mathematics

MDPI

Article

Sensitivity Analysis and Simulation of a Multiserver Queueing System with Mixed Service Time Distribution

Evsey Morozov [1,2,3,†], Michele Pagano [4,†] and Irina Peshkova [1,*,†,‡] and Alexander Rumyantsev [1,2,†]

1 Department of Applied Mathematics and Cybernetics, Petrozavodsk State University, 185035 Petrozavodsk, Russia; emorozov@karelia.ru (E.M.); ar0@krc.karelia.ru (A.R.)
2 Institute of Applied Mathematical Research, Karelian Research Centre, Russian Academy of Sciences, 185910 Petrozavodsk, Russia
3 Moscow Center for Fundamental and Applied Mathematics, Moscow State University, 119991 Moscow, Russia
4 Department of Information Engineering, University of Pisa, 56126 Pisa, Italy; michele.pagano@iet.unipi.it
* Correspondence: iaminova@petrsu.ru; Tel.: +7-8142-71-3261
† These authors contributed equally to this work.
‡ Current address: Lenina Str., 33, 185910 Petrozavodsk, Russia.

Received: 9 June 2020; Accepted: 24 July 2020; Published: 3 August 2020

Abstract: The motivation of mixing distributions in communication/queueing systems modeling is that some input data (e.g., service time in queueing models) may follow several distinct distributions in a single input flow. In this paper, we study the sensitivity of performance measures on proximity of the service time distributions of a multiserver system model with two-component Pareto mixture distribution of service times. The theoretical results are illustrated by numerical simulation of the $M/G/c$ systems while using the perfect sampling approach.

Keywords: pareto mixture distribution; multiserver system; uniform distance; perfect simulation

1. Introduction

Mixtures of distributions arise in complex stochastic systems and they are extensively used for statistical analysis in many real fields, such as lifetime modeling, ageing or failure processes, engineering reliability [1], and survival theory [2], where data are assumed to be heterogeneous. The application of the mixture of distributions in the modeling of queueing systems is often induced by diverse structure of the customers in the system, e.g., by various service time requirements of multiple classes of customers that arrive into the system (for instance, the transmission time of IP datagrams with different lengths), or by the noisy/biased measurements that induce the so-called contaminated distributions. Ignoring such a diversity at the modeling phase may lead to significant deviation of system performance at practical implementation phase as compared to the modelled values. This motivates various types of analysis, including the analysis of continuity, robustness, monotonicity, stability, and sensitivity. In this regard, we mention the fundamental result obtained for telecommunication system models by B. A. Sevast'yanov [3], and the basic monographs [4,5].

The authors would like to use this opportunity to pay tribute to Professor Vladimir Zolotarev and to note his outstanding role as the founder of the International Seminar on Stability Problems for Stochastic Models. One of the authors had a great pleasure to communicate with Professor Zolotarev over many years, and all of the authors actively participated in the seminar he has founded. In the

context of this paper, it is especially appropriate to emphasize an important role of Professor Zolotarev in the study of the stability and monotonicty of queuieng processes, see [6–8].

Information flows in modern telecommunications and computing systems have the form of a superposition of some sequential-parallel structures [9]. Ranging from small personal devices up to large scale high-performance computing systems, all of these may be modeled as multiserver queueing systems. Thus, it is highly important to study the performance of such systems and, in particular, the sensitivity of stationary performance indexes with respect to the variability of input parameters. However, direct output analysis of queueing systems is often tricky (see e.g., [10,11]), and explicit expressions for the distributions of steady-state performance indexes of a multiserver system are, in general, hardly available and, beyond classic models, known in some special cases only. In some cases, the analysis may be performed by obtaining asymptotic upper bounds, as in the paper [12], or studying the continuity of the process, as in [8], or stochastic stability of the queueing process, like in [4,13], or by means of simulation. In the present paper, we utilize the latter approach.

This paper is dedicated to sensitivity analysis of a steady-state performance index of a multiserver system with respect to service time distribution having the form of the so-called finite mixture [14]. However, instead of studying the direct parametric sensitivity, we focus on a more delicate analysis of the (combined) effect of the service time distribution on the steady-state performance estimate. That is, we compare the basic system to a disturbed one, using a sensitivity measure (Kolmogorov–Smirnov distance) both for the service time distributions, and for the steady-state performance estimate (queue size). The service time distribution perturbation is performed by changing the mixing coefficient and parameters defining the mixture components. We formalize this at the end of Section 3.

In general, the output distributions are hardly analytically available and, in this case, we must be able to obtain the steady-state performance indexes by simulation. As a basic model, we consider the classical $M/G/c$ model, where the steady-state distribution of the vector workload process is unknown as well; however, it can be estimated by means of the recently developed method of regenerative perfect simulation [15]. In more detail, as the target (perturbed) service time distribution we take the two-component mixture of Type-II Pareto distributions with support on the positive axis, which is known as Lomax distributions, as well as two-component exponential (hyperexponential) distribution. Such a choice also allows for obtaining some analytical expressions. Our interest to Pareto distribution is caused by the heavy tailed property of this distribution that is frequently observed in models of file size and flow duration [16].

This paper continues the study performed in [17] in the context of monotonicity. The key idea of the present paper is to study qualitatively the sensitivity of the steady-state distribution of the system performance index (steady-state queue size) to the variability of service time distribution by means of simulation. We also apply the auxiliary results on the failure rates comparison, which allows us to characterize the monotonicity of some stationary performance measures.

The structure of the paper is as follows. In Section 2, we introduce the two-component mixture of distributions and discuss some properties that are used in the subsequent analysis. Subsequently, we define the uniform distance between the mixture and the corresponding parent distribution. In Section 3, some known stochastic monotonicity properties of the multiserver system are collected, which further are specified for the considered mixture distributions. In Section 4, we describe the perfect sampling algorithm that is then used to sample from exact (but unknown) steady-state distribution of a multiserver queue $M/G/c$. The results of simulation are presented in Section 5. We study the sensitivity of the steady-state queue size distribution with respect to (w.r.t.) the shape parameters of mixture and the mixing coefficient and illustrate stochastic monotonocity of the system performance. The discussion of the simulation results finalizes the paper in the concluding Section 6.

2. Two-Component Mixture Distributions

The goal of this Section is to derive the uniform distance between the two-component mixture distribution and its parent distribution. First, we introduce the two-component mixture, and then give

a few properties, including the stochastic monotonicity. This property is further used to obtain the monotonicity of the corresponding output queueing process.

Let X_i be independent random variables having mean EX_i, density f_i, tail distribution function (d.f.) $\overline{F}_i(x) = 1 - F_i(x)$, and failure rate

$$r_i(x) = \frac{f_i(x)}{\overline{F}_i(x)}, \ i = 1, 2,$$

defined for such x that $\overline{F}_i(x) > 0$. We assume that $F_1 \neq F_2$ to avoid trivial case. Let I be a Bernoulli random variable independent of X_i, with success probability $P(I = 1) = p$. Subsequently, it is called the random variable

$$X_M = IX_1 + (1 - I)X_2,$$

has the two-component mixture distribution [18] (we use the index M to denote the mixture). The mean EX_M and density, f_M of X_M equal, respectively,

$$EX_M = pEX_1 + (1 - p)EX_2, \tag{1}$$

$$f_M(x) = pf_1(x) + (1 - p)f_2(x), \tag{2}$$

and it is easy to see that the tail distribution is

$$\overline{F}_M(x) = p\overline{F}_1(x) + (1 - p)\overline{F}_2(x). \tag{3}$$

Note that the d.f. F_i may belong to the same family of distributions but have other parameters. In reliability analysis, such a mixture may be interpreted as a contaminated distribution [19], where $1 - p$ is, as a rule, small enough. F_1 is called the parent distribution and F_2 is the contaminating distribution. In this Section, we focus on the distance between the mixture and its parent distribution.

