TIME-SCALING LIMITS FOR MARKOV-MODULATED INFINITE-SERVER QUEUES

J. BLOM [⋆], M. MANDJES [•,⋆], H. THORSDOTTIR [⋆,•]

ABSTRACT. In this paper we study semi-Markov modulated M/M/∞ queues, which are to be understood as infinite-server systems in which the Poisson input rate is modulated by a Markovian background process (where the times spent in each of its states are assumed deterministic), and the service times are exponential. Two specific scalings are considered, both in terms of transient and steady-state behavior. In the former the transition times of the background process are divided by N, and then N is sent to ∞; a Poisson limit is obtained. In the latter both the transition times and the Poissonian input rates are scaled, but the background process is sped up more than the arrival process; here a central-limit type regime applies. The accuracy and convergence rate of the limiting results are demonstrated with numerical experiments.

[•] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, the Netherlands. [⋆] CWI, Amsterdam, the Netherlands. M. Mandjes is also with EURANDOM (Eindhoven University of Technology, the Netherlands).

1. INTRODUCTION

Adding the effect of a random environment to the classical M/M/∞ queue provides us with a natural framework to model various real-life phenomena. In this model an infinite server queue is fed by a Poisson arrival stream whose rate is modulated by a Markovian background process, and the service times are exponential. The resulting model, usually called a *semi-Markov-modulated inifinite-server queue*, is a suitable candidate for several applications, for instance in telecommunication network engineering, where the arrival rates of customers vary between different modes [9]. Another example is the synthesis and later degradation of mRNA strings in cells, after transcription of the DNA which tends to occur in a clustered fashion [8].

In our work we focus on the following variant of the semi-Markov-modulated inifinite-server queue. The Poisson arrivals to the queue have rate λ_i depending on the state i of an external Markovian background process. In our results it is assumed that the service rate μ is not affected by the background process, but in the last section we comment on the case where it is. The background process stays in state i for a deterministic time t_i (this time is usually referred to as *transition time*) — we do indicate, though, how the analysis should be adapted to allow other transition time distributions.

Date: July 10, 2012.

Key words and phrases. Markov-modulated Poisson process, queues, diffusion limit.

Other variants of the semi-Markov-modulated queues, such as the single-server counterpart, have been widely studied (see for example the discussion in the introduction of [4]); considerably less attention has been paid to the infinite-server case. Without aiming at giving a full account of the existing literature, we mention that results for the system's steady-state behavior (mostly in terms of factorial moments) have already been available for some time; see e.g. [1, 6]. Also Markov-modulated infinite server queues with Erlang or hyperexponentially distributed service times have been addressed [3] .
A new line of research started in [4], where time-scalings are imposed so as to obtain explicit expressions for the resulting limiting distribution. The main result of that paper relates to speeding up the transition times by a factor N; the resulting arrival process turns out to be a Poisson process (with a rate λ_∞ that can be given explicitly in terms of the λ_i and the invariant associated with the generator of the background process).
In the present paper we apply two scalings, to which we refer (for obvious reasons) the *Poisson regime* and the *CLT regime.* The first one amounts to dividing the t_i by N, as described above. In the second scaling the rate of arrivals to the system is scaled linearly with N (that is, the arrival rates become $N\lambda_i$), but the transition times are scaled *superlinearly* (that is, they become $t_i/N^{1+\varepsilon}$, for some $\varepsilon > 0$).

The contributions and organization of our paper are as follows. After formally describing our model in Section 2, we study both steady-state and transient behavior of our infinite-server queue, in the Poisson regime. as well as the CLT regime. In more detail, the following results are derived.

- Whereas in [4] we proved that under the Poisson scaling the input process converges to a Poisson process with rate λ_∞, we show in Section 3 that the steady-state number of customers in the system converges to a Poisson distribution with mean μ. The transient variant of this result is established in Section 4. At the methodological level, the argument used is (i) set up a system of equations for the probability generating function (pgf) of the quantity of interest, (ii) then send N to ∞, and obtain a differential equation for the pgf, (iii) and finally conclude the stated from solving the differential equation. Note that the differential equation is in terms of the argument of the pgf in the steady-state case, and in time in the transient case.
- Under the CLT scaling, we obtain results of the diffusion-type. Essentially two effects are combined. By scaling the transition times by $N^{1+\varepsilon}$, by virtue of the findings in [4], the input process converges to a Poisson process with rate λ_∞. The effect of scaling the λ_is as well is that a central-limit type of regime kicks in, as described in e.g. [7, Section 6.6]. As a consequence, the number of customers minus its expected value, divided by $\sqrt{N}$ converges to a zero-mean random variable. Sections 5 and 6 establish the transient and steady-state version of this result, respectively. The underlying argumentation, although considerably more delicate, resembles the one developed for the Poisson regime.

- We have extensively tested the resulting approximation in a set of numerical experiments, to confirm the speed of convergence to the limiting distribution. In Section 7 we present results, showing that the asymptotics lead to quite accurate approximations, already for relatively low N.
- In a discussion section, we comment on a number of extensions: (i) state-dependent service rate, (ii) general transition times, and (iii) large deviations results.

2. Model description

This paper studies an infinite-server queue with semi-Markov-modulated Poisson arrivals and exponential service times. The model can be described as follows.

Consider an irreducible semi-Markov process $X(t)$ on a finite state space $\{1, \dots, d\}$, with $d \in \mathbb{N}$. Its transition matrix is given by $P = (p_{ij})_{i,j=1}^{d}$, where p_{ii} need not necessarily be zero. The time spent in state i is distributed as a non-negative random variable T_i (to be referred to as a *transition time*). The subsequent transition times in state i, say $(T_{i,j})_{j\in\mathbb{N}}$, constitute a sequence of i.i.d. random variables; in addition the sequences $(T_{i,j})_{j\in\mathbb{N}}$, for various $i \in \{1, \dots, d\}$, are assumed independent. There is also independence between the jumps of the semi-Markov process and the transition times. While the process $X(t)$, often referred to as the *background process*, is in state i, customers arrive according to a Poisson process with rate $\lambda_i \geq 0$. The service times are assumed to be exponentially distributed with mean $1/\mu$, irrespective of the state of the background process.

