorithm* introduced in [14].

Consider a stable M/G/c-FCFS system Σ with arrival rate λ, Poisson arrival times $\{t_n : n \geq 0\}$ ($t_0 := 0$). Let the independent and identically distributed (iid) service times $\{S_n\}$ have distribution function (d.f.) $G(x) = P(S \leq x)$, $x \geq 0$, and finite mean $ES = 1/\mu$. The workload vector $\boldsymbol{V}(t) = (V^{(1)}(t), V^{(2)}(t), \ldots, V^{(c)}(t))$, $t \geq 0$, contains the residual work at the corresponding servers at time t, in ascending order, $V^{(1)}(t) \leq V^{(2)}(t) \leq \cdots \leq V^{(c)}(t)$. The value $\boldsymbol{V}(t)$ just before the arrival epoch is known as the Kiefer-Wolfowitz vector, $\boldsymbol{W}_n = \boldsymbol{V}(t_n-)$, and satisfies the recursion

$$\boldsymbol{W}_{n+1} = R\big(\boldsymbol{W}_n + S_n\boldsymbol{e} - T_n\mathbf{1}\big)^+ \tag{1}$$

where $T_n = t_{n+1} - t_n$ are iid interarrival times, $\boldsymbol{W}_n = (W_n^{(1)}, \ldots, W_n^{(c)})$, $\boldsymbol{e} = (1, 0, \ldots, 0)$, $\mathbf{1} = (1, 1, \ldots, 1)$, R places the coordinates in ascending order, and $(\cdot)^+$ denotes the positive part of each coordinate. If $\rho := \lambda/\mu < c$, with $n \to \infty$, the vector $\boldsymbol{W}_n$ converges in distribution to the stationary workload vector $\boldsymbol{W}$ having distribution which coincides with the time-stationary limiting distribution of $\boldsymbol{V}(t)$ as $t \to \infty$ by the property PASTA.

The original algorithm for sampling exactly from the stationary workload vector $\boldsymbol{W}$ is proposed in [14] for the case

$$\rho < 1. \tag{2}$$

In this case, the system Σ is called super-stable. The algorithm is based on the following stochastic comparison result between the workload in the system Σ and the system M/G/1 operating under the processor sharing (PS) discipline (we denote it by $\hat{\Sigma}$). It is known (see e.g. [14]) that the workload in M/G/1-FCFS is an upper bound for the summary workload in Σ. Moreover, the workload in the system M/G/1 is invariant under changes of work-conserving disciplines [16], that is, workload has exactly the same sample paths under PS as it does under FCFS. Thus, the $\hat{\Sigma}$ serves as the upper bound of Σ [14]. The choice of PS discipline in the single-server queue is motivated by the fact that in this case the workload process turns out to be time reversible [17]. Note that under condition (2), the system $\hat{\Sigma}$ is stable. To apply the algorithm, it is necessary to have the steady-state distribution explicitly available, or we should have the possibility to draw from the statistical equilibrium. To ensure the finite completion time of the algorithm, we assume that $ES^2 < \infty$ [14] (hereafter, unless stated otherwise, we omit the index for a typical member of the iid sequence).

The algorithm can be briefly described in the following steps:

1. From the stationary state, simulate the system $\hat{\Sigma}$ backward in time until it empties.
2. Invert the time axis and using the data of customer departure epochs (as arrivals) and corresponding work amounts from the trajectory of $\hat{\Sigma}$, construct the workload vectors in Σ by recursion (1).

The original approach is summarized in Algorithm 1. To simplify its comprehension, let's clarify some points. In the first step of the algorithm, we must start the workload process in $\hat{\Sigma}$ from the steady state. The stationary workload for the M/G/1 queue can be constructed from the Pollaczek-Khintchine formula as follows,

$$\sum_{i=1}^{Q} S_e^{(i)}, \tag{3}$$

where the $\left\{S_e^{(i)}\right\}_{i\geq 1}$ are iid residual *work amounts* having distribution function given by

$$G_e(x) = \frac{1}{\mathrm{E}S}\int_0^x \mathrm{P}(S > y)dy, \quad x \geq 0, \tag{4}$$

and, independently, Q has a geometric distribution

$$P(Q = k) = \rho^k(1-\rho), \quad k \geq 0. \tag{5}$$

However, in order to move backwards in time, and to properly restore the workload vector for the Σ by recursion (1) forward in time, we also need the complete (the sum of completed and residual) work amounts for the Q customers present in the system $\hat{\Sigma}$ at origin. Each such complete work amount H has the stationary spread distribution [18] with tail d.f. given by

$$P(H > x) = \frac{1}{\mathrm{E}S}x\overline{G}(x) + \overline{G}_e(x). \tag{6}$$

Note that in the system $\hat{\Sigma}$ we must distinguish the work amount of customer (which is defined upon arrival) and the service time (which indeed depends on the number of customers in the system during service). Indeed, in the $\hat{\Sigma}$ the customer starts being served immediately upon arrival, at a rate inversely proportional to the number of customers in the system. in Σ, on the contrast, service times and work amounts coincide.

Using a complete work amount H distributed as (6), we can generate a residual work amount S_e (having d.f. (4)) by the following trick [14]: take U uniform in $[0,1]$ and take $S_e = UH$. Using this trick, we generate Q iid copies H_i and iid U_i, and set $S_e^{(i)} = U_iH_i$, $i = 1,\ldots,Q$ to initialize the steady-state workload vector in $\hat{\Sigma}$. We also store the values $\{H_i\}_{i=1,\ldots,Q}$ used for the workload reconstruction in Σ at the second step.

After initialization, we simulate the workload in $\hat{\Sigma}$, say, by the discrete-event simulation. In fact, this simulation is done in reverse time, so we need to keep track of the customer departure epochs $-t_i$ and the corresponding work amounts S_i. Note that, after time inversion the customer departure epochs $-t_i$ will become arrival epochs and work amounts S_i have the d.f. G except the Q customers present in the system at origin, if any. We stop simulation at the time epoch $-t_n$ with $n = \min\{k \geq 0 : Q(-t_k) = 0\}$, where $Q(t)$ is the number of customers in $\hat{\Sigma}$ at time t. Technically this simulation can be implemented in "forward" time and then time reversal can be performed (care must be taken regarding the indices).

