Math Meth Oper Res (2016) 83:351–372
DOI 10.1007/s00186-016-0531-7

Functional central limit theorems for Markov-modulated infinite-server systems

J. Blom[2] · K. De Turck[3] · M. Mandjes[1,2,4,5]

Received: 30 April 2015 / Accepted: 10 January 2016 / Published online: 11 March 2016

Abstract In this paper we study the Markov-modulated M/M/∞ queue, with a focus on the correlation structure of the number of jobs in the system. The main results describe the system's asymptotic behavior under a particular scaling of the model parameters in terms of a functional central limit theorem. More specifically, relying on the martingale central limit theorem, this result is established, covering the situation in which the arrival rates are sped up by a factor N and the transition rates of the background process by N^α, for some $\alpha > 0$. The results reveal an interesting dichotomy, with crucially different behavior for $\alpha > 1$ and $\alpha < 1$, respectively. The limiting Gaussian process, which is of the Ornstein–Uhlenbeck type, is explicitly identified, and it is shown to be in accordance with explicit results on the mean, variances and covariances of the number of jobs in the system.

Keywords Queues · Infinite-server systems · Markov modulation · Central limit theorems

M. Mandjes' research is partly funded by the NWO Gravitation project NETWORKS, Grant No. 024.002.003.

✉ K. De Turck
koen.deturck@gmail.com

[1] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

[2] CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[3] Laboratoire Signaux et Systèmes (L2S, CNRS UMR8506), École CentraleSupélec, Université Paris Saclay, 3 Rue Joliot Curie, Plateau de Moulon, 91190 Gif-sur-Yvette, France

[4] EURANDOM, Eindhoven University of Technology, Eindhoven, The Netherlands

[5] IBIS, Faculty of Economics and Business, University of Amsterdam, Amsterdam, The Netherlands

1 Introduction

This paper studies the infinite-server queue modulated by a finite-state irreducible continuous-time Markov chain J; when the so-called *background process* J is in state i, jobs arrive according to a Poisson process with rate λ_i, while the departure rate is given by μ_i. The resulting Markov-modulated infinite-server queue has attracted some attention during the past decades; see e.g. the early contributions (D'Auria 2008; Keilson and Servi 1993; O'Cinneide and Purdue 1986). In these papers the main results were in terms of systems of (partial) differential equations characterizing probability generating functions related to the system's transient behavior, and recursions enabling the evaluation of the corresponding moments.

In a series of recent papers (Anderson et al. 2015; Blom et al. 2013a, b, 2014a, b, 2015; Blom and Mandjes 2013), substantial attention has been paid to the asymptotic behavior of Markov-modulated infinite-server queues in specific scaling regimes. In these parameter scalings the arrival rates are typically inflated by a factor N, while the transition rates of the background process are sped up by a factor N^α for some $\alpha \geq 0$. The objective is to analyze the transient distribution of the number of jobs in the system at time t, to be denoted by $M^{(N)}(t)$, in the limiting regime that N grows large.

The asymptotic results derived come in three flavors: (1) large deviations (LD) results, describing the tail probabilities $\mathbb{P}(M^{(N)}(t)/N \geq a)$ for N large; (2) central-limit-theorem (CLT) type of results, describing the convergence of $M^{(N)}(t)$ (after centering and normalization) to a Normally distributed random variable; and (3) functional central limit theorems (FCLTs), describing the convergence of the process $M^{(N)}(\cdot)$ to an appropriate Gaussian process.

Importantly, two model variants can be distinguished, with their own specific departure processes.

- In the first, to be referred to as Model I, each job present is experiencing a departure rate μ_i when J is in state i; as a consequence, this hazard rate may change during the job's sojourn time (when the background process makes a transition).
- In the second, Model II, the job's sojourn time is sampled upon arrival: when the background process is then in state i, it has an exponential distribution with mean $1/\mu_i$, and hence the corresponding hazard rate is constant over its lifetime.

Figure 1 summarizes the results that have been established so far. In the LD domain, the papers (Blom et al. 2013b, 2014a; Blom and Mandjes 2013) cover, for both models, the regime in which the background process is relatively slow (more specifically, $\alpha = 0$) as well as the regime in which it is essentially faster than the arrival process ($\alpha > 1$). Also in the CLT regime the picture is complete, with results for Models I and II, and with both slow ($\alpha < 1$) and fast ($\alpha > 1$) switching of the background process. In terms of FCLTs, however, not all cases are covered: the only result derived so far (Anderson et al. 2015) concerns the case that $\mu_i = \mu$ for all i, i.e., the case in which Models I and II actually coincide; we may refer to this model as to 'Model 0'. The main contribution of the present paper is the derivation of FCLTs for Models I and II; this is done in Sects. 5 and 6, respectively. These findings, with a limiting Gaussian process of the Ornstein–Uhlenbeck type, turn out to be in accordance with explicit

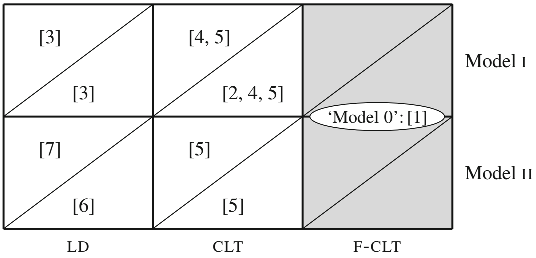

Fig. 1 Graphical illustration of literature on Markov-modulated infinite-server queues. *Upper-left triangle* fast regime; *lower-right triangle* slow regime. *White area* regimes covered by earlier work; *shaded areas* regimes not covered yet

expressions for means, variances, and covariances in these models, as we present in Sects. 3 and 4. We conclude in Sect. 7 with some numerical experiments.

2 Notation, preliminaries

Let $J(t)$ denote an irreducible continuous-time Markov chain on the (finite) state space $\{1,\dots,d\}$, with transition rate matrix $Q = (q_{ij})_{i,j=1}^{d}$ and (unique) invariant probability measure $\boldsymbol{\pi}$. In addition, we let $p_{ij}(t) := \mathbb{P}(J(t) = j \mid J(0) = i)$. It is assumed that $J(0)$ is distributed according to $\boldsymbol{\pi}$.

The process $J(t)$ is referred to as the *background process*, and regulates an infinite-server queue. When $J(t)$ is in state i, jobs arrive at the queueing resource according to a Poisson process of rate λ_i. Regarding the way in which these jobs are handled, two variants are distinguished:

- In Model I the hazard rate of jobs leaving is μ_i when the background process is in state i. Observe that this hazard rate may change during the lifetime of the job, when the background process jumps.
- In Model II job durations are sampled upon arrival: they are drawn from an exponential distribution with mean $1/\mu_i$ if the background process is in state i when the job enters the system.

Throughout this paper we write $\boldsymbol{\lambda} := (\lambda_1,\dots,\lambda_d)^{\mathrm{T}}$ and $\Lambda := \mathrm{diag}\{\boldsymbol{\lambda}\}$, and likewise $\boldsymbol{\mu} := (\mu_1,\dots,\mu_d)^{\mathrm{T}}$ and $\mathcal{M} := \mathrm{diag}\{\boldsymbol{\mu}\}$. We also define $\lambda_\infty := \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{\lambda}$ and $\mu_\infty := \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{\mu}$.

In Sects. 3 and 4 we consider explicit expressions for the means, variances and covariances in the *unscaled system*. There we denote by $M(t)$ the number of jobs present at time t, for $t \geq 0$. For simplicity, it is assumed that the system starts empty at time 0, i.e., $M(0) = 0$.

In Sects. 3 and 4 we also analyze the obtained expressions for the mean, variance and covariance in a specific parameter scaling, viz. we replace the arrival rates $\boldsymbol{\lambda}$ by $N\boldsymbol{\lambda}$, and the generator matrix Q by $N^\alpha Q$, for some $\alpha > 0$, and let N grow large. It is in this asymptotic regime that we also establish our FCLTs in Sects. 5 and 6. For these scaled models we write $M^{(N)}(t)$ for the number of jobs at time t, to emphasize the dependence on the scaling parameter.

In the sequel, we use the concept of *deviation matrices*. Define the (i,j)-th element of the *exponentially weighted deviation matrix* $D^{(\boldsymbol{\gamma})}$, as a function of the vector $\boldsymbol{\gamma} \in \mathbb{R}_+^d$, by

$$D_{ij}^{(\boldsymbol{\gamma})} := \int_0^\infty e^{-\gamma_i t}\left(p_{ij}(t) - \pi_j\right)\mathrm{d}t.$$

The matrix $D := D^{(\mathbf{0})}$ is the canonical *deviation matrix*. In the sequel, also the matrix $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$ plays a role, as well as the fundamental matrix $F := D + \Pi$. A number of identities hold: $QF = FQ = \Pi - I$, $\Pi F = F\Pi = \Pi$, and $F\mathbf{1} = \Pi\mathbf{1} = \mathbf{1}$.

3 Mean, variance, and covariance for model I

The first part of this section presents explicit formulae for the mean, variance, and covariance in Model I. In the second part these turn out to allow for a more explicit characterization in particular asymptotic regimes.