We use bold fonts to denote vectors; for instance $\boldsymbol{\lambda} \equiv (\lambda_1, ..., \lambda_d)$. We denote the invariant distribution corresponding to the transition matrix P by $\boldsymbol{\pi}$.

The main objective of this paper is to analyze the distribution of the number of customers in the system, and in particular under specific scalings. In our analysis, we primarily focus on the case that the T_is equal a deterministic number $t_i > 0$ (unless stated otherwise).

3. Steady-state, Poisson regime

Let, following [4], M_i denote the steady-state number of customers in the system when the background process enters state i. We denote by $\gamma_i(\cdot)$ the probability generating function of M_i: $\gamma_i(z) := \mathbb{E}z^{M_i}$. The probabilities of the time-reversed process, that is of coming from state j, given that the process just jumped to state i, are denoted by $\tilde{p}_{ij} = p_{ji}\pi_j/\pi_i$. Then, from [4, Thm. 2], for the case of deterministic transition times,

$$\gamma_i(z) = \sum_{j=1}^{d} \tilde{p}_{ij} g_j(z) \gamma_j(h_j(z)), \tag{1}$$

with

$$h_j(z) := 1 - e^{-\mu t_j}(1-z), \quad g_j(z) := \exp\left(-\lambda_j \frac{1-e^{-\mu t_j}}{\mu}(1-z)\right).$$

We now scale, as in [4, Section 4.2], $t_i \mapsto t_i/N$ (in self-evident notation), and study the solution for $\gamma_i(z)$ in the above fixed point relation (1). Intuitively, this scaling means that the background process moves fast between the states in the state space, so that it is conceivable that the particle arrival process tends to a Poisson process as N grows large. It was shown in [4, Section 4.2] that this is indeed the case, with associated arrival rate

$$\lambda_\infty := \frac{\sum_{j=1}^d \pi_j \lambda_j t_j}{\sum_{j=1}^d \pi_j t_j}.$$

This means that, under this scaling, the queueing system will resemble an M/M/∞ queue, with arrival rate λ_∞ and departure rate μ; such a queue has a steady-state distribution that is Poisson with mean λ_∞/μ. In this section, we verify this property, predominantly relying on Taylor expansion.

To this end, first observe that, up to and including $O(1/N)$-terms,

$$h_j(z) = z + (1-z)\frac{t_j\mu}{N}, \quad g_j(z) = 1 - \frac{\lambda_j t_j}{N}(1-z).$$

We thus obtain (using the superscript $^{(N)}$ to indicate the dependence on N):

$$(2) \quad \gamma_i^{(N)}(z) = \sum_{j=1}^d \tilde{p}_{ij}\left(1 - \frac{\lambda_j t_j}{N}(1-z)\right)\left(\gamma_j^{(N)}(z) + \left(\gamma_j^{(N)}\right)'(z)\cdot(1-z)\frac{t_j\mu}{N}\right) + O\left(\frac{1}{N^2}\right).$$

Letting $N \to \infty$, we obtain that (provided the limits exist)

$$\lim_{N\to\infty} \gamma_i^{(N)}(z) = \lim_{N\to\infty} \sum_{j=1}^d \tilde{p}_{ij}\gamma_j^{(N)}(z).$$

We conclude that $\lim_{N\to\infty}\gamma_i^{(N)}(z) = \gamma(z)$ for a pgf $\gamma(\cdot)$ that does not depend on i.

Now multiply Eqn. (2) by N, multiply by π_i and sum over i:

$$\begin{aligned} N\sum_{i=1}^d \pi_i \gamma_i^{(N)}(z) &= N\sum_{i=1}^d \pi_i \sum_{j=1}^d \tilde{p}_{ij}\gamma_j^{(N)}(z) \\ &+ \quad \sum_{i=1}^d\sum_{j=1}^d \pi_i\tilde{p}_{ij}\left(-\lambda_j t_j(1-z))\gamma_j^{(N)}(z) + \left(\gamma_j^{(N)}\right)'(z)\cdot(1-z)t_j\mu\right) + O\left(\frac{1}{N}\right). \end{aligned}$$

Note that, for any vector $\boldsymbol{\zeta}$,

$$(3) \qquad \sum_{i=1}^d\sum_{j=1}^d \pi_i\tilde{p}_{ij}\zeta_j = \sum_{j=1}^d \pi_j\zeta_j\sum_{i=1}^d p_{ji} = \sum_{j=1}^d \pi_j\zeta_j,$$

Combining the previous three displays and letting $N \to \infty$ we obtain the differential equation

$$\sum_{j=1}^d \pi_j\lambda_j t_j\gamma(z) = \sum_{j=1}^d \pi_j t_j\mu\gamma'(z).$$

With the requirement that $\gamma(1) = 1$, it is trivial to deduce that

$$\gamma(z) = \exp\left(\frac{\lambda_\infty}{\mu}(z-1)\right),$$

corresponding to the Poisson distribution with mean $\varrho := \lambda_\infty/\mu$. The following result summarizes the findings of this section; $\mathbb{P}\mathrm{ois}(\nu)$ denotes a Poisson random variable with mean ν.