We also note that the service speed in system $\hat{\Sigma}$ changes both at departure and arrival epochs. Actually, the server processes work at unit rate, which is divided equally among all customers present in the system. That is, whenever there are Q customers in the system, each will receive service at rate $r = 1/Q$.

At the second step, we construct the Kiefer-Wolfowitz vector for Σ. We need to correctly define the interarrival times in Σ and the initial conditions for the workload vector $\boldsymbol{W}_1$. To do so,

we firstly define the interdeparture times (in reverse time) in $\hat{\Sigma}$ as $T_i = t_i - t_{i-1}$, $1 \le i \le n$, where $t_0 = 0$, and write down the corresponding work amounts $S_1, \dots, S_n$. Then time reversal of $\hat{\Sigma}$ is performed by resetting $(T_1, \dots, T_n) := (T_n, \dots, T_1)$ and $(S_1, \dots, S_n) := (S_n, \dots, S_1)$, respectively. Thus, sequences $(T_1, \dots, T_n)$ and $(S_1, \dots, S_n)$ are now the sample path in $\hat{\Sigma}$ from an empty state at time in the past $-t_n$, $t_n \ge 0$. Starting from $\boldsymbol{W}_1 = \boldsymbol{0}$, we reconstruct the sample path in Σ up to $\boldsymbol{W}_{n+1}$ which is taken as the sample from the steady state. Indeed, if at time $-t_n$ the first customer arrives in an empty system and until time 0 we have n arrivals, then at time 0 the system will be observed by a customer with the number $n+1$.

Algorithm 1 The original approach

Require: Service time d.f. G, arrival rate λ
Set modeling time $t = t_0 = 0$
Set $n = 0$
Generate the number of customers, Q, in the system $\hat{\Sigma}$ using (5).
if $Q = 0$ **then**
  Set $\boldsymbol{V}(0) = \boldsymbol{0}$
else
  Sample iid H_i, $i = 1, \dots, Q$ from the stationary spread distribution, using (6)
  Sample iid U_i, $i = 1, \dots, Q$, uniformly on $[0, 1]$
  Take $\{Q, U_1 H_1, \dots, U_Q H_Q\}$ as the stationary workload in $\hat{\Sigma}$ system at origin
  Sample interarrival time $T \sim Exp(\lambda)$
  while $Q > 0$ **do** ▷ Discrete-event simulation of $\hat{\Sigma}$
    if the next event is an arrival **then**
      Generate a work amount $S \sim G$ and keep a record of the value
      Sample interarrival time $T \sim Exp(\lambda)$
      Reset $Q = Q + 1$ and set $r = 1/Q$
    else if the next event is a departure **then**
      Reset $n = n + 1$
      Record the departure time t_n of the customer
      Reset $Q = Q - 1$ and set $r = 1/Q$
    end if
    Update the times and residual work amounts
  end while
  Define the interdeparture times $T_i = t_i - t_{i-1}$, $1 \le i \le n$
  Reset $(S_1, \dots, S_n) = (S_n, \dots, S_1)$ and $(T_1, \dots, T_n) = (T_n, \dots, T_1)$
  Set $\boldsymbol{W}_1 = \boldsymbol{0}$
  for $i = 1, 2, \dots, n$ **do** ▷ Reconstruct the sample path in Σ
    $\boldsymbol{W}_{i+1} = R\big(\boldsymbol{W}_i + S_i \boldsymbol{e} - T_i \boldsymbol{1}\big)^+$
  end for
  Set $\boldsymbol{V}(0) = \boldsymbol{W}_{n+1}$
end if
return $\boldsymbol{V}(0)$

3. Modified Algorithm for Heterogeneous MJM

Let us describe a MJM with c servers operating under FCFS discipline [4, 19, 20, 6]. In such a system, class-k customer requires k (available) servers which are seized and released simultaneously, $1 \le k \le c$. The (random) service time S_i of customer i is thus identical at each of the N_i (random number of) servers. Note that we consider a heterogeneous MJM system (denoted by Σ) where $\{S_i\}_{i \ge 1}$ may have a class-dependent distribution.

The input of class-k customers follows a Poisson process with rate λ_k. Thus, the interarrival

times $\{T_i\}_{i\geq 1}$ have an exponential distribution with rate $\lambda = \lambda_1 + \cdots + \lambda_c$, and

$$p_k = \mathrm{P}\{N = k\} = \lambda_k/\lambda \tag{7}$$

may be treated as the probability of class-k customer arrival.

In MJM, servers can be idle when the head-of-queue customer blocks the subsequent customers in the queue, which is known as a non-work-conserving property. After appropriate modification of the Kiefer–Wolfowitz recursion (1), the workload vector can be calculated as follows [20]:

$$\boldsymbol{W}_{i+1} = R\big((W_{i,N_i} + S_i)\boldsymbol{e}_{1:N_i} + \boldsymbol{W}_i \circ (\mathbf{1} - \boldsymbol{e}_{1:N_i}) - T_i\mathbf{1}\big)^+, \tag{8}$$

where $\boldsymbol{e}_{1:k}$, for $k = 1, \ldots, c$, is a vector having ones on the first k elements and zeroes elsewhere and $\circ$ is Hadamard (componentwise) multiplication. Alternatively, (8) reads

$$\boldsymbol{W}_{i+1} = R\big(\underbrace{W_{i,N_i} + S_i - T_i, \ldots, W_{i,N_i} + S_i - T_i}_{N_i \text{ components}}, W_{i,N_i+1} - T_i, \ldots, W_{i,c} - T_i\big)^+.$$

We can see that the first N_i components of the vector are the same, since customer i waits for the release of exactly N_i servers and seizes them simultaneously for the same time S_i. Thus W_{i,N_i} is the delay of the customer i, and W_N is the (stationary) delay of a generic customer of class N. By this reason, each coordinate in the stationary workload vector $\boldsymbol{W}$ gives the stationary delay of the customer of the corresponding class.