A straightforward analysis shows that the failure rate of the mixture has the following form [20]:

$$r_M(x) = \frac{pf_1(x) + (1 - p)f_2(x)}{p\overline{F}_1(x) + (1 - p)\overline{F}_2(x)} = a(x)r_1(x) + (1 - a(x))r_2(x), \tag{4}$$

where

$$a(x) = \frac{p\overline{F}_1(x)}{p\overline{F}_1(x) + (1 - p)\overline{F}_2(x)}, \quad x \geq 0.$$

In particular, it follows from Equation (4) that

$$r_M(x) \geq \min(r_1(x), r_2(x)), \quad x \geq 0. \tag{5}$$

It is worth mentioning that the mixture preserves the monotonicity of failure rate in the following way: if both rates $r_i(x)$ are non-increasing, that is d.f.'s $F_i(x)$ are decreasing failure rate distributions (DFR), then the mixture $F_M(x)$ is DFR distribution as well [21]. Indeed, one can check that

$$r'_M(x) = a(x)r'_1(x) + (1 - a(x))r'_2(x) - a(x)(1 - a(x))(r_1(x) - r_2(x))^2, \tag{6}$$

also see [1]. Subsequently, if $r'_i(x) < 0$, $i = 1, 2$, it follows from Equation (6) that $r'_M(x) < 0$, since $a(x) \in [0, 1]$ for any $x \in (0, \infty)$. In particular, it follows from Equation (6) that the mixture of two exponential distributions is DFR (note that the exponential distribution has constant failure rate).

Another example of a DFR distribution is the Type-II Pareto distribution, denoted by $Pareto(\alpha_i, x_0)$, having d.f. (see e.g., [22])

$$F_i(x) = 1 - \left(\frac{x_0}{x_0 + x}\right)^{\alpha_i}, \ x \geq 0, \ x_0 > 0, \ \alpha_i > 0, \ i = 1, 2.$$

The failure rate of $Pareto(\alpha_i, x_0)$ equals

$$r_i(x) = \frac{\alpha_i}{x_0 + x}, \ x \geq 0, \ \ i = 1, 2, \tag{7}$$

and it is monotonically decreases to 0 as $x \to \infty$. As has been noted above, the two-component mixture of Pareto distributions is DFR distribution. However, the failure rate of finite mixtures, in general, is a complicated function [20].

The uniform distance between distributions F and G, defined as [12]

$$\Delta(F, G) = \sup_x |F(x) - G(x)|, \tag{8}$$

is a recognised measure, which is actively used in the sensitivity analysis [12]. It is easy to see that the uniform distance $\Delta(F_M, F_1)$ between the mixture distribution Equation (3) and its parent distribution is

$$\Delta(F_M, F_1) = \sup_{x \geq 0} |pF_1(x) + (1-p)F_2(x) - F_1(x)| = (1-p) \sup_{x \geq 0} |F_1(x) - F_2(x)|. \tag{9}$$

Note that, if the densities f_i exist, and there exists x^* that delivers the supremum in Equation (9),

$$\Delta(F_M, F_1) = |F_1(x^*) - F_2(x^*)|,$$

then x^* satisfies the equality

$$f_1(x^*) = f_2(x^*). \tag{10}$$

By definition of the failure rates, r_i, it then follows that

$$r_1(x^*)\overline{F}_1(x^*) = r_2(x^*)\overline{F}_2(x^*).$$

Thus, expression Equation (9) can be written in the following convenient form

$$\Delta(F_M, F_1) = (1-p)\frac{|r_2(x^*) - r_1(x^*)|}{r_2(x^*)}\overline{F}_1(x^*) = (1-p)\frac{|r_1(x^*) - r_2(x^*)|}{r_1(x^*)}\overline{F}_2(x^*). \tag{11}$$

Note that Equation (11) allows obtaining the following upper bound for the distance $\Delta(F_M, F_1)$:

$$\Delta(F_M, F_1) \leq (1-p)\frac{|r_2(x^*) - r_1(x^*)|}{r_1(x^*)} =: \delta(x^*). \tag{12}$$

In particular, for the hyperexponential distribution, that is for two-component mixture of exponential distributions with densities $f_i(x) = \lambda_i e^{-\lambda_i x}$, $i = 1, 2$, it follows from Equation (10), that

$$x^* = \frac{\log \lambda_1 - \log \lambda_2}{\lambda_1 - \lambda_2},$$

and in this case expression Equation (11) becomes

$$\Delta(F_M, F_1) = (1-p)\frac{|\lambda_2 - \lambda_1|}{\lambda_2}\left(\frac{\lambda_1}{\lambda_2}\right)^{-\frac{\lambda_1}{\lambda_1 - \lambda_2}} \leq (1-p)\frac{|\lambda_2 - \lambda_1|}{\lambda_2}. \tag{13}$$

Note that the last inequality in Equation (13) is a particular case of Equation (12). Expression Equation (13) is consistent with a more general result for the so-called univariate scale mixture X_M having form [2]

$$X_M \stackrel{d}{=} \frac{X_1}{Y}, \tag{14}$$

with d.f.

$$\hat{F}_M(x) = \int_0^\infty F_1(\theta x) dG(\theta),$$

where F_1 is the parent distribution of the random variable X_1 and G is the distribution of a mixing random variable $Y \geq 0$. It is clear that the transformation Equation (14) is a scale change, and if $Y \in \{y_1, \ldots, y_m\}$ is a discrete random variable, then Equation (14) becomes

$$X_M = \sum_{i=1}^{m} \frac{1}{y_i} I(Y = y_i) X_1, \tag{15}$$

which is a finite mixture, where I is an indicator function. The aforementioned general result for the univariate scale mixture states that if F_1 is an exponential d.f., then an upper bound for the uniform distance may be obtained as follows [2,23]:

$$\Delta(\hat{F}_M, F_1) \leq \mathrm{E}|Y - 1|. \tag{16}$$

To show that Equation (16) indeed coincides with Equation (13) for the two-component scale mixture case, let Y have point masses at 1 and λ_2/λ_1 with probabilities p and $1 - p$, respectively. It immediately follows from Equations (15) and (16) that

$$\mathrm{E}(Y - 1) = (1 - p) \frac{|\lambda_1 - \lambda_2|}{\lambda_2}.$$

Now, we return to the two-component $Pareto(\alpha_i, x_0)$ mixture F_M. It follows from Equation (11) that in this case

$$\Delta(F_M, F_1) = (1 - p) \frac{|\alpha_1 - \alpha_2|}{\alpha_2} \left(\frac{\alpha_2}{\alpha_1} \right)^{\frac{\alpha_1}{\alpha_1 - \alpha_2}}, \tag{17}$$

where the value x^* satisfying equality Equation (10) equals

$$x^* = x_0 \left(\left(\frac{\alpha_1}{\alpha_2} \right)^{\frac{1}{\alpha_1 - \alpha_2}} - 1 \right).$$

Note that the r.h.s. of Equation (17) is similar to the r.h.s. of Equation (13). This similarity is caused by the specific shape of the failure rate of the distribution $Pareto(\alpha_i, x_0)$. Moreover, in such a case, the quantity $\delta(x^*)$ defined in Equation (12) does not depend on x^* and, thus, for Pareto mixture, it readily follows from Equation (12) that

$$\Delta(F_M, F_1) \leq (1 - p) \frac{|\alpha_1 - \alpha_2|}{\alpha_2} =: \delta(\alpha_1, \alpha_2). \tag{18}$$

In Figure 1, to illustrate the dependence of the uniform distance on the parameter α_2 of the contaminating distribution, we depict $\Delta(F_M, F_1)$ jointly with $\delta(\alpha_1, \alpha_2)$ for fixed $\alpha_1 = 2$ and $p = 0.9$ by varying α_2 in the interval $(1, 5)$.

