Efficient simulation of tail probabilities in a queueing model with heterogeneous servers

Kuhn, J.; Mandjes, M.

DOI
10.1080/15326349.2018.1458629

Publication date
2018

Document Version
Final published version

Published in
Stochastic Models

License
CC BY-NC-ND

Link to publication

Citation for published version (APA):
Kuhn, J., & Mandjes, M. (2018). Efficient simulation of tail probabilities in a queueing model with heterogeneous servers. *Stochastic Models*, *34*(2), 239-267. https://doi.org/10.1080/15326349.2018.1458629

STOCHASTIC MODELS
2018, VOL. 34, NO. 2, 239–267
https://doi.org/10.1080/15326349.2018.1458629

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Efficient simulation of tail probabilities in a queueing model with heterogeneous servers

Julia Kuhn[a,b] and Michel Mandjes[a]

[a]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands;
[b]School of Mathematics and Physics, The University of Queensland, St. Lucia, QLD, Australia

ABSTRACT

This paper considers a multi-server queue with Markov-modulated Poisson input and server-dependent phase-type service times. We develop an efficient rare-event simulation technique to estimate the probability that the number of customers in this system reaches a high value. Relying on explicit bounds on the probability under consideration as well as the associated likelihood ratio, we succeed in proving that the proposed estimator is of bounded relative error. Simulation experiments illustrate the significant speed-up that can be achieved by the proposed algorithm.

ARTICLE HISTORY
Received April 2017
Accepted March 2018

KEYWORDS
Importance sampling; multi-server queues; overdispersion; rare events

MATHEMATICS SUBJECT CLASSIFICATION
60K25; 60F10

1. Introduction

The multi-server queue is a well-studied object in operations research with widespread applications, for example in the modeling of call centres [11] and health-care systems[9]. In many situations, the system needs to be designed in such a way that the service level offered is sufficiently high. This is usually translated into the requirement that the probability of the backlog exceeding some critical value should be below a given threshold value.

For the case of homogeneous servers (meaning that the service times at the various servers have a common distribution), a strand of research focuses on evaluating the probability that the number of customers waiting exceeds some high level K. A key result in this area concerns the situation in which the service-time distribution has a finite moment generating function around the zero (implying that all moments exist): it was proven by Sadowsky[20] that for such GI/GI/m queues the tail of the probability of interest decays effectively exponentially, cf. also the earlier paper by Takahashi[21] for the setting with phase-type interarrival times and phase-type service times. In addition, Ref.[20] provides a fast (importance-sampling based) simulation procedure to estimate this probability with provable optimality properties. More specifically, it was shown that the estimator is logarithmically

CONTACT Michel Mandjes m.r.h.mandjes@uva.nl Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 105, 1098 XH Amsterdam, The Netherlands.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/lstm.

efficient; this entails that the number of runs needed to obtain an estimate with a given precision grows sub-exponentially in the level K.

In the above literature, it was assumed that the servers are homogeneous; this implies, for example, that each service entity serves customers at the same average speed. In many practical situations, however, this assumption is overly restrictive as has been recognized in the work of, e.g., Refs.[1,14] (as well as in other references that deal with the problem of routing in systems with heterogeneous servers). Not much is known, however, about the tail distribution of such heterogeneous multi-server systems.

Another aspect that is hardly covered in the importance-sampling literature concerns the incorporation of *overdispersion*. Traditionally the arrival process is modeled as a Poisson process (or more generally a renewal process), implying that the mean and variance of the number of arrivals in an interval of given length coincide. Empirical studies, however, indicate that the Poisson process is not always appropriate to model the arrival stream's intrinsic variability. Arrival data turns out to be often "overdispersed," meaning that that the variance of the number of arrivals in an interval of given length is significantly larger than the corresponding mean value; see, e.g., Refs.[10,11,15,23]. This phenomenon is better captured by a *Cox process*, which is a Poisson process but now with a randomly evolving (rather than fixed) parameter. The traditional example of such an overdispersed arrival process is the Markov-modulated Poisson (MMP) process. For an MMP process the arrival rate is λ_i when an independently evolving continuous-time, finite-state Markov chain (typically referred to as the *background process*) is in state i. For results on queues with Markov-modulated input we refer to, e.g., Ref.,[2 Ch. XI].

Motivated by the above considerations, the object of study in this paper is the multi-server queue with *MMP input* and *server-dependent* phase-type service times. The main contribution is that we devise efficient simulation techniques for the purpose of estimating the tail distribution of the number of customers in the system. In more detail, our work extends the existing literature on importance sampling for multi-server queues are as follows.

- In our set-up we allow the servers to be heterogeneous, whereas[20,21] assume server-homogeneity. We remark that Ref.[20] considers light-tailed service-time distributions, whereas we focus on the subclass of phase-type distributions. Recall, however, that general non-negative distributions can be approximated arbitrarily closely by phase-type distributions so that in practical terms hardly any generality is lost; see, e.g., Refs.[4,22] and Ref.,[2 Thm. III.4.2]. The focus is still on light-tailed distributions as for heavy-tailed distributions the number of phases needed to adequately model the tails may be excessively large.
- In addition, we allow for the arrival process to be overdispersed. We focus on the case of MMP arrivals, but, as we will point out, other types of arrival processes can be treated with similar techniques (such as the renewal processes that were studied in Ref.[20]).

- We show that our proposed importance-sampling estimator is *strongly efficient*, (or, equivalently, has *bounded relative error*). This means that the number of runs needed to obtain an estimate with given precision remains bounded (i.e., is smaller than some constant that does not depend on K). Recall that in Ref.[20] just logarithmic efficiency was proven (implying that the number of runs needed grows sub-exponentially).

In summary, our model can be viewed as a generalization of that of Ref.[20] in that we allow for heterogeneous servers as well as overdispersed arrival processes; the (minor) sacrifice that we make is that we assume the service times to be of phase-type, rather than just light-tailed. In more detail, the results obtained are the following.

(i) In the first place, for the queue under study, we propose efficient simulation algorithms for the estimation of the probability that the backlog (that is, the number of customers or jobs waiting in the queue) during a busy cycle (during which the system is non-empty) exceeds a given level K. The procedure can be modified for the estimation of related quantities such as the fraction of customers or jobs entering the system while the backlog is larger than K, or the fraction of customers lost in the corresponding model with a waiting room of finite size K.

The algorithms are based on importance sampling, that is, the model is simulated under an alternative measure, under which the event under consideration is *not* rare. We identify an efficient change of measure by solving a particular eigensystem. As it turns out, this change of measure provides us with upper and lower bounds on the probability of interest that are both exponential (with decay rate $\vartheta^\star$) in the exceedance level K (and that match up to a multiplicative constant). This property can be used to prove that our importance-sampling estimator is strongly efficient.

(ii) Since the eigensystem to be solved depends on both arrival and departure processes jointly, it can become prohibitively large when the dimension of the background process and/or the dimensions of the phase-type distributed service times grow large. In order to resolve this issue, we exploit the fact that there is an alternative way to compute the decay rate $\vartheta^\star$ that relies on properties of the arrival process and service processes in a decoupled fashion.

(iii) Finally, we point out how the change of measure can be found for various variants of the arrival and service processes.

The organization of this paper is as follows. In Section 2, we introduce the model and formulate our objectives in greater detail. In Section 3, we propose the change of measure that is to be used in the importance-sampling based procedure. We then establish bounds on the probability of interest, which we use to prove that the importance-sampling algorithm has bounded relative error. In Section 4, we show that the same change of measure can be obtained when considering the arrival and service processes separately, thus drastically reducing the computational effort needed to compute the change of measure. Section 5 contains illustrative numerical experiments that give an impression of the typically achievable speed-up. We conclude in Section 6 by discussing how the importance sampling algorithm can be adapted to estimate related quantities.

2. Framework

2.1. Model

In this paper, we primarily focus on the following MMP/Ph/m with heterogeneous servers. We now introduce the arrival process and service processes used.

Arrival process. Consider the following MMP arrival process. The transition rate matrix of the (finite-state) background process $(I_t)_{t\geq 0}$ is $Q = (q_{ij})_{i,j=1}^{d}$; define $q_i := -q_{ii} = \sum_{j\neq i} q_{ij}$. When the background process (assumed to be irreducible) is in state i arrivals occur according to a Poisson process with rate $\lambda_i \geqslant 0$. Let the mean arrival rate be $\lambda := \boldsymbol{\pi}'\boldsymbol{\lambda}$, with $\boldsymbol{\pi}$ the invariant probability measure of the background process, $\boldsymbol{\pi}'$ its transpose, and $\boldsymbol{\lambda} := (\lambda_1, \dots, \lambda_d)'$.

Service processes. There are m heterogeneous servers. Service times at server $\ell \in \{1, \dots, m\}$ are i.i.d. samples distributed as the non-negative random variable $B^{(\ell)}$. We let $B^{(\ell)}$ be of phase-type[2, Ch. III] with initial distribution $\boldsymbol{\alpha}^{(\ell)}$ concentrated on the transient states, and transition rate matrix

$$T^{(\ell)} = \left(t_{ij}^{(\ell)}\right)_{i,j=1}^{D^{(\ell)}+1} = \begin{pmatrix} S^{(\ell)} & \boldsymbol{s}^{(\ell)} \\ \boldsymbol{0} & 0 \end{pmatrix}, \tag{1}$$

for some $D^{(\ell)} \in \mathbb{N}$. We impose the usual requirement that $t_{ij}^{(\ell)} \geqslant 0$ for $i \in \{1, \dots, D^{(\ell)}\}$ and $j \in \{1, \dots, D^{(\ell)} + 1\}$ with $i \neq j$, and $t_{D^{(\ell)}+1,j} = 0$ for $j \in \{1, \dots, D^{(\ell)} + 1\}$; in addition we define

$$t_i^{(\ell)} := -t_{ii}^{(\ell)} = \sum_{j=1, j\neq i}^{D^{(\ell)}+1} t_{ij}^{(\ell)},$$

so that the row sums are zero. In words, the above means that the service time at server ℓ stays in phase i for an exponentially distributed amount of time with mean $(t_i^{(\ell)})^{-1}$, and then jumps to state $j \neq i$ with probability $t_{ij}^{(\ell)}/t_i^{(\ell)}$.

Thus, the evolution of the system is recorded by the following trivariate process:

(i) The state of the background process $(I_t)_{t\geqslant 0}$ taking values in $\mathscr{I} := \{1, \dots, d\}$.

(ii) The state vector $(\boldsymbol{J}_t)_{t\geqslant 0}$ of the phase-type distributions of the customers in service; with † indicating that the corresponding server is idle, this takes values in

$$\mathscr{D} := \{1, \dots, D^{(1)}, \dagger\} \times \cdots \times \{1, \dots, D^{(m)}, \dagger\}.$$

We will sometimes use the suggestive notation $t_{i,\dagger}^{(\ell)} := t_{i,D^{(\ell)}+1}^{(\ell)}$.

(iii) The number of customers in the system, $(N_t)_{t\geqslant 0}$, taking values in $\mathbb{N} = \{0, 1, \ldots\}$. We stress that this number includes the customers in service: when $N_t = m + n$, then all servers are occupied, and n customers are waiting.

Observe that $(I_t, \boldsymbol{J}_t, N_t)_{t\geqslant 0}$ is a continuous-time Markov chain on the state space $\mathscr{I} \times \mathscr{D} \times \mathbb{N}$. Throughout the queue is assumed to be stable, i.e., we impose the condition

$$\lambda < \sum_{\ell=1}^{m} \frac{1}{\mathbb{E}\, B^{(\ell)}},$$

where $\mathbb{E}\, B^{(\ell)}$ can be evaluated in terms of $\boldsymbol{\alpha}^{(\ell)}$ and $T^{(\ell)}$ as in Ref.[2, Prop. III.4.1]. This stability criterion can be interpreted as the average number of clients arriving to the multi-server queue per unit of time should be strictly majorized by the average number of clients that can be served (by the m queues together) per unit of time.

Since servers are heterogeneous, we shall assume that the free server with the lowest index serves the next customer arriving or waiting in the queue. In practice, one may wish to prioritize faster servers; a service policy of this type can be achieved by labeling servers in increasing order according to their average service times.

2.2. Objective and methodology

Our first objective is to estimate the probability that the backlog, $\max\{N_t - m, 0\}$, exceeds a given level $K \in \mathbb{N}$ during a *busy cycle*, which we define as an uninterrupted period during which the system has systematically been non-empty. Such a period is initiated by the arrival of a customer to an empty system, and ends by the departure of the last customer (leaving all servers idle). Notice that in our model busy cycles are *not* i.i.d., as the state of the background process at the beginning of the busy cycle has impact on its evolution (as opposed to for instance the situation with renewal-type arrivals that was studied in Ref.[20]). We denote by $\mathscr{F}_i$ the event that a busy cycle started when the background process was in state i. We focus on estimating the probability $\varrho_i(K)$ that in a busy cycle the number of customers in the queue exceeds the value K conditional on $\mathscr{F}_i$.