Theorem 1. *Under the scaling $t_i \mapsto t_i/N$,*

$$M_i^{(N)} \xrightarrow{\mathrm{d}} \mathbb{P}\mathrm{ois}(\varrho).$$

4. Transient, Poisson regime

Again, we let the sojourn times be t_i/N. The number still present at time t out of the initial population $x_0 \in \mathbb{N}$ does *not* depend on the background process (as the departure rate is state-independent). This random variable, say $\check{M}^{(N)}(t)$, has a binomial distribution with parameters x_0 and $e^{-\mu t}$. We therefore focus on the number of customers arriving in $(0,t]$ that are still present at time t, of which we evidently know that it is independent of $\check{M}^{(N)}(t)$. Let $\bar{M}_i^{(N)}(t)$ be the number of these, given the modulating process is in state i at time 0, and let

$$\bar{\gamma}_i^{(N)}(z,t) := \mathbb{E}z^{\bar{M}_i^{(N)}(t)}$$

be the corresponding pgf. The primary objective of this section is to show that $\bar{M}_i^{(N)}(t)$ converges in distribution to a Poisson random variable with mean $\varrho(1-e^{-\mu t})$, thus identifying the limiting distribution of $M_i^{(N)}(t) := \check{M}^{(N)}(t) + \bar{M}_i^{(N)}(t)$ as $N \to \infty$.
Define $\varrho_t := (\lambda_\infty/\mu)\cdot(1-e^{-\mu t})$. An elementary conditioning argument yields that

$$\bar{\gamma}_i^{(N)}(z,t) = \sum_{k=0}^{\infty} e^{-\lambda_i t_i/N}\frac{(\lambda_i t_i/N)^k}{k!}(p_i^{(N)}(z,t))^k \sum_{j=1}^{d} p_{ij}\bar{\gamma}_j^{(N)}\left(z, t-\frac{t_i}{N}\right). \tag{4}$$

The $p_i^{(N)}(z,t)$ are pgf s of random variables that are alternatively distributed on 0 and 1, with the probability of equalling 1 being, up to and including $O(N^{-1})$-terms,

$$\int_0^{t_i/N}\frac{1}{t_i/N}\int_{t-u}^{\infty}\mu e^{-\mu v}\mathrm{d}v\mathrm{d}u = e^{-\mu t} + \frac{1}{2}e^{-\mu t}\frac{\mu t_i}{N}.$$

We thus obtain, neglecting terms of order $O(N^{-2})$,

$$\begin{aligned}\bar{\gamma}_i^{(N)}(z,t) &= \exp\left(\frac{\lambda_i t_i}{N}e^{-\mu t}(z-1)\right)\sum_{j=1}^{d}p_{ij}\left(\bar{\gamma}_j^{(N)}(z,t) - \frac{t_i}{N}\frac{\mathrm{d}}{\mathrm{d}t}\bar{\gamma}_j^{(N)}(z,t)\right)\\ &= \left(1+\frac{\lambda_i t_i}{N}e^{-\mu t}(z-1)\right)\sum_{j=1}^{d}p_{ij}\left(\bar{\gamma}_j^{(N)}(z,t) - \frac{t_i}{N}\frac{\mathrm{d}}{\mathrm{d}t}\bar{\gamma}_j^{(N)}(z,t)\right). \qquad (5)\end{aligned}$$

Letting $N \to \infty$, we conclude again that $\lim_{N\to\infty}\gamma_i^{(N)}(z,t) = \gamma(z,t)$ for a pgf $\gamma(\cdot,t)$ that does not depend on i. Now multiply Eqn. (5) by N, multiply by π_i and sum over i. Due to the 'p_{ij} analogue' of (3) the $O(N)$-terms cancel. By virtue of the state independence of $\gamma(\cdot,t)$, we obtain when sending $N \to \infty$,

$$\lambda_\infty e^{-\mu t}(z-1)\gamma(z,t) = \frac{\mathrm{d}}{\mathrm{d}t}\gamma(z,t),$$

which leads, in conjunction with the requirement $\gamma(z,0)=1$, to

$$\gamma(z,t)=\exp\left(\frac{\lambda_\infty}{\mu}(1-e^{-\mu t})(z-1)\right),$$

corresponding to a Poisson distribution with mean ϱ_t. We have thus derived the following limiting distribution for the transient number of customers in the system.

Theorem 2. *Under the scaling $t_i\mapsto t_i/N$,*

$$M_i^{(N)}(t)\xrightarrow{\rm d}\mathbb{B}{\rm in}(x_0,e^{-\mu t})+\mathbb{P}{\rm ois}(\varrho_t),$$

where the random variables in the right-hand side are independent.

5. Steady-state, CLT regime

In, e.g., [7, Section 6.6] an M/M/∞ queue is observed under a linear scaling of the arrival rate λ, that is, one scales $\lambda\mapsto\lambda N$. With Nx_0 customers being present at time 0, a central-limit-theorem (CLT) type of result is proven. More specifically, it is shown that the number of customers in this M/M/∞ system at time t, minus its expected value $N\,m(\lambda,\mu)$, and divided by $\sqrt{N}$, tends to a zero-mean Normal random variable, with variance $v(\mu,\lambda)$; here

$$m(\lambda,\mu):=x_0e^{-\mu t}+N\lambda/\mu\cdot(1-e^{-\mu t}),\quad v(\lambda,\mu):=x_0e^{-\mu t}(1-e^{-\mu t})+\lambda/\mu\cdot(1-e^{-\mu t}).$$

(In fact, [7, Thm. 6.14] provides us with a considerably more refined result: a functional central limit theorem. More precisely, there is weak convergence of the queueing process to a specific Gaussian process, viz. an Ornstein-Uhlenbeck process.)

The idea of this section (as well as the next section) is that we scale the arrival rate linearly ($\lambda_i\mapsto\lambda_iN$, that is), but we scale the transition times *superlinearly* ($t_i\mapsto t_i/N^{1+\varepsilon}$, for some $\varepsilon>0$). The effect of this scaling is that we combine the convergence of the particle arrival process to a Poisson process of rate λ_∞ (essentially as in [4, Section 4.2]) with the CLT regime kicking in (as in [7, Section 6.6]). As a consequence, one would expect convergence of the steady-state number of customers, minus $N\,m(\lambda_\infty,\mu)$, divided by $\sqrt{N}$, to a zero-mean Normal random variable with variance $v(\lambda_\infty,\mu)$. The objective of this section is to verify the steady-state counterpart of this claim (in which the variance is λ_∞/μ), where the transient version is established in the next section.

