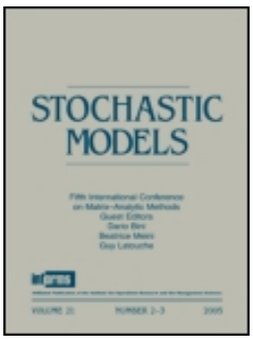

Taylor & Francis
Taylor & Francis Group

Stochastic Models

ISSN: 1532-6349 (Print) 1532-4214 (Online) Journal homepage: http://www.tandfonline.com/loi/lstm20

Markov-modulated infinite-server queues driven by a common background process

Michel Mandjes & Koen De Turck

To cite this article: Michel Mandjes & Koen De Turck (2016) Markov-modulated infinite-server queues driven by a common background process, Stochastic Models, 32:2, 206-232, DOI: 10.1080/15326349.2015.1100085

To link to this article: http://dx.doi.org/10.1080/15326349.2015.1100085

Published online: 14 Dec 2015.

STOCHASTIC MODELS
2016, VOL. 32, NO. 2, 206–232
http://dx.doi.org/10.1080/15326349.2015.1100085

Taylor & Francis
Taylor & Francis Group

Markov-modulated infinite-server queues driven by a common background process

Michel Mandjes[a,b] and Koen De Turck[c,d]

[a]Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, the Netherlands; [b]CWI, Amsterdam, the Netherlands; [c]TELIN, Ghent University, Gent, Belgium; [d]Laboratoire Signaux et Systémes, École CentraleSupélec, Université Paris Saclay, Gif-sur-Yvette, France

ABSTRACT

This paper studies a system with multiple infinite-server queues that are modulated by a *common* background process. If this background process, being modeled as a finite-state continuous-time Markov chain, is in state j, then the arrival rate into the i-th queue is $\lambda_{i,j}$, whereas the service times of customers present in this queue are exponentially distributed with mean $\mu_{i,j}^{-1}$; at each of the individual queues all customers present are served in parallel (thus reflecting their infinite-server nature).

Three types of results are presented: in the first place (i) we derive differential equations for the probability-generating functions corresponding to the distributions of the transient and stationary numbers of customers (jointly in all queues), then (ii) we set up recursions for the (joint) moments, and finally (iii) we establish a central limit theorem in the asymptotic regime in which the arrival rates as well as the transition rates of the background process are simultaneously growing large.

ARTICLE HISTORY
Received March 2015
Accepted September 2015

KEYWORDS
Markov-modulation; infinite-server queues; central limit theorems

MATHEMATICS SUBJECT CLASSIFICATION
60K25; 60F05; 60J60; 60J27

1. Introduction

Markov-modulated queueing systems are resources at which customers arrive and depart, but with the special feature that the corresponding interarrival times and service times depend on the state of an external Markovian process, usually referred to as "background process". In most studies, such a background process is represented by a finite-state irreducible continuous-time Markov chain. Markov-modulated queues have been studied intensively over the past, say, four decades, with a primary focus on developing techniques to determine the underlying stationary distribution. For further background, we refer to the monographs by Asmussen[2] and Neuts[15]; see also, e.g., Refs.[9,13,16].

In the case of Markov-modulated single-server queues, in which the arrival rates and services rates do not depend on the number of customers present (i.e., they are affected by the state of the background process only), the stationary distribution

CONTACT Michel Mandjes m.r.h.mandjes@uva.nl Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands.

of the number of costumers, jointly with the state of the background process, is of *matrix-geometric* form. It is noted that this property can be considered a true matrix-counterpart of the scalar M/M/1 queue (in which the stationary distribution has a scalar-geometric distribution).

The corresponding Markov-modulated *infinite-server* queue allows considerably less explicit results. In Ref.[14] a system of partial (ordinary) differential equations is derived for the probability-generating function of the transient (stationary, respectively) number of customers in the system (jointly with the state of the background process). These differential equations can then be exploited to set up a recursive procedure that facilitates the computation of all moments. Importantly, the stationary number of customers does *not* have some sort of "matrix Poisson distribution", and in this sense the queue cannot be seen as a direct generalization of its scalar-counterpart, the ordinary M/M/∞ queue.

When stochastic systems do not allow any explicit analysis, a common procedure to gain insight into the system is to impose a particular parameter scaling, and to then consider the resulting asymptotic regime. In a series of more recent articles[1,4,6], such an approach has been followed; in particular, by scaling the arrival rates as well as the transition rates of the background process, it is shown that the (transient and stationary) number of customers obeys a central limit theorem (CLT). If the background process evolves faster than the arrival process, the system essentially behaves as a scalar M/M/∞ queue in diffusion scaling, whereas in the opposite regime, the resulting Gaussian process has a more refined structure, in which the deviation matrix (associated with the background process) plays a crucial role.

The key novelty of the present paper is that it considers a system with *multiple* Markov-modulated infinite-server queues, which are *driven by the same background process*—this common background process is denoted by J throughout this paper. The motivation behind studying this model lies in the fact that in many practical situations, individual queues react to the same "outer world"; one could, for instance, think of a wireless network, in which users react to the same channel conditions, or a road traffic network in which all drivers are affected by the same weather conditions.

More concretely, in this paper we study a queueing model in which the arrival rate of the i-th queue is $\lambda_{i,j}$ if J is in state j, while the service times of all individual customers present in the i-th queue are then exponentially distributed with mean $\mu_{i,j}^{-1}$. At each of the queues all customers present are served in parallel. To keep the notation light, we focus on the situation with $i \in \{1, 2\}$, but the analysis naturally extends to any finite number of Markov-modulated infinite-server queues.

It is important to realize that for single-server models, this type of coupled model typically does not allow any explicit analysis. This is primarily due to discontinuities that arise when (at least) one of the queues is idle: when J is in state j, the service rate in queue i is $\mu_{i,j}$ as long as the number of customers in this queue, say k, is in $\{1, 2 \ldots\}$, and 0 if $k = 0$. It is observed, however, that for their infinite-server counterparts such discontinuity does *not* exist: the service rate $k\mu_{i,j}$ applies to *any* $k \in \{0, 1, \ldots\}$. As we show in this paper, it is an immediate consequence of this fact

that coupled Markov-modulated infinite-server queues are essentially as complex as their *non-coupled* counterpart. It is noted that some related results for Markov-modulated Ornstein-Uhlenbeck processes (driven by a common background process) have recently been reported in Ref.[12]. In addition, related results on multiple queues driven by the same underlying continuous-time Markov chain have been reported in Ref.[3].

We now detail the contributions of this paper. At a high level, the main objective is to extend the results of Refs.[4,14] for non-coupled Markov-modulated infinite-server queues to their coupled counterpart. More specifically, the following three types of results are presented.

(i) In the first place we set up systems of differential equations for the probability-generating function of the (joint) distribution of the numbers of customers in both queues; these are partial differential equations when considering the transient distribution, and ordinary differential equations for its stationary counterpart. The results are in terms of *systems* of equations, as they cover the number of customers present, jointly with the state of the background process.

(ii) In the second place we develop recursions for the (joint) moments, for both the transient and stationary distribution. In addition, we give explicit expressions for means, variances, and covariances, which turn out to simplify drastically in various particular limiting regimes.

(iii) We finally establish a CLT in the asymptotic regime in which a scaling is imposed on the arrival rates as well as the transition rates of the background process J. Importantly, following the ideas presented in Ref.[4], the arrival rates are inflated by a factor N, whereas the transition rates of J are scaled as N^f for some $f > 0$; as N grows large, one ends up in different limiting regimes, depending on the value of f. For $f > 1$ it is concluded that the resulting system behaves essentially as the diffusion version of two independently operating M/M/∞ queues, while for $f < 1$ one obtains a Gaussian process in which the effect of the common background process becomes explicitly visible.

As pointed out in detail in Ref.[4], the Markov-modulated infinite-server queue comes in two variants, in this paper systematically referred to as Model I and Model II. In the former model, the departure rates at any point in time are determined by the current state of the background process; as a consequence, this rate may (possibly multiply) change during a customer's stay in the system. In the latter model, however, the departure time is determined by the state of the background process that the customer sees upon arrival (and can therefore be sampled the moment the customer enters the system). We provide a detailed description of these two variants in Section 2.

The rest of the paper is organized as follows. Sections 3 and 4 characterize the probability-generating functions related to the (transient and stationary) numbers of customers at both queues, as well as corresponding moments, for Model I and Model II, respectively. Then these results are used to explicitly find, for both models, variances and covariances in Sections 5 and 6. Central limit theorems (when

imposing particular scalings on the arrival rates and the transition rates of the background process) are established in Sections 7 and 8. A numerical illustration is presented in Section 9. In Section 10 the paper is concluded by a brief discussion of the applicability of the results, as well as an outlook.

2. Model and preliminaries

We start this section by giving a detailed model description of the coupled system of Markov-modulated infinite-server queues. A first component of this model is the so-called *background process* $(J(t))_{t \geq 0}$, which is an irreducible, finite-state Markov process on a finite state space $\{1, \ldots, d\}$. Let the corresponding transition rates be given through the transition rate matrix $Q = (q_{ij})_{i,j=1}^{d}$; throughout, $q_{ij} \geq 0$ for $i \neq j$, and $q_i := -q_{ii} = \sum_{j \neq i} q_{ij}$. In addition, the (unique) invariant distribution is denoted by (the column vector) $\boldsymbol{\pi}$. We adopt here and in the sequel the convention that we write vectors in bold fonts; vectors are consistently understood as *column* vectors, unless stated otherwise.

In the setting studied in this paper we suppose that the process $J(\cdot)$ modulates *two* inifinite-server systems; as mentioned in the introduction, all results can be straightforwardly extended to the case of three or more queues, but for reasons of transparency we have chosen to leave this out. While $J(\cdot)$ is in state $j \in \{1, \ldots, d\}$, the process that describes the number of jobs present in system $i \in \{1, 2\}$, in this paper denoted by $(M_i(t))_{t \geq 0}$, locally behaves as an infinite-server queue fed by a Poisson process of rate $\lambda_{i,j}$, while the service times of each of the customers present in the i-th system are exponentially distributed with mean $\mu_{i,j}^{-1}$. For ease, we let both systems start off empty: $M_i(0) = 0$, for $i = 1, 2$. Also, we let M_i denote the stationary version of $M_i(t)$.

As pointed out in the introduction, two variants are to be distinguished. They can be described as follows.

- In the first variant (in the sequel referred to as Model I), all jobs present at a certain time instant t are subject to a hazard rate determined by the state of background chain at time t, regardless of when they arrived. In other words, when k customers are present in queue i and J is in state j, the infinitesimal transition rate corresponding to a customer leaving from this queue is $k\mu_{i,j}$.
- In the second variant (to be referred to as Model II), the service rate is determined by the background state as seen by the job upon its arrival. This means that if there are k customers in queue i that have entered when J was in state j, the infinitesimal transition rate corresponding to one of these customers leaving is $k\mu_{i,j}$.

For notational convenience, we introduce the $d \times d$ matrices $\Delta(\boldsymbol{\lambda}_i) := \mathrm{diag}\{\boldsymbol{\lambda}_i\}$ and $\Delta(\boldsymbol{\mu}_i) := \mathrm{diag}\{\boldsymbol{\mu}_i\}$. In the sequel we frequently use the "time-average arrival rates" and "time average departure rates", being defined by

$$\lambda_{i,\infty} := \sum_{j=1}^{d} \pi_j \lambda_{i,j} = \boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{\lambda}_i, \qquad \mu_{i,\infty} := \sum_{j=1}^{d} \pi_j \mu_{i,j} = \boldsymbol{\pi}^{\mathrm{T}} \boldsymbol{\mu}_i,$$

respectively. We let $\boldsymbol{a}$ be the column vector corresponding to the initial distribution of the background process: $a_i := \mathbb{P}(J(0) = i)$ for $i = 1, \ldots, d$; in addition, we denote $P(t) := (p_{ij}(t))_{i,j=1}^d$, with $p_{ij}(t)$ denoting the transient probabilities $\mathbb{P}(J(t) = j \,|\, J(0) = i) = (e^{Qt})_{ij}$.

An important concept in this paper is the so-called *deviation matrix*, see, e.g., Ref.[10] for more background. Recall that the deviation matrix $D = (D_{ij})_{i,j=1}^d$ of the finite-state Markov chain $J(\cdot)$ is defined through

$$D_{ij} := \int_0^\infty (p_{ij}(t) - \pi_j)\mathrm{d}t,$$

or, in matrix notation, $D = \int_0^\infty (e^{Qt} - \Pi)\mathrm{d}t$, with $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$. The *fundamental matrix* F is given by $F := D + \Pi$. A number of standard identities play a role below, in particular $QF = FQ = \Pi - I$, $\Pi F = F\Pi = \Pi$, and $F\mathbf{1} = \mathbf{1}$.