In practice, the probability $\varrho_i(K)$ is usually required to be small, which makes estimating it by crude Monte Carlo simulation inefficient Ref.[3, Ch. VI]. We are therefore interested in an estimation procedure that relies on importance sampling Ref.[3, Section V.1] in order to limit the required simulation effort. Importance sampling is based on imposing a "change of measure" with respect to the original measure $\mathbb{P}$. More concretely, the simulation is performed under a different probability measure $\mathbb{Q}$, and the simulation output is corrected by the "likelihood ratios" $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$ evaluated at the observed outcome in order to retain an unbiased estimation procedure. The challenge is to find an alternative measure $\mathbb{Q}$ that effectively reduces the variance of the estimator. This typically means that $\mathbb{Q}$ should be such that the event under consideration becomes more likely to occur, but in addition it is required that the likelihood ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$ on the event of interest should have a low variance. This is made explicit in Ref.[3, Section VI.1], where various efficiency measures for importance-sampling estimators are discussed.

Compared to the efficient simulation of tail probabilities in an M/M/m queue, a number of complications arises in our set-up. In the first place, as mentioned above, busy cycles are not independent. Furthermore, since servers are heterogeneous, one needs to keep track not only of the number of busy servers but also of their indices and current phases. In addition, the arrival rate is not fixed but depends on the state of the background processes.

Observe, however, that during periods in which N_t is larger than m, the dynamics of the process $(I_t, \boldsymbol{J}_t, N_t)_{t\geqslant 0}$ depend on I_t and $\boldsymbol{J}_t$ only; one could say it is "level-homogeneous." This motivates that we split each busy cycle into subintervals in which $N_t \in \{m+1, m+2, \ldots\}$ (i.e., the queue is not empty; we refer to these intervals as *fully* busy periods), and periods in which $N_t \in \{1, \ldots, m\}$ (i.e., the queue is empty; we call these intervals *partially* busy periods). Thus, during a busy cycle the system alternates between partially and fully busy periods until the system becomes idle again.

Based on the above, we can decompose $\varrho_i(K)$ as follows. With $\varrho_i(K, n)$ the probability that the number of customers attains $m + K + 1$ for the first time in the nth fully busy period (conditional on $\mathscr{F}_i$), we can write

$$\varrho_i(K) = \sum_{n=1}^{\infty} \varrho_i(K, n).$$

With this decomposition in mind, we first consider the following approach to estimate $\varrho_i(K)$, which will be detailed in Section 3. During the fully busy periods, in which the system is level-homogeneous, we use an alternative probability measure $\mathbb{Q}$ under which the queueing system is unstable (so that the rare event under study will occur frequently). During partially busy periods, in which the system is not level-homogeneous, we use the original measure $\mathbb{P}$. To establish particular efficiency properties, the number of fully busy periods (per busy cycle) in which $\mathbb{Q}$ is applied should be bounded by an arbitrary constant $C \in \mathbb{N}$; we return to this subtlety in Section 3. Based on the insights gained in Section 3, we then show in Section 4 how to obtain a change of measure that can be applied throughout the entire busy cycle.

3. Importance sampling procedure and its efficiency properties

In this section, we describe an importance sampling routine for estimating the quantity $\varrho_i(K)$. As this probability relates to the event that a given number of customers is reached before that number returns to 0, it suffices to track the evolution of the *embedded* discrete-time Markov process, i.e., of the continuous-time Markov chain $(I_t, \boldsymbol{J}_t, N_t)_{t\geqslant 0}$ observed at its transition epochs. With a mild abuse of notation, we refer to the embedded process as

$$(I_n, \boldsymbol{J}_n, N_n)_{n\in\mathbb{N}} \in \mathscr{I} \times \mathscr{D} \times \mathbb{N}, \qquad (2)$$

where n enumerates the epochs at which any of the three processes makes a transition. Note that at each transition epoch n of this embedded process typically only one of the state vector components changes, the exception being the occurrence of a departure (in which case $\boldsymbol{J}_n$ may change, and N_n decreases by one).

Define σ_k to be the first time that $(N_n)_{n\in\mathbb{N}}$ reaches level k. Assuming that a busy cycle starts with the arrival of a first customer, the backlog exceeds K within that cycle if and only if $\overline{\sigma}_K := \sigma_{m+K+1} < \sigma_0$. The objective of this section is to find an efficient algorithm for estimating the probability $\varrho_i(K)$ that in a busy cycle the number of customers in the queue exceeds the value K given that the background process is in $i \in \mathscr{I}$ at the start of the busy cycle; that is,

$$\varrho_i(K) = \mathbb{P}(\overline{\sigma}_K < \sigma_0 \mid \mathscr{F}_i)$$

where $\mathscr{F}_i$ corresponds to the event that $I_0 = i$, $J_0^{(1)}$ is sampled according to $\boldsymbol{\alpha}^{(1)}$, $J_0^{(2)} = \cdots = J_0^{(m)} = \dagger$, and $N_0 = 1$ (recall from Section 2.1 that the first customer is attended to by the server with the lowest index).

The remainder of this section is organized as follows.

- First, in Section 3.1, we focus on a fully busy period; we conveniently shift time, such that the start of the busy period corresponds to time 0. We fix a state $(i, \boldsymbol{j}) \in \mathscr{I} \times \mathscr{D}$, and consider the probability

$$\begin{aligned} r_{i,\boldsymbol{j}}(K) &:= \mathbb{P}_{i,\boldsymbol{j}}(\overline{\sigma}_K < \sigma_m \,\big|\, N_0 = m+1) \\ &:= \mathbb{P}(\overline{\sigma}_K < \sigma_m \,\big|\, I_0 = i, \boldsymbol{J}_0 = \boldsymbol{j}, N_0 = m+1). \end{aligned} \tag{3}$$

  Observe that $r_{i,\boldsymbol{j}}(K)$ can be interpreted as the probability that the backlog exceeds K within a fully busy period given that such a period has started when the background process and the service times were in state $(i, \boldsymbol{j})$. Relying on the fact that during the fully busy period the system is level-homogeneous, we define a change of measure for estimating $r_{i,\boldsymbol{j}}(K)$. We then propose an importance sampling algorithm for estimating $\varrho_i(K)$ which applies this change of measure during the first $C \in \mathbb{N}$ fully busy periods. (In Section 4, we will see how the rates can be twisted in general, without the restriction of changing the measure only during the fully busy periods.)
- In Section 3.2, we investigate efficiency properties of the proposed estimators. In the first place, we show that the importance sampling procedure for estimating $r_{i,\boldsymbol{j}}(K)$ has bounded relative error. In addition, we show that $\varrho_i(K)$ can be bounded in terms of $r_{i,\boldsymbol{j}}(K)$, so that the procedure for estimating $\varrho_i(K)$ has bounded relative error as well.
- Section 3.3 presents a numerical example, in which the new measure $\mathbb{Q}$ is computed. It gives rise to a decomposition property, formalized in Section 3.4, which drastically reduces the computational efforts required to evaluate the measure $\mathbb{Q}$.

3.1. Change of measure

In this subsection, we focus on the estimation of $r_{i,\boldsymbol{j}}(K)$ as defined in (3), with fixed i and $\boldsymbol{j}$. We first sketch the reasoning we followed to come up with a guess for a promising change of measure. Later in Section 3.2, we prove that this change of measure indeed has the desired properties.

Observe that in order to decide whether or not $\overline{\sigma}_K < \sigma_m$, we consider a time interval during which the value of N_n has not dropped below $m+1$. In other words, the transition matrix of $(I_n, \boldsymbol{J}_n, N_n)$ does not depend on N_n during that interval. It is essentially this property that enables the following construction of the alternative measure, which mimics the construction in Ref.[19] for the easier case of the Markov fluid queue.

The discrete-time Markov chain $(\overline{I}_n, \overline{\boldsymbol{J}}_n, \overline{N}_n) \in \mathscr{I} \times \mathscr{D} \times \mathbb{Z}$ that behaves as $(I_n, \boldsymbol{J}_n, N_n)$ during the fully busy period is characterized by the following transition probabilities. Define

$$\varphi_{i,\boldsymbol{j}} := \lambda_i + \sum_{\ell=1}^{m} t_{j_\ell}^{(\ell)} + q_i.$$

Let $\boldsymbol{e}_\ell$ be a vector of dimension m with a one on position ℓ and zeros otherwise. Then the probability of moving from $(i, \boldsymbol{j}, n)$ to $(i, \boldsymbol{j}, n+1)$ is $\lambda_i/\varphi_{i,\boldsymbol{j}}$ (this corresponds to an arrival); the probability of moving from $(i, \boldsymbol{j}, n)$ to $(i', \boldsymbol{j}, n)$ is $q_{ii'}/\varphi_{i,\boldsymbol{j}}$ (this corresponds to a transition of the background process); the probability of moving from $(i, \boldsymbol{j}, n)$ to $(i, \boldsymbol{j} + (k - j_\ell)\boldsymbol{e}_\ell, n)$ is $t_{j_\ell,k}^{(\ell)}/\varphi_{i,\boldsymbol{j}}$ (this corresponds to a transition in the phase of one of the service times, without a departure); and the probability of moving from $(i, \boldsymbol{j}, n)$ to $(i, \boldsymbol{j} + (k - j_\ell)\boldsymbol{e}_\ell, n-1)$ is $\overline{t}_{j_\ell,k}^{(\ell)}/\varphi_{i,\boldsymbol{j}}$ with $\overline{t}_{j_\ell,k}^{(\ell)} := t_{j_\ell,\dagger}^{(\ell)}\alpha_k^{(\ell)}$ (this corresponds to a transition in the phase of one of the service times, but now with a departure).

We now point out how the alternative measure $\mathbb{Q}$ is constructed. Let, for $(i, \boldsymbol{j}) \neq (i^\star, \boldsymbol{j}^\star)$, $\xi_{i,\boldsymbol{j}}$ denote the net increase of the number of customers $\overline{N}_n$ from an epoch that $(\overline{I}_n, \overline{\boldsymbol{J}}_n)$ is in $(i, \boldsymbol{j})$ until $(\overline{I}_n, \overline{\boldsymbol{J}}_n)$ arrives at a given reference state $(i^\star, \boldsymbol{j}^\star)$, whereas $\xi_{i^\star,\boldsymbol{j}^\star}$ is the net increase of $\overline{N}_n$ between two subsequent visits of $(\overline{I}_n, \overline{\boldsymbol{J}}_n)$ to $(i^\star, \boldsymbol{j}^\star)$; we show below that the choice of the reference state does not affect the resulting new measure $\mathbb{Q}$. Let $x_{i,\boldsymbol{j}}(\vartheta) := \mathbb{E}e^{\vartheta\xi_{i,\boldsymbol{j}}}$ denote the moment generating function (MGF) of $\xi_{i,\boldsymbol{j}}$.