To appropriately modify the original algorithm for the MJM, we need to find a suitable upper bound $\hat{\Sigma}$. Firstly, we need the following intermediate result. It was shown in [20] that for the MJM in homogeneous case (class-independent service times), an upper bound $\tilde{\Sigma}$ can be constructed as follows. By coupling, take the same interarrival times $\tilde{T}_i = T_i$ and service times $\tilde{S}_i = S_i$, $i \geq 1$. In the c-server system $\tilde{\Sigma}$, force each customer to use all the c servers simultaneously. In such a case, workload $\tilde{\boldsymbol{W}}_i$ at each of the c servers upon arrival of customer i essentially follows the Lindley recursion,

$$\tilde{\boldsymbol{W}}_{i+1} = \big(\tilde{\boldsymbol{W}}_i + S_i\mathbf{1} - T_i\mathbf{1}\big)^+. \tag{9}$$

It can be seen that this c-server system is in fact equivalent to an M/G/1 system operating under FCFS discipline. Moreover, the proof in [20] does not depend on the assumption of homogeneity for the service times. However, a more delicate coupling is needed in the heterogeneous case, which we provide below.

Take c iid sequences $\{S_i^{(k)}\}_{i\geq 1}$, where the generic element $S^{(k)}$ has d.f. $G^{(k)}$, $k = 1, \ldots, c$. The service times in Σ are taken in accordance with the class N_i of the customer i as $S_i^{(N_i)}$, i.e. the i-th element of the appropriate sequence. In this case, the upper bound for the workload sequence $\{W_i\}_{i\geq 1}$ is provided by the c-server system $\tilde{\Sigma}$ in which the workload sequence $\{\tilde{W}_i\}_{i\geq 1}$ follows the recursion (9), where the service time of the i-th customer is given as a *finite mixture*

$$\tilde{S}_i = \sum_{k=1}^{c} I_i^{(k)} S_i^{(k)}, \tag{10}$$

and $I_i^{(k)}$ is the indicator that the i-th customer has class k. Note that $I_i^{(N_i)} = 1$ implies $\tilde{S}_i = S_i^{(N_i)}$, so the service times of customers in systems Σ and $\hat{\Sigma}$ coincide. Thus the system $\tilde{\Sigma}$ can be used as an upper bound for the system Σ (in which we indeed use a coupling to have relations w.p.1 between random variables) under the conditions given in the following Lemma.

Lemma 1. Assume that in the systems Σ and $\tilde{\Sigma}$, the initial workload vectors are ordered as $\boldsymbol{W}_1 \leq \tilde{\boldsymbol{W}}_1$, and the governing variables are related as $T_i = \tilde{T}_i$, $\hat{S}_i = S_i^{(N_i)}$, $i = 1, 2, \ldots$. Then $\boldsymbol{W}_i \leq \tilde{\boldsymbol{W}}_i$.

Proof. Assume by induction that $\boldsymbol{W}_i \leq \tilde{\boldsymbol{W}}_i$ for some $i \geq 1$. Rewrite (8) as follows,

$$\boldsymbol{W}_{i+1} = \left\{ \begin{array}{c} W_{i+1,1} \\ \cdots \\ W_{i+1,c} \end{array} \right\} = R \left\{ \begin{array}{c} W_{i,N_i} + S_i^{(N_i)} - T_i \\ \cdots \\ W_{i,N_i} + S_i^{(N_i)} - T_i \\ W_{i,N_i+1} - T_i \\ \cdots \\ W_{i,c} - T_i \end{array} \right\}^+ .$$

Since the operator $R(\cdot)^+$ keeps the ordering between components (e.g. see Lemma 1 [20]) and using (9), verify

$$\boldsymbol{W}_{i+1} = \left\{ \begin{array}{c} W_{i+1,1} \\ \cdots \\ W_{i+1,c} \end{array} \right\} = R \left\{ \begin{array}{c} W_{i,N_i} + S_i^{(N_i)} - T_i \\ \cdots \\ W_{i,N_i} + S_i^{(N_i)} - T_i \\ W_{i,N_i+1} - T_i \\ \cdots \\ W_{i,c} - T_i \end{array} \right\}^+ \leq R \left\{ \begin{array}{c} W_{i,c} + S_i^{(N_i)} - T_i \\ \cdots \\ W_{i,c} + S_i^{(N_i)} - T_i \\ W_{i,c} - T_i \\ \cdots \\ W_{i,c} - T_i \end{array} \right\}^+ \leq$$

$$R \left\{ \begin{array}{c} W_{i,c} + S_i^{(N_i)} - T_i \\ \cdots \\ W_{i,c} + S_i^{(N_i)} - T_i \end{array} \right\}^+ \leq \left\{ \begin{array}{c} \tilde{W}_i + S_i^{(N_i)} - T_i \\ \cdots \\ \tilde{W}_i + S_i^{(N_i)} - T_i \end{array} \right\}^+ = \left\{ \begin{array}{c} \tilde{W}_{i+1} \\ \cdots \\ \tilde{W}_{i+1} \end{array} \right\}^+ = \tilde{\boldsymbol{W}}_{i+1}.$$

■

We have shown that $\tilde{\Sigma}$ is a componentwise upper bound for Σ in terms of the workload. At the same time, $\tilde{\Sigma}$ is equivalent to M/G/1-FCFS which is insensitive to the work-conserving queueing discipline. Thus, we can take the workload sequence $\{\hat{W}_i\}_{i\geq 1}$ in the M/G/1-PS system $\hat{\Sigma}$ as an upper bound for each component of the workload vector sequence $\{\boldsymbol{W}_i\}_{i\geq 1}$.

Now we need to construct the steady state of the system $\hat{\Sigma}$. Denote the class-k service rate by $\mu_k = 1/ES^{(k)}$ and the class-k load by $\rho_k = \lambda_k/\mu_k$, $k = 1, \ldots, c$. Let $H^{(k)}$ have the stationary spread distribution (6) for each class k (in obvious notation). A construction of the spread distribution in $\hat{\Sigma}$ is given in the following Lemma.