3. Model I: Distribution and moments

In this section we consider the stationary and transient distribution associated to Model I, focusing on setting up a system of differential equations for the corresponding probability-generating functions, and developing a recursion for all moments; for Model II similar computations are done in the next section.

3.1. *Stationary behavior*

Our objective is to find the steady-state distribution $(\boldsymbol{p}_{k,\ell})_{k,\ell=1}^\infty$, where each $\boldsymbol{p}_{k,\ell}$ is a vector in $\mathbb{R}^d$, whose j-th entry is defined as

$$[\boldsymbol{p}_{k,\ell}]_j := \mathbb{P}(M_1 = k, M_2 = \ell, J = j),$$

with $j = 1, \ldots, d$. The vector-valued probability-generating function (pgf) $\boldsymbol{p}(w, z)$ is given by, with $|w|, |z| \le 1$, and $j = 1, \ldots, d$,

$$[\boldsymbol{p}(w,z)]_j := \mathbb{E}\left(w^{M_1} z^{M_2} 1_{\{J=j\}}\right) = \sum_{k=0}^\infty \sum_{\ell=0}^\infty [\boldsymbol{p}_{k,\ell}]_j w^k z^\ell.$$

It is noted that in Model I the trivariate process $(M_1(t), M_2(t), J(t))_{t\ge 0}$ is a continuous-time Markov chain, attaining values in $\mathbb{N} \times \mathbb{N} \times \{1, \ldots, d\}$.

To study $\boldsymbol{p}_{k,\ell}$, we first define its transient counterpart through, for $j = 1, \ldots, d$,

$$[\boldsymbol{p}_{k,\ell}(t)]_j := \mathbb{P}(M_1(t) = k, M_2(t) = \ell, J(t) = j).$$

As an immediate consequence of the Chapman-Kolmogorov equation, it follows that

$$\begin{aligned}\frac{\partial \boldsymbol{p}_{k,\ell}(t)}{\partial t} &= \boldsymbol{p}_{k-1,\ell}(t)\cdot\Delta(\boldsymbol{\lambda}_1)\\ &\quad+\boldsymbol{p}_{k,\ell-1}(t)\cdot\Delta(\boldsymbol{\lambda}_2)+\boldsymbol{p}_{k,\ell}(t)\cdot(Q-\Delta(\boldsymbol{\lambda}_1)-\Delta(\boldsymbol{\lambda}_2)\\ &\quad-k\Delta(\boldsymbol{\mu}_1)-\ell\Delta(\boldsymbol{\mu}_2))\\ &\quad+\boldsymbol{p}_{k+1,\ell}(t)\cdot(k+1)\Delta(\boldsymbol{\mu}_1)+\boldsymbol{p}_{k,\ell+1}(t)\cdot(\ell+1)\Delta(\boldsymbol{\mu}_2)\end{aligned}\tag{1}$$

for $k, \ell = 0, 1, \ldots$ (where we put $\boldsymbol{p}_{-1,\ell}(t) = \boldsymbol{p}_{k,-1}(t) = \mathbf{0}$).

This identity is to be equated to 0 to obtain the stationary distribution $(\boldsymbol{p}_{k,\ell})_{k,\ell=1}^{\infty}$; note that in this case we need to set $\boldsymbol{p}_{-1,\ell} = \boldsymbol{p}_{k,-1} = \mathbf{0}$. Now multiply the equation by $w^k z^\ell$ and sum over k and ℓ, so as to obtain, relying on standard properties of pgf s, the following differential equation for $\boldsymbol{p}(w, z)$:

$$\begin{aligned}w\boldsymbol{p}(w,z)\cdot\Delta(\boldsymbol{\lambda}_1)+z\boldsymbol{p}(w,z)\cdot\Delta(\boldsymbol{\lambda}_2)+\boldsymbol{p}(w,z)\cdot(Q-\Delta(\boldsymbol{\lambda}_1)-\Delta(\boldsymbol{\lambda}_2))\\ -(w-1)\frac{\partial\boldsymbol{p}}{\partial w}\cdot\Delta(\boldsymbol{\mu}_1)-(z-1)\frac{\partial\boldsymbol{p}}{\partial z}\cdot\Delta(\boldsymbol{\mu}_2)=\mathbf{0}^{\mathrm{T}};\end{aligned}$$

here we tacitly assumed that the pgf s are row vectors. The differential equation can be rewritten in the following compact form.

Proposition 3.1.1. *The pgf $\boldsymbol{p}(w, z)$ satisfies the differential equation*

$$\begin{aligned}\boldsymbol{p}(w,z)\,Q+(w-1)\left(\boldsymbol{p}(w,z)\,\Delta(\boldsymbol{\lambda}_1)-\frac{\partial\boldsymbol{p}}{\partial w}\,\Delta(\boldsymbol{\mu}_1)\right)\\ +(z-1)\left(\boldsymbol{p}(w,z)\,\Delta(\boldsymbol{\lambda}_2)-\frac{\partial\boldsymbol{p}}{\partial z}\,\Delta(\boldsymbol{\mu}_2)\right)=\mathbf{0}^{\mathrm{T}}.\end{aligned}$$

Our next objective is to use the differential equation for the pgf to develop an algorithm for computing all (joint) moments. It relies on the property that differentiating the pgf and inserting the argument 1 yields the so-called factorial moments.

It takes some elementary calculus to verify that, for any "sufficiently differentiable" function $\varphi(\cdot, \cdot)$,

$$\frac{\partial^{k+\ell}}{\partial w^k\partial z^\ell}(w-1)\varphi(w,z)=(w-1)\,\frac{\partial^{k+\ell}\varphi(w,z)}{\partial w^k\partial z^\ell}+k\,\frac{\partial^{k+\ell-1}\varphi(w,z)}{\partial w^{k-1}\partial z^\ell}.\tag{2}$$

Define the (row-)vectors of the mixed factorial moments by $\boldsymbol{\Gamma}_{k,\ell}\in\mathbb{R}^d$; its j-th entry equals

$$[\boldsymbol{\Gamma}_{k,\ell}]_j := \mathbb{E}\left((M_1)_k\,(M_2)_\ell\cdot 1_{\{J=j\}}\right),$$

using the Pochhammer notation for the falling factorial, i.e.,

$$(N)_k := \frac{N!}{(N-k)!} = N(N-1)\cdots(N-k+1).$$

The next step is to combine Proposition 3.1 with (2). It is a matter of applying standard rules for pgf s to obtain

$$\boldsymbol{\Gamma}_{k,\ell}Q = k\boldsymbol{\Gamma}_{k,\ell}\,\Delta(\boldsymbol{\mu}_1)-k\boldsymbol{\Gamma}_{k-1,\ell}\,\Delta(\boldsymbol{\lambda}_1)+\ell\boldsymbol{\Gamma}_{k,\ell}\,\Delta(\boldsymbol{\mu}_2)-\ell\boldsymbol{\Gamma}_{k,\ell-1}\,\Delta(\boldsymbol{\lambda}_2),$$

so that we have established the validity of the following iterative procedure.

Proposition 3.1.2. *The factorial moments* $\mathbf{\Gamma}_{k,\ell}$ *satisfy the recursion*

$$\mathbf{\Gamma}_{k,\ell} = \left(k\mathbf{\Gamma}_{k-1,\ell}\,\Delta(\boldsymbol{\lambda}_1) + \ell\mathbf{\Gamma}_{k,\ell-1}\,\Delta(\boldsymbol{\lambda}_2)\right)(k\Delta(\boldsymbol{\mu}_1) + \ell\Delta(\boldsymbol{\mu}_2) - Q)^{-1},$$

to be initialized with $\mathbf{\Gamma}_{0,0} = \boldsymbol{\pi}^{\mathrm{T}}$.

For $k = 0$ or $\ell = 0$, this yields precisely the recursion found in O'Cinneide and Purdue[14] (covering the case of a single Markov-modulated infinite-server queue).

3.2. *Transient behavior*

Where the previous subsection studied the stationary behavior of Model I, we now consider the corresponding transient behavior. As will turn out, the system of ordinary differential equations becomes a system of partial differential equations (as was of course to be expected). In addition, each iteration in the recursion for the factorial moments now requires solving a system of non-homogeneous linear differential equations.

We first focus on characterizing the pgf $\boldsymbol{p}(t, w, z)$, defined in the obvious way. In the same manner as before, from the Chapman-Kolmogorov equation (1) we find the following system of partial differential equations.

Proposition 3.2.1. *The pgf* $\boldsymbol{p}(t, w, z)$ *satisfies the differential equation*

$$\begin{aligned}\boldsymbol{p}(t,w,z)\,Q + (w-1)\left(\boldsymbol{p}(t,w,z)\,\Delta(\boldsymbol{\lambda}_1) - \frac{\partial \boldsymbol{p}}{\partial w}\,\Delta(\boldsymbol{\mu}_1)\right)\\ + (z-1)\left(\boldsymbol{p}(t,w,z)\,\Delta(\boldsymbol{\lambda}_2) - \frac{\partial \boldsymbol{p}}{\partial z}\,\Delta(\boldsymbol{\mu}_2)\right) = \frac{\partial \boldsymbol{p}}{\partial t}.\end{aligned}$$

Let $\mathbf{\Gamma}_{k,\ell}(t)$ be the time-dependent counterpart of $\mathbf{\Gamma}_{k,\ell}$. It is a matter of straightforward calculus to obtain that

$$\begin{aligned}\mathbf{\Gamma}_{k,\ell}(t)\,Q - \mathbf{\Gamma}'_{k,\ell}(t) = k\mathbf{\Gamma}_{k,\ell}(t)\,\Delta(\boldsymbol{\mu}_1) - k\mathbf{\Gamma}_{k-1,\ell}(t)\,\Delta(\boldsymbol{\lambda}_1) + \ell\mathbf{\Gamma}_{k,\ell}(t)\,\Delta(\boldsymbol{\mu}_2)\\ -\ell\mathbf{\Gamma}_{k,\ell-1}(t)\,\Delta(\boldsymbol{\lambda}_2),\end{aligned}$$

or, equivalently,

$$\begin{aligned}\mathbf{\Gamma}'_{k,\ell}(t) = \mathbf{\Gamma}_{k,\ell}(t)\,(Q - k\Delta(\boldsymbol{\mu}_1) - \ell\Delta(\boldsymbol{\mu}_2)) + k\mathbf{\Gamma}_{k-1,\ell}(t)\,\Delta(\boldsymbol{\lambda}_1)\\ +\ell\mathbf{\Gamma}_{k,\ell-1}(t)\,\Delta(\boldsymbol{\lambda}_2).\end{aligned}$$

We thus conclude that for $\mathbf{\Gamma}_{k-1,\ell}(t)$ and $\mathbf{\Gamma}_{k,\ell-1}(t)$ given, $\mathbf{\Gamma}_{k,\ell}(t)$ can be determined by solving a non-homogeneous system of linear differential equations; cf. Ref. [14, Theorem 3.2] for the case of a single Markov-modulated infinite-server system. As a consequence, this provides us with a recursive scheme to evaluate the transient factorial moments $\mathbf{\Gamma}_{k,\ell}(t)$; recall that we assumed that $M_i(0) = 0$ for $i = 1, 2$.

Proposition 3.2.2. *The factorial moments* $\mathbf{\Gamma}_{k,\ell}(t)$ *satisfy the recursion*

$$\begin{aligned}\mathbf{\Gamma}'_{k,\ell}(t) = \mathbf{\Gamma}_{k,\ell}(t)\,(Q - k\Delta(\boldsymbol{\mu}_1) - \ell\Delta(\boldsymbol{\mu}_2)) + k\mathbf{\Gamma}_{k-1,\ell}(t)\,\Delta(\boldsymbol{\lambda}_1)\\ +\ell\mathbf{\Gamma}_{k,\ell-1}(t)\,\Delta(\boldsymbol{\lambda}_2), \quad \mathbf{\Gamma}_{k,\ell}(0) = \mathbf{0}^{\mathrm{T}},\end{aligned}$$

to be initialized with $\mathbf{\Gamma}_{0,0}(t) = \boldsymbol{a}^{\mathrm{T}}P(t)$.

4. Model II: Distribution and moments

As we did for Model I in the previous section, we now analyze the stationary and transient distributions associated with Model II, again by setting up differential equations for the probability-generating functions, as well as a recursive procedure that generates all moments.

