MARKOV-MODULATED INFINITE-SERVER QUEUES WITH GENERAL SERVICE TIMES

J. BLOM *, O. KELLA †, M. MANDJES $^{\bullet,\star}$, H. THORSDOTTIR $^{\star,\bullet}$

ABSTRACT. This paper studies an infinite-server queue in a Markov environment, that is, an infinite-server queue with an arrival rate that equals λ_i when an external Markov process is in state i. The service times have a general distribution that depends on the state of the background process upon arrival. 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 customers at an exponentially distributed time, as well as the steady-state moments. 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 $t \geq 0$ is asymptotically Normally distributed; in addition convergence of finite-dimensional distributions is proven.

KEYWORDS. Queues $\star$ infinite-server systems $\star$ Markov modulation $\star$ Laplace transforms $\star$ fluid and diffusion scaling

$^{\bullet}$ Korteweg-de Vries Institute for Mathematics, University of Amsterdam, the Netherlands. * CWI, Amsterdam, the Netherlands. $\dagger$ Department of Statistics, The Hebrew University of Jerusalem, Israel (partially supported by The Vigevani Chair in Statistics). M. Mandjes is also with EURANDOM (Eindhoven University of Technology, the Netherlands).

1. INTRODUCTION

Owing to its wide applicability, the infinite-server queue has proven to be an extremely useful model. It 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 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 $1/\mu$, the so-called M/G/∞ queue, it is known that the stationary number in the system has a Poisson distribution with mean λ/μ. Also the transient behavior of such an M/G/∞ queue is well understood, see e.g. [19, p. 355].

By broadening the assumptions of the M/G/∞ queue, many interesting variants have been studied. Attention has been paid to the case of renewal (rather than Poisson) arrivals [6, 7], but in the present paper we aim at introducing some sort of 'burstiness' in the arrivals. Here the arrivals occur according to a Poisson process, but the arrival rate is determined by the state of an external Markov process, also referred to as the 'background process'. Put somewhat more precisely, with $X(t)$ denoting an irreducible continuous-time Markov process defined on a finite state space $\{1, \ldots, d\}$, the arrival rate at time t is given by $\lambda_{X(t)}$, where $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_d)$ is a vector with non-negative entries. Throughout it is assumed that the time a customer remains in the system, the

Key words and phrases. Markov-modulated Poisson process, queues, general service times.

service time, has some general distribution with mean $1/\mu_i$ depending on the state of the background process upon arrival.
The resulting model is called a *Markov-modulated* M/G/∞ *queue*, or an infinite-server queue in a Markov-modulated random environment. This type of systems can be used in several application domains, ranging from telecommunication networks, where the arrival rate of customers may vary between times of the day, to biology: mRNA strings are transcripted and degraded in a cell, where these transcriptions typically tend to occur in clustered fashion. The proposed model therefore captures the key characteristics of this mechanism well, as argued in [18].

A variety of results exists on Markov-modulated single- and many-server queues, whereas the literature on infinite-server variants is surprisingly scarce. In the case of a single server, the stationary distribution of the number of customers is of matrix-geometric form [13]; in this sense this system can be viewed as a matrix generalization of the normal M/M/1 queue where the stationary distribution is scalar-geometric. In [15] the stationary distribution for the case of infinitely many servers is considered; the results are in terms of the factorial moments of the numbers of customers (and in addition, 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 [9]. The case of Markov-modulated *renewal* (rather than Poisson) arrivals, but exponential service times, is covered in [14]. Related results can be found in [11] as well, where special attention is paid to the autocorrelations in infinite-server systems of various types.
Using the same model, D'Auria [3] finds a recursion for the factorial moments of the number of particles in the system for the case that the sojourn times of the background process are not necessarily exponential. He relies on the observation that the number of customers present has, in stationarity, a Poisson distribution with random parameter. The computation of this distribution requires quite careful analysis though. Fralix and Adan [5] also focus on the situation in which the service times are not necessarily exponential, but rather Erlang or hyperexponential; this can then be used to address the case with general service times. In [8] it was shown that if the sojourn 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 reduces to an M/G/∞ system.
While the above results focus on Markov-modulated infinite-server queues in stationarity, literature on their transient behavior is much less prominent. In [2], we studied both the transient and stationary behavior of the model in question with exponential service rates and a Markovian background process with deterministic transition times, under two scaling regimes. In the former, the transition times are sped up by a factor N leading asymptotically to the modulated queue resembling a Poisson process. In the latter, the arrival rates are scaled by a factor N and the transition times by a factor $N^{1+\varepsilon}$, eventually leading to a central limit result. Here we continue with this scaling regime while generalizing the distribution of the service times and the background process.