As mentioned above, we now let the arrival rates be λ_iN, and the sojourn times $t_i/N^{1+\varepsilon}$. Define, with $\varrho:=\lambda_\infty/\mu$,

$$\delta_i^{(N)}(\vartheta):=\mathbb{E}\left(\exp\left(\frac{\vartheta M_i^{(N)}-\vartheta N\varrho}{\sqrt{N}}\right)\right)=\gamma_i^{(N)}\left(e^{\vartheta/\sqrt{N}}\right)\cdot e^{-\vartheta\sqrt{N}\varrho},$$

where $\gamma_i^{(N)}(\cdot)$ is the probability generating function of $M_i^{(N)}$. From [4, Thm. 2], for deterministic transition times,

$$\delta_i^{(N)}(\vartheta) = \sum_{j=1}^{d} \tilde{p}_{ij} g_j \left(e^{\vartheta/\sqrt{N}}\right) \gamma_j^{(N)} \left(h_j \left(e^{\vartheta/\sqrt{N}}\right)\right) \cdot e^{-\vartheta\sqrt{N}\varrho}.$$

Elementary Taylor computations yield:

$$\begin{aligned} g_j \left(e^{\vartheta/\sqrt{N}}\right) &= \exp\left(-N\lambda_j \left(\frac{t_j}{N^{1+\varepsilon}} - \frac{\mu t_j^2}{2N^{2+2\varepsilon}}\right)\left(-\frac{\vartheta}{\sqrt{N}} - \frac{\vartheta^2}{2N}\right)\right) \\ &= 1 + \frac{\lambda_j t_j \vartheta}{N^{\frac{1}{2}+\varepsilon}} + \frac{\lambda_j t_j \vartheta^2}{2N^{1+\varepsilon}} + O\left(\frac{1}{N^{\frac{3}{2}+2\varepsilon}}\right), \end{aligned}$$

Likewise,

$$h_j \left(e^{\vartheta/\sqrt{N}}\right) = e^{\vartheta/\sqrt{N}} + \frac{\mu t_j}{N^{1+\varepsilon}}\left(1 - e^{\vartheta/\sqrt{N}}\right) + O\left(\frac{1}{N^{\frac{5}{2}+2\varepsilon}}\right).$$

Using the latter Taylor approximation, we obtain after elementary computations that

$$\gamma_j^{(N)}\left(h_j\left(e^{\vartheta/\sqrt{N}}\right)\right) \cdot e^{-\vartheta\sqrt{N}\varrho} = \delta_j^{(N)}(\vartheta) + \left[\left(\gamma_j^{(N)}\right)'(e^{\vartheta/\sqrt{N}})\right] \cdot \frac{\mu t_j}{N^{1+\varepsilon}}\left(1 - e^{\vartheta/\sqrt{N}}\right) e^{-\vartheta\sqrt{N}\varrho}.$$

Obviously,

$$(\delta_j^{(N)})'(\vartheta) = \frac{1}{\sqrt{N}} e^{\vartheta/\sqrt{N}} \left[\left(\gamma_j^{(N)}\right)'(e^{\vartheta/\sqrt{N}})\right] \cdot e^{-\vartheta\sqrt{N}\varrho} - \sqrt{N}\varrho\delta_j^{(N)}(\vartheta);$$

equivalently,

$$\left[\left(\gamma_j^{(N)}\right)'(e^{\vartheta/\sqrt{N}})\right] \cdot e^{-\vartheta\sqrt{N}\varrho} = \sqrt{N}e^{-\vartheta/\sqrt{N}}(\delta_j^{(N)})'(\vartheta) + N\varrho e^{-\vartheta/\sqrt{N}}\delta_j^{(N)}(\vartheta).$$

Combining the above, we arrive at (neglecting $O(N^{-3/2-2\varepsilon})$-terms)

$$\begin{aligned} (6) \quad \delta_i^{(N)}(\vartheta) = \sum_{j=1}^{d} \Bigg[\tilde{p}_{ij}\left(1 + \frac{\lambda_j t_j \vartheta}{N^{\frac{1}{2}+\varepsilon}} + \frac{\lambda_j t_j \vartheta^2}{2N^{1+\varepsilon}}\right) \\ \left(\delta_j^{(N)}(\vartheta) + \left(\sqrt{N}e^{-\vartheta/\sqrt{N}}(\delta_j^{(N)})'(\vartheta) + N\varrho e^{-\vartheta/\sqrt{N}}\delta_j^{(N)}(\vartheta)\right)\frac{\mu t_j}{N^{1+\varepsilon}}\left(1 - e^{\vartheta/\sqrt{N}}\right)\right)\Bigg]. \end{aligned}$$

Now multiply (6) with $N^{1+\varepsilon}\pi_i$ and sum over i to obtain, relying on Eqn. (3) and some Tayloring,

$$\begin{aligned} N^{1+\varepsilon}\sum_{i=1}^{d} \pi_i \delta_i^{(N)}(\vartheta) = \sum_{j=1}^{d} \pi_j \left(1 + \frac{\lambda_j t_j \vartheta}{N^{\frac{1}{2}+\varepsilon}} + \frac{\lambda_j t_j \vartheta^2}{2N^{1+\varepsilon}}\right) \\ \left(N^{1+\varepsilon}\delta_j^{(N)}(\vartheta) + \left(\sqrt{N}(\delta_j^{(N)})'(\vartheta) + N\varrho\delta_j^{(N)}(\vartheta)\right)\mu t_j\left(-\frac{\vartheta}{\sqrt{N}} + \frac{\vartheta^2}{2N}\right)\right). \end{aligned}$$