3.1 Explicit formulae

Our goal is to devise a method to compute $\mathbb{C}\mathrm{ov}(M(t), M(t+u))$. To this end, the object that we study first is, for $u \geq 0$ fixed, the bivariate probability generating function

$$\Xi_{ij}(z, w, t, u) := \mathbb{E}\left(z^{M(t)} w^{M(t+u)}\, 1\{J(t) = i, J(t+u) = j\}\right),$$

which implicitly contains all information about the joint distribution of $M(t)$ and $M(t+u)$. In matrix notation, we obtain in "Appendix 1", suppressing the arguments for ease of notation,

$$\frac{\partial\,\Xi}{\partial t} = (z-1)\Lambda\,\Xi + (w-1)\Xi\Lambda - (z-1)\mathcal{M}\frac{\partial\,\Xi}{\partial z} - (w-1)\frac{\partial\,\Xi}{\partial w}\mathcal{M} + Q^{\mathrm{T}}\Xi + \Xi Q. \tag{1}$$

We now point out how to compute the covariance between $M(t)$ and $M(t+u)$ from this system of partial differential equations.

To this end, we first define the three matrices

$$\begin{aligned}
E(t,u) &\equiv (E_{ij}(t,u))_{i,j=1}^d, &&\text{where } E_{ij}(t,u) := \mathbb{E}M(t)1\{J(t)=i, J(t+u)=j\}\\
G(t,u) &\equiv (G_{ij}(t,u))_{i,j=1}^d, &&\text{where } G_{ij}(t,u) := \mathbb{E}M(t+u)1\{J(t)=i, J(t+u)=j\}\\
C(t,u) &\equiv (C_{ij}(t,u))_{i,j=1}^d, &&\text{where } C_{ij}(t,u) := \mathbb{E}M(t)M(t+u)1\{J(t)=i, J(t+u)=j\}
\end{aligned} \tag{2}$$

It follows from the moment-generating property of generating functions that

$$E_{ij}(t,u) = \lim_{z,w\uparrow 1}\frac{\partial\,\Xi_{ij}}{\partial z},\quad G_{ij}(t,u) = \lim_{z,w\uparrow 1}\frac{\partial\,\Xi_{ij}}{\partial w},\quad C_{ij}(t,u) = \lim_{z,w\uparrow 1}\frac{\partial^2\,\Xi_{ij}}{\partial z\partial w}. \tag{3}$$

From the partial differential equation (1) that defines Ξ, we can find the following systems of ordinary differential equations for the matrices $E(t,u)$, $G(t,u)$ and $C(t,u)$.

We demonstrate how this is done for the equation involving $E(t,u)$. Differentiate (1) with respect to z, and take the limit of $w, z \uparrow 1$. Recalling that $J(0)$ is distributed according to $\boldsymbol{\pi}$, it is straightforward to obtain, with $K_{ij}(u) := \pi_i p_{ij}(u)$,

$$E'(t,u) = \Lambda K(u) - \mathcal{M}E(t,u) + Q^{\mathrm{T}}E(t,u) + E(t,u)Q,$$

where the derivative in the left-hand side is again with respect to t. We can derive the ODEs for $G(t,u)$ in the same manner,

$$G'(t,u) = K(u)\Lambda - G(t,u)\mathcal{M} + Q^{\mathrm{T}}G(t,u) + G(t,u)Q.$$

Similarly, for $C(t,u)$ we have:

$$C'(t,u) = \Lambda G(t,u) + E(t,u)\Lambda - \mathcal{M}C(t,u) - C(t,u)\mathcal{M} + Q^{\mathrm{T}}C(t,u) + C(t,u)Q,$$

The above differential equations are *matrix-valued* systems of linear differential equations, which can be converted into *vector-valued* systems of linear differential equations, relying on the concept of ‘vectorization’. We show this idea for the matrix $E(t,u)$. We take the columns of $E(t,u)$, and put them into a vector $\boldsymbol{e}(t,u)$ of dimension d^2, such that the first d entries are $E_{11}(t,u)$ up to $E_{d1}(t,u)$, entries $d+1$ up to $2d$ correspond to $E_{12}(t,u)$ up to $E_{d2}(t,u)$, etc.; we write $\boldsymbol{e}(t,u) := \mathrm{vec}(E(t,u))$. Likewise, $\boldsymbol{g}(t,u) := \mathrm{vec}(G(t,u))$, $\boldsymbol{c}(t,u) := \mathrm{vec}(C(t,u))$ and $\boldsymbol{k}(u) := \mathrm{vec}(K(u))$.

For $d \times d$ matrices A, B, and C, and with as usual $A \otimes B$ denoting the Kronecker product and $A \oplus B := A \otimes I + I \otimes B$ the Kronecker sum of the matrices A and B, recall

$$\mathrm{vec}(AB) = (I \otimes A)\mathrm{vec}(B) = \left(B^{\mathrm{T}} \otimes I\right)\mathrm{vec}(A) \text{ and } \mathrm{vec}(ABC) = \left(C^{\mathrm{T}} \otimes A\right)\mathrm{vec}(B),$$

with I the $d \times d$ identity matrix. We thus obtain the following equations in terms of Kronecker sums and products:

$$\boldsymbol{e}'(t,u) = (I \otimes \Lambda)\boldsymbol{k}(u) - (I \otimes \mathcal{M})\boldsymbol{e}(t,u) + \left(Q^{\mathrm{T}} \oplus Q^{\mathrm{T}}\right)\boldsymbol{e}(t,u).$$

An equation for $\boldsymbol{g}(t,u)$ can be found analogously:

$$\boldsymbol{g}'(t,u) = (\Lambda \otimes I)\boldsymbol{k}(u) - (\mathcal{M} \otimes I)\boldsymbol{g}(t,u) + \left(Q^{\mathrm{T}} \oplus Q^{\mathrm{T}}\right)\boldsymbol{g}(t,u).$$

Along the same lines we obtain

$$\boldsymbol{c}'(t,u) = (I \otimes \Lambda)\boldsymbol{g}(t,u) + (\Lambda \otimes I)\boldsymbol{e}(t,u) - (\mathcal{M} \oplus \mathcal{M})\boldsymbol{c}(t,u) + (Q^{\mathrm{T}} \oplus Q^{\mathrm{T}})\boldsymbol{c}(t,u),$$

the derivatives in the left-hand sides being again with respect to t. Observe that $Q \oplus Q$ is again a transition rate matrix, and $\mathcal{M} \oplus \mathcal{M}$ a diagonal matrix with non-negative entries.

The systems describing $\boldsymbol{e}(t,u)$ and $\boldsymbol{g}(t,u)$ are standard systems of non-homogeneous linear differential equations, which can be solved with standard techniques. Then the solution can be plugged into the differential equation describing $\boldsymbol{c}(t,u)$, which is then also a system of non-homogeneous linear differential equations. We summarize the results in the following proposition.

Proposition 1 *The matrix-valued functions $E(t,u)$, $G(t,u)$ and $C(t,u)$ satisfy the following* ODE*s:*

$$\begin{aligned}
E'(t,u) &= \Lambda K(u) - \mathcal{M}E(t,u) + Q^{\mathrm{T}}E(t,u) + E(t,u)Q,\\
G'(t,u) &= K(u)\Lambda - G(t,u)\mathcal{M} + Q^{\mathrm{T}}G(t,u) + G(t,u)Q.\\
C'(t,u) &= \Lambda G(t,u)+E(t,u)\Lambda-\mathcal{M}C(t,u) - C(t,u)\mathcal{M}+Q^{\mathrm{T}}C(t,u)+C(t,u)Q,
\end{aligned}$$

Moreover, the vectorized versions $\boldsymbol{e}(t,u)$, $\boldsymbol{g}(t,u)$ and $\boldsymbol{c}(t,u)$, of the matrices $E(t,u)$, $G(t,u)$ and $C(t,u)$ satisfy the following linear differential equations.

$$\begin{aligned}
\boldsymbol{e}'(t,u) &= (I\otimes\Lambda)\boldsymbol{k}(u) - (I\otimes\mathcal{M})\boldsymbol{e}(t,u) + \left(Q^{\mathrm{T}}\oplus Q^{\mathrm{T}}\right)\boldsymbol{e}(t,u).\\
\boldsymbol{g}'(t,u) &= (\Lambda\otimes I)\boldsymbol{k}(u) - (\mathcal{M}\otimes I)\boldsymbol{g}(t,u) + \left(Q^{\mathrm{T}}\oplus Q^{\mathrm{T}}\right)\boldsymbol{g}(t,u).\\
\boldsymbol{c}'(t,u) &= (I\otimes\Lambda)\boldsymbol{g}(t,u) + (\Lambda\otimes I)\boldsymbol{e}(t,u) - (\mathcal{M}\oplus\mathcal{M})\boldsymbol{c}(t,u)\\
&\quad + \left(Q^{\mathrm{T}}\oplus Q^{\mathrm{T}}\right)\boldsymbol{c}(t,u).
\end{aligned}$$

All occurring derivatives are with respect to t.

