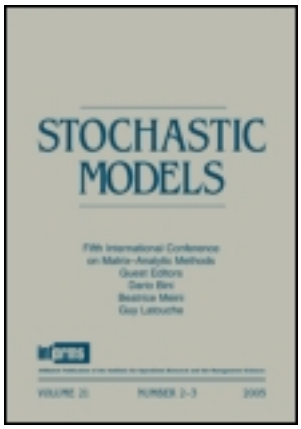
Semi-Markov-modulated infinite-server queues: approximations by time-scaling

Hellings, T.; Mandjes, M.; Blom, J.

DOI
10.1080/15326349.2012.699759
Publication date
2012
Document Version
Final published version
Published in
Stochastic Models



Stochastic Models
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lstm20

Semi-Markov-Modulated Infinite-Server Queues: Approximations by Time-Scaling

Ton Hellings [a], Michel Mandjes [a] & Joke Blom [a]
[a] Centram Wiskunde and Information, Amsterdam, The Netherlands
Version of record first published: 01 Aug 2012.

Stochastic Models, 28:452–477, 2012

ISSN: 1532-6349 print/1532-4214 online
DOI: 10.1080/15326349.2012.699759

Taylor & Francis
Taylor & Francis Group

SEMI-MARKOV-MODULATED INFINITE-SERVER QUEUES: APPROXIMATIONS BY TIME-SCALING

Ton Hellings, Michel Mandjes, and Joke Blom

Centram Wiskunde and Information, Amsterdam, The Netherlands

□ *This article studies an infinite-server queue in a semi-Markov environment: the queue's input rate is modulated by a semi-Markovian background process, and the service times are assumed to be exponentially distributed. The primary objective of this article is to propose approximations for the queue-length distribution, based on time-scaling arguments. The analysis starts with an explicit analysis of the cases in which the transition times of the modulating semi-Markov process are either all deterministic or all exponential. We use these results to obtain approximations under time-scalings; both a quasi-stationary regime (in which time is slowed down) and a fluid-scaling regime (in which time is sped up) are considered. Notably, in the latter regime, the limiting distribution of the number of customers present is Poisson, irrespective of the distribution of the transition times. The accuracy of the resulting approximations is illustrated by several numerical experiments, that moreover give an indication of the speed of convergence in both regimes, for various distributions of the transition times. The last section derives conditions under which the distribution of the number of customers present is Poisson (in an exact sense, i.e., not in a limiting regime).*

Keywords Infinite-server systems; Laplace transforms; Markov modulation; Queues.

Mathematics Subject Classification Primary 60K25, 60K37; Secondary 44A10.

1. INTRODUCTION

The infinite-server queue has proven to be an extremely useful model, being applicable in many contexts. It describes units of work ('customers', in queueing language) 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 mean $1/\mu$

Received February 03, 2011; Accepted October, 22 2011
Address correspondence to Ton Hellings, CWI, P.O. Box 94079, NL-1090 GB, Amsterdam, The Netherlands; E-mail: tonhellings@gmail.com

(the so-called M/G/∞ queue), it is known that the stationary number of customers in the system has a Poisson distribution with mean λ/μ. In fact, the transient behaviour of this M/G/∞ queue is well understood: conditional on the number of customers present at time 0, the distribution of the number of customers at time $t > 0$ is known[6].

The analysis complicates considerably if the model assumptions are relaxed. If the arrival process is of the renewal type, for instance, the steady-state distribution of the resulting GI/G/∞ queue cannot be explicitly computed. Various limiting results are available though, in terms of a central limit theorem under a specific scaling, see Ref.[4], as well as large-deviations results, see Ref.[3].

Another relevant variant, on which we focus in the present article, allows some 'burstiness' in the arrivals. The arrivals occur according to a Poisson process, but the arrival rate is determined by the state of an external semi-Markov process, which we also refer to as the 'background process.' More precisely, with $X(t)$ denoting an irreducible continuous-time semi-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 'service time') has an exponential distribution. Here 'semi-Markov' refers to the class of processes in which the transition times (i.e., the sojourn times in the individual states of the background process) can stem from *any* distribution on $\mathbb{R}_+$ (i.e., not necessarily the exponential distribution), while the process jumps between these states in a Markovian manner.

The resulting model could be called a *semi-Markov-modulated* M/M/∞ *queue*, or an infinite-server queue in a semi-Markov-modulated random environment (for ease we often leave out 'semi' in the sequel). This type of system can be used in several application domains. Suppose for instance that users of a specific service in a communication network occupy one unit of resource while being present (to be thought of as a telephone line, or a given amount of bandwidth); if the arrival rate of these customers alternates between various modes, which is typically the case, the model presented could be used. Another example relates to biology: mRNA strings are synthesized after transcription of the DNA and later degraded in a cell, where the transcription typically tends to occur in a clustered fashion. The proposed model therefore captures the key characteristics of this mechanism well, as argued in Ref.[10].

There is surprisingly little literature on the Markov-modulated infinite-server queue and its variants, compared to the huge literature on Markov-modulated single- and many-server queues. Notably, in the case of exponential transition times and a *single* server, the stationary distribution of the number of customers in the system is of matrix-geometric form[8];

in this sense that system can be viewed as a matrix generalization of the normal M/M/1 queue where the stationary distribution is 'scalar-geometric.' In Ref.[9] the case of exponential transition times and *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 Ref.[5].

The most general result is by D'Auria[1], who finds a recursion for the factorial moments for *general* transition times, i.e., for the semi-Markov-modulated M/M/∞ queue we introduced above. He relies on the observation that the number of customers present has, in the stationary regime, a Poisson distribution *with random parameter*—the computation of this distribution requires a substantial amount of careful analysis though. Fralix and Adan[2] 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.

As mentioned before, the results obtained so far are primarily in terms of (factorial) moments of the queue-length distribution. To facilitate practical use, however, one should get a handle on the distribution itself.

This article proposes approximations for the stationary distribution of the number of clients present in the queueing system, based on two limiting time-scaling regimes. This is done for general transition time distributions, and we furthermore present exact results for deterministic and exponential distributions. The first two sections introduce the problem. In Section 2 the model is defined. Section 3 starts by considering the special case in which the transition times are state-specific but *deterministic*. A very elementary argument provides the factorial moments of the stationary number of customers present; this means that for this special case we do not have to go through the procedure followed in Ref.[1] to get to the same results. Later we also address the case of exponential transition times. This leads to explicit formulae for the factorial moments, in line with those presented in Ref.[9]. Phase-type transition times can be dealt with analogously. The major contributions of the article are the following.

In Section 4 generally distributed transition times are analysed using time-scaling. Both the so-called *quasi-stationary* and *fluid-scaling* regimes are considered. In the former regime, the transition times are divided by a factor n, and then the limiting system corresponding to $n \to 0$ is considered. Our findings indicate that the stationary distribution of the number of customers is 'mixed Poisson', i.e., it is Poisson with mean λ_i/μ with some probability π_i, where π_i is the steady-state probability that the

modulating Markov chain is in state i. Notice that this is a conceivable property, as, due to the time scaling enforced in the quasi-stationary regime, while being in state i the system looks like an ordinary $M/M/\infty$ queue with arrival rate λ_i.

In the latter regime (fluid scaling), the transition times are sped up by a factor n, and n is sent to ∞. The limiting arrival process then turns out to be a Poisson process, with a rate λ_∞ that is a weighted combination of the λ_i. Importantly, this result can be regarded as an insensitivity property, as it holds for *arbitrary* transition time distributions (only the *mean* transition times end up in the expression for λ_∞).

The next section contains a series of numerical experiments for the above regimes. The experiments indicate that there is a rapid convergence to the quasi-stationary and fluid-scaling limits for various distributions of the transition times.

As mentioned above, in Ref.[1] it is shown that the number of customers in the system has, in the stationary regime, a Poisson distribution with random mean. We also mentioned that we obtain a Poisson distribution in the fluid-scaling regime, and this is also the case when $d = 1$. This raises the question: under what conditions is the steady-state distribution Poisson? It is observed that, with X being some non-negative random variable, under the assumption that the random variable Z has a Poisson distribution with (random) mean X,

$$\mathbb{V}\mathrm{ar}[Z] = \mathbb{E}\left[X^2\right] + \mathbb{E}[X] - \mathbb{E}[X]^2 = \mathbb{V}\mathrm{ar}[X] + \mathbb{E}[X] \geq \mathbb{E}[X] = \mathbb{E}[Z],$$

with equality only when X is deterministic. This inequality indicates that approximating the distribution by a Poisson distribution tends to be too optimistic (as it underestimates the variance). In Section 6 we identify conditions under which the Poisson distribution is indeed justified, in that the number of customers has exactly a Poisson distribution (i.e., not in a limiting regime, like in the fluid-scaling studied in Section 4).

2. MODEL DESCRIPTION

In this article we consider an infinite-server queue with semi-Markov-modulated Poisson arrivals and exponential service times. More precisely, the model can be described as follows.

Consider an irreducible semi-Markov process $X(t)$ on a finite state space $\{1, \ldots, d\}$, with $d \in \mathbb{N}$. Its transition matrix is given by $P = \left(p_{ij}\right)_{i,j=1}^{d}$, 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 $i \in \{1,\ldots,d\}$, 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 $\boldsymbol{\lambda} \equiv (\lambda_1,\ldots,\lambda_d)$. We denote the invariant distribution corresponding to the transition matrix P by $\boldsymbol{\pi}$.