Simplifying we arrive at

$$\text{(7)}\quad \sum_{j=1}^{d}\pi_j\mu t_j\vartheta(\delta_j^{(N)})'(\vartheta)$$
$$=\sum_{j=1}^{d}\pi_j\left(\mu t_j\vartheta^2\frac{\varrho}{2}+\frac{\lambda_j t_j\vartheta^2}{2}-\sqrt{N}(\mu t_j\vartheta\varrho-\lambda_j t_j\vartheta)\right)\delta_j^{(N)}(\vartheta)+O\left(\frac{1}{N^{\varepsilon}}\right)+O\left(\frac{1}{\sqrt{N}}\right).$$

Note that, on multiplying Eqn. (6) with $\sqrt{N}$, we obtain that

$$\lim_{N\to\infty}\sqrt{N}\left(\delta_i^{(N)}(\vartheta)-\sum_{j=1}^{d}\tilde{p}_{ij}\delta_j^{(N)}(\vartheta)\right)=0;$$

we thus conclude that also $\sqrt{N}\delta_i^{(N)}(\vartheta)$ is independent of i in the limit $N\to\infty$. In addition, due to the very definition of ϱ, the $\sqrt{N}$ terms of Eqn. (7) cancel when sending N to ∞. We consequently obtain the differential equation

$$\vartheta\delta(\vartheta)\left(\sum_{j=1}^{d}\frac{\pi_j t_j\lambda_j}{2}+\frac{\pi_j t_j\mu\varrho}{2}\right)=\delta'(\vartheta)\sum_{j=1}^{d}\pi_j t_j\mu,$$

which reduces to $\varrho\vartheta\delta(\vartheta)=\delta'(\vartheta)$. Solving this ordinary differential equation with $\delta(0)=1$ yields $\delta(\vartheta)=e^{\frac{1}{2}\varrho\vartheta^2}$. We conclude that, as $N\to\infty$, irrespective of i,

$$\frac{M_i^{(N)}-N\varrho}{\sqrt{N}}$$

converges to a Normally distributed random variable with mean 0 and variance ϱ. Denoting by $\mathbb{N}\mathrm{orm}(\nu,\sigma^2)$ a Normal random variable with mean ν and variance σ^2, we have established the following result.

Theorem 3. *Under the scaling $t_i\mapsto t_i/N^{1+\varepsilon}$ and $\lambda_i\mapsto\lambda_i N$,*

$$\frac{M_i^{(N)}-N\varrho}{\sqrt{N}}\xrightarrow{\mathrm{d}}\mathbb{N}\mathrm{orm}(0,\varrho).$$

6. TRANSIENT, CLT REGIME

As in the previous section, we let the arrival rates be $\lambda_i N$, and the sojourn times $t_i/N^{1+\varepsilon}$. We already observed that the number $\check{M}^{(N)}(t)$ of customers still present at time t, out of the initial population of size Nx_0, is not affected by the evolution of the background process (as the departure rate is state-independent). This random variable has a binomial distribution with parameters Nx_0 and $e^{-\mu t}$, and therefore

$$\frac{\check{M}^{(N)}(t)-Nx_0e^{-\mu t}}{\sqrt{N}}\to\mathbb{N}\mathrm{orm}\left(0,x_0e^{-\mu t}(1-e^{-\mu t})\right).$$

In the light of the above remark, we can focus on the number of customers 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 $\bar{M}_i^{(N)}(t)$; as mentioned earlier, $\bar{M}_i^{(N)}(t)$ is independent of $\check{M}^{(N)}(t)$. The objective of the present section is to analyze $\bar{M}_i^{(N)}(t)$ in the CLT regime.

Recall that $\varrho_t := (\lambda_\infty/\mu)\cdot(1-e^{-\mu t})$. Then, with $\bar{\gamma}_i^{(N)}(\cdot,t)$ now denoting the moment generating function of $\bar{M}_i^{(N)}(t)$, we define

$$\bar{\delta}_i^{(N)}(\vartheta,t) := \mathbb{E}\left(\exp\left(\frac{\vartheta \bar{M}_i^{(N)}(t) - \vartheta N\varrho_t}{\sqrt{N}}\right)\right) = \bar{\gamma}_i^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t\right)\cdot e^{-\vartheta\sqrt{N}\varrho_t}. \tag{8}$$

As before,

$$\bar{\gamma}_i^{(N)}(\vartheta,t) = \sum_{k=0}^{\infty} e^{-\lambda_i t_i N^{-\varepsilon}}\frac{(\lambda_i t_i N^{-\varepsilon})^k}{k!}(p_i^{(N)}(\vartheta,t))^k \sum_{j=1}^{d} p_{ij}\bar{\gamma}_j^{(N)}\left(\vartheta, t-\frac{t_i}{N^{1+\varepsilon}}\right).$$

The $p_i^{(N)}(\vartheta,t)$ are mgf s of random variables that are alternatively distributed on 0 and 1, with the probability of equalling 1 being

$$\int_0^{t_i/N^{1+\varepsilon}} \frac{1}{t_i/N^{1+\varepsilon}}\int_{t-u}^{\infty}\mu e^{-\mu v}\mathrm{d}v\mathrm{d}u = e^{-\mu t} + \frac{1}{2}e^{-\mu t}\frac{\mu t_i}{N^{1+\varepsilon}} + O\left(\frac{1}{N^{2+2\varepsilon}}\right).$$

Consequently,

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

and

$$\begin{aligned}&\sum_{k=0}^{\infty} e^{-\lambda_i t_i N^{-\varepsilon}}\frac{(\lambda_i t_i N^{-\varepsilon})^k}{k!}\left(p_i^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t\right)\right)^k\\ &\quad= 1+\frac{\lambda_i t_i}{N^{\varepsilon}}e^{-\mu t}\left(\frac{\vartheta}{\sqrt{N}}+\frac{\vartheta^2}{2N}\right) + O\left(\frac{1}{N^{\frac{3}{2}+\varepsilon}}\right) + O\left(\frac{1}{N^{1+2\varepsilon}}\right).\end{aligned}$$