The main contributions of our paper are the following. In the first place we develop in Section 2 expressions for the transient mean and variance for the number of particles in the system at time $t \geq 0$. In Section 3 we focus on 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 includes a recursive formula to compute the higher moments. Section 4 considers the regime in which 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 $\varepsilon > 0$. We prove that under this scaling the transient number of particles in the system obeys a central limit theorem for finite dimensions. This is explicitly shown for exponentially distributed service times, after which we present the corresponding result for generally distributed service times, without details. Finally, Section 5 contains examples demonstrating analytically and numerically the results from Sections 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 $Q = (q_{ij})_{i,j=1}^{d}$. 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 $\boldsymbol{\pi}Q = \mathbf{0}$). Denote $q_i := -q_{ii} = \sum_{j \neq i} q_{ij}$.

Now consider the embedded discrete-time Markov chain that corresponds to the jump epochs of $X(t)$. It has a probability transition matrix $P = (p_{ij})_{i,j=1}^{d}$, 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 (in self-evident notation) $D_Q := \mathrm{diag}\{\boldsymbol{q}\}$. 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 $F_i(x) := \mathbb{P}(B_i \leq x)$. The service times are independent of the background process $X(t)$ and the arrival process. 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.

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* input process with rate function $\lambda(s)$, 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 have that $M(t)$ has a Poisson distribution:

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

and we note for later that

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

After this general observation, we return to the initial context. Instead of taking a deterministic $\lambda(t)$, we now consider $\lambda_{X(t)}$ and $F_{X(t)}(\cdot)$, the latter being the distribution function of particles arriving in the state $X(t)$. 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 $\int_0^t \bar{F}_{f(t-s)}(s)\lambda_{f(t-s)}{\rm d}s$. Then by unconditioning, i.e., returning to the random process $X(t)$, and using that since $M(t)$ is Poisson, its probability generating function (pgf) equals the moment generating function (mgf) of its random parameter, evaluated at $(z-1)$ (see e.g. [3], p.226):

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

Since $X(\cdot)$ is stationary, we have the distributional equality $\{X(t+u)|\, u\in\mathbb{R}\} \overset{\rm d}{=} \{X(u)|\, u\in\mathbb{R}\}$, so that

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

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

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

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. [16, Thm. 2.1],

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

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

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

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

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

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

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

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

where for $u < s$

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

We now make the expressions more explicit for the case that t tends to ∞. With $D_\pi = \mathrm{diag}\{\boldsymbol{\pi}\}$, Q and $\hat{Q} = D_\pi Q^{\mathrm{T}} D_\pi^{-1}$ are the transition rate matrices of X and $\hat{X}$, respectively. Let us denote 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

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

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 that

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

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; 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

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

It now follows that

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

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,

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

In the case of $d=2$, we have that $\pi_1=q_2/\bar q=1-\pi_2$, 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

$$e^{Qs}=\left(\begin{array}{cc}\pi_1+\pi_2e^{-\bar q s} & \pi_2-\pi_2e^{-\bar q s}\\ \pi_1-\pi_1e^{-\bar q s} & \pi_2+\pi_1e^{-\bar q s}\end{array}\right),$$

and also

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

Elementary calculus now yields that (3) equals

$$\frac{q_1q_2}{\bar q^2}\left(\frac{\lambda_1^2}{\mu_1}\cdot\frac{1}{\bar q+\mu_1}+\frac{\lambda_2^2}{\mu_2}\cdot\frac{1}{\bar q+\mu_2}-2\frac{\lambda_1\lambda_2}{\mu_1+\mu_2}\left(\frac{1}{\bar q+\mu_1}+\frac{1}{\bar q+\mu_2}\right)\right).$$

3. Exponential service times

In this section now consider the case of exponential service times in greater detail. The number of particles in the system at time t, conditional on the background process being in state i at time 0, is denoted by $M_i(t)$. 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).

