# Analysis of Markov-modulated infinite-server queues in the central-limit regime

Joke Blom, Koen de Turck, Michel Mandjes

▶ **To cite this version:**

Joke Blom, Koen de Turck, Michel Mandjes. Analysis of Markov-modulated infinite-server queues in the central-limit regime. Probability in the Engineering and Informational Sciences, 2015, 10.1017/S026996481500008X . hal-01327062

**HAL Id: hal-01327062**
**https://hal.science/hal-01327062**

Submitted on 6 Jun 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# ANALYSIS OF MARKOV-MODULATED INFINITE-SERVER QUEUES IN THE CENTRAL-LIMIT REGIME

JOKE BLOM [⋆], KOEN DE TURCK [†], MICHEL MANDJES [•,⋆]

Abstract. This paper focuses on an infinite-server queue modulated by an independently evolving finite-state Markovian background process, with transition rate matrix $Q \equiv (q_{ij})_{i,j=1}^{d}$. Both arrival rates and service rates are depending on the state of the background process. The main contribution concerns the derivation of central limit theorems for the number of customers in the system at time $t \geq 0$, in the asymptotic regime in which the arrival rates $\lambda_i$ are scaled by a factor $N$, and the transition rates $q_{ij}$ by a factor $N^\alpha$, with $\alpha \in \mathbb{R}^+$. The specific value of $\alpha$ has a crucial impact on the result: (i) for $\alpha > 1$ the system essentially behaves as an M/M/$\infty$ queue, and in the central limit theorem the centered process has to be normalized by $\sqrt{N}$; (ii) for $\alpha < 1$, the centered process has to be normalized by $N^{1-\alpha/2}$, with the deviation matrix appearing in the expression for the variance.

Keywords. Infinite-server queues ⋆ Markov modulation ⋆ central limit theorem ⋆ deviation matrices

Work done while K. de Turck was visiting Korteweg-de Vries Institute for Mathematics, University of Amsterdam, the Netherlands, with greatly appreciated financial support from *Fonds Wetenschappelijk Onderzoek / Research Foundation – Flanders*. He is also a Postdoctoral Fellow of the same foundation.

• Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands.

⋆ CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands.

† TELIN, Ghent University, St.-Pietersnieuwstraat 41, B9000 Gent, Belgium.

M. Mandjes is also with Eurandom, Eindhoven University of Technology, Eindhoven, the Netherlands, and IBIS, Faculty of Economics and Business, University of Amsterdam, Amsterdam, the Netherlands.

joke.blom@cwi.nl, kdeturck@telin.ugent.be, M.R.H.Mandjes@uva.nl

*Date*: September 18, 2014.

## 1. Introduction

The infinite-server queue has been intensively studied, perhaps owing to its wide applicability and attractive computational features. In these systems jobs arrive according to a given arrival process, go into service immediately, are served in parallel, and leave when their service is completed. An important feature of this model is that there is *no waiting*: jobs do not interfere with each other. The infinite-server queue was originally developed to analyze the probabilistic properties of the number of calls in progress in a trunk in a communication network, as an approximation of the corresponding system with many servers. More recently, however, various other application domains have been identified, such as road traffic [17] and biology [15].

In the most standard variant of the infinite-server model, known as the M/M/$\infty$ model, jobs arrive according to a Poisson process with a fixed rate $\lambda$, where the service times are i.i.d. samples from an exponential distribution with mean $\mu^{-1}$ (independent of the job arrival process). A classical result states that the stationary number of jobs in the system has a Poisson distribution with mean $\lambda/\mu$. Also the transient behavior of this queueing system is well understood.

In many practical situations, however, the assumptions underlying the standard infinite-server model are not valid. The arrivals often tend to be 'clustered' (so that the assumption of a fixed arrival rate does not apply), while also the service distribution may vary over time. This explains the interest in *Markov-modulated* infinite-server queues, so as to incorporate 'burstiness' into the queue's input process. In such queues, the input process is modulated by a finite-state (of dimension $d \in \mathbb{N}$) irreducible continuous-time Markov process $(J(t))_{t\in\mathbb{R}}$, often referred to as the *background process* or *modulating process*, with transition rate matrix $Q \equiv (q_{ij})_{i,j=1}^{d}$. If $J(t)$ is in state $i$, the arrival process is (locally) a Poisson process with rate $\lambda_i$ and the service times are exponential with mean $\mu_i^{-1}$ (while the obvious independence assumptions are assumed to be fulfilled).

The Markov-modulated infinite-server queue has attracted some attention over the past decades (but the number of papers on this type of system is relatively modest, compared to the vast literature on Markov-modulated single-server queues). The main focus in the literature so far has been on characterizing the steady-state number of jobs in the system; see e.g. [6, 8, 9, 12, 14] and references therein. Interestingly, there are hardly any explicit results on the probability distribution of the (transient or stationary) number of jobs present: the results are in terms of recursive schemes to determine all moments, and implicit characterizations of the probability generating function.

An idea to obtain more explicit results for the distribution of the number of jobs in the system, is by applying specific *time-scalings*. In [2, 10] a time-scaling is studied in which the transitions of the background process occur at a faster rate than the Poisson arrivals. As a consequence, the limiting input process becomes essentially Poisson (with an arrival rate being an average of the

$\lambda_i$ s); a similar property applies for the service times. Under this scaling, one gets in the limit the Poisson distribution for the stationary number of jobs present. Recently, related transient results have been obtained as well, under specific scalings of the arrival rates and transition times of the background process [2, 4].

