Queueing Syst (2014) 76:403–424
DOI 10.1007/s11134-013-9368-4

Markov-modulated infinite-server queues with general service times

J. Blom · O. Kella · M. Mandjes · H. Thorsdottir

Received: 6 January 2013 / Revised: 17 May 2013 / Published online: 5 July 2013

Abstract This paper analyzes several aspects of the Markov-modulated infinite-server queue. In the system considered (i) particles arrive according to a Poisson process with rate λ_i when an external Markov process ("background process") is in state i, (ii) service times are drawn from a distribution with distribution function $F_i(\cdot)$ when the state of the background process (as seen at arrival) is i, (iii) there are infinitely many servers. We start by setting up explicit formulas for the mean and variance of the number of particles in the system at time $t \geq 0$, given the system started empty. The special case of exponential service times is studied in detail, resulting in a recursive scheme to compute the moments of the number of particles at an exponentially distributed time, as well as their steady-state counterparts. Then we consider an asymptotic regime in which the arrival rates are sped up by a factor N, and the transition times by a factor $N^{1+\varepsilon}$ (for some $\varepsilon > 0$). Under this scaling it turns out that the number of customers at time $t \geq 0$ obeys a central limit theorem; the convergence of the finite-dimensional distributions is proven.

J. Blom (✉) · M. Mandjes · H. Thorsdottir
CWI, Amsterdam, The Netherlands
e-mail: joke.blom@cwi.nl

O. Kella
Department of Statistics, The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: Offer.Kella@huji.ac.il

M. Mandjes · H. Thorsdottir
Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands
e-mail: M.R.H.Mandjes@uva.nl

H. Thorsdottir
e-mail: halldora@cwi.nl

Keywords Markov-modulated Poisson process · General service times · Queues · Infinite-server systems · Markov modulation · Laplace transforms · Fluid and diffusion scaling

Mathematics Subject Classification Primary: 60K25 · 60K37 · Secondary: 60F05

1 Introduction

Owing to its wide applicability and its attractive mathematical features, the infinite-server queue has been intensively studied. Such a system describes units of work, e.g., particles or customers, arriving at a resource, that stay present for some random duration that is independent of other customers (in that there is no waiting). In the special case that these customers arrive according to a Poisson process with rate λ, and the sojourn times are i.i.d. random variables with finite mean $1/\mu$—a system commonly referred to as the M/G/∞ queue—it is known that the stationary number of particles in the system has a Poisson distribution with mean λ/μ. Also the transient behavior of such an M/G/∞ queue is well understood; e.g. [24, p. 355].

When relaxing the model assumptions mentioned above, several interesting variants arise. In one branch of the literature, for instance, attention has been paid to the case of renewal (rather than Poisson) arrivals [10,11]. In the present paper, however, we consider a variant in which we introduce some sort of "burstiness" in the arrivals and service times, using the concept of *Markov modulation*. This means that both the arrival process and the service-time distributions are driven by an external Markov process ("background process"), in the following manner. Let $X(t)$ denote an irreducible continuous-time Markov process defined on a finite state space $\{1, \ldots, d\}$. When $X(t) = i$, then the (Poissonian) arrival rate at time t equals λ_i, where $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_d)$ is a vector with nonnegative entries. In addition, it is assumed that the time a particle remains in the system, the service time, has some general distribution with distribution function $F_i(\cdot)$ that depends on the state of the background process as seen upon arrival by the particle.

The resulting model could be called a *Markov-modulated* M/G/∞ *queue*, or an infinite-server queue in a Markov-modulated environment. This type of system is relevant in a broad variety of application domains, ranging from telecommunication networks to biology. The rationale behind using this model in a communication networks setting is that the arrival rate and service times of customers may vary during the day, or on shorter timescales. In the biological context, one could think of mRNA strings being transcribed and degraded in a cell, where these transcriptions typically tend to occur in a clustered fashion; the proposed model captures the key characteristics of this mechanism well, as argued in [23].

A variety of results exist on Markov-modulated single- and many-server queues, whereas the literature on their infinite-server counterpart is, surprisingly, considerably scarcer. In the case of a single server, the key result is that the stationary distribution of the number of customers is of matrix-geometric form [17], so this system can be viewed as a "matrix generalization" of the normal M/M/1 queue where the stationary distribution is scalar-geometric. In [20] the stationary distribution for the case of infinitely many servers is considered; the results are in terms of the factorial moments

of the number of customers. More particularly, it is shown that the corresponding distribution is *not* of matrix-Poisson type; in other words: this system is not the "matrix generalization" of the M/M/∞, which has a scalar-Poisson distribution. A somewhat more general model that includes retrials has been studied in [13].

The case of Markov-modulated *renewal* (rather than Poisson) arrivals, but exponential service times, is covered in [18]. Related results can be found in [16] as well, where special attention is paid to the autocorrelations in infinite-server systems of various types. Steady-state results for the infinite-server queue with modulated service rates have been derived in [2]. Falin [8] furthermore considers the simultaneous modulation of arrival and service rates and finds the mean number of customers in steady-state.

It should also be noted that introducing burstiness using a Markovian background process is by no means the only way to incorporate a nonhomogeneous arrival rate. Willmot and Drekic [25] apply bulk arrivals with a random bulk size, whereas Economou and Fakinos [7] study arrivals generated by a compound Poisson process, both to find the transient distribution of the number of customers in the system using a generating functions based approach.

D'Auria [5] studies the same model as we do in the present paper. Among several other results, he finds a recursion for the factorial moments of the stationary number of particles in the system. A key observation in his analysis is that the number of particles present has, in stationarity, a Poisson distribution with random parameter. Fralix and Adan [9] focus on the situation that the service times have specific phase-type distributions. In Hellings *et al.* [12] it was shown that if the transition times of the background process are sped up by a factor N, then the arrival process tends (as $N \to \infty$) to a Poisson process; the queue under consideration then essentially behaves as an M/G/∞ system.

While the above results focus on Markov-modulated infinite-server queues in stationarity, there are considerably fewer results on the associated transient behavior. In [3], we studied both the transient and stationary behavior of a model similar to the one studied in the present paper, viz. the one with exponential service times and a Markovian background process with deterministic transition times. The main focus of [3] lies on specific time scalings. In the first scaling, just the background process' transition times are sped up by a factor N; then it turns out that the distribution of the resulting queueing system converges to that of an appropriate M/M/∞ queue (which has, in steady-state, a Poisson distribution). In the second scaling, the background process jumps at a faster rate than the arrival process: the arrival rates are scaled by a factor N and the transition times by a factor $N^{1+\varepsilon}$ for some $\varepsilon > 0$. Under this scaling a central limit result was proven, for both the transient and stationary distribution.

The main contributions of our paper are the following. In the first place we develop in Sect. 2 expressions for the transient mean and variance for the number of particles in the system at time $t \geq 0$. For exponential service times the resulting expressions simplify considerably. In Sect. 3 we exclusively consider the special case of exponential service times: we develop a differential equation that describes the moment generating function of the number of particles in the system, and show how this differential equation facilitates the computation of moments (at an exponentially distributed time epoch, as well as in steady-state). This section also includes a recursive scheme to compute the higher moments. Section 4 considers one of the scalings studied in [3]: the arrival rates

λ_i are replaced by $N\lambda_i$, while the transition times of the background Markov process are sped up by a factor $N^{1+\varepsilon}$, for some $\varepsilon > 0$, where N grows large. The objective is to prove a central limit theorem for the number of particles in the system in a finite-dimensional setting, that is, at multiple points in time. The result is established by first setting up a system of differential equations for the number of particles in the system at multiple points in time in the *non*-scaled system, then applying the scaling, and then deriving (using Taylor approximations) a limiting differential equation (as $N \to \infty$) which eventually provides us with the claimed multivariate central limit theorem. Finally, Sect. 5 contains examples demonstrating analytically and numerically the results from Sects. 3 and 4.

2 General results

In full detail, the model can be described as follows. Consider an irreducible continuous-time Markov process $X(t)$ on a finite state space $\{1, \ldots, d\}$, with $d \in \mathbb{N}$. $X(t)$, often referred to as the *background process*, has a transition rate matrix given by $Q = \left(q_{ij}\right)_{i,j=1}^{d}$. The steady-state distribution of $X(t)$ is given by $\boldsymbol{\pi}$, being a d-dimensional vector with non-negative entries summing to 1, solving $\boldsymbol{\pi} Q = \mathbf{0}$. Define $q_i := -q_{ii} = \sum_{j \neq i} q_{ij}$.