In the case that the μ_i are identical, $\check M(t)$ 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 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 $\bar M_i(t)$:

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

Note that

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

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 $(t-\Delta t,t)$ is still present at t. It is seen that the value 1 occurs with probability

$$\begin{aligned}\int_0^{\Delta t}\frac{1}{\Delta t}\int_{t-u}^{\infty}\mu_i e^{-\mu_i v}\mathrm{d}v\mathrm{d}u &= \frac{1}{\Delta t}\int_0^{\Delta t}\left[-e^{-\mu_i v}\right]_{t-u}^{\infty}\mathrm{d}u = \frac{e^{-\mu_i t}}{\Delta t}\int_0^{\Delta t}e^{\mu_i u}\mathrm{d}u \\ &= \frac{e^{-\mu_i t}}{\mu_i\Delta t}(e^{\mu_i\Delta t}-1) = e^{-\mu_i t}+O\left(\Delta t\right).\end{aligned}$$

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

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

The usual 'infinitesimal argument' yields

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

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. [10, Thm. 3] for a related result.

3.2. **Mean.** To compute $\mathbb{E}\bar{M}_i(\tau_\alpha)$, with $\tau_\alpha\sim\exp(\alpha)$, we differentiate the above expression 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

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

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 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 that

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

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,

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

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. Time-scaling

In this section we first consider exponentially distributed service times, like in the previous section, but now we scale $q_{ij} \mapsto N^{1+\varepsilon} q_{ij}$ and $\lambda_i \mapsto N\lambda_i$; here $\varepsilon > 0$. The idea is that the state of the background moves at a faster time scale than the arrival processes (so that the arrival process is effectively a Poisson process as $N \to \infty$), while the arrival process is sped up by a factor N (so that a central limit regime kicks in).

It will turn out later that the line of reasoning below can also be followed for *arbitrarily* distributed service times B_i, but as the exponential case is notationally considerably more concise, we start by examining the exponential case to later explain how to extend it to the general case. We further establish the convergence of the process in finite dimension to a Normal distribution.

4.1. Exponential case.

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. 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 $N \to \infty$

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

In light of this, we can focus on the number of particles arriving in $(0, t]$ that are still present at time t. Let, as before, in case the modulating process is in state i at time 0, this number be denoted by $\bar{M}_i^{(N)}(t)$.

The main point of this section is that we can essentially replace our Markov-modulated infinite server system, as $N \to \infty$, by an M/M/∞ queue, in that, irrespective of the initial state i, $\bar{M}_i^{(N)}(t)$ can be approximated by a Poisson distribution with parameter $N\varrho_t$, where

$$\varrho_t := \sum_{i=1}^{d} \pi_i \frac{\lambda_i}{\mu_i}(1-e^{-\mu_i t}).$$

In our later analysis we also need the covariance between $\bar{M}_i^{(N)}(t)$ and $\bar{M}_i^{(N)}(t+u)$; this is a standard computation that we include for the sake of completeness. Using standard results for the M/M/∞ queue, this number will be roughly the order of $N^2(\varrho_{t+u} - \varrho_t)$, as illustrated by the following. Let $N(t)$ be the number of particles in an M/M/∞ system (with arrival rate λ and service rate μ; define $\varrho := \lambda/\mu$, and here locally $\varrho_t := \varrho(1-e^{-\mu t})$) at time $t > 0$, starting empty. Then $N(t+u)$ can be written as the sum of the particles that were already present at time t and that are still present at time $t+u$, and the ones that have arrived in $(t,t+u]$ and that are still present at time $t+u$ (which we denote by $N_t(t+u)$). The former quantity being independent of $N(t)$, we have

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

Then

$$\begin{aligned}
\mathbb{E}N(t)N_t(t+u) &= \sum_{k=0}^{\infty}\sum_{\ell=0}^{k} k\ell\,\mathbb{P}(N(t)=k,N_t(t+u)=\ell)\\
&= \sum_{k=0}^{\infty}\sum_{\ell=0}^{k} k\ell\, e^{-\varrho_t}\frac{\varrho_t^k}{k!}\binom{k}{\ell}(e^{-\mu u})^{\ell}(1-e^{-\mu u})^{k-\ell}\\
&= e^{-\mu u}\sum_{k=0}^{\infty}k^2 e^{-\varrho_t}\frac{\varrho_t^k}{k!}=(\varrho_t^2+\varrho_t)e^{-\mu u},
\end{aligned}$$

whereas $\mathbb{E}N(t)=\varrho_t$ and $\mathbb{E}N_t(t+u)=\varrho_t e^{-\mu u}$. We conclude that

$$\mathbb{C}\mathrm{ov}(N(t),N_t(t+u))=\varrho_t e^{-\mu u}=\varrho_{t+u}-\varrho_u.$$