*Contribution.* Our work considers a time-scaling featuring in [2, 4] as well. In this scaling, the arrival rates $\lambda_i$ are inflated by a factor $N$, while the background process $(J(t))_{t\in\mathbb{R}}$ is sped up by a factor $N^\alpha$, for some $\alpha \in (0,\infty)$. The primary focus is on the regime in which $N$ grows large.
The object of study is the number of jobs in the scaled system at time $t$, in the sequel denoted by $M^{(N)}(t)$. More specifically, we aim at deriving a central limit theorem (CLT) for $M^{(N)}(t)$, as well as for its stationary counterpart $M^{(N)}$. Interestingly, we find different scaling regimes, based on the value of $\alpha$. The rationale behind these different regimes lies in the fact that for $\alpha > 1$ the variances of $M^{(N)}(t)$ and $M^{(N)}$ grow essentially linearly in $N$, while for $\alpha < 1$ they grow as $N^{2-\alpha}$.
It is important to notice that there are actually two variants of this Markov-modulated infinite-server queue. In the first (to be referred to as 'Model I') the service times of jobs present at time $t$ are subject to a hazard rate that is determined by the state $J(t)$ of the background process at time $t$. In the second variant (referred to as 'Model II') the service times are determined by the state of the modulating process at the job's arrival epoch (and hence can be sampled upon arrival).
The main contribution of our work is that we develop a unified approach to prove the CLTs for both Model I and Model II for the scalings given above, for arbitrary $\alpha \in (0,\infty)$, and for both the transient and stationary regimes. The technique used can be summarized as follows. We first derive differential equations for the probability generating functions (pgfs) of the transient number of jobs in the system $M^{(N)}(t)$ as well as its stationary counterpart $M^{(N)}$ (for both models). The next step is to establish laws of large numbers: we identify $\varrho(t)$ ($\varrho$, respectively) to which $N^{-1}M^{(N)}(t)$ ($N^{-1}M^{(N)}$, respectively) converges as $N \to \infty$. This result indicates how $M^{(N)}(t)$ and $M^{(N)}$ should be centered in a CLT. The thus obtained centered random variables are then normalized (that is, divided by $N^\gamma$, for an appropriately chosen $\gamma$), so as to obtain the CLT. As suggested by the asymptotic behavior of the variance of $M^{(N)}(t)$ and $M^{(N)}$, as we pointed out above, the appropriate choice of the parameter $\gamma$ in the normalization is $\gamma = \frac{1}{2}$ for $\alpha > 1$, and $\gamma = 1 - \frac{\alpha}{2}$ for $\alpha < 1$. The proofs rely on (non-trivial) manipulations of the differential equations that underly the pgfs. For $\alpha < 1$ the *deviation matrix* [7] appears in the CLT in the expression for the variance.

*Relation to previous work.* In our preliminary conference paper [3] we just covered Model I, with an approach similar to the one featuring in the present paper. In [2] the transient regime of Model II is analyzed, but just for $\alpha > 1$, relying on a different and more elaborate methodology. New results

of this paper are: (i) Model II for $\alpha \leq 1$, (ii) the CLT for the stationary number of jobs $M^{(N)}$ in Model II, (iii) results on the correlation across time. The main contribution, however, concerns the unified approach: where earlier work has been using ad hoc solutions for the scenario at hand, we now have a general 'recipe' to derive CLT s of this kind. Current work in progress aims at functional versions of the CLT s for the *process* $(M^{(N)}(t))_{t\in\mathbb{R}}$; [1] covers the special case of uniform service rates, which constitutes the intersection of Model I and Model II.

*Organization.* The organization of this paper is as follows. In Section 2, we explain the model in detail and introduce the notations used throughout the paper. Section 3 provides a systematic explanation of our technique for proving this kind of CLT s; in addition, we demonstrate the approach for a special case, viz. the transient analysis for the model with uniform service rates (in which Model I and Model II coincide). In Section 4 we recall the results for Model I as derived in the precursor paper [3]. Then in Section 5, we state and prove for Model II the CLT s, both for the stationary and transient distribution. The single-dimensional convergence can be extended to convergence of the finite-dimensional distributions (viz. at different points in time); see Section 6. In Section 7, we provide some numerical examples so as to get insight into the speed of convergence to the various limiting regimes. The final section of the paper, Section 8, contains a discussion and concluding remarks.

## 2. Model description and preliminaries

In this section, we first provide a detailed model description. We then give a number of explicit calculations for the mean and variance of $M^{(N)}(t)$, that indicate how this random variable should be centered and normalized so as to obtain a CLT. We conclude by presenting a number of preliminary results (e.g., a number of standard results on deviation matrices).

### 2.1. Model description, scaling.

The main objective of this paper is to study an infinite-server queue with Markov-modulated Poisson arrivals and exponential service times. In full detail, the model is described as follows.

*Model.* Consider an irreducible continuous-time Markov process $(J(t))_{t\in\mathbb{R}}$ on a finite state space $\{1,\ldots,d\}$, with $d \in \mathbb{N}$. Let its transition rate matrix be given by $Q \equiv (q_{ij})_{i,j=1}^{d}$; here the rates $q_{ij}$ are nonnegative if $i \neq j$, whereas $q_{ii} = -\sum_{j\neq i} q_{ij}$ (so that the row sums are 0). Let $\pi_i$ be the stationary probability that the background process is in state $i$, for $i = 1,\ldots,d$; due to the irreducibility assumption there is a unique stationary distribution. The time spent in state $i$ (often referred to as the *transition time*) has an exponential distribution with mean $1/q_i$, where $q_i := -q_{ii}$. Let $M(t)$ denote the number of jobs in the system at time $t$, and $M$ its steady-state counterpart. The dynamics of the process $(M(t))_{t\in\mathbb{R}}$ can be described as follows. While the process $(J(t))_{t\in\mathbb{R}}$,

usually referred to as the *background process* or *modulating process*, is in state $i \in \{1, \ldots, d\}$, jobs arrive at the queue according to a Poisson process with rate $\lambda_i \geq 0$. The service times are assumed to be exponentially distributed with rate $\mu_i$, however, more importantly this statement can be interpreted in two ways:

Model I: In the first variant of our model, the service times of all jobs present at a certain time instant $t$ are subject to a hazard rate determined by the state $J(t)$ of background chain at time $t$, regardless of when they arrived. Informally, if the system is in state $i$, then the probability of an arbitrary job leaving the system in the next $\Delta t$ time units is $\mu_i \, \Delta t$.

Model II: In the second variant the service rate is determined by the background state as seen by the job upon its arrival. If the background process was in state $i$, the service time is sampled from an exponential distribution with mean $\mu_i^{-1}$.