We have now devised a procedure to compute the covariance $\mathbb{C}\mathrm{ov}(M(t), M(t+u))$. To this end, first realize that, with $e(t) := \mathbb{E}M(t)$ and $\mathbf{1}$ denoting here a d^2-dimensional all-ones vector,

$$e(t) = \mathbf{1}^{\mathrm{T}}\boldsymbol{e}(t,u), \quad e(t+u) = \mathbf{1}^{\mathrm{T}}\boldsymbol{g}(t,u).$$

As a consequence,

$$\mathbb{C}\mathrm{ov}(M(t), M(t+u)) = \mathbf{1}^{\mathrm{T}}\boldsymbol{c}(t,u) - e(t)\,e(t+u).$$

3.2 Two specific limiting regimes

In this subsection, we consider two particular limiting regimes, in which the expressions simplify considerably.

▷ Let us first consider the behavior for $t\to\infty$. It is readily verified that

$$\boldsymbol{e}(\infty,u) := \lim_{t\to\infty}\boldsymbol{e}(t,u) = \left((I\otimes\mathcal{M}) - \left(Q^{\mathrm{T}}\oplus Q^{\mathrm{T}}\right)\right)^{-1}(I\otimes\Lambda)\boldsymbol{k}(u),$$

and hence

$$e(\infty) = \mathbf{1}^{\mathrm{T}}\boldsymbol{e}(\infty, u) = \mathbf{1}^{\mathrm{T}}\left((I \otimes \mathcal{M}) - \left(Q^{\mathrm{T}} \oplus Q^{\mathrm{T}}\right)\right)^{-1}(I \otimes \Lambda)\boldsymbol{k}(u) = \mathbf{1}^{\mathrm{T}}\boldsymbol{g}(\infty, u).$$

For $u = 0$ we obtain the solution from O'Cinneide and Purdue (1986, Thm. 3.1).

Along the same lines,

$$\boldsymbol{c}(\infty, u) := \lim_{t\to\infty} \boldsymbol{c}(t, u) = \left((\mathcal{M} \oplus \mathcal{M}) - \left(Q^{\mathrm{T}} \oplus Q^{\mathrm{T}}\right)\right)^{-1}((I \otimes \Lambda)\boldsymbol{g}(\infty, u) \\ + (\Lambda \otimes I)\boldsymbol{e}(\infty, u)).$$

We have thus derived an expression for the limit of $\mathbb{C}\mathrm{ov}(M(t), M(t+u))$ as $t \to \infty$:

$$\lim_{t\to\infty} \mathbb{C}\mathrm{ov}(M(t), M(t+u)) = \mathbf{1}^{\mathrm{T}}\boldsymbol{c}(\infty, u) - (e(\infty))^2.$$

▷ Next, we consider the following scaling: we replace $\boldsymbol{\lambda} \mapsto N\boldsymbol{\lambda}$ and $Q \mapsto N^{\alpha}Q$, for $\alpha > 0$. In this regime, the pace with which the arrival process is sped up, differs from that corresponding to the background process. As we will see below, the situation $\alpha > 1$ crucially differs from $\alpha < 1$; this was already observed earlier in e.g. Anderson et al. (2015) and Blom et al. (2015). As mentioned before, to stress the dependence on N, we write $M^{(N)}(t)$ rather than $M(t)$. It is this scaling that is imposed in Sect. 5, and under which an FCLT is established. We now identify the associated mean and (co-)variance, relying on elementary techniques.

Let $\boldsymbol{m}^{(N)}(t) \equiv \boldsymbol{m}(t)$ the d-dimensional *row*-vector, with $\mathbb{E}M^{(N)}(t)1\{J(t) = i\}$ on the i-th position. According to O'Cinneide and Purdue (1986, Thm. 3.2), $\boldsymbol{m}(t)$ satisfies the following non-homogeneous linear differential equation:

$$\boldsymbol{\pi}^{\mathrm{T}}N\Lambda - \boldsymbol{m}(t)(\mathcal{M} - N^{\alpha}Q) = \boldsymbol{m}'(t). \tag{4}$$

In Blom et al. (2015) we proved that, with $\varrho^{(i)} := \lambda_\infty/\mu_\infty$,

$$\mathbb{E}M^{(N)}(t) = N\,\varrho^{(i)}\,(1 - e^{-\mu_\infty t}) + o(N). \tag{5}$$

Now define $\varrho^{(i)}(t) := \varrho^{(i)}(1 - e^{-\mu_\infty t})$ and

$$\varsigma^{(i)}(t) := 2\int_0^t e^{-2\mu_\infty(t-s)}\boldsymbol{\pi}^{\mathrm{T}}\left(\Lambda - \mathcal{M}\varrho^{(i)}(s)\right)D\left(\Lambda - \mathcal{M}\varrho^{(i)}(s)\right)\mathbf{1}\,\mathrm{d}s. \tag{6}$$

In "Appendix 2" it is shown that

$$\lim_{N\to\infty} \frac{\mathbb{C}\mathrm{ov}\left(M^{(N)}(t), M^{(N)}(t+u)\right)}{N^{\max\{1,2-\alpha\}}} \\ = v^{(i)}(t, u) := e^{-\mu_\infty u}\left(\varsigma^{(i)}(t)1_{\{\alpha\le 1\}} + \varrho^{(i)}(t)1_{\{\alpha\ge 1\}}\right). \tag{7}$$

We conclude that under this parameter scaling the covariance exhibits the same dichotomy as the one observed in Blom et al. (2015) for the variance, i.e., behaving crucially different for $\alpha < 1$ and $\alpha > 1$. In the latter regime, the system essentially behaves as a (non-modulated) M/M/∞ queue, with arrival rate λ_∞ and service rate μ_∞, whereas for $\alpha < 1$ the full transition rate matrix Q plays a role (as $\varsigma^{(i)}(t)$ involves the deviation matrix D).

4 Mean, variance, and covariance for model II

As we saw above, for Model I the mean, variance and covariance can be determined by solving specific non-homogeneous linear differential equations; for Model II, however, the analysis is simpler, and can be performed by relying on the law of total (co-)variance, as shown in Sect. 4.1. Focusing on the same limiting regimes as we have studied for Model I, the expressions become more explicit; see Sect. 4.2.

4.1 Explicit formulae

The mean of $M(t)$ for Model II was already determined in Blom et al. (2014b); recalling from e.g. D'Auria (2008) the observation that $M(t)$ obeys a Poisson distribution with the random parameter $\mathbb{E}(M(t) \mid J)$, we conclude that

$$\begin{aligned}\mathbb{E}M(t) &= \mathbb{E}\left(\mathbb{E}(M(t) \mid J)\right) = \mathbb{E}\left(\int_0^t \lambda_{J(s)} e^{-\mu_{J(s)}(t-s)} \mathrm{d}s\right) \\ &= \sum_{i=1}^{d} \pi_i \frac{\lambda_i}{\mu_i}\left(1 - e^{-\mu_i t}\right) =: \varrho^{(ii)}(t),\end{aligned}$$

with $J \equiv (J(s))_{s=0}^{t}$.

Now concentrate on the evaluation of the covariance between $M(t)$ and $M(t+u)$; assume, without loss of generality, that $u \geq 0$. The 'law of total covariance' entails that

$$\begin{aligned}\mathbb{C}\mathrm{ov}(M(t), M(t+u)) = {} & \mathbb{E}(\mathbb{C}\mathrm{ov}(M(t), M(t+u) \mid J)) \\ & + \mathbb{C}\mathrm{ov}(\mathbb{E}(M(t) \mid J), \mathbb{E}(M(t+u) \mid J)).\end{aligned} \tag{8}$$

In "Appendix 4", we evaluate both terms, so as to obtain

$$\mathbb{C}\mathrm{ov}(M(t), M(t+u)) = \sum_{i=1}^{d} \pi_i \frac{\lambda_i}{\mu_i}\left(1 - e^{-\mu_i t}\right) e^{-\mu_i u} + \boldsymbol{\lambda}^{\mathrm{T}} \mathscr{K}(t,u) \boldsymbol{\lambda} + \boldsymbol{\lambda}^{\mathrm{T}} \mathscr{L}(t,u) \boldsymbol{\lambda}; \tag{9}$$

the precise form of the matrices $\mathscr{K}(t,u)$ and $\mathscr{L}(t,u)$ is given in "Appendix 4" as well.

4.2 Two specific limiting regimes

In this subsection, we consider the two particular limiting regimes that we studied earlier, in Sect. 3.2, for Model I. As it turns out, in these regimes the expressions simplify considerably.