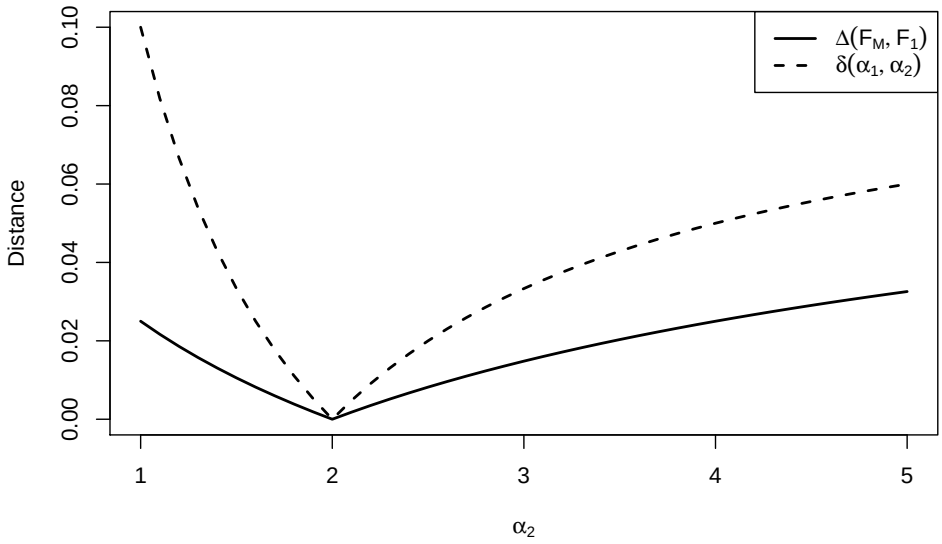

Figure 1. The distance $\Delta(F_M, F_1)$ with mixing parameter $p = 0.9$ and an upper bound $\delta(\alpha_1, \alpha_2)$ vs. parameter α_2.

3. Multiserver System Sensitivity

In this Section, we formalize our main goal for the numerical experiments conducted and discussed in Section 5. We then demonstrate how stochastic and failure rate ordering can be applied to multiserver systems with mixed service time distribution. The numerical experiments equipped with the stochastic comparison technique not only allow for obtaining the absolute value, but also characterizing the monotonicity of performance indexes.

Consider a classical First-Come-First-Served (FCFS) c-server $M/G/c$ queueing system that is fed by a Poisson input with rate λ, arrival instants $\{t_i,\ i \geq 1\}$ with $t_1 = 0$, independent and identically distributed (iid) interarrival times $T_i = t_{i+1} - t_i$ and iid service times $\{S_i,\ i \geq 1\}$. Note that $\lambda = 1/\mathrm{E}T$, where T is generic interarrival time. Now, we consider the c-dimensional vector of the remaining workload process in such a system,

$$W_i = (W_{i,1}, \ldots, W_{i,c}),$$

where $W_{i,k}$ is the kth smallest component of the vector which is observed by the ith arrival [24]. Thus, the vector components are kept in ascending order,

$$W_{i,1} \leq \cdots \leq W_{i,c},$$

and the quantity $W_{i,j}$, "observed" by the arriving customer i, equals the unfinished work which must be done by server j provided no new work arrives after arrival instant t_i of customer i; $j = 1, \ldots, c$. If there are no idle servers upon arrival of customer i, then s/he waits in a common infinite capacity queue until the server with minimal work, $W_{i,1}$, becomes free. It is easy to see that $W_{i,1}$ is the waiting time of customer i which starts being served at time $t_i + W_{i,1}$. It is well-known that the workload vector sequence follows the celebrated stochastic Kiefer–Wolfowitz recursion [25]:

$$W_{i+1} = R(W_i + e_1 S_i - \mathbf{1}T_i)^+, \tag{19}$$

where $e_1 = (1, 0, \ldots, 0)$ and $\mathbf{1}$ is the vector of ones, operator R puts the components in an ascending order, and operation $(\cdot)^+ = \max(0, \cdot)$ is applied componentwise (we omit the sub-index for a generic element of a sequence). In what follows, we assume that the stability condition holds [25],

$$\rho := \lambda \mathrm{E}S < c. \tag{20}$$

Define the departure instant of customer i by $d_i = t_i + W_{i,1} + S_i$. Now define the process

$$Q_n = \sum_{j \geq 1, j \neq n} I(t_j \leq t_n < d_j), \tag{21}$$

counting the queue size (number of customers in the system) at the arrival instant t_n. Under condition Equation (20), Q_n converges in distribution, as $n \to \infty$, to the steady-state queue size Q, with stationary distribution

$$\pi_n = \mathrm{P}(Q = n),\ n \geq 0.$$

Note that when service times S_i are exponential, the steady-state queue size distribution, $\pi_n, n \geq 0$, is well known [26]:

$$\pi_n = \begin{cases} \left(\sum\limits_{k=0}^{c-1} \dfrac{\rho^k}{k!} + \dfrac{\rho^c}{(c-1)!(c-\rho)} \right)^{-1}, & n = 0, \\ \pi_0 \dfrac{\rho^n}{n!}, & 1 \leq n \leq c, \\ \pi_0 \dfrac{\rho^n}{c! c^{n-c}}, & n > c. \end{cases} \tag{22}$$

The operators $R(\cdot)$ and $(\cdot)^+$ in Equation (19) preserve ordering, and it allows for us to establish the monotonicity of the workload sequence in the multiserver system in the case when the driving sequences $\{T_n^{(i)}, S_n^{(i)}, n \geq 1\}$, $i = 1, 2$ satisfy stochastic order. We recall that the stochastic order $X_2 \leq_{st} X_1$ between two random variables X_1, X_2 means that the tail d.f.'s satisfy inequality

$$\mathrm{P}(X_2 > x) \leq \mathrm{P}(X_1 > x),\ \ x \geq 0. \tag{23}$$

In is known [27] that, in two c-server systems with stochastically ordered input sequences, $T^{(2)} \geq_{st} T^{(1)}$ and $S^{(2)} \leq_{st} S^{(1)}$, the workload sequences $\{W_n^{(i)}\}, i = 1, 2$, are (componentwise) ordered in the following way

$$W_n^{(2)} \leq_{st} W_n^{(1)}\ \ n \geq 1. \tag{24}$$

It also holds for the steady-state workloads:

$$W^{(2)} \leq_{st} W^{(1)}.$$

If the input in both systems is the same, which is $T^{(2)} =_{st} T^{(1)}$, then the the queue length process at the arrival instants satisfy similar ordering both in path-wise sense and in steady-state [27]

$$Q^{(2)} \leq_{st} Q^{(1)}. \tag{25}$$

The stochastic ordering $\leq_{st}$ can be transformed into the ordering with probability 1 by the coupling technique [28]. In the context of this work, it is worth mentioning that the sufficient condition for the stochastic ordering $S^{(2)} \leq_{st} S^{(1)}$ is the failure rate ordering [29]:

$$r_2(x) \geq r_1(x),\ x \geq 0, \tag{26}$$

where r_i is the failure rate of r.v. $S^{(i)}$, $i = 1, 2$. We summarize the discussion in the following lemma which is a straightforward result of [27].