Lemma 2. Let H have the stationary spread distribution in $\hat{\Sigma}$, then

$$P(H > x) = \sum_{k=1}^{c} \frac{\rho_k}{\rho} P(H^{(k)} > x). \tag{11}$$

Proof. Indeed, it follows from (4), (6) and (10), after simple algebra,

$$P(H > x) = \mu x \overline{G}(x) + \overline{G}_e(x) = \mu x \sum_{k=1}^{c} p_k \overline{G}^{(k)}(x) + \mu \sum_{k=1}^{c} \frac{p_k}{\mu_k} \overline{G}_e^{(k)}(x) =$$
$$= \sum_{k=1}^{c} \frac{\rho_k}{\rho} \left(x\mu_k \overline{G}^{(k)}(x) + \overline{G}_e^{(k)}(x) \right) = \sum_{k=1}^{c} \frac{\rho_k}{\rho} P(H^{(k)} > x).$$

■

It follows from (11) that the spread distribution in $\hat{\Sigma}$ is a finite mixture of the corresponding class-k r.v. $H^{(k)}$, with mixing probabilities

$$\left\{ \frac{\rho_1}{\rho}, \ldots, \frac{\rho_c}{\rho} \right\}. \tag{12}$$

Thus, the stationary workload in $\hat{\Sigma}$ is constructed as

$$\left\{ Q, U_1 H_1^{(N_1)}, \ldots, U_Q H_Q^{(N_Q)} \right\}, \tag{13}$$

where Q has a geometric distribution (5), N_i are iid samples from (12), $H_i^{(N_i)}$ are iid samples from the corresponding stationary spread distributions, using (6) (in obvious notation), and U_i are iid samples from uniform distribution on $[0,1]$, $i = 1, \dots, Q$.

We are ready to describe the modified algorithm. At the first step, the steady state of the M/G/1-PS system $\hat{\Sigma}$ is constructed as (13). The system is simulated backwards in time until it empties, and the classes of customers are stored.

At the second step, invert the time axis, and using the data of customer departure epochs (as arrivals) and corresponding work amounts and customer classes from the trajectory of $\hat{\Sigma}$, construct the workload vectors in Σ by recursion (8).

The modified algorithm for class-dependent (heterogeneous) MJM is presented as a pseudocode in Algorithm 2.

The procedure is simplified in the homogeneous MJM where the service times of customers $\{S_i\}_{i\geq 1}$ are iid and have arbitrary class-independent distribution G. In this case the first step of the original algorithm in Section 2 remains unchanged. At the second step, the sample path in Σ is restored using the recursion (8), where classes of customers are *sampled* from discrete distribution (7). The homogeneous version of the modified algorithm is presented as a pseudocode in Algorithm 3.

4. Validation and Numerical Results

To validate the modified algorithm, we use the explicit results obtained by Brill and Green [4] in the M/M/2-type MJM. In this model, class-k customers arrive at rate λ_k, $k = 1, 2$, and service times have exponential distribution with class-independent rate μ. A notable proximity of the exact solution given in Example 3 [4] to the sample obtained by Algorithm 3 is shown in Figure 1.

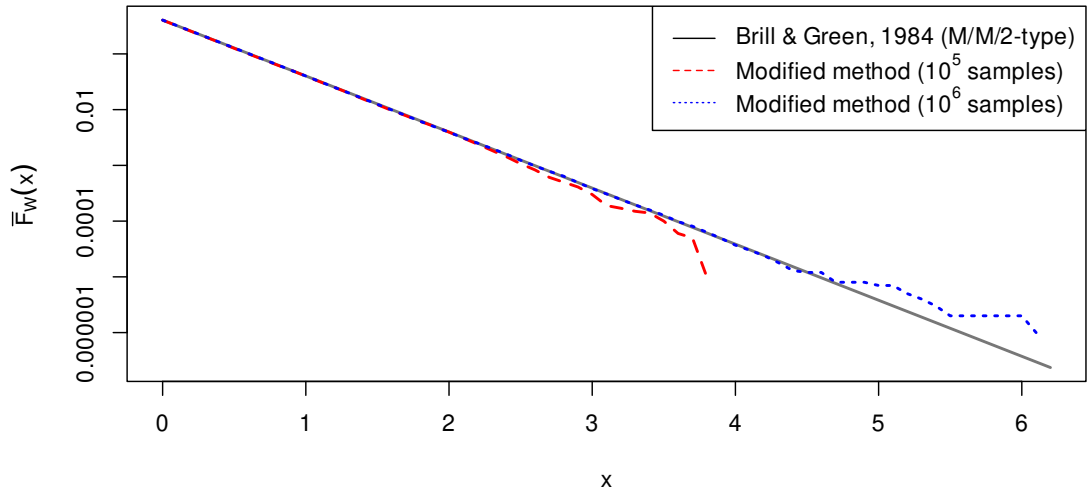


Figure 1: *M/M/2-type FCFS, $\lambda_1 = 1, \lambda_2 = 1, \mu = 4, \rho = 0.5$. Exact solution for tail d.f. waiting time (Example 3 [4]) and estimated c.d.f. from 10^5 and 10^6 samples obtained by Algorithm 3.*

In the following experiments we take the service time distribution as the second-type Pareto, ParetoII(α,β) having density and d.f.,respectively

$$f_S(x) = \frac{\alpha}{\beta}\left(\frac{\beta}{x+\beta}\right)^{\alpha+1}, \quad F_S(x) = 1 - \left(\frac{\beta}{x+\beta}\right)^{\alpha}. \tag{14}$$