Now consider the embedded discrete-time Markov chain that corresponds to the jump epochs of $X(t)$. It has a probability transition matrix $P = \left(p_{ij}\right)_{i,j=1}^{d}$, with diagonal elements equalling 0 and $p_{ij} := q_{ij}/q_i$. Let $\hat{\pi}_i$ be the stationary probability vector at the jump epochs of $X(t)$; it solves (after normalization to 1) the linear system $\hat{\boldsymbol{\pi}} D_Q^{-1} Q = 0$, with $D_Q := \mathrm{diag}\{\boldsymbol{q}\}$. The time spent by $X(t)$ in state i, denoted T_i, is referred to as *transition time*. T_i has an exponential distribution with mean $1/q_i$. There is the following obvious relation between $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$:

$$\pi_i := \frac{\hat{\pi}_i \mathbb{E}T_i}{\sum_{j=1}^{d} \hat{\pi}_j \mathbb{E}T_j} = \frac{\hat{\pi}_i / q_i}{\sum_{j=1}^{d} \hat{\pi}_j / q_j}.$$

While the process $X(t)$ is in state i, particles arrive according to a Poisson process with rate $\lambda_i \geq 0$, for $i = 1, \ldots, d$. The service times are assumed to be i.i.d. samples distributed as a random variable B_i with mean $1/\mu_i$ if the client was generated when the background process was in state i; the corresponding distribution function is $F_i(x) := \mathbb{P}(B_i \leq x)$, with $x \geq 0$. The service times are independent of the background process $X(t)$ and the arrival process. The system we consider is an *infinite-server queue*, meaning that each particle stays in the system just for the duration of its service time (that is, there is no waiting). In the rest of this section we focus on analyzing the probabilistic properties of the number of particles in the system at given points in time, starting empty. It is assumed that the background process is in stationarity at time 0.

We start by considering a somewhat different model than the one introduced above, where the relation with our model becomes clear soon. Consider an M/G/∞ queue with (i) a *nonhomogeneous* arrival process with rate function $\lambda(\cdot)$ (such that the Poissonian arrival rate is $\lambda(s)$ at time s), and (ii) a *time dependent* distribution function $F(s, \cdot)$

(to be interpreted as the probability that a customer that arrives at time s leaves before time $t+s$ is $F(s,t)$). Observe that, conditional on the event that there are n arrivals by time t, the joint distribution of the arrival times is that of the order statistics taken from independent random variables with density

$$\frac{\lambda(s)}{\Lambda(t)} 1_{[0,t]}(s),$$

where $\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$. It now follows that if $M(t)$ is the number of particles in the system at time t, starting with an empty system, then with $\bar{F}(\cdot) := 1 - F(\cdot)$ we find that $M(t)$ has a Poisson distribution:

$$M(t) \stackrel{\mathrm{d}}{=} \mathbb{P}\mathrm{ois}\left(\int_0^t \bar{F}(s, t-s)\lambda(s)\mathrm{d}s\right),$$

and we note for later that

$$\int_0^t \bar{F}(s, t-s)\lambda(s)\mathrm{d}s = \int_0^t \bar{F}(t-s, s)\lambda(t-s)\mathrm{d}s.$$

After this general observation, we return to the initial context. Whereas we so far assumed that the input rate function and service-time distribution function were deterministic, we now assume that they are stochastic. More specifically, we use λ_i for the arrival rate when the background process $X(\cdot)$ is in state i, and $F_i(\cdot)$ for the distribution function of particles arriving while the background process is in the state i.

By conditioning on the sample path of the background process, say $X(s) = f(s)$, we find that $M(t)$ is Poisson distributed with parameter $\int_0^t \bar{F}_{f(t-s)}(s)\lambda_{f(t-s)}\mathrm{d}s$. Then by unconditioning, i.e., averaging over all paths $f(\cdot)$ of the background process, its probability generating function (pgf) equals the moment generating function (mgf) of its random parameter, evaluated at $(z-1)$:

$$\mathbb{E}\, z^{M(t)} = \mathbb{E}\exp\left(-(1-z)\int_0^t \bar{F}_{X(t-s)}(s)\lambda_{X(t-s)}\mathrm{d}s\right);$$

see for example [5, p. 226]. Recalling that $X(\cdot)$ is assumed to be stationary, we have the distributional equality $\{X(t+u)|\ u \in \mathbb{R}\} \stackrel{\mathrm{d}}{=} \{X(u)|\ u \in \mathbb{R}\}$, so that

$$\mathbb{E}\, z^{M(t)} = \mathbb{E}\exp\left(-(1-z)\int_0^t \bar{F}_{X(-s)}(s)\lambda_{X(-s)}\mathrm{d}s\right),$$

or, denoting by $\hat{X}(\cdot)$ the time-reversed version of $X(\cdot)$, with $a_i(s) := \lambda_i \bar{F}_i(s)$,

$$\begin{aligned}\mathbb{E}\, z^{M(t)} &= \mathbb{E}\exp\left(-(1-z)\int_0^t \bar{F}_{\hat{X}(s)}(s)\lambda_{\hat{X}(s)}\mathrm{d}s\right)\\ &= \mathbb{E}\exp\left(-(1-z)\int_0^t a_{\hat{X}(s)}(s)\mathrm{d}s\right).\end{aligned}$$

This probability generating function allows us to analyze the mean and variance of $M(t)$. It is immediate that the mean of $M(t)$ equals, cf. [21, Thm. 2.1],

$$\mathbb{E}M(t) = \mathbb{E}\int_0^t a_{\hat{X}(s)}(s)\mathrm{d}s = \int_0^t \mathbb{E}a_{\hat{X}(s)}(s)\mathrm{d}s = \sum_{i=1}^d \pi_i\lambda_i \int_0^t \bar{F}_i(s)\mathrm{d}s. \tag{1}$$

This evidently converges to $\sum_{i=1}^d \pi_i\varrho_i$ as $t\to\infty$, where $\varrho_i := \lambda_i\int_0^\infty \bar{F}_i(s)\mathrm{d}s$ is the traffic intensity when in state i.

The variance can be computed as well, as follows. We start with the standard equality (commonly known as the "law of total variance")

$$\mathbb{V}\mathrm{ar}(M(t)) = \mathbb{E}\left[\mathbb{V}\mathrm{ar}(M(t)|\hat{X})\right] + \mathbb{V}\mathrm{ar}\left[\mathbb{E}(M(t)|\hat{X})\right].$$

First notice that $\mathbb{V}\mathrm{ar}(M(t)|\hat{X}) = \mathbb{E}(M(t)|\hat{X}) = \int_0^t a_{\hat{X}(s)}(s)\mathrm{d}s$ because $(M(t)\mid\hat{X})$ has a Poisson distribution (as was noted above). Hence,

$$\mathbb{E}\left[\mathbb{V}\mathrm{ar}(M(t)|\hat{X})\right] = \mathbb{E}\left[\mathbb{E}(M(t)\mid\hat{X})\right] = \mathbb{E}M(t) = \sum_{i=1}^d \pi_i\lambda_i\int_0^t \bar{F}_i(s)\mathrm{d}s.$$

The only quantity that remains to be computed is now $\mathbb{V}\mathrm{ar}[\mathbb{E}(M(t)|\hat{X})]$. That is done as follows:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\int_0^t a_{\hat{X}(s)}(s)\mathrm{d}s\right) &= \int_0^t\int_0^t \mathbb{C}\mathrm{ov}\left(a_{\hat{X}(u)}(u), a_{\hat{X}(s)}(s)\right)\mathrm{d}u\,\mathrm{d}s\\ &= \sum_{i,j=1}^d \int_0^t\int_0^t a_i(u)a_j(s)\,\mathbb{C}\mathrm{ov}\left(1\{\hat{X}(u)=i\}, 1\{\hat{X}(s)=j\}\right)\mathrm{d}u\,\mathrm{d}s,\end{aligned}$$

where for $u<s$

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(1\{\hat{X}(u)=i\}, 1\{\hat{X}(s)=j\}\right) &= \pi_i\left(e^{\hat{Q}(s-u)}\right)_{ij} - \pi_i\pi_j\\ &= \pi_j\left(e^{Q(s-u)}\right)_{ji} - \pi_i\pi_j.\end{aligned} \tag{2}$$

We now make the expressions more explicit for the case where t tends to ∞. With $D_\pi = \mathrm{diag}\{\boldsymbol{\pi}\}$, Q and $\hat{Q} = D_\pi^{-1}Q^{\mathrm{T}}D_\pi^{-1}$ are the transition rate matrices of X and $\hat{X}$, respectively. Let us define the matrix $\Sigma(s) = (\sigma_{ij}(s))_{i,j=1}^d$ through

$$\sigma_{ij}(s) := \pi_j\left(e^{Qs}\right)_{ji} - \pi_i\pi_j.$$

Letting $t \to \infty$, we obtain

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\int_0^\infty a_{\hat{X}(s)}(s)\mathrm{d}s\right) &= \sum_{i,j=1}^{d}\int_0^\infty\int_0^\infty a_i(u)a_j(s)\left(\sigma_{ij}(s-u)1\{s>u\}\right.\\ &\quad\left.+\sigma_{ji}(u-s)1\{s<u\}\right)\mathrm{d}u\,\mathrm{d}s\\ &= \sum_{i,j=1}^{d}\int_0^\infty\int_0^\infty \left(a_i(u)a_j(u+s)\sigma_{ij}(s)\right.\\ &\quad\left.+a_i(u+s)a_j(u)\sigma_{ji}(s)\right)\mathrm{d}u\,\mathrm{d}s\\ &= 2\sum_{i,j=1}^{d}\int_0^\infty\int_0^\infty a_i(u)a_j(u+s)\sigma_{ij}(s)\mathrm{d}u\,\mathrm{d}s.\end{aligned}$$

When the service-time distributions are exponential, that is, $\bar{F}_i(t) = e^{-\mu_i t}$, so that $a_i(t) = \lambda_i e^{-\mu_i t}$, we have

$$\mathbb{V}\mathrm{ar}\left(\int_0^\infty a_{\hat{X}(s)}(s)\mathrm{d}s\right) = 2\sum_{i,j}\frac{\lambda_i\lambda_j}{\mu_i+\mu_j}\int_0^\infty e^{-\mu_j s}\sigma_{ij}(s)\mathrm{d}s. \tag{3}$$

We summarize (some of) our findings.