Returning to the modulated process, the above reasoning motivates us to expect that, for all α_1,α_2, as $N\to\infty$,

$$\xi_i^{(N)}(t,u):=\frac{\alpha_1\bar{M}_i^{(N)}(t)+\alpha_2\bar{M}_i^{(N)}(t+u)-N(\alpha_1\varrho_t+\alpha_2\varrho_{t+u})}{\sqrt{N}}\xrightarrow{\mathrm{d}}\mathbb{N}\mathrm{orm}(0,v(t,u)). \tag{8}$$

where $v(t,u):=\alpha_1^2\varrho_t+2\alpha_1\alpha_2(\varrho_{t+u}-\varrho_u)+\alpha_2^2\varrho_{t+u}$. The objective of this section is to prove this property.

To this end, we let $\gamma_i^{(N)}(\vartheta,t,u)$ denote the moment generating function of the random variable $\alpha_1\bar{M}_i^{(N)}(t)+\alpha_2\bar{M}_i^{(N)}(t+u)$ conditional on the background process being in state i at time 0. With $\varrho(t,u):=\alpha_1\varrho_t+\alpha_2\varrho_{t+u}$, we define the mgf of $\xi_i^{(N)}(t,u)$,

$$\delta_i^{(N)}(\vartheta,t,u):=\mathbb{E}\left(\exp\left(\vartheta\xi_i^{(N)}(t,u)\right)\right)=\gamma_i^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t,u\right)\cdot e^{-\vartheta\sqrt{N}\varrho(t,u)}. \tag{9}$$

Now consider the first $1/N^{1+2\varepsilon}$ time units; due to the usual Markovian arguments, the background process has either zero jumps (with probability $1-q_i/N^{\varepsilon}+O(N^{-1-2\varepsilon})$), or a jump to state $j\neq i$ (with probability $q_{ij}/N^{\varepsilon}+O(N^{-1-2\varepsilon})$). The number of arrivals up to time $1/N^{1+2\varepsilon}$ has a Poisson distribution with mean $\lambda_i N^{-2\varepsilon}$. It then follows that, neglecting $O(N^{-1-2\varepsilon})$ terms,

$$\begin{aligned}
\gamma_i^{(N)}(\vartheta,t,u) &= \sum_{k=0}^{\infty}e^{-\lambda_i N^{-2\varepsilon}}\frac{(\lambda_i N^{-2\varepsilon})^k}{k!}(p_i^{(N)}(\vartheta,t,u))^k\times\\
&\quad\left(\sum_{j\neq i}\frac{q_{ij}}{N^{\varepsilon}}\gamma_j^{(N)}\left(\vartheta,t-\frac{1}{N^{1+2\varepsilon}},u\right)+\left(1-\frac{q_i}{N^{\varepsilon}}\right)\gamma_i^{(N)}\left(\vartheta,t-\frac{1}{N^{1+2\varepsilon}},u\right)\right)\\
&= \sum_{k=0}^{\infty}e^{-\lambda_i N^{-2\varepsilon}}\frac{(\lambda_i N^{-2\varepsilon})^k}{k!}(p_i^{(N)}(\vartheta,t,u))^k\times\\
&\quad\left(\sum_{j=1}^{d}\frac{q_{ij}}{N^{\varepsilon}}\gamma_j^{(N)}\left(\vartheta,t-\frac{1}{N^{1+2\varepsilon}},u\right)+\gamma_i^{(N)}\left(\vartheta,t-\frac{1}{N^{1+2\varepsilon}},u\right)\right);
\end{aligned}$$

here $p_i^{(N)}(\vartheta,t,u)$ represents the mgf of a random variable $\alpha_1 I(t)+\alpha_2 I(t+u)$, where $I(t)$ is the indicator function of the event that a particle that arrived in $[0,N^{-1-2\varepsilon}]$ is still present at time t. Again neglecting $O(N^{-1-2\varepsilon})$ terms,

$$p_i^{(N)}(\vartheta,t,u)=1+e^{-\mu_i t}(e^{\vartheta\alpha_1}-1)+e^{-\mu_i(t+u)}e^{\vartheta\alpha_1}(e^{\vartheta\alpha_2}-1).$$