Relying on the usual "Markovian reasoning," the MGFs satisfy, for any $(i, \boldsymbol{j}) \neq (i^\star, \boldsymbol{j}^\star)$,

$$x_{i,\boldsymbol{j}}(\vartheta) = \frac{\lambda_i}{\varphi_{i,\boldsymbol{j}}} x_{i,\boldsymbol{j}}(\vartheta)\, \mathrm{e}^{\vartheta} + \sum_{i'=1, i' \neq i}^{d} \frac{q_{ii'}}{\varphi_{i,\boldsymbol{j}}} x_{i',\boldsymbol{j}}(\vartheta) + \sum_{\ell=1}^{m} \sum_{k=1, k \neq j_\ell}^{D^{(\ell)}} \frac{t_{j_\ell,k}^{(\ell)}}{\varphi_{i,\boldsymbol{j}}} x_{i,\boldsymbol{j}+(k-j_\ell)\boldsymbol{e}_\ell}(\vartheta)$$
$$+ \sum_{\ell=1}^{m} \sum_{k=1}^{D^{(\ell)}} \frac{\overline{t}_{j_\ell,k}^{(\ell)}}{\varphi_{i,\boldsymbol{j}}} x_{i,\boldsymbol{j}+(k-j_\ell)\boldsymbol{e}_\ell}(\vartheta)\, \mathrm{e}^{-\vartheta}, \qquad (4)$$

wherever $x_{i^\star,\boldsymbol{j}^\star}(\vartheta)$ appears in the right-hand side of (4), it equals 1 (as the process $(\bar{I}_n, \bar{\boldsymbol{J}}_n)$ has arrived in the reference state). The system of Equation (4) can be interpreted as follows: the first term on the right-hand side corresponds to an arrival (hence, the factor e^{ϑ}), the second to a jump of the background process, the third to a phase-transition of one of the service times (but not a departure), and the fourth to a departure and simultaneously the start of a new service (hence, the factor $\mathrm{e}^{-\vartheta}$). For $(i, \boldsymbol{j}) = (i^\star, \boldsymbol{j}^\star)$ we obtain, by the same token,

$$\begin{aligned} x_{i^\star,\boldsymbol{j}^\star}(\vartheta) &= \frac{\lambda_{i^\star}}{\varphi_{i^\star,\boldsymbol{j}^\star}}\,\mathrm{e}^{\vartheta}\,x_{i^\star,\boldsymbol{j}^\star}(\vartheta) + \sum_{i'=1,i'\neq i^\star}^{d} \frac{q_{i^\star i'}}{\varphi_{i^\star,\boldsymbol{j}^\star}} x_{i',\boldsymbol{j}^\star}(\vartheta) \\ &\quad + \sum_{\ell=1}^{m} \sum_{k=1,k\neq j^\star_\ell}^{D^{(\ell)}} \frac{t^{(\ell)}_{j^\star_\ell,k}}{\varphi_{i^\star,\boldsymbol{j}^\star}} x_{i^\star,\boldsymbol{j}^\star+(k-j^\star_\ell)\boldsymbol{e}_\ell}(\vartheta) \\ &\quad + \sum_{\ell=1}^{m} \sum_{k=1}^{D^{(\ell)}} \frac{\bar{t}^{(\ell)}_{j^\star_\ell,k}}{\varphi_{i^\star,\boldsymbol{j}^\star}} x_{i^\star,\boldsymbol{j}^\star+(k-j^\star_\ell)\boldsymbol{e}_\ell}(\vartheta)\,\mathrm{e}^{-\vartheta}, \end{aligned} \tag{5}$$

For any value of ϑ, the MGFS $x_{i^\star,\boldsymbol{j}^\star}(\vartheta)$ now follow from solving the above system of linear equations.

Recall that $\xi_{i^\star,\boldsymbol{j}^\star}$ is the net increase of the number of customers between two subsequent visits of $(\bar{I}_n, \bar{\boldsymbol{J}}_n)$ to the reference state $(i^\star, \boldsymbol{j}^\star)$. From Ref.[3, Section VI.2a], we know that the alternative measure obtained by an exponential twist of the original measure $\mathbb{P}$ should be such that the MGF of $\xi_{i^\star,\boldsymbol{j}^\star}$ evaluated in ϑ under $\mathbb{Q}$ matches the MGF of $\xi_{i^\star,\boldsymbol{j}^\star}$ evaluated in $\vartheta + \vartheta^\star$ under $\mathbb{P}$. In self-evident notation

$$\mathbb{E}^{\mathbb{Q}}\,\mathrm{e}^{\vartheta\xi_{i^\star,\boldsymbol{j}^\star}} = \mathbb{E}\,\mathrm{e}^{(\vartheta+\vartheta^\star)\xi_{i^\star,\boldsymbol{j}^\star}},$$

with the value of $\vartheta^\star$ such that $x_{i^\star,\boldsymbol{j}^\star}(\vartheta^\star) \equiv \mathbb{E}\,\mathrm{e}^{\vartheta^\star\xi_{i^\star,\boldsymbol{j}^\star}} = 1$; $\vartheta^\star$ can thus be interpreted as the *Cramér root*[3, Section VI.2a] associated with the random variable $\xi_{i^\star,\boldsymbol{j}^\star}$. In the set of equations (4)–(5) at all places where $x_{i^\star,\boldsymbol{j}^\star}(\vartheta^\star)$ appears it has been replaced by 1, and as a consequence it can be written in terms of an eigensystem $A(\vartheta^\star)\boldsymbol{x}(\vartheta^\star) = \boldsymbol{x}(\vartheta^\star)$ with $x_{i^\star,\boldsymbol{j}^\star}(\vartheta^\star) = 1$; hence $A(\vartheta^\star)$ has eigenvalue/eigenvector pair $(1, \boldsymbol{x}(\vartheta^\star))$. For the sake of readability in the sequel we write $A := A(\vartheta^\star)$, $\boldsymbol{x} := \boldsymbol{x}(\vartheta^\star)$ and $x_{i,\boldsymbol{j}} := x_{i,\boldsymbol{j}}(\vartheta^\star)$.

Two remarks are in place now. In the first place, the representation $A(\vartheta^\star)\boldsymbol{x}(\vartheta^\star) = \boldsymbol{x}(\vartheta^\star)$ reveals that the choice of the reference state has no impact: picking another reference state leads to (i) the same $\vartheta^\star$, and (ii) (up to a multiplicative constant) the same vector $\boldsymbol{x}(\vartheta^\star)$ (recall that the component corresponding to the reference state equals 1). Second, as the components of the vector $\boldsymbol{x}(\vartheta^\star)$ have the interpretation of MGFS, the vector is necessarily componentwise positive.

In practical terms, $\vartheta^\star$ and the corresponding eigenvector can be found as follows. We have already pointed out that for given ϑ evaluating $u(\vartheta) := \mathbb{E}\,\mathrm{e}^{\vartheta\xi_{i^\star,\boldsymbol{j}^\star}}$ boils down to solving a system of linear equations. Observe that $u(0) = 1$, $u'(0) < 0$ (due to the stability condition), $u(\cdot)$ is convex and increasing to ∞. As a consequence, $\vartheta^\star > 0$

(solving $u(\vartheta^\star) = 1$) exists and is unique (and can be numerically found by e.g. an elementary bisection procedure).

Inspired by the above eigensystem, we now propose the following new measure $\mathbb{Q}$ corresponding to an exponential twist of the distribution of ξ, to be used to estimate $r_{i,\boldsymbol{j}}(K)$: when $(\bar{I}_n, \bar{\boldsymbol{J}}_n) = (i, \boldsymbol{j})$,

$$\lambda_i^\circ = \lambda_i \,\mathrm{e}^{\vartheta^\star}, \quad q_{ii'}^\circ = q_{ii'} \frac{x_{i',\boldsymbol{j}}}{x_{i,\boldsymbol{j}}}, \quad \left(t_{j_\ell,k}^{(\ell)}\right)^\circ = t_{j_\ell,k}^{(\ell)} \frac{x_{i,\boldsymbol{j}+(k-j_\ell)\boldsymbol{e}_\ell}}{x_{i,\boldsymbol{j}}},$$
$$\left(\bar{t}_{j_\ell,k}^{(\ell)}\right)^\circ = \bar{t}_{j_\ell,k}^{(\ell)} \frac{x_{i,\boldsymbol{j}+(k-j_\ell)\boldsymbol{e}_\ell}}{x_{i,\boldsymbol{j}}} \,\mathrm{e}^{-\vartheta^\star}. \tag{6}$$

In the remainder of this subsection, we evaluate the likelihood ratio that results from this change of measure when estimating $r_{i,\boldsymbol{j}}(K)$; as it turns out, this has a surprisingly simple form. To this end, we consider an arbitrary path of the process $(\bar{I}_n, \bar{\boldsymbol{J}}_n)$ starting when the fully busy period commences (that is, we have $\bar{I}_0 = i$, $\bar{\boldsymbol{J}}_0 = \boldsymbol{j}$, and $\bar{N}_0 = m + 1$), and ending at time $\tau = \min\{\bar{\sigma}_K, \sigma_m\}$, visiting states $(i_n, \boldsymbol{j}_n)$, where n denotes the nth transition epoch of the process (2). Let $\mathscr{N}_+$ denote the $n \in \mathscr{S} := \{1, \dots, \tau\}$ corresponding to arrivals, $\mathscr{N}_\rightsquigarrow$ the $n \in \mathscr{S}$ corresponding to transitions of the background process, $\mathscr{N}_\circlearrowleft^{(\ell)}$ the $n \in \mathscr{S}$ corresponding to a phase-transition of the service time at server ℓ (not being a service completion), and $\mathscr{N}_-^{(\ell)}$ corresponding to a service completion at server ℓ. The likelihood (under $\mathbb{P}$) of such a path is thus given by, with $i = i_0$ and $\boldsymbol{j} = \boldsymbol{j}_0$,

$$\prod_{n\in\mathscr{N}_+} \frac{\lambda_{i_n}}{\varphi_{i_n,\boldsymbol{j}_n}} \prod_{n\in\mathscr{N}_\rightsquigarrow} \frac{q_{i_n,i_{n+1}}}{\varphi_{i_n,\boldsymbol{j}_n}} \prod_{\ell=1}^m \prod_{n\in\mathscr{N}_\circlearrowleft^{(\ell)}} \frac{t_{j_n,j_{n+1}}^{(\ell)}}{\varphi_{i_n,\boldsymbol{j}_n}} \prod_{\ell=1}^m \prod_{n\in\mathscr{N}_-^{(\ell)}} \frac{\bar{t}_{j_n,j_{n+1}}^{(\ell)}}{\varphi_{i_n,\boldsymbol{j}_n}}. \tag{7}$$

The likelihood of the same path under the new measure $\mathbb{Q}$ has the same form, except that all probabilities in (7) are replaced by their counterparts under $\mathbb{Q}$, where, due to (4) and the definition of the new rates,

$$\begin{aligned}\varphi_{i,\boldsymbol{j}}^\circ &:= \lambda_i^\circ + \sum_{\ell=1}^m \left(t_{j_\ell}^{(\ell)}\right)^\circ + q_i^\circ \\ &= \lambda_i^\circ + \sum_{\ell=1}^m \sum_{k=1,k\neq j_\ell}^{D^{(\ell)}+1} \left(t_{j_\ell,k}^{(\ell)}\right)^\circ + \sum_{i'=1,i'\neq i}^d q_{ii'}^\circ = \varphi_{i,\boldsymbol{j}}.\end{aligned}$$

It follows that the likelihood ratio over the path takes the form

$$\begin{aligned}L = &\prod_{n\in\mathscr{N}_+} \frac{\lambda_{i_n}/\varphi_{i_n,\boldsymbol{j}_n}}{\lambda_{i_n}^\circ/\varphi_{i_n,\boldsymbol{j}_n}^\circ} \prod_{n\in\mathscr{N}_\rightsquigarrow} \frac{q_{i_n,i_{n+1}}/\varphi_{i_n,\boldsymbol{j}_n}}{q_{i_n,i_{n+1}}^\circ/\varphi_{i_n,\boldsymbol{j}_n}^\circ} \\ &\times \prod_{\ell=1}^m \prod_{n\in\mathscr{N}_\circlearrowleft^{(\ell)}} \frac{t_{j_n,j_{n+1}}^{(\ell)}/\varphi_{i_n,\boldsymbol{j}_n}}{\left(t_{j_n,j_{n+1}}^{(\ell)}\right)^\circ/\varphi_{i_n,\boldsymbol{j}_n}^\circ} \prod_{\ell=1}^m \prod_{n\in\mathscr{N}_-^{(\ell)}} \frac{\bar{t}_{j_n,j_{n+1}}^{(\ell)}/\varphi_{i_n,\boldsymbol{j}_n}}{\left(\bar{t}_{j_n,j_{n+1}}^{(\ell)}\right)^\circ/\varphi_{i_n,\boldsymbol{j}_n}^\circ}.\end{aligned}$$

Because $\varphi_{i,j} = \varphi^{\circ}_{i,j}$, by (6) this reduces to the "telescopic product"

$$L = \prod_{n \in \mathcal{N}_+} \mathrm{e}^{-\vartheta^\star} \prod_{n \in \mathcal{N}_{\rightsquigarrow}} \frac{x_{i_n,j_n}}{x_{i_{n+1},j_{n+1}}} \prod_{\ell=1}^{m} \prod_{n \in \mathcal{N}^{(\ell)}_{\circlearrowleft}} \frac{x_{i_n,j_n}}{x_{i_{n+1},j_{n+1}}} \prod_{\ell=1}^{m} \prod_{n \in \mathcal{N}^{(\ell)}_{-}} \frac{x_{i_n,j_n}}{x_{i_{n+1},j_{n+1}}} \mathrm{e}^{\vartheta^\star}$$

$$= \mathrm{e}^{-\vartheta^\star \Sigma_+} \frac{x_{i_0,j_0}}{x_{i_\tau,j_\tau}} \mathrm{e}^{\vartheta^\star \Sigma_-}, \tag{8}$$

where Σ_+ is the number of arrivals in $\mathscr{S}$, and Σ_- the number of departures. Observe that at the end of each fully busy period we either have $\Sigma_- - \Sigma_+ = -K$ if $\tau = \overline{\sigma}_K$, or $\Sigma_- - \Sigma_+ = 1$ if $\tau = \sigma_m$. We find the following identities.