Proposition 1 *The transient mean of the number of particles is*

$$\mathbb{E}M(t) = \mathbb{E}\int_0^t a_{\hat{X}(s)}(s)\mathrm{d}s = \int_0^t \mathbb{E}a_{\hat{X}(s)}(s)\mathrm{d}s = \sum_{i=1}^{d}\pi_i\lambda_i\int_0^t \bar{F}_i(s)\mathrm{d}s,$$

whereas the stationary variance is

$$\mathbb{V}\mathrm{ar}M(\infty) = \sum_{i=1}^{d}\pi_i\frac{\lambda_i}{\mu_i} + 2\sum_{i,j=1}^{d}\int_0^\infty\int_0^\infty a_i(u)a_j(u+s)\sigma_{ij}(s)\mathrm{d}u\,\mathrm{d}s,$$

provided that the system started empty.

We finish this section by performing some explicit calculations for the case that X is reversible and exponential service times; later on we further focus on the situation of $d = 2$. Due to the reversibility, $\pi_i q_{ij} = \pi_j q_{ji}$ for all $i, j \in \{1,\ldots,d\}$. As a consequence $D_\pi Q = Q^{\mathrm{T}} D_\pi$, so that the matrix

$$D_\pi^{1/2} Q D_\pi^{-1/2}$$

is symmetric, and can be written as $G(-\Delta)G^{\mathrm{T}}$, where G is a (real-valued) orthogonal matrix, and $\Delta = \mathrm{diag}\{\boldsymbol{\delta}\}$ is a (real-valued) diagonal matrix (where it is noted that,

owing to the background process' irreducibility all but one entries of $\boldsymbol{\delta}$ are strictly positive). It follows that

$$Q = (D_\pi^{-1/2} G)(-\Delta)(D_\pi^{-1/2} G)^{-1},$$

and therefore

$$\begin{aligned} e^{Qs} &= (D_\pi^{-1/2} G)(e^{-\Delta s})(D_\pi^{-1/2} G)^{-1} = D_\pi^{-1/2} G\, e^{-\Delta s}\, G^{\mathrm{T}} D_\pi^{1/2}; \\ (e^{Qs})^{\mathrm{T}} &= D_\pi^{1/2} G\, e^{-\Delta s}\, G^{\mathrm{T}} D_\pi^{-1/2}. \end{aligned}$$

It now follows that

$$\Sigma(s) = (e^{Qs})^{\mathrm{T}} D_\pi - \boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}} = D_\pi^{1/2} G\, e^{-\Delta s}\, G^{\mathrm{T}} D_\pi^{1/2} - \boldsymbol{\pi}\boldsymbol{\pi}^{\mathrm{T}}$$

is symmetric, and hence for each $i, j \in \{1, \ldots, d\}$ we can write $\sigma_{ij}(s) = \sum_{k=1}^{d} c_{ijk} e^{-\delta_k s} - \pi_i \pi_j$. As a consequence,

$$\begin{aligned} \mathbb{V}\mathrm{ar}\left(\int_0^\infty a_{\hat{X}(s)}(s) \mathrm{d}s \right) &= 2 \sum_{i,j} \frac{\lambda_i \lambda_j}{\mu_i + \mu_j} \int_0^\infty e^{-\mu_j s} \sigma_{ij}(s) \mathrm{d}s \\ &= 2 \sum_{i,j,k} \frac{\lambda_i \lambda_j}{\mu_i + \mu_j} \left(\frac{c_{ijk}}{\mu_j + \delta_k} - \frac{\pi_i \pi_j}{\mu_j} \right). \end{aligned}$$

In the case of $d = 2$, we have $\pi_1 = q_2/\bar{q} = 1 - \pi_2$, with $\bar{q} := q_1 + q_2$. It is readily verified that $\delta_1 = 0$ and $\delta_2 = \bar{q}$. It requires a standard computation to verify that

$$e^{Qs} = \begin{pmatrix} \pi_1 + \pi_2 e^{-\bar{q}s} & \pi_2 - \pi_2 e^{-\bar{q}s} \\ \pi_1 - \pi_1 e^{-\bar{q}s} & \pi_2 + \pi_1 e^{-\bar{q}s} \end{pmatrix},$$

and also

$$\int_0^\infty \Sigma(s) \begin{pmatrix} e^{-\mu_1 s} & 0 \\ 0 & e^{-\mu_2 s} \end{pmatrix} \mathrm{d}s = \pi_1 \pi_2 \begin{pmatrix} (\bar{q} + \mu_1)^{-1} & -(\bar{q} + \mu_2)^{-1} \\ -(\bar{q} + \mu_1)^{-1} & (\bar{q} + \mu_2)^{-1} \end{pmatrix}.$$

Elementary calculus now yields that (3) equals

$$\frac{q_1 q_2}{\bar{q}^2} \left(\frac{\lambda_1^2}{\mu_1} \cdot \frac{1}{\bar{q} + \mu_1} + \frac{\lambda_2^2}{\mu_2} \cdot \frac{1}{\bar{q} + \mu_2} - 2 \frac{\lambda_1 \lambda_2}{\mu_1 + \mu_2} \left(\frac{1}{\bar{q} + \mu_1} + \frac{1}{\bar{q} + \mu_2} \right) \right).$$

3 Exponential service times

In this section we analyze the special case of exponential service times in greater detail. We set up a system of differential equations for the moment generating function of the transient number of particles in the system. Then this is used to determine the mean and higher moments after an exponential amount of time.

We start this section with some preliminaries and additional notation. Here and in the remaining sections we denote by $M_i(t)$ the number of particles in the system at time t, conditional on the background process being in state i at time 0. It is evident that $M_i(t)$ can be written as the sum of two independent components: the number of particles still present at time t out of the original population of size x_0 (in the sequel denoted by $\check{M}(t)$), increased by the number of particles that arrived in $(0, t]$ that is still present at time t (in the sequel denoted by $\bar{M}_i(t)$ in case the background process is in state i at time 0).

Due to the assumption that the service times are exponentially distributed, there are positive numbers μ_i (for $i = 1, \ldots, d$) such that $\bar{F}_i(t) = e^{-\mu_i t}$. In the case that the μ_i are identical (say equal to $\mu > 0$), $\check{M}(t)$ follows a binomial distribution with parameters x_0 and $e^{-\mu t}$. In the case the μ_i are not identical, we need to know the number $x_{0,i}$ particles present at time 0 that were generated while the background process was in state i. The resulting (independent) random variables $\check{M}_i(t)$ follow binomial distributions with parameters $x_{0,i}$ and $e^{-\mu_i t}$; indeed, $\check{M}(t) = \sum_i \check{M}_i(t)$.

Given these observations we concentrate in the remainder of this section on the more complicated component of $M(t)$, that is $\bar{M}_i(t)$.

3.1 Differential equation

Recall that we write, for ease of notation, $q_i := 1/\mathbb{E}T_i$, and $q_{ij} := p_{ij} q_i$ (where $i \neq j$), with $q_{ii} = -q_i$. The main quantity in this subsection is the moment generating function of $\bar{M}_i(t)$:

$$\Lambda_i(\vartheta, t) := \mathbb{E} e^{\vartheta \bar{M}_i(t)}.$$

Consider a small time period Δt, and focus on all terms of magnitude $O(\Delta t)$ or larger. In our continuous-time Markov setting, the background process has either zero jumps (with probability $1 - q_i \Delta t + o(\Delta t)$), or a jump to state $j \neq i$ (with probability $q_{ij} \Delta t + o(\Delta t)$); the probability of more than one transition is $o(\Delta t)$ (see for instance [19, Thm. 2.8.2]).