It is now a matter of straightforward algebra that, up to $O(N^{-2\varepsilon})$-terms,

$$\sum_{k=0}^{\infty}e^{-\lambda_i N^{-2\varepsilon}}\frac{(\lambda_i N^{-2\varepsilon})^k}{k!}p_i^{(N)}(\vartheta,t,u)^k=1+\frac{\lambda_i}{N^{2\varepsilon}}\left(e^{-\mu_i t}(e^{\vartheta\alpha_1}-1)+e^{-\mu_i(t+u)}e^{\vartheta\alpha_1}(e^{\vartheta\alpha_2}-1)\right).$$

and as a consequence, up to $O(N^{-1-2\varepsilon})$-terms,

$$\begin{aligned}&\sum_{k=0}^{\infty}e^{-\lambda_i N^{-2\varepsilon}}\frac{(\lambda_i N^{-2\varepsilon})^k}{k!}\left(p_i^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t,u\right)\right)^k\\ &=\;1+\frac{\lambda_i}{N^{2\varepsilon}}\left(e^{-\mu_i t}\left(\frac{\vartheta\alpha_1}{\sqrt{N}}+\frac{\vartheta^2\alpha_1^2}{2N}\right)+e^{-\mu_i(t+u)}\left(\frac{\vartheta\alpha_2}{\sqrt{N}}+\frac{\vartheta^2(2\alpha_1+\alpha_2)\alpha_2}{2N}\right)\right),\end{aligned}$$

In addition, up to $O(N^{-1-2\varepsilon})$-terms,

$$\gamma_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t-\frac{1}{N^{1+2\varepsilon}},u\right)=\gamma_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t,u\right)-\frac{1}{N^{1+2\varepsilon}}\frac{\mathrm{d}}{\mathrm{d}t}\gamma_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t,u\right).$$

Upon combining the above,

$$\begin{aligned}&\gamma_i^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t,u\right)\\ &=\;\left(1+\frac{\lambda_i}{N^{2\varepsilon}}\left(e^{-\mu_i t}\left(\frac{\vartheta\alpha_1}{\sqrt{N}}+\frac{\vartheta^2\alpha_1^2}{2N}\right)+e^{-\mu_i(t+u)}\left(\frac{\vartheta\alpha_2}{\sqrt{N}}+\frac{\vartheta^2(2\alpha_1+\alpha_2)\alpha_2}{2N}\right)\right)\right)\\ &\quad\times\left(\sum_{j=1}^{d}\frac{q_{ij}}{N^{\varepsilon}}\gamma_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t,u\right)+\gamma_i^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t,u\right)\right.\\ &\qquad\left.-\frac{1}{N^{1+2\varepsilon}}\frac{\mathrm{d}}{\mathrm{d}t}\gamma_i^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t,u\right)\right)+O\left(\frac{1}{N^{1+2\varepsilon}}\right).\end{aligned}$$

It is clear that

$$\frac{\mathrm{d}}{\mathrm{d}t}\gamma_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t,u\right)e^{-\vartheta\sqrt{N}\varrho(t,u)}=\frac{\mathrm{d}}{\mathrm{d}t}\delta_j^{(N)}(\vartheta,t,u)+\vartheta\sqrt{N}\delta_j^{(N)}(\vartheta,t,u)\cdot\frac{\mathrm{d}}{\mathrm{d}t}\varrho(t,u).$$