4.1. *Stationary behavior*

First observe that for Model II the trivariate process $(M_1(t), M_2(t), J(t))_{t\geq 0}$ is *not* Markov, as for each customer one needs to know what state J was in when it arrived. This is why we here use a description with a slightly more general state space: we keep track of the number of jobs present of each type, where "type" refers to the state of the background process as seen by the customer upon arrival. To this end, we work with the d-dimensional stochastic process

$$\boldsymbol{M}_i(t) = (M_{i,1}(t), \ldots, M_{i,d}(t))_{t\geqslant 0},$$

where the k-th entry of this vector denotes the number of customers of type k in the i-th system at time t, for $i = 1, 2$; the vector $\boldsymbol{M}_i = (M_{i,1}, \ldots, M_{i,d})$ is its stationary counterpart. The transient total number of customers in queue i is (obviously) equal to $M_i(t) := \sum_{m=1}^{d} M_{i,m}(t)$, and the stationary total number equal to $M_i := \sum_{m=1}^{d} M_{i,m}$.

The j-th entry of the pgf $\boldsymbol{p}(t, \boldsymbol{w}, \boldsymbol{z})$ is defined by, for $j = 1, \ldots, d$ and $|w_m|, |z_m| < 1$,

$$[\boldsymbol{p}(t, \boldsymbol{w}, \boldsymbol{z})]_j = \mathbb{E}\left(\prod_{m=1}^{d} w_m^{M_{1,m}(t)}\, z_m^{M_{2,m}(t)} 1_{\{J(t)=j\}}\right).$$

In addition, E_m is a matrix for which $[E_m]_{mm} = 1$, and whose other entries are zero (or, in other words, the matrix E_m equals $\mathrm{diag}\{\boldsymbol{e}_m\}$, where $\boldsymbol{e}_m$ is the m-th unit vector, having a one on the m-th position and zeros elsewhere); the multiplication $\boldsymbol{p}E_m$ thus results in a (row-)vector that leaves the m-th entry of the row-vector $\boldsymbol{p}$ unchanged while the other entries become zero.

With the pgf $\boldsymbol{p}(\boldsymbol{w}, \boldsymbol{z})$ defined in the obvious way, the system of differential equations for the stationary case turns out to be the following.

Proposition 4.1.1. *The pgf $\boldsymbol{p}(\boldsymbol{w},\boldsymbol{z})$ satisfies the differential equation*

$$\boldsymbol{p}(\boldsymbol{w},\boldsymbol{z})\,Q+\sum_{m=1}^{d}(w_m-1)\left(\lambda_{1,m}\,\boldsymbol{p}(\boldsymbol{w},\boldsymbol{z})\,E_m+\mu_{1,m}\,\frac{\partial \boldsymbol{p}}{\partial w_m}\right)$$
$$+\sum_{m=1}^{d}(z_m-1)\left(\lambda_{2,m}\,\boldsymbol{p}(\boldsymbol{w},\boldsymbol{z})\,E_m+\mu_{2,m}\,\frac{\partial \boldsymbol{p}}{\partial z_m}\right)=\mathbf{0}^{\mathrm{T}}.$$

The proof of this proposition is straightforward and follows the same lines as before: we consider the generator of the Markov process and transform the Chapman-Kolmogorov equation.

Also, the corresponding moments can be computed as before. To this end, we first define the factorial moments using the Pochhammer notation introduced earlier:

$$[\boldsymbol{\Gamma}_{\boldsymbol{k},\boldsymbol{\ell}}]_j:=\mathbb{E}\left(\prod_{m=1}^{d}(M_{1,m})_{k_m}\cdot\prod_{m=1}^{d}(M_{2,m})_{\ell_m}\cdot 1_{\{J=j\}}\right),$$

as well as the differential operator $\mathscr{D}(\boldsymbol{k},\boldsymbol{\ell})[\cdot]$:

$$\mathscr{D}(\boldsymbol{k},\boldsymbol{\ell})[f(\boldsymbol{w},\boldsymbol{z})]:=\frac{\partial^{k_1+\cdots+k_d+\ell_1+\cdots+\ell_d}}{\partial w_1^{k_1}\cdots\partial w_d^{k_d}\,\partial z_1^{\ell_1}\cdots\partial z_d^{\ell_d}}f(\boldsymbol{w},\boldsymbol{z}).$$

Clearly, $\boldsymbol{\Gamma}_{\boldsymbol{k},\boldsymbol{\ell}}=\mathscr{D}(\boldsymbol{k},\boldsymbol{\ell})[\boldsymbol{p}(\mathbf{1},\mathbf{1})]$. Now apply the operator $\mathscr{D}(\boldsymbol{k},\boldsymbol{\ell})$ to the differential equation in Proposition 4.1.1. Abbreviate $\boldsymbol{d}_{\boldsymbol{k},\boldsymbol{\ell}}\equiv\boldsymbol{d}_{\boldsymbol{k},\boldsymbol{\ell}}(\boldsymbol{w},\boldsymbol{z}):=\mathscr{D}(\boldsymbol{k},\boldsymbol{\ell})[\boldsymbol{p}(\boldsymbol{w},\boldsymbol{z})]$. We thus obtain

$$\boldsymbol{d}_{\boldsymbol{k},\boldsymbol{\ell}}Q+\sum_{m=1}^{d}(w_m-1)\left(\lambda_{1,m}\boldsymbol{d}_{\boldsymbol{k},\boldsymbol{\ell}}E_m+\mu_{1,m}\boldsymbol{d}_{\boldsymbol{k}+\boldsymbol{e}_m,\boldsymbol{\ell}}\right)$$
$$+\sum_{m=1}^{d}k_m\left(\lambda_{1,m}\boldsymbol{d}_{\boldsymbol{k}-\boldsymbol{e}_m,\boldsymbol{\ell}}E_m+\mu_{1,m}\boldsymbol{d}_{\boldsymbol{k},\boldsymbol{\ell}}\right)$$
$$+\sum_{m=1}^{d}(z_m-1)\left(\lambda_{2,m}\boldsymbol{d}_{\boldsymbol{k},\boldsymbol{\ell}}E_m+\mu_{2,m}\boldsymbol{d}_{\boldsymbol{k}+\boldsymbol{e}_m,\boldsymbol{\ell}}\right)$$
$$+\sum_{m=1}^{d}\ell_m\left(\lambda_{2,m}\boldsymbol{d}_{\boldsymbol{k},\boldsymbol{\ell}-\boldsymbol{e}_m}E_m+\mu_{2,m}\boldsymbol{d}_{\boldsymbol{k},\boldsymbol{\ell}}\right)=\mathbf{0}^{\mathrm{T}}.$$

Now plugging in $\boldsymbol{w}=\boldsymbol{z}=\mathbf{1}$ yields the relation

$$\boldsymbol{\Gamma}_{\boldsymbol{k},\boldsymbol{\ell}}Q+\sum_{m=1}^{d}k_m\left(\lambda_{1,m}\boldsymbol{\Gamma}_{\boldsymbol{k}-\boldsymbol{e}_m,\boldsymbol{\ell}}E_m+\mu_{1,m}\boldsymbol{\Gamma}_{\boldsymbol{k},\boldsymbol{\ell}}\right)$$
$$+\sum_{m=1}^{d}\ell_m\left(\lambda_{2,m}\boldsymbol{\Gamma}_{\boldsymbol{k},\boldsymbol{\ell}-\boldsymbol{e}_m}E_m+\mu_{2,m}\boldsymbol{\Gamma}_{\boldsymbol{k},\boldsymbol{\ell}}\right)=\mathbf{0}^{\mathrm{T}}.$$

Define $\Lambda_{i,m}:=\lambda_{i,m}\,\mathrm{diag}\{\boldsymbol{e}_m\}$ and $\mathcal{M}_{i,m}:=\mu_{i,m}I$. We obtain the following recursion.

Proposition 4.1.2. *The factorial moments* $\mathbf{\Gamma}_{\boldsymbol{k},\boldsymbol{\ell}}$ *satisfy the recursion*

$$\mathbf{\Gamma}_{\boldsymbol{k},\boldsymbol{\ell}} = \left(\sum_{m=1}^{d} k_m \mathbf{\Gamma}_{\boldsymbol{k}-\boldsymbol{e}_m,\boldsymbol{\ell}} \Lambda_{1,m} + \sum_{m=1}^{d} \ell_m \mathbf{\Gamma}_{\boldsymbol{k},\boldsymbol{\ell}-\boldsymbol{e}_m} \Lambda_{2,m} \right) \times \left(\sum_{m=1}^{d} k_m \mathcal{M}_{1,m} + \sum_{m=1}^{d} \ell_m \mathcal{M}_{2,m} - Q \right)^{-1},$$

to be initialized with $\mathbf{\Gamma}_{\mathbf{0},\mathbf{0}} = \boldsymbol{\pi}^{\mathrm{T}}$.

4.2. *Transient behavior*

We now shift our attention from the steady-state distribution to the corresponding transient behavior. As in Model I, the factorial moments can be found by a recursion, where in each step a non-homogeneous system of linear differential equations needs to be solved.

The following differential equation has been derived in a similar way as the other differential equations that we presented so far.

Proposition 4.2.1. *The pgf* $\boldsymbol{p}(t, \boldsymbol{w}, \boldsymbol{z})$ *satisfies the differential equation*

$$\boldsymbol{p}(t, \boldsymbol{w}, \boldsymbol{z})Q + \sum_{k=1}^{d} (w_k - 1) \left(\lambda_{1,k} \boldsymbol{p}(t, \boldsymbol{w}, \boldsymbol{z}) E_k + \mu_{1,k} \frac{\partial \boldsymbol{p}}{\partial w_k} \right) + \sum_{k=1}^{d} (z_k - 1) \left(\lambda_{2,k} \boldsymbol{p}(t, \boldsymbol{w}, \boldsymbol{z}) E_k + \mu_{2,k} \frac{\partial \boldsymbol{p}}{\partial z_k} \right) = \frac{\partial \boldsymbol{p}}{\partial t}.$$

The moments can be in principle derived in the same way as for Model I; it leads to a recursive scheme of inhomogeneous linear differential equations. There is a more compact alternative though, based on a different system of differential equations. Precisely as is done in Ref.[6] for the case of a single Markov-modulated infinite-server system, we can derive the following result. We define

$$[\bar{\boldsymbol{p}}(t, w, z)]_j := \mathbb{E}\left(w^{N_1(t)} z^{N_2(t)} \,\middle|\, J(0) = j \right),$$

which is now assumed to be a column vector. Define

$$\Delta(\boldsymbol{\mu}_i, t) := \mathrm{diag}\{e^{-\mu_{i,1}t}, \ldots, e^{-\mu_{i,d}t}\}.$$

Proposition 4.2.2. *The pgf* $\bar{\boldsymbol{p}}(t, w, z)$ *satisfies the differential equation*

$$Q\,\bar{\boldsymbol{p}}(t, w, z) + (w - 1)\Delta(\boldsymbol{\lambda}_1)\,\Delta(\boldsymbol{\mu}_1, t)\,\bar{\boldsymbol{p}}(t, w, z) + (z - 1)\Delta(\boldsymbol{\lambda}_2)\,\Delta(\boldsymbol{\mu}_2, t)\,\bar{\boldsymbol{p}}(t, w, z) = \frac{\partial \bar{\boldsymbol{p}}}{\partial t}.$$

Observe that this system of differential equations just implicitly provides us with information about the stationary behavior, as sending $t \to \infty$ yields $\mathbf{0} = \mathbf{0}$.

The column vector $\bar{\mathbf{\Gamma}}_{k,\ell}(t)$ is defined as

$$[\bar{\mathbf{\Gamma}}_{k,\ell}(t)]_j := \mathbb{E}\left((M_1(t))_k\,(M_2(t))_\ell \,|\, J(0)=j\right).$$

It takes a basic computation to verify the following recursion.

Proposition 4.2.3. *The factorial moments $\mathbf{\Gamma}_{k,\ell}(t)$ satisfy the recursion*

$$\begin{aligned}\bar{\mathbf{\Gamma}}'_{k,\ell}(t) &= Q\bar{\mathbf{\Gamma}}_{k,\ell}(t) + k\Delta(\boldsymbol{\lambda}_1)\Delta(\boldsymbol{\mu}_1,t)\bar{\mathbf{\Gamma}}_{k-1,\ell}(t)\\ &\quad +\ell\Delta(\boldsymbol{\lambda}_2)\Delta(\boldsymbol{\mu}_2,t)\bar{\mathbf{\Gamma}}_{k,\ell-1}(t), \qquad \bar{\mathbf{\Gamma}}_{k,\ell}(0)=\mathbf{0}^{\mathrm{T}},\end{aligned}$$

to be initialized with $\bar{\mathbf{\Gamma}}_{0,0}(t)=\mathbf{1}^{\mathrm{T}}$.