Note that

$$\begin{aligned}\Lambda_i(\vartheta, t) = \sum_{k=0}^{\infty} e^{-\lambda_i \Delta t} \frac{(\lambda_i \Delta t)^k}{k!} \left(p_i(\vartheta, t)\right)^k \\ \times \left(\sum_{j \neq i} q_{ij} \Delta t \, \Lambda_j(\vartheta, t - \Delta t) + \left(1 - \sum_{j \neq i} q_{ij} \Delta t\right) \Lambda_i(\vartheta, t - \Delta t) \right) \\ + O\left((\Delta t)^2\right); \end{aligned} \tag{4}$$

here $p_i(\vartheta, t)$ is the mgf of a random variable distributed on $\{0, 1\}$, indicating whether a particle arriving in the time period $(0, \Delta t)$ is still present at t. It is seen that the value 1 occurs with probability

$$\int_0^{\Delta t} \frac{1}{\Delta t}\left(\int_{t-u}^{\infty} \mu_i e^{-\mu_i v}\mathrm{d}v\right)\mathrm{d}u = \frac{e^{-\mu_i t}}{\Delta t}\int_0^{\Delta t} e^{\mu_i u}\mathrm{d}u = \frac{e^{-\mu_i t}}{\mu_i \Delta t}(e^{\mu_i \Delta t}-1)$$
$$= e^{-\mu_i t} + O(\Delta t).$$

Hence, $p_i(\vartheta, t) = 1 + e^{-\mu_i t}\left(e^{\vartheta}-1\right) + O(\Delta t)$, and thus

$$\sum_{k=0}^{\infty} e^{-\lambda_i \Delta t}\frac{(\lambda_i \Delta t)^k}{k!}\left(p_i(\vartheta,t)\right)^k = e^{-\lambda_i \Delta t}\exp\left[\lambda_i \Delta t\, p_i(\vartheta,t)\right]$$
$$= 1 + \lambda_i \Delta t\left(e^{\vartheta}-1\right)e^{-\mu_i t} + O\left((\Delta t)^2\right).$$

Now $q_i = \sum_{j\neq i} q_{ij}$ yields

$$\begin{aligned}
\Lambda_i(\vartheta,t) &= \left(1+\lambda_i\,\Delta t\,(e^{\vartheta}-1)e^{-\mu_i t}\right)\\
&\quad\times\left(\sum_{j\neq i} q_{ij}\Delta t\,\Lambda_j(\vartheta,t-\Delta t) + (1-q_i\Delta t)\,\Lambda_i(\vartheta,t-\Delta t)\right) + O\left((\Delta t)^2\right)\\
&= \left(1+\lambda_i\,\Delta t\,(e^{\vartheta}-1)e^{-\mu_i t}\right)\\
&\quad\times\left(\sum_{j\neq i} q_{ij}\Delta t\,\Lambda_j(\vartheta,t) + \Lambda_i(\vartheta,t) - \Delta t\,\Lambda_i'(\vartheta,t) - q_i\Delta t\,\Lambda_i(\vartheta,t)\right) + O\left((\Delta t)^2\right)\\
&= \left(1+\lambda_i\,\Delta t\,(e^{\vartheta}-1)e^{-\mu_i t}\right)\\
&\quad\times\left(\sum_{j=1}^{d} q_{ij}\Delta t\,\Lambda_j(\vartheta,t) + \Lambda_i(\vartheta,t) - \Delta t\,\Lambda_i'(\vartheta,t)\right) + O\left((\Delta t)^2\right),
\end{aligned}$$

where the derivative is with respect to t. We have found the following system of differential equations.

Proposition 2 *The mgfs $\Lambda_i(\vartheta,t)$ satisfy*

$$\lambda_i\,(e^{\vartheta}-1)e^{-\mu_i t}\Lambda_i(\vartheta,t) = \Lambda_i'(\vartheta,t) - \sum_{j=1}^{d} q_{ij}\Lambda_j(\vartheta,t). \tag{5}$$

Now define $\psi_i(\alpha,\vartheta) := \int_0^{\infty}\alpha e^{-\alpha t}\Lambda_i(\vartheta,t)\mathrm{d}t$. Then, by integrating,

$$\int_0^{\infty}\alpha e^{-\alpha t}\Lambda_i'(\vartheta,t)\mathrm{d}t = \alpha(\psi_i(\alpha,\vartheta)-1).$$

We thus obtain

$$\lambda_i(e^{\vartheta}-1)\frac{\alpha}{\alpha+\mu_i}\psi_i(\alpha+\mu_i,\vartheta) = \alpha(\psi_i(\alpha,\vartheta)-1) - \sum_{j=1}^{d} q_{ij}\psi_j(\alpha,\vartheta); \tag{6}$$

cf. [1, Eq. (4.6), Cor. 1] in the one-dimensional case and [15, Thm. 3] in the network case for equations that resemble (5) for Markov-modulated shot-noise models. These may be viewed as continuous state-space analogs or weak limits of the infinite-server queue (see [14] regarding a general framework that includes both for the network version in the non-modulated case).

3.2 Mean

To compute $\mathbb{E}\bar{M}_i(\tau_\alpha)$, with $\tau_\alpha \sim \exp(\alpha)$, we differentiate Eq. (6) with respect to ϑ and let $\vartheta \downarrow 0$, thus obtaining

$$\lambda_i \frac{\alpha}{\alpha+\mu_i}\psi_i(\alpha+\mu_i,0) = \alpha \cdot \lim_{\vartheta\downarrow 0}\frac{\mathrm{d}}{\mathrm{d}\vartheta}\psi_i(\alpha,\vartheta) - \sum_{j=1}^{d} q_{ij} \cdot \lim_{\vartheta\downarrow 0}\frac{\mathrm{d}}{\mathrm{d}\vartheta}\psi_j(\alpha,\vartheta),$$

or

$$\begin{aligned}\lambda_i \frac{\alpha}{\alpha+\mu_i} &= \alpha\int_0^\infty \alpha e^{-\alpha t}\mathbb{E}\bar{M}_i(t)\mathrm{d}t - \sum_{j=1}^{d} q_{ij}\int_0^\infty \alpha e^{-\alpha t}\mathbb{E}\bar{M}_j(t)\mathrm{d}t \\ &= \alpha\mathbb{E}\bar{M}_i(\tau_\alpha) - \sum_{j=1}^{d} q_{ij}\mathbb{E}\bar{M}_j(\tau_\alpha). \end{aligned} \tag{7}$$

Now consider the special case that the background process is in equilibrium at time 0. It turns out that the expressions simplify significantly. We have, due to (7), using the fact that $\sum_i \pi_i q_{ij} = 0$,

$$\sum_{i=1}^{d}\pi_i\mathbb{E}\bar{M}_i(\tau_\alpha) = \sum_{i=1}^{d}\pi_i\lambda_i\frac{1}{\alpha+\mu_i}.$$

Laplace inversion yields

$$\sum_{i=1}^{d}\pi_i\mathbb{E}\bar{M}_i(t) = \sum_{i=1}^{d}\frac{\pi_i\lambda_i}{\mu_i}(1-e^{-\mu_i t}),$$

in line with (1). Now consider steady-state behavior, that is, we let $\alpha \downarrow 0$. From the above, we obtain an expression that could as well have been found by applying Little's law:

$$\sum_{i=1}^{d}\pi_i\mathbb{E}\bar{M}_i(\infty) = \sum_{i=1}^{d}\pi_i\frac{\lambda_i}{\mu_i}.$$

3.3 Higher moments

A second differentiation of (6) yields

$$2\lambda_i \frac{\alpha}{\alpha+\mu_i}\mathbb{E}\bar{M}_i(\tau_{\alpha+\mu_i}) + \lambda_i \frac{\alpha}{\alpha+\mu_i} = \alpha\mathbb{E}\bar{M}_i^2(\tau_\alpha) - \sum_{j=1}^{d} q_{ij}\mathbb{E}\bar{M}_j^2(\tau_\alpha).$$

In other words, once we know the $\mathbb{E}\bar{M}_i(\tau_\alpha)$ for all $\alpha > 0$, we can compute the associated second moment as well.