▷ In the first regime, we consider $\mathbb{C}\mathrm{ov}(M(t), M(t+u))$ for $t \to \infty$. Going through the calculations, relying on the explicit expressions for $\mathscr{K}(t,u)$ and $\mathscr{L}(t,u)$ as given in "Appendix 4", we obtain

$$\begin{aligned}\lim_{t\to\infty} &\mathbb{C}\mathrm{ov}(M(t), M(t+u))\\ &= \sum_{i=1}^{d} \pi_i \frac{\lambda_i}{\mu_i} e^{-\mu_i u} + \sum_{i=1}^{d}\sum_{j=1}^{d} \pi_i \frac{\lambda_i\lambda_j}{\mu_i+\mu_j} e^{-\mu_j u} D_{ij}^{(\boldsymbol{\mu})}\\ &\quad + \sum_{i=1}^{d}\sum_{j=1}^{d} \pi_j \frac{\lambda_i\lambda_j}{\mu_i+\mu_j} \int_0^u e^{-\mu_j u+\mu_i w}(p_{ji}(w)-\pi_i)\mathrm{d}w\\ &\quad + \sum_{i=1}^{d}\sum_{j=1}^{d} \pi_j \frac{\lambda_i\lambda_j}{\mu_i+\mu_j} \int_u^\infty e^{\mu_i u-\mu_j w}(p_{ji}(w)-\pi_i)\mathrm{d}w,\end{aligned}$$

also entailing that

$$\lim_{t\to\infty} \mathbb{V}\mathrm{ar} M(t) = \sum_{i=1}^{d} \pi_i \frac{\lambda_i}{\mu_i} + 2\sum_{i=1}^{d}\sum_{j=1}^{d} \pi_i \frac{\lambda_i\lambda_j}{\mu_i+\mu_j} D_{ij}^{(\boldsymbol{\mu})}.$$

▷ In the second scaling, we replace $\boldsymbol{\lambda} \mapsto N\boldsymbol{\lambda}$ and $Q \mapsto N^\alpha Q$, for $\alpha > 0$. The FCLT under this scaling is proven in Sect. 6; we here find the corresponding mean and (co-)variance. It turns out that, for N large,

$$\mathbb{C}\mathrm{ov}\left(M^{(N)}(t), M^{(N)}(t+u)\right) \sim N\sum_{i=1}^{d} e^{-\mu_i u}\varrho_i^{(ii)}(t) + N^{2-\alpha}\sum_{i=1}^{d} e^{-\mu_i u}\varsigma_i^{(ii)}(t),$$

with $\varrho_i^{(ii)} := \pi_i\lambda_i/\mu_i$ and $\varrho_i^{(ii)}(t) := \varrho_i^{(ii)} \cdot (1-e^{-\mu_i t})$ and

$$\varsigma_i^{(ii)}(t) := \sum_{j=1}^{d} \frac{\lambda_i\lambda_j}{\mu_i+\mu_j}\left(1-e^{-(\mu_i+\mu_j)t}\right)\left(\pi_j D_{ji} + \pi_i D_{ij}\right).$$

We conclude that

$$\begin{aligned}\lim_{N\to\infty} \frac{\mathbb{C}\mathrm{ov}(M^{(N)}(t), M^{(N)}(t+u))}{N^{\max\{1,2-\alpha\}}} &= v^{(ii)}(t,u)\\ := \sum_{i=1}^{d} e^{-\mu_i u}&\left(\varsigma_i^{(ii)}(t)1_{\{\alpha\le 1\}} + \varrho_i^{(ii)}(t)1_{\{\alpha\ge 1\}}\right). \qquad (10)\end{aligned}$$

We observe that the same dichotomy applies as the one we have observed for Model I: for $\alpha > 1$ the number of jobs in the system behaves 'Poissonian', with mean and variance scaling essentially linearly with N, both with proportionality constant $\varrho^{(ii)}(t)$. For $\alpha < 1$, as seen earlier in e.g. Blom et al. (2015), the variance grows superlinearly with N, with a proportionality constant that involves the deviation matrix D.

5 Functional central limit theorem for Model I

In Sect. 3.2 we considered the covariance of the number of jobs in the system under a specific scaling: $\boldsymbol{\lambda} \mapsto N\boldsymbol{\lambda}$ and $Q \mapsto N^{\alpha}Q$, for $\alpha > 0$. In this section, we prove that for a given t the random variable $M^{(N)}(t)$ obeys a central limit theorem; moreover, we prove the stronger property that after centering and normalizing the process $M^{(N)}(t)$, there is weak convergence to a specific Gaussian process. We essentially adopt the methodology used in Huang et al. (2014); some steps that are fully analogous to those in Huang et al. (2014) are described concisely. In the sequel, we let $Z_i^{(N)}(t)$ be the indicator function of the event $\{J^{(N)}(t) = i\}$, where $J^{(N)}(t)$ is a Markov chain with transition rate matrix $N^{\alpha}Q$.

First observe that, with $P_1(\cdot)$ and $P_2(\cdot)$ two independent unit-rate Poisson processes, it is straightforward to see that $M^{(N)}(t)$ can be written as

$$M^{(N)}(t) = P_1\left(N\int_0^t \sum_{i=1}^{d} \lambda_i Z_i^{(N)}(s)\mathrm{d}s\right) - P_2\left(\int_0^t \sum_{i=1}^{d} \mu_i M^{(N)}(s) Z_i^{(N)}(s)\mathrm{d}s\right). \tag{11}$$

Now impose the following centering and normalization, with $\beta := \max\{1, 2-\alpha\}/2$,

$$\tilde{M}^{(N)}(t) := N^{-\beta}\left(M^{(N)}(t) - N\varrho^{(i)}(t)\right),$$

where $\varrho^{(i)}(t) := \varrho^{(i)}(1 - e^{-\mu_\infty t})$; the objective of this section is to establish the convergence of $\tilde{M}^{(N)}(\cdot)$ to a specific Gaussian process, essentially relying on the martingale central limit theorem; see for background on the martingale central limit theorem e.g. Jacod and Shiryayev (1987) and Whitt (2007).

It is first realized that, as a direct implication of (11), for some martingale $\kappa^{(N)}(\cdot)$,

$$\mathrm{d}M^{(N)}(t) = N\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}^{(N)}(t)\mathrm{d}t - \boldsymbol{\mu}^{\mathrm{T}}\mathbf{Z}^{(N)}(t)\, M^{(N)}(t)\mathrm{d}t + \mathrm{d}\kappa^{(N)}(t).$$

Then we rewrite this equation in terms of one for $\tilde{M}^{(N)}(t)$:

$$\begin{aligned}
\mathrm{d}\tilde{M}^{(N)}(t) &= N^{1-\beta}\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}^{(N)}(t)\mathrm{d}t - N^{-\beta}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{Z}^{(N)}(t)\, M^{(N)}(t)\mathrm{d}t \\
&\quad + N^{-\beta}\mathrm{d}\kappa^{(N)}(t) - N^{1-\beta}\left(\varrho^{(i)}\right)'(t)\mathrm{d}t \\
&= N^{1-\beta}\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}^{(N)}(t)\mathrm{d}t - \boldsymbol{\mu}^{\mathrm{T}}\mathbf{Z}^{(N)}(t)\, \tilde{M}^{(N)}(t)\mathrm{d}t - N^{1-\beta}\boldsymbol{\mu}^{\mathrm{T}}\mathbf{Z}^{(N)}(t)\varrho^{(i)}(t)\mathrm{d}t \\
&\quad + N^{-\beta}\mathrm{d}\kappa^{(N)}(t) - N^{1-\beta}\left(\varrho^{(i)}\right)'(t)\mathrm{d}t.
\end{aligned}$$

Following the ideas of Huang et al. (2014), we now introduce

$$Y^{(N)}(t) := \exp\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\zeta}^{(N)}(t)\right)\tilde{M}^{(N)}(t), \quad \text{where} \quad \boldsymbol{\zeta}^{(N)}(t) := \int_0^t \boldsymbol{Z}^{(N)}(s)\mathrm{d}s.$$

It thus follows that, using standard stochastic differentiation rules,

$$\mathrm{d}Y^{(N)}(t) = \exp\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\zeta}^{(N)}(t)\right)\left(N^{1-\beta}\left(\boldsymbol{\lambda} - \boldsymbol{\mu}\varrho^{(i)}(t)\right)^{\mathrm{T}}\boldsymbol{Z}^{(N)}(t)\mathrm{d}t\right.$$
$$\left. + N^{-\beta}\mathrm{d}\kappa^{(N)}(t) - N^{1-\beta}\left(\varrho^{(i)}\right)'(t)\mathrm{d}t\right).$$

Now observe that, from the definition of the function $\varrho^{(i)}(t)$, we find that

$$\left(\boldsymbol{\lambda} - \boldsymbol{\mu}\varrho^{(i)}(t)\right)^{\mathrm{T}}\boldsymbol{\pi} = \lambda_\infty e^{-\mu_\infty t} = \left(\varrho^{(i)}\right)'(t),$$

and hence it is obtained that

$$\mathrm{d}Y^{(N)}(t) = \exp\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\zeta}^{(N)}(t)\right)$$
$$\left(N^{1-\beta}\left(\boldsymbol{\lambda} - \boldsymbol{\mu}\varrho^{(i)}(t)\right)^{\mathrm{T}}\left(\boldsymbol{Z}^{(N)}(t) - \boldsymbol{\pi}\right)\mathrm{d}t + N^{-\beta}\mathrm{d}\kappa^{(N)}(t)\right).$$

We now analyze the two terms in the previous display separately.