5. Model I: Explicit calculation of mean, variance, and covariance

In this section we further analyze the mean and variance of the (transient and stationary) numbers of customers in both infinite-server queues, as well as the covariance between them.

According to Proposition 3.2.2, the mean of $M_k(t)$ can be found by solving a non-homogeneous linear differential equation. With (row vector!)

$$\boldsymbol{m}_k(t) := \left(\mathbb{E}(M_k(t)1_{\{J(t)=1\}}),\ldots,\mathbb{E}(M_k(t)1_{\{J(t)=d\}})\right),$$

we are to solve

$$\boldsymbol{m}'_k(t) = \boldsymbol{m}(t)\ (Q-\Delta(\boldsymbol{\mu}_k)) + \boldsymbol{a}^{\mathrm{T}}P(t)\,\Delta(\boldsymbol{\lambda}_k).$$

This can be done by standard techniques; we do not include the explicit expression here. It is noted that we evidently have that $\mathbb{E}M_k(t)=\boldsymbol{m}_k(t)\mathbf{1}$. Using the resulting expression for the $\boldsymbol{m}_k(t)$, we can also identify, again using Proposition 3.2.2, $\mathbb{V}\mathrm{ar}\,M_k(t)$ and $\mathbb{C}\mathrm{ov}\,(M_1(t),M_2(t))$.

Stationarity. The expressions drastically simplify in stationarity. It is readily checked from Proposition 3.1.2 that, in accordance with the results of Ref.[14], for $k=1,2$,

$$\mathbb{E}M_k = \boldsymbol{\pi}^{\mathrm{T}}\Delta(\boldsymbol{\lambda}_k)(\Delta(\boldsymbol{\mu}_k)-Q)^{-1}\mathbf{1},$$

whereas

$$\mathbb{E}M_k(M_k-1) = 2\boldsymbol{\pi}^{\mathrm{T}}\Delta(\boldsymbol{\lambda}_k)(\Delta(\boldsymbol{\mu}_k)-Q)^{-1}\Delta(\boldsymbol{\lambda}_k)(2\Delta(\boldsymbol{\mu}_k)-Q)^{-1}\mathbf{1}.$$

The covariance $\mathbb{C}\mathrm{ov}\,(M_1,M_2)=\mathbb{E}M_1M_2-\mathbb{E}M_1\,\mathbb{E}M_2$ between the stationary number of jobs in both systems can be easily computed, too; realize that

$$\begin{aligned}\mathbb{E}M_1M_2 = \boldsymbol{\pi}^{\mathrm{T}}&\left(\Delta(\boldsymbol{\lambda}_2)(\Delta(\boldsymbol{\mu}_2)-Q)^{-1}\Delta(\boldsymbol{\lambda}_1)+\Delta(\boldsymbol{\lambda}_1)(\Delta(\boldsymbol{\mu}_1)-Q)^{-1}\Delta(\boldsymbol{\lambda}_2)\right)\\ &\times(\Delta(\boldsymbol{\mu}_1)+\Delta(\boldsymbol{\mu}_2)-Q)^{-1}\mathbf{1}.\end{aligned}$$

The formula for $\mathbb{C}\mathrm{ov}\,(M_1,M_2)$ further simplifies if $\Delta(\boldsymbol{\mu}_i)=m_iI$ (that is, for each of the two infinite-server queues there are uniform departure rates). To this end,

define the entries of the exponentially γ-weighted (for $\gamma > 0$) deviation matrix [10, Section 4] by

$$D_{ij}(\gamma) := \int_0^\infty e^{-\gamma v} \left(p_{ij}(v) - \pi_j\right) \mathrm{d}v,$$

and let $\check{D}_{ij}(\gamma) := D_{ij}(\gamma) + \pi_j/\gamma$. Integration by parts yields, for $\gamma > 0$,

$$\begin{aligned} Q\check{D}(\gamma) = \int_0^\infty QP(v)e^{-\gamma v}\mathrm{d}v = \int_0^\infty P'(v)e^{-\gamma v}\mathrm{d}v = -I \\ + \int_0^\infty \gamma P(v)e^{-\gamma v}\mathrm{d}v = -I + \gamma\check{D}(\gamma). \end{aligned}$$

As a consequence, $-(Q - \gamma I)\check{D}(\gamma) = I$, so that $(m_i I - Q)^{-1} = -\check{D}(m_i)$. In addition, for any α,

$$(\alpha I - Q)^{-1}\mathbf{1} = \frac{1}{\alpha}\sum_{i=0}^{\infty}\frac{1}{\alpha^i}Q^i\mathbf{1} = \frac{1}{\alpha}\mathbf{1}.$$

It is now concluded that

$$\begin{aligned} \mathbb{C}\mathrm{ov}\,(M_1, M_2) = -\frac{1}{m_1 + m_2}\boldsymbol{\pi}^{\mathrm{T}}\left(\Delta(\boldsymbol{\lambda}_2)\check{D}(m_2)\Delta(\boldsymbol{\lambda}_1) + \Delta(\boldsymbol{\lambda}_1)\check{D}(m_1)\Delta(\boldsymbol{\lambda}_2)\right)\mathbf{1} \\ - \left(\frac{\boldsymbol{\pi}^{\mathrm{T}}\Delta(\boldsymbol{\lambda}_1)\mathbf{1}}{m_1}\right)\left(\frac{\boldsymbol{\pi}^{\mathrm{T}}\Delta(\boldsymbol{\lambda}_2)\mathbf{1}}{m_2}\right). \end{aligned}$$

It requires elementary algebra to verify that this expression equals

$$\mathbb{C}\mathrm{ov}\,(M_1, M_2) = \frac{\boldsymbol{\pi}^{\mathrm{T}}\left(\Delta(\boldsymbol{\lambda}_2)D(m_2)\Delta(\boldsymbol{\lambda}_1) + \Delta(\boldsymbol{\lambda}_1)D(m_1)\Delta(\boldsymbol{\lambda}_2)\right)\mathbf{1}}{m_1 + m_2}. \tag{3}$$

Time scalings. Under a specific parameter scaling, the expressions for the transient mean and variance can be computed in closed form. We include these computations, as they directly relate to those that we use later when establishing central limit theorems.

We focus on the regime in which we speed up the background process by a factor N^f (for some $f > 0$), meaning that we replace Q by $N^f Q$, and at the same time the arrival rates by N, meaning that we replace $\boldsymbol{\lambda}_i$ by $N\boldsymbol{\lambda}_i$ for $i = 1, 2$. In this context, we write $M_k^{(N)}(t)$ rather than $M_k(t)$ to reflect the dependence on N; the background process becomes $J^{(N)}(\cdot)$. Below we work with

$$\left[\boldsymbol{m}_k^{(N)}(t)\right]_j := \frac{1}{N}\mathbb{E}\left(M_k^{(N)}(t)1_{\{J^{(N)}(t)=j\}}\right).$$

From Proposition 3.2.2, we immediately have

$$\left(\boldsymbol{m}_k^{(N)}\right)'(t) = \boldsymbol{m}_k^{(N)}(t)(N^f Q - \Delta(\boldsymbol{\mu}_k)) + \boldsymbol{a}^{\mathrm{T}}P(N^f t)\Delta(\boldsymbol{\lambda}_k).$$

Postmultiply the equation by the fundamental matrix F and N^{-f}, so as to obtain

$$\begin{aligned} \boldsymbol{m}_k^{(N)}(t) = \boldsymbol{m}_k^{(N)}(t)\Pi - (\boldsymbol{m}_k^{(N)})'(t)FN^{-f} - \boldsymbol{m}_k^{(N)}(t)\Delta(\boldsymbol{\mu}_k)FN^{-f} \\ + \boldsymbol{a}^{\mathrm{T}}P(N^f t)\Delta(\boldsymbol{\lambda}_k)\,FN^{-f}. \end{aligned}$$

Iterate this relation once, and realize that due to $\Pi = \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$, it follows that $\boldsymbol{m}_k^{(N)}(t)\Pi = \bar{m}_k^{(N)}(t)\,\boldsymbol{\pi}^{\mathrm{T}}$ for some (single-dimensional) function $\bar{m}_k^{(N)}(\cdot)$. We thus obtain

$$\left(\bar{m}_k^{(N)}\right)'(t)\boldsymbol{\pi}^{\mathrm{T}}N^{-f} = -\bar{m}_k^{(N)}(t)\boldsymbol{\pi}^{\mathrm{T}}\,\Delta(\boldsymbol{\mu}_k)FN^{-f} + \boldsymbol{a}^{\mathrm{T}}P\left(N^f t\right)\,\Delta\left(\boldsymbol{\lambda}_k\right)FN^{-f} + o\left(N^{-f}\right),$$

where it is also used that $\Pi F = \Pi$. Now postmultiply by $\mathbf{1}N^f$, recalling that $F\mathbf{1} = \mathbf{1}$, and observing that $\boldsymbol{a}^{\mathrm{T}}P(N^f t) \to \boldsymbol{\pi}^{\mathrm{T}}$, we arrive when sending $N \to \infty$ at the differential equation

$$\bar{m}_k'(t) = -\bar{m}_k(t)\mu_{k,\infty} + \lambda_{k,\infty},$$

with $\bar{m}_k(t)$ defined as $\lim_{N\to\infty} m_k^{(N)}(t)$. This trivial differential equation is evidently solved by $\bar{m}_k(t) = (\lambda_{k,\infty}/\mu_{k,\infty})\,(1 - e^{-\mu_{k,\infty}t})$. We conclude that

$$\lim_{n\to\infty} \frac{\mathbb{E}M_k^{(N)}(t)}{N} = \varrho_k^{(\mathrm{I})}(t) := \frac{\lambda_{k,\infty}}{\mu_{k,\infty}}\,(1 - e^{-\mu_{k,\infty}t}).$$

Essentially the same procedure can be followed to determine the asymptotics of the variances and covariances related to the $M_K^{(N)}(t)$. After considerable algebra (which is left out here), it eventually turns out that, with $\beta := \max\{1/2,\, 1 - f/2\}$, as $N \to \infty$,

$$\frac{1}{N^{2\beta}}\begin{pmatrix} \mathbb{V}\mathrm{ar}\, M_1^{(N)}(t) & \mathbb{C}\mathrm{ov}\left(M_1^{(N)}(t), M_2^{(N)}(t)\right) \\ \mathbb{C}\mathrm{ov}\left(M_1^{(N)}(t), M_2^{(N)}(t)\right) & \mathbb{V}\mathrm{ar}\, M_2^{(N)}(t) \end{pmatrix} \to \Sigma^{(\mathrm{I})}(t),$$

with the covariance matrix $\Sigma^{(\mathrm{I})}(t)$ to be defined in (8). From the form of $\Sigma^{(\mathrm{I})}(t)$, as given in (8), we observe that the system behaves crucially different for $f > 1$ and $f < 1$:

- For $f > 1$, we have $\beta = \frac{1}{2}$: the variances grow essentially linearly, but the covariance sublinearly. This reflects that, when the background process jumps at a faster timescale than the arrival processes, the individual queues roughly behave as two independent M/M/∞ systems. It suggests that in the CLT we have to normalize by the usual $\sqrt{N}$.
- For $f < 1$, on the other hand, all entries of the covariance matrix grow like N^{2-f}, that is, superlinearly. As a consequence, in this scaling the two queues behave dependently, and in the CLT a normalization by $N^{1-f/2}$ is anticipated.

In the next section it is shown that the variances and covariances in Model II have the same qualitative behavior. It is this dichotomy that plays an important role in the central limit theorems that we derive later in this paper.

6. Model II: Explicit calculation of mean, variance, and covariance

For Model II, the mean and variance of the numbers of customers have been explicitly found in Ref.[6]. In this section, we show that, with computations resembling those featuring in Ref.[6], one can also find the covariance between the numbers of jobs present in both systems. The underlying type of reasoning heavily relies on

the representation of the number of customers present as a Poisson random variable with stochastic parameter, as observed in Ref.[11]. The reasoning behind it, however, provides intuition as to why deviation matrices appear in variances and covariances under certain scalings, and that is why we have chosen to include these computations here.