The difference between the two models is nicely illustrated by the following alternative representation [6]. In Model I $M(t)$ has a Poisson distribution with random parameter $\psi(J)$, while in Model II it is Poisson with random parameter $\varphi(J)$, where $J \equiv (J(s))_{s\in[0,t]}$, and

$$\psi(f) := \int_0^t \lambda_{f(s)} e^{-\int_s^t \mu_{f(r)} \mathrm{d}r} \mathrm{d}s, \qquad \varphi(f) := \int_0^t \lambda_{f(s)} e^{-\mu_{f(s)}\,(t-s)} \mathrm{d}s, \tag{1}$$

with $f : [0, t] \mapsto \{1, \ldots, d\}$.

*Scaling.* In this paper, we consider a scaling in which both (i) the arrival process, and (ii) the background process are sped up, at a possibly distinct rate. More specifically, the arrival rates are scaled linearly, that is, as $\lambda_i \mapsto N\lambda_i$, whereas the background chain is scaled as $q_{ij} \mapsto N^\alpha q_{ij}$, for some positive $\alpha$. We call the resulting process $(M^{(N)}(t))_{t\in\mathbb{R}}$, to stress the dependence on the scaling parameter $N$; the corresponding background process is denoted by $(J^{(N)}(t))_{t\in\mathbb{R}}$.

The main objective of this paper is the derivation of CLTs for the number of jobs in the system, as $N$ grows large. As mentioned in the introduction, the parameter $\alpha$ plays an important role here: it turns out to matter whether $\alpha$ is assumed smaller than, equal to, or larger than 1. Letting the system start off empty at time 0, we consider the number of jobs present at time $t$, denoted by $M^{(N)}(t)$; we write $M^{(N)}$ for its stationary counterpart.

Our main result is a 'non-standard CLT': for a deterministic function $\varrho(t)$,

$$\frac{M^{(N)}(t) - N\varrho(t)}{N^\gamma} \tag{2}$$

converges in distribution to a zero-mean Normal distribution with a certain variance, say, $\sigma^2(t)$. It is important to note that in the case $\alpha > 1$ we have that the parameter $\gamma$ equals the usual $\frac{1}{2}$, while for $\alpha \leq 1$ it has the uncommon value $1 - \frac{\alpha}{2}$. A similar dichotomy holds for the stationary

counterpart $M^{(N)}$. In the next subsection, we present explicit calculations for the mean and variance of $M^{(N)}(t)$ and $M^{(N)}$ that explain the reason behind this dichotomy.

2.2. **Explicit calculations for the mean and variance.** We now present a number of explicit calculations for the mean and variance of the number of jobs present; for ease we consider the case that $\mu_i = \mu$ for all $i \in \{1, \ldots, d\}$, so that Models I and II coincide. We assume $J(0)$ is distributed according to the stationary distribution of the Markov chain $J(t)$. Directly from, e.g., [2], for any $N \in \mathbb{N}$,

$$\frac{\mathbb{E}M^{(N)}(t)}{N} = \varrho(t) := \frac{1-e^{-\mu t}}{\mu}\sum_{i=1}^{d}\pi_i\lambda_i, \qquad \frac{\mathbb{E}M^{(N)}}{N} = \varrho := \frac{1}{\mu}\sum_{i=1}^{d}\pi_i\lambda_i.$$

We now concentrate on the corresponding variance; we first consider the non-scaled system, to later explore the effect of the time-scaling. In the sequel we use the notation $p_{ij}(t) := \mathbb{P}(J(t) = j \,|\, J(0) = i)$. The 'law of total variance', with $J \equiv (J(s))_{s=0}^{t}$, entails that

$$\mathbb{V}\mathrm{ar}\, M(t) = \mathbb{E}\,\mathbb{V}\mathrm{ar}(M(t) \,|\, J) + \mathbb{V}\mathrm{ar}\,\mathbb{E}(M(t) \,|\, J). \tag{3}$$

We first recall from (1) that $M(t)$ obeys a Poisson distribution with the *random* parameter $\varphi(J)$. As a result, the second term on the right of (3) can be written as

$$\mathbb{V}\mathrm{ar}\varphi(J) = \mathbb{V}\mathrm{ar}\left(\int_0^t \lambda_{J(s)}e^{-\mu\,(t-s)}\mathrm{d}s\right) = \int_0^t\int_0^t \mathbb{C}\mathrm{ov}\left(\lambda_{J(u)}e^{-\mu\,(t-u)}\lambda_{J(v)}e^{-\mu\,(t-v)}\right)\mathrm{d}u\mathrm{d}v,$$

which can be decomposed into $I_1 + I_2$, where

$$I_1 := \sum_{i=1}^{d}\sum_{j=1}^{d}\lambda_i\lambda_j K_{ij}, \text{ with } K_{ij} := \int_0^t\int_0^v e^{-\mu(t-u)}e^{-\mu(t-v)}\pi_i\left(p_{ij}(v-u)-\pi_j\right)\mathrm{d}u\mathrm{d}v,$$

$$I_2 := \sum_{i=1}^{d}\sum_{j=1}^{d}\lambda_i\lambda_j L_{ij}, \text{ with } L_{ij} := \int_0^t\int_v^t e^{-\mu(t-u)}e^{-\mu(t-v)}\pi_j\left(p_{ji}(u-v)-\pi_i\right)\mathrm{d}u\mathrm{d}v.$$

Let us first evaluate $K_{ij}$. To this end, substitute $w := v - u$ (i.e., replace $u$ by $v - w$), and then interchange the order of integration, so as to obtain

$$K_{ij} = e^{-\mu t}\pi_i\int_0^t\left(\int_w^t e^{2\mu v}\mathrm{d}v\right)e^{-\mu(t+w)}\left(p_{ij}(w)-\pi_j\right)\mathrm{d}w.$$

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

$$K_{ij} = \frac{1}{2\mu}e^{-\mu t}\pi_i\int_0^t\left(e^{\mu(t-w)}-e^{-\mu(t-w)}\right)\left(p_{ij}(w)-\pi_j\right)\mathrm{d}w.$$