- We first concentrate on the first term. In Huang et al. (2014), relying on the methodology developed in Jacod and Shiryayev (1987), it was shown that the following weak convergence holds:

  $$\int_0^{\cdot} N^{\alpha/2}\exp\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\zeta}^{(N)}(s)\right)\left(\boldsymbol{\lambda} - \boldsymbol{\mu}\varrho^{(i)}(s)\right)^{\mathrm{T}}\left(\boldsymbol{Z}^{(N)}(s) - \boldsymbol{\pi}\right)\mathrm{d}s \to \int_0^{\cdot} e^{\mu_\infty s}\mathrm{d}\mathcal{G}(s),$$

  where the stochastic process $\mathcal{G}(\cdot)$ is such that

  $$\langle\mathcal{G}\rangle_t = V(t) := \int_0^t \left(\boldsymbol{\lambda} - \boldsymbol{\mu}\varrho^{(i)}(s)\right)^{\mathrm{T}}\left(\mathrm{diag}\{\boldsymbol{\pi}\}D + D^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{\pi}\}\right)\left(\boldsymbol{\lambda} - \boldsymbol{\mu}\varrho^{(i)}(s)\right)\mathrm{d}s$$
  $$= 2\int_0^t \boldsymbol{\pi}^{\mathrm{T}}\left(\Lambda - \mathcal{M}\varrho^{(i)}(s)\right)D\left(\Lambda - \mathcal{M}\varrho^{(i)}(s)\right)\mathbf{1}\,\mathrm{d}s;$$

  cf. Eq. (6) [it is noted that in Huang et al. (2014) the background process was sped up by a factor N rather than N^α; this explains that there the growth rate $\sqrt{N}$ was found, while in our setup we have $N^{\alpha/2}$].

  Importantly, from the above we conclude that the full first term in $\mathrm{d}Y^{(N)}(t)$ behaves essentially proportional to $N^{1-\beta-\alpha/2}$, which converges to a constant if $\alpha \le 1$, and vanishes otherwise.

- We now consider the second term. We note that, recalling the fact that $P_1(\cdot)$ and $P_2(\cdot)$ are independent unit-rate Poisson processes in combination with standard properties for pure jump processes,

$$\frac{\mathrm{d}}{\mathrm{d}t}\langle\kappa^{(N)}\rangle_t = N\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}^{(N)}(t) + \boldsymbol{\mu}^{\mathrm{T}}\mathbf{Z}^{(N)}(t)\, M^{(N)}(t),$$

and consequently

$$\frac{1}{N}\langle\kappa^{(N)}\rangle_t = \int_0^t \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}^{(N)}(s)\mathrm{d}s + \int_0^t \boldsymbol{\mu}^{\mathrm{T}}\mathbf{Z}^{(N)}(s)\frac{M^{(N)}(s)}{N}\mathrm{d}s.$$

Using the ergodic theorem, the first integral in the right-hand side of the previous display converges to $\boldsymbol{\lambda}^{\mathrm{T}}\boldsymbol{\pi}\cdot t = \lambda_\infty t$. Likewise, the second integral converges to

$$\lim_{N\to\infty}\frac{1}{N}\int_0^t \sum_{i=1}^d \mu_i\,\mathbb{E}\left(M^{(N)}(s)1\{J(s)=i\}\right)\mathrm{d}s,$$

which, due to arguments similar to those underlying (5), turns out to equal

$$\int_0^t \sum_{i=1}^d \mu_i\pi_i\varrho^{(i)}(1-e^{-\mu_\infty s})\mathrm{d}s.$$

Hence $N^{-1}\langle\kappa^{(N)}\rangle_t$ converges, as $N\to\infty$, to

$$W(t) := \lambda_\infty t + \int_0^t \mu_\infty\varrho^{(i)}(1-e^{-\mu_\infty s})\mathrm{d}s.$$

We conclude from the above that $\kappa^{(N)}(\cdot)/\sqrt{N}$ converges to an appropriately scaled Brownian motion.

In addition, this second term in $\mathrm{d}Y^{(N)}(t)$ is essentially proportional to $N^{1/2-\beta}$, i.e., converging to a constant if $\alpha \geq 1$, and vanishes otherwise.

Summarizing, we have that $Y^{(N)}(t)$ converges weakly to a process $Y(t)$ which is the solution to the following stochastic differential equation:

$$\mathrm{d}Y(t) = \sqrt{V'(t)1_{\{\alpha\leq 1\}} + W'(t)1_{\{\alpha\geq 1\}}}\,\mathrm{d}B(t),$$

where we used the property that for a standard Brownian motion B and a differentiable function f, we have that $B(f(t))$ is equal in distribution to $\sqrt{f'(t)}\hat{B}(t)$, where $\hat{B}$ denotes another Brownian motion, but with the same distribution. Also, due to the ergodic theorem we have that $\exp\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\zeta}^{(N)}(t)\right)$ converges $\exp(\mu_\infty t)$. From the definition of $Y^{(N)}(t)$, we thus conclude the following weak convergence: $\tilde{M}^{(N)}(\cdot)\to\tilde{M}(\cdot)$, where $\tilde{M}(\cdot)$ solves the stochastic differential equation

$$\mathrm{d}\tilde{M}(t) = -\mu_\infty \tilde{M}(t)\mathrm{d}t + \sqrt{V'(t)1_{\{\alpha\leq 1\}} + W'(t)1_{\{\alpha\geq 1\}}}\,\mathrm{d}B(t),$$

for a standard Brownian motion $B(\cdot)$. Its solution is that the limiting process $\tilde{M}(\cdot)$ is a centered Gaussian process of the Ornstein–Uhlenbeck type, characterized by its covariance $v^{(i)}(t,u)$, as given in (7).

Theorem 5.1 *Consider Model* I. *As $N \to \infty$, the process $\tilde{M}^{(N)}(\cdot)$ converges weakly to a centered Gaussian process, with covariance structure $v^{(i)}(\cdot,\cdot)$ given in* (7).

6 Functional central limit theorem for Model II

We now shift our attention from Model I to Model II. Essentially the same approach can be followed, with an important difference being that now one has to keep track of the number of jobs present *of each type*, to be denoted by $M_i^{(N)}(t)$ for type i, where 'type' refers to the state the background process was in upon arrival of the job. We use an approach similar to the one used in the previous section, but it is noted that a viable alternative is to adapt the approach followed in Anderson et al. (2015) for the case that the departure rates are state-independent, to that of Model II.

As in the previous section, we start by writing the $M_i^{(N)}(t)$, for $i = 1,\ldots,d$ in terms of unit-rate Poisson processes; in self-evident notation, we now have

$$M_i^{(N)}(t) = P_{1,i}\left(N\int_0^t \lambda_i Z_i^{(N)}(s)\mathrm{d}s\right) - P_{2,i}\left(\int_0^t \mu_i M_i^{(N)}(s)\mathrm{d}s\right).$$

As before, we apply centering and normalization, in that we will study, recalling that $\varrho_i^{(ii)} := \pi_i\lambda_i/\mu_i$ and $\varrho_i^{(ii)}(t) := \varrho_i^{(ii)}\cdot(1-e^{-\mu_i t})$,

$$\tilde{M}_i^{(N)}(t) := N^{-\beta}\left(M_i^{(N)}(t) - N\varrho_i^{(ii)}(t)\right),$$

where, as in the previous section, $\beta := \max\{1, 2-\alpha\}/2$. Also we have that, for martingales $\kappa_i^{(N)}(t)$ (with $i = 1,\ldots,d$),

$$\mathrm{d}M_i^{(N)}(t) = N\lambda_i Z_i^{(N)}(t)\mathrm{d}t - \mu_i M_i^{(N)}(t)\mathrm{d}t + \mathrm{d}\kappa_i^{(N)}(t),$$

which we can express in terms of $\mathrm{d}\tilde{M}_i^{(N)}(t)$:

$$\begin{aligned}\mathrm{d}\tilde{M}_i^{(N)}(t) &= N^{-\beta}\mathrm{d}M_i^{(N)} - N^{1-\beta}\left(\varrho_i^{(ii)}\right)'(t)\,\mathrm{d}t\\ &= N^{1-\beta}\lambda_i Z_i^{(N)}(t)\mathrm{d}t - \mu_i\tilde{M}_i^{(N)}(t)\mathrm{d}t - N^{1-\beta}\mu_i\varrho_i^{(ii)}(t)\mathrm{d}t\\ &\quad + N^{-\beta}\mathrm{d}\kappa_i^{(N)}(t) - N^{1-\beta}\left(\varrho_i^{(ii)}\right)'(t)\,\mathrm{d}t.\end{aligned}$$

Using the definition of $\varrho_i^{(ii)}(t)$, after some calculus we eventually obtain the stochastic differential equation

$$\mathrm{d}\tilde{M}_i^{(N)}(t) = -\mu_i \tilde{M}_i^{(N)}(t)\mathrm{d}t + N^{1-\beta}\lambda_i \left(Z_i^{(N)}(t) - \pi_i\right)\mathrm{d}t + N^{-\beta}\mathrm{d}\kappa_i^{(N)}(t).$$

Mimicking the ideas used in the previous section, we study the last two terms appearing in the right-hand side of the previous display separately.