For ease, we assume the background process starts off in equilibrium at time 0, but it can be verified that this is not necessary. In Ref.[6] it was observed that, with $J \equiv (J(s)\colon s \in [0, t])$,

$$(\mathbb{E}(M_1(t) \mid J) = \int_0^t \lambda_{k,J(s)} e^{-\mu_{k,J(s)}(t-s)} \mathrm{d}s.$$

In line with what was found in Ref.[6], the mean $\mathbb{E}M_k(t)$ is therefore given by, for $k = 1, 2$,

$$\varrho_k^{(\mathrm{II})}(t) := \mathbb{E}M_k(t) = \sum_{i=1}^{d} \pi_i \frac{\lambda_{k,i}}{\mu_{k,i}} \left(1 - e^{-\mu_{k,i}t}\right).$$

Now focus on the evaluation of $\mathbb{C}\mathrm{ov}\,(M_1(t), M_2(t))$. The law of total covariance entails that

$$\mathbb{C}\mathrm{ov}\,(M_1(t), M_2(t)) = \mathbb{E}(\mathbb{C}\mathrm{ov}((M_1(t), M_2(t))|J)) + \mathbb{C}\mathrm{ov}(\mathbb{E}(M_1(t)|J), \mathbb{E}(M_2(t)|J)).$$

The first of these terms cancels: given the path of J, there is no systematic effect of the $M_i(t)$ on each other. Plugging in expressions we found earlier for $\mathbb{E}(M_i(t) \mid J)$, the second term equals

$$\mathbb{C}\mathrm{ov}\left(\int_0^t \lambda_{1,J(s)} e^{-\mu_{1,J(s)}(t-s)} \mathrm{d}s, \int_0^t \lambda_{2,J(s)} e^{-\mu_{2,J(s)}(t-s)} \mathrm{d}s\right),$$

which can be rewritten as

$$\int_0^t \int_0^t \mathbb{C}\mathrm{ov}\,\left(\lambda_{1,J(r)} e^{-\mu_{1,J(r)}(t-r)}, \lambda_{2,J(s)} e^{-\mu_{2,J(s)}(t-s)}\right) \mathrm{d}r\, \mathrm{d}s.$$

Now we split the double integral into the cases $r < s$ and $r \geq s$. The contribution of the first of these two cases is

$$\begin{aligned}
&\int_0^t \int_0^s \sum_{i=1}^{d} \sum_{j=1}^{d} \lambda_{1,i}\lambda_{2,j} e^{-\mu_{1,i}(t-r)} e^{-\mu_{2,j}(t-s)}\, \mathbb{C}\mathrm{ov}\,\left(1_{\{J(r)=i\}}, 1_{\{J(s)=j\}}\right) \mathrm{d}r\, \mathrm{d}s \\
&\quad = \int_0^t \int_0^s \sum_{i=1}^{d} \sum_{j=1}^{d} \lambda_{1,i}\lambda_{2,j} e^{-\mu_{1,i}(t-r)} e^{-\mu_{2,j}(t-s)} \pi_i \left(p_{ij}(s-r) - \pi_j\right) \mathrm{d}r\, \mathrm{d}s.
\end{aligned}$$

Using elementary algebra (put $v := s - r$ and interchange the order of the integrals), we find that this equals

$$\sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\lambda_{1,i}\lambda_{2,j}}{\mu_{1,i} + \mu_{2,j}} \int_0^t \left(e^{-\mu_{1,i}v} - e^{-(\mu_{1,i}+\mu_{2,j})t + \mu_{2,j}v}\right) \pi_i \left(p_{ij}(v) - \pi_j\right) \mathrm{d}v. \qquad (4)$$

It is verified that the contribution due the other case ($r \geq s$, that is) equals (4), but with the roles of the two processes interchanged. We thus end up with the following result:

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(M_1(t), M_2(t)\right) = &\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\lambda_{1,i}\lambda_{2,j}}{\mu_{1,i}+\mu_{2,j}}\int_0^t \left(e^{-\mu_{1,i}\upsilon} - e^{-(\mu_{1,i}+\mu_{2,j})t+\mu_{2,j}\upsilon}\right)\pi_i \\ &\times \left(p_{ij}(\upsilon)-\pi_j\right)\mathrm{d}\upsilon \\ &+\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\lambda_{1,j}\lambda_{2,i}}{\mu_{1,j}+\mu_{2,i}}\int_0^t \left(e^{-\mu_{2,j}\upsilon} - e^{-(\mu_{2,j}+\mu_{1,i})t+\mu_{1,j}\upsilon}\right)\pi_i \\ &\times \left(p_{ij}(\upsilon)-\pi_j\right)\mathrm{d}\upsilon,\end{aligned}$$

which simplifies to

$$\begin{aligned}\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\lambda_{1,i}\lambda_{2,j}}{\mu_{1,i}+\mu_{2,j}}\int_0^t &\left(e^{-\mu_{1,i}\upsilon} - e^{-(\mu_{1,i}+\mu_{2,j})t+\mu_{2,j}\upsilon}\right) \\ &\times \left(\pi_i\left(p_{ij}(\upsilon)-\pi_j\right)+\pi_j\left(p_{ji}(\upsilon)-\pi_i\right)\right)\mathrm{d}\upsilon.\end{aligned}$$

As mentioned above, in Ref.[6] an expression for the variance of the transient distribution was already established: relying on the law of total variance it is found that, for $k = 1, 2$,

$$\begin{aligned}\mathbb{V}\mathrm{ar}\, M_k(t) = \varrho_k(t) + 2\sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\lambda_{k,i}\lambda_{k,j}}{\mu_{k,i}+\mu_{k,j}}\int_0^t &\left(e^{-\mu_{k,j}\upsilon} - e^{-2\mu_{k,j}t+\mu_{k,j}\upsilon}\right)\pi_i \\ &\times \left(p_{ij}(\upsilon)-\pi_j\right)\mathrm{d}\upsilon.\end{aligned}$$

As we did in the previous section, we now consider a few special cases that provide us with interesting insights. In the first special case we let t grow large, while in the second special case we scale the arrival rates and the transition rates of the background process in a particular manner.

Stationarity. In stationarity we obtain

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(M_1, M_2\right) = \sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\lambda_{1,i}\lambda_{2,j}}{\mu_{1,i}+\mu_{2,j}}\int_0^\infty &\left(e^{-\mu_{1,i}\upsilon}\pi_i\left(p_{ij}(\upsilon)-\pi_j\right)\right. \\ &\left.+e^{-\mu_{2,j}\upsilon}\pi_j\left(p_{ji}(\upsilon)-\pi_i\right)\right)\mathrm{d}\upsilon.\end{aligned}$$

Recalling the definition of the γ-weighted deviation matrix, we obtain the appealing expression

$$\mathbb{C}\mathrm{ov}\left(M_1, M_2\right) = \sum_{i=1}^{d}\sum_{j=1}^{d}\frac{\lambda_{1,i}\lambda_{2,j}}{\mu_{1,i}+\mu_{2,j}}\left(\pi_i D_{ij}(\mu_{1,i}) + \pi_j D_{ji}(\mu_{2,j})\right),$$

whereas, for $k = 1, 2$,

$$\mathbb{V}\mathrm{ar}\, M_k = \sum_{i=1}^{d} \pi_i \frac{\lambda_{k,i}}{\mu_{k,i}} + 2 \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\lambda_{k,i}\lambda_{k,j}}{\mu_{k,i} + \mu_{k,j}} \pi_i D_{ij}(\mu_{k,j}).$$

It takes a short, direct computation to verify that the expression for $\mathbb{C}\mathrm{ov}\,(M_1, M_2)$ coincides with (3) in case $\Delta(\boldsymbol{\mu}_i) = m_i I$.

Time scalings. We again consider the regime in which we speed up the background process by a factor N^f (for some $f > 0$), meaning that we replace Q by $N^f Q$, and the arrival rates by N, meaning that we replace $\boldsymbol{\lambda}_i$ by $N\boldsymbol{\lambda}_i$ for $i = 1, 2$; as before, we write $M_k^{(N)}(t)$ rather than $M_k(t)$. It is readily verified that, with $D := D(0)$ the (ordinary, non-weighted) deviation matrix, for $k = 1, 2$,

$$\mathbb{V}\mathrm{ar}\, M_k^{(N)}(t) := N\varrho_k^{(\mathrm{II})}(t) + N^{2-f} v_k^{(\mathrm{II})}(t),$$

with $\varrho_k(t)$ as before, and

$$v_k^{(\mathrm{II})}(t) := 2 \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\lambda_{k,i}\lambda_{k,j}}{\mu_{k,i} + \mu_{k,j}} \left(1 - e^{-(\mu_{k,i}+\mu_{k,j})t}\right) \pi_i D_{ij}, \tag{5}$$

whereas the covariance equals

$$\mathbb{C}\mathrm{ov}\left(M_1^{(N)}(t), M_2^{(N)}(t)\right) = N^{2-f} c^{(\mathrm{II})}(t)$$

with

$$c^{(\mathrm{II})}(t) := \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\lambda_{1,i}\lambda_{2,j}}{\mu_{1,i} + \mu_{2,j}} \left(1 - e^{-(\mu_{1,i}+\mu_{2,j})t}\right) \left(\pi_i D_{ij} + \pi_j D_{ji}\right). \tag{6}$$

Just like we have seen in Model I, for $f > 1$ the variances grow linearly, while the covariance behaves sublinearly. As a consequence the two processes effectively decouple; it is therefore expected that in the CLT we need to normalize by the usual $\sqrt{N}$. For $f < 1$, on the contrary, the entire covariance matrix behaves as N^{2-f}, so that it is anticipated that in the CLT we have to scale by $N^{1-f/2}$. In the next sections we study CLT results for both models.

7. Model I: Central limit theorem

In this and the next section, our aim is to derive a CLT under the scaling of the transition rate matrix and arrival rates that we have considered earlier in this paper, that is, $Q \mapsto N^f Q$, $\boldsymbol{\lambda}_i \mapsto N\boldsymbol{\lambda}_i$. As before, we add the superscript $^{(N)}$ to the random variables $M_i(t)$ and M_i, to express the dependence of these objects on the scaling.

In principle, we could analyze CLT s for all four variants discussed earlier in this paper: Model I and II, and stationary and transient regimes. Such an analysis, however, by and large follows the approach carried out in Ref.[4] for the case of a *single* (non-coupled, that is) infinite-server system with Markov-modulated input, and also the results strongly resemble those presented in Ref.[4]. To prove the CLT s, in

Ref.[4] the "single-system counterparts" of Propositions 3.1.1, 3.2.1, 4.1.1, and 4.2.1 are intensively relied on.

Motivated by the above considerations, we present in this section and the next section the full analyses for just the transient cases of both models. More precisely, the contents of these sections are:

- In this section we treat Model I with a derivation that mimics the one used to analyze the single-system counterpart in Ref.[4]; as it turns out, the stationary result follows directly from the transient result.
- The next section gives a detailed analysis of the transient of Model II but relies on the characterization of the pgf featuring in Proposition 4.2.2 instead of the one appearing in Proposition 4.2.1; this means that the type of argumentation used now has not been presented in Ref.[4]. The choice of relying on Proposition 4.2.2, instead of Proposition 4.2.1, has the advantage that we have to deal with a system of *ordinary* differential equations (with respect to time), rather than a system of *partial* differential equations, which makes the analysis slightly easier. Formally, the CLT for the stationary number of jobs in the system for Model II does not follow directly from the transient result; it is pointed out how the stationary result should be rigorously derived (and this stationary result is also stated).

The procedure, as followed in this and the next section, can be summarized as follows. In the CLT s it is established that a centered and scaled (or normalized) version of $(M_1^{(N)}(t), M_2^{(N)}(t))$ converges to a bivariate Normally distributed random variable. The first step is to use the systems of (partial) differential equations, as presented in Sections 3 and 4, that relate to the non-centered and non-scaled model, to set up the corresponding differential equations for the centered and scaled model, under the scaling under consideration. Then Taylor approximations are used to study their behavior for large N. The resulting (single-dimensional) differential equation can be solved and yields the claimed Normality. After having established the claim for the transient distribution, we can also identify its stationary counterpart.

Importantly, the CLT s featuring in this and the next section are nonstandard in the sense that the normalization imposed is not necessarily the classical $\sqrt{N}$ scaling: if $f > 1$, then we should indeed use $\sqrt{N}$, but if $f < 1$, we have to scale by $N^{1-f/2}$, as indicated earlier.