In the sequel, we let M_i denote the random variable describing the stationary number of customers present when the background process *enters* state i. The primary objective of this article is to analyze the distribution of M_i for $i = 1,\ldots,d$, and in particular after time-scaling has taken place. For $d = 1$ it will immediately be seen that the process described is actually a classical M/M/∞-queue, and hence M_1 has a Poisson distribution with mean λ_1/μ. In our analysis, special attention is paid to the case that the T_is equal a deterministic number $t_i > 0$ (Section 3); these results are then used to also tackle the case of exponential transition times, while they also facilitate analysis of the quasi-stationary and fluid-scaling regimes for *general* transition times.

3. FIXED-POINT RELATIONS FOR DETERMINISTIC AND EXPONENTIAL TRANSITION TIMES

In this section the probability generating function (PGF) of the M_i, for $i = 1,\ldots,d$, is first analyzed for deterministic transition times; recall that the time the background process spends in state j is t_j. This is done by expressing the PGF of M_i in terms of the PGFs of M_j with $j = 1,\ldots,d$, conditioning on the state from which the background process jumped to state i. This leads to a fixed-point equation that enables the calculation of all moments. Later on in this section the PGF of M_j with exponential transition times will be derived from the deterministic case.

To find the PGF of M_j, we need the probabilities of coming from state j, given that the process just jumped to state i; these are the transition probabilities of the time-reversed process, denoted by $\tilde{p}_{ij} = p_{ji}\pi_j/\pi_i$. Let Y denote the state the semi-Markov process was in *prior to its visit to state* i. This leads to

$$\gamma_i(z) := \mathbb{E}\left[z^{M_i}\right] = \sum_{j=1}^{d} \tilde{p}_{ij}\mathbb{E}\left[z^{M_i} \mid Y = j\right]$$

$$= \sum_{j=1}^{d}\sum_{n=0}^{\infty} \tilde{p}_{ij}\mathbb{E}\left[z^{N_j} \mid M_j = n\right]\mathbb{P}\left[M_j = n\right]; \qquad (1)$$

here $\mathbb{E}\left[z^{N_j} \mid M_j = n\right]$ is the PGF associated with the number of customers present in a *birth-death process* with arrival rate λ_j and service rate μ (per customer), after a time interval of length t_j, conditional on n customers being present at the start of this interval.

Lemma 1. *With* $h_j(z) := 1 - e^{-\mu t_j}(1-z)$ *and* $g_j(z) := e^{-\lambda_j \frac{1-e^{-\mu t_j}}{\mu}(1-z)}$, *we have*

$$\mathbb{E}\left[z^{N_j} \mid M_j = n\right] = (h_j(z))^n g_j(z).$$

Proof. First observe that $(N_j \mid M_j = n)$ can be written as the sum of two independent components: $N_{j,1}$, i.e., the number of the initial n customers that is still present after t_j units of time, and $N_{j,2}$, i.e., the number of arrivals during the period of length t_j that are still in service at the end of this time period. Note that $N_{j,2}$ obviously does not depend on the initial population n.

It is elementary that $(N_{j,1} \mid M_j = n)$ has a binomial distribution with parameters n and $e^{-\mu t_j}$, so that

$$\mathbb{E}\left[z^{N_{j,1}} \mid M_j = n\right] = (h_j(z))^n.$$

We now focus on $N_{j,2}$. First recall that the number of arrivals in the interval has a Poisson distribution with mean $\lambda_j t_j$. Conditional on the number of arrivals, each of them arrives at an epoch uniformly distributed on the interval of length t_j; hence the probability that a given customer is still present at time t_j equals $q(t_j)$, with

$$q(t) := \int_0^t \frac{1}{t} e^{-\mu(t-u)} \mathrm{d}u = \frac{1 - e^{-\mu t}}{\mu t}.$$

It now follows that

$$\mathbb{E}\left[z^{N_{j,2}}\right] = \sum_{k=0}^{\infty} \frac{e^{-\lambda_j t_j} \left(\lambda_j t_j\right)^k}{k!} \sum_{m=0}^{k} z^m \binom{k}{m} \left(q(t_j)\right)^m \left(1 - q(t_j)\right)^{k-m};$$

basic computations show that this equals $g_j(z)$. □

We observe that the PGF $\mathbb{E}\left[z^{N_{j,2}}\right] = g_j(z)$ corresponds to a Poisson random variable with mean $\lambda_j t_j q(t_j)$, which can be understood as follows. The arrival process is a Poisson process with rate λ_j, so that over a period of length t_j the number of arrivals is Poisson distributed with mean $\lambda_j t_j$. However, each of these arrivals is still present after a time interval of length t_j with probability $q(t_j)$. This results in an 'effective mean' of $\lambda_j t_j q(t_j)$.

Eq. (1) and Lemma 1 immediately lead to the following system of fixed-point equations for the $\gamma_i(z)$.

Theorem 2. *For $i = 1, \ldots, d$,*

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

Proof. Observe that

$$\gamma_i(z) = \sum_{j=1}^{d} \tilde{p}_{ij} \sum_{n=0}^{\infty} g_j(z) \left(h_j(z)\right)^n \mathbb{P}\left[M_j = n\right],$$

and the stated follows directly. □

The means of the M_i can be found by differentiating the fixed-point equation of Theorem 2 and inserting $z = 1$. We obtain the following linear system:

$$\mathbb{E}\left[M_i\right] = \sum_{j=1}^{d} \tilde{p}_{ij} \left(e^{-\mu t_j} \mathbb{E}\left[M_j\right] + \left(1 - e^{-\mu t_j}\right) \frac{\lambda_j}{\mu} \right).$$

In fact *all* moments can be derived in this manner. Relying on the standard identity

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k}\left(f(x)g(x)\right) = \sum_{m=0}^{k} \binom{k}{m} f^{(m)}(x) g^{(k-m)}(x).$$

and

$$\frac{\mathrm{d}^k}{\mathrm{d}z^k} \gamma_j\left(h_j(z)\right) = \gamma_j^{(k)}\left(h_j(z)\right) \left(h_j'(z)\right)^k$$

(where it is used that $h_j^{(2)}(z) = 0$), it follows that

$$\gamma_i^{(k)}(z) = \sum_{j=1}^{d} \tilde{p}_{ij} \left(\sum_{m=0}^{k} \binom{k}{m} g_j^{(m)}(z) \gamma_j^{(k-m)}\left(h_j(z)\right) \left(h_j'(z)\right)^{k-m} \right).$$

We thus obtain

$$\gamma_i^{(k)}(1) = \sum_{j=1}^{d} \tilde{p}_{ij} e^{-k\mu t_j} \sum_{m=0}^{k} \binom{k}{m} \left(\frac{\lambda_j}{\mu} \left(e^{\mu t_j} - 1\right) \right)^m \gamma_j^{(k-m)}(1).$$

Abbreviating

$$b_{ij}^{(k)} := \tilde{p}_{ij} e^{-k\mu t_j}, \quad a_i^{(k)} := \sum_{j=1}^{d} b_{ij}^{(k)} \sum_{m=1}^{k} \binom{k}{m} \left(\frac{\lambda_j}{\mu} \left(e^{\mu t_j} - 1 \right) \right)^m \gamma_j^{(k-m)}(1),$$

the value of the *factorial moment*

$$M_i^{(k)} := \gamma_i^{(k)}(1) = \mathbb{E}\left[M_i(M_i - 1) \cdots (M_i - k + 1) \right] = \mathbb{E}\left[\frac{M_i!}{(M_i - k)!} \right]$$

can be computed through $M_i^{(k)} = a_i^{(k)} + \sum_{j=1}^{N} b_{ij}^{(k)} M_j^{(k)}$. This leads to a procedure that enables the computation of $\boldsymbol{M}^{(k)}$ recursively from $\boldsymbol{M}^{(1)}$ up to $\boldsymbol{M}^{(m-1)}$, based on the relation (in self-evident notation)

$$\boldsymbol{M}^{(k)} = \left(I_d - B^{(k)} \right)^{-1} \boldsymbol{a}^{(k)}. \tag{2}$$

Using Stirling's numbers of the second kind, denoted as $\mathcal{S}(n, k)$, the raw moments can be found from factorial moments:

$$\mathbb{E}\left[M_i^n \right] = \sum_{k=0}^{n} \mathcal{S}(n, k) \mathbb{E}\left[\frac{M_i!}{(M_i - k)!} \right], \text{ with } \mathcal{S}(n, k) := \frac{1}{k!} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^n.$$

Remark 3. The service rate μ can easily be made state-dependent, by writing μ_j instead of μ, so that $h_j(t) = 1 - e^{-\mu_j t}(1 - z)$ and $g_j(t) = \exp(-\lambda_i/\mu_i(1 - \exp(-\mu_i t))(1 - z))$.

Below, the results for deterministic transition times will be used to analyse the case where the background process is a Markov process, meaning that the transition times are exponentially distributed. The following classical result, featuring the notion of characteristic function (CF), is needed, see Ref.[12]; '$\xrightarrow{\mathrm{d}}$' means convergence in distribution.

Proposition 4 (Lévy's Convergence Theorem)**.** *Consider a sequence of random variables* $X_1, X_2, \ldots$, *with* CF*s* $\phi_1(s), \phi_2(s), \ldots$, *so that* $\phi_n(s) = \mathbb{E}[e^{\mathrm{i}sX_n}]$. *If*

$$\lim_{n \to \infty} \phi_n(s) = \phi(s)$$

for some function $\phi(s)$ *for all* $s \in \mathbb{R}$, *and furthermore* $\phi(s)$ *is continuous at* $s = 0$, *then*

$$\lim_{n \to \infty} X_n \xrightarrow{\mathrm{d}} X,$$

where X *has* CF $\phi(s)$.