Corollary 3.1. *Let $\bar{I}_0 = i$, $\bar{\boldsymbol{J}}_0 = \boldsymbol{j}$, and $\overline{N}_0 = m + 1$. For any $K \in \mathbb{N}$,*

$$L\,\mathbb{1}\{\tau = \overline{\sigma}_K\} = \mathrm{e}^{-\vartheta^\star K} \frac{x_{i,j}}{x_{i_\tau,j_\tau}}, \quad L\,\mathbb{1}\{\tau = \sigma_m\} = \mathrm{e}^{\vartheta^\star} \frac{x_{i,j}}{x_{i_\tau,j_\tau}}.$$

Remark 3.1. It is reassuring to note that the proposed change of measure satisfies Juneja's "equi-probable cycle" condition, which should hold for an asymptotically optimal change of measure[12]. Namely, if $(\bar{I}_n, \bar{\boldsymbol{J}}_n, \overline{N}_n)_{n\in\mathbb{N}}$ visits a specific state multiple times, the contribution to the likelihood ratio of the interval between two such subsequent visits is equal to one.

Based on the analysis in this section, we propose to change to the alternative measure $\mathbb{Q}$ during the fully busy periods, whereas during the partially busy periods $\mathbb{P}$ should be used; implementation details are provided in the appendix. If $\mathbb{Q}$ is applied only during the first C fully busy periods, we prove in Section 3.2 that the procedure has bounded relative error. In practice, however, C can be chosen arbitrarily large without compromising the estimator's performance; see Section 5 for a numerical assessment.

3.2. Bounds and relative error

In this subsection, we derive bounds on $\varrho_i(K)$ in terms of the probability $r_{i,j}(K)$, which we then use to prove that the proposed importance sampling estimation procedure leads to bounded relative error.

The partially busy period before each fully busy period commences either when 1 customer is in the system (at the beginning of the busy cycle), or when m customers are in the system (when we just exited a fully busy period). Accordingly, we define, in self-evident notation,

$$\overline{p}_i := \mathbb{P}\left(\sigma_{m+1} < \sigma_0 \mid \mathscr{F}_i\right), \quad p_{i,j} := \mathbb{P}_{i,j}(\sigma_{m+1} < \sigma_0 \mid N_0 = m).$$

Observe that the number of fully busy periods in a busy cycle is bounded from above by a geometric random variable G with success probability

$$p_+ := \max\left\{\max_i \overline{p}_i,\ \max_{i,j} p_{i,j}\right\} = \max_{i,j} p_{i,j} < 1. \tag{9}$$

The equation holds because the fully busy period is initiated with the $(m+1)$-th customer; evidently, the probability of this event is larger given $N_0 = m$ than given $N_0 = 1$.

In every one of these fully busy periods, level $m + K + 1$ is reached with a probability that is bounded above by

$$r_+(K) := \max_{i,\boldsymbol{j}} r_{i,\boldsymbol{j}}(K).$$

Supposing that $G = k$, in each of the k attempts the level $m + K + 1$ can be reached. The union bound then yields the following upper bound: uniformly in $i \in \mathscr{I}$,

$$\varrho_i(K) \leqslant \sum_{k=1}^{\infty} p_+^k (1 - p_+) k \, r_+(K) = \frac{r_+(K)}{p_+}, \tag{10}$$

where the equation follows by evaluating the expectation of a geometric random variable. Now focus on establishing a lower bound based on the probability of reaching $m + K + 1$ in the first fully busy period. To this end, we define

$$p_- := \min_i \mathbb{P}(\sigma_{m+1} < \sigma_0 \mid \mathscr{F}_i), \quad r_-(K) := \min_{i,\boldsymbol{j}} r_{i,\boldsymbol{j}}(K).$$

Then, it follows directly that, uniformly in $i \in \mathscr{I}$,

$$\varrho_i(K) \geqslant p_- r_-(K). \tag{11}$$

In order to make the bounds on $\varrho_i(K)$ more explicit, we now show how $r_{i,\boldsymbol{j}}(K)$ as defined in (3) can be bounded. These bounds are derived by using the change of measure $\mathbb{Q}$ that we identified above. Denoting, as before, the likelihood ratio in the fully busy period by L, we have

$$\begin{aligned} r_{i,\boldsymbol{j}}(K) &= \mathbb{E}^{\mathbb{Q}}_{i,\boldsymbol{j}} [L \, \mathbb{1}\{\overline{\sigma}_K < \sigma_m\} \mid N_0 = m+1] \\ &= \mathbb{E}^{\mathbb{Q}}_{i,\boldsymbol{j}} [L \mid \overline{\sigma}_K < \sigma_m, N_0 = m+1] \, \mathbb{Q}_{i,\boldsymbol{j}} \left(\overline{\sigma}_K < \sigma_m \mid N_0 = m+1\right), \end{aligned}$$

where the subscript $i, \boldsymbol{j}$ denotes conditioning on the initial states $I_0 = i$ and $\boldsymbol{J}_0 = \boldsymbol{j}$ as before. Using Equation (8) (or Corollary 3.1), we thus conclude

$$\eta_- \mathrm{e}^{-\vartheta^\star K} \leqslant \mathbb{E}^{\mathbb{Q}}_{i,\boldsymbol{j}} [L \mid \overline{\sigma}_K < \sigma_m, N_0 = m+1] \leqslant \eta_+ \mathrm{e}^{-\vartheta^\star K}, \tag{12}$$

with the constants η_- and η_+ defined by

$$\eta_- := \min_{i,i' \in \mathscr{I}, \boldsymbol{j}, \boldsymbol{j}' \in \mathscr{D}} \frac{x_{i,\boldsymbol{j}}}{x_{i',\boldsymbol{j}'}}, \quad \eta_+ := \max_{i,i' \in \mathscr{I}, \boldsymbol{j}, \boldsymbol{j}' \in \mathscr{D}} \frac{x_{i,\boldsymbol{j}}}{x_{i',\boldsymbol{j}'}}.$$

Since $\mathscr{I}$ and $\mathscr{D}$ are finite, and recalling that $\boldsymbol{x}$ is componentwise positive, η_- and η_+ are positive and finite.

Due to the fact that under $\mathbb{Q}$ the queueing system is unstable[3, Section VI.2], we have that, as $K \to \infty$,

$$\mathbb{Q}_{i,\boldsymbol{j}} \left(\overline{\sigma}_K < \sigma_m \middle| N_0 = m+1\right) \downarrow \mathbb{Q}_{i,\boldsymbol{j}} \left(\overline{\sigma}_\infty < \sigma_m \middle| N_0 = m+1\right) > 0.$$

Furthermore, note that (12) holds for any fixed $(i, \boldsymbol{j})$; thus, in particular, we can take the minimum or the maximum over such states. We have thus shown that there exist positive and finite numbers κ_- and κ_+, such that

$$\kappa_- \mathrm{e}^{-\vartheta^\star K} \leqslant r_-(K) \leqslant r_+(K) \leqslant \kappa_+ \mathrm{e}^{-\vartheta^\star K}.$$

Combining this with (10) and (11), we have established the following result.

Proposition 3.1. For any $K \in \mathbb{N}$, $i \in \mathscr{I}$, and $\boldsymbol{j} \in \mathscr{D}$, uniformly in $i \in \mathscr{I}$,

$$p_-\kappa_- \mathrm{e}^{-\vartheta^\star K} \leqslant \varrho_i(K) \leqslant \frac{\kappa_+}{p_+} \mathrm{e}^{-\vartheta^\star K}.$$

Proposition 3.1 provides us with a lower and upper bound on $\varrho_i(K)$, which are valid across all $K \in \mathbb{N}$, and which differ just by a multiplicative constant. We now use these bounds to assess the estimator's efficiency properties. We now use the above proposition to assess the estimator's efficiency properties. Since $\vartheta^\star$ is the Cramér root, it is positive, and hence both bounds tend to zero as K grows large. Below we establish an upper bound on the relative error that is independent of K.

The probability $\varrho_i(K)$ is estimated by using $\mathbb{Q}$ during the fully busy periods, and the original measure $\mathbb{P}$ otherwise. Denoting this "composite measure" by $\overline{\mathbb{Q}}$, we rely on the identity

$$\varrho_i(K) = \mathbb{E}^{\overline{\mathbb{Q}}}[L\mathbb{1}\{\overline{\sigma}_K < \sigma_0\} \mid \mathscr{F}_i],$$

with the event $\mathscr{F}_i$ as defined above; cf. Ref.[3, Section V.1a]. The *relative error* of an estimator is defined by the ratio of the estimator's standard deviation to its mean. Noting that the estimator is unbiased, we obtain that its per-sample standard deviation can be written as

$$\nu^\circ = \sqrt{\mathbb{E}^{\overline{\mathbb{Q}}}\left[L^2\mathbb{1}\left\{\overline{\sigma}_K < \sigma_0\right\} \mid \mathscr{F}_i\right] - \varrho_i(K)^2}.$$

Thus, the relative error based on n simulation runs is

$$\frac{\nu^\circ}{\sqrt{n}\,\varrho_i(K)} = \sqrt{\frac{\mathbb{E}^{\overline{\mathbb{Q}}}\left[L^2\mathbb{1}\left\{\overline{\sigma}_K < \sigma_0\right\} \mid \mathscr{F}_i\right]}{n\varrho_i(K)^2} - \frac{1}{n}} \leqslant \sqrt{\frac{\mathbb{E}^{\overline{\mathbb{Q}}}\left[L^2\mathbb{1}\left\{\overline{\sigma}_K < \sigma_0\right\} \mid \mathscr{F}_i\right]}{n\varrho_i(K)^2}}. \tag{13}$$

Observe that the number of fully busy periods that are performed under $\overline{\mathbb{Q}}$ is bounded from above by a geometric random variable with success probability p_+. Invoking Corollary 1, we thus obtain, for any $i \in \mathscr{I}$,

$$\mathbb{E}^{\overline{\mathbb{Q}}}\left[L^2\mathbb{1}\left\{\overline{\sigma}_K < \sigma_0\right\} \mid \mathscr{F}_i\right] \leqslant \mathrm{e}^{-2\vartheta^\star K} \sum_{k=1}^{\infty} \mathrm{e}^{2\vartheta^\star k} p_+^k (1-p_+)\eta_+^{2k}. \tag{14}$$

If $\mathrm{e}^{2\vartheta^\star} p_+ \eta_+^2 < 1$, then the geometric series converges. Combining this with the lower bound $\varrho_i(K) \geqslant p_-\kappa_-\mathrm{e}^{-\vartheta^\star K}$, the bounded relative error follows: the expression in the left-hand side of (13) is bounded above by a finite expression that does not depend on K.

Proposition 3.2. *If $\eta^\circ := \mathrm{e}^{2\vartheta^\star} p_+ \eta_+^2 < 1$, then the estimator for $\varrho_i(K)$ based on $\overline{\mathbb{Q}}$ has bounded relative error:*

$$\frac{\nu^\circ}{\sqrt{n}\,\varrho_i(K)} \leqslant \frac{1}{p_-\kappa_-}\sqrt{\frac{(1-p_+)\eta^\circ}{n(1-\eta^\circ)}}$$

The question arises what can be done if the condition $\eta^\circ < 1$ is not fulfilled (in which case the geometric series in (14) does not converge). Various approaches are possible, of which we present two possibilities.

Approach 1. This approach is based on a truncation. We replace the probability $\varrho_i(K)$ by a corresponding probability $\varrho_{i,C}(K)$: this $\varrho_{i,C}(K)$ is defined precisely as $\varrho_i(K)$ but with the additional requirement that the number of customers exceeds K *in one of the first C fully busy periods*. It is clear that by choosing C appropriately large, the bias can be made arbitrarily small. Indeed, by observing that $\varrho_i(K)$ and $\varrho_{i,C}(K)$ only differ in scenarios in which there are more than C fully busy periods in the busy cycle, and that after these C fully busy periods the number of customers should exceed K before the busy cycle ends, the absolute bias can be bounded as follows:

$$\varrho_i(K) - \varrho_{i,C}(K) \leqslant (p_+)^C \cdot \frac{\kappa_+}{p_+}\mathrm{e}^{-\vartheta^\star K};$$

recognize the factor $(\kappa_+/p_+)\,\mathrm{e}^{-\vartheta^\star K}$ from Proposition 3.1. Suppose now that a relative bias $\delta > 0$ is allowed; then, again by Proposition 3.1 it suffices to choose C such that

$$\frac{(p_+)^C \cdot (\kappa_+/p_+)\,\mathrm{e}^{-\vartheta^\star K}}{p_-\kappa_-\mathrm{e}^{-\vartheta^\star K}} \leqslant \delta,$$

or, equivalently,

$$C \geqslant \frac{\log(\delta\, p_- p_+ \kappa_+/\kappa_-)}{\log p_+}.$$

Obviously, the lower δ, the larger the number of fully busy periods C.