Lemma 1. *Consider two c-server systems with stochastically equivalent input, $T^{(2)} =_{st} T^{(1)}$, and failure rate ordered service time distributions, $r_2(x) \geq r_1(x)$, $x \geq 0$. Subsequently, Equation (25) holds.*

Now, we consider two $M/G/c$ queueing systems, denoted by $\Sigma^{(1)}$ and $\Sigma^{(M)}$, fed by (stochastically) identical Poisson process with rate λ. Let the first system $\Sigma^{(1)}$ have the service time

distribution F_1. We refer below to the first system as being basic. In the second (contaminated) system $\Sigma^{(M)}$, we use service time distribution F_M defined by Equation (3), with the same F_1 and some F_2 and $p \in (0,1)$. Let now $Q^{(1)}$ (with d.f. $F_{Q^{(1)}}$) be the steady-state queue size in the first system. Define similarly $Q^{(M)}$ and $F_{Q^{(M)}}$ for the system $\Sigma^{(M)}$. We are interested in studying the sensitivity of the uniform distance

$$\Delta(F_{Q^{(M)}}, F_{Q^{(1)}}) = \sup_{x \geq 0} |F_{Q^{(1)}}(x) - F_{Q^{(M)}}(x)|. \tag{27}$$

More formally, we study the effect of $\Delta(F_M, F_1)$ given by Equation (9), on the steady-state performance $\Delta(F_{Q^{(M)}}, F_{Q^{(1)}})$ defined in Equation (27), by varying the mixing coefficient p and parameters defining the mixture components F_1 and F_2. However, since the distributions $F_{Q^{(M)}}$ and $F_{Q^{(1)}}$ are not available explicitly in general, we use simulation to obtain the corresponding estimates. As such, we study a combined effect of the service time distribution on the steady-state performance estimate.

The generic service time $S^{(M)}$ in the contaminated system $\Sigma^{(M)}$ has a two-component mixture d.f. F_M and, thus, it follows from Equation (5) that the conditions of Lemma 1 are satisfied, since M, where r_M is the failure rate of $S^{(M)}$. In particular, this means that the basic system $\Sigma^{(1)}$ is heavier loaded than the contaminated system $\Sigma^{(M)}$. It then follows from Lemma 1 and Equation (23) that the difference $F_{Q^{(1)}}(x) - F_{Q^{(M)}}(x)$ (see Equation (27)) is negative for all $x \geq 0$. In Section 5, we study the distance Equation (27) numerically.

4. Exact Steady-State Simulation by Regenerative Approach

In general, there are no closed form expressions for the steady-state distribution of the queue length and vector workload process in an $M/G/c$ system. Although a number of approximations exist [30–33], in general the accuracy of such methods is a point of discussion [34], especially when the service times distribution is heavy-tailed. Thus, to study the sensitivity we need to rely on simulation. A contribution of this work is that unlike classical discrete-event simulation (crude Monte-Carlo), which always has the so-called transient (warm-up) period during which an influence of initial conditions exists, we use the perfect simulation technique that allows exact sampling from the (unknown) steady-state distribution. In what follows, we rely on the regenerative approach designed for the $M/G/c$ system in the work [15] (although there are recently developed more sophisticated techniques based on backward coupling, for instant [35], which are valid for a more general $G/G/c$ system). Below, we outline the approach from [15].

This approach uses the so-called a Random Assignment (RA) system $M/G/c$ as a majorant for the original $M/G/c$ system. In the RA system, each new customer is assigned to arbitrary server randomly (that is with probability $1/c$). As a result, the remaining workload in server j that customer n meets, denoted by $V_{n,j}$, satisfies recursion

$$V_{n+1,j} = [V_{n,j} + I(U_n = j)S_n - T_n]^+, \quad j = 1, \ldots, c, \tag{28}$$

where iid random variables $\{U_n\}$ are uniformly distributed over $\{1, \ldots, c\}$, and $I(U_n = j) = 1$ means that customer n is routed to server j. The RA system is indeed is a collection of $M/G/1$ systems, each with Poisson input with rate λ/c. As a result, in each, such a system the stationary workload, D, is distributed in accordance with the following version of the Pollaczek–Khintchine formula [15]

$$D = \sum_{i=1}^{L} S_i^{(e)}, \tag{29}$$

where L has geometric distribution

$$P(L = k) = (\frac{\rho}{c})^k (1 - \frac{\rho}{c}), \tag{30}$$

and $S^{(e)}$ has the so-called equilibrium (integrated tail) distribution,

$$\mathrm{P}\left(S^{(e)} > x\right) = \frac{1}{\mathrm{E}S}\int_x^\infty \overline{F}_S(t)dt, \tag{31}$$

where F_S is the d.f. of original service time S. It is well-known that both workload process and queue size process in the RA system dominate the corresponding process in the original $M/G/c$ system [5,24,27]. Applying coupling, this dominance holds with probability (w.p.) 1. In particular, the regenerations of RA system (the instants when customers meet totally idle system) are also regeneration instants of the original system $M/G/c$. These results then are used to sample from the steady-state distribution of the RA system as follows:

1. sample the values $L_i, i = 1, \ldots, c$ according to geometric distribution Equation (30);
2. sample $S_1^{(e)}, \ldots, S_{L_i}^{(e)}, i = 1, \ldots, c$ according to integrated tail distribution Equation (31); and,
3. construct the (stationary) components D_i for $i = 1, \ldots, c$, by formula Equation (29).

Subsequently, starting from the steady-state vector $V_1 = (D_1, \ldots, D_c)$ containing iid components, the recursion Equation (28) is applied to each separate queue in the RA system until the event

$$V_{\tau_e} = (V_{\tau_e,1}, \ldots, V_{\tau_e,c}) = \mathbf{0}$$

happens at the (arrival) instant of some customer τ_e. Thus, τ_e is the length of equilibrium (steady-state) remaining regeneration period. Note that by construction, at each step of this recursion, the workload vector has steady-state distribution in the RA system. Omitting unnecessary details, the remaining steps of algorithm are as follows [15]:

1. sample stochastic copies $V^{(k)} = (V_1^{(k)}, V_2^{(k)}, \ldots)$, $k = 1, 2, \ldots$ of the sequence of workload vectors using recursion Equation (28); each sequence starts with $V_1^{(k)} = 0$ and lasts until the event $V_{\tau(k)}^{(k)} = 0$ happens at some instant $\tau(k)$; note that $\{\tau(k)\}$ are iid random variables distributed as a generic regeneration period τ of RA system;
2. repeat previous step until the event $\tau(j) > \tau_e$ happens in some sample $V^{(j)} = (V_1^{(j)}, V_2^{(j)}, \ldots)$; and,
3. the value $V_{\tau_e}^{(j)}$ of the workload vector $V^{(j)}$ at instant τ_e, has the target steady-state distribution of the workload in the original $M/G/c$ system.

We note that, although this approach allows to sample exactly from the steady-state distribution, the regeneration period in the dominated RA system can be very large in practice, and, thus, can lead to unacceptable long simulation. For further details on perfect sampling, see [15,35–37].

Now we explain how to sample from the equilibrium distribution of a two-component mixture. Let $\overline{F}_M$ be the tail of a two-component mixture Equation (3). Subsequently, it follows from Equation (3) that

$$\overline{F}_M^{(e)}(x) = \frac{1}{\mathrm{E}X_M}\int_x^\infty \overline{F}_M(u)du = \frac{p\mathrm{E}X_1}{\mathrm{E}X_M}\overline{F}_1^{(e)}(x) + \frac{(1-p)\mathrm{E}X_2}{\mathrm{E}X_M}\overline{F}_2^{(e)}(x). \tag{32}$$

It is clearly seen from Equation (32) that the equilibrium distribution of a mixture is itself a two-component mixture of equilibrium distributions of the components. Thus, to sample from the equilibrium distribution (32), we sample from $F_1^{(e)}$ w.p. $q = p\mathrm{E}X_1/\mathrm{E}X_M$, and sample from $F_2^{(e)}$ w.p. $1-q$. Finally, note that, as easy to see, if the original distributions are $Pareto(\alpha_i, x_0)$, then $\overline{F}_i^{(e)}$ are also $Pareto(\alpha_i - 1, x_0), i = 1, 2$ (also see [38]).