Algorithm 2 The algorithm for heterogeneous MJM

Require: Service time d.f. G^k, arrival rates λ_k, $k = 1, \dots, c$
Set modeling time $t = t_0 = 0$
Set $n = 0$ ▷ order number of departure
Generate the number of customers, Q, in the system $\hat{\Sigma}$ using (5).
Set $m = Q$ ▷ order number of arrival
if $Q = 0$ **then**
    Set $\boldsymbol{V}(0) = \mathbf{0}$
else
    Sample iid N_i, $i = 1, \dots, m$ from discrete distribution $\{\rho_1/\rho, \dots, \rho_c/\rho\}$
    Sample iid $H_i^{(N_i)}$, $i = 1, \dots, m$ from the stationary spread distribution, using (6)
    Sample iid U_i, $i = 1, \dots, m$, uniformly on $[0, 1]$
    Take $\{Q, U_1 H_1^{(N_1)}, \dots, U_m H_m^{(N_m)}\}$ as the stationary workload in $\hat{\Sigma}$ system at origin
    Reset $m = m + 1$
    Sample customer class N_m from discrete distribution $\{p_1, \dots, p_c\}$, record this value
    Sample a work amount $S_m \sim G^{(N_m)}$ and keep a record of the value
    Sample next interarrival time $T_m \sim Exp(\lambda_{N_m})$
    while $Q > 0$ **do** ▷ Discrete-event simulation of $\hat{\Sigma}$
        if the next event is an arrival **then**
            Reset $m = m + 1$
            Sample class of arrival customer N_m from discrete distribution $\{p_1, \dots, p_c\}$, record this value
            Sample a work amount $S_m \sim G^{(N_m)}$ and keep a record of the value
            Sample the next interarrival time $T_m \sim Exp(\lambda_{N_m})$
            Reset $Q = Q + 1$ and set $r = 1/Q$
        else if the next event is a departure **then**
            Reset $n = n + 1$
            Record the departure time t_n of the customer
            Reset $Q = Q - 1$ and set $r = 1/Q$
        end if
    end while
    Define the interdeparture times $T_i = t_i - t_{i-1}, 1 \le i \le n$, with $t_0 = 0$
    Reset $(S_1, \dots, S_n) = (S_n, \dots, S_1)$, $(T_1, \dots, T_n) = (T_n, \dots, T_1)$ and $(N_1, \dots, N_n) = (N_n, \dots, N_1)$
    for $i = 1, 2, \dots, n$ **do** ▷ Reconstruct the sample path in Σ
        $\boldsymbol{W}_{i+1} = R\big((W_{i,N_i} + S_i)\boldsymbol{e}_{1:N_i} + \boldsymbol{W}_i \circ (\mathbf{1} - \boldsymbol{e}_{1:N_i}) - T_i\mathbf{1}\big)^+$
    end for
    Set $\boldsymbol{V}(0) = \boldsymbol{W}_{n+1}$
    Sample class of arrival customer N_{n+1} from discrete distribution $\{p_1, \dots, p_c\}$
end if
return $(\boldsymbol{V}(0), N_{n+1})$

Algorithm 3 The algorithm for homogeneous MJM

Require: Service time d.f. G, arrival rates $\lambda_1, \dots, \lambda_c$

Set modeling time $t = t_0 = 0$
Set $n = 0$
Generate the number of customers, Q, in the system $\hat{\Sigma}$ using (5).
if $Q = 0$ **then**
    Set $\boldsymbol{V}(0) = \mathbf{0}$
else
    Sample iid H_i, $i = 1, \dots, Q$ from the stationary spread distribution, using (6)
    Sample iid U_i, $i = 1, \dots, Q$, uniformly on $[0, 1]$
    Take $\{Q, U_1 H_1, \dots, U_Q H_Q\}$ as the stationary workload in $\hat{\Sigma}$ system at origin
    Sample interarrival time $T \sim Exp(\lambda)$
    while $Q > 0$ **do** ▷ Discrete-event simulation of $\hat{\Sigma}$
        if the next event is an arrival **then**
            Sample a work amount $S \sim G$ and keep a record of the value
            Sample interarrival time $T \sim Exp(\lambda)$
            Reset $Q = Q + 1$ and set $r = 1/Q$
        else if the next event is a departure **then**
            Reset $n = n + 1$
            Record the departure time t_n of the customer
            Reset $Q = Q - 1$ and set $r = 1/Q$
        end if
        Update the times and residual work amounts
    end while
    Define the interdeparture times $T_i = t_i - t_{i-1}, 1 \le i \le n$
    Reset $(S_1, \dots, S_n) = (S_n, \dots, S_1)$ and $(T_1, \dots, T_n) = (T_n, \dots, T_1)$
    Sample classes of arrival customers $(N_1, \dots, N_{n+1})$ in Σ from discrete distribution $\{p_1, \dots, p_c\}$
    Set $\boldsymbol{W}_1 = 0$
    for $i = 1, 2, \dots, n$ **do** ▷ Reconstruct the sample path in Σ
        $\boldsymbol{W}_{i+1} = R\big((W_{i,N_i} + S_i)\boldsymbol{e}_{1:N_i} + \boldsymbol{W}_i \circ (\mathbf{1} - \boldsymbol{e}_{1:N_i}) - T_i \mathbf{1}\big)^+$
    end for
    Set $\boldsymbol{V}(0) = \boldsymbol{W}_{n+1}$
end if
return $(\boldsymbol{V}(0), N_{n+1})$

The stationary waiting time distribution in M/ParetoII(α,β)/1 is obtained in [21] in explicit form,

$$F_W(x) = 1 - \rho(1-\rho)\int_0^\infty \frac{u^{\alpha-2}e^{-(1+x/\beta)u}}{\Gamma(\alpha-1)H_1(u,\alpha,\rho)}du, \tag{15}$$

where

$$H_1(u,\alpha,\rho) = \begin{cases} (1+(\alpha-1)\rho e^{-u}Ei_\alpha(u))^2 + (\rho I(u,\alpha-1))^2 \text{ if } \alpha = 2,3,... \\ (1-\rho R(u,\alpha-1))^2 + (\rho I(u,\alpha-1))^2 \text{ if } \alpha > 1 \cup \alpha \neq 2,3,..., \end{cases}$$

$Ei_\alpha(u)$ is Ramsay's generalization of the exponential integral,

$$I(u,\alpha) = \frac{\pi u^\alpha e^{-u}}{\Gamma(\alpha)}, \quad R(u,\alpha) = {}_1F_1(1,1-\alpha,-u) - I(u,\alpha)\cot(\pi\alpha),$$

and ${}_1F_1(1,1-\alpha,-u)$ is the hypergeometric function of the first kind.