The integral $L_{ij}$ can be evaluated similarly:

$$\begin{aligned} L_{ij} &= e^{-\mu t}\pi_j\int_0^t\left(\int_0^{t-w} e^{2\mu v}\mathrm{d}v\right)e^{-\mu(t-w)}\left(p_{ji}(w)-\pi_i\right)\mathrm{d}w \\ &= \frac{1}{2\mu}e^{-\mu t}\pi_j\int_0^t\left(e^{\mu(t-w)}-e^{-\mu(t-w)}\right)\left(p_{ji}(w)-\pi_i\right)\mathrm{d}w = K_{ji}. \end{aligned}$$

The first term in the right hand side of (3) is easily evaluated, again relying on the fact that $M(t)$ has a Poisson distribution, conditional on $J$:

$$\mathbb{E}\,\mathbb{V}\mathrm{ar}(M(t)\,|\,J) = \sum_{i=1}^{d} \pi_i \lambda_i \int_0^t e^{-\mu s} \mathrm{d}s = \frac{1-e^{-\mu t}}{\mu} \sum_{i=1}^{d} \pi_i \lambda_i = \varrho(t).$$

Now we study the effect of the time-scaling: we replace $\lambda_i$ by $N\lambda_i$ (for $i = 1, \ldots, d$) and $p_{ij}(w)$ by $p_{ij}(N^\alpha w)$ (for $i, j = 1, \ldots, d$). Introduce the *deviation matrix* $D$, by

$$[D]_{ij} := \int_0^\infty (p_{ij}(t) - \pi_j)\,\mathrm{d}t;$$

see e.g. [7]. Combining the above results, it is a matter of some elementary algebra to verify that, in obvious notation,

$$\mathbb{V}\mathrm{ar}\, M^{(N)}(t) \sim N\varrho(t) + N^{2-\alpha}\, \frac{1-e^{-2\mu t}}{\mu} \sum_{i=1}^{d}\sum_{j=1}^{d} \pi_i \lambda_i \lambda_j [D]_{ij}.$$

From this relation, the above mentioned dichotomy becomes clear. It is observed that for $\alpha > 1$ the variance of $M^{(N)}(t)$ grows linearly in $N$, and is essentially equal to the corresponding mean, viz. $N\varrho(t)$. The intuition here is that in this regime the background process jumps faster than the arrival process, so that the arrival stream is nearly Poisson with parameter $\sum_{i=1}^d \pi_i\lambda_i$. The resulting system behaves therefore, as $N \to \infty$, essentially as an M/M/$\infty$. If $\alpha < 1$ the background process is slower than the arrival process. The variance of $M^{(N)}(t)$ now grows like $N^{2-\alpha}$, proportionally to a constant that is a linear combination of the entries of the deviation matrix $D$.

The above computations were done for the transient number of jobs $M^{(N)}(t)$, but obviously an analogous reasoning applies to its stationary counterpart $M^{(N)}$.

2.3. **Preliminaries on deviation matrices, additional notation.** In this subsection, we recall a number of key properties of deviation matrices; for more detailed treatments we refer to e.g. the standard texts [11, 13, 16], as well as the compact survey [7]. We also introduce additional notation, which is intensively used later on.

We define the diagonal matrices $\Lambda$ and $\mathcal{M}$, where $[\Lambda]_{ii} = \lambda_i$ and $[\mathcal{M}]_{ii} = \mu_i$. We denote the invariant distribution corresponding to the transition matrix $Q$ by the vector $\boldsymbol{\pi}$; we follow the convention that vectors are column vectors unless stated otherwise, and that they are written in bold fonts. As $\boldsymbol{\pi}$ denotes the invariant distribution, we have $\boldsymbol{\pi}^{\mathrm{T}} Q = \mathbf{0}^{\mathrm{T}}$ and $\boldsymbol{\pi}^{\mathrm{T}}\mathbf{1} = 1$, where $\mathbf{0}$ and $\mathbf{1}$ denote vectors of zeros and ones, respectively. In the sequel we frequently use the 'time-average arrival rate' $\lambda_\infty := \sum_{i=1}^d \pi_i\lambda_i = \boldsymbol{\pi}^{\mathrm{T}}\Lambda\mathbf{1}$, and the 'time average departure rate' $\mu_\infty := \sum_{i=1}^d \pi_i\mu_i = \boldsymbol{\pi}^{\mathrm{T}}\mathcal{M}\mathbf{1}$.

We recall some concepts pertaining to the theory of deviation matrices of Markov processes; see e.g. [7]. In particular, we let $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$ denote the *ergodic matrix.* We also define the *fundamental matrix* $F := (\Pi - Q)^{-1}$. It turns out that the deviation matrix $D$, introduced above, satisfies

$D = F - \Pi$. We will frequently use the identities $QF = FQ = \Pi - I$, as well as the facts that $\Pi D = D\Pi = 0$ (here 0 is to be read as an all-zeros $d \times d$ matrix) and $F\mathbf{1} = \mathbf{1}$.

We use the following three vector-valued generating functions throughout the paper: $\boldsymbol{p}$ denotes the unscaled probability generating function (pgf); $\bar{\boldsymbol{p}} \equiv \bar{\boldsymbol{p}}^{(N)}$ denotes the corresponding moment generating function (mgf) under the law-of-large-numbers scaling; and $\tilde{\boldsymbol{p}} \equiv \tilde{\boldsymbol{p}}^{(N)}$ denotes the mgf centered and normalized appropriately for the central limit theorem at hand. For the transient cases, these generating functions involve an extra argument $t$ to incorporate time. Importantly, all three generating functions are *vectors* of dimension $d$ as we consider distributions jointly with the state of the background process; to make the notation easier, we assume that these vectors are *row* vectors. Lastly, $\phi \equiv \phi^{(N)}$ denotes the scalar mgf under the centering and normalization (obtained by summing the elements of $\tilde{\boldsymbol{p}}$).

## 3. Outline of CLT proofs

In this section we point out how we set up our CLT proofs. In the next two sections this 'recipe' is then applied to analyze Model I and Model II, covering both the transient and stationary number of jobs in the system. We use a fairly classical approach to proving the CLTs for centered and normalized sequences of random variables of the type (2). More specifically, our objective is to show that under the appropriate normalization (i.e., an appropriate choice of $\gamma$), the moment generating function of (2) converges to that of the Normal distribution; the same is done for the stationary counterpart of (2).