- We first concentrate on the middle term of the right-hand side of the previous display. To this end, we define

$$I_i^{(N)}(t) := \int_0^t \left(Z_i^{(N)}(s) - \pi_i\right)\mathrm{d}s.$$

  In e.g. Anderson et al. (2015, Prop. 3.2) it was shown that the following weak convergence holds:

$$N^{\alpha/2}\boldsymbol{I}^{(N)}(\cdot) \to \boldsymbol{B}(\cdot),$$

  where $\boldsymbol{B}(\cdot)$ denotes a zero-mean d-dimensional Brownian motion with covariance matrix $\mathrm{diag}\{\boldsymbol{\pi}\}D + D^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{\pi}\}$. The fact that this matrix is nonnegative definite has been proven in Huang et al. (2014, Prop. 3.2), and hence it allows a Cholesky decomposition. We, in addition, obtain the weak convergence of $\boldsymbol{H}^{(N)}(\cdot)$, with $H_i^{(N)}(t) := \lambda_i I_i^{(N)}(t)$, to a zero-mean d-dimensional Brownian motion with covariance matrix (and thus also allowing a Cholesky decomposition)

$$V := \Lambda\left(\mathrm{diag}\{\boldsymbol{\pi}\}D + D^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{\pi}\}\right)\Lambda. \tag{12}$$

  It also follows that this term behaves essentially proportional to $N^{1-\beta-\alpha/2}$; more specifically, it converges to a constant if $\alpha \leq 1$, and vanishes otherwise.
- We now consider the second term. We note that

$$\frac{\mathrm{d}}{\mathrm{d}t}\left\langle \kappa_i^{(N)}\right\rangle_t = N\lambda_i Z_i^{(N)}(t) + \mu_i M_i^{(N)}(t),$$

  and consequently

$$\frac{1}{N}\left\langle \kappa_i^{(N)}\right\rangle_t = \lambda_i \int_0^t Z_i^{(N)}(s)\mathrm{d}s + \mu_i \int_0^t \frac{M_i^{(N)}(s)}{N}\mathrm{d}s,$$

  which we can prove, using standard arguments (such as the ergodic theorem), to converge to

$$w_i(t) := \lambda_i \pi_i t + \mu_i \int_0^t \varrho_i^{(ii)}(1 - e^{-\mu_i s})\mathrm{d}s.$$

We thus find that $\kappa_i^{(N)}(\cdot)/\sqrt{N}$ converges to an appropriately scaled one-dimensional Brownian motion.

By doing similar steps for $\langle\kappa_i^{(N)}+\kappa_j^{(N)}\rangle_t$, for $i\neq j$, we find that the quadratic covariation between $\kappa_i^{(N)}(\cdot)/\sqrt{N}$ and $\kappa_j^{(N)}(\cdot)/\sqrt{N}$ equals 0. We conclude from the above that $\boldsymbol{\kappa}^{(N)}(\cdot)/\sqrt{N}$ converges to an appropriately scaled d-dimensional Brownian motion; the variance of component i at time t is $w_i(t)$, and the covariances are all 0.

It also follows that this term is essentially proportional to $N^{1/2-\beta}$, which is a constant if $\alpha\geq 1$, and vanishes otherwise.

Define $\tilde{\boldsymbol{M}}(t) := (\tilde{M}_1(t),\dots,\tilde{M}_d(t))^{\mathrm{T}}$; we also write

$$W_i(t) := \lambda_i\pi_i + \mu_i\varrho_i^{(ii)}(t) = 2\lambda_i\pi_i - \lambda_i\pi_i e^{-\mu_i t}.$$

Based on the above, we obtain that $\tilde{M}^{(N)}$ converges as $N\to\infty$ to the solution $\tilde{\boldsymbol{M}}(t)$ of the stochastic differential equation, in self-evident notation,

$$\mathrm{d}\tilde{\boldsymbol{M}}(t) = -\mathcal{M}\,\tilde{\boldsymbol{M}}(t)\mathrm{d}t + \sqrt{V1_{\{\alpha\leq 1\}}+\mathrm{diag}\{\boldsymbol{W}(t)\}1_{\{\alpha\geq 1\}}}\,\mathrm{d}\boldsymbol{B}(t).$$

From this stochastic differential equation it follows by applying standard techniques that the resulting limiting process is a centered Gaussian process, with

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(\tilde{M}_i(t),\tilde{M}_j(t)\right) &= e^{-\mu_i t-\mu_j t}\int_0^t e^{\mu_i s+\mu_j s}\lambda_i\lambda_j\left(\pi_i D_{ij}+\pi_j D_{ji}\right)\mathrm{d}s\\ &= \frac{\lambda_i\lambda_j}{\mu_i+\mu_j}\left(1-e^{-(\mu_i+\mu_j)t}\right)\left(\pi_i D_{ij}+\pi_j D_{ji}\right)\end{aligned}$$

if $\alpha\leq 1$. If $\alpha\geq 1$, on the contrary, the covariance is 0 if $i\neq j$, and $\varrho_i^{(ii)}(t)$ if $i=j$.

Now consider the limiting distribution of the *total* population of the system; from the above, it immediately follows that we have the weak convergence

$$\sum_{i=1}^d \tilde{M}_i^{(N)}(t)\to\tilde{M}(t) := \sum_{i=1}^d \tilde{M}_i(t),$$

which is a centered Gaussian process of the Ornstein–Uhlenbeck type, characterized by its covariance $v^{(ii)}(t,u)$, as given in (10).

Theorem 6.1 *Consider Model* II. *As $N\to\infty$, the process $\tilde{M}^{(N)}(\cdot)$ converges weakly to a centered Gaussian process with covariance structure $v^{(ii)}(\cdot,\cdot)$ given in* (9).

7 Numerical experiments

In this section, we illustrate the results with two plots. In all cases, we consider a Model I scenario with a two-state Markov chain. We assume that $q_{12} = q_{21} = 5$, and $\boldsymbol{\lambda} = [20\,10]$.

In the first figure, Fig. 2, we plot the covariance of a system starting in stationarity for two scenarios. In the first scenario (dashed line) $\boldsymbol{\mu} = [2\ 1]$, whereas in the second scenario (full line) $\boldsymbol{\mu} = [1\ 2]$.

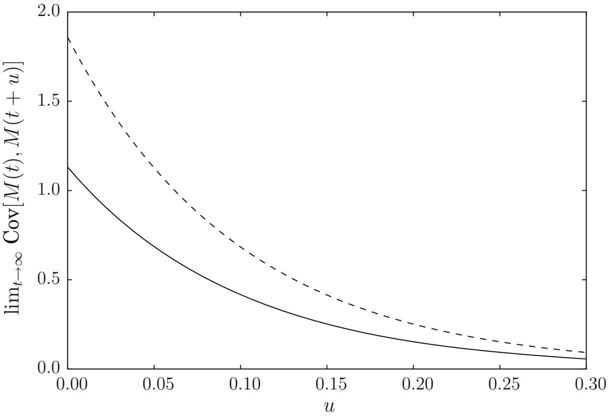

Fig. 2 The limiting covariance versus time for $\boldsymbol{\mu} = [2\ 1]$ (*dashed line*), and, $\boldsymbol{\mu} = [1\ 2]$ (*full line*)

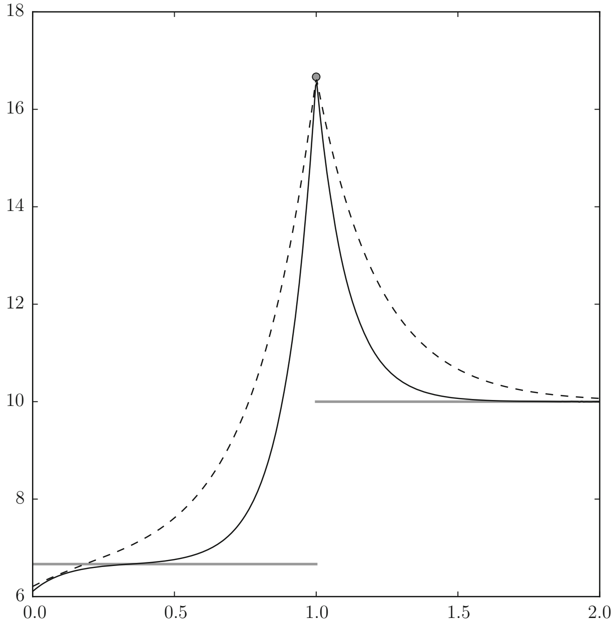

Fig. 3 The scaled variance of $M^{(N)}$

For the second plot, we assume $\boldsymbol{\mu} = [1\ 2]$ and apply the scaling $\boldsymbol{\lambda} \mapsto N\boldsymbol{\lambda}$ and $Q \mapsto N^{\alpha} Q$. Figure 3 shows the stationary variance of the number of jobs. We divide this variance by the theoretically predicted growth factor $N^{2\gamma}$, and plot it against α. The dashed line corresponds to $N = 100$, the full line to $N = 100{,}000$. We plot the limiting curve in gray.

Acknowledgments The authors thank Peter Spreij (Korteweg-de Vries Institute for Mathematics, University of Amsterdam) for useful suggestions.