Observe that the exponential distribution can be approximated by a geometric number of 'short' deterministic times, where the success probability of this geometric distribution is 'small.' This idea is formalized in the following well-known lemma.

Lemma 5. *Let G_t have a geometric distribution with success probability $(1 - pt)$, that is, $\mathbb{P}[G_t = i] = (1 - pt)^{i-1}pt$. Then $tG_t \xrightarrow{\mathrm{d}} H$ as $t \downarrow 0$, where H has an exponential distribution with mean $1/p$.*

This way the Markov process (with d states) can be discretized; we use the transition probabilities

$$p_{ij} = \begin{cases} r_{ij}t & \text{for } i \neq j \\ 1 - \sum_{j \neq i} r_{ij}t & \text{for } i = j. \end{cases}$$

Here r_{ij} is the transition rate from state i to j of the Markov process; $t < (\max_i \sum_{j \neq i} r_{ij})^{-1}$. When the intervals between the transitions (which are possibly self-transitions) are of length t (deterministically) and taking $t \downarrow 0$, the resulting discrete-time Markov chain matches with the original Markov process, according to Lemma 5.

However, it has to be noted that the random variables M_i (with $i = 1, \ldots, N$) denote the population at epochs that a state is entered, but entering happens increasingly often when $t \downarrow 0$. Since self-transitions are allowed (and occur each time with probability close to 1), the corresponding discrete process re-enters this state continuously during an exponential staying time in a state. Therefore, the variable M_i denotes the stationary distribution of the population at arbitrary moments in which the system is in state i (rather than the stationary distribution at the epoch the Markov process enters i).

Bearing in mind Eq. (2), we now consider subsequently $I_d - B^{(k)}$ and $\boldsymbol{a}^{(k)}$ at $t \downarrow 0$.

First the entries of the matrix $I_d - B^{(k)}$ will be analysed at $t \downarrow 0$. First consider $i \neq j$. Using the definition, it is immediately seen that

$$\left(I_d - B^{(k)}\right)_{ij} = \tilde{p}_{ij}e^{-k\mu t} = -r_{ji}\frac{\pi_j}{\pi_i}t(1 - k\mu t + O(t^2)) = -r_{ji}\frac{\pi_j}{\pi_i}t + O(t^2).$$

For $\left(I_d - B^{(k)}\right)_{ii}$ something similar can be done:

$$\left(I_d - B^{(k)}\right)_{ii} = \tilde{p}_{ii}e^{-k\mu t} = 1 - \left(1 - \sum_{j \neq i} r_{ji}t\right)(1 - k\mu t + O(t^2))$$

$$= \left(\sum_{j \neq i} r_{ji} + k\mu \right) t + O(t^2).$$

The analysis of $\boldsymbol{a}^{(k)}$ requires a few more calculations:

$$\begin{aligned}
a_i^{(k)} &= \sum_{j=1}^{d} b_{ij}^{(k)} \sum_{m=1}^{k} \binom{k}{m} \left(\frac{\lambda_j}{\mu} \left(e^{\mu t} - 1 \right) \right)^m \gamma_j^{(k-m)}(1) \\
&= \sum_{j=1}^{d} b_{ij}^{(k)} \sum_{m=1}^{k} \binom{k}{m} \lambda_j^m \left(t + O(t^2) \right)^m \gamma_j^{(k-m)}(1) \\
&= \sum_{j=1}^{d} b_{ij}^{(k)} \sum_{m=1}^{k} \left(\binom{k}{m} (\lambda_j t)^m \gamma_j^{(k-m)}(1) + O\left(t^{m+1}\right) \right) \\
&= \sum_{j \neq i} \left(r_{ji} \frac{\pi_j}{\pi_i} t + O(t^2) \right) \sum_{m=1}^{k} \left(\binom{k}{m} (\lambda_j t)^m \gamma_j^{(k-m)}(1) + O\left(t^{m+1}\right) \right) \\
&\quad + \left(1 - \left(\sum_{j \neq i} r_{ji} + k\mu \right) t + O(t^2) \right) \\
&\quad \times \sum_{m=1}^{k} \left(\binom{k}{m} (\lambda_i t)^m \gamma_i^{(k-m)}(1) + O\left(t^{m+1}\right) \right) \\
&= O\left(t^2\right) + (1 + O(t)) \left(k\lambda_i t \gamma_i^{(k-1)}(1) + O\left(t^2\right) \right) \\
&= \left(k\lambda_i \gamma_i^{(k-1)}(1) \right) t + O\left(t^2\right).
\end{aligned}$$

It is concluded that both $I_d - B^{(k)}$ and $\boldsymbol{a}^{(k)}$ have a linear term in t in all coefficients and no constant term, as $t \downarrow 0$. Since the matrix is inverted in (2), they cancel each other out, and the terms with $O(t^2)$ vanish as $t \downarrow 0$. It follows that

$$\boldsymbol{M}^{(k)} = k\, C_k^{-1} \Lambda \boldsymbol{M}^{(k-1)},$$

with $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$, and $C_k = \left(c_{ij}(k) \right)_{i,j=1}^{N}$, in which

$$c_{ij}(k) := \begin{cases} -r_{ji} \dfrac{\pi_j}{\pi_i} & \text{if } i \neq j, \\ \displaystyle\sum_{h \neq i} r_{hi} + k\mu & \text{else.} \end{cases}$$

This leads to the explicit expression

$$\boldsymbol{M}^{(k)} = k!\left[C_k^{-1}\Lambda C_{k-1}^{-1}\Lambda\cdots C_2^{-1}\Lambda C_1^{-1}\Lambda\right](1,\ldots,1)^{\mathrm{T}}. \tag{3}$$

Note that this result is in line with Theorem 3.1 in Ref.[9].

Example 6. Consider the 2-state system with $p_{11} = 1-\kappa_1 t$, $p_{12} = \kappa_1 t$, $p_{21} = \kappa_2 t$, and $p_{22} = 1-\kappa_2 t$, where $t < (\max\{\kappa_1,\kappa_2\})^{-1}$. It is readily verified that this results in $\pi_i = \kappa_{3-i}/(\kappa_1+\kappa_2)$ for $i = 1,2$. Take $\mathbf{t} = (t,t)$ and $\mathbf{l} = (\lambda_1,\lambda_2)$.

Now it turns out that $\tilde{p}_{ij} = p_{ij}$ (the matrix P corresponds to a reversible discrete-time Markov chain). From (2) it can be found that

$$\mathbb{E}_t[M_i] = \frac{\lambda_i}{\mu}\cdot\frac{1-e^{-\mu t}-\kappa_i t+(\kappa_1+\kappa_2-2\kappa_1\kappa_2 t)te^{-\mu t}}{1-e^{-\mu t}+(\kappa_1+\kappa_2)te^{-\mu t}} + \frac{\lambda_{3-i}}{\mu}\cdot\frac{\kappa_i t(1-2e^{-\mu t}+2\kappa_{3-i}te^{-\mu t})}{1-e^{-\mu t}+(\kappa_1+\kappa_2)te^{-\mu t}}.$$

Now taking the limit $t\downarrow 0$, the Markov chain becomes equivalent to the Markov process with rates $r_{12} = \kappa_1$ and $r_{21} = \kappa_2$. As we explained above, M_i does not denote the population at the epoch of entering state i in this case; instead it is the population found at a random moment while being in state i. The mean is found to equal

$$\mathbb{E}[M_i] = \lim_{t\downarrow 0}\mathbb{E}_t[M_i] = \frac{\lambda_i}{\mu}\cdot\frac{\mu+\kappa_{3-i}}{\mu+\kappa_1+\kappa_2} + \frac{\lambda_{3-i}}{\mu}\cdot\frac{\kappa_i}{\mu+\kappa_1+\kappa_2}.$$

The average population over all time is now

$$\begin{aligned}\mathbb{E}[M] &= \pi_1\mathbb{E}[M_1]+\pi_2\mathbb{E}[M_2]\\ &= \frac{\lambda_1}{\mu}\cdot\frac{\kappa_2}{\kappa_1+\kappa_2} + \frac{\lambda_2}{\mu}\cdot\frac{\kappa_1}{\kappa_1+\kappa_2} = \frac{\lambda_1}{\mu}\pi_1+\frac{\lambda_2}{\mu}\pi_2.\end{aligned} \tag{4}$$

The latter result can be obtained more easily from Little's law, which says that $\mathbb{E}[M]$ equals the product of the mean arrival rate and the expected time spent in the system; the latter quantity is obviously $1/\mu$.

Using the Stirling numbers introduced in Section 3, the following result is found for exponentially distributed transition times:

$$\left(\mathbb{E}[M_1^n],\ldots,\mathbb{E}[M_d^n]\right)^{\mathrm{T}} = \sum_{k=0}^{n}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n \times\left[C_k^{-1}\Lambda C_{k-1}^{-1}\Lambda\cdots C_2^{-1}\Lambda C_1^{-1}\Lambda\right](1,\ldots,1)^{\mathrm{T}}.$$

4. GENERALLY DISTRIBUTED TRANSITION TIMES: LIMITING REGIMES

In this section we study two well-known limiting regimes. In the first one, all transition times are divided by n, and then n is sent to 0. As can be expected, in this *quasi-stationary regime* the number of customers present is a mixture of Poisson random variables: time-scaling makes sure that when the modulating Markov process is in state i, the process locally behaves as an M/M/∞ system with arrival rate λ_i.