Let $\mathbb{Q}_C$ be the change of measure in which $\varrho_{i,C}(K)$ is estimated using $\mathbb{Q}$ during the fully busy periods (where it is recalled that there are at most C of them), and the original measure $\mathbb{P}$ otherwise. We thus obtain that

$$\mathbb{E}^{\mathbb{Q}_C}\left[L^2 \mathbb{1}\left\{\overline{\sigma}_K < \sigma_0\right\} \mid \mathscr{F}_i\right] \leqslant \mathrm{e}^{-2\vartheta^\star K} \sum_{k=1}^{C} \mathrm{e}^{2\vartheta^\star k} p_+^k (1-p_+)\eta_+^{2k}.$$

This leads to the following result, with ν_C° denoting the counterpart of ν° for the measure $\mathbb{Q}_C$, which holds irrespective of the value of η°.

Theorem 3.1. *The estimator for $\varrho_{i,C}(K)$ based on $\mathbb{Q}_C$ has bounded relative error:*

$$\frac{\nu_C^\circ}{\sqrt{n}\,\varrho_{i,C}(K)} \leqslant \frac{1}{p_-\kappa_-}\sqrt{\frac{1}{n}\sum\nolimits_{k=1}^{C} \mathrm{e}^{2\vartheta^\star k} p_+^k (1-p_+)\eta_+^{2k}}$$

Approach 2. A pragmatic alternative is the following. We switch on the importance sampling (i.e., use measure $\mathbb{Q}$) only during the first C fully busy periods, and use the original measure $\mathbb{P}$ otherwise. Evidently, in this approach no bias has to be taken care of. For $C = \infty$, it coincides with the measure $\mathbb{Q}$. Pseudocode for this approach is presented in the appendix (but it can easily be adapted to Approach 1). In Section 5, we empirically study the impact of the choice of C.

3.3. Numerical example

In this subsection, we present a small numerical illustration. Consider a two-server system, with $d = 2$ and $D^{(1)} = D^{(2)} = 3$. The arrival rates, initial probabilities and transition rate matrices are

$$Q = \begin{pmatrix} -0.5 & 0.5 \\ 0.1 & -0.1 \end{pmatrix}, \quad \boldsymbol{\alpha}^{(1)} = \begin{pmatrix} 0.5 \\ 0.3 \\ 0.2 \\ 0 \end{pmatrix}, \quad T^{(1)} = \begin{pmatrix} -0.9 & 0.2 & 0.1 & 0.6 \\ 0.5 & -1.5 & 0.5 & 0.5 \\ 0.2 & 0.2 & -0.8 & 0.4 \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\boldsymbol{\lambda} = \begin{pmatrix} 0.1 \\ 0.5 \end{pmatrix}, \quad \boldsymbol{\alpha}^{(2)} = \begin{pmatrix} 0.5 \\ 0.2 \\ 0.3 \\ 0 \end{pmatrix}, \quad T^{(2)} = \begin{pmatrix} -1 & 0.2 & 0.2 & 0.6 \\ 1 & -2 & 0.5 & 0.5 \\ 0.2 & 0.2 & -1 & 0.6 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

We order the states lexicographically and solve the eigensystem (4) by first using bisection to find a value of ϑ such that the largest eigenvalue of the matrix A defining the eigensystem equals one. One thus obtains $\vartheta^\star = 0.86$. We normalize the eigenvector corresponding to this eigenvalue such that its first entry is one, and call the resulting vector $\boldsymbol{x}$.

For example, for states $i = 1, i' = 2, \boldsymbol{j} = (1, 1), \boldsymbol{j}' = (2, 3)$, we then have $x_{i,\boldsymbol{j}} = 1$, $x_{i,\boldsymbol{j}'} \approx 1.09$, $x_{i',\boldsymbol{j}} \approx 1.99$, $x_{i',\boldsymbol{j}'} \approx 2.17$ (rounded to two decimal digits). We observe the remarkable property that (up to the rounding error) $x_{i',\boldsymbol{j}'}/x_{i',\boldsymbol{j}} = x_{i,\boldsymbol{j}'}/x_{i,\boldsymbol{j}}$. Generally, it turns out that the obtained MGFs $\boldsymbol{x}$ are such that

$$\frac{x_{i,\boldsymbol{j}}}{x_{i,\boldsymbol{j}'}} = \frac{x_{i',\boldsymbol{j}}}{x_{i',\boldsymbol{j}'}} \quad \text{and} \quad \frac{x_{i,\boldsymbol{j}}}{x_{i',\boldsymbol{j}}} = \frac{x_{i,\boldsymbol{j}'}}{x_{i',\boldsymbol{j}'}}, \quad \text{for any} \quad i \neq i', \boldsymbol{j} \neq \boldsymbol{j}';$$

indicating that there is a certain decoupling among the servers as well as between servers and arrivals. The decoupling means that under $\mathbb{Q}$ (as was the case under the original measure $\mathbb{P}$), (i) the transition rates of the background process do not depend on the phases the customers in service are in, (ii) the service-time distribution at any particular server does not depend on the state of the background process nor the phases the other customers are in.

3.4. The structure of A

The observed decoupling can be formally established as follows. Note that if the states $(i, j_1, \ldots, j_m)$ are ordered lexicographically, then the matrix A defining the eigensystem given in (4) can be decomposed as

$$A \circ \boldsymbol{\varphi}\mathbf{1} = \Lambda\, \mathrm{e}^{\vartheta} \otimes I_D + \underline{Q} \otimes I_D + I_d \otimes R = \left[\Lambda\, \mathrm{e}^{\vartheta} + \underline{Q}\right] \oplus R, \tag{15}$$

where $\circ$ denotes the Hadamard product, $\boldsymbol{\varphi}$ is the column vector with entries $\varphi_{i,j}$, $\mathbf{1}$ is a row vector of ones, $\Lambda := \mathrm{diag}\{\boldsymbol{\lambda}\}$; $D := \prod_{\ell=1}^{m} D^{(\ell)}$; I_D is the $D \times D$-identity matrix, and I_d the $d \times d$-identity matrix; $\oplus$ and $\otimes$ denote the Kronecker sum and product, respectively; $\underline{Q}$ denotes the matrix $Q - Q \circ I_d$; and the "remainder term" R is of dimension $D \times D$ and depends on the matrices

$$T^{(\ell)} = \left(t^{(\ell)}_{i,j}\right)_{i,j} \quad \text{and} \quad \overline{T}^{(\ell)} = \left(\overline{t}^{(\ell)}_{i,j}\right)_{i,j}$$

but not on $\boldsymbol{\lambda}$ or Q. Because the eigenvalues of a Kronecker sum are given by the sums of the eigenvalues of each Kronecker summand[17, Thm. 13.16], this decomposition shows that the eigensystem can be split up into a part corresponding to arrivals and a part corresponding to services.

Let us now consider the remainder term R, which corresponds to the service processes. We note that R contains $D^{-} := \prod_{\ell=1}^{m-1} D^{(\ell)}$ block matrices of size $D^{(m)} \times D^{(m)}$, which have the following structure.

(i) The block matrices on the diagonal are of the form

$$\underline{T}^{(m)} + \overline{T}^{(m)}\, \mathrm{e}^{-\vartheta} + I_{D^{(m)}} \mathrm{e}^{-\vartheta} \sum_{\ell=1}^{m-1} \overline{t}^{(\ell)}_{j_\ell, j_\ell}.$$

where $\underline{T}^{(m)}$ denotes the matrix $T^{(m)} - T^{(m)} \circ I_{D^{(m)}}$.

(ii) Off-diagonal block matrices are either of the form $[t^{(\ell)}_{j_\ell, j'_\ell} + \overline{t}^{(\ell)}_{j_\ell, j'_\ell}\mathrm{e}^{-\vartheta}] I_{D^{(m)}}$ with $\ell < m$, or they are zero.

As it turns out, R can thus be decomposed as

$$R = \bigoplus_{\ell=1}^{m} \left(\underline{T}^{(\ell)} + \overline{T}^{(\ell)}\, \mathrm{e}^{-\vartheta}\right).$$

Inserting this expression in (15) we see (e.g., from Ref.[17, Thm. 13.16]) that the MGF $\boldsymbol{x}$ that we found as an eigenvector of A can be computed as the Kronecker product of eigenvectors of $m + 1$ decoupled eigensystems, corresponding with the arrivals and the services for each of the m servers. In this way, while the dimension of A is $d\,D$, the measure $\mathbb{Q}$ can be found by solving a system of dimension just $d + \sum_{\ell=1}^{m} D^{(\ell)}$; in the above example this would yield a reduction of dimension 18 to dimension 8. We detail such an alternative approach in the next section.

4. Efficient computation of change of measure

As mentioned above, an intrinsic problem of the change of measure defined in Section 3 is that the underlying eigensystem may become prohibitively large, and as a result the computation of $\mathbb{Q}$ becomes problematic. Already for the small example of Section 3.3, the length of the vector $\boldsymbol{x}$ is 18; if one has 10 servers with 3-dimensional phase-type distributions, and if $d = 4$, the dimension of the matrix A is as high as $4 \cdot 3^{10} = 2.36 \cdot 10^{5}$. This explains why we explore an alternative approach to compute $\mathbb{Q}$, in which the twist of the arrival processes and the service times do not interrelate. In the above example with 10 servers, this means that the alternative measure can be found by solving a system of dimension 34. The "catch" is that the decoupling-based approach requires function evaluations that are more involved, and thus for low-dimensional problems the approach of Section 3 may be preferred; see also Section 5, where we discuss simulation examples.

In the following, we explain how we can compute the (exponentially twisted) change of measure without having to solve a large eigensystem. This we do by relying on an auxiliary model that allows us to efficiently compute the alternative measure $\mathbb{Q}$, but which, importantly, does not correspond directly to the setup of Section 3. We will get back to this issue below. Our reasoning in this section relies on specific earlier results from large deviations theory; we therefore start our reasoning by recalling these.

4.1. Preliminaries

Consider a sequence of i.i.d. positive random variables $(R_n)_{n\in\mathbb{N}}$ (with a bounded MGF around 0), and its associated counting process

$$R(t) := \sup\left\{n \in \mathbb{N} : \sum_{i=1}^{n} R_i \leqslant t\right\}.$$

Define the associated limiting logarithmic MGF (ll-MGF):

$$\mathscr{R}(\vartheta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}e^{\vartheta R(t)}.$$

We now introduce our auxiliary setup that helps us identifying the measure $\mathbb{Q}$ for our model. Let c be some number larger than $\mathbb{E}R$ (with R being distributed as R_1). Consider a (stable) queue that drains at a constant rate c, where unit-sized jobs arrive with interarrival times $(R_n)_{n\in\mathbb{N}}$. From, e.g., Ref.[5], we have the logarithmic decay rate of the probability $\mathbb{P}(W > u)$ that the stationary workload W exceeds u,

$$\lim_{u\to\infty} \frac{1}{u} \log \mathbb{P}(W > u) = -\vartheta,$$

obeys $\mathscr{R}(\vartheta) - c\vartheta = 0$. The crucial insight, however, is that there is a second way to compute the same decay rate, viz. as the solution of $r(-c\vartheta)e^{\vartheta} = 1$, with $r(\vartheta) := \mathbb{E}e^{\vartheta R}$; see, e.g., Ref.[7]. This entails that both approaches lead to the same (c, ϑ)-pairs. A minor computation then yields that

$$\mathscr{R}(\vartheta) = -r^{-1}(e^{-\vartheta}). \tag{16}$$

We have thus expressed the ll-MGF $\mathscr{R}(\cdot)$ in terms of the MGF of R. For instance, for R having an exponential distribution with mean μ^{-1}, this yields $\mathscr{R}(\vartheta) = \mu(e^{\vartheta} - 1)$, as it should (recall that in this case $R(t)$ is Poisson with mean μt). See[6] for more background on this type of inversion result. In the following section, we point out how we can use this principle to identify $\mathbb{Q}$.

4.2. Twist of our model

After having introduced the auxiliary system in the previous subsection, we now return to our queueing model. In Section 3, we developed a way to determine the decay rate $-\vartheta^\star$ of the probability of our interest, leading to the alternative measure $\mathbb{Q}$; as mentioned, the intrinsic drawback of this approach is that it requires a possibly large eigensystem. In this section, we present an alternative way to compute $-\vartheta^\star$ that, importantly, decouples the arrival and service processes, thus leading to a substantial computational gain.