5. Simulation Results

As a sanity check of the perfect sampling $M/G/c$ model, we validate the algorithm via the $M/M/c$ system having input rate $\lambda = 7.5$, service rate $\mu = 1.5$, $c = 10$ servers, and $\rho = \lambda/\mu = 5$.

We run $N = 5000$ samples from steady-state process using perfect simulation and build the empirical queue size distribution vs. theoretical values that were obtained from Equation (22). We depict the results of validation on Figure 2. Note that the uniform distance between the empirical and theoretical distributions is 0.0091.

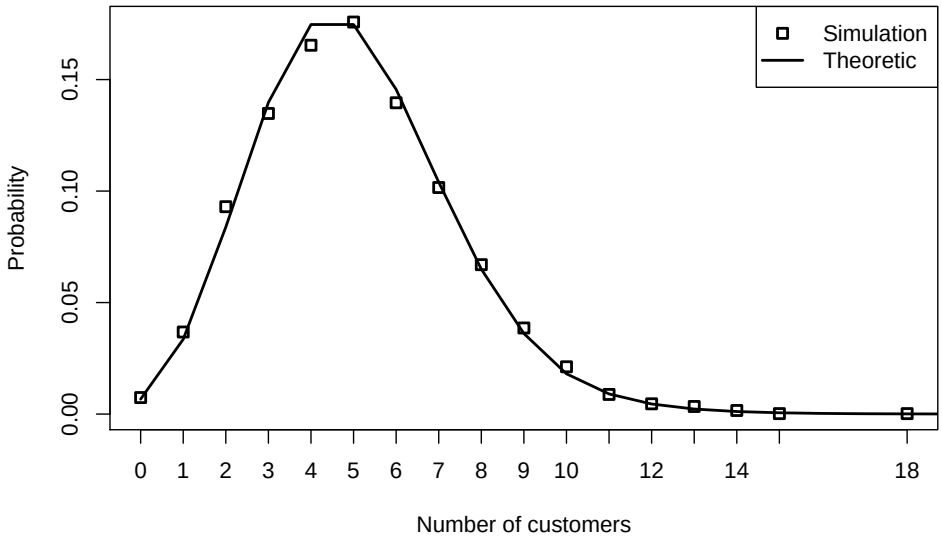

Figure 2. Theoretical distribution of the steady-state queue size in an $M/M/10$ system vs.empirical distribution ($N = 5000$ samples), with input rate $\lambda = 7.5$, service rate $\mu = 1.5$. The uniform distance between the theoretical and estimated queue size distributions equals $\Delta = 0.0091$.

5.1. Experiment 1: Hyperexponential Case

Now, we step away from the basic Markovian case, $M/M/c$, having service time distribution $F_1(x) = 1 - e^{-\mu_1 x}$, by introducing a contaminated system $M/G/c$ with generic service time, $S^{(M)}$, having two-state hyperexponential distribution, H_2, with $F_i(x) = 1 - e^{-\mu_i x}$, and mixing coefficient $p \in (0,1)$. Note that such a case has computationally tractable solution, see [39]. However, we use the perfect sampling algorithm to check the accuracy of the sensitivity analysis. We fix

$$\mu_1 = 2,\ c = \lambda = 5,\ p = 0.7,$$

and vary μ_2 over range $(2, 8]$ with step 0.4. We obtain the empirical queue size d.f., $\hat{F}_{Q^{(1)}}$, in the basic, and $\hat{F}_{Q^{(M)}}$ in the contaminated system, and construct Equation (27) for each combination of the parameters while using N = 10,000 samples from the steady-state distribution. The linear dependence of $\Delta(\hat{F}_{Q^{(M)}}, \hat{F}_{Q^{(1)}})$ on $\Delta(F_M, F_1)$ is clearly seen in Figure 3.

5.2. Experiment 2a: Pareto Case, Sensitivity to Mixing Parameter

In the following experiments, we use an $M/G/c$ system with $c = 4$, load $\rho = 0.5$, and $Pareto(\alpha_1, 1)$ service time d.f., with $\alpha_1 = 2.1$ as the basic system for comparison. The input rate of the basic system is taken as $\lambda = \rho c(\alpha_1 - 1)$ so as to guarantee the desired load $\rho = 0.5$. Note that, to the best of our knowledge, there is no explicit expression for the steady-state queue size in such a system, and thus simulation is used to obtain the corresponding estimates of the steady-state queue size d.f. To obtain such an estimate, N = 10,000 samples from the corresponding steady-state distribution are obtained by the perfect sampling technique described in Section 4.

In the first experiment, we study the steady-state queue size distribution sensitivity to the mixing parameter, p. The mixing coefficient is iterated over the discrete values $p = 0.95, 0.9, \ldots, 0.25$, and the empirical steady-state queue size d.f., $F_{Q_p^{(M)}}$, is constructed for the disturbed system with mixture service time d.f., F_M given in Equation (3), consisting of $Pareto(\alpha_1, 1)$ and $Pareto(\alpha_2, 1)$ with mixing

parameter p, where $\alpha_2 = 4.9$. The input rate λ is fixed at the level $\lambda = 2.2$, so as to guarantee the load $\rho = 0.5$ in the basic system. Note that the parameter p is varied in such a way that the mixing proportion of $Pareto(\alpha_2, 1)$ distribution becomes larger with smaller p, and dominates the $Pareto(\alpha_1, 1)$, for $p < 0.5$. Finally, we plot the values $\Delta(F_M, F_1)$ vs. $\Delta(\hat{F}_{Q_p^{(M)}}, \hat{F}_{Q^{(1)}})$ for the values p given. The results are depicted in Figure 4. Note that the dependence of the distance is approximately linear in mixing probability, p.

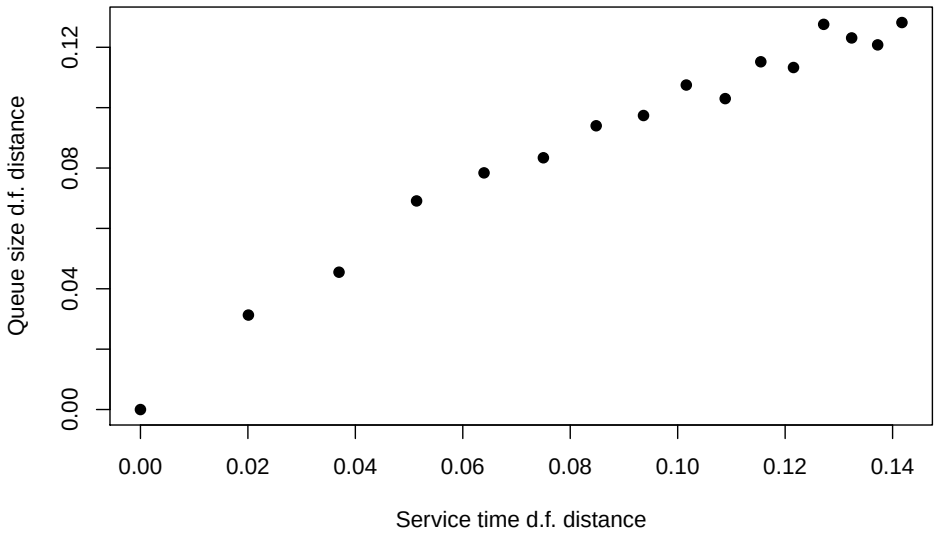

Figure 3. Distance, $\Delta(\hat{F}_{Q^{(M)}}, \hat{F}_{Q^{(1)}})$, between the empirical queue size d.f. in a basic $M/M/5$ system with input rate $\lambda = 5$, service rate $\mu_1 = 2$, compared to a contaminated $M/H_2/5$ system with input rate $\lambda = 5$ and hyperexponential service times being a mixture with $\mu_1 = 2$ and $\mu_2 = 2, 2.4, \ldots, 8$, $p = 0.7$, obtained from N = 10,000 samples, vs. service time d.f. distance, $\Delta(F_M, F_1)$.