In the second regime (the so-called *fluid-scaling regime*) time is sped up by a factor n, and the behaviour for $n \to \infty$ is considered. In this case, it turns out that the limiting arrival process is a Poisson process with rate, say, λ_∞. Remarkably, this property holds for transition times T_i $(i = 1, \ldots, d)$ with *arbitrary* distributions, in the sense that λ_∞ depends on the transition times only through $(\mathbb{E}[T_1], \ldots, \mathbb{E}[T_d])$, see Corollary 9.

In Section 5 we show that the limiting regimes already yield reasonable approximations for n relatively close to 1.

4.1. Quasi-Stationary Behavior

First we consider the situation that the transition times are slowed down, that is, divided by a factor n, where n is then sent to 0.

Theorem 7. *As $n \to 0$, in the infinite-server system with transition times t_i/n, the random variable M_i has a 'mixed Poisson distribution', i.e., a Poisson distribution with parameter λ_j/μ with probability $\tilde{p}_{ij}$ for $j = 1, \ldots, d$.*

Proof. Note that taking transition times t_i/n with $n \to 0$, it formally becomes problematic to speak about the system in steady state. Therefore, instead of slowing down the transition process, we speed up the arrival and departure processes, $\boldsymbol{\lambda} \mapsto \boldsymbol{\lambda}/n$ and $\mu \mapsto \mu/n$, obviously resulting in exactly the same distribution of M_i. Using Taylor expansions, the equivalents of $g_j(z)$ and $h_j(z)$ obey

$$\begin{aligned} g_j^{(n)}(z) &= \exp\left(-\frac{\lambda_j}{\mu}\left(1 - e^{-\mu t_j/n}\right)(1-z)\right) \\ &= e^{-\frac{\lambda_j}{\mu}(1-z)}\left(1 + O\left(e^{-\mu t_j/n}\right)\right); \\ h_j^{(n)}(z) &= 1 - e^{-\mu t_j/n}(1-z). \end{aligned}$$

This leads for the corresponding PGF, say $\gamma^{(n)}(z)$, to

$$\gamma_i^{(n)}(z) = \sum_{j=1}^{d} \tilde{p}_{ij} e^{-\frac{\lambda_j}{\mu}(1-z)}\left(1 + O\left(\left(e^{-\mu t_j}\right)^{\frac{1}{n}}\right)\right)\gamma_j^{(n)}\left(1 - e^{-\mu t_j/n}(1-z)\right)$$

so that $\gamma_i^{(n)}(z)$ converges, as $n \to 0$, to

$$\lim_{n\to 0} \sum_{j=1}^{d} \tilde{p}_{ij} e^{-\frac{\lambda_j}{\mu}(1-z)} \left(1 + O\left(\left(e^{-\mu t_j}\right)^{\frac{1}{n}}\right)\right) \gamma_j^{(n)} \left(1 - e^{-\mu t_j/n}(1-z)\right) = \sum_{j=1}^{d} \tilde{p}_{ij} e^{-\frac{\lambda_j}{\mu}(1-z)},$$

which concludes the proof. □

For the distribution of the number of customers at an arbitrary transition epoch, say M, we obtain

$$\mathbb{E}\left[z^M\right] = \sum_{i=1}^{d} \pi_i \mathbb{E}\left[z^{M_i}\right] = \sum_{i=1}^{d} \pi_i \sum_{j=1}^{d} \tilde{p}_{ij} e^{-\frac{\lambda_j}{\mu}(1-z)} = \sum_{j=1}^{d} \pi_j e^{-\frac{\lambda_j}{\mu}(1-z)}.$$

In other words: M has a Poisson distribution with parameter λ_j/μ with probability π_j, as expected. This property will carry over to the case that the T_is have an arbitrary distribution on $\mathbb{R}_+$.
In Section 5 we will use the symbol M^0 for this limiting variable M, while M^∞ corresponds to the same quantity under the fluid scaling, to be discussed in the next subsection.

4.2. Fluid-Scaling Behavior

In case all transition times are divided by a factor n, while $\boldsymbol{\lambda}$ and μ remain unchanged, we show in this subsection that in the limit ($n \to \infty$) the inter-arrival times become exponential. This is first shown for the case of deterministic transition times (as was discussed in Section 3), and then we argue that the result carries over to general (finite-mean) transition times.

Theorem 8. *Consider the system with deterministic transition times t_i/n. As $n \to \infty$, the time until the first arrival converges in distribution to an exponential random variable M^∞ with mean $1/\lambda_\infty$, where*

$$\lambda_\infty := \frac{\sum_{i=1}^{d} \pi_i t_i \lambda_i}{\sum_{i=1}^{d} \pi_i t_i}. \tag{5}$$

Proof. Let $\phi_i(s)$ denote the CF of X_i $(i = 1, \ldots, d)$, where X_i is the time until the next arrival when entering state i, in the limiting situation of $n \to \infty$ (assumed to exist); $\boldsymbol{\phi}(s)$ is the d-dimensional vector with entries $\phi_i(s)$ $(i = 1, \ldots, d)$. The counterparts of these notions in the pre-limit situation are $\mathbf{X}^{(n)}$ and $\boldsymbol{\phi}^{(n)}(s)$. Our objective is to prove that the X_i have an exponential distribution with mean $1/\lambda_\infty$, irrespective of i. As the proof is

rather lengthy, we have tried to make it more transparent by breaking it up in a number of steps.

Step I. We first derive a fixed point equation for the CF of $\mathbf{X}^{(n)}$. Standard arguments yield that

$$\begin{aligned}
\phi_i^{(n)}(s) &= s\mathbb{E}\left[e^{\mathrm{i}sX_i^{(n)}}\right] \\
&= \mathbb{P}\left[X_i^{(n)} < t_i/n\right]\mathbb{E}\left[e^{\mathrm{i}}sX_i^{(n)} \mid X_i^{(n)} < t_i/n\right] \\
&\quad + \mathbb{P}\left[X_i^{(n)} > t_i/n\right]\mathbb{E}\left[e^{\mathrm{i}}sX_i^{(n)} \mid X_i^{(n)} > t_i/n\right] \\
&= \left(1 - e^{-\lambda_i t_i/n}\right)\frac{\lambda_i}{\lambda_i - \mathrm{i}s} \cdot \frac{1 - e^{-(\lambda_i - \mathrm{i}s)t_i/n}}{1 - e^{-\lambda_i t_i/n}} \\
&\quad + e^{-\lambda_i t_i/n}e^{\mathrm{i}}st_i/n \sum_{j=1}^{d} p_{ij}\mathbb{E}\left[e^{\mathrm{i}}sX_j^{(n)}\right] \\
&= \left(1 - e^{-(\lambda_i - \mathrm{i}s)t_i/n}\right)\frac{\lambda_i}{\lambda_i - \mathrm{i}s} + e^{-(\lambda_i - \mathrm{i}s)t_i/n}\sum_{j=1}^{d} p_{ij}\phi_j^{(n)}(s). \qquad (6)
\end{aligned}$$

In vector notation, this is written as

$$\boldsymbol{\phi}^{(n)}(s) = (I - D)(\Lambda - \mathrm{i}sI)^{-1}\boldsymbol{\lambda} + DP\boldsymbol{\phi}^{(n)}(s),$$

where $D := \mathrm{diag}\left(e^{-(\lambda_1 - \mathrm{i}s)t_1/n}, \ldots, e^{-(\lambda_d - \mathrm{i}s)t_d/n}\right)$, which depends on both n and s, $\Lambda := \mathrm{diag}(\boldsymbol{\lambda})$. Provided $\det(I - DP) \neq 0$, this yields:

$$\boldsymbol{\phi}^{(n)}(s) = (I - DP)^{-1}(I - D)(\Lambda - \mathrm{i}sI)^{-1}\boldsymbol{\lambda}. \qquad (7)$$

Step II. We now 'Taylorize' expression (7). First define $T = \mathrm{diag}(\mathbf{t})$. Using standard Taylor expansions, the elements of D can be rewritten as $d_{ii} = 1 - (\lambda_i - \mathrm{i}s)t_i/n + O\left(n^{-2}\right)$, so that

$$D = I - \frac{1}{n}T(\Lambda - \mathrm{i}sI) + \frac{1}{n^2}R_1,$$

for some matrix R_1 with $R_1/n \to 0$ as $n \to \infty$. With R_2 being a matrix with the same property as R_1, expression (7) is rewritten as

$$\boldsymbol{\phi}^{(n)}(s) = \left(I - P + \frac{1}{n}T(\Lambda - \mathrm{i}sI)P - \frac{1}{n^2}R_1P\right)^{-1}\left(\frac{1}{n}T + \frac{1}{n^2}R_2\right)\boldsymbol{\lambda}.$$

Note that R_1 and R_2 are the only matrices that depend on s.