Appendix 1

In this appendix we characterize the probability generating functions $\Xi_{ij}(\cdot,\cdot,\cdot,\cdot)$ by setting up a system of partial differential equations. The starting point for this is the system of Kolmogorov equations related to the transient probabilities of the number of jobs present (jointly with the background state) at two time epochs:

$$p_{ij}(m,n,t,u) := \mathbb{P}(M(t) = m, M(t+u) = n, J(t) = i, J(t+u) = j);$$

we suppress the u as this is held fixed for the moment. Standard arguments from Markov chain theory immediately yield the equations, for $i, j \in \{1,\ldots,d\}$, $m, n \in \{0, 1, 2\ldots\}$, and $q_i := -q_{ii}$,

$$\begin{aligned} p_{ij}(m,n,t+\Delta t,u) = {} & p_{ij}(m,n,t,u)\left(1 - \left(\lambda_i + \lambda_j + m\,\mu_i + n\,\mu_j + q_i + q_j\right)\Delta t\right) \\ & + p_{ij}(m-1,n,t,u)\lambda_i\Delta t \ + \ p_{ij}(m,n-1,t,u)\lambda_j\Delta t \\ & + p_{ij}(m+1,n,t,u)\,(m+1)\mu_i\Delta t \\ & + p_{ij}(m,n+1,t,u)\,(n+1)\mu_j\Delta t \\ & + \sum_{k\neq i} p_{kj}(m,n,t,u)q_{ki}\Delta t \ + \ \sum_{k\neq j} p_{ik}(m,n,t,u)q_{kj}\Delta t \\ & + o(\Delta t); \end{aligned}$$

here $p_{ij}(-1,n,t,u)$ and $p_{ij}(m,-1,t,u)$ are to be understood as 0. As a consequence, with $p'_{ij}(m,n,t,u)$ denoting the derivative of $p_{ij}(m,n,t,u)$ with respect to t, it is readily obtained that the transient probabilities satisfy the following system of (ordinary) differential equations:

$$\begin{aligned} p'_{ij}(m,n,t,u) = {} & \lambda_i\left(p_{ij}(m-1,n,t,u) - p_{ij}(m,n,t,u)\right) \\ & + \lambda_j\left(p_{ij}(m,n-1,t,u) - p_{ij}(m,n,t,u)\right) \\ & + \mu_i\left((m+1)p_{ij}(m+1,n,t,u) - mp_{ij}(m,n,t,u)\right) \\ & + \mu_j\left((n+1)p_{ij}(m,n+1,t,u) - np_{ij}(m,n,t,u)\right) \\ & + \sum_{k=1}^{d} p_{kj}(m,n,t,u)q_{ki} \ + \ \sum_{k=1}^{d} p_{ik}(m,n,t,u)q_{kj}. \end{aligned}$$

Our goal is to transform these differential equations into a system of partial differential equations for the corresponding probability generating functions. To this end, multiply both sides of the equation by $z^m w^n$, and sum over $m, n = 0, 1, 2, \ldots$. This results in the following equation:

$$\begin{aligned}\frac{\partial}{\partial t}\Xi_{ij}(z,w,t,u) &= \big(\lambda_i(z-1)+\lambda_j(w-1)\big)\cdot\Xi_{ij}(z,w,t,u)\\ &\quad -\mu_i(z-1)\frac{\partial}{\partial z}\Xi_{ij}(z,w,t,u)-\mu_j(w-1)\frac{\partial}{\partial w}\Xi_{ij}(z,w,t,u)\\ &\quad +\sum_{k=1}^{d}\Xi_{kj}(z,w,t,u)q_{ki}\;+\sum_{k=1}^{d}\Xi_{ik}(z,w,t,u)q_{kj},\end{aligned}$$

which in matrix notation coincides with Eq. (1).

Appendix 2

As will be proven in "Appendix 3" below, the statement (5) can be refined to, for some constant κ (whose precise form is irrelevant here),

$$\mathbb{E}M^{(N)}(t)=\left(N\varrho^{(i)}+N^{1-\alpha}\kappa\right)\left(1-e^{-\mu_\infty t}\right)+o(1). \tag{13}$$

Likewise, for any $x \geq 0$,

$$\mathbb{E}\left(M^{(N)}(t)\mid M^{(N)}(0)=x\right)=\left(N\varrho^{(i)}+N^{1-\alpha}\kappa\right)\left(1-e^{-\mu_\infty t}\right)+xe^{-\mu_\infty t}+o(1). \tag{14}$$

In the sequel we write $a(N) := N\varrho^{(i)} + N^{1-\alpha}\kappa$. Applying an elementary time shift, we obtain that

$$\begin{aligned}&\mathbb{E}\left(M^{(N)}(t)M^{(N)}(t+u)\right)\\ &\quad=\int_0^\infty \mathbb{E}\left(M^{(N)}(t)M^{(N)}(t+u)\mid M^{(N)}(t)=x\right)\mathbb{P}\left(M^{(N)}(t)\in \mathrm{d}x\right)\\ &\quad=\int_0^\infty x\mathbb{E}\left(M^{(N)}(u)\mid M^{(N)}(0)=x\right)\mathbb{P}\left(M^{(N)}(t)\in \mathrm{d}x\right).\end{aligned}$$

By plugging in (14), the expression in the last display equals

$$\begin{aligned}&\int_0^\infty x\left(a(N)(1-e^{-\mu_\infty u})+xe^{-\mu_\infty u}+o(N^{1-\alpha})\right)\mathbb{P}\left(M^{(N)}(t)\in \mathrm{d}x\right)\\ &\quad=\left(a(N)(1-e^{-\mu_\infty u})+o(N^{1-\alpha})\right)\mathbb{E}M^{(N)}(t)+e^{-\mu_\infty u}\,\mathbb{E}\left(\left(M^{(N)}(t)\right)^2\right)\\ &\quad=(\xi_N(u)+o(N^{1-\alpha}))(\xi_N(t)+o(N^{1-\alpha}))+e^{-\mu_\infty u}\,\mathbb{E}\left(\left(M^{(N)}(t)\right)^2\right),\end{aligned}$$

where $\xi_N(u) := a(N)(1-e^{-\mu_\infty u})$. Now note that, due to the computations underlying (Blom et al. 2015, Thm. 2), with $\varrho^{(i)}(t)$ and $\varsigma^{(i)}(t)$ as defined in Section 3.2, and with $\psi(N) = o(N^{\max\{1,2-\alpha\}})$, that

$$\mathbb{E}\left((M^{(N)}(t))^2\right) - \left(\mathbb{E}M^{(N)}(t)\right)^2 = N^{2-\alpha}\varsigma^{(i)}(t)1_{\{\alpha\leq 1\}} + N\varrho^{(i)}(t)1_{\{\alpha\geq 1\}} + \psi(N).$$

We now turn our attention to characterizing the covariance between $M^{(N)}(t)$ and $M^{(N)}(t+u)$. Based on the above we have

$$\begin{aligned}
&\mathbb{C}\mathrm{ov}(M^{(N)}(t), M^{(N)}(t+u)) = \mathbb{E}\left(M^{(N)}(t)M^{(N)}(t+u)\right)\\
&\quad -\mathbb{E}\left(M^{(N)}(t)\right)\mathbb{E}\left(M^{(N)}(t+u)\right)\\
&\quad = (\xi_N(u) + o(N^{1-\alpha}))(\xi_N(t) + o(N^{1-\alpha})) + e^{-\mu_\infty u}\left(\xi_N(t) + o(N^{1-\alpha})\right)^2\\
&\quad\quad + e^{-\mu_\infty u}\left(N^{2-\alpha}\varsigma^{(i)}(t)1_{\{\alpha\leq 1\}} + N\varrho^{(i)}(t)1_{\{\alpha\geq 1\}} + \psi(N)\right)\\
&\quad\quad - \left(\xi_N(t) + o(N^{1-\alpha})\right)\left(\xi_N(t+u)) + o(N^{1-\alpha})\right).
\end{aligned}$$

A direct computation now yields that the zero-order terms cancel, and that we end up with (7).

Appendix 3

In this appendix, we establish (13). The idea is to manipulate differential equation (4), so as to characterize the behavior of $\boldsymbol{m}(t)$ for N large. The first step is to postmultiply the equation by the fundamental matrix $F := D + \Pi$. We obtain, multiplying the equation by $N^{-\alpha}$ as well,

$$\boldsymbol{m}(t) = \boldsymbol{m}(t)\Pi - N^{-\alpha}\boldsymbol{m}(t)\mathcal{M}F + N^{1-\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda F - N^{-\alpha}\boldsymbol{m}'(t)F.$$

Iterate this equation once, we obtain

$$\begin{aligned}
\boldsymbol{m}(t) &= \boldsymbol{m}(t)\Pi - N^{-\alpha}\boldsymbol{m}(t)\mathcal{M}\Pi + N^{1-\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda\Pi - N^{-\alpha}\boldsymbol{m}'(t)\Pi\\
&\quad - N^{-\alpha}\boldsymbol{m}(t)\Pi\mathcal{M}F - N^{1-2\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda F\mathcal{M}F + N^{1-\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda F - N^{-\alpha}\boldsymbol{m}'(t)\Pi + o(N^{-\alpha}).
\end{aligned}$$

Iterating once again to replace all occurrences of $\boldsymbol{m}(t)$ by $\boldsymbol{m}(t)\Pi$, we obtain, with $\boldsymbol{n}(t) := \boldsymbol{m}(t)\Pi$,

$$\begin{aligned}
\boldsymbol{m}(t) &= \boldsymbol{n}(t) - N^{-\alpha}\boldsymbol{n}(t)\Pi\mathcal{M}\Pi - N^{1-2\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda F\mathcal{M}\Pi + N^{1-\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda\Pi - N^{-\alpha}\boldsymbol{n}'(t)\\
&\quad - N^{-\alpha}\boldsymbol{n}(t)\Pi\mathcal{M}F - N^{1-2\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda F\mathcal{M}F + N^{1-\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda F - N^{-\alpha}\boldsymbol{n}'(t) + o(N^{-\alpha}).
\end{aligned}$$