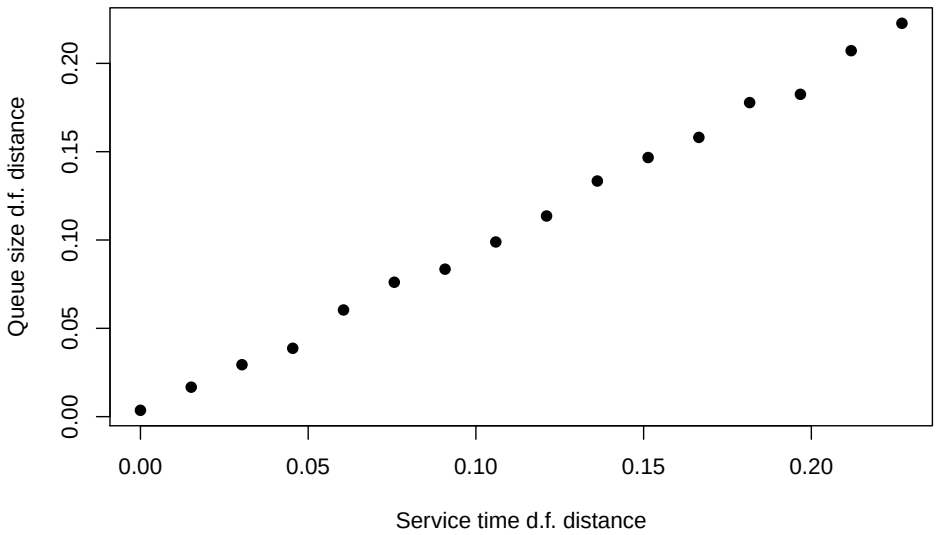

Figure 4. Distance between the empirical queue size d.f. in a basic $M/G/c$ system with $c = 4$, $\rho = 0.5$, F_1 being $Pareto(2.1, 1)$ service time d.f. and $\lambda = 2.2$, and system with a mixture, F_M of $Pareto(2.1, 1)$ and $Pareto(4.9, 1)$ service time d.f. vs. the distance between F_1 and F_M, for varying $p = 1, 0.95, \ldots, 0.25$.

5.3. Experiment 2b: Pareto Case, Sensitivity to Contaminating Distribution

In the following experiment, we study the sensitivity of the steady-state queue size distribution on the parameter α_2 of the mixture. Now $p = 0.7$ is fixed, and α_2 is iterated over the discrete set $\alpha_2 \in \{2.1, 2.3, \ldots, 4.9\}$, ceteris paribus. As in the previous experiment, we build the empirical

steady-state queue size distribution of the basic system, $\hat{F}_{Q^{(1)}}$, by exact sampling from steady state using the method described in Section 4. We plot the values $\Delta(F_M, F_1)$ vs. $\Delta(\hat{F}_{Q^{(M)}_{\alpha_2}}, \hat{F}_{Q^{(1)}})$ for the given values of α_2 as a parametric functions of α_2. The results are depicted in Figure 5, where, unlike the previous scenarios, the nonlinear dependence on α_2 is clear.

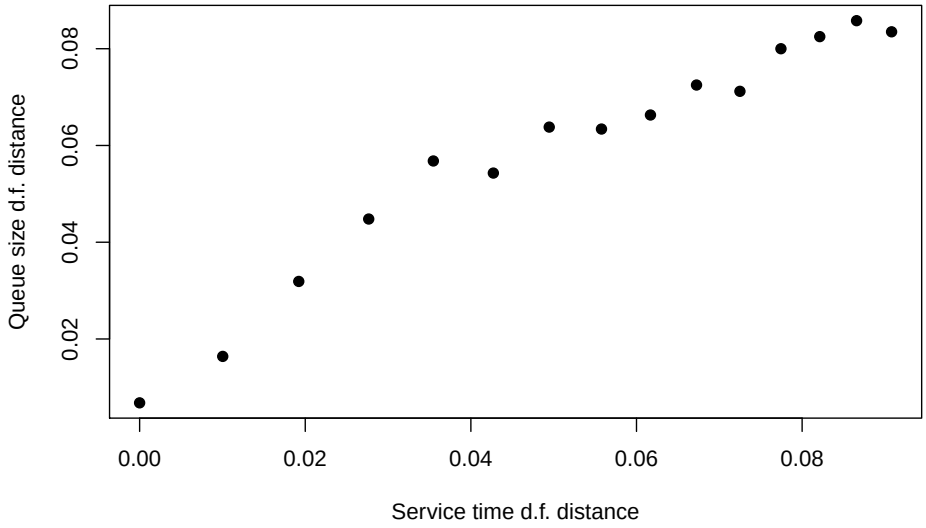

Figure 5. Distance between the empirical queue size d.f. in a basic $M/G/c$ system with $c = 4$, $\rho = 0.5$, F_1 being $Pareto(2.1, 1)$ service time d.f. and $\lambda = 2.2$, and system with a mixture, F_M of $Pareto(2.1, 1)$ and $Pareto(\alpha_2, 1)$ service time d.f. vs. the distance between F_1 and F_M, for fixed $p = 0.7$ and varying $\alpha_2 = 2.1, 2.3, \ldots, 4.9$.

Note that the non-linear dependence of $\Delta(F_{Q^{(M)}}, F_{Q^{(1)}})$ on α_2 may be caused by the non-linear dependence of the distance of service time distributions, $\Delta(F_M, F_1)$, on α_2, see Figure 1. Moreover, the mean service time, $S^{(M)}$, also differs from mean service time of the basic system, which causes appropriate changes in the load, ρ, in the disturbed system, see Figure 6.

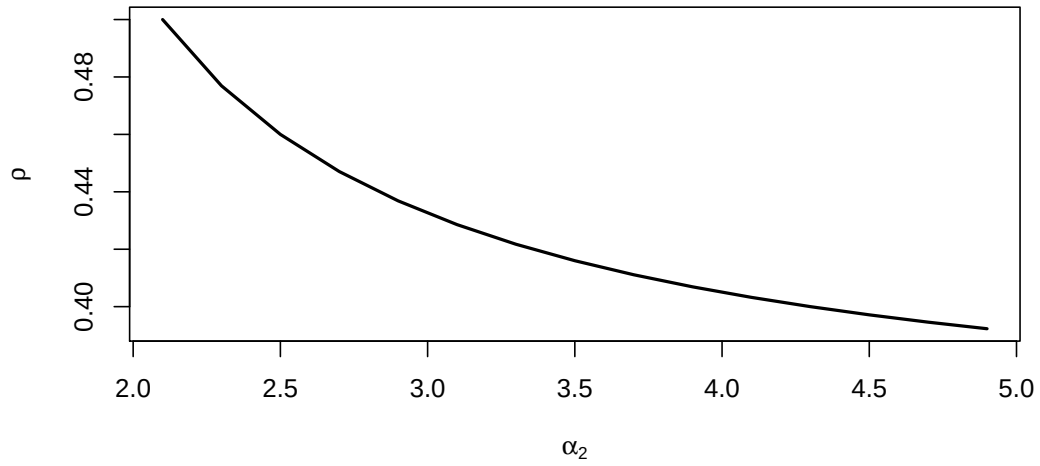

Figure 6. Dependence of the system load, ρ, on the parameter $\alpha_2 = 2.1, 2.3, \ldots, 4.9$ of the mixture distribution in an $M/G/c$ system with $c = 4$, $\lambda = 2.2$, mixture, F_m of $Pareto(2.1, 1)$ and $Pareto(\alpha_2, 1)$ service time d.f. with mixing coefficient $p = 0.7$.