The inverse of matrix $I - P + T(\Lambda - \mathrm{i}sI)P/n + R_1P/n^2 = A + B/n$ is to be determined now, with $A := I - P$ and $B = T(\Lambda - \mathrm{i}sI)P + O(1/n)$. This equals (under the assumption $\det(I - DP) \neq 0$)

$$\mathrm{inv}\left(A + \frac{1}{n}B\right) = \frac{1}{\det(A + \frac{1}{n}B)}\mathrm{adj}\left(A + \frac{1}{n}B\right),$$

which is a direct result from Cramer's rule[11]. Since P is a probability matrix, we have that $\det(A) = \det(I - P) = 0$. Hence, using the common permutations description of the determinant,

$$\begin{aligned}
\det\left(A + \frac{1}{n}B\right) &= \det\left(A + \frac{1}{n}B\right) - \det(A) \\
&= \sum_{\sigma\in S_d}\mathrm{sgn}(\sigma)\left(\prod_{i=1}^{d}\left(a_{i,\sigma(i)} + \frac{1}{n}b_{i,\sigma(i)}\right) - \prod_{i=1}^{d}a_{i,\sigma(i)}\right) \\
&= \sum_{\sigma\in S_d}\mathrm{sgn}(\sigma)\left(\prod_{i=1}^{d}a_{i,\sigma(i)} + \frac{1}{n}\sum_{i=1}^{d}b_{i,\sigma(i)}\left(\prod_{j=1,j\neq i}^{d}a_{j,\sigma(j)}\right) + O\left(\frac{1}{n^2}\right) - \prod_{i=1}^{d}a_{i,\sigma(i)}\right) \\
&= \frac{1}{n}\sum_{\sigma\in S_d}\mathrm{sgn}(\sigma)\sum_{i=1}^{d}b_{i,\sigma(i)}\left(\prod_{j=1,j\neq i}^{d}a_{j,\sigma(j)}\right) + O\left(\frac{1}{n^2}\right),
\end{aligned}$$

with S_d denoting all permutations. As $b_{i,\sigma(i)} = p_{i,\sigma(i)}t_i(\lambda_i - \mathrm{i}s) + O(1/n)$, we obtain

$$\begin{aligned}
\det(A + \tfrac{1}{n}B) &= \frac{1}{n}\sum_{\sigma\in S_N}\mathrm{sgn}(\sigma)\sum_{i=1}^{d}p_{i,\sigma(i)}t_i(\lambda_i - \mathrm{i}s)\left(\prod_{j=1,j\neq i}^{d}a_{j,\sigma(j)}\right) + O\left(\frac{1}{n^2}\right) \\
&= \frac{1}{n}(q - \mathrm{i}rs) + O\left(\frac{1}{n^2}\right),
\end{aligned}$$

for a positive q and r, as we will show in the next step.

Step III. We now show that both q and r are positive. Observe that with c_i, for $i = 1, \ldots, d$, defined suitably,

$$\begin{aligned}
q &= \sum_{\sigma\in S_d}\mathrm{sgn}(\sigma)\sum_{i=1}^{d}p_{i,\sigma(i)}\left(\prod_{j=1,j\neq i}^{d}a_{j,\sigma(j)}\right)t_i\lambda_i \\
&= \sum_{i=1}^{d}\sum_{\sigma\in S_d}\mathrm{sgn}(\sigma)p_{i,\sigma(i)}\left(\prod_{j=1,j\neq i}^{d}a_{j,\sigma(j)}\right)t_i\lambda_i = \sum_{i=1}^{d}c_it_i\lambda_i,
\end{aligned}$$

and likewise, $r = \sum_{i=1}^{d} c_i t_i$ with the same coefficients c_i. Here $\lambda_i \geq 0$ for all i with strict inequality for at least one i and $t_i > 0$ for all i. Showing $c_i > 0$ for all $i = 1, 2, \ldots, d$ is therefore sufficient to prove both $q > 0$ and $r > 0$. Without loss of generality, we focus on $i = 1$.

Since

$$c_1 = \sum_{\sigma \in S_d} \operatorname{sgn}(\sigma) p_{1,\sigma(1)} \left(\prod_{j=2}^{d} a_{j,\sigma(j)} \right),$$

it is the determinant of a matrix of which the bottom $d-1$ rows equal those of $A = I - P$, while the upper row equals $(p_{11}, \ldots, p_{1d})$. Standard algebraic manipulations yield

$$\begin{aligned} c_1 &= -\sum_{\sigma \in S_d} \operatorname{sgn}(\sigma) \left(-p_{1,\sigma(1)}\right) \left(\prod_{j=2}^{d} a_{j,\sigma(j)} \right) \\ &= \sum_{\sigma \in S_d} \operatorname{sgn}(\sigma) \delta_{1,\sigma(1)} \left(\prod_{j=2}^{d} a_{j,\sigma(j)} \right) - \sum_{\sigma \in S_d} \operatorname{sgn}(\sigma) \left(\prod_{j=1}^{d} a_{j,\sigma(j)} \right) \\ &= \det(A^{(1,1)}) - \det(A) = \det(A^{(1,1)}); \end{aligned} \tag{8}$$

here $A^{(i,j)}$ is the $(d-1) \times (d-1)$ submatrix of A with the ith row and jth column omitted, of which the determinant is called the (i,j)th minor. Since $A = I - P$ with P a transition probability matrix for an irreducible finite-state discrete-time Markov chain, $A^{(1,1)}$ is strictly diagonally dominant. Indeed,

$$\begin{aligned} \left| \left(A^{(1,1)}\right)_{ii} \right| = 1 - p_{i+1,i+1} &= \sum_{j=1, j \neq i+1}^{d} p_{i+1,j} \\ &\geq \sum_{j=2, j \neq i+1}^{d} \left| -p_{i+1,j} \right| = \sum_{j=1, j \neq i}^{d-1} \left| \left(A^{(1,1)}\right)_{ij} \right|, \end{aligned}$$

with for at least some i the weak inequality sign being a strict inequality, since $p_{i,1} > 0$ for some $i = 2, \ldots, d$. Strict diagonal dominance implies that the determinant is non-zero. The fact that all diagonal elements are positive entails all eigenvalues to be positive as well, which results in $\det\left(A^{(1,1)}\right) > 0$, since the determinant is the product of the eigenvalues. Therefore $c_i > 0$ for all $i = 1, \ldots, d$, and thus also $q > 0$ and $r > 0$. Note that both constants q and r are independent of n and s. As expected, $\det(A + B/n) \to 0$ as $n \to \infty$.

Step IV. Next, $M^{[n]} := adj(A + B/n)$ is determined as the transpose of the matrix of cofactors of $A + B/n$. Since the only arithmetic operations used to calculate the cofactor are addition, subtraction, and multiplication, one concludes that if A has some non-zero cofactor, then $\lim_{n\to\infty} M^{[n]}$ equals $M := adj(A)$. Inspecting the diagonal immediately shows that this claim holds, because the $(1,1)$th cofactor equals $c_1 > 0$ as was found in (8), and likewise the (i,i)th cofactor equals $c_i > 0$ for $i = 2, \ldots, d$. This entails that M is independent of s, and $\lim_{n\to\infty} M^{[n]} = M$. The inverse is now calculated as

$$\mathrm{inv}(A + \tfrac{1}{n}B) = \frac{n}{q - \mathrm{i}rs + O\left(\frac{1}{n}\right)} M^{[n]},$$

so that

$$\begin{aligned}\boldsymbol{\phi}^{(n)}(s) &= \frac{n}{q - \mathrm{i}rs + O\left(\frac{1}{n}\right)} M^{[n]} \left(\frac{1}{n}T + \frac{1}{n^2}R_2\right)\boldsymbol{\lambda} \\ &= \frac{1}{q - \mathrm{i}rs + O\left(\frac{1}{n}\right)} M^{[n]} \left(T + \frac{1}{n}R_2\right)\boldsymbol{\lambda}.\end{aligned}$$

Now the limit $n \to \infty$ can finally be taken:

$$\begin{aligned}\boldsymbol{\phi}(s) &= \lim_{n\to\infty} \frac{1}{q - \mathrm{i}rs + O\left(\frac{1}{n}\right)} M^{[n]} \left(T + \frac{1}{n}R_2\right)\boldsymbol{\lambda} \\ &= \frac{1}{q - \mathrm{i}rs} MT\boldsymbol{\lambda} = \frac{1}{q - \mathrm{i}rs}\mathbf{m},\end{aligned}$$

for some d-dimensional vector $\mathbf{m}$. From $\phi_i(0) = 1$ follows $\mathbf{m} = q\mathbf{1}$, where $\mathbf{1}$ is the all-one column vector of appropriate dimension, and thus $\phi_i^{(n)}(s)$ converges to the characteristic function of an exponential distribution with parameter q/r, for every $i = 1, 2, \ldots, d$. Since $\boldsymbol{\phi}(s)$ is continuous at $s = 0$, Proposition 4 yields the desired convergence in distribution of $X_i^{(n)}$ to X_i, with $X_i \sim \mathrm{Exp}(q/r)$ for all $i = 1, \ldots, d$.

Step V. It is left to show that λ_∞ is the only candidate for q/r. This can be done by only looking at the first moment. For a Poisson process $Y(t)$ with rate λ, the number of arrivals obeys

$$\lim_{t\to\infty} \frac{Y(t)}{t} = \lambda,$$

almost surely. On the long run the process spends $\pi_i t_i / \sum \pi_j t_j$ part of the time in state i (independent of n), so the contribution to the number of

arrivals done in state i, named $Y_i(t)$ will be

$$\lim_{t\to\infty} \frac{Y_i(t)}{t} = \frac{\pi_i t_i \lambda_i}{\sum_{j=1}^{d} \pi_j t_j},$$

almost surely. This implies

$$\lambda_\infty = \lim_{t\to\infty} \frac{Y(t)}{t} = \lim_{t\to\infty} \sum_{i=1}^{d} \frac{Y_i(t)}{t} = \sum_{i=1}^{d} \frac{\pi_i t_i \lambda_i}{\sum_{j=1}^{d} \pi_j t_j},$$

which is indeed (5). □

This theorem extends to systems with arbitrary transition times, where the t_i in (5) should be replaced by $\mathbb{E}[T_i]$, the expected value of the random transition time T_i, as shown in the following corollary.