In addition, neglecting higher-order terms as before,

$$\bar{\gamma}_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}}, t-\frac{t_i}{N^{1+\varepsilon}}\right) = \bar{\gamma}_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t\right) - \frac{\mathrm{d}}{\mathrm{d}t}\bar{\gamma}_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t\right)\cdot\frac{t_i}{N^{1+\varepsilon}}.$$

Upon combining the above,

$$\begin{aligned}\bar{\delta}_i^{(N)}(\vartheta,t) &= \left(1+\frac{\lambda_i t_i}{N^{\varepsilon}}e^{-\mu t}\left(\frac{\vartheta}{\sqrt{N}}+\frac{\vartheta^2}{2N}\right)\right)\times\\ &\qquad \sum_{j=1}^{d} p_{ij}\left(\bar{\gamma}_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t\right) - \frac{\mathrm{d}}{\mathrm{d}t}\bar{\gamma}_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t\right)\cdot\frac{t_i}{N^{1+\varepsilon}}\right)e^{-\vartheta\sqrt{N}\varrho_t}\\ &= \left(1+\frac{\lambda_i t_i}{N^{\varepsilon}}e^{-\mu t}\left(\frac{\vartheta}{\sqrt{N}}+\frac{\vartheta^2}{2N}\right)\right)\times\\ &\qquad \sum_{j=1}^{d} p_{ij}\left(\bar{\delta}_j^{(N)}(\vartheta,t) - \frac{\mathrm{d}}{\mathrm{d}t}\bar{\gamma}_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t\right)\cdot\frac{t_i}{N^{1+\varepsilon}}e^{-\vartheta\sqrt{N}\varrho_t}\right).\end{aligned}$$

From the definition of $\bar{\delta}_i^{(N)}(\vartheta, t)$ we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\bar{\gamma}_j^{(N)}\left(\frac{\vartheta}{\sqrt{N}}, t\right) \cdot \frac{t_i}{N^{1+\varepsilon}} e^{-\vartheta\sqrt{N}\varrho_t} = \frac{\mathrm{d}}{\mathrm{d}t}\bar{\delta}_j^{(N)}(\vartheta, t)\frac{t_i}{N^{1+\varepsilon}} + \bar{\delta}_j^{(N)}(\vartheta, t)\, \vartheta \varrho_t' \frac{t_i}{N^{\frac{1}{2}+\varepsilon}}.$$

Similar to the analysis presented in the previous section, we note that

$$\lim_{N\to\infty} \sqrt{N}\left(\bar{\delta}_i^{(N)}(\vartheta, t) - \sum_{j=1}^{d} p_{ij}\bar{\delta}_j^{(N)}(\vartheta, t)\right) = 0.$$

In line with the preceding sections, we multiply the equation by $N^{1+\varepsilon}\pi_i$ and sum over i. Due to the 'p_{ij}-analogue' of Eqn. (3) we obtain

$$\begin{aligned} &\sum_{j=1}^{d}\sum_{i=1}^{d} \pi_i p_{ij} t_i \frac{\mathrm{d}}{\mathrm{d}t}\bar{\delta}_j^{(N)}(\vartheta, t) \\ &= \vartheta \sum_{j=1}^{d}\sum_{i=1}^{d} \pi_i p_{ij} t_i \left(\lambda_i e^{-\mu t}\frac{\vartheta}{2} + \lambda_i e^{-\mu t}\sqrt{N} - \varrho_t'\sqrt{N}\right)\bar{\delta}_j^{(N)}(\vartheta, t), \end{aligned} \tag{9}$$

in addition to terms that are vanishing as $N \to \infty$ (where it can be verified that the dominating terms of those are of the order of either $N^{-\varepsilon}$ or $N^{-\frac{1}{2}}$, as will be confirmed in the numerical experiments reported on in Section 7).

Combining the above, realizing that $\sqrt{N}\bar{\delta}_i^{(N)}(\vartheta, t)$ is independent of i in the limit $N \to \infty$, and remarking that the $\sqrt{N}$ terms cancel due to the definition of ϱ_t, we obtain:

$$\bar{\delta}(\vartheta, t) \cdot \frac{\vartheta^2}{2} e^{-\mu t} \sum_{i=1}^{d} \pi_i \lambda_i t_i = \frac{\mathrm{d}}{\mathrm{d}t}\bar{\delta}(\vartheta, t) \cdot \sum_{i=1}^{d} \pi_i t_i.$$

Solving this differential equation, and using that $\bar{\delta}(\vartheta, 0) = 1$, we obtain $\bar{\delta}(\vartheta, t) = e^{\frac{1}{2}\varrho_t\vartheta^2}$. Conclude that, as $N \to \infty$, irrespective of i, the random variable

$$\frac{\bar{M}_i^{(N)}(t) - N\varrho_t}{\sqrt{N}}$$

converges to a Normally distributed random variable with mean 0 and variance ϱ_t. Taking into account the contribution of the Nx_0 customers that were already present at time 0, our findings can be summarized in the following statement.