Now postmultiply this equation by $N^{\alpha}\Pi\mathbf{1}$. Recalling that $F\mathbf{1} = \mathbf{1}$ and $\Pi\mathbf{1} = \mathbf{1}$, we obtain

$$\boldsymbol{n}'(t)\mathbf{1} = -\boldsymbol{n}(t)\Pi\mathcal{M}\mathbf{1} + N\boldsymbol{\pi}^{\mathrm{T}}\Lambda\mathbf{1} - N^{1-\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda F\mathcal{M}\mathbf{1} + o(1).$$

which leads to, using $\boldsymbol{n}(t)\mathbf{1} := \phi(t)$ and $\Pi = \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$,

$$\phi'(t) = -\mu_{\infty}\phi(t) + N\lambda_{\infty} - N^{1-\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda F\mathcal{M}\mathbf{1} + o(1).$$

We find that, with $\phi(0) = 0$,

$$\mathbb{E}M^{(N)}(t) = \left(N\frac{\lambda_{\infty}}{\mu_{\infty}} - \frac{1}{\mu_{\infty}}N^{1-\alpha}\boldsymbol{\pi}^{\mathrm{T}}\Lambda F\mathcal{M}\mathbf{1}\right)(1 - e^{-\mu_{\infty}t}) + o(1).$$

Appendix 4

We first focus on the first term in the right hand side of (8). To this end, consider the following decomposition:

$$M(t) := M^{(1)}(t, t+u) + M^{(2)}(t, t+u),$$
$$M(t+u) := M^{(2)}(t, t+u) + M^{(3)}(t, t+u),$$

where $M^{(1)}(t, t+u)$ are the jobs that arrived in $[0, t)$ that are still present at time t but have left at time $t+u$, $M^{(2)}(t, t+u)$ the jobs that have arrived in $[0, t)$ that are still present at time $t+u$, and $M^{(3)}(t, t+u)$ the jobs that have arrived in $[t, t+u)$ that are still present at time $t+u$. Observe that, conditional on J, these three random quantities are independent. As a result,

$$\mathbb{E}(\mathbb{C}\mathrm{ov}(M(t), M(t+u)) \mid J)) = \mathbb{E}(\mathbb{V}\mathrm{ar}\, M^{(2)}(t, t+u) \mid J).$$

Mimicking the arguments used in D'Auria (2008), it is immediate that $M^{(2)}(t, t+u)$, conditional on J, has a Poisson distribution with parameter

$$\int_0^t \lambda_{J(s)} e^{-\mu_{J(s)}(t+u-s)}\mathrm{d}s.$$

We conclude that

$$\begin{aligned}\mathbb{E}(\mathbb{C}\mathrm{ov}(M(t), M(t+u)) \mid J)) &= \mathbb{E}\left(\int_0^t \lambda_{J(s)} e^{-\mu_{J(s)}(t+u-s)}\mathrm{d}s\right)\\ &= \sum_{i=1}^{d}\pi_i\lambda_i\int_0^t e^{-\mu_i(t+u-s)}\mathrm{d}s\\ &= \sum_{i=1}^{d}\pi_i\frac{\lambda_i}{\mu_i}\left(1 - e^{-\mu_i t}\right)e^{-\mu_i u}.\end{aligned}$$

Now analyze the second term in the right hand side of (8). First observe that it can be written as

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

This decomposes into $I_1 + I_2$, where

$$\begin{aligned} I_1 &:= \sum_{i=1}^{d}\sum_{j=1}^{d} \lambda_i\lambda_j \mathscr{K}_{ij}, \quad \text{where } \mathscr{K}_{ij} \\ &:= \int_0^t \int_0^s e^{-\mu_i(t-r)} e^{-\mu_j(t+u-s)} \pi_i \left(p_{ij}(s-r) - \pi_j\right) \mathrm{d}r\mathrm{d}s, \\ I_2 &:= \sum_{i=1}^{d}\sum_{j=1}^{d} \lambda_i\lambda_j \mathscr{L}_{ij}, \quad \text{where } \mathscr{L}_{ij} \\ &:= \int_0^t \int_s^{t+u} e^{-\mu_i(t-r)} e^{-\mu_j(t+u-s)} \pi_j \left(p_{ji}(r-s) - \pi_i\right) \mathrm{d}r\mathrm{d}s. \end{aligned}$$

Let us first evaluate $\mathscr{K}_{ij} \equiv \mathscr{K}_{ij}(t,u)$. To this end, substitute $w := s - r$ (i.e., replace r by $s - w$), and then interchange the order of integration, so as to obtain

$$\mathscr{K}_{ij} = e^{-\mu_j(t+u)}\pi_i \int_0^t \left(\int_w^t e^{(\mu_i+\mu_j)s}\mathrm{d}s\right) e^{-\mu_i(t+w)} \left(p_{ij}(w) - \pi_j\right)\mathrm{d}w.$$

Performing the inner integral (i.e., the one over s) leads to

$$\mathscr{K}_{ij} = \frac{1}{\mu_i+\mu_j} e^{-\mu_j(t+u)}\pi_i \int_0^t \left(e^{-\mu_i w+\mu_j t} - e^{-\mu_i t+\mu_j w}\right)\left(p_{ij}(w) - \pi_j\right)\mathrm{d}w.$$

For $\mathscr{L}_{ij} \equiv \mathscr{L}_{ij}(t,u)$, again by a substitution and by interchanging the order of integration, we obtain $\mathscr{L}_{ij}^{(1)} + \mathscr{L}_{ij}^{(2)}$, where

$$\begin{aligned} \mathscr{L}_{ij}^{(1)} &:= e^{-\mu_j(t+u)}\pi_j \int_0^u \left(\int_0^t e^{(\mu_i+\mu_j)s}\mathrm{d}s\right) e^{-\mu_i(t-w)}\left(p_{ji}(w) - \pi_i\right)\mathrm{d}w, \\ \mathscr{L}_{ij}^{(2)} &:= e^{-\mu_j(t+u)}\pi_j \int_u^{t+u} \left(\int_0^{t+u-w} e^{(\mu_i+\mu_j)s}\mathrm{d}s\right) e^{-\mu_i(t-w)}\left(p_{ji}(w) - \pi_i\right)\mathrm{d}w, \end{aligned}$$

which reduce to

$$\begin{aligned} \mathscr{L}_{ij}^{(1)} &:= \frac{1}{\mu_i+\mu_j} e^{-\mu_j(t+u)}\pi_j \left(e^{\mu_j t} - e^{-\mu_i t}\right) \int_0^u e^{\mu_i w}\left(p_{ji}(w) - \pi_i\right)\mathrm{d}w, \\ \mathscr{L}_{ij}^{(2)} &:= \frac{1}{\mu_i+\mu_j} e^{-\mu_i t}\pi_j \int_u^{t+u} \left(e^{\mu_i(t+u)-\mu_j w} - e^{\mu_i w-\mu_j(t+u)}\right)\left(p_{ji}(w) - \pi_i\right)\mathrm{d}w. \end{aligned}$$

Now Eq. (9) follows.

References

Anderson D, Blom J, Mandjes M, Thorsdottir H, de Turck K (2015) A functional central limit theorem for a Markov-modulated infinite-server queue. Methodol Comput Appl Probab. doi:10.1007/s11009-014-9405-8

Blom J, Mandjes M (2013) A large-deviations analysis of Markov-modulated inifinite-server queues. Oper Res Lett 41:220–225

Blom J, de Turck K, Mandjes M (2013a) A central limit theorem for Markov-modulated infinite-server queues. In: Dudin A, de Turck K (eds) Proceedings ASMTA 2013. Lecture Notes in Computer Science (LNCS) series, vol 7984, Ghent, Belgium, pp 81–95

Blom J, de Turck K, Mandjes M (2013b) Rare event analysis of Markov-modulated infinite-server queues: a Poisson limit. Stoch Models 29:463–474

Blom J, Kella O, Mandjes M, de Turck K (2014a) Tail asymptotics of a Markov-modulated infinite-server queue. Queueing Syst 78:337–357

Blom J, Kella O, Mandjes M, Thorsdottir H (2014b) Markov-modulated infinite server queues with general service times. Queueing Syst 76:403–424

Blom J, de Turck K, Mandjes M (2015) Analysis of Markov-modulated infinite-server queues in the central-limit regime. Probab Eng Inf Sci, FirstView, http://journals.cambridge.org/article_S026996481500008X

D'Auria B (2008) M/M/∞ queues in semi-Markovian random environment. Queueing Syst 58:221–237

Huang G, Jansen HM, Mandjes M, Spreij P, de Turck K (2014) Markov-modulated Ornstein–Uhlenbeck processes. Adv Appl Probab 48(1):235–254

Jacod J, Shiryayev A (1987) Limit theorems for stochastic processes. Springer, Berlin

Keilson J, Servi L (1993) The matrix M/M/∞ system: retrial models and Markov modulated sources. Adv Appl Probab 25:453–471

O'Cinneide C, Purdue P (1986) The M/M/∞ queue in a random environment. J Appl Probab 23:175–184

Whitt W (2007) Proofs of the martingale FCLT. Probab Surv 4:268–302