Corollary 9. *Consider the system with arbitrary transition times; assume* $\mathbb{E}[T_i] < \infty$. *As* $n \to \infty$, *the time until the first arrival converges in distribution to an exponential random variable with mean* $1/\lambda_\infty$, *where*

$$\lambda_\infty = \frac{\sum_{i=1}^{d} \pi_i \mathbb{E}[T_i]\, \lambda_i}{\sum_{i=1}^{d} \pi_i \mathbb{E}[T_i]}. \tag{9}$$

Proof. Writing the time until the next arrival when switching to state i as $X_i^{(n)}$ for finite n, the corresponding CF can be expressed as

$$\mathbb{E}\left[e^{\mathrm{i}sX_i^{(n)}}\right] = I_1 + I_2 \sum_{j=1}^{d} p_{ij} \mathbb{E}\left[e^{\mathrm{i}sX_j^{(n)}}\right].$$

with

$$I_1 := \int_0^\infty \int_0^t f_{T_i/n}(u) \lambda_i e^{-\lambda_i t} e^{\mathrm{i}su}\, \mathrm{d}u\, \mathrm{d}t,$$
$$I_2 := \int_0^\infty \int_t^\infty f_{T_i/n}(u) \lambda_i e^{-\lambda_i t} e^{\mathrm{i}st}\, \mathrm{d}u\, \mathrm{d}t;$$

here $f_X(\cdot)$ denotes the density of some random variable X. Now these two integrals I_1 and I_2 are evaluated separately. The first one reduces to

$$\begin{aligned} I_1 &= \int_0^\infty \int_0^t f_{T_i/n}(u) \lambda_i e^{-\lambda_i t} e^{\mathrm{i}su}\, \mathrm{d}u\, \mathrm{d}t \\ &= \int_0^\infty e^{\mathrm{i}su} f_{T_i/n}(u) \int_u^\infty \lambda_i e^{-\lambda_i t}\, \mathrm{d}t\, \mathrm{d}u \\ &= \int_0^\infty e^{-(\lambda_i - \mathrm{i}s)u} f_{T_i/n}(u)\, \mathrm{d}u = \mathscr{M}_{T_i/n}(\lambda_i - \mathrm{i}s), \end{aligned}$$

whereas the second can be evaluated as

$$\begin{aligned} I_2 &= \int_0^\infty \int_t^\infty f_{T_i/n}(u)\lambda_i e^{-\lambda_i t} e^{\mathrm{i}st}\,\mathrm{d}u\,\mathrm{d}t \\ &= \int_0^\infty f_{T_i/n}(u) \int_u^\infty \lambda_i e^{-(\lambda_i - \mathrm{i}s)t}\,\mathrm{d}t\,\mathrm{d}u \\ &= \int_0^\infty f_{T_i/n}(u) \frac{\lambda_i}{\lambda_i - \mathrm{i}s}\left(1 - e^{-(\lambda_i - \mathrm{i}s)u}\right) du \\ &= \frac{\lambda_i}{\lambda_i - \mathrm{i}s}\left(1 - \mathscr{M}_{T_i/n}(\lambda_i - \mathrm{i}s)\right), \end{aligned}$$

where $\mathscr{M}_{T_i/n}(\cdot)$ denotes the CF of T_i/n. Since

$$\mathscr{M}_{T_i/n}(\lambda_i - \mathrm{i}s) = 1 - \frac{1}{n}\mathbb{E}[T_i](\lambda_i - \mathrm{i}s) + O\left(\frac{1}{n^2}\right),$$

it follows that the calculations in the proof of Theorem 8 are identical for the case of generally distributed transition times, since the Taylor expansion of matrix D is the same after t_i is replaced by $\mathbb{E}[T_i]$. The rest of the argument is identical. □

Remark 10. The technique that we used to prove Theorem 8 can be adapted to show that not only the time till the first arrival is exponential (with mean $1/\lambda_\infty$), but also that the limiting arrival process is a Poisson process with rate λ_∞. We here sketch how this can be done; we focus on deterministic transition times but this can be generalized to arbitrary (finite-mean) distributions in the way described above. To this end, let $Z_i^{(n)}(\tau_1, \tau_2)$ denote the number of arrivals in the time interval $[\tau_1, \tau_2)$ (with $\tau_1 < \tau_2$), given the background process just entered state i at time τ_1. Define

$$\psi_i^{(n)}(s; \tau_1, \tau_2) := \mathbb{E}\left[e^{\mathrm{i}sZ_i^{(n)}(\tau_1,\tau_2)}\right],$$

and $\psi_i(s; \tau_1, \tau_2)$ the corresponding limiting value as $n \to \infty$. For $\tau_2 - \tau_1 > t_i/n$,

$$\psi_i^{(n)}(s; \tau_1, \tau_2) = \exp\left(-\lambda_i \frac{t_i}{n}(1 - e^{\mathrm{i}s})\right) \sum_{j=1}^{d} p_{ij}\psi_j^{(n)}(s; \tau_1 + t_i/n, \tau_2).$$

This system of equations can be regarded as the counterpart of (6). Write

$$\psi_j^{(n)}(s; \tau_1 + t_i/n, \tau_2) = \psi_j^{(n)}(s; \tau_1, \tau_2) + \frac{t_i}{n}(\psi_j^{(n)})'(s; \tau_1, \tau_2) + O\left(\frac{1}{n^2}\right),$$

where the differentiation is with respect to τ_1. Where in the proof of Theorem 8 the vector $\boldsymbol{\phi}^{(n)}(s)$ could be found by solving a system of linear equations, we now have to solve a system of linear differential equations to identify $\boldsymbol{\psi}^{(n)}(s;\tau_1,\tau_2)$. From that point on, we can follow precisely the same steps as in the proof of Theorem 8. It eventually follows that $(n \to \infty)$

$$\psi_i^{(n)}(s;\tau_1,\tau_2) \to \exp\bigl(-\lambda_\infty(\tau_2-\tau_1)(1-e^{\mathrm{i}s})\bigr),$$

proving that the limiting distribution of $Z_i^{(n)}(\tau_1,\tau_2)$ is Poisson with mean $\lambda_\infty(\tau_2-\tau_1)$. Likewise, the bivariate CF of the number of arrivals in $[\tau_1,\tau_2)$ and the number of arrivals in $[\tau_3,\tau_4)$ (with $\tau_1<\tau_2<\tau_3<\tau_4$) converges to

$$\exp\bigl(-\lambda_\infty(\tau_2-\tau_1)(1-e^{\mathrm{i}s_1})\bigr)\exp\bigl(-\lambda_\infty(\tau_4-\tau_3)(1-e^{\mathrm{i}s_2})\bigr),$$

irrespective of the state of the background process at time τ_1. We have then shown that the number of arrivals in non-overlapping intervals follow independent Poisson distributions, from which it follows that the limiting arrival process is a Poisson process. (As an aside we mention that that last argument was also used by Khinchine[7] to prove that the superposition of n renewal processes, after a time-scaling by a factor n, converges to a Poisson process.) We eventually find the following result.

Theorem 11. *Consider the system with arbitrary transition times; assume* $\mathbb{E}[T_i]<\infty$. *As* $n\to\infty$, *the arrival process converges to a Poisson process with rate* λ_∞.

Remark 12. One of the referees suggested an alternative proof for the property that, in the fluid scaling regime, $Z_i^{(n)}(0,t)$ converges for $n\to\infty$ to a Poisson random variable with mean λ_∞, irrespective of i. In this alternative proof the CF of $Z_i^{(n)}(0,t)$ is first written in terms of an integral over the arrival rate, that is $\int_0^t \lambda_{X_n(s)}\,ds$, with $X_n(\cdot)$ the background process in the n-scaled model. Relying on the renewal reward theorem and elementary properties of Markov chains, it is shown that the CF of interest converges to the postulated one.

5. TIME SCALING SIMULATIONS

In the previous section results have been found for limiting regimes, in which time was either sped up or slowed down. In this section we numerically study how fast these regimes are reached.

We have chosen various simulation settings which all gave similar results. One representative example will be shown in detail below.

The simulation setting was chosen to be a 3-STATE Markov chain with transition matrix

$$P = \begin{pmatrix} 1/5 & 2/5 & 2/5 \\ 0 & 1/5 & 4/5 \\ 1 & 0 & 0 \end{pmatrix},$$

which results in the steady-state distribution $\boldsymbol{\pi} = \frac{1}{23}(10, 5, 8)$ and thus

$$\tilde{P} = \begin{pmatrix} 1/5 & 0 & 4/5 \\ 4/5 & 1/5 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

Now take $\boldsymbol{\lambda} = (1, 3, 8)$, $\mu = 1$, and let the random variables $\mathbf{T}$ be generally distributed with means $\mathbb{E}[T] = (2, 7, 1)$. When dividing the transition times by a factor n, where $n \to 0$, and given $\mathbb{P}[T_i > 0] = 1$ for all i, the distributions turn out to be

$$\begin{aligned} M_1^0 &\sim (A_1 \mathbb{P}\text{ois}(1) + (1 - A_1)\mathbb{P}\text{ois}(8)), \\ M_2^0 &\sim (A_2 \mathbb{P}\text{ois}(1) + (1 - A_2)\mathbb{P}\text{ois}(3)), \\ M_3^0 &\sim (A_3 \mathbb{P}\text{ois}(1) + (1 - A_3)\mathbb{P}\text{ois}(3)), \end{aligned}$$

where $(A_1, A_2, A_3) \sim (\mathbb{B}\text{er}(1/5), \mathbb{B}\text{er}(4/5), \mathbb{B}\text{er}(1/2))$, where $\mathbb{B}\text{er}(p)$ denotes a Bernouilli random variable with success probability p, i.e., $\mathbb{P}[\mathbb{B}\text{er}(p) = 1] = 1 - \mathbb{P}[\mathbb{B}\text{er}(p) = 0] = p$. For speeding up the transition times the outcome distribution is $\mathbb{P}\text{ois}(\lambda_\infty)$ with rate $\lambda_\infty = 3$, using the notation of (5). Denote these limiting random variables as M_i^∞, with $i = 1, 2, 3$.