In the first experiment, we apply the modified method to illustrate the stochastic monotonicity established in Lemma 1. We plot the tails of the per-class (as well as generic) waiting time distributions in a heterogeneous 2-server MJM. We take $\rho = 0.4$, $\lambda_1 = 0.832$, $\lambda_2 = 0.208$, and use class-dependent ParetoII(α_k,β_k) service time distribution for class-k customer, $k = 1,2$, where $\alpha_1 = 3.5$, $\alpha_2 = 4.1$, $\beta_1 = \beta_2 = 1$. Figure 2 shows that the tail of the waiting time distribution of a generic customer in a heterogeneous MJM 2-server system is sandwiched between the waiting time distributions of class 1 and class 2 customers. It is also shown that the classical M/G/2-FCFS system with a finite mixture service time distribution is the lower bound, and M/G/1-FCFS is the upper bound for the 2-server MJM. The lower bound d.f. estimate is built from samples obtained by the original method, and the upper bound obtained explicitly using (15).

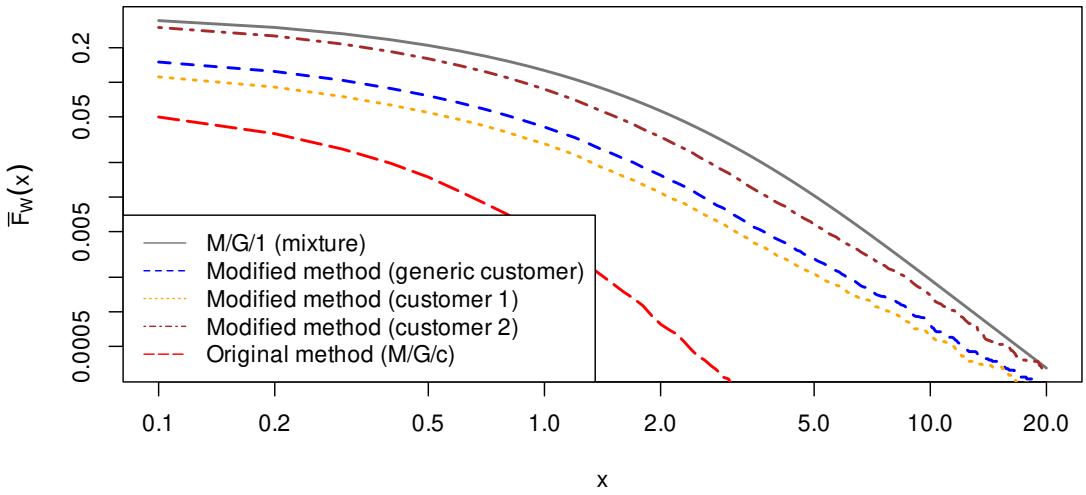


Figure 2: *Waiting time tail distribution for the case $c = 2, \rho = 0.4, \lambda_1 = 0.832, \lambda_2 = 0.208, \alpha_1 = 3.5, \alpha_2 = 4.1, \beta_1 = \beta_2 = 1, p_1 = 0.8, p_2 = 0.2$, where service time has distribution Pareto II. Explicit solution for M/G/1 for mixture vs. samples by modified method (over 10^5 samples) vs. original method (over 10^5 samples). Modified method (generic customer) is the tail c.d.f. delay which see a generic customer, modified method (customer 1) - customer of class 1, modified method (customer 2) - customer of class 2. Note the logarithmic scale on both axes.*

In the subsequent experiments we use the homogeneous MJM, when the service times are class-independent.

Figure 3 shows the results of an experiment with an MJM system having two servers, with $\rho = 0.9$, input rate $\lambda = 2.25$, and ParetoII(α,β) service time distribution with $\alpha = 3.5$, $\beta =$

1. The results of sampling from the stationary delay in a classical M/G/2 system using the original method (Algorithm 1) and delay in MJM 2-server system using the modified method (Algorithm 3), as well as the sample path-based estimates for the delays (using corresponding stochastic recursions (1) and (8)), are shown. The workload distribution for the upper bound M/G/1-FCFS system is obtained explicitly using (15). In our example, single-server and two-server customers are equally likely to arrive, $p_1 = p_2 = 0.5$. Interestingly, despite the fact that only about half of the customers require simultaneous service on two servers, the waiting time for a generic customer is close to the waiting time in a single-server system.

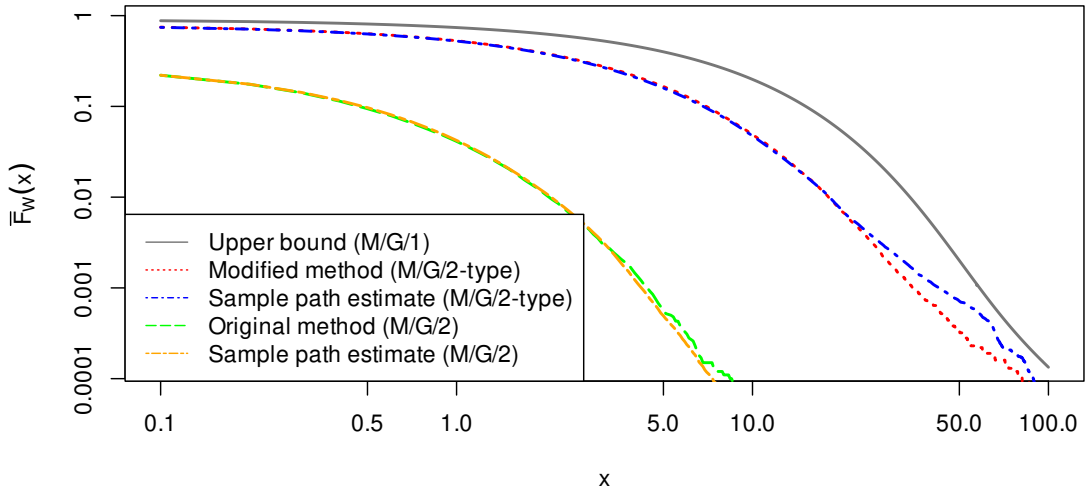


Figure 3: *Waiting time tail distribution for the case $c = 2$, $\rho = 0.9$, $\lambda = 2.25$, $\alpha = 3.5$, $\beta = 1$, where service time has distribution* (14)*. Explicit solution* (15) *for M/G/1 vs. samples by modified method (10^5 samples) vs. sample path of MJM (10^6 arrivals) vs. original method for classic M/G/2 model (10^5 samples) vs. sample path estimates for classic M/G/2 system (10^6 arrivals). Note the logarithmic scale on both axes.*