Our technique consists of the following steps.

(a) Derive a differential equation for the pgf $\boldsymbol{p}$ of the random quantities $M(t)$ and $M$.

(b) Establish the 'mean behavior' $\varrho(t)$ ($\varrho$, respectively) of $M^{(N)}(t)$ ($M^{(N)}$, respectively). This law of large numbers follows by manipulating the mgf $\bar{\boldsymbol{p}} \equiv \bar{\boldsymbol{p}}^{(N)}$, obtaining a scalar limit solution $\exp(\vartheta\varrho(t))$ in the transient case, and $\exp(\vartheta\varrho)$ in the stationary case.

(c) Reformulate the differential equation for the uncentered and unnormalized pgf $\boldsymbol{p}$ into a recurrence relation for the centered and normalized mgf $\tilde{\boldsymbol{p}} \equiv \tilde{\boldsymbol{p}}^{(N)}$.

(d) Manipulate and iterate this equation, approximate by suitable Taylor expansions, to obtain a differential equation for the scalar mgf $\phi$ under the chosen centering and normalization.

(e) Discard asymptotically vanishing terms, so as to obtain a unique limit solution, viz., $\phi(\vartheta) = \exp(\vartheta^2\sigma^2(t))$ in the transient case and $\phi(\vartheta) = \exp(\vartheta^2\sigma^2)$ in the stationary case. We explicitly identify $\sigma^2(t)$ and $\sigma^2$.

This limit solution resulting from the last step corresponds to a zero-mean Normal distribution. Due to Lévy's continuity theorem, this pointwise convergence of characteristic functions implies

convergence in distribution to the zero-mean Normal random variable, so that we have derived the CLT.

Issues related to the uniqueness of the solution of the differential equation are dealt with in Appendix A. Below we demonstrate this proof technique for the special case that the service rates in each of the states are identical, i.e., $\mathcal{M} = \mu I$ for some $\mu > 0$, so that Models I and II coincide. Importantly, Prop. 1 in Section 3.1 holds for general $\mathcal{M}$.

3.1. **Differential equations for the pgf $\boldsymbol{p}$.** First we derive a system of differential equations for the pgf of the number of jobs in the system, jointly with the background state. We consider the bivariate process $(M(t), J(t))_{t\in\mathbb{R}}$, which is an ergodic Markov process on the state space $\{1,\ldots,d\}\times\mathbb{N}$. With the states of this process enumerated in the obvious way, it has the (infinite-dimensional) transition rate matrix

$$\begin{pmatrix} Q-\Lambda & \Lambda & & & & \\ \mathcal{M} & Q-\mathcal{M}-\Lambda & \Lambda & & & \\ & 2\mathcal{M} & Q-2\mathcal{M}-\Lambda & \Lambda & & \\ & & 3\mathcal{M} & Q-3\mathcal{M}-\Lambda & \Lambda & \\ & & & \ddots & \ddots & \ddots \end{pmatrix}.$$

We set out to find the transient distribution $(\boldsymbol{p}_k(t))_{k=0}^{\infty}$, where $\boldsymbol{p}_k(t)$ is a $d$-dimensional row-vector whose entries are defined by $[\boldsymbol{p}_k(t)]_j := \mathbb{P}(M(t)=k, J(t)=j)$. The (row-vector-)pgf $\boldsymbol{p}(t,z)$ is then defined through

$$\boldsymbol{p}(t,z) := \sum_{k=0}^{\infty} \boldsymbol{p}_k(t) z^k,$$

such that

$$[\boldsymbol{p}(t,z)]_j = \mathbb{E}\left(z^{M(t)} 1_{\{J(t)=j\}}\right).$$

**Proposition 1.** *The pgf $\boldsymbol{p}(t,z)$ satisfies the following differential equation:*

$$\frac{\partial \boldsymbol{p}(t,z)}{\partial t} = \boldsymbol{p}(t,z)\, Q + (z-1)\left(\boldsymbol{p}(t,z)\,\Lambda - \frac{\partial \boldsymbol{p}(t,z)}{\partial z}\mathcal{M}\right).$$

*Proof.* The result follows from classical arguments. By virtue of the Chapman-Kolgomorov equation, we have that

$$\frac{\mathrm{d}\boldsymbol{p}_k(t)}{\mathrm{d}t} = \boldsymbol{p}_{k-1}(t)\Lambda + \boldsymbol{p}_k(t)(Q-\Lambda-k\mathcal{M}) + (k+1)\boldsymbol{p}_{k+1}(t)\mathcal{M}, \tag{4}$$

for all $k\in\mathbb{N}$, where we put $\boldsymbol{p}_{-1}(t) := 0$ for all $t \geq 0$.

From the standard relations

$$\sum_{k=0}^{\infty}(k+1)\boldsymbol{p}_{k+1}(t)z^k = \frac{\partial \boldsymbol{p}(t,z)}{\partial z}, \text{ and } \sum_{k=0}^{\infty} k\boldsymbol{p}_k(t)z^k = z\frac{\partial \boldsymbol{p}(t,z)}{\partial z},$$

we obtain by multiplying both sides of (4) by $z^k$ and summing over $k \in \mathbb{N}$,

$$\frac{\partial \boldsymbol{p}(t,z)}{\partial t} = z\boldsymbol{p}(t,z)\Lambda + \boldsymbol{p}(t,z)(Q-\Lambda) - z\frac{\partial \boldsymbol{p}(t,z)}{\partial z}\mathcal{M} + \frac{\partial \boldsymbol{p}(t,z)}{\partial z}\mathcal{M}.$$

The claim follows directly. □

We assume that at time 0 the system starts off empty. Under the scaling $\Lambda \mapsto N\Lambda$ and $Q \mapsto N^\alpha Q$, Prop. 1 implies that we have the following system of partial differential equations governing $(M^{(N)}(t), J^{(N)}(t))$:

$$\frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial t} = N^\alpha \boldsymbol{p}^{(N)}(t,z)\, Q + (z-1)\left(N\boldsymbol{p}^{(N)}(t,z)\Lambda - \frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial z}\mathcal{M}\right) \tag{5}$$

describing the pgf $\boldsymbol{p}^{(N)}$ of the number of jobs in the scaled system.