With n chosen to be the acceleration of the process, meaning $\mathbf{T} \mapsto \mathbf{T}/n$, the simulation is first run for deterministic transition times and $n = 10^{-4}, 0.5, 1, 2, 10^4$, so that two cases with only small changes in transition times are covered, as well as both limiting cases. The outcome distributions are shown in Figure 1. As expected the steady-state distributions for $n = 10^{-4}$ and $n = 10^4$ closely follow the limiting distributions, which are shown in the figures as dotted lines. It is clearly visible that small accelerations and slow-downs already lead to close approximations of the limiting processes.

This is further examined in Figures 2 through 4, where the Kullback–Leibler divergences with respect to the limiting distributions are shown for both exponentially and uniformly distributed transition times and for deterministic transition times, all having the same mean. The Kullback–Leibler divergence between the random variables X and Y is given by

$$\mathrm{KL}(X, Y) = \sum_n \mathbb{P}[X = n] \log\left(\frac{\mathbb{P}[X = n]}{\mathbb{P}[Y = n]}\right).$$

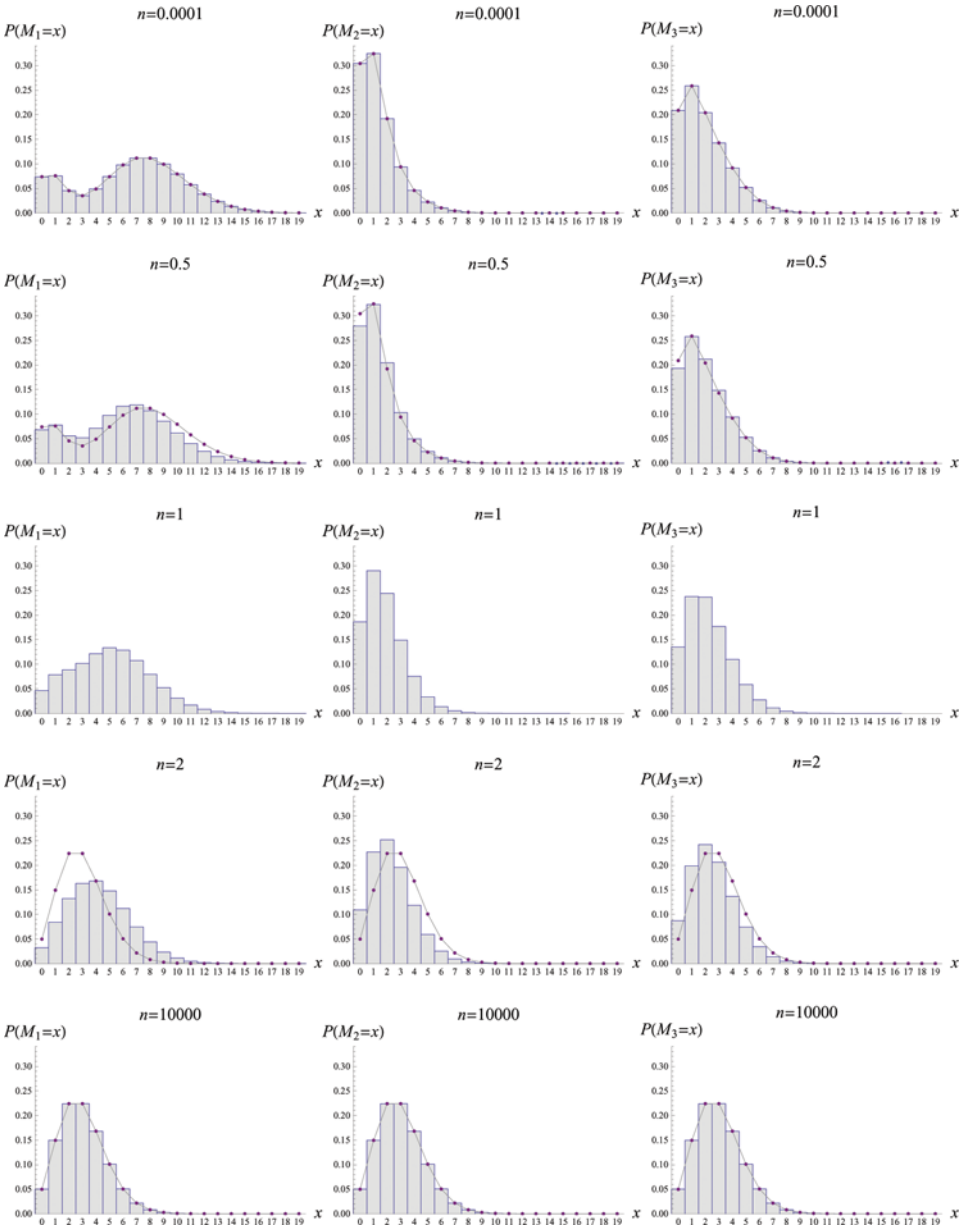


FIGURE 1 Steady-state distributions of M_i, $i = 1, 2, 3$, with deterministic transition times t_i/n for different values of n. The dotted lines depict the expected limiting distributions, for either the quasi-stationary or the fluid-scaling regime (color figure available online).

Informally, the smaller the KL-divergence is, the closer the distributions are to each other. In Figures 2 and 4 the $\mathbb{U}\text{nif}(0, 2\mathbb{E}[T_i])$ distributions, with mean $\mathbb{E}[T_i]$ and 0 as the infinum of the support, are chosen to represent

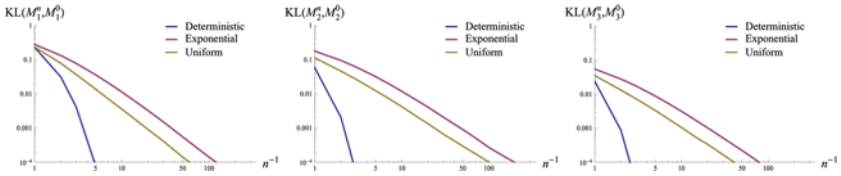


FIGURE 2 Kullback–Leibler divergences of M_i^n and M_i^0, $i = 1, 2, 3$, for different values of $n \leq 1$. For the uniform distribution $\mathbb{U}\text{nif}(0, 2\mathbb{E}[T_i])$ has been chosen (color figure available online).

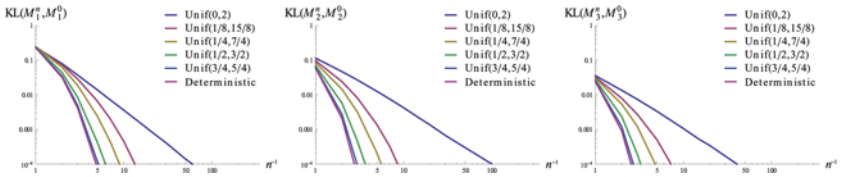


FIGURE 3 Kullback–Leibler divergences of M_i^n and M_i^0, $i = 1, 2, 3$, for different values of $n \leq 1$ and various values of α in $\mathbb{E}[T_i] \cdot \mathbb{U}\text{nif}(\alpha, 2 - \alpha)$ (color figure available online).

the uniform case, while in Figure 3 multiple uniform distributions with mean $\mathbb{E}[T_i]$ are evaluated.

The largest difference between the transition time distributions is seen in Figure 2, which shows the quasi-stationary regime. The transition time distribution has a significant influence on the convergence speed. The exponential distribution appears to be the slowest, followed by the uniform distribution, and the deterministic one has the fastest convergence rate. This is also the descending order of variance for these three distributions. This can be explained intuitively since the higher variance also implies the higher probability of the transition time being close to zero. For the quasi-stationary case the length of the previous intervals is important. Say the order of visiting states has been $k \to j \to i$ at a certain instance, then for the quasi-stationary case $M_i \sim \mathbb{P}\text{ois}(\lambda_j/\mu)$. However, if the probability of having spent a very short time in j is significant, M_i is still too strongly influenced by the time spent in state k.

This effect is further analysed in Figure 3, where multiple uniform distributions have been simulated, namely $\mathbb{E}[T_i] \cdot \mathbb{U}\text{nif}(\alpha, 2 - \alpha)$ for $\alpha = 0, 1/8, 1/4, 1/2, 3/4, 1$, where $\mathbb{U}\text{nif}(1, 1)$ reduces evidently to the

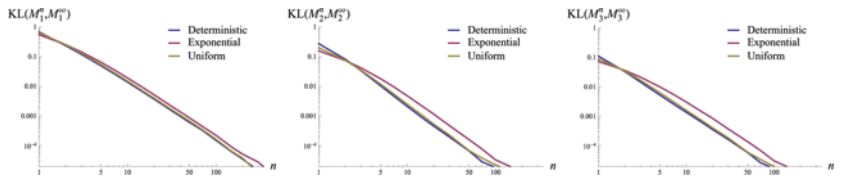


FIGURE 4 Kullback–Leibler divergences of M_i^n and M_i^∞, $i = 1, 2, 3$, for different values of $n \geq 1$. For the uniform distribution $\mathbb{U}\text{nif}(0, 2\mathbb{E}[T_i])$ has been chosen (color figure available online).

deterministic case. The biggest difference is between for $\alpha = 0$ and $\alpha = 1/8$, since the infinum of the support is suddenly greater than 0.