We thus obtain, neglecting the higher order terms,

$$\begin{aligned}&\delta_i^{(N)}(\vartheta,t,u)\\ &=\;\left(1+\frac{\lambda_i}{N^{2\varepsilon}}\left(e^{-\mu_i t}\left(\frac{\vartheta\alpha_1}{\sqrt{N}}+\frac{\vartheta^2\alpha_1^2}{2N}\right)+e^{-\mu_i(t+u)}\left(\frac{\vartheta\alpha_2}{\sqrt{N}}+\frac{\vartheta^2(2\alpha_1+\alpha_2)\alpha_2}{2N}\right)\right)\right)\\ &\quad\times\left(\sum_{j=1}^{d}\frac{q_{ij}}{N^{\varepsilon}}\delta_j^{(N)}(\vartheta,t,u)+\delta_i^{(N)}(\vartheta,t,u)-\frac{1}{N^{1+2\varepsilon}}\frac{\mathrm{d}}{\mathrm{d}t}\delta_i^{(N)}(\vartheta,t,u)\right.\\ &\qquad\left.-\frac{\vartheta}{N^{\frac{1}{2}+2\varepsilon}}\delta_i^{(N)}(\vartheta,t,u)\cdot\frac{\mathrm{d}}{\mathrm{d}t}\varrho(t,u)-\sum_{j=1}^{d}\frac{\vartheta q_{ij}}{N^{\frac{1}{2}+3\varepsilon}}\delta_j^{(N)}(\vartheta,t,u)\cdot\frac{\mathrm{d}}{\mathrm{d}t}\varrho(t,u)\right).\end{aligned}\tag{10}$$

We now proceed by multiplying Eqn. (10) by $\pi_i N^{1+2\varepsilon}$ and sum over i. With the resulting equation we let $N\to\infty$ to obtain a differential equation. The q_{ij}-terms in the resulting equation vanish since for any vector ζ holds that $\sum_i \pi_i \sum_j q_{ij}\zeta_j = \sum_j \zeta_j \sum_i \pi_i q_{ij} = 0$. The term

$$\sqrt{N}\sum_{i=1}^{d}\pi_i\left(-\vartheta\delta_i\frac{\mathrm{d}}{\mathrm{d}t}\varrho(t,u)+\delta_i(\lambda_i e^{-\mu_i t}\vartheta\alpha_1+\lambda_i e^{-\mu_i(t+u)}\vartheta\alpha_2)\right)$$

vanishes since $\sqrt{N}\delta_i^{(N)}(\vartheta,t,u)\to\sqrt{N}\delta(\vartheta,t,u)$ irrespective of the initial state i. This can be seen by multiplying Eqn. (10) by $N^{\frac{1}{2}+\varepsilon}$ and taking the limit $N\to\infty$. We then obtain the differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\delta(\vartheta,t,u)=\vartheta^2\delta(\vartheta,t,u)\sum_{i=1}^{d}\pi_i\lambda_i\left(e^{-\mu_i t}\frac{\alpha_1^2}{2}+e^{-\mu_i(t+u)}\frac{(2\alpha_1+\alpha_2)\alpha_2}{2}\right).$$

A separation of variables argument yields that

$$\delta(\vartheta,t,u)=\exp\left(-\frac{\vartheta^2}{2}\sum_{i=1}^{d}\pi_i\frac{\lambda_i}{\mu_i}\left(e^{-\mu_i t}\alpha_1^2+e^{-\mu_i(t+u)}(2\alpha_1+\alpha_2)\alpha_2\right)\right)K(\vartheta,u),$$

for some function $K(\vartheta,u)$ 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 $K(\vartheta,u)$. We obtain

$$\begin{aligned}\delta(\vartheta,t,u) &= \exp\left(\frac{\vartheta^2}{2}\sum_{i=1}^{d}\pi_i\frac{\lambda_i}{\mu_i}\left((1-e^{-\mu_i t})\alpha_1^2+(1-e^{-\mu_i t})e^{-\mu_i u}2\alpha_1\alpha_2+(1-e^{-\mu_i(t+u)})\alpha_2^2\right)\right)\\ &= \exp\left(\frac{\vartheta^2}{2}v(t,u)\right),\end{aligned}$$

as desired. We have proven the convergence (8).

4.2. **General case.** In the M/G/∞ queue, starting empty, 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. As before we first realize that it suffices to compute $\mathbb{C}\mathrm{ov}(N(t),N_t(t+u))$. First define

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

the first of these quantities can be interpreted as the probability that an arbitrary particle that has arrived in $[0,t)$ is still present 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) $q^{\rm C}\,\lambda t + q^{\rm C}(1-q^{\rm A})\lambda^2 t^2$. As $\mathbb{E}N(t) = (1-q^{\rm A})\lambda t$ and $\mathbb{E}N_t(t+u) = q^{\rm C}\,\lambda t$, it follows that

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

This computation provides us with the candidate for the central limit result in the case of general service times. Define in this context

$$\varrho_t = \sum_{i=1}^d \pi_i\lambda_i\int_0^t \bar{F}_i(v){\rm d}v,\quad c_{t_1,t_2} = \sum_{i=1}^d \pi_i\lambda_i\int_0^t \bar{F}_i(v+u){\rm d}v.$$

Analogously to the case of exponential service times we can now prove the following result; the structure of the proof is exactly identical, but the notation is more cumbersome. Notice that the extension from the bivariate case (time epochs t and u) to a general dimension (time epochs t_1 up to t_K) is straightforward and essentially a matter of careful bookkeeping. The final result now reads as follows.