Theorem 4. *Under the scaling $t_i \mapsto t_i/N^{1+\varepsilon}$ and $\lambda_i \mapsto \lambda_i N$,*

$$\frac{M_i^{(N)}(t) - Nx_0e^{-\mu t} - N\varrho_t}{\sqrt{N}} \xrightarrow{\mathrm{d}} \mathbb{N}\mathrm{orm}\left(0, x_0e^{-\mu t}(1 - e^{-\mu t}) + \varrho_t\right).$$

7. Computational results

This section contains numerical results corresponding to the limiting regimes studied in Sections 4 and 6. We compare the resulting approximations with the explicit solutions to Eqns. (4) and (8), for a range of values of the scaling parameter N. To enable easy numerical evaluation of (4) and (8), we assume that the deterministic transition times are

equal: $t_i = 1$ for all $i \in \{1, \dots, d\}$; as a consequence, computing the pgf of the transient number of customers present reduces to elementary matrix multiplications.
We consider a two state system, with arrival rates $\lambda_1 = 1$ and $\lambda_2 = 2$, and service rate $\mu = 1$. The probability transition matrix is

$$P = \begin{pmatrix} 0.2 & 0.8 \\ 0.7 & 0.3 \end{pmatrix}.$$

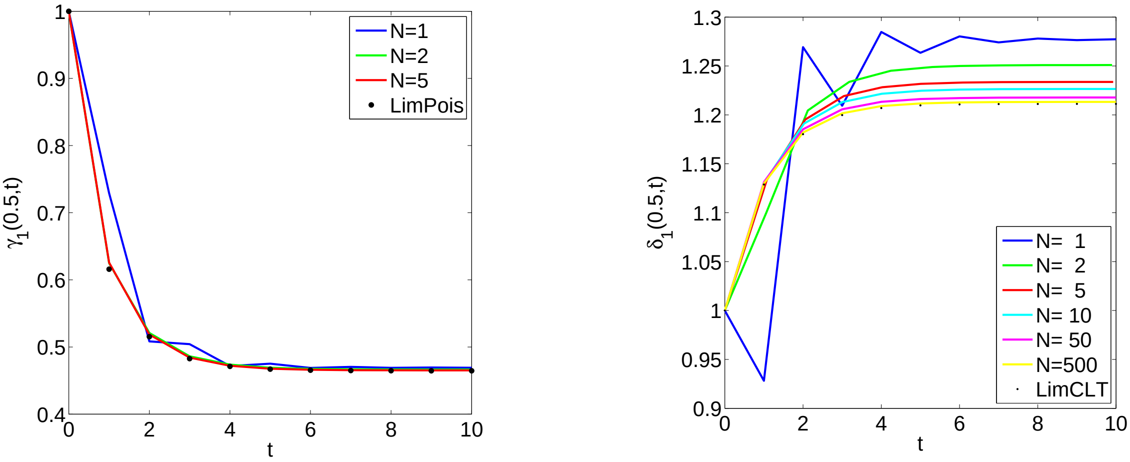

Figure 1: (*Left panel*) The transient Poisson regime converging to its limiting curve, $z = 0.5$. (*Right panel*) Convergence of the transient CLT regime to its limiting curve, $\vartheta = 0.5$, $\varepsilon = 0.5$.

In Fig. 1 we present the results for the Poisson regime (left) and the CLT regime (right). The two graphs demonstrate the convergence behavior to the limiting curve. In the left panel we see that the Poisson regime converges very quickly (at $N = 5$ the maximum error is just $O(10^{-3})$), whereas from the right panel it is observed that in the CLT regime a substantially larger value of N is required to reach the same accuracy level.
For the CLT regime we note that the solution curve, $\bar{\delta}_1^{(N)}(\vartheta, t)$, corresponding to $N = 1$ exhibits large jumps. These can be explained by the specific choice of the matrix P, for which jumping between the two states is highly probable. In fact, the complementing solution curve for $\bar{\delta}_2^{(N)}(\vartheta, t)$ (not depicted here), exhibits jumps in the opposite direction at the early stages.

To get insight into the rate of convergence we look at the maximum difference between the transient pgf and mgf on the one hand, and the limiting curve on the other hand, over the computed time periods for the two limiting regimes, that is $\max_t |X_i(t) - L(t)|$, where $X_i(t)$ is $\bar{\gamma}_i^{(N)}(z, t)$ in the case of the Poisson regime, and $\bar{\delta}_i^{(N)}(\vartheta, t)$ in the case of the CLT regime, and where $L(t)$ denotes the limiting curve. The maximum difference is depicted as a function of N in the top panel of Fig. 2. For the CLT regime in particular we note how