Figure 4 shows the Kullback–Leibler divergence between $M_i^\infty \sim \mathbb{P}\text{ois}(3)$ and the steady-state distributions of M_i for various values of $n \geq 1$, i.e., the fluid-scaling regime. It is seen that the transition time distribution has little influence on the convergence speed, compared to the quasi-stationary regime. It seems that queue lengths for deterministic transition times converge slightly faster than uniformly distributed transition times, with exponential just behind these two. This is the same order as found for quasi-stationary scaling. It is seen that among states $1, 2, 3$ the convergence rate is comparable but the off-set varies.

6. CONDITIONS FOR M_i HAVING A POISSON DISTRIBUTION

In the standard M/M/∞-queue, the stationary number of customers present has a Poisson distribution. This corresponds to our case with $d = 1$. Also, in Section 4 it has been shown that M_i admits to a Poisson distribution in several limiting cases. Both when transition times tend to infinity and when transition times tend to zero, the steady-state distribution of M_i follows a (combination of) Poisson distribution(s). This raises the following question: under what conditions does the steady-state population (exactly) obey a Poisson distribution? In this section explicit conditions are identified.

In Ref.[1] it has been shown that the number M_i follows a Poisson distribution, but with a possibly random parameter. We briefly sketch the distribution of this parameter; for details we refer to Ref.[1]. Denoting the random arrival rate by $\Gamma(t) := \lambda_{X(t)}$, for $t \in \mathbb{R}$, as the mapping from time to rate λ, with as before $X(t)$ the state of the Markov chain at time t, we have

$$M_i \sim \mathbb{P}\text{ois}\left(\left|A_\Gamma\right|_\Gamma\right), \tag{10}$$

where

$$\left|A_\Gamma\right|_\Gamma := \sum_{h<0} \mathbb{P}\left[\text{Exp}(\mu) > -u_h \,|\, \Gamma \text{ with a transition to } i \text{ happening at } t = 0\right].$$

Here u_h are all the (random) arrival epochs that happened before time 0, so that $\left|A_\Gamma\right|_\Gamma$ equals the sum of the probabilities for every individual customer to still be present at time 0. It is concluded that this parameter depends on Γ, and is therefore random for all distributions of Γ, except the deterministic one. In that case $\Gamma(t)$ corresponds to a deterministic cycle through the states, and knowledge of a realization Γ would not contribute anything.

In other words: if the rate λ is non-random at all times, then the random variables M_i, with $i = 1, 2, \ldots, N$, have a Poisson distribution. Recalling that we have taken the service rate μ to be state-independent, we conclude

that the information about the arrival rate is effectively the only factor of importance of random environment $\Gamma(t)$, $t \in \mathbb{R}$. The process in[10] is of this type with two states and $p_{12} = p_{21} = 1$, and thus results in a Poisson process.

However, this does not imply that we only have a Poisson distribution in case of cyclic routing through the d states. For example, consider the case with three states so that $\lambda_2 = \lambda_3$ and $t_2 = t_3$, and $p_{12} = p_{13} = 1/2$, and $p_{21} = p_{31} = 1$. The Markov chain will alternate between state 1 (having λ_1 for a time t_1), and either state 2 or 3 (having $\lambda_2 = \lambda_3$ for a time $t_2 = t_3$), so that the function $\Gamma(t)$ will still be deterministic, given that the Markov-chain enters some state i at time 0. Note, however, that this expansion of the state space results essentially in the same sample path behaviour.

This idea is formalized in the following theorem, where we recall the model description from Section 2. In this theorem a partition of the states of the background process is demanded such that (1) every element of the partition has the same arrival rate λ, (2) the routing through the elements of this partition is cyclic, and (3) the time the Markov chain spends each of these elements is fixed.

Theorem 13. *Consider the Markov chain comprising d states with transition probabilities p_{ij}. Denote $P_i^+ = \{j \mid p_{ij} > 0\}$ and $P_i^- = \{j \mid p_{ji} > 0\}$ for all $i = 1, 2, \ldots, d$. Then every M_i has a Poisson steady-state distribution, if and only if a partition $N_1, \ldots, N_k$ of the d states can be made so that*

(1) *$\lambda_i = \lambda_j$ for all $i, j \in N_n$ with $n = 1, \ldots, k$. Call these rates $\Lambda_1, \ldots, \Lambda_k$.*
(2) *For all $i \in N_n$, $P_i^+ / N_n \subset N_{n+1}$, with $n = 1, \ldots, k$.*
(3) *For every sequence $i_1, i_2, \ldots, i_m \in N_n$ so that $i_{j+1} \in P_{i_j}^+$ for all $j = 1, 2, \ldots, m-1$ and $P_{i_1}^- \cap N_{n-1} \neq \emptyset$ and $P_{i_m}^+ \cap N_{n+1} \neq \emptyset$, $\sum_{j=1}^{m} t_{i_j} = T_n$ for some constant T_n.*

Note that indices N_n are to be understood modulo k.

Proof. First it will be shown that the three criteria result in a deterministic $\Gamma(t)$ for all t. After that, the reverse will be proven.

- For $k = 1$, by virtue of the first point the same Λ_1 holds for every vertex, so that $\Gamma(t) = \Lambda_1$ for all t and is consequently deterministic.
  Now let $k \geq 2$. Recall that we assumed the Markov chain to be irreducible and positive recurrent. Then, since $k \geq 2$, no class of states can be absorbing. By the second point the order of going through the classes is determined, namely from n to $n+1$, not skipping any class. This also means $\Gamma(t)$ has a cyclic pattern of $\Lambda_1, \ldots, \Lambda_k$.
  Every sequence $i_1, i_2, \ldots, i_m \in N_n$ as described in the third point of the theorem is a possible way to travel through the states of N_n. If N_n is entered at some time T_0, then it must be left at time $T_0 + T_n$. This is due

to the third point, which tells us that the total time spent in a class of states is always fixed. Therefore $\Gamma(t)$ is always deterministic.

- Now assume $\Gamma(t)$ is deterministic. In case $\Gamma(t) = \Lambda_1$ is constant for all t, take $k = 1$ and group all states into one class. Then the first point is satisfied and the second and third are trivial.
  In the other case $\Gamma(t)$ does not always hold the same value. Without loss of generality, assume that it jumps to Λ_1 at time 0. Since $\Gamma(t)$ depends on the state of a finite-state Markov chain, it must go through a fixed cyclic pattern with period T, by virtue of the Markov property. This means $\Gamma(t) = \Lambda_1$ for all $0 < t < T_1$ for some T_1, then $\Gamma(t) = \Lambda_2$ for all $T_1 < t < T_1 + T_2$ for some T_2, and so on until $\Gamma(t) = \Lambda_k$ for all $\sum_{n=1}^{k-1} T_n < t < \sum_{n=1}^{k} T_n =: T$, after which the cycle is repeated. Regardless of the structure of the Markov chain, it starts at time 0 at some state with rate Λ_1 and after time T_1 it switches to a state with rate Λ_2. Group together all states it can be in at times t with $0 < t \pmod T < T_1$ as N_1, the ones for $0 < (t - T_1) \pmod T < T_2$ as N_2, and so on. The first point is satisfied. The third point is then satisfied since every path through the N_n takes time T_n. The second point is also satisfied since from N_n one must travel to another state within N_n (maintaining rate Λ_n) or switch to a state in N_{n+1}, which will have rate Λ_{n+1}. □

ACKNOWLEDGMENT

The authors thank the anonymous referees for their careful assessment of our manuscript.

REFERENCES

1. D'Auria, B. M/M/∞ queues in semi-Markovian random environment. *Queueing Syst.* **2008**, *58*, 221–237.
2. Fralix, B.H.; Adan, I.J.B.F. An infinite-server queue influenced by a semi-Markovian environment. Queueing Syst. **2009**, *61*, 65–84.
3. Glynn, P.W. Large deviations for the infinite server queue in heavy traffic. *Inst. Math. Appl.* **1995**, *71*, 387–394.
4. Glynn, P.W.; Whitt, W. A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv. Appl. Probab.* **1991**, *23*, 188–209.
5. Keilson, J.; Servi, L.D. The matrix M/M/∞ system: Retrial models and Markov modulated sources. *Adv. Appl. Probab.* **1993**, *25*, 453–471.
6. Kelly, F.P. *Reversibility and Stochastic Networks*; John Wiley and Sons Ltd.: New York, 1979.
7. Khinchine, A. *Mathematical Models in the Theory of Queueing*; Griffin: London, UK, 1960.
8. Neuts, M.F. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*; Johns Hopkins University Press, 1981.
9. O'Cinneide, C.A.; Purdue, P. The M/M/∞ queue in a random environment. Journal of Applied Probability **1986**, *23* (1), 175–184.
10. Schwabe, A.; Dobrzynski, M.; Rybakova, K.N.; Verschure, P.J.; Bruggeman, F.J. Origins of stochastic intracellular processes and consequences for cell-to-cell variability and cellular survival strategies. Methods in Enzymology **2011**, *500*, 597–625.
11. Strang, G. *Linear Algebra and Its Applications*, 3rd Ed.; Thomson Learning, Inc.: San Diego, CA, 1988.
12. Williams, D. *Probability with Martingales*; Cambridge University Press: Cambridge, UK, 1991.