7.1. *Model I, transient case*

In the CLT setting it is more convenient to work with moment-generating functions (mgf s) rather than probability-generating functions. For that reason, introduce the bivariate mgf $\check{\boldsymbol{p}}(t, \boldsymbol{\vartheta})$, with $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)^{\mathrm{T}}$. It is an elementary exercise that the partial differential equation in Proposition 3.2.1 translates into

$$\check{\boldsymbol{p}}(t, \boldsymbol{\vartheta})\, Q + \sum_{j=1}^{2} \left((e^{\vartheta_j} - 1) \check{\boldsymbol{p}}(t, \boldsymbol{\vartheta})\, \Delta(\boldsymbol{\lambda}_j) - (1 - e^{-\vartheta_j}) \frac{\partial \check{\boldsymbol{p}}(t, \boldsymbol{\vartheta})}{\partial \vartheta_j} \Delta(\boldsymbol{\mu}_j) \right) = \frac{\partial \check{\boldsymbol{p}}(t, \boldsymbol{\vartheta})}{\partial t}.$$

The scaling amounts to replacing Q by $N^f Q$ and $\Delta(\boldsymbol{\lambda}_j)$ by $N\Delta(\boldsymbol{\lambda}_j)$; to stress the dependence of the mgf on the scaling parameter N, we write $\check{\boldsymbol{p}}^{(N)}(t, \boldsymbol{\vartheta})$ rather than $\check{\boldsymbol{p}}(t, \boldsymbol{\vartheta})$.

Recall that $\varrho_j^{(\mathrm{I})}(t) = \varrho_j^{(\mathrm{I})} \cdot (1 - e^{-\mu_{j,\infty} t})$ with $\varrho_j^{(\mathrm{I})} := \lambda_{j,\infty}/\mu_{j,\infty}$, and consider the random variable, with $\beta := \max\{1/2, 1 - f/2\}$,

$$\vartheta_1 \left(\frac{M_1^{(N)}(t) - N\varrho_1^{(\mathrm{II})}(t)}{N^\beta} \right) + \vartheta_2 \left(\frac{M_2^{(N)}(t) - N\varrho_2^{(\mathrm{II})}(t)}{N^\beta} \right), \tag{7}$$

with mgf $g^{(N)}(t, \boldsymbol{\vartheta})$ (jointly with the event $J^{(N)}(t) = i$, for $i = 1, \ldots, d$, so that $g^{(N)}(t, \boldsymbol{\vartheta})$ is a d-dimensional row vector). It is readily verified that

$$\begin{aligned}
\frac{\partial \boldsymbol{g}^{(N)}(t, \boldsymbol{\vartheta})}{\partial t} &= \frac{\partial \check{\boldsymbol{p}}^{(N)}(t, \boldsymbol{\vartheta}/N^\beta)}{\partial t} \exp\left(-\sum_{j=1}^{2} \vartheta_j \varrho_j^{(\mathrm{I})}(t) \right) \\
&\quad - \boldsymbol{g}^{(N)}(t, \boldsymbol{\vartheta}) N^{1-\beta} \sum_{j=1}^{2} \vartheta_j (\varrho_j^{(\mathrm{I})})'(t), \\
\frac{\partial \boldsymbol{g}^{(N)}(t, \boldsymbol{\vartheta})}{\partial \vartheta_i} &= N^{-\beta} \frac{\partial \check{\boldsymbol{p}}^{(N)}(t, \boldsymbol{\vartheta}/N^\beta)}{\partial \vartheta_i} \exp\left(-\sum_{j=1}^{2} \vartheta_j \varrho_j^{(\mathrm{I})}(t) \right) \\
&\quad - \boldsymbol{g}^{(N)}(t, \boldsymbol{\vartheta}) N^{1-\beta} \varrho_i^{(\mathrm{I})}(t).
\end{aligned}$$

We thus arrive at, suppressing the arguments of $\boldsymbol{g}^{(N)}(t, \boldsymbol{\vartheta})$,

$$\begin{aligned}
&\sum_{j=1}^{2} \left(N \left(e^{\vartheta_j/N^\beta} - 1 \right) \boldsymbol{g}^{(N)} \, \Delta(\boldsymbol{\lambda}_j) - \left(1 - e^{-\vartheta_j/N^\beta} \right) \right. \\
&\qquad \left. \times \left(N^\beta \frac{\partial \boldsymbol{g}^{(N)}}{\partial \vartheta_j} + N \boldsymbol{g}^{(N)} \varrho_j^{(\mathrm{I})}(t) \right) \Delta(\boldsymbol{\mu}_j) \right) \\
&\quad = \frac{\partial \boldsymbol{g}^{(N)}}{\partial t} + N^{1-\beta} \boldsymbol{g}^{(N)} \sum_{j=1}^{2} \vartheta_j (\varrho_j^{(\mathrm{I})})'(t) - \boldsymbol{g}^{(N)} Q N^f.
\end{aligned}$$

Now replace the exponential functions by the first two terms of their Taylor expansions, and postmultiply with F, to obtain

$$\begin{aligned}
\boldsymbol{g}^{(N)} &= \boldsymbol{g}^{(N)} \Pi - N^{-f} \frac{\partial \boldsymbol{g}^{(N)}}{\partial t} F - N^{1-f-\beta} \boldsymbol{g}^{(N)} F \cdot \sum_{j=1}^{2} \vartheta_j (\varrho_j^{(\mathrm{I})})'(t) \\
&\quad + N^{-f} \sum_{j=1}^{2} \left(N \left(\frac{\vartheta_j}{N^\beta} + \frac{\vartheta_j^2}{2N^{2\beta}} \right) \boldsymbol{g}^{(N)} \Delta(\boldsymbol{\lambda}_j) \right. \\
&\quad \left. - \left(\frac{\vartheta_j}{N^\beta} - \frac{\vartheta_j^2}{2N^{2\beta}} \right) \left(N^\beta \frac{\partial \boldsymbol{g}^{(N)}}{\partial \vartheta_j} + N \boldsymbol{g}^{(N)} \varrho_j^{(\mathrm{I})}(t) \right) \Delta(\boldsymbol{\mu}_j) \right) F + o(N^{1-f-2\beta}).
\end{aligned}$$

Now the next steps (which resemble those that will be used when analyzing the CLT for Model II) are: first we iterate this equation, and then postmultiply by $\mathbf{1} \cdot N^f$, leading to four relevant terms, viz. of orders 1, $N^{1-\beta}$, $N^{2-f-2\beta}$, and $N^{1-2\beta}$. Let $h^{(N)}$ denote $\boldsymbol{g}^{(N)}\mathbf{1}$, so that $\boldsymbol{g}^{(N)}\Pi = h^{(N)} \cdot \boldsymbol{\pi}^{\mathrm{T}}$. The term of order 1 is (use, e.g., $F\mathbf{1} = \mathbf{1}$)

$$-\frac{\partial h^{(N)}}{\partial t} - \sum_{j=1}^{2} \vartheta_j \frac{\partial h^{(N)}}{\partial \vartheta_j} \mu_{j,\infty}.$$

The term of order $N^{1-\beta}$ cancels, due to

$$\begin{aligned} &\boldsymbol{g}^{(N)}\Pi \left(\Delta(\boldsymbol{\lambda}_j)\mathbf{1} - \mathbf{1} \cdot (\varrho_j^{(\mathrm{I})})'(t) - \Delta(\boldsymbol{\mu}_j) F\mathbf{1} \cdot \varrho_j^{(\mathrm{I})}(t) \right) \\ &\quad = h^{(N)} \left(\lambda_{j,\infty} - (\varrho_j^{(\mathrm{I})})'(t) - \varrho_j^{(\mathrm{I})}(t)\mu_{j,\infty} \right) = 0. \end{aligned}$$

The term of order $N^{2-f-2\beta}$ has the form $h^{(N)}(t, \boldsymbol{\vartheta}) \cdot k(t, \boldsymbol{\vartheta})$, with

$$k(t, \boldsymbol{\vartheta}) := \boldsymbol{\pi}^{\mathrm{T}} \left(\sum_{j=1}^{2} \vartheta_j A_j(t) \right) F \left(\sum_{j=1}^{2} \vartheta_j A_j(t) \right) \mathbf{1},$$

where $A_j(t) := -(\varrho_j^{(\mathrm{I})})'(t) I + \Delta(\boldsymbol{\lambda}_j) - \varrho_j^{(\mathrm{I})}(t)\Delta(\boldsymbol{\mu}_j)$. A simplification can be made: using, e.g., $F = \Pi + D$ and $\boldsymbol{\pi}^{\mathrm{T}} D = \mathbf{0}^{\mathrm{T}}$, it is straightforward to conclude that

$$k(t, \boldsymbol{\vartheta}) := \boldsymbol{\pi}^{\mathrm{T}} \left(\sum_{j=1}^{2} \vartheta_j B_j(t) \right) D \left(\sum_{j=1}^{2} \vartheta_j B_j(t) \right) \mathbf{1},$$

where $B_j(t) := \Delta(\boldsymbol{\lambda}_j) - \varrho_j^{(\mathrm{I})}(t)\Delta(\boldsymbol{\mu}_j)$. Finally, the term of order $N^{1-2\beta}$ equals $h^{(N)}(t, \boldsymbol{\vartheta}) \cdot \ell(t, \boldsymbol{\vartheta})$, with

$$\ell(t, \boldsymbol{\vartheta}) := \sum_{j=1}^{2} \vartheta_j^2 \lambda_{j,\infty} \left(1 - \frac{1}{2} e^{-\mu_{j,\infty} t} \right).$$

We obtain the limiting partial differential equation (as $N \to \infty$)

$$\frac{\partial h(t, \boldsymbol{\vartheta})}{\partial t} + \sum_{j=1}^{2} \vartheta_j \frac{\partial h(t, \boldsymbol{\vartheta})}{\partial \vartheta_j} \mu_{j,\infty} = h(t, \boldsymbol{\vartheta}) \cdot \left(k(t, \boldsymbol{\vartheta}) 1_{\{f \leqslant 1\}} + \ell(t, \boldsymbol{\vartheta}) 1_{\{f \geqslant 1\}} \right).$$

Now two cases need to be distinguished: $f > 1$ and $f < 1$ (with $f = 1$ corresponding to a boundary case that needs to be handled separately).

- Now try for $f \leq 1$ the solution $h_+(t, \boldsymbol{\vartheta}) = \exp(\vartheta_1^2 v_1^{(\mathrm{I})}(t)/2 + \vartheta_1 \vartheta_2 c^{(\mathrm{I})}(t) + \vartheta^2 v_2^{(\mathrm{I})}(t)/2)$. After straightforward calculus, we obtain that, for $k = 1, 2$,

$$\begin{aligned} v_k^{(\mathrm{I})}(t) &= 2\boldsymbol{\pi}^{\mathrm{T}} \left(\int_0^t e^{-2\mu_{k,\infty}(t-s)} B_k(s) D B_k(s) \, \mathrm{d}s \right) \mathbf{1}, \\ c^{(\mathrm{I})}(t) &= \boldsymbol{\pi}^{\mathrm{T}} \left(\int_0^t e^{-(\mu_{1,\infty}+\mu_{2,\infty})(t-s)} \left(B_1(s) D B_2(s) + B_2(s) D B_1(s) \right) \, \mathrm{d}s \right) \mathbf{1}. \end{aligned}$$

- The case $f \geq 1$ is solved analogously (and obviously does not have a cross term):

$$h_+(t) := \exp\left(\frac{1}{2}\left(\varrho_1^{(\mathrm{I})}(t)\vartheta_1^2 + \varrho_2^{(\mathrm{I})}(t)\vartheta_2^2\right)\right).$$

- In case $f = 1$, it is seen that both terms should be taken into account; we thus find $h(t) = h_-(t) + h_+(t)$.

Define

$$\sum^{(\mathrm{I})}(t) := \begin{pmatrix} v_1^{(\mathrm{I})}(t) & c^{(\mathrm{I})}(t) \\ c^{(\mathrm{I})}(t) & v_2^{(\mathrm{I})}(t) \end{pmatrix} 1_{\{f \leqslant 1\}} + \begin{pmatrix} \varrho_1^{(\mathrm{I})}(t) & 0 \\ 0 & \varrho_2^{(\mathrm{I})}(t) \end{pmatrix} 1_{\{f \geqslant 1\}}. \tag{8}$$

Theorem 7.1.1. *Consider Model I. For any $t \geq 0$, the random variable*

$$\left(\frac{M_1^{(N)}(t) - N\varrho_1^{(\mathrm{I})}(t)}{N^\beta}, \frac{M_2^{(N)}(t) - N\varrho_2^{(\mathrm{I})}(t)}{N^\beta}\right)$$

converges to a bivariate Normal distribution with mean $\mathbf{0}$ *and covariance matrix* $\Sigma^{(\mathrm{I})}(t)$ *as* $N \to \infty$.