Using the results of Experiment 2b, we illustrate the stochastic monotonicity property Equation (25) for selected values of parameter α_2. Figure 7 depicts the results.

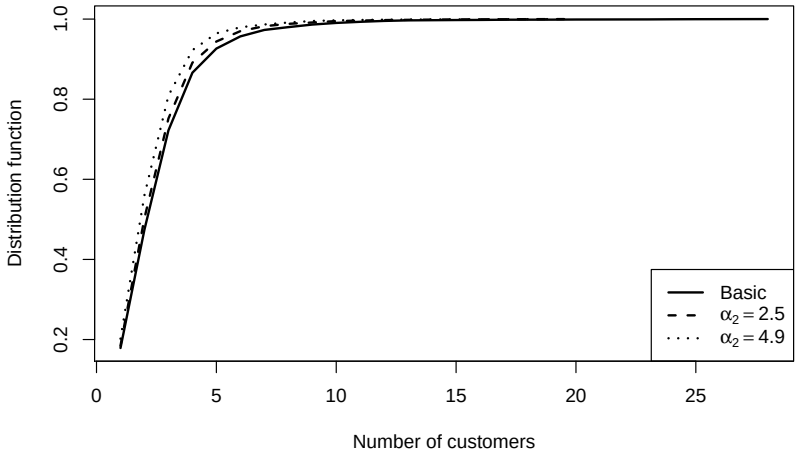

Figure 7. Stochastic monotonicity of the system output, in terms of steady-state queue size d.f., on the parameter $\alpha_2 = 2.1, 2.5, 4.9$ of the mixture distribution in an $M/G/c$ system with $c = 4$, $\lambda = 2.2$, mixture, F_M of $Pareto(2.1, 1)$ and $Pareto(\alpha_2, 1)$ service time d.f. with mixing coefficient $p = 0.7$.

5.4. Experiment 2c: Pareto Case, Constant Load

In the final experiment, we study the joint effect of both the parent and the contaminating (Pareto) distributions. To do so, we change $\alpha_1 = 2.1, 2.5, \ldots, 4.9$, vary $\alpha_2 = \alpha_1, \alpha_1 + 0.4, \ldots, 4.9$. To mitigate the effect of changing load illustrated by Figure 6, we simultaneously change the parameter λ, so as to guarantee constant load $\rho = 0.5$ for all systems, keeping $p = 0.7, c = 4$ constant. The comparison is done to the system with the parent distribution of $\alpha_1 = 2.1$ of the service times. Each point is obtained then by N = 10,000 samples by the perfect sampling technique. Figure 8 depicts the results, where the color reflects the parent distribution parameter, α_1, and size of a dot is proportional to α_2. With increasing distance of the parent distribution from the contaminating distribution, the distance changes in a linear manner. Moreover, increased α_1 changes the starting point (which in all lines corresponds to the parent distribution with parameter α_1), and increasing α_2 for fixed α_1 increases the distance both in the input (for the mixture) and performance index (queue size distribution distance). Interestingly, for the lower line that corresponds to the fixed $\alpha_1 = 2.1$ and varying α_2, there seems to be a slightly negative slope, which is likely to be the result of an increasing variance and, hence, decreasing accuracy. However, this effect might be interesting to study separately in the future.

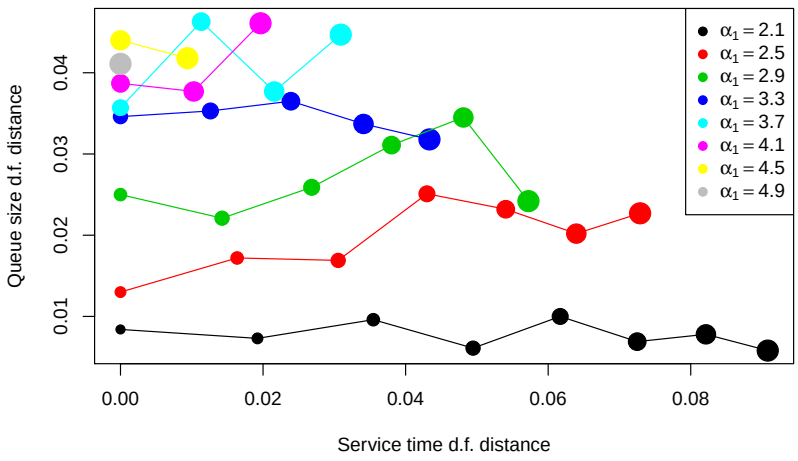

Figure 8. Distance between the empirical queue size d.f. in a basic $M/G/c$ system with $c = 4$, F_1 being $Pareto(\alpha_1, 1)$ service time d.f., and system with a mixture, F_M of $Pareto(\alpha_1, 1)$ and $Pareto(\alpha_2, 1)$ service time d.f. vs. the distance between F_1 and F_M, for fixed $p = 0.7$, fixed $\rho = 0.5$, varying $\alpha_1 = 2.1, 2.4, \ldots, 4.9$ (color), varying $\alpha_2 = \alpha_1, \alpha_1 + 0.4, \ldots, 4.9$ (dot size), and varying λ, so as to fix the load, ρ.

Finally, we note that, to speedup the computation, we used parallel computation of the uniform distance for various system configurations using the resources of the High-Performance Datacenter of Karelian Research Centre of Russian Academy of Sciences.

6. Conclusions and Discussion

In this paper, the effect of the service time distribution perturbation on the steady-state performance measures of a multiserver queueing system is studied. The explicit form for the sensitivity measure (Kolmogorov-Smirnov distance) between the service time distribution functions was obtained, and the performance estimates were obtained by the regenerative perfect simulation technique. The simulation results outline the qualitative nature of the sensitivity, which is, in most cases, linear (possibly after appropriate scaling of the input rate to guarantee the constant load).

The approach to sensitivity analysis that is presented in this paper can be applied to more sophisticated, and more practically oriented systems, such as the simultaneous service multiserver system [40], which would result, though, in an increased dimension of the system state. However, we note that the steady-state exact sampling by regenerative simulation has several serious drawbacks. First, the average working time of the algorithm may be infinite [36], e.g., in a system with large number of servers (which indeed depends on the regenerative cycle length). This problem can be solved either by the coupling-from-the-past technique [35] (which, although, is rather technically tricky), or by non traditional regenerative techniques, such as the artificial regeneration [41] or regenerative envelopes [40]. Finally, the study may be extended to larger classes of service time distributions. At that, we leave these as opportunities for future research.