Along the same lines,

$$\lambda_i \frac{\alpha}{\alpha+\mu_i}\sum_{k=0}^{n-1}\binom{n}{k}\cdot\lim_{\vartheta\downarrow 0}\frac{\mathrm{d}^k}{\mathrm{d}\vartheta^k}\psi_i(\alpha+\mu_i,\vartheta) = \lambda_i \frac{\alpha}{\alpha+\mu_i}\sum_{k=0}^{n-1}\binom{n}{k}\mathbb{E}\bar{M}_i^k(\tau_{\alpha+\mu_i})$$
$$= \alpha\mathbb{E}\bar{M}_i^n(\tau_\alpha) - \sum_{j=1}^{d} q_{ij}\mathbb{E}\bar{M}_j^n(\tau_\alpha).$$

As a consequence, these higher moments (at exponentially distributed epochs) can be recursively determined. Again there is a simplification if the background process is in equilibrium at time 0. Then we have the equation

$$\sum_{i=1}^{d}\pi_i\mathbb{E}\bar{M}_i^n(\tau_\alpha) = \sum_{i=1}^{d}\pi_i\lambda_i\frac{1}{\alpha+\mu_i}\sum_{k=0}^{n-1}\binom{n}{k}\mathbb{E}\bar{M}_i^k(\tau_{\alpha+\mu_i}).$$

For the steady-state we obtain, cf. [1],

$$\sum_{i=1}^{d}\pi_i\mathbb{E}\bar{M}_i^n(\infty) = \sum_{i=1}^{d}\pi_i\frac{\lambda_i}{\mu_i}\sum_{k=0}^{n-1}\binom{n}{k}\mathbb{E}\bar{M}_i^k(\tau_{\mu_i}).$$

4 Asymptotic normality for general service times

In this section we consider our Markov-modulated infinite-server system, but, as opposed to the setting discussed in the previous section, now with *generally* distributed service times. The main result is a central limit theorem (for $N\to\infty$) under the scaling $q_{ij}\mapsto N^{1+\varepsilon}q_{ij}$ and $\lambda_i\mapsto N\lambda_i$; here $\varepsilon>0$. The intuitive idea behind this scaling is that the state of the background process moves at a faster time scale than the arrival processes (so that the arrival process is effectively a single Poisson process as $N\to\infty$), while this arrival process is sped up by a factor N (so that a central limit regime kicks in).

Remark 1 We already observed that the number $\check{M}_i^{(N)}(t)$ of particles still present at time t, out of the initial population of size Nx_0 and that arrived while the background process was in state i, is not affected by the evolution of the background process, as the

departure rate has been determined upon arrival. Specializing to the case of exponential service times (with mean μ_i^{-1} if the particle under consideration had entered while the background process was in state i), the corresponding random variables have independent binomial distributions with parameters $Nx_{0,i}$ and $e^{-\mu_i t}$. $Nx_{0,i}$ denotes the number of particles present at time 0 that arrived while the background was in state i. Therefore, as $N \to \infty$

$$\frac{\check{M}_i^{(N)}(t) - Nx_{0,i}e^{-\mu_i t}}{\sqrt{N}} \xrightarrow{\mathrm{d}} \mathbb{N}\mathrm{orm}\left(0, x_{0,i}e^{-\mu_i t}(1 - e^{-\mu_i t})\right).$$

For other service-time distributions the quantities $e^{-\mu_i t}$ have to be replaced by the appropriate survival probability associated with the residual lifetime of a particle that is present at time 0 and that had arrived while the background process was in i.

In light of the above, it suffices to focus on establishing a central limit theorem for the number of particles that arrived in $(0, t]$ that are still present at time t. Let, in line with earlier definitions, this number be denoted by $\bar{M}_i^{(N)}(t)$ in case the modulating process is in state i at time 0. ◊

One of the leading intuitions of this section is that, due to the fact that the timescale of the background process is faster than that of the arrival process, we can essentially replace our Markov-modulated infinite-server system, as $N \to \infty$, by an M/G/∞ queue. This effectively means that, irrespective of the initial state i, $\bar{M}_i^{(N)}(t)$ can be approximated by a Poisson distribution with parameter $N\varrho_t$. The candidate for ϱ_t can be easily identified using the theory of Sect. 2:

$$\varrho_t := \sum_{i=1}^{d} \pi_i \lambda_i \int_0^t \bar{F}_i(s)\,\mathrm{d}s. \tag{8}$$

Let us now focus on identifying a candidate for the limiting covariance between $\bar{M}_i^{(N)}(t)$ and $\bar{M}_i^{(N)}(t+u)$; this is a rather elementary computation that we include for the sake of completeness. Let $N(t)$ be the number present in an M/G/∞ queue that started off empty at time 0; the arrival rate is λ and the distribution function of the service times is denoted by $F(\cdot)$. In this system it is possible to compute the covariance between $N(t)$ and $N(t+u)$ explicitly in terms of the arrival rate and the distribution function $F(\cdot)$ of the service times. Realize that $N(t+u)$ can be written as the sum of the particles that were already present at time t and that are still present at time $t+u$ (which we denote by $N_t(t+u)$), and the ones that have arrived in $(t, t+u]$ and that are still present at time $t+u$. The latter quantity being independent of $N(t)$, we have

$$\mathbb{C}\mathrm{ov}(N(t), N(t+u)) = \mathbb{C}\mathrm{ov}(N(t), N_t(t+u)).$$

It thus suffices to compute the quantity $\mathbb{C}\mathrm{ov}(N(t), N_t(t+u))$. To this end, define

$$q^{\mathrm{A}} \equiv q^{\mathrm{A}}_{u,t} := \int_0^t \frac{1}{t} F(t-v)\mathrm{d}v = \int_0^t \frac{1}{t} F(v)\mathrm{d}v,$$
$$q^{\mathrm{B}} \equiv q^{\mathrm{B}}_{u,t} := \int_0^t \frac{1}{t}(F(t+u-v) - F(t-v))\mathrm{d}v = \int_0^t \frac{1}{t}(F(v+u) - F(v))\mathrm{d}v,$$
$$q^{\mathrm{C}} \equiv q^{\mathrm{C}}_{u,t} := \int_0^t \frac{1}{t}(1 - F(t+u-v))\mathrm{d}v = \int_0^t \frac{1}{t}(1 - F(v+u))\mathrm{d}v;$$

the first of these quantities can be interpreted as the probability that an arbitrary particle that has arrived in $[0, t)$ has already left the system at time t, the second as the probability that it is still present at time t but not at $t+u$ anymore, and the third as the probability that it is still present at time $t+u$. It now follows that

$$\begin{aligned}\mathbb{E}N(t)\,N_t(t+u) &= \sum_{k=0}^{\infty}\sum_{\ell=0}^{k} k\ell\,\mathbb{P}(N(t)=k, N_t(t+u)=\ell)\\ &= \sum_{m=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^m}{m!}\sum_{k=0}^{m}\sum_{\ell=0}^{k} k\ell\binom{m}{k,\ell}(q^{\mathrm{A}})^{m-k}(q^{\mathrm{B}})^{k-\ell}(q^{\mathrm{C}})^{\ell},\end{aligned}$$

which turns out to equal (after some elementary computations) $q^{\mathrm{C}}\,\lambda t + q^{\mathrm{C}}(1-q^{\mathrm{A}})\lambda^2 t^2$. As $\mathbb{E}N(t) = (1-q^{\mathrm{A}})\lambda t$ and $\mathbb{E}N_t(t+u) = q^{\mathrm{C}}\,\lambda t$, it follows that

$$\mathbb{C}\mathrm{ov}(N(t), N(t+u)) = q^{\mathrm{C}}\,\lambda t = \lambda\int_0^t (1 - F(v+u))\mathrm{d}v.$$

This computation provides us with the candidate for the central limit result in the case of general service times. Define in this context, for $t_1 \le t_2$,

$$c_{t_1,t_2} := \sum_{i=1}^{d} \pi_i\lambda_i \int_0^{t_1} \bar{F}_i(v + t_2 - t_1)\mathrm{d}v$$

(while if $t_2 < t_1$ we put $c_{t_1,t_2} = c_{t_2,t_1}$).

The following result covers the asymptotic multivariate normality. In the proof we consider the bivariate case (time epochs t and $t+u$), but the extension to a general dimension (time epochs t_1 up to t_K with, without loss of generality, $t_1 \le \ldots \le t_K$) is straightforward and essentially a matter of careful bookkeeping.

Theorem 1 *For any $\boldsymbol{\alpha} \in \mathbb{R}^K$ and $\boldsymbol{t} \in \mathbb{R}^K$ (with $t_1 \le \ldots \le t_K$), and general service times, as $N \to \infty$,*

$$\frac{\sum_{k=1}^K \alpha_k \bar{M}_i^{(N)}(t_k) - N\sum_{k=1}^K \alpha_k \varrho_{t_k}}{\sqrt{N}} \xrightarrow{d} N(0, \sigma^2),$$

with

$$\sigma^2 := \sum_{k=1}^{K} \alpha_k^2 \varrho_{t_k} + 2 \sum_{k=1}^{K-1} \sum_{\ell=1}^{k-1} \alpha_k \alpha_\ell c_{t_k,t_\ell}.$$

This theorem shows convergence of the finite-dimensional distributions to a multivariate Normal distribution. A next step would be to prove convergence *at the process level*, viz. convergence of

$$\left(\frac{\bar{M}_i^{(N)}(t) - N\varrho_t}{\sqrt{N}} \right)_{t \geq 0}$$

to a Gaussian process with a specific correlation structure. Such a result has been proven for the regular (that is, non-modulated) M/M/∞ queue in which the Poisson arrival rate is scaled by N; the limiting process is then an Ornstein-Uhlenbeck process—see, for example [22]. The proofs of such weak convergence results typically consist of three steps: single-dimensional convergence, finite-dimensional convergence, and a tightness argument, where the tightness step tends to be relatively complicated. In our setup (that is, the Markov-modulated M/G/∞ queue) we have proven the first two steps; the third step (tightness) is beyond the scope of this paper.