7.2. Model I, stationary case

Recall $\varrho_k^{(\mathrm{I})} = \lim_{t\to\infty} \varrho_k^{(\mathrm{I})}(t) = \lambda_{k,\infty}/\mu_{k,\infty}$. In addition, we introduce the notation $\Sigma^{(\mathrm{I})} := \lim_{t\to\infty} \Sigma^{(\mathrm{I})}(t)$; it takes a bit of calculus to verify that

$$\Sigma^{(\mathrm{I})} := \begin{pmatrix} v_1^{(\mathrm{I})} & c^{(\mathrm{I})} \\ c^{(\mathrm{I})} & v_2^{(\mathrm{I})} \end{pmatrix} 1_{\{f \leqslant 1\}} + \begin{pmatrix} \varrho_1^{(\mathrm{I})} & 0 \\ 0 & \varrho_2^{(\mathrm{I})} \end{pmatrix} 1_{\{f \geqslant 1\}},$$

with $B_j := \Delta(\boldsymbol{\lambda}_j) - \varrho_j^{(\mathrm{I})}\Delta(\boldsymbol{\mu}_j)$ and, for $k = 1, 2$,

$$v_k^{(\mathrm{I})} := \frac{1}{\mu_{k,\infty}} \cdot \boldsymbol{\pi}^{\mathrm{T}} B_k D B_k \mathbf{1}, \qquad c^{(\mathrm{I})} := \frac{1}{\mu_{1,\infty} + \mu_{2,\infty}} \cdot \boldsymbol{\pi}^{\mathrm{T}} \left(B_1 D B_2 + B_2 D B_1\right) \mathbf{1}.$$

The following result is shown just like Theorem 7.1, ignoring in the proof the partial derivative with respect to time.

Theorem 7.2.1. *Consider Model I. The random variable*

$$\left(\frac{M_1^{(N)} - N\varrho_1^{(\mathrm{I})}}{N^\beta}, \frac{M_2^{(N)} - N\varrho_2^{(\mathrm{I})}}{N^\beta}\right)$$

converges to a bivariate Normal distribution with mean $\mathbf{0}$ *and covariance matrix* $\Sigma^{(\mathrm{I})}$ *as* $N \to \infty$.

8. Model II: Central limit theorem

In this section the CLT s for Model II are established. The first subsection treats the transient case and relies on the system of (ordinary) differential equations presented in Proposition 4.2.2. In the second subsection it is pointed out how the corresponding stationary CLT can be found.

8.1. *Model II, transient case*

To derive the CLT, we are to analyze the limiting behavior ($N \to \infty$) of the random variable, with again $\beta := \max\{1/2, 1 - f/2\}$,

$$\vartheta_1 \left(\frac{M_1^{(N)}(t) - N\varrho_1^{(\mathrm{II})}(t)}{N^\beta} \right) + \vartheta_2 \left(\frac{M_2^{(N)}(t) - N\varrho_2^{(\mathrm{II})}(t)}{N^\beta} \right), \tag{9}$$

conditional on the background process starting in state $i \in \{1, \ldots, d\}$. This random variable has moment the generating function (being a d-dimensional column vector — the values of ϑ_1 and ϑ_2 are held fixed throughout this derivation, and therefore suppressed)

$$\boldsymbol{g}^{(N)}(t) = \bar{\boldsymbol{p}}\left(t, e^{\vartheta_1/N^\beta}, e^{\vartheta_2/N^\beta}\right) \exp\left(-N^{1-\beta}\vartheta_1\varrho_1^{(\mathrm{II})}(t) - N^{1-\beta}\vartheta_2\varrho_2^{(\mathrm{II})}(t)\right);$$

here the pgf $\bar{\boldsymbol{p}}$ is the one featuring in Proposition 4.2.2. A straightforward application of the chain rule yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}^{(N)}(t) = &\left(\frac{\mathrm{d}}{\mathrm{d}t}\,\bar{\boldsymbol{p}}\left(t, e^{\vartheta_1/N^\beta}, e^{\vartheta_2/N^\beta}\right)\right) \exp\left(-N^{1-\beta}\vartheta_1\varrho_1^{(\mathrm{II})}(t) - N^{1-\beta}\vartheta_2\varrho_2^{(\mathrm{II})}(t)\right)\\ &- \left(N^{1-\beta}\vartheta_1(\varrho_1^{(\mathrm{II})})'(t) + N^{1-\beta}\vartheta_2(\varrho_2^{(\mathrm{II})})'(t)\right)\boldsymbol{g}^{(N)}(t).\end{aligned}$$

Define

$$\Delta_{j,t} := \mathrm{diag}\left\{\lambda_{j,1}e^{-\mu_{j,1}t}, \ldots, \lambda_{j,d}e^{-\mu_{j,d}t}\right\}.$$

Now take the differential equation for the pgf from Proposition 4.2.2., apply the scaling introduced above, and rewrite the resulting equation in terms of the moment generating function $\boldsymbol{g}^{(N)}(t)$, to obtain

$$N^f Q\boldsymbol{g}^{(N)}(t) + \sum_{j=1}^{2}\left(N(e^{\vartheta_j/N^\beta} - 1)\Delta_{j,t} - N^{1-\beta}\vartheta_j(\varrho_j^{(\mathrm{II})})'(t)\right)\boldsymbol{g}^{(N)}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}^{(N)}(t).$$

Let D be the deviation matrix introduced earlier, and F the corresponding fundamental matrix, defined through $F := D + \Pi$, with $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$. Now premultiply the above differential equation by $N^{-f}F$; recall the standard property of the fundamental matrix[10] that $FQ = QF = \Pi - I$. In addition, we define

$$\Delta_{j,t} := \mathrm{diag}\left\{\lambda_{j,1}e^{-\mu_{j,1}t}, \ldots, \lambda_{j,d}e^{-\mu_{j,d}t}\right\}.$$

Using a Taylor expansion, the resulting differential equation can be rewritten as

$$\begin{aligned}\boldsymbol{g}^{(N)}(t) = \Pi\boldsymbol{g}^{(N)}(t) &+ N^{1-f-\beta}F\left(\sum_{j=1}^{2}\vartheta_j\left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t)\right)\right)\boldsymbol{g}^{(N)}(t)\\ &+ N^{1-f-2\beta}F\left(\sum_{j=1}^{2}\frac{\vartheta_j^2}{2}\Delta_{j,t}\right)\boldsymbol{g}^{(N)}(t) - N^{-f}\,F\,\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}^{(N)}(t) + o(N^{1-f-2\beta}).\end{aligned}$$

Iterating this relation, we obtain

$$\begin{aligned}\boldsymbol{g}^{(N)}(t) &= \Pi\boldsymbol{g}^{(N)}(t) + N^{1-f-\beta}F\left(\sum_{j=1}^{2}\vartheta_j\left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t)\right)\right)\Pi\boldsymbol{g}^{(N)}(t)\\ &\quad + N^{2-2f-2\beta}F\left(\sum_{j=1}^{2}\vartheta_j\left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t)\right)\right)F\\ &\quad \times\left(\sum_{j=1}^{2}\vartheta_j\left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t)\right)\right)\boldsymbol{g}^{(N)}(t) \qquad (10)\\ &\quad + N^{1-f-2\beta}F\left(\sum_{j=1}^{2}\frac{\vartheta_j^2}{2}\Delta_{j,t}\right)\Pi\boldsymbol{g}^{(N)}(t)\\ &\quad - N^{-f}F\Pi\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}^{(N)}(t) + o(N^{2-2f-2\beta}) + o(N^{1-f-2\beta}).\end{aligned}$$

It is noticed that this relation remains valid with $\boldsymbol{g}^{(N)}(t)$ is replaced by $\Pi\boldsymbol{g}^{(N)}(t)$ in the term (10); this is seen when iterating the relation once more. Premultiply the resulting relation with $\mathbf{1}^{\mathrm{T}}\Pi\cdot N^f = \boldsymbol{\pi}^{\mathrm{T}}N^f$. Observing that immediately from the definition of $\varrho_j^{(\mathrm{II})}(t)$

$$\mathbf{1}^{\mathrm{T}}\Pi F\left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t)\right)\Pi = \mathbf{1}^{\mathrm{T}}\Pi\left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t)\right)\mathbf{1}\boldsymbol{\pi} = 0,$$

using $\Pi F = F\Pi = \Pi$ (see, e.g., [10]), we thus obtain

$$\begin{aligned}0 &= N^{2-f-2\beta}\boldsymbol{\pi}^{\mathrm{T}}\left(\sum_{j=1}^{2}\vartheta_j\left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t)\right)\right)F\\ &\quad\times\left(\sum_{j=1}^{2}\vartheta_j\left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t)\right)\right)\Pi\boldsymbol{g}^{(N)}(t)\\ &\quad + N^{1-2\beta}\boldsymbol{\pi}^{\mathrm{T}}\left(\sum_{j=1}^{2}\frac{\vartheta_j^2}{2}\Delta_{j,t}\right)\Pi\boldsymbol{g}^{(N)}(t) - \boldsymbol{\pi}^{\mathrm{T}}\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}^{(N)}(t)\\ &\quad + o(N^{2-f-2\beta}) + o(N^{1-2\beta}).\end{aligned}$$

Now remark that $\Pi\boldsymbol{g}^{(N)}(t)$ can be written as $\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{g}^{(N)}(t) = \mathbf{1}h^{(N)}(t)$ for a scalar moment-generating function $h^{(N)}(t)$. We now compute $h(t)$, defined as $\lim_{N\to\infty}h^{(N)}(t)$. Again, two cases need to be distinguished: $f > 1$ and $f < 1$ (with, as before, $f = 1$ being a boundary case that needs to be handled separately).

- If $f < 1$, then $\beta = 1 - f/2 > 1/2$. As $N \to \infty$, the above equation becomes

$$\boldsymbol{\pi}^{\mathrm{T}} \left(\sum_{j=1}^{2} \vartheta_j \left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t) \right) \right) F \times \left(\sum_{j=1}^{2} \vartheta_j \left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t) \right) \right) \mathbf{1} \cdot h(t) = h'(t).$$

  It is readily verified that, using $F = D + \Pi$ and the definitions of $\varrho_j^{(\mathrm{II})}(t)$ and $\Delta_{j,t}$, for $i, j = 1, 2$,

$$\begin{aligned} &\boldsymbol{\pi}^{\mathrm{T}} \left(\Delta_{i,t} - (\varrho_i^{(\mathrm{II})})'(t) \right) F \left(\Delta_{j,t} - (\varrho_j^{(\mathrm{II})})'(t) \right) \mathbf{1} \\ &\quad = \boldsymbol{\pi}^{\mathrm{T}} \Delta_{i,t} F \Delta_{j,t} \mathbf{1} - (\varrho_i^{(\mathrm{II})})'(t) \cdot (\varrho_j^{(\mathrm{II})})'(t) = \boldsymbol{\pi}^{\mathrm{T}} \Delta_{i,t} D \Delta_{j,t} \mathbf{1}. \end{aligned}$$

  Recalling the definitions of $v_k^{(\mathrm{II})}(t)$ and $c^{(\mathrm{II})}(t)$ from (5) and (6), respectively, and taking into account the obvious boundary conditions, it is now verified that the above differential equation is solved by

$$h_-(t) := \exp\left(\frac{1}{2} \left(v_1^{(\mathrm{II})}(t) \vartheta_1^2 + 2c^{(\mathrm{II})}(t) \vartheta_1 \vartheta_2 + v_2^{(\mathrm{II})}(t) \vartheta_2^2 \right) \right).$$

- If $f > 1$, then $\beta = 1/2$, and we obtain

$$\boldsymbol{\pi}^{\mathrm{T}} \left(\sum_{j=1}^{2} \frac{\vartheta_j^2}{2} \Delta_{j,t} \right) \mathbf{1} \cdot h(t) = h'(t).$$

  Imposing the appropriate boundary conditions, it is elementary to check that this differential equation is solved by

$$h_+(t) := \exp\left(\frac{1}{2} \left(\varrho_1^{(\mathrm{II})}(t) \vartheta_1^2 + \varrho_2^{(\mathrm{II})}(t) \vartheta_2^2 \right) \right).$$

- In case $f = 1$, both terms contribute, leading to $h(t) = h_-(t) + h_+(t)$.