A crucial idea is that the probability of our interest (i.e., the probability that the number of customers in the queue exceeds K in a busy cycle) has the same decay rate (denoted by $\vartheta^\star$) as the probability the steady-state queue length exceeds K[7]. This equivalence is due to the fact that, in a large-deviations sense, both events correspond to the same most likely path toward the rare event (along which the process essentially behaves as being generated by the measure $\mathbb{Q}$). In this subsection, we further elaborate on this relation.

Let $\mathscr{A}(\vartheta)$ be the ll-MGF corresponding with the interarrival times $(A_n)_{n\in\mathbb{N}}$, and $\mathscr{B}^{(\ell)}(\vartheta)$ the ll-MGF corresponding with the service times $(B_n^{(\ell)})_{n\in\mathbb{N}}$ (in case there would always be jobs to serve). That is,

$$\mathscr{A}(\vartheta) := \lim_{t\to\infty}\frac{1}{t}\log\mathbb{E}e^{\vartheta A(t)},\quad \mathscr{B}^{(\ell)}(\vartheta) := \lim_{t\to\infty}\frac{1}{t}\log\mathbb{E}e^{\vartheta B^{(\ell)}(t)},$$

where $A(t)$ and $B^{(\ell)}(t)$ are defined analogously to $R(t)$ above. Then, based on the theory of Section 4.1, the decay rate of the probability $\mathbb{P}(W_Q > K)$ that the stationary queue W_Q exceeds K, i.e.,

$$\lim_{K\to\infty}\frac{1}{K}\log\mathbb{P}(W_Q > K) = -\vartheta^\star,$$

is the solution $\vartheta^\star$ of

$$\mathscr{A}(\vartheta) + \sum_{\ell=1}^{m}\mathscr{B}^{(\ell)}(-\vartheta) = 0.$$

We have thus found an alternative way to determine the decay rate $-\vartheta^\star$; observe that it uses the characteristics of the arrival and service processes in a decoupled way (which is a crucial advantage over the approach presented in Section 3, where the arrival and service processes were coupled). As we mentioned at the beginning of the section, a reference for this type of large deviations result is Ref.[7] (see in particular Equation (1.19) in that paper) and Ref.[5].

Invoking (16) we conclude that $\vartheta^\star$ solves

$$\mathscr{A}(\vartheta) = \sum_{\ell=1}^{m}\left(b^{(\ell)}\right)^{-1}(e^{\vartheta}), \tag{17}$$

where, with $S^{(\ell)}$ and $\boldsymbol{s}^{(\ell)}$ as in Equation (1),

$$b^{(\ell)}(\vartheta) = \boldsymbol{\alpha}^{(\ell)}\left(-\vartheta I_{D^{(\ell)}} - S^{(\ell)}\right)^{-1}\boldsymbol{s}^{(\ell)} \tag{18}$$

is the MGF associated with server ℓ (see, e.g., Ref.[4, Thm. 4.3]). Now that we have determined the decay rate $-\vartheta^\star$, the next goal is to find the corresponding exponentially twisted version of the arrival and service processes.

Regarding the arrival times this means that to obtain an exponentially twisted importance sampling distribution (with twist $\vartheta^\star$), we have to find an MMP process such that its ll-MGF is $\mathscr{A}^\circ(\vartheta^\star) = \mathscr{A}(\vartheta + \vartheta^\star) - \mathscr{A}(\vartheta^\star)$. As can be found in e.g. Kesidis, Walrand, and Chang[13], $\mathscr{A}(\vartheta)$ equals $\Xi(Q + (\mathrm{e}^\vartheta - 1)\,\mathrm{diag}\{\boldsymbol{\lambda}\})$, where $\Xi(M)$ denotes the largest eigenvalue of M.

Regarding the service times, exponential twisting of $\mathscr{B}^{(\ell)}$ requires us to construct $\mathbb{Q}$ such that under this new measure the ll-MGF of the service times at server ℓ equals

$$\left(\mathscr{B}^{(\ell)}\right)^\circ(\vartheta^\star) = -\left(b^{(\ell)}\right)^{-1}(\mathrm{e}^{\vartheta+\vartheta^\star}) + \left(b^{(\ell)}\right)^{-1}(\mathrm{e}^{\vartheta^\star}). \tag{19}$$

We now point out how the corresponding changes of measure can be performed; Section 4.3 focuses on the arrival process and Section 4.4 on the service processes.

4.3. Twist of the arrival process

As mentioned, $\mathscr{A}(\vartheta)$ equals $\Xi(Q + \Lambda(\mathrm{e}^\vartheta - 1))$, and hence we wish to find $\Lambda^\circ := \mathrm{diag}\{\boldsymbol{\lambda}^\circ\}$ and Q° such that $\mathscr{A}^\circ(\vartheta^\star) = \mathscr{A}(\vartheta + \vartheta^\star) - \mathscr{A}(\vartheta^\star)$, i.e.,

$$\Xi\left(Q^\circ + \Lambda^\circ(\mathrm{e}^\vartheta - 1)\right) = \Xi\left(Q + \Lambda(\mathrm{e}^{\vartheta+\vartheta^\star} - 1)\right) - \mathscr{A}(\vartheta^\star),$$

where it is noted that the right hand side of the previous expression can alternatively be written as, with I_d denoting the $d \times d$-identity matrix, $\Xi\big(Q + \Lambda(\mathrm{e}^{\vartheta+\vartheta^\star} - 1) - I_d\mathscr{A}(\vartheta^\star)\big)$.

To construct the exponentially twisted version of the arrival process, we again work with an auxiliary system. In this auxiliary system, the input MPP process, which is served at a constant rate c (larger than the mean input rate of the MMP process). As we argued earlier in this section, the decay rate $\vartheta^\star$ of the auxiliary system can be evaluated by solving $\mathscr{A}(\vartheta) = c\vartheta$. Alternatively, we can find a system of equations that $\vartheta^\star$ should satisfy, similar to the approach in Section 3.1. Let $z_i(\vartheta)$ denote the MGF of the net increase in the number of customers in the auxiliary system during a period in which the background process transitions from i to an arbitrary reference state. Mimicking the way the change of measure was constructed in Section 3.1, we find that $(\vartheta^\star, z_i)$ (with $z_i := z_i(\vartheta^\star)$) should satisfy

$$z_i = \sum_{j\neq i} \frac{q_{i,j}}{q_i} z_j \int_0^\infty q_i\,\mathrm{e}^{-q_i t}\mathrm{e}^{-\vartheta ct}\mathrm{e}^{\lambda_i(\mathrm{e}^\vartheta - 1)}\mathrm{d}t = \sum_{j\neq i} \frac{q_{ij}}{q_i - \lambda_i(\mathrm{e}^\vartheta - 1) + c\vartheta} z_j,$$

which can be rewritten as

$$\left(-\lambda_i(\mathrm{e}^\vartheta - 1) + c\vartheta\right)z_i = \sum_{j=1}^d q_{ij}z_j.$$

Inserting $c\vartheta^\star = \mathscr{A}(\vartheta^\star)$, we conclude that for $\vartheta = \vartheta^\star$ there exists a componentwise positive vector $\boldsymbol{z}$ such that

$$\big(-(\mathrm{e}^{\vartheta^\star}-1)\Lambda + I_d\mathscr{A}(\vartheta^\star)\big)\boldsymbol{z} = Q\boldsymbol{z}. \tag{20}$$

Now, let Z denote diag$\{\boldsymbol{z}\}$. Observe that any eigenvalue of $Q + (\mathrm{e}^{\vartheta+\vartheta^\star}-1)\,\Lambda - I_d\mathscr{A}(\vartheta^\star)$, is eigenvalue of

$$\begin{aligned}
&Z^{-1}\big(Q + (\mathrm{e}^{\vartheta+\vartheta^\star}-1)\,\Lambda - I_d\mathscr{A}(\vartheta^\star)\big)Z\\
&\quad = Z^{-1}QZ + (\mathrm{e}^{\vartheta+\vartheta^\star}-1)\,\Lambda - I_d\mathscr{A}(\vartheta^\star)\\
&\quad = Z^{-1}QZ + (\mathrm{e}^{\vartheta^\star}-1)\,\Lambda - I_d\mathscr{A}(\vartheta^\star) + (\mathrm{e}^{\vartheta}-1)\Lambda^\circ
\end{aligned}$$

(with $\Lambda^\circ := \Lambda\mathrm{e}^{\vartheta^\star}$) as well. Now note that, by virtue of (20),

$$Q^\circ := Z^{-1}QZ + (\mathrm{e}^{\vartheta^\star}-1)\,\Lambda - I_d\mathscr{A}(\vartheta^\star)$$

is a generator matrix. We have thus found that the desired change of measure is

$$\lambda_i^\circ := \lambda_i\mathrm{e}^{\vartheta^\star}, \quad q_{ij}^\circ := q_{ij}\frac{z_j}{z_i}, \quad q_i^\circ := q_i - \lambda_i(\mathrm{e}^{\vartheta^\star}-1) + \mathscr{A}(\vartheta^\star). \tag{21}$$

4.4. Twist of the service times

This subsection points out how to determine the exponential twist of the service processes. We start by pointing out that realizing the desired change of measure such that the ll-MGF becomes (19) amounts to exponentially twisting the service times at server ℓ by some $\zeta_\ell^\star$ that we specify below. We wish to find service times (with MGF $\overline{b}^{(\ell)}(\cdot)$) such that (19) equals $-(\overline{b}^{(\ell)})^{-1}(\mathrm{e}^{\vartheta})$. Observe that (under the usual regularity conditions) $f^{-1}(yu) = g^{-1}(y) + v$ (for all y) is equivalent to $g(x) = f(x+v)/u$ (for all x). This means that we have to identify a $\overline{b}(\cdot)$ such that

$$\overline{b}^{(\ell)}(\zeta) = \frac{b^{(\ell)}\big(\zeta + (b^{(\ell)})^{-1}(\mathrm{e}^{\vartheta^\star})\big)}{\mathrm{e}^{\vartheta^\star}},$$

which corresponds to exponentially twisting the service times at the ℓth server with twist $\zeta_\ell^\star := (b^{(\ell)})^{-1}(\mathrm{e}^{\vartheta^\star})$.

We proceed by explaining how the change of measure can be found for each server. Consider a generic server with phase-type distributed service times B, parameterized by the initial distribution $\boldsymbol{\alpha}$, the transition matrix T, and the dimension $D+1$. The twisted measure should satisfy

$$\mathbb{E}^{\mathbb{Q}}\mathrm{e}^{\zeta B} = \frac{\mathbb{E}\mathrm{e}^{(\zeta+\zeta^\star)B}}{\mathbb{E}\mathrm{e}^{\zeta^\star B}}, \tag{22}$$

where $\zeta^\star = b^{-1}(\mathrm{e}^{\vartheta^\star})$ as argued above. Consider all paths $(i_0, i_1, \dots, i_{\tau+1})$ of the underlying Markov process, starting from i_0 (sampled according to $\boldsymbol{\alpha}$) and ending at $i_{\tau+1} = D+1$. Let h_j be the time spent between states i_j and i_{j+1}. The probability of such a realization is

$$\alpha_{i_0}\frac{t_{i_0 i_1}}{t_0}t_0 \mathrm{e}^{-t_{i_0}h_0}\cdots\frac{t_{i_\tau i_{\tau+1}}}{t_{i_\tau}}t_{i_\tau}\mathrm{e}^{-t_{i_\tau}h_\tau}\,\mathrm{d}h_0\cdots\mathrm{d}h_\tau.$$

Following the line of reasoning developed in Ref.[19], the right-hand side of (22) can be written as, with $\mathbb{E}\mathrm{e}^{\zeta^\star B}=b(\zeta^\star)=\mathrm{e}^{\vartheta^\star}$,

$$\mathrm{e}^{-\vartheta^\star}\sum_{\text{all paths}}\int_0^\infty\int_{h_j:\sum_{j=0}^{\tau}h_j=h}\alpha_{i_0}t_{i_0 i_1}\mathrm{e}^{-t_{i_0}h_0}\cdots t_{i_\tau i_{\tau+1}}\mathrm{e}^{-t_{i_\tau}h_\tau}\mathrm{e}^{(\zeta+\zeta^\star)h}\mathrm{d}h_0\cdots\mathrm{d}h_\tau\,\mathrm{d}h,$$

whereas the left hand side reads

$$\sum_{\text{all paths}}\int_0^\infty\int_{h_j:\sum_{j=0}^{\tau}h_j=h}\alpha^\circ_{i_0}t^\circ_{i_0 i_1}\mathrm{e}^{-t^\circ_{i_0}h_0}\cdots t^\circ_{i_\tau i_{\tau+1}}\mathrm{e}^{-t^\circ_{i_\tau}h_\tau}\mathrm{e}^{\zeta h}\mathrm{d}h_0\cdots\mathrm{d}h_\tau\,\mathrm{d}h.$$

We wish to identify $\boldsymbol{\alpha}^\circ$ and T° such that both expressions match. To this end, solve the following eigensystem:

$$-\zeta^\star y_i=\sum_{j=1}^{D+1}t_{ij}y_j\quad\text{for }\ i=1,\ldots,D,\qquad \mathrm{e}^{\vartheta^\star}y_{D+1}=\sum_{i=1}^{D}\alpha_i y_i.$$

Then, define

$$\alpha^\circ_i:=\frac{\alpha_i}{\mathrm{e}^{\vartheta^\star}}\frac{y_i}{y_{D+1}},\qquad t^\circ_{ij}:=t_{ij}\frac{y_j}{y_i},\qquad t^\circ_i:=t_i-\zeta^\star. \tag{23}$$

The following two observations are crucial:

- ○ $(\boldsymbol{\alpha}^\circ,T^\circ)$ corresponds to a phase-type distribution. To this end, note that, by definition of the vector $\boldsymbol{y}$, the new initial distribution $\boldsymbol{\alpha}^\circ$ is non-negative and sums to 1. In addition,

$$\sum_{j\neq i}t^\circ_{ij}=\sum_{j\neq i}t_{ij}\frac{y_j}{y_i}=t_i+\zeta^\star=t^\circ_i.$$

- ○ It is an easy verification that for the above defined $(\boldsymbol{\alpha}^\circ,T^\circ)$ both MGFs match, as desired.