Theorem 1. *For any $\boldsymbol{\alpha}\in\mathbb{R}^K$ and $\boldsymbol{t}\in\mathbb{R}^K$, and general state-dependent service times, as $N\to\infty$,*

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

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) infinite-server queue in which the Poisson arrival rate is scaled by N; the limiting process is then an Ornstein-Uhlenbeck process — see e.g. [17]. 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 (Markov modulated M/G/∞ queue) we have proven the first two steps; the third step (tightness) is beyond the scope of this paper.

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)

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

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 [8].

We now focus on computing the second moment; for ease we consider the stationary case. From Section 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 $q := q_1 + q_2$,

$$\mathbb{E}\bar{M}^2(\infty) = \frac{\pi_1 r_1}{\mu-q}\left((\mu-q_2)r_1-q_1r_2\right)+\pi_1r_1+\frac{\pi_2r_2}{\mu-q}\left((\mu-q_1)r_2-q_2r_1\right)+\pi_2r_2.$$

We now claim that, with R denoting the stationary mean $\pi_1r_1 + \pi_2r_2$, the stationary variance is larger than this R, or equivalently

$$\mathbb{E}\bar{M}^2(\infty) \geq R^2 + R, \tag{11}$$

with equality only if $\lambda_1 = \lambda_2$. This can be shown as follows. Writing $r_1 = ar_2$, the above claim reduces to verifying that, for all $a \in (0, \infty)$,

$$a^2(f_1-\pi_1)\pi_1+a(f_2-\pi_2)\pi_1+a(g_1-\pi_1)\pi_2+(g_2-\pi_2)\pi_2 \geq 0, \tag{12}$$

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 (12) 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, (12) 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 (11) thus follows. We conclude that $\mathbb{V}\text{ar}\bar{M}(\infty) \geq \mathbb{E}\bar{M}(\infty)$, 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 [8] that this entails that $\mathbb{V}\text{ar}\bar{M}(\infty) \geq \mathbb{E}\bar{M}(\infty)$; 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 Figure 5.2 is the limiting behavior of Eqn. (9), obtained by solving the differential equation (5) with the mgf parameter $\vartheta = 0.5$ and $\varepsilon = 0.5$. The limiting curve derived at the end of Section 4.1 is plotted as well. As in the case with deterministic transition times [2], we observe loglinear convergence, with the solution curve closely following the limiting curve for $N = 1000$. Tweaking the parameters results in the same converging behavior.