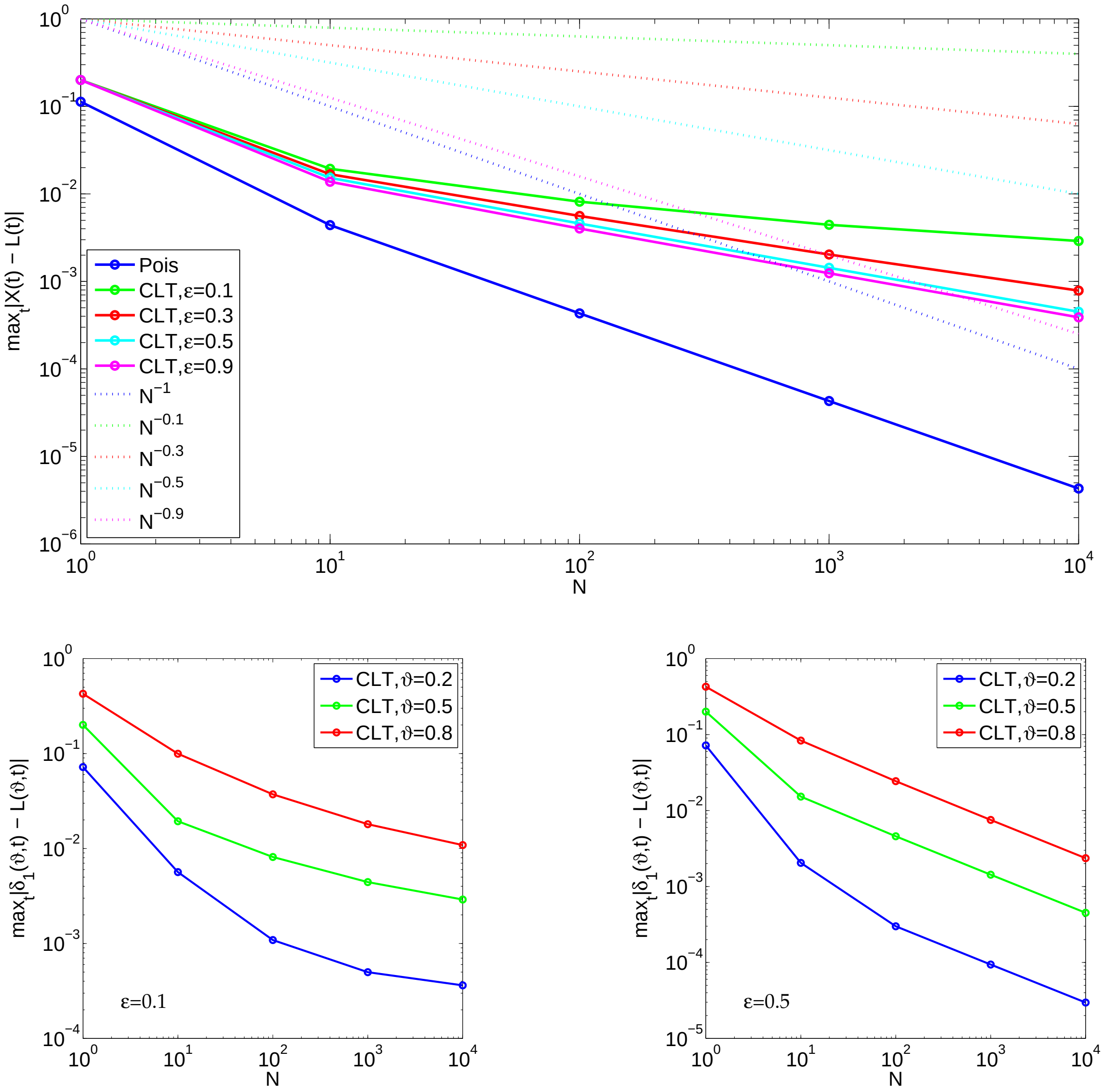

Figure 2: (*Top panel*) Maximum error for both regimes with varying ε as a function of N. $X(t)$ represents the solution curve δ_1 or γ_1 for the CLT and the Poisson regime, respectively. The superimposed dotted lines demonstrate the convergence rate. (*Bottom panels*) Maximum error for varying ϑ, (*left panel*) $\varepsilon = 0.1$, (*right panel*) $\varepsilon = 0.5$ in the CLT regime.

the convergence rate grows with ε until $\varepsilon = 0.5$; from that point on the $\sqrt{N}$ error term takes over, as noted in Section 6).

In the bottom panels of Fig. 2 we see the effect of the choice of ϑ in the mgf for the CLT regime. It is seen that the accuracy improves as ϑ gets smaller, but the convergence rate remains the same for all ϑ.

8. Discussion and concluding remarks

We conclude this paper by discussing a number of extensions: (i) state-dependent service rates, (ii) general transition times, and (iii) large deviations results.

It is not hard to verify that state-dependent service rates can be incorporated. Some care needs to be taken, though. As mentioned in [4], in the steady-state regime we can let the service time depend on the state the background process is currently in. The analysis remains essentially the same; the μ in the definition of $h_j(z)$ is to be replaced by μ_j. Inspection of the proofs however, reveals that it is not straightforward to incorporate this type of state-dependence in the transient cases; it is *not* hard, though, to let in these transient cases the service times depend on the state of the background process upon arrival of the particle (it essentially means that the μ in the definition of $p_i^{(N)}(z,t)$ should be replaced by μ_i).

In [4] it is indicated how the case of general transition times can be addressed in the Poisson regime. The intuition behind the argument is that the probability that a next transition occurs in a small time interval is essentially proportional to the reciprocal of the mean transition time. As a consequence, the same limiting random variable apply, but with the deterministic transition times t_i replaced by the mean transition times $\mathbb{E}T_i$.

The proof method applied in this paper can be used to obtain large deviation properties of the customers in the system. By establishing the existence of the limit of the appropriate moment generating function, the Gärtner-Ellis theorem [2, Thm. 2.3.6] can be applied to the random variable under consideration. A simpler variant of the model studied in this paper, is the one in which the arrivals result from N i.i.d. Markov-modulated input streams with transition times scaled with $1/N^{\varepsilon}$ for $\varepsilon > 0$; for ease we let the system start empty. Then the crucial observation is that the number of customers present in this system at time t, say $\bar{M}^{(N)}(t)$, can be written as the sum of N i.i.d. contributions. The corresponding limiting cumulant function can easily be derived following the method of the previous sections; we eventually find for $a \geq \varrho_t$,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(\frac{\bar{M}^{(N)}(t)}{N} \geq a\right) = a - \varrho_t - a \log \frac{a}{\varrho_t};$$

recognize the large-deviations rate function of the Poisson distribution.

REFERENCES

[1] B. D'AURIA (2008) M/M/∞ queues in semi-Markovian random environment. *Queueing Syst.*, **58**, pp. 221-237.

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

[3] B. FRALIX and I. ADAN (2008). An infinite-server queue influenced by a semi-Markovian environment. *Queueing Syst.*, **61**, pp. 65–84.

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

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

[6] C. O'CINNEIDE and P. PURDUE (1986). The M/M/∞ queue in a random environment. *J. Appl. Prob.*, **23**, pp. 175–184.

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

[8] A. SCHWABE, M. DOBRZYŃSKI, K. RYBAKOVA, P. VERSCHURE, and F.J. BRUGGEMAN (2011). Origins of stochastic intracellular processes and consequences for cell-to-cell variability and cellular survival strategies. *Methods in Enzymology*, **500**, pp. 597-625.

[9] W. WHITT (2002). *Stochastic-Process Limits.* Springer, New York.

E-mail address: joke.blom@cwi.nl, M.R.H.Mandjes@uva.nl, halldora@cwi.nl