Thus, the proposed twist of the service times corresponds to a valid change of measure.

Note that the twisted rates we found in this and the previous subsection take the same form as those in Section 3.1, the only difference being the MGFs involved. The counterpart to the likelihood ratio given in Equation (8) is

$$L=\left(\mathrm{e}^{-\vartheta^\star\Sigma_+}\frac{z_{i_0}}{z_{i_\tau}}\right)\times\left(\mathrm{e}^{\vartheta^\star\Sigma_-}\prod_{\ell=1}^{m}\frac{y^{(\ell)}_{j_0}}{y^{(\ell)}_{j_\tau}}\right),$$

where, with a slight abuse of notation, z_{i_0} and z_{i_τ} correspond to the first and last state of the background process, and $y^{(\ell)}_{j_0}$ and $y^{(\ell)}_{j_\tau}$ correspond to the first and last phase of server ℓ, respectively, given that we observe a path of length τ as defined in Section 3.1. In line with the Kronecker decomposition found in Section 3.4, we thus see that $x_{i,j}=z_i\prod_{\ell=1}^{m}y^{(\ell)}_{j_\ell}$.

Remark 4.1. Recall that in the set-up of Section 3, we did not quite find the change of measure of the individual service times, as we only identified the twisted version of $\bar{t}^{(\ell)}_{j_\ell,k} := t^{(\ell)}_{j_\ell,\dagger}\alpha^{(\ell)}_k$ rather than the twisted version of $t^{(\ell)}_{j_\ell,\dagger}$ and $\alpha^{(\ell)}_k$ individually. In this section, we showed how to twist the rates for each server independently. In other words, we can now use the twisted rates found in this section *throughout* the entire busy cycle, i.e., also outside of fully busy periods. Note that outside of the fully busy period the (total) rates of leaving state $(i, \boldsymbol{j})$, that is, the counterparts to $\varphi_{i,\boldsymbol{j}}$ and $\varphi^{\circ}_{i,\boldsymbol{j}}$ defined in Section 3.1, are *not* equal so that the corresponding terms in the likelihood ratio do not cancel. This means that to evaluate the likelihood ratio, it needs to be continuously updated when the process is not in a fully busy period as it does not have a clean form of the type (8).

Remark 4.2. There is one important situation in which all expressions simplify considerably: that of no modulation and identical servers. In fact, in this case the interarrival times do not need to be exponential, but any renewal sequence works. To see this, suppose the interarrival times have MGF $a(\cdot)$ and the service times (at each of the servers) have MGF $b(\cdot)$. Then, (17) reads

$$-a^{-1}(\mathrm{e}^{-\vartheta}) = mb^{-1}(\mathrm{e}^{\vartheta}),$$

which is solved by $\vartheta^\star := -\log a(-m\alpha)$, where α is such that

$$\log a(-m\alpha) + \log b(\alpha) = 0,$$

which coincides with Equation (1.7) in Ref.[20].

Remark 4.3. To compute the change of measure, first $\vartheta^\star$ is found by solving (17); once $\vartheta^\star$ has been determined, we can twist the arrival and service processes as in (21) and (23), respectively. In order to solve (17), the following steps need to be performed:

- In the first place, it requires the evaluation of $\mathscr{A}(\vartheta)$ and $(b^{(\ell)})^{-1}(\mathrm{e}^{\vartheta})$ (for $\ell = 1, \dots, m$) which typically cannot be done in closed-form so that we have to resort to numerics. (In the numerical example, in Section 5, we use a bisection procedure.) To determine $\mathscr{A}(\vartheta)$ for every ϑ the eigenvalues of a d-dimensional matrix need to be found; to determine $(b^{(\ell)})^{-1}(\mathrm{e}^{\vartheta})$ the inverse of the function $b^{(\ell)}(\zeta)$ as defined in (18) is to be evaluated.
- In the second place, a numerical solver needs to be used to solve (17) (we again use bisection to perform this step in Section 5).

5. Simulation examples

In this section, we investigate a number of numerical examples to assess the efficacy of the proposed procedure, and to learn about various aspects of the rare-event behavior of the multi-server queue under study.

We start by evaluating the impact of heterogeneity among servers on the speed of decay and the relative error of the estimation procedure. We consider a two-server system with Erlang distributed service times. The service times of Server 1 are

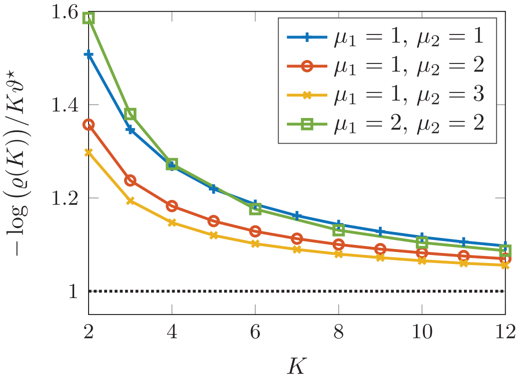

Figure 1. Ratio of $\log(\widehat{\varrho}_1(K))/K$ and its limit $-\vartheta^\star$ for a two-server system with service times that are Erlang-distributed with shape parameter 3 and rate parameters μ_1 and μ_2, respectively. The horizontal line indicates a ratio of one.

distributed with shape parameter 3 and rate parameter μ_1, while the service times of Server 2 are distributed with shape parameter 3 and rate parameter μ_2; the initial phase is distributed as $\boldsymbol{\alpha} = (0.5, 0.2, 0.3)$ in both cases. The arrival process is Poisson with rate 0.1 (i.e., not modulated). We estimate $\varrho_1(K)$ using the change of measure proposed in Section 3.1 during practically all fully busy periods; that is, we employ the algorithm stated in the appendix with C chosen very large.

Figure 1 shows that the convergence of each scaled logarithmic importance sampling estimator (calculated from 10^7 samples) to its corresponding limit $-\vartheta^\star$ appears to be faster when servers are more heterogeneous. For μ_1 and μ_2 fixed, it turns out that the corresponding relative error values are roughly independent of K (in line with our theoretical findings). More precisely, we obtained that for $\mu_1 = \mu_2 = 1$ the relative error of a generic sample is approximately 2 (across a wide range of K-values, independently of n), for $\mu_1 = 1$, $\mu_2 = 2$ it is 2.82, for $\mu_1 = 1$, $\mu_2 = 3$ it is 3.96, and for $\mu_1 = \mu_2 = 2$ it is 7.7. Thus, as one would expect, the deviation from the mean is larger when servers are more heterogeneous. The comparison of the two homogeneous examples suggests that faster service tends to have a negative impact on the relative error performance.

Furthermore, we can check numerically that, as it should be, the twisted rates are the same as those obtained in the way we described in Section 4.

For a small example as the ones just discussed, to find the alternative measure $\mathbb{Q}$, the method of Section 3 may be preferred for its conceptual simplicity. For example, of a larger dimension on the other hand, the methodology of Section 3 quickly becomes slow or even infeasible due to memory constraints, and the method of Section 4 is to be preferred.

We now discuss such a large-scale example; it is too large to be able to efficiently find the measure $\mathbb{Q}$ using the methodology of Section 3. We again assume that service times have an Erlang distribution with shape parameter 3. In a system with 10 servers, we set $\boldsymbol{\alpha}^{(\ell)} = (0.5, 0.2, 0.3)$, and choose the Erlang rate parameter as $\ell/3$, for $\ell = 1, \ldots, m$. We set $d = 4$, $\boldsymbol{\lambda} = (1, 2, 3, 4)$, and let Q have off-diagonal entries 0.1 (and diagonal entries -0.3). Recall that the dimension of the matrix A defining

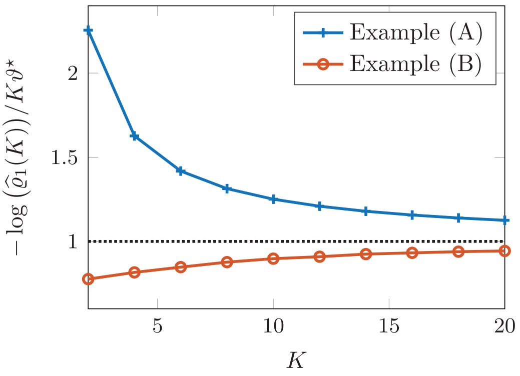

Figure 2. Ratio of $\log(\widehat{\varrho}_1(K))/K$ and its limit $-\vartheta^\star$ for (A) the example from Section 3.3, and (B) a large-scale example with 10 servers. The horizontal line indicates a ratio of one.

the eigensystem (4) is as large as $2.36 \cdot 10^5$. Despite this dimension, it turns out that with the methodology of Section 4 the change of measure can be computed in less than a second. In Figure 2, we show the ratio of the scaled logarithmic importance sampling estimator and its limit $-\vartheta^\star$ obtained in 10^5 simulation runs for (A) the small example from Section 3.3, and (B) the large-scale example with 10 servers; in both cases the change of measure is evaluated as described in Section 4, and applied during all fully busy periods.

In the approach, we have developed, we use the alternative measure $\mathbb{Q}$ only during fully busy periods. As mentioned in Remark 4.1, thanks to the decoupling of servers that we described in Section 4, in principle the twisted rates can also be applied throughout – during the entire busy cycle – rather than only during fully busy periods. Considering again the example of Section 3.3, we compare the sample confidence interval obtained under crude Monte Carlo estimation to that achieved when the change of measure is applied either throughout, or only during fully busy periods (i.e., the measure $\overline{\mathbb{Q}}$ featuring in Proposition 3.2, coinciding with Approach 2 in Section 3.2 with $C = \infty$). Figure 3 shows the scaled logarithmic unbiased estimate of $\varrho_1(K)$ averaged over 10^5 runs. The limit $-\vartheta^\star$ is indicated by the horizontal line. The

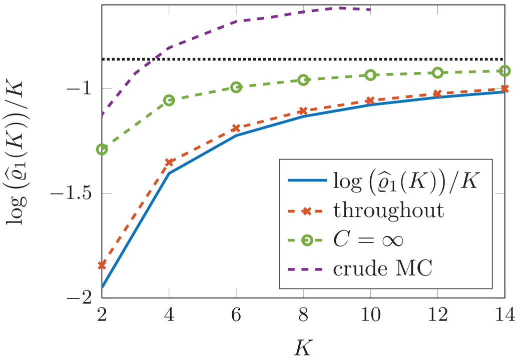

Figure 3. Depicted is $\log(\widehat{\varrho}_1(K))/K$ obtained for the example of Section 3.3. The horizontal dotted line indicates the limit value $-\vartheta^\star$. The change of measure is evaluated using the methodology developed in Section 4, and applied either throughout the entire busy cycle, only during fully busy periods ($C = \infty$), or never (crude MC), yielding the (scaled logarithmic) upper 95% confidence bounds that are indicated by the dashed lines.