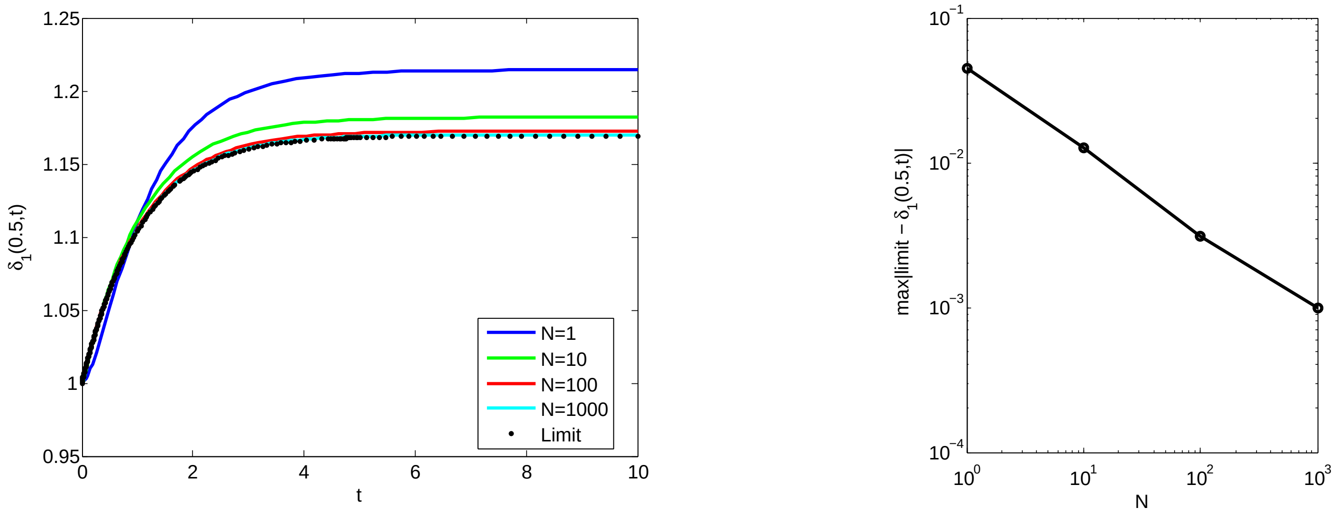

Figure 1: (left) The scaled process approaches the limiting curve as N grows larger. (right) Maximum error as a function of N shows loglinear convergence.

References

[1] S. Asmussen and O. Kella (1996). Rate modulation in dams and ruin problems. *Journal of Applied Probability*, **33**, pp. 523–535.

[2] J. Blom, M. Mandjes and H. Thorsdottir (2012). Time-scaling limits for Markov-modulated infinite-server queues. To appear in *Stochastic Models*.

[3] B. D'Auria (2008). M/M/∞ queues in semi-Markovian random environment. *Queueing Systems*, **58**, 221–237.

[4] A. Dembo and O. Zeitouni (1998). *Large Deviations Techniques and Applications,* 2nd edition. Springer, New York.

[5] B. FRALIX and I. ADAN (2009). An infinite-server queue influenced by a semi-Markovian environment. *Queueing Systems*, **61**, 65–84.

[6] P. GLYNN (1995). Large deviations for the infinite server queue in heavy traffic. *Institute for Mathematics and Its Applications*, **71**, 387–394.

[7] P. GLYNN AND W. WHITT (1991). A new view of the heavy-traffic limit theorem for infinite-server queues. *Advances in Applied Probability*, **23**, 188–209.

[8] T. HELLINGS, M. MANDJES, and J. BLOM (2012). Semi-Markov-modulated infinite-server queues: approximations by time-scaling. *Stochastic Models*, **28**, 452–477.

[9] J. KEILSON and L. SERVI (1993). The matrix $M/M/\infty$ system: retrial models and Markov modulated sources. *Advances in Applied Probability*, **25**, 453–471.

[10] O. KELLA and W. STADJE (2002). Markov-modulated linear fluid networks with Markov additive input. *Journal of Applied Probability*, **39**, pp. 413–420.

[11] L. LIU and J. TEMPLETON (1993). Autocorrelations in infinite server batch arrival queues *Queueing Systems,* **14**, pp. 313–337.

[12] M. MANDJES and A. RIDDER (2001). A large deviations approach to the transient of the Erlang loss model. *Performance Evaluation*, **43**, pp. 181–198.

[13] M. NEUTS (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach.* Johns Hopkins University Press.

[14] M. NEUTS and S. CHEN (1972). The infinite server queue with semi-Markovian arrivals and negative exponential services. *Journal of Applied Probability*, **9**, pp. 178–184.

[15] C. O'CINNEIDE and P. PURDUE (1986). The $M/M/\infty$ queue in a random environment. *Journal of Applied Probability*, **23**, pp. 175–184.

[16] P. PURDUE and D. LINTON (1981). An infinite-server queue subject to an extraneous phase process and related models. (English) *Journal of Applied Probability*, **18**, pp. 236–244.

[17] PH. ROBERT (2003). *Stochastic Networks and Queues.* Springer, Berlin.

[18] A. SCHWABE, K.N. RYBAKOVA, and F.J. BRUGGEMAN (2012). Transcription Stochasticity of Complex Gene Regulation Models. *Biophysical Journal*, **103**, pp. 1152–1161.

[19] W. WHITT (2001). *Stochastic-process Limits.* Springer, New York.

E-mail address: joke.blom@cwi.nl, Offer.Kella@huji.ac.il, M.R.H.Mandjes@uva.nl, halldora@cwi.nl