We prove Thm. 1 for the case of $K = 2$, with $t_1 = t$ and $t_2 = t + u$ (for $t, u \geq 0$); higher dimensions can be dealt with fully analogously but these require substantially more administration. Our starting point is to set up a system of differential equations for the non-scaled process. This system is derived in the very same way as the differential equations for the univariate exponential case (see Prop. 2). Define, for fixed scalars α_1, α_2, and for $u \geq 0$ given,

$$\Lambda_i(\vartheta, t) := \mathbb{E} \exp(\vartheta \alpha_1 \bar{M}_i(t) + \vartheta \alpha_2 \bar{M}_i(t+u)).$$

In addition, let

$$\begin{aligned} p_i(\vartheta, t) &:= F_i(t) + e^{\vartheta \alpha_1}(F_i(t+u) - F_i(t)) + e^{\vartheta(\alpha_1+\alpha_2)}(1 - F_i(t+u)) \\ &= e^{\vartheta(\alpha_1+\alpha_2)} - (e^{\vartheta\alpha_1} - 1)F_i(t) - e^{\vartheta\alpha_1}(e^{\vartheta\alpha_2} - 1)F_i(t+u). \end{aligned}$$

Proposition 3 *The mgfs $\Lambda_i(\vartheta, t)$ satisfy*

$$\bar{p}_i(\vartheta, t)\Lambda_i(\vartheta, t) = \Lambda_i'(\vartheta, t) - \sum_{j=1}^{d} q_{ij}\Lambda_j(\vartheta, t),$$

where $\bar{p}_i(\vartheta, t) := \lambda_i(p_i(\vartheta, t) - 1)$.

Proof Let $I_i(t)$ be the indicator function of the event that a particle arriving in $(0, \Delta t]$ (while the background process was in state i) is still in the system at time t, and consider the random variable $\alpha_1 I_i(t) + \alpha_2 I_i(t+u)$. Similarly to what we did earlier in this

section, $\alpha_1 I_i(t) + \alpha_2 I_i(t+u)$ can be split into three contributions; one corresponding to the event that a particle that arrived in $(0, \Delta t]$ has already left the system at time t, one corresponding to the event that it is still present at time t but not anymore at time $t+u$, and finally one corresponding to the event that it is still present at time $t+u$. With some standard calculus it is readily obtained that

$$\mathbb{E}\exp\left(\vartheta\alpha_1 I_i(t) + \vartheta\alpha_2 I_i(t+u)\right) = p_i(\vartheta, t) + O(\Delta t).$$

This means that we obtain

$$\begin{aligned}\Lambda_i(\vartheta, t) = {} & \lambda_i \Delta t \cdot p_i(\vartheta, t)\Lambda_i(\vartheta, t-\Delta t) \\ & + \sum_{j\neq i} q_{ij}\Delta t \cdot \Lambda_j(\vartheta, t-\Delta t) + (1 - \lambda_i\Delta t - q_i\Delta t)\,\Lambda_i(\vartheta, t-\Delta t) + o(\Delta t).\end{aligned}$$

Now subtracting $\Lambda_i(\vartheta, t-\Delta t)$ from both sides, dividing by Δt, and letting $\Delta t \downarrow 0$ leads to the desired system of differential equations. □

Proof of Thm. 1. Now we are ready to prove the bivariate asymptotic normality for the case of general service times. The idea behind the proof is to (i) start off with the differential equations for the non-scaled system as derived in Prop. 3; (ii) incorporate the scaling in the differential equations, and apply the centering and normalization corresponding to the central limit regime; (iii) use Taylor expansions (for large N); (iv) obtain a limiting differential equation (as $N \to \infty$). This limiting differential equation finally yields the claimed central limit theorem.

We first "center" the random variable $\alpha_1 \bar{M}_i^{(N)}(t) + \alpha_2 \bar{M}_i^{(N)}(t+u)$; to this end we subtract $N\varrho(t,u)$ from this random variable, with

$$\varrho(t,u) := \alpha_1\varrho_t + \alpha_2\varrho_{t+u},$$

and ϱ_t defined as in Eq. (8). At this point we impose the scaling, that is, we replace q_{ij} by $N^{1+\varepsilon}q_{ij}$, and λ_i by $N\lambda_i$. With these parameters, we now study the appropriately centered and scaled random variable

$$\frac{\vartheta\alpha_1\bar{M}_i^{(N)}(t) + \vartheta\alpha_2\bar{M}_i^{(N)}(t+u) - N\vartheta\varrho(t)}{\sqrt{N}},$$

where we suppress the argument u in $\varrho(t,u)$ (as u is held fixed throughout the proof). It means that we study the "centered and scaled mgf"

$$\tilde{\Lambda}_i^{(N)}(\vartheta, t) := \Lambda_i\left(\frac{\vartheta}{\sqrt{N}}, t\right)\exp\left(-\sqrt{N}\vartheta\varrho(t)\right), \tag{9}$$

where, due to Prop. 3, $\Lambda_i(\vartheta/\sqrt{N}, t)$ satisfies

$$N\bar{p}_i\left(\frac{\vartheta}{\sqrt{N}}, t\right)\Lambda_i\left(\frac{\vartheta}{\sqrt{N}}, t\right) = \Lambda_i'\left(\frac{\vartheta}{\sqrt{N}}, t\right) - N^{1+\varepsilon}\sum_{j=1}^{d} q_{ij}\Lambda_j\left(\frac{\vartheta}{\sqrt{N}}, t\right).$$

Realize that, as a straightforward application of the chain rule,

$$\left(\tilde{\Lambda}_i^{(N)}\right)'(\vartheta,t) = \Lambda_i'\left(\frac{\vartheta}{\sqrt{N}},t\right)\exp\left(-\sqrt{N}\vartheta\varrho(t)\right) - \sqrt{N}\vartheta\varrho'(t)\tilde{\Lambda}_i^{(N)}(\vartheta,t).$$

Upon combining the above, we find a relation which is completely in terms of the centered/scaled mgf $\tilde{\Lambda}_i^{(N)}(\vartheta,t)$:

$$\begin{aligned} N\bar{p}_i\left(\frac{\vartheta}{\sqrt{N}},t\right)\tilde{\Lambda}_i^{(N)}(\vartheta,t) &= \left(\tilde{\Lambda}_i^{(N)}\right)'(\vartheta,t) \\ &\quad +\sqrt{N}\vartheta\varrho'(t)\tilde{\Lambda}_i^{(N)}(\vartheta,t) - N^{1+\varepsilon}\sum_{j=1}^{d} q_{ij}\tilde{\Lambda}_j^{(N)}(\vartheta,t). \end{aligned} \tag{10}$$

We now study the solution of this system of differential equations for N large by "Tayloring" the function $\bar{p}_i(\vartheta/\sqrt{N},t)$ with respect to N. It is an elementary exercise to check that

$$\bar{p}_i\left(\frac{\vartheta}{\sqrt{N}},t\right) = \frac{h_{1,i}(\vartheta,t)}{\sqrt{N}} + \frac{h_{2,i}(\vartheta,t)}{N} + O(N^{-\frac{3}{2}}),$$

with

$$\begin{aligned} h_{1,i}(\vartheta,t) &:= \lambda_i\left(\vartheta\alpha_1\bar{F}_i(t) + \vartheta\alpha_2\bar{F}_i(t+u)\right), \\ h_{2,i}(\vartheta,t) &:= \frac{\lambda_i}{2}\left(\vartheta^2\alpha_1^2\bar{F}_i(t) + \vartheta^2(\alpha_2(2\alpha_1+\alpha_2))\bar{F}_i(t+u)\right). \end{aligned}$$