3.2. **Mean behavior.** In the remainder of this section, we assume that all service rates are identical: $\mathcal{M} = \mu I$. To obtain the limiting behavior of $N^{-1}\, M^{(N)}(t)$ when $N$ grows large, it turns out to be convenient to take the following steps.

(i) Rewrite the differential equation (5) as a recurrence relation for $\boldsymbol{p}^{(N)}$ involving the fundamental matrix $F$; recall from Section 2.3 the relation $QF = \Pi - I$.

(ii) Translate this into a recurrence relation in terms of the mgf $\bar{\boldsymbol{p}}^{(N)}$ of $N^{-1}\, M^{(N)}(t)$, using a Taylor expansion for $z = \exp(\vartheta/N)$.

(iii) Sum over the possible background states by postmultiplying with $\mathbf{1}$, so as to obtain a scalar mgf; in this step we make use of the identity $F\mathbf{1} = \mathbf{1}$.

(iv) Obtain the limiting differential equation by taking the limit for $N \to \infty$. This equation has a closed solution. This is the mgf of the limiting constant $\varrho(t)$.

In this way we have proven the convergence in distribution of $N^{-1}M^{(N)}(t)$ to $\varrho(t)$; as this limit is a constant, convergence in probability follows immediately.

Let us go through the procedure in full detail now. Postmultiplication of Eqn. (5) with $F$ and $N^{-\alpha}$, using $QF = \Pi - I$, results in the recurrence relation

$$\begin{aligned} \boldsymbol{p}^{(N)}(t,z) &= \boldsymbol{p}^{(N)}(t,z)\,\Pi + N^{-\alpha}(z-1)\left(N\boldsymbol{p}^{(N)}(t,z)\Lambda - \frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial z}\mathcal{M}\right)F \\ &\quad - N^{-\alpha}\frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial t}F. \end{aligned} \tag{6}$$

We are now set to state and prove the mean behavior of $M^{(N)}(t)$. Define $\varrho(t) := \varrho\,(1-e^{-\mu t})$, with $\varrho = \lambda_\infty/\mu$.

**Lemma 1.** *$N^{-1}M^{(N)}(t)$ converges in probability to $\varrho(t)$, as $N \to \infty$.*

*Proof.* We introduce the transient scaled moment generating function $\bar{\boldsymbol{p}}^{(N)}(t,\vartheta)$:

$$\bar{\boldsymbol{p}}^{(N)}(t,\vartheta) := \boldsymbol{p}^{(N)}(t,z),$$

with $z \equiv z^{(N)}(\vartheta) = \exp(\vartheta/N)$. Evidently,

$$\frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t} = \frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial t}, \qquad \frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial \vartheta} = \frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}\vartheta} = \frac{z}{N}\,\frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial z}.$$

Substituting these expressions in Eqn. (6) and noting that $z^{\pm 1} = 1 \pm \vartheta N^{-1} + O(N^{-2})$, we obtain

$$\begin{aligned}\bar{\boldsymbol{p}}^{(N)}(t,\vartheta) \;\; = \;\; & \bar{\boldsymbol{p}}^{(N)}(t,\vartheta)\Pi + N^{-\alpha}\left(\vartheta\,\bar{\boldsymbol{p}}^{(N)}(t,\vartheta)\,\Lambda - \vartheta\,\frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial \vartheta}\mu I\right. \\ & \left. - \frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t}\right) F + o(N^{-\alpha}).\end{aligned}$$

The above implies that $\bar{\boldsymbol{p}}^{(N)}(t,\vartheta) = \bar{\boldsymbol{p}}^{(N)}(t,\vartheta)\Pi + O(N^{-\alpha})$, and the same holds for the partial derivatives of $\bar{\boldsymbol{p}}^{(N)}(t,\vartheta)$, so all $\bar{\boldsymbol{p}}^{(N)}(t,\vartheta)$ between the brackets can be replaced by $\bar{\boldsymbol{p}}^{(N)}(t,\vartheta)\,\Pi$. Postmultiplying by $\mathbf{1}N^{\alpha}$ and using the identities $\Pi\mathbf{1} = \mathbf{1}$ and $F\mathbf{1} = \mathbf{1}$, yields

$$0 = \left(\vartheta\lambda_\infty \bar{\boldsymbol{p}}^{(N)}(t,\vartheta)\,\mathbf{1} - \mu\vartheta\,\frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial \vartheta}\mathbf{1} - \frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t}\mathbf{1}\right) + o(1);$$

recall the definitions $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$ and $\lambda_\infty := \boldsymbol{\pi}^{\mathrm{T}}\Lambda\mathbf{1}$. Define $\bar{\boldsymbol{p}}(t,\vartheta)\mathbf{1}$ as the limit of $\bar{\boldsymbol{p}}^{(N)}(t,\vartheta)\mathbf{1}$ as $N \to \infty$. Now multiply the differential equation with $N^{\alpha}$ and let $N \to \infty$. We thus obtain a scalar partial differential equation in $\bar{\boldsymbol{p}}(t,\vartheta)\mathbf{1}$

$$\frac{\partial(\bar{\boldsymbol{p}}(t,\vartheta)\mathbf{1})}{\partial t} = \vartheta\lambda_\infty(\bar{\boldsymbol{p}}(t,\vartheta)\mathbf{1}) - \mu\vartheta\frac{\partial(\bar{\boldsymbol{p}}(t,\vartheta)\mathbf{1})}{\partial \vartheta}.$$

It is straightforward to check that $\bar{\boldsymbol{p}}(t,\vartheta)\mathbf{1} = \exp(\vartheta\varrho(t))$ satisfies the equation as well as the boundary conditions $\bar{\boldsymbol{p}}(t,0)\mathbf{1} = 1$ and $\bar{\boldsymbol{p}}(0,\vartheta)\mathbf{1} = 1$. Now the stated follows directly. □

3.3. **Recurrence relations for the centered and normalized mgf $\tilde{\boldsymbol{p}}^{(N)}$.** Now that we have derived the weak law of large numbers, we introduce in the next step the centered and normalized mgf $\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)$, that is, centered around $N\varrho(t)$ and normalized by $N^{-\gamma}$, with the scalar $\gamma$ yet to be determined. We perform a change of variables in the recurrence relation for $\boldsymbol{p}^{(N)}$, Eqn. (6), so as to obtain the recurrence relation for the centered and normalized mgf $\tilde{\boldsymbol{p}}^{(N)}$.