Define

$$\Sigma^{(\mathrm{II})}(t) := \begin{pmatrix} v_1^{(\mathrm{II})}(t) & c^{(\mathrm{II})}(t) \\ c^{(\mathrm{II})}(t) & v_2^{(\mathrm{II})}(t) \end{pmatrix} 1_{\{f \leqslant 1\}} + \begin{pmatrix} \varrho_1^{(\mathrm{II})}(t) & 0 \\ 0 & \varrho_2^{(\mathrm{II})}(t) \end{pmatrix} 1_{\{f \geqslant 1\}}.$$

We have proven the following result.

Theorem 8.1.1. *Consider Model II. For any $t \geq 0$, the random variable*

$$\left(\frac{M_1^{(N)}(t) - N\varrho_1{}^{(\mathrm{II})}(t)}{N^\beta}, \frac{M_2^{(N)}(t) - N\varrho_2{}^{(\mathrm{II})}(t)}{N^\beta} \right)$$

converges to a bivariate Normal distribution with mean $\mathbf{0}$ *and covariance matrix* $\Sigma^{(\mathrm{II})}(t)$ *as* $N \to \infty$.

8.2. *Model II, stationary case*

As could be anticipated on the basis of Theorem. 8.1.1, the CLT for the stationary case is as follows. Define

$$\Sigma^{(\mathrm{II})} := \lim_{t\to\infty} \Sigma^{(\mathrm{II})}(t) = \begin{pmatrix} v_1^{(\mathrm{II})} & c^{(\mathrm{II})} \\ c^{(\mathrm{II})} & v_2^{(\mathrm{II})} \end{pmatrix} 1_{\{f\leqslant 1\}} + \begin{pmatrix} \varrho_1^{(\mathrm{II})} & 0 \\ 0 & \varrho_2^{(\mathrm{II})} \end{pmatrix} 1_{\{f\geqslant 1\}},$$

with

$$v_k^{(\mathrm{II})} := 2\sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\lambda_{k,i}\lambda_{k,j}}{\mu_{k,i}+\mu_{k,j}}\pi_i D_{ij}, \quad c^{(\mathrm{II})} := \sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\lambda_{1,i}\lambda_{2,j}}{\mu_{1,i}+\mu_{2,j}}\left(\pi_i D_{ij} + \pi_j D_{ji}\right).$$

Theorem 8.2.1. *Consider Model II. The random variable*

$$\left(\frac{M_1^{(N)} - N\varrho_1{}^{(\mathrm{II})}}{N^\beta}, \frac{M_2^{(N)} - N\varrho_2{}^{(\mathrm{II})}}{N^\beta}\right)$$

converges to a bivariate Normal distribution with mean $\mathbf{0}$ *and covariance matrix* $\Sigma^{(\mathrm{II})}$ *as* $N\to\infty$.

It is important to notice that this result does *not* follow directly from Theorem 8.1, as that would involve interchanging the limits $t\to\infty$ and $N\to\infty$, for which a formal justification is lacking. The way to rigorously prove this result is analogous to the corresponding result for the single-system case in Ref.[4], viz. using the differential equations featuring in Proposition 4.1.1. We omit the full derivation of this result.

9. Numerical illustration

As a numerical illustration of the dichotomy, we plot for Model I the variance and covariance of the system contents; these are computed using the results from Section 3. The numerics correspond to the stationary numbers of jobs in the system, imposing the scaling studied in detail in Section 7, i.e., $M_1^{(N)}$ and $M_2^{(N)}$, in the regime $N\to\infty$.

In the experiment the background Markov chain has two states, with transition rates $q_{12}=2$ and $q_{21}=3$. The (unscaled) arrival and departure rates are as follows:

$$\boldsymbol{\lambda}_1 = [2\ \ 1], \quad \boldsymbol{\lambda}_2 = [1\ \ 2], \quad \boldsymbol{\mu}_1 = [1\ \ 5], \quad \boldsymbol{\mu}_2 = [5\ \ 1].$$

As is directly seen from Figures 1 and 2, using the scaling $\boldsymbol{\lambda}_i \mapsto N\boldsymbol{\lambda}_i$ for $i=1,2$, and $Q\mapsto N^f Q$, we indeed observe an intrinsically different limit behavior for $f<1$ and $f>1$. The (normalized) variance peaks at $f=1$, in line with the spike that the limiting variance has at $f=1$; see Theorem 7.2.1. The covariance is negative for $f<1$ and vanishes for $f>1$ (as $N\to\infty$), as desired.

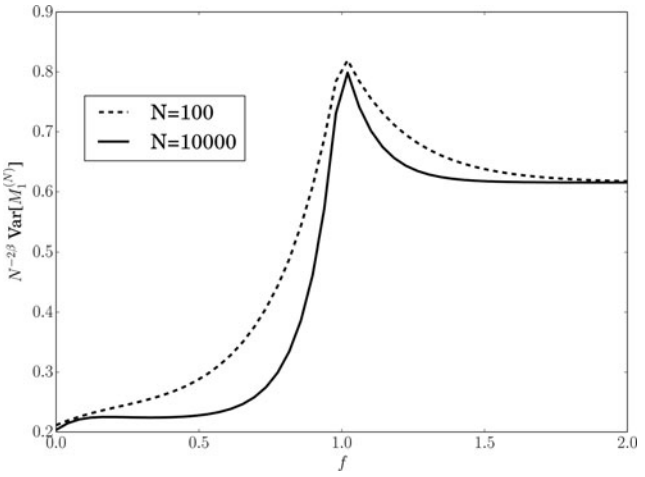

Figure 1. The scaled variance of $M_1^{(N)}$.

10. Discussion and concluding remarks

This paper has extended the results of Refs.[4,14] to the situation of *multiple* Markov-modulated infinite-server queues driven by a common background process. These results concern the probability-generating function for the transient and stationary distributions, recursive procedures to generate the corresponding moments, and central limit theorems under a specific scaling.

The model that we analyzed has the potential to be applied in a wide variety of settings. For instance, in the context of mathematical finance, a key problem concerns the composition of portfolios. A portfolio consists of a set of, typically correlated, financial assets, such as stocks and bonds, or potentially also options. The objective is to compose a portfolio such that the revenue is maximized, while the corresponding risk is kept at an acceptable level. Noticing that the asset prices are (partly) affected by the same economic forces, it becomes clear that models in

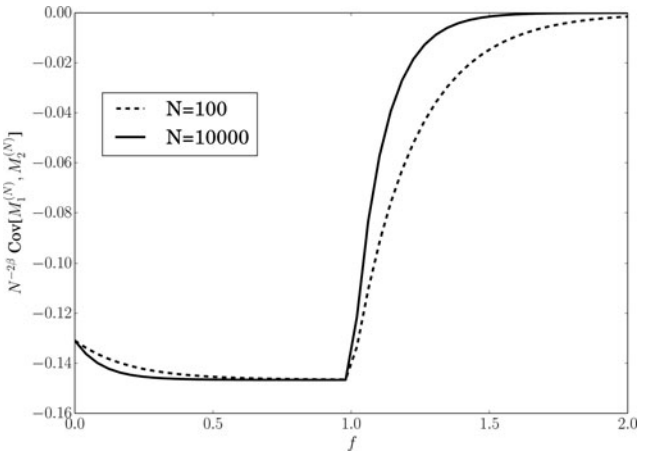

Figure 2. The scaled covariance between $M_1^{(N)}$ and $M_2^{(N)}$.

the spirit of the one discussed in this paper can be used; see also the exposition in Ref.[12].

A second example can be found in biology. As argued in, e.g., Ref.[17], the infinite-server model can be used to describe the concentration of mRNA in cells: molecules are generated, and they remain present for some random duration. The generation and decay processes, however, are subject to external factors, such as temperature; those factors can be captured by imposing Markov modulation. Clearly, when studying multiple "nearby" cells, which react to the same external factors, our model can be used.

A third example concerns wireless communication networks. The channel conditions in adjacent cells are typically highly correlated, which could be described by Markov modulation. Modelling the number of clients in the individual cells as infinite-server queues (as an approximation to queues that can accommodate a finite but relatively large number of clients), our model can be used to study the joint distribution of the number of users present.

In the first part of this paper we have derived differential equations that characterize the probability-generating function of the numbers of jobs in both queues. In principle, these (ordinary or partial) differential equations uniquely define the probabilistic properties of our queueing system, but they do not allow an explicit solution (except in very special cases). As is often done in such situations, we consider scalings under which closed-form asymptotic results can be derived. In our setup we scale both the arrival rates and the transition rates of the modulating Markov process. Scaling the arrival rates by a factor N, for N large, can be interpreted as considering a system that is used by a large superposition of users. Interestingly, we speed up the transition rates by a *different* factor, i.e., N^f; this allows us to obtain insight into the effect of these different speeds.

Possible topics for follow-up research include (i) functional versions of the central limit theorems, in the spirit of Ref.[1], (ii) networks of Markov-modulated infinite-server queues (where the output of one queue can serve as input for a next queue), (iii) large deviations results under the scaling we have considered in this paper, similar to those derived in Refs.[5,7,8] for non-coupled Markov-modulated infinite-server queues.

Acknowledgements

This research was partly performed when K. De Turck was a Postdoctoral Fellow of Fonds Wetenschappelijk Onderzoek / Research Foundation–Flanders. The authors thank Peter Taylor (University of Melbourne) for suggesting this problem, and for stimulating discussions.

Funding

M. Mandjes is also affiliated to Eurandom, Eindhoven University of Technology, Eindhoven, the Netherlands, and IBIS, Faculty of Economics and Business, University of Amsterdam, Amsterdam, the Netherlands. His research is partly funded by the NWO Gravitation project Networks, grant number 024.002.003.

References

[1] Anderson, D.; Blom, J.; Mandjes, M.; Thorsdottir, H.; de Turck, K. A functional central limit theorem for a Markov-modulated infinite-server queue. Methodology and Computing in Applied Probability (in press).

[2] Asmussen, S. *Applied Probability and Queues*, 2nd edition; Springer: New York, 2003.

[3] Bean, N.; O'Reilly, M. A stochastic fluid model driven by an uncountable-state process, which is a stochastic fluid model itself. Stochastic Proc. Appl. **2014**, *124*, 1741–1772.

[4] Blom, J.; de Turck, K.; Mandjes, M. Analysis of Markov-modulated infinite-server queues in the central-limit regime. Probab. Eng. Info. Sci. **2015**, *29*, 433–459. (A short version has appeared in: *Proceedings ASMTA 2013*, Ghent, Belgium. Lecture Notes in Computer Science (LNCS) Series, Vol. 7984, 81–95.)

[5] Blom, J.; de Turck, K.; Mandjes, M. Rare-event analysis of Markov-modulated infinite-server queues: A Poisson limit. Stochastic Models **2013**, *29*, 463–474.

[6] Blom, J.; Kella, O.; Mandjes, M.; Thorsdottir, H. Markov-modulated infinite-server queues with general service times. Queueing Syst. **2013**, *76*, 403–424.

[7] Blom, J.; Kella, O.; Mandjes, M.; de Turck, K. Tail asymptotics of a Markov-modulated infinite-server queue. Queueing Syst. **2014**, *78*, 337–357.

[8] Blom, J.; Mandjes, M. A large-deviations analysis of Markov-modulated infinite-server queues. Oper. Research Letters **2012**, *41*, 220–225.

[9] Bright, L.; Taylor, P. Calculating the equilibrium distribution in level dependent quasi-birth-and-death processes. Stochastic Models **1995**, *11*, 497–526.

[10] Coolen-Schrijner, P.; van Doorn, E. The deviation matrix of a continuous-time Markov chain. Probab. Eng. Infor. Sci. **2002**, *16*, 351–366.

[11] D'Auria, B. M/M/∞ queues in semi-Markovian random environment. Queue. Syst. **2008**, *58*, 221–237.

[12] Huang, G.; Jansen, H. M.; Mandjes, M.; Spreij, P.; de Turck, K. Markov-modulated Ornstein-Uhlenbeck processes. J. Appl. Probab. (in press).

[13] Latouche, G.; Ramaswami, V. *Introduction to Matrix Analytic Methods in Stochastic Modelling*. ASA/SIAM Series on Statistics and Applied Probability; Philadelphia PA, 1999.

[14] O'Cinneide, C.; Purdue, P. The M/M/∞ queue in a random environment. J. Appl. Probab. **1986**, *23*, 175–184.

[15] Neuts, M. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*; Courier Dover: New York, 1981.

[16] Ramaswami, V.; Taylor, P.G. Some properties of the rate matrices in level dependent quasi-birth-and-death processes with a countable number of phases. Stochastic Models **1996**, *12*, 143–164.

[17] Schwabe, A.; Rybakova, K.; Bruggeman, F. Transcription stochasticity of complex gene regulation models. Biophysical Journal **2012**, *103*, 1152–1161.