In Figure 4, we depict the estimated tail d.f. of the stationary delay (of a generic customer) in an MJM model with $c = 100$ servers, $\lambda = 2.25$, and ParetoII$(3.5, 1)$ service time distribution where all classes are sampled uniformly. A notable proximity of the d.f. with the explicit d.f. for M/G/1 model (15) is observed, which is in sharp contrast to the classical multiserver model. In fact, we performed several experiments with M/G/$\tilde{c}$ classical multiserver models, varying $\tilde{c}$ up to 10, and starting from $\tilde{c} = 7$ onwards, the delays are almost negligible. This illustrates that the condition $\rho < 1$ is not very restrictive for MJM (in contrast to the super-stability in the classical system).

It can be seen from Figures 3 and 4 that the sample path estimates give rather inaccurate results in the tail part. To study the reason for this inaccuracy, we compare the tail d.f. of the delays using an explicit solution (15), sample path estimate (10^6 customers), and simulation of stationary delay using formula (3) (10^5 samples) in M/G/1 system. The results depicted in Figure 5 show the same pattern which demonstrates that the number of samples for the sample path estimates should be taken large enough to mitigate the dependence of the members of a delay sequence.

5. Conclusion

In this paper we introduced a modification of the exact simulation approach suggested in [14]. Adaptation of this approach to heterogeneous MJM by using stochastic recursion in the form (8) allowed us to study the performance of this rather sophisticated model. The proposed algorithm allows one to obtain the unbiased estimates of the system performance, as well as the per-class

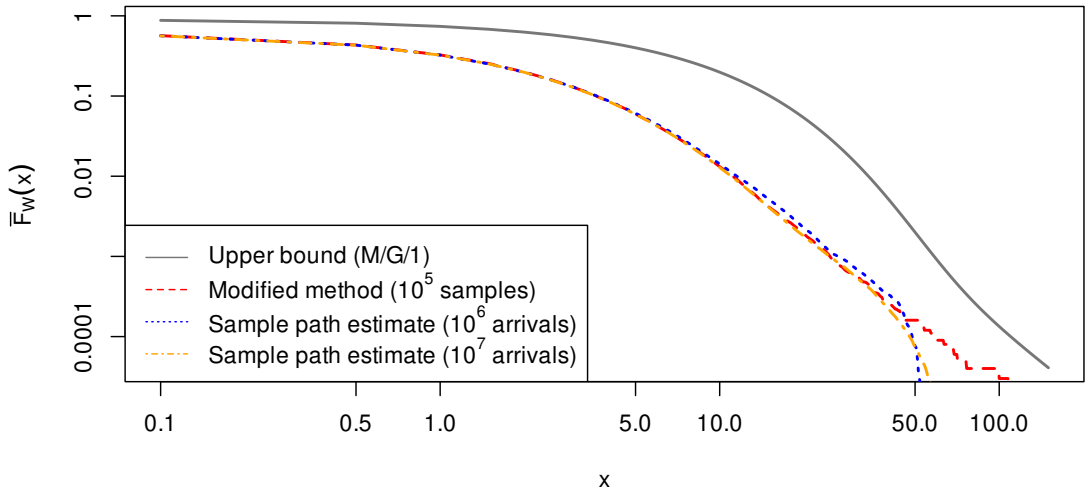


Figure 4: *Waiting time tail distribution for the M/G/100-type, where service time has distribution Pareto II,* $\rho = 0.9$, $\lambda = 2.25$, $\alpha = 3.5$, $\beta = 1$. *Explicit solution* (15) *for M/G/1 vs. samples by modified method* (10^5 *samples) vs. sample path estimates* (10^6 *and* 10^7 *arrivals). Note the logarithmic scale on both axes.*

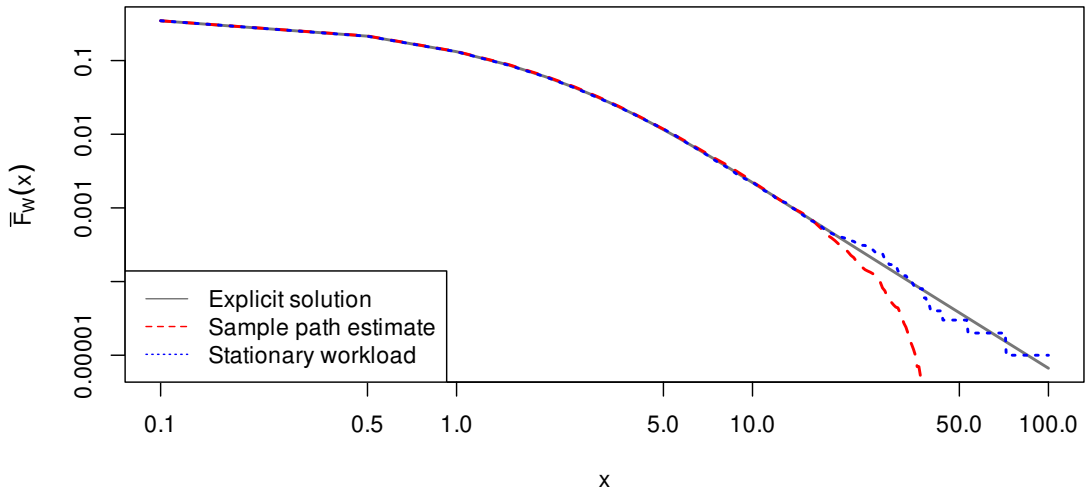


Figure 5: *Tail of the waiting time for M/G/1 model, where service time has distribution ParetoII*(α, β), $\rho = 0.4$, $\lambda = 1$, $\alpha = 3.5$, $\beta = 1$. *Explicit solution* (15) *vs. sample path estimate* (10^6 *arrivals) vs. simulation by* (3) (10^5 *samples). Note the logarithmic scale on both axes.*