The pgf $\boldsymbol{p}^{(N)}$ can be expressed in the normalized and centered mgf $\tilde{\boldsymbol{p}}^{(N)}$ using

$$\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta) = \exp(-N\varrho(t)\vartheta/N^{\gamma})\,\boldsymbol{p}^{(N)}\left(t,\exp(\vartheta/N^{\gamma})\right),$$

which can be written as

$$\boldsymbol{p}^{(N)}(t,z) = \exp(\varrho(t)\vartheta N^{1-\gamma})\,\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta),$$

with $z \equiv z^{(N)}(\vartheta) = \exp(\vartheta N^{-\gamma})$. It is readily verified that

$$\begin{aligned}\frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial z}\frac{\mathrm{d}z}{\mathrm{d}\vartheta} &= \exp(\varrho(t)\vartheta N^{1-\gamma})\left(\varrho(t)N^{1-\gamma}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)+\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(\vartheta)}{\partial\vartheta}\right);\\ \frac{\mathrm{d}z}{\mathrm{d}\vartheta} &= N^{-\gamma}\exp(\vartheta N^{\gamma}) = N^{-\gamma}z,\end{aligned}$$

so the derivatives of $\boldsymbol{p}^{(N)}$ can be expressed in terms of the corresponding derivatives of $\tilde{\boldsymbol{p}}^{(N)}$:

$$\begin{aligned}\frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial t} &= \exp(\varrho(t)\vartheta N^{1-\gamma})\left(\varrho'(t)\vartheta N^{1-\gamma}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)+\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t}\right),\\ \frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial z} &= \frac{1}{z}\exp(\varrho(t)\vartheta N^{1-\gamma})\left(N\varrho(t)\,\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)+N^{\gamma}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial\vartheta}\right).\end{aligned}$$

Now perform the change of variables and substitute the expressions for $\boldsymbol{p}^{(N)}(t,z)$ and its partial derivatives into Eqn. (6). Dividing by $\exp(\varrho(t)\vartheta N^{1-\gamma})$ yields the following recurrence relation for $\tilde{\boldsymbol{p}}^{(N)}$:

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta) &= \tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\Pi + N^{1-\alpha}\left(z^{(N)}(\vartheta)-1\right)\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\Lambda F\\ &\quad -N^{1-\alpha}\left(1-\frac{1}{z^{(N)}(\vartheta)}\right)\varrho(t)\,\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\mathcal{M}F\\ &\quad -N^{\gamma-\alpha}\left(1-\frac{1}{z^{(N)}(\vartheta)}\right)\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial\vartheta}\mathcal{M}F\\ &\quad -N^{1-\alpha-\gamma}\varrho'(t)\vartheta\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)F - N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t}F. \qquad (7)\end{aligned}$$

3.4. **Differential equation for the scalar, centered and normalized, mgf $\phi^{(N)}$.** The next step is to expand $z$ in a Taylor series. Assuming certain restrictions on $\gamma$ (that we later justify) we delete all terms of order smaller than $N^{-\alpha}$. The resulting recurrence relation is iterated and manipulated until all terms in the right-hand side contain $\tilde{\boldsymbol{p}}^{(N)}\Pi$. Next we postmultiply this system of partial differential equations by $\mathbf{1}$, so as to obtain a scalar partial differential equation in terms of $\phi^{(N)}(t,\vartheta) := \tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\mathbf{1}$. In this step we make use of the definition of $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$ and the identities $\Pi\mathbf{1} = \mathbf{1}$ and $F\mathbf{1} = \mathbf{1}$.

The Taylor expansions of $z$ and $z^{-1}$ are

$$z^{\pm 1} = 1 \pm \vartheta N^{-\gamma} + \frac{1}{2}\vartheta^2 N^{-2\gamma} + O(N^{-3\gamma}),$$

Applying these to Eqn. (7) results in

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta) &= \tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\Pi + \vartheta N^{1-\alpha-\gamma}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)(\Lambda-\varrho(t)\mathcal{M}-\varrho'(t)I)F\\ &\quad+\frac{\vartheta^2}{2}N^{1-\alpha-2\gamma}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)(\Lambda+\varrho(t)\mathcal{M})F\\ &\quad-\vartheta N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial\vartheta}\mathcal{M}F - N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t}F + O(N^{1-\alpha-3\gamma}) + O(N^{-\alpha-\gamma}). \qquad (8)\end{aligned}$$

Under the assumption that $\gamma > 1/3$ (to be justified later) the order terms can be replaced by $o(N^{-\alpha})$.

Next we iterate Eqn. (8) until all terms in the right-hand side either contain $\tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Pi$ or are of $O(N^{-\alpha})$. For the latter we assume a second restriction, viz., $\gamma \geq 1-\alpha/2$ (also justified later). We thus obtain

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta) &= \tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\Pi\\ &\quad+\vartheta N^{1-\alpha-\gamma}\Big(\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\Pi + \vartheta N^{1-\alpha-\gamma}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)(\Lambda-\varrho(t)\mathcal{M}-\varrho'(t)I)F+\\ &\qquad O(N^{1-\alpha-2\gamma}) + O(N^{-\alpha})\Big)(\Lambda-\varrho(t)\mathcal{M}-\varrho'(t)I)F\\ &\quad+\frac{\vartheta^2}{2}N^{1-\alpha-2\gamma}\Big(\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\Pi + O(N^{1-\alpha-\gamma}) + O(N^{-\alpha})\Big)(\Lambda+\varrho(t)\mathcal{M})F\\ &\quad-\vartheta N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial\vartheta}\mathcal{M}F - N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t}F + o(N^{-\alpha});\end{aligned}$$

here we remark that in the $O(N^{-\alpha})$-terms $\tilde{\boldsymbol{p}}^{(N)}$ can be replaced by $\tilde{\boldsymbol{p}}^{(N)}\Pi$ as an immediate consequence of the fact that Eqn. (8) implies $\tilde{\boldsymbol{p}}^{(N)} = \tilde{\boldsymbol{p}}^{(N)}\Pi + o(1)$, while the same applies to its derivatives. The above equation can be rewritten as