We thus obtain the differential equation

$$\begin{aligned} &\left(\sqrt{N}\left(h_{1,i}(\vartheta,t) - \vartheta\varrho'(t)\right) + h_{2,i}(\vartheta,t) + O(N^{-\frac{1}{2}})\right)\tilde{\Lambda}_i^{(N)}(\vartheta,t) \\ &= \left(\tilde{\Lambda}_i^{(N)}\right)'(\vartheta,t) - N^{1+\varepsilon}\sum_{j=1}^{d} q_{ij}\tilde{\Lambda}_j^{(N)}(\vartheta,t), \end{aligned}$$

or in self-evident matrix/vector notation,

$$\begin{aligned} N^{1+\varepsilon}Q\tilde{\boldsymbol{\Lambda}}^{(N)}(\vartheta,t) &= \left(\tilde{\boldsymbol{\Lambda}}^{(N)}\right)'(\vartheta,t) - \sqrt{N}\left(H_1(\vartheta,t) - \vartheta\varrho'(t)\right)\tilde{\boldsymbol{\Lambda}}^{(N)}(\vartheta,t) \\ &\quad - H_2(\vartheta,t)\tilde{\boldsymbol{\Lambda}}^{(N)}(\vartheta,t) + O(N^{-\frac{1}{2}}). \end{aligned}$$

Now premultiply this equation by the so-called *fundamental matrix* $\mathcal{F} := (\Pi - Q)^{-1}$, where $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$. It holds that $\Pi^2 = \Pi$, $\mathcal{F}\Pi = \Pi\mathcal{F} = \Pi$, and $Q\mathcal{F} = \mathcal{F}Q = \Pi - I$;

see for these properties and more background on fundamental matrices and deviations matrices e.g. [4]. We then obtain

$$N^{1+\varepsilon}\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta,t) = N^{1+\varepsilon}\Pi\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta,t) - \mathcal{F}\left(\tilde{\mathbf{\Lambda}}^{(N)}\right)'(\vartheta,t) + \sqrt{N}\mathcal{F}\left(H_1(\vartheta,t) - \vartheta\varrho'(t)\right)\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta,t) + \mathcal{F}H_2(\vartheta,t)\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta,t) + O(N^{-\frac{1}{2}}).$$

Iterating this identity once, we obtain

$$N^{1+\varepsilon}\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta,t) = N^{1+\varepsilon}\Pi\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta,t) - \Pi\mathcal{F}\left(\tilde{\mathbf{\Lambda}}^{(N)}\right)'(\vartheta,t) + \sqrt{N}\mathcal{F}\left(H_1(\vartheta,t) - \vartheta\varrho'(t)\right)\Pi\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta,t) + \mathcal{F}H_2(\vartheta,t)\Pi\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta,t) + O(N^{-\frac{1}{2}}) + O(N^{-\varepsilon}).$$

Now premultiply the equation by $d\boldsymbol{\pi}^{\mathrm{T}} = \mathbf{1}^{\mathrm{T}}\Pi$. Recalling the identity $\Pi\mathcal{F} = \Pi$ and noting that it follows from the definition of $\varrho(t)$ that

$$\mathbf{1}^{\mathrm{T}}\Pi\left(H_1(\vartheta,t) - \vartheta\varrho'(t)\right)\mathbf{1} = 0,$$

all terms of $O(N^{\alpha})$ with $\alpha > 0$ cancel. For $\lim_{N\to\infty}\boldsymbol{\pi}^{\mathrm{T}}\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta,t) =: \tilde{\Lambda}(\vartheta,t)$ we thus obtain the following differential equation:

$$\tilde{\Lambda}'(\vartheta,t) = \left(\sum_{i=1}^{d}\pi_i h_{2,i}(\vartheta,t)\right)\tilde{\Lambda}(\vartheta,t).$$

Using the technique of separation of variables, it follows that

$$\tilde{\Lambda}(\vartheta,t) = \exp\left(\int_0^t\sum_{i=1}^{d}\pi_i h_{2,i}(\vartheta,s)\mathrm{d}s\right)\kappa(\vartheta,u),$$

or

$$\tilde{\Lambda}(\vartheta,t) = \exp\left(\frac{\vartheta^2}{2}\sum_{i=1}^{d}\pi_i\int_0^t\left(\lambda_i\left(\alpha_1^2\bar{F}_i(s) + (2\alpha_1+\alpha_2)\alpha_2\bar{F}_i(s+u)\right)\right)\mathrm{d}s\right)\kappa(\vartheta,u),$$

for some function $\kappa(\vartheta,u)$ that is independent of t. Now note that this expression should not depend on α_1 if $t = 0$. In addition, if we insert $u = 0$, then α_1 and α_2 should appear in the expression as $\alpha_1+\alpha_2$. This enables us to identify $\kappa(\vartheta,u)$. We eventually obtain

$$\tilde{\Lambda}(\vartheta,t) = \exp\left(\frac{\vartheta^2}{2}\left(\alpha_1^2\varrho_t + 2\alpha_1\alpha_2 c_{t,t+u} + \alpha_2^2\varrho_{t+u}\right)\right), \tag{11}$$

as desired. We have proven the claimed convergence.

Remark 2 It is remarked that the central limit theorem does not carry over to the case $\varepsilon \in (-1, 0]$, as then the term of order $N^{1-2\varepsilon}$ cannot be neglected relative to the term of order $N^{1-\varepsilon}$. As a result, in that situation the variance featuring in the central limit theorem will contain the fundamental matrix $\mathcal{F}$ for these values of ε. ◊

5 Examples

5.1 Two-state model

In this example we consider the case $d = 2$, and exponential sojourn times of the background process, that is, the time spent in state i is exponential with mean $1/q_i \in (0, \infty)$. From $\mathbb{E}\bar{\boldsymbol{M}}(\tau_\alpha) = (A(\alpha))^{-1}\boldsymbol{\varphi}(\alpha)$ we obtain for the mean number in the system after an exponential time with mean $1/\alpha$ (ignoring the effect of an initial population)

$$\begin{pmatrix} \mathbb{E}\bar{M}_1(\tau_\alpha) \\ \mathbb{E}\bar{M}_2(\tau_\alpha) \end{pmatrix} = \frac{1}{q_1+q_2+\alpha} \begin{pmatrix} q_2+\alpha & q_1 \\ q_2 & q_1+\alpha \end{pmatrix} \begin{pmatrix} \dfrac{\lambda_1}{\alpha+\mu_1} \\ \dfrac{\lambda_2}{\alpha+\mu_2} \end{pmatrix}$$
$$= \begin{pmatrix} \dfrac{\alpha+q_2}{\alpha+q_1+q_2}\dfrac{\lambda_1}{\alpha+\mu_1} + \dfrac{q_1}{\alpha+q_1+q_2}\dfrac{\lambda_2}{\alpha+\mu_2} \\ \dfrac{\alpha+q_1}{\alpha+q_1+q_2}\dfrac{\lambda_2}{\alpha+\mu_2} + \dfrac{q_2}{\alpha+q_1+q_2}\dfrac{\lambda_1}{\alpha+\mu_1} \end{pmatrix}$$

When sending α to ∞, we indeed obtain that $\mathbb{E}\bar{M}_i(\tau_\infty) = 0$; when sending α to 0, the resulting formula is consistent with the long-term mean number in the system, as found earlier. Replacing q_i by Nq_i (for $i = 1, 2$), we obtain that both components of $\mathbb{E}\bar{\boldsymbol{M}}(\tau_\alpha)$ converge (as $N \to \infty$) to

$$\pi_1 \frac{\lambda_1}{\alpha+\mu_1} + \pi_2 \frac{\lambda_2}{\alpha+\mu_2},$$

which is for $\mu_1 = \mu_2$ in line with the findings in [12].

We now focus on computing the second moment; for ease we consider the stationary case. From Sect. 3.3, we have

$$\sum_{i=1}^{d} \frac{2\pi_i\lambda_i}{\alpha+\mu_i}\mathbb{E}\bar{M}_i(\tau_{\alpha+\mu_i}) + \sum_{i=1}^{d} \frac{\pi_i\lambda_i}{\alpha+\mu_i} = \sum_{i=1}^{d} \pi_i \mathbb{E}\bar{M}_i^2(\tau_\alpha),$$

which becomes after sending α to 0,

$$\mathbb{E}\bar{M}^2(\infty) := \sum_{i=1}^{d} \pi_i \mathbb{E}\bar{M}_i^2(\infty) = \sum_{i=1}^{d} 2\pi_i \frac{\lambda_i}{\mu_i}\mathbb{E}\bar{M}_i(\tau_{\mu_i}) + \sum_{i=1}^{d} \pi_i \frac{\lambda_i}{\mu_i};$$

obviously, $\pi_1 = 1 - \pi_2 = q_2/(q_1+q_2)$.