Author Contributions: Conceptualization, writing—original draft preparation, I.P.; simulation and visualization, A.R.; writing—review and editing, M.P.; supervision, project administration—E.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research is supported by Russian Foundation for Basic Research, projects No. 19-57-45022, 19-07-00303, 18-07-00156, 18-07-00147.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

d.f. distribution function
w.p. with probability

References

1. Al-Hussaini, E.K.; Sultan, K.S. Reliability and hazard based on finite mixture models. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2001; Volume 20, pp. 139–183. [CrossRef]
2. Shaked, M.; Spizzichino, F. Mixtures and monotonicity of failure rate functions. In *Handbook of Statistics*; Elsevier: Amsterdam, The Netherlands, 2001; Volume 20, pp. 185–198. [CrossRef]
3. Sevast'yanov, B.A. An Ergodic Theorem for Markov Processes and Its Application to Telephone Systems with Refusals. *Theory Probab. Its Appl.* **1957**, *2*, 104–112. [CrossRef]
4. Kalashnikov, V.V. Stability Analysis of in Queueing Problems by a Method of Trial Functions. *Theory Probab. Its Appl.* **1977**, *22*, 86–103. [CrossRef]
5. Müller, A.; Stoyan, D. *Comparison Methods for Stochastic Models and Risks*; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2002.
6. Zolotarev, V.M. On the stochastic continuity of the queuing systems of type G | G | 1. *Theory Probab. Its Appl.* **1977**, *21*, 250–269. [CrossRef]
7. Zolotarev, V.M. Quantitative estimates for the continuity property of queueing systems of type G | G | ∞. *Theory Probab. Its Appl.* **1978**, *22*, 679–691. [CrossRef]
8. Zolotarev, V.M. Qualitative Estimates in Problems of Continuity of Queuing Systems. *Theory Probab. Its Appl.* **1975**, *20*, 211–213. [CrossRef]

9. Batrakova, D.; Korolev, V.; Shorgin, S. A new method for the probabilistic and statistical analysis of information flows in telecommunication networks. *Inform. Appl.* **2007**, *1*, 40–53.
10. Daley, D.J. Queueing Output Processes. *Adv. Appl. Probab.* **1976**, *8*, 395. [CrossRef]
11. Daley, D.J. Revisiting queueing output processes: A point process viewpoint. *Queueing Syst.* **2011**, *68*, 395–405. [CrossRef]
12. Korolev, V.Y.E.; Krylov, V.A.; Kuz'min, V.Y.E. Stability of finite mixtures of generalized Gamma-distributions with respect to disturbance of parameters. *Inform. Appl.* **2011**, *5*, 31–38.
13. Kalashnikov, V.V.; Tsitsiashvili, G.S. Stability analysis of queueing systems. *J. Sov. Math.* **1981**, *17*, 2238–2255. [CrossRef]
14. McLachlan, G.J.; Lee, S.X.; Rathnayake, S.I. Finite Mixture Models. *Annu. Rev. Stat. Its Appl.* **2019**, *6*, 355–378. [CrossRef]
15. Sigman, K. Exact simulation of the stationary distribution of the FIFO M/G/c queue: the general case for $\rho < c$. *Queueing Syst.* **2012**, *70*, 37–43. [CrossRef]
16. Feitelson, D.G. *Workload Modeling for Computer Dystems Performance Evaluation*; Cambridge University Press: Cambridge, UK, 2014.
17. Morozov, E.; Peshkova, I.; Rumyantsev, A. On Failure Rate Comparison of Finite Multiserver Systems. In *Distributed Computer and Communication Networks*; Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V., Eds.; Springer International Publishing: Cham, Germany, 2019; Volume 11965, pp. 419–431._32. [CrossRef]
18. Marshall, A.W.; Olkin, I. *Life Distributions: Structure of Nonparametric, Semiparametric, and Parametric Families*; Springer Series in Statistics; Springer: New York, NY, USA; London, UK, 2007.
19. Goldstein, M. Contamination Distributions. In *The Annals of Statistics*; Institute of Mathematical Statistics: Beachwood, OH, USA, 1982; Volume 10, pp. 174–183.
20. Block, H.W. The Failure Rates of Mixtures. In *Advances in Distribution Theory, Order Statistics, and Inference*; Balakrishnan, N., Sarabia, J.M., Castillo, E., Eds.; Birkhäuser Boston: Boston, MA, USA, 2006; pp. 267–277._17. [CrossRef]
21. Barlow, R.E.; Proschan, F. *Mathematical Theory of Reliability*; Classics in Applied Mathematics; Society for Industrial and Applied Mathematics: Philadelphia, PA, USA, 1996. [CrossRef]
22. Goldie, C.M.; Klüppelberg, C. Subexponential Distributions. In *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*; Birkhauser Boston Inc.: Cambridge, MA, USA, 1998; pp. 435–459.
23. Shaked, M. Bounds on the Distance of a Mixture from Its Parent Distribution. *J. Appl. Probab.* **1981**, *18*, 853–863. [CrossRef]
24. Asmussen, S. *Applied Probability and Queues*; Springer: New York, NY, USA, 2003.
25. Kiefer, J.D.; Wolfowitz, J. On the theory of queues with many servers. *Trans. Am. Math. Soc.* **1955**, *78*, 1–18. [CrossRef]
26. Kleinrock, L. *Theory, Volume 1, Queueing Systems*; Wiley-Interscience: Hoboken, NJ, USA, 1975.
27. Whitt, W. Comparing counting processes and queues. *Adv. Appl. Probab.* **1981**, *13*, 207–220. [CrossRef]
28. Thorrison, H. *Coupling, Stationarity, and Regeneration*; Springer: New York, NY, USA, 2000.
29. Shaked, M.; Shanthikumar, J.G. *Stochastic Orders*; Springer Series in Statistics; Springer: New York, NY, USA, 2007.
30. Whitt, W. Approximations for the GI/G/M Queue. *Prod. Oper. Manag.* **1993**, *2*, 114–161. [CrossRef]
31. Van Hoorn, M.; Tijms, H. Approximations for the waiting time distribution of the M/G/c queue. *Perform. Eval.* **1982**, *2*, 22–28. [CrossRef]
32. Ma, B.N.W.; Mark, J.W. Approximation of the Mean Queue Length of an M/G/c Queueing System. *Oper. Res.* **1995**, *43*, 158–165. [CrossRef]
33. Kimura, T. Approximations for multi-server queues: system interpolations. *Queueing Syst.* **1994**, *17*, 347–382. [CrossRef]
34. Gupta, V.; Harchol-Balter, M.; Dai, J.G.; Zwart, B. On the inapproximability of M/G/K: Why two moments of job size distribution are not enough. *Queueing Syst.* **2010**, *64*, 5–48. [CrossRef]
35. Blanchet, J.; Pei, Y.; Sigman, K. Exact sampling for some multi-dimensional queueing models with renewal input. *Adv. Appl. Probab.* **2019**, *51*, 1179–1208. [CrossRef]
36. Xiong, Y. Perfect and Nearly Perfect Sampling of Work-Conserving Queues. Ph.D. Thesis, The School of Graduate and Postdoctoral Studies, The University of Western Ontario, London, ON, Canada, 2015.
37. Blanchet, J.; Dong, J.; Pei, Y. Perfect Sampling of GI/GI/c Queues. *arXiv* **2015**, arXiv: 1508.02262.

38. Nair, N.U.; Preeth, M. On some properties of equilibrium distributions of order n. *Stat. Methods Appl.* **2009**, *18*, 453–464. [CrossRef]
39. de Smit, J.H. A numerical solution for the multi-server queue with hyper-exponential service times. *Oper. Res. Lett.* **1983**, *2*, 217–224. [CrossRef]
40. Morozov, E.; Peshkova, I.; Rumyantsev, A. On Regenerative Envelopes for Cluster Model Simulation. In Proceedings of the Distributed Computer and Communication Networks: 19th International Conference, DCCN 2016, Moscow, Russia, 21–25 November 2016; Vishnevskiy, V.M., Samouylov, K.E., Kozyrev, D.V., Eds.; Springer International Publishing: Cham, Germany, 2016; pp. 222–230._20. [CrossRef]
41. Rumyantsev, A.; Peshkova, I. Artificial Regeneration Based Regenerative Estimation of Multiserver System with Multiple Vacations Policy. In *Information Technologies and Mathematical Modelling. Queueing Theory and Applications*; Dudin, A., Nazarov, A., Moiseev, A., Eds; Springer International Publishing: Berlin/Heidelberg, Germany, 2019; Volume 1109, pp. 38–50._4. [CrossRef]