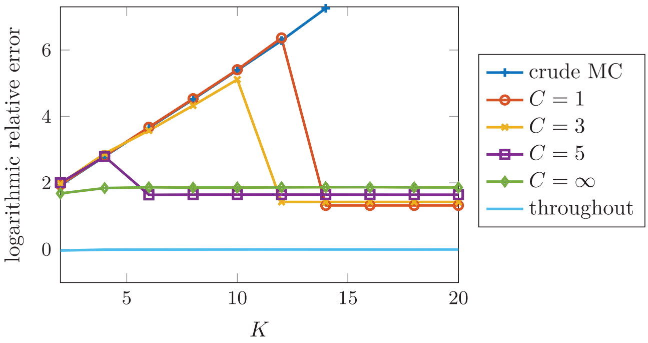

Figure 4. Logarithmic relative error values obtained for crude Monte Carlo (MC) estimation of $\varrho_1(K)$ compared to the values obtained under importance sampling when the change of measure is applied during the first $C > 0$ fully busy periods or throughout; the rates are chosen as in Section 3.3.

dashed lines indicate the scaled logarithmic upper bounds of the 95% standard normal confidence intervals. As one would expect, the change of measure significantly improves the accuracy of the estimation procedure for a fixed number of runs. In addition, we observe that when the change of measure is applied throughout (rather than only during fully busy periods) the confidence is noticeably more narrow.

We now investigate, for the same example, the impact of the choice of C when using Approach 2 in Section 3.2. In Figure 4, we compare the relative errors obtained for various values of C in 10^7 runs, where C denotes the number of fully busy periods during which the change of measure is applied. The values obtained when the twisted rates are used either never (crude Monte Carlo) or throughout the entire busy cycle are also shown. The relative error obtained for the crude Monte Carlo estimator (corresponding to $C = 0$) increases exponentially as K grows large. For large C indeed we see that relative error is independent of K. For smaller C instead, the relative error does increase with K until it drops sharply for K large enough. It appears that for small K if the event did not occur in the first period, then it may still occur afterwards even though the original measure is used, which causes a large variance.

Note that for large K the relative error corresponding to $C = 1$ appears to be the smallest. For an explanation, recall that each fully busy period during which the change of measure is applied contributes to the likelihood ratio by a factor between $e^{\vartheta^\star}\zeta_-$ and $e^{\vartheta^\star}\zeta_+$ (cf. Corollary 3.1), and thus potentially increases the variance of the estimator (where it should be kept in mind that $\zeta_- < 1$, $\zeta_+ > 1$, and $\vartheta^\star > 0$). In this sense, each additional fully busy period may have a negative impact on the quality of the estimator. Choosing a good value for C amounts to finding a proper balance between increasing the likelihood of the event of interest and minimizing the possible additional contribution to the variance of the likelihood ratio.

In the experiments that we performed, if the change of measure is applied throughout, then the relative error is remarkably low at about 0.99, substantially lower than when it is applied only during fully busy periods. We see a similar improvement in terms of estimation accuracy for the other examples discussed in this section when $\mathbb{Q}$ is applied throughout.

6. Discussion and concluding remarks

In this paper, we developed an algorithm for estimating the probability that the number of customers in a multi-server queueing system reaches a high value. The input is MPP, whereas the service-times have server-dependent phase-type distributions. We have identified explicit bounds on the probability under consideration as well as the associated likelihood ratio, which help quantifying the relative error. In particular, we have proven that the relative error of our estimator is bounded. We also develop a technique to efficiently compute the alternative measure to be used in our importance-sampling based algorithm, which remains tractable even when the dimension of the system (in terms of the number of servers and the dimensions of the phase-type distributions) is large.

A couple of experiments provide us with indications of the significant speed-up that can be achieved by the proposed algorithm relative to naive simulation. The focus is on estimating $\varrho_i(K)$, i.e., the probability that the backlog (that is, the number of customers or jobs waiting in the queue) during a busy cycle exceeds a given level K (with the background process being in state i at the beginning of the busy cycle). The method, however, directly extends to a procedure for estimating the fraction of customers or jobs entering the system while the backlog is larger than K. To this end, first note that this quantity can be written as the ratio of the mean number of customers that entered the system while the backlog is larger than K during a busy cycle, and the mean total number of customers that entered during a busy cycle. Then, the idea is to estimate the numerator and denominator of the ratio separately. The denominator does not contain a rare event, and hence can be estimated using the original measure. The numerator *does* involve a rare event, but simulating under $\mathbb{Q}$ (corresponding to a positive drift) would mean that *terminating* the busy cycle would become a rare event. Following, e.g., Ref.[8], this issue can be remedied by applying a *measure-specific dynamic importance sampling* approach, where $\mathbb{Q}$ is switched off as soon as K has been reached. Along the same lines, one could set up a procedure to estimate the fraction of customers lost in the corresponding model with a waiting room of finite size K, as was done for a similar system in Ref.[18].

In this paper, we considered specific arrival and service processes, but we anticipate that importance sampling procedures for related processes can be developed with the same techniques. As we argued, the MMP is suitable for modeling overdispersion, but there are various other processes that could be used to this end, such as the Cox processes advocated in Ref.[16].

Appendix A. Importance sampling algorithm for the embedded Markov chain

We provide pseudo-code for a single run of the importance sampling algorithm as suggested in Approach 2 in Section 3. Note that in the description of the algorithm, q_i, λ_i, and $t_{i,j}^{(\ell)}$ denote the *current* rates, that is, they may correspond to either $\mathbb{P}$ or $\mathbb{Q}$ depending on how the rates were set in the previous step of the algorithm.

If the change of measure is computed as in Section 4 instead, then we need to replace $x_{i,j}$ by $z_i \prod_{\ell=1}^{m} y_{j_\ell}^{(\ell)}$. Thanks to the decoupling, it is then also possible to apply the change of measure throughout the entire busy cycle (the algorithm needs be modified accordingly with due regard to Remark 4.1).

Algorithm. One run of the importance sampling algorithm that applies the change of measure only during the first C fully busy periods.

```
Set $N = 1, L = 1, c = 0$. Set $i$ as the initial state of the background process. Generate $j_1 \sim \boldsymbol{\alpha}^{(1)}$, and set $j_\ell = \dagger$ for $\ell = 2, \dots, m$.
while $N \in \{1, \dots, m + K\}$ do
    if $N \leqslant m$ then
        Set all rates to the original rates. Let $\varphi_{i,j} = \lambda_i + \sum_{\ell:\, j_\ell > 0} t_{j_\ell}^{(\ell)} + q_i$. Set $\boldsymbol{p}$ to be the vector with entries $\lambda_i$, $q_{ii'}$ for all $i' \neq i$, and $t_{j_\ell,k}^{(\ell)}$ for all $\ell$ such that $j_\ell \neq \dagger$ and $k \in \{1, \dots, D^{(\ell)}, \dagger\}$. Generate the next event from the discrete distribution $\boldsymbol{p}/\varphi_{i,j}$.
        if Arrival then
            $N \leftarrow N + 1$
            if $N > m$ then
                $c \leftarrow c + 1$
                if $c \le C$ then
                    $L \leftarrow L\, x_{i,j}$
                end if
            end if
        else if Transition of the background process then
            $i \leftarrow i'$, where $i'$ corresponds to entry $q_{ii'}$ of $\boldsymbol{p}$
        else if Phase transition at server $\ell$ to $k \leqslant D^{(\ell)}$ then
            $j_\ell \leftarrow k$
        else if Phase transition at server $\ell$ to $\dagger$ then
            $j_\ell \leftarrow \dagger$, and $N \leftarrow N - 1$
        end if
    else if $N > m$ then
        if $c \leqslant C$ then
            Set all rates as in (6).
        end if
        Update $\varphi_{i,j}$, and set $\boldsymbol{p}$ as for the partially busy period but including rates $\bar{t}_{j_\ell,k}^{(\ell)}$. Generate the next event from the discrete distribution $\boldsymbol{p}/\varphi_{i,j}$.
        if Phase transition at server $\ell$ with departure then
            $j_\ell \leftarrow k$, where $k$ corresponds to entry $\bar{t}_{j_\ell,k}^{(\ell)}$ of $\boldsymbol{p}$, and $N \leftarrow N - 1$
            if $N = m$ and $c \leqslant C$ then
                $L \leftarrow L\, \mathrm{e}^{\vartheta^\star}/x_{i,j}$
            end if
        else if Other transition then
            Proceed as for the partially busy period.
        end if
    end if
end while
if $N > m + K$ and $c \leqslant C$ then
    $L \leftarrow L\, \mathrm{e}^{-K\vartheta^\star}/x_{i,j}$
end if
return $L\, \mathbb{1}\{N > m + K\}$
```

Acknowledgments

The authors wish to thank Thomas Taimre for his helpful comments on earlier drafts of this paper, and Leonardo Rojas-Nandayapa for valuable comments concerning the presentation of Theorem 1.

Funding

Julia Kuhn is supported by the Australian Research Council (ARC), Grant DP130100156. Michel Mandjes' research is partly funded by the NWO Gravitation Project NETWORKS, Grant 024002003.

References

[1] Armony, M.; Ward, A. R. Fair dynamic routing in large-scale heterogeneous-server systems. Oper. Res. **2010**, *58*, 624–637.

[2] Asmussen, S. *Applied Probability and Queues*, 2nd ed.; Springer-Verlag: New York, 2008; vol. 51.

[3] Asmussen, S.; Glynn, P. W. Stochastic simulation: algorithms and analysis. In *Stochastic Modelling and Applied Probability*; Asmussen, S.; Glynn, P. W., Eds.; Springer: New York, 2007.

[4] Bladt, M. A review on phase-type distributions and their use in risk theory. ASTIN Bull. **2005**, *35*, 145–161.

[5] Duffield, N.; O'Connell, N. Large deviations and overflow probabilities for the general single-server queue, with applications. Math. Proc. Cambridge Philos. Soc. **1995**, *118*, 363–374.

[6] Duffield, N. G.; Whitt, W. Large deviations of inverse processes with nonlinear scalings. Ann. Appl. Prob. **1998**, *8*, 995–1026.

[7] Glynn, P. W.; Whitt, W. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. J. Appl. Prob. **1994**, *31*, 131–156.

[8] Goyal, A.; Shahabuddin, P., Heidelberger, P., Nicola, V.; Glynn, P. A. Unified framework for simulating markovian models of highly reliable systems. IEEE Trans. Comput. **1992**, *41*, 36–51.

[9] Green, L. Queueing analysis in healthcare. In *Patient Flow: Reducing Delay in Healthcare Delivery*; Hall, R. W., Ed.; Springer: Boston, MA, USA, **2006**; pp. 281–307.

[10] Heemskerk, M.; van Leeuwaarden, J.; Mandjes, M. Scaling limits for infinite-server systems in a random environment. Stoch. Syst.`arXiv:1602.00499`, in press, **2016**. https://projecteuclid.org/euclid.ssy/1495785616

[11] Jongbloed, G.; Koole, G. Managing uncertainty in call centers using Poisson mixtures. Appl. Stoch. Model. Bus. Ind. **2001**, *17*, 307–318.

[12] Juneja, S. Importance sampling and the cyclic approach. Oper. Res. **2001**, *49*, 900–912.

[13] Kesidis, G.; Walrand, J.; Chang, C.-S. Effective Bandwidths for Multiclass Markov Fluids and other ATM Sources. IEEE/ACM Trans. Netw. **1993**, *1*, 424–428.

[14] Kim, J.; Ahn, H.; Righter, R. Managing queues with heterogeneous servers. J. Appl. Prob. **2011**, *48*, 435–452.

[15] Kim, S.; Whitt, W. Are call center and hospital arrivals well modeled by nonhomogeneous Poisson processes? Manuf. Serv. Oper. Manag. **2014**, *16*, 464–480.

[16] Koops, D.; Boxma, O.; Mandjes, M. Networks of $\cdot/G/\infty$ Queues with shot-noise-driven arrival intensities. Queueing Syst.`arXiv:1608.04924`, in press, **2017**. https://link.springer.com/article/10.1007/s11134-017-9520-7

[17] Laub A. J. *Matrix Analysis for Scientists and Engineers*; SIAM: Philadelphia, PA, USA, 2005.

[18] Mandjes, M. Rare event analysis of batch-arrival queues. Telecommun. Syst. **1996**, *6*, 161–180.

[19] Mandjes, M.; Ridder, A. Finding the conjugate of Markov fluid processes. Prob. Eng. Inf. Sci. **1995**, *9*, 297–315.

[20] Sadowsky, J. S. Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/*m* queue. IEEE Trans. Autom. Control **1991**, *36*, 1383–1394.

[21] Takahashi, Y. Asymptotic exponentially of the tail of the waiting-time distribution in a Ph/Ph/*c* queue. Adv. Appl. Prob. **1981**, *4*, 619–630.

[22] Tijms, H. *A First Course in Stochastic Models*; Wiley: Chichester, 2003.

[23] Whitt, W., Green, L. V.; Kolesar, P. J. Coping with time-varying demand when setting staffing requirements for a service system. Prod. Oper. Manag. **2007**, *16*, 13–39.