We now find a lower bound on the variance of the stationary number of particles in the system. Restricting ourselves to the case $\mu_i \equiv \mu$ for all $i = 1, \ldots, d$, elementary computations yield, with $r_i := \lambda_i/\mu$ and $q := q_1 + q_2$,

$$\mathbb{E}\bar{M}^2(\infty) = \frac{\pi_1 r_1}{\mu - q}\left((\mu - q_2)r_1 - q_1 r_2\right) + \pi_1 r_1 + \frac{\pi_2 r_2}{\mu - q}\left((\mu - q_1)r_2 - q_2 r_1\right) + \pi_2 r_2.$$

We now claim that, with R denoting the stationary mean $\pi_1 r_1 + \pi_2 r_2$, the stationary variance is larger than this R, or equivalently

$$\mathbb{E}\bar{M}^2(\infty) \geq R^2 + R, \tag{12}$$

with equality only if $\lambda_1 = \lambda_2$. This can be shown as follows. Writing $r_1 = a r_2$, the above claim reduces to verifying that, for all $a \in (0, \infty)$,

$$a^2(f_1 - \pi_1)\pi_1 + a(f_2 - \pi_2)\pi_1 + a(g_1 - \pi_1)\pi_2 + (g_2 - \pi_2)\pi_2 \geq 0, \tag{13}$$

with equality only if $a = 1$; here

$$f_1 = 1 - f_2 := \frac{\mu - q_2}{\mu - q}, \quad g_2 := 1 - g_1 := \frac{\mu - q_1}{\mu - q}.$$

Observe that $f_1 > \pi_1$, so that the left-hand side of (13) has a minimum. Now realize that $f_1 - \pi_1 = -(f_2 - \pi_2)$ and $g_2 - \pi_2 = -(g_1 - \pi_1)$. As a result, (13) reduces to

$$(a - 1)(a(f_1 - \pi_1)\pi_1 - (g_2 - \pi_2)\pi_2) \geq 0,$$

which, due to $(f_1 - \pi_1)\pi_1 = (g_2 - \pi_2)\pi_2$ can be rewritten as $(f_1 - \pi_1)\pi_1(a-1)^2 \geq 0$. Claim (12) thus follows. We conclude that $\mathbb{V}\mathrm{ar}\bar{M}(\infty) \geq \mathbb{E}\bar{M}(\infty)$, with equality if and only if $\lambda_1 = \lambda_2$.

This result can be intuitively understood. As argued before, $\bar{M}(\infty)$ is distributed as a Poisson random variable with a *random* parameter. We showed with an elementary argument in the introduction of [12] that this entails that $\mathbb{V}\mathrm{ar}\bar{M}(\infty) \geq \mathbb{E}\bar{M}(\infty)$; informally, this says that Markov modulation increases the variability of the stationary distribution. We have now shown that for $d = 2$ this inequality is in fact strict, unless the λ_i match (and equal, say λ). In fact, then the queue is just an M/M/∞ system which has the Poisson(λ/μ) distribution as the equilibrium distribution, for which mean and variance coincide (and have the value λ/μ). In other words, for $d = 2$ there are no other ways to obtain a Poisson stationary distribution than letting all λ_i be equal.

5.2 Computational results

We include computational results demonstrating the converging behavior of the two-state scaled process in one dimension (i.e., $K = 1$ in Thm. 1). Unscaled, the parameters are $\boldsymbol{\lambda} = (1, 2)$, $\boldsymbol{\mu} = (1, 1)$, and $\boldsymbol{q} = (1, 3)$. Depicted in Fig. 1 is the limiting behavior of Eq. (9) assuming exponential service times, obtained by solving the scaled version

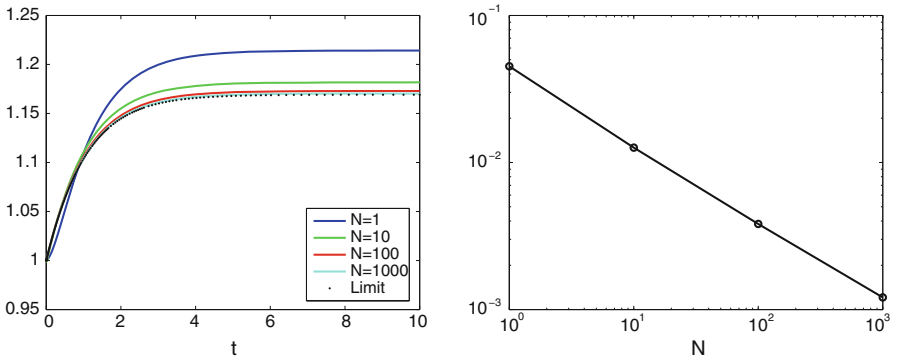

Fig. 1 (*left*) The scaled process, $\tilde{\Lambda}^{(N)}(0.5, t)$, approaches the limiting curve as N grows larger. (*right*) Maximum error as a function of N shows loglinear convergence.

of the differential equation (5) with the mgf parameter $\vartheta = 0.5$ and $\varepsilon = 0.5$. The corresponding limiting curve from Eq. (11) is plotted as well. As in the case with deterministic transition times [3], we observe loglinear convergence, with the solution curve closely following the limiting curve for $N = 1000$. Tweaking the parameters results in the same convergence behavior.

Acknowledgements The authors like to thank Koen de Turck (Ghent University) for helpful discussions. O. Kella is partially supported by The Vigevani Chair in Statistics. M. Mandjes is also with EURANDOM (Eindhoven University of Technology, the Netherlands). Part of this work was done while M. Mandjes was visiting The Hebrew University

References

1. Asmussen, S., Kella, O.: Rate modulation in dams and ruin problems. J. Appl. Probab. **33**, 523–535 (1996)
2. Baykal-Gursoy, M., Xiao, W.: Stochastic decomposition in M/M/∞ queues with Markov-modulated service rates. Queueing Syst. **48**, 75–88 (2004)
3. Blom, J., Mandjes, M., Thorsdottir, H.: Time-scaling limits for Markov-modulated infinite-server queues. Stoch. Models **29**, 112–127 (2012)
4. Coolen-Schrijner, P., van Doorn, E.: The deviation matrix of a continuous-time Markov chain. Probab. Eng. Inform. Sci. **16**, 351–366 (2002)
5. D'Auria, B.: M/M/∞ queues in semi-Markovian random environment. Queueing Syst. **58**, 221–237 (2008)
6. Dembo, A., Zeitouni, O.: Large Deviations Techniques and Applications, 2nd edn. Springer, New York (1998)
7. Economou, A., Fakinos, D.: The infinite server queue with arrivals generated by a non-homogeneous compound Poisson process and heterogeneous customers. Commun. Stat.: Stoch. Models **15**, 993–1002 (1999)
8. Falin, G.: The M/M/ ∞ queue in random environment. Queueing Syst. **58**, 65–76 (2008)
9. Fralix, B., Adan, I.: An infinite-server queue influenced by a semi-Markovian environment. Queueing Syst. **61**, 65–84 (2009)
10. Glynn, P.: Large deviations for the infinite server queue in heavy traffic. Inst. Math. Appl. **71**, 387–394 (1995)
11. Glynn, P., Whitt, W.: A new view of the heavy-traffic limit theorem for infinite-server queues. Adv. Appl. Probab. **23**, 188–209 (1991)

12. Hellings, T., Mandjes, M., Blom, J.: Semi-Markov-modulated infinite-server queues: approximations by time-scaling. Stoch. Models **28**, 452–477 (2012)
13. Keilson, J., Servi, L.: The matrix M/M/∞ system: retrial models and Markov modulated sources. Adv. Appl. Probab. **25**, 453–471 (1993)
14. Kella, O., Whitt, W.: Linear stochastic fluid networks. J. Appl. Probab. **36**, 244–260 (1999)
15. Kella, O., Stadje, W.: Markov-modulated linear fluid networks with Markov additive input. J. Appl. Probab. **39**, 413–420 (2002)
16. Liu, L., Templeton, J.: Autocorrelations in infinite server batch arrival queues. Queueing Syst. **14**, 313–337 (1993)
17. Neuts, M.: Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach. Johns Hopkins University Press, Baltimore (1981)
18. Neuts, M., Chen, S.: The infinite server queue with semi-Markovian arrivals and negative exponential services. J. Appl. Probab. **9**, 178–184 (1972)
19. Norris, J.: Markov Chains. Cambridge University Press, Cambridge (1997)
20. O'Cinneide, C., Purdue, P.: The M/M/∞ queue in a random environment. J. Appl. Probab. **23**, 175–184 (1986)
21. Purdue, P., Linton, D.: An infinite-server queue subject to an extraneous phase process and related models. J. Appl. Probab. **18**, 236–244 (1981)
22. Robert, P: Stochastic Networks and Queues. Springer, Berlin (2003)
23. Schwabe, A., Rybakova, K., Bruggeman, F.: Transcription stochasticity of complex gene regulation models. Biophysical J. **103**, 1152–1161 (2012)
24. Whitt, W.: Stochastic-Process Limits. Springer, New York (2001)
25. Willmot, G.E., Drekic, S.: Time-dependent analysis of some infinite server queues with bulk Poisson arrivals. INFOR **47**, 297–303 (2009)