stationary waiting times of customers. Heavy-tailed (type-II Pareto) distribution was used in the numerical experiments. As a further study, other service time distributions, such as exponential or phase-type distributions, can be considered, as well as another distribution for the customer class, such as the Zipf distribution, which may have practical value for other applications. In some situations, we are very interested in examining rare but critical events, such as the probability of very long waiting times. We suppose that exact sampling allows us to simulate these rare events, and, through a large number of samples, to accurately estimate the probability that such events will occur.

References

[1] L. Gimpelson. "Analysis of Mixtures of Wide- and Narrow-Band Traffic". en. In: *IEEE Transactions on Communications* 13.3 (1965), pp. 258–266. DOI: 10.1109/TCOM.1965.1089121.

[2] Eric Wolman. "The Camp-On Problem for Multiple-Address Traffic". en. In: *Bell System Technical Journal* 51.6 (1972), pp. 1363–1422. DOI: 10.1002/j.1538-7305.1972.tb02657.x.

[3] E. Arthurs and J. S. Kaufman. "Sizing a Message Store Subject to Blocking Criteria". In: *Proceedings of the Third International Symposium on Modelling and Performance Evaluation of Computer Systems: Performance of Computer Systems*. Amsterdam, The Netherlands, The Netherlands: North-Holland Publishing Co., 1979, pp. 547–564. URL: http://dl.acm.org/citation.cfm?id=647409.724583.

[4] Percy H. Brill and Linda Green. "Queues in which customers receive simultaneous service from a random number of servers: a system point approach". In: *Management Science* 30.1 (1984), pp. 51–68. DOI: 10.1287/mnsc.30.1.51.

[5] D. Filippopoulos and H. Karatza. "An M/M/2 parallel system model with pure space sharing among rigid jobs". In: *Mathematical and Computer Modelling* 45.5–6 (2007), pp. 491–530. DOI: 10.1016/j.mcm.2006.06.007.

[6] Alexander Rumyantsev and Evsey Morozov. "Stability criterion of a multiserver model with simultaneous service". In: *Annals of Operations Research* 252.1 (2017), pp. 29–39. DOI: 10.1007/s10479-015-1917-2.

[7] Mor Harchol-Balter. "The multiserver job queueing model". en. In: *Queueing Systems* 100.3-4 (2022), pp. 201–203. DOI: 10.1007/s11134-022-09762-x.

[8] Diletta Olliaro et al. "The Impact of Service Demand Variability on Data Center Performance". en. In: *IEEE Transactions on Parallel and Distributed Systems* 36.2 (2025), pp. 120–132. DOI: 10.1109/TPDS.2024.3497792.

[9] Eugene Furman et al. *Capacity Allocation for Clouds with Parallel Processing, Batch Arrivals, and Heterogeneous Service Requirements*. 2022. arXiv: 2209.08820 [stat.AP]. URL: https://arxiv.org/abs/2209.08820.

[10] Alexander Rumyantsev et al. "Three-level modeling of a speed-scaling supercomputer". en. In: *Annals of Operations Research* (2022). DOI: 10.1007/s10479-022-04830-0. (Visited on 06/21/2022).

[11] Mor Harchol-Balter. "Open problems in queueing theory inspired by datacenter computing". en. In: *Queueing Systems* 97.1-2 (2021), pp. 3–37. DOI: 10.1007/s11134-020-09684-6.

[12] Mark L. Huber. *Perfect Simulation*. Chapman and Hall/CRC, Jan. 2016. ISBN: 9780429165269. DOI: 10.1201/b19235.

[13] A. Golovin, A. Rumyantsev, and S. Chakravarthy. *Regenerative Exact Simulation for Multiserver Job Model*. Unpublished.

[14] Karl Sigman. "Exact Simulation of the Stationary Distribution of the FIFO M/G/c Queue". In: *Journal of Applied Probability* 48A (2011). Publisher: Applied Probability Trust, pp. 209–213. DOI: 10.1239/jap/1318940466.

[15] Wilfrid Kendall. "Geometric Ergodicity and Perfect Simulation". In: *Electronic Communications in Probability* 9.none (2004). DOI: 10.1214/ecp.v9-1117.

[16] Erol Gelenbe and Isi Mitrani. *Analysis and Synthesis of Computer Systems*. IMPERIAL COLLEGE PRESS, 2010. ISBN: 9781848163966. DOI: 10.1142/p643.

[17] Sheldon M. Ross. *Stochastic processes*. 2nd ed. Wiley series in probability and statistics. New York: Wiley, 1996. ISBN: 978-0-471-12062-9.

[18] Karl Sigman. "Stationary Marked Point Processes". In: *Springer Handbook of Engineering Statistics*. Springer London, 2006, pp. 137–152. DOI: 10.1007/978-1-84628-288-1_8.

[19] G. Y. Fletcher, H. G. Perros, and W. J. Stewart. "A queueing system where customers require a random number of servers simultaneously". In: *European Journal of Operational Research* 23.3 (1986), pp. 331–342. DOI: 10.1016/0377-2217(86)90299-7.

[20] Evsey Morozov, Alexander Rumyantsev, and Irina Peshkova. "Monotonicity and stochastic bounds for simultaneous service multiserver systems". In: *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT), 2016 8th International Congress on*. IEEE, 2016, pp. 294–297. DOI: 10.1109/ICUMT.2016.7765374.

[21] Colin M. Ramsay. "Exact waiting time and queue size distributions for equilibrium M/G/1 queues with Pareto service". en. In: *Queueing Systems* 57.4 (2007), pp. 147–155. DOI: 10.1007/s11134-007-9052-7.