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta) &= \tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\Pi + \vartheta N^{1-\alpha-\gamma}\,\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\,\Pi(\Lambda-\varrho(t)\mathcal{M}-\varrho'(t)I)F\\ &\quad+\vartheta^2 N^{2-2\alpha-2\gamma}\,\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\,\Pi(\Lambda-\varrho(t)\mathcal{M}-\varrho'(t)I)F(\Lambda-\varrho(t)\mathcal{M}-\varrho'(t)I)F\\ &\quad+\frac{\vartheta^2}{2}N^{1-\alpha-2\gamma}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\,\Pi(\Lambda+\varrho(t)\mathcal{M})F\\ &\quad-\vartheta N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial\vartheta}\,\Pi\mathcal{M}F - N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t}\,\Pi F + o(N^{-\alpha}). \qquad (9)\end{aligned}$$

Now postmultiply Eqn. (9) by $\mathbf{1}\,N^{\alpha}$; using the identities $\Pi\mathbf{1} = \mathbf{1}$ and $F\mathbf{1} = \mathbf{1}$, and the definition $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$. We obtain

$$\begin{aligned}0 &= \vartheta N^{1-\gamma}\phi^{(N)}(t,\vartheta)\,\boldsymbol{\pi}^{\mathrm{T}}\left(\Lambda-\varrho(t)\mathcal{M}-\varrho'(t)I\right)\mathbf{1}\\ &\quad+\vartheta^2 N^{2-\alpha-2\gamma}\,\phi^{(N)}(t,\vartheta)\,\boldsymbol{\pi}^{\mathrm{T}}(\Lambda-\varrho(t)\mathcal{M}-\varrho'(t)I)F(\Lambda-\varrho(t)\mathcal{M}-\varrho'(t)I)\mathbf{1}\\ &\quad+\frac{\vartheta^2}{2}N^{1-2\gamma}\phi^{(N)}(t,\vartheta)\,\boldsymbol{\pi}^{\mathrm{T}}(\Lambda+\varrho(t)\mathcal{M})\mathbf{1} - \vartheta\mu\frac{\partial\phi^{(N)}(t,\vartheta)}{\partial\vartheta} - \frac{\partial\phi^{(N)}(t,\vartheta)}{\partial t} + o(1).\end{aligned}$$

Directly from the definition of $\varrho(t)$, it is seen that the first term on the right-hand side vanishes. In addition, it takes some elementary algebra to check that

$$\boldsymbol{\pi}^{\mathrm{T}}(\Lambda-(\varrho(t)\mu+\varrho'(t))I)F(\Lambda-(\varrho(t)\mu+\varrho'(t)I)\mathbf{1}=\boldsymbol{\pi}^{\mathrm{T}}\Lambda D\Lambda\mathbf{1}=:U,$$

where we used $F=D+\Pi=D+\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$, and

$$\frac{1}{2}\boldsymbol{\pi}^{\mathrm{T}}(\Lambda+\varrho(t)\mu I)\mathbf{1}=\lambda_\infty\left(1-\frac{e^{-\mu t}}{2}\right).$$

This results in the partial differential equation

$$\begin{aligned}&\frac{\partial\phi^{(N)}(t,\vartheta)}{\partial t}+\vartheta\mu\frac{\partial\phi^{(N)}(t,\vartheta)}{\partial\vartheta}\\ &\quad=\ \vartheta^2\phi^{(N)}(t,\vartheta)\left(N^{2-\alpha-2\gamma}U+\frac{1}{2}N^{1-2\gamma}\lambda_\infty(1-\frac{1}{2}e^{-\mu t})\right)+o(1).\end{aligned}\tag{10}$$

3.5. **Limit solution.** The last step in our proof is to obtain the limiting differential equation for $\phi(t,\vartheta)$, being the limit of $\phi^{(N)}(\vartheta,t)$. Its unique solution corresponds to a normal distribution $\mathcal{N}(0,\sigma^2(t))$.

First, note that if we choose $\gamma$ larger than both $1-\alpha/2$ and $1/2$, we do not obtain a CLT, but rather that the random variable under study converges in distribution to the constant 0. Hence, we take $\gamma=\max\{1-\alpha/2,1/2\}$, in which case the largest term dominates in (10), with both terms contributing if $\alpha=1$. Note that this choice is consistent with the restrictions on $\gamma$ we used during our proof. We obtain by sending $N\to\infty$,

$$\frac{\partial\phi(t,\vartheta)}{\partial t}+\vartheta\mu\frac{\partial\phi(t,\vartheta)}{\partial\vartheta}=\vartheta^2\phi(t,\vartheta)\,g(t),\tag{11}$$

with $g(t):=U\,1_{\{\alpha\le1\}}+(\lambda_\infty(1-e^{-\mu t}/2))\,1_{\{\alpha\ge1\}}$.

We propose the *ansatz*

$$\phi(t,\vartheta)=\exp\left(\frac{1}{2}\vartheta^2e^{-2\mu t}f(t)\right),$$

for some unknown function $f(t)$; recognize the mgf associated with the Normal distribution. This leads to the following ordinary differential equation for $f(t)$:

$$f'(t)=2e^{2\mu t}g(t),$$

which is obviously solved by integrating the right-hand side. From this we immediately find the expression for the variance $\sigma^2(t)$ of the Normal distribution.

With this last step we have proven our claim. It is instructive to compare the findings with the expressions obtained in Section 2.2.

**Theorem 1.** *Consider Model* I *or* II *with* $\mu_i = \mu$ *for all* $i \in \{1, \ldots, d\}$. *The random variable*

$$\frac{M^{(N)}(t) - N\varrho(t)}{N^\gamma}$$

*converges to a Normal distribution with zero mean and variance* $\sigma^2(t)$ *as* $N \to \infty$*; here the parameter* $\gamma$ *equals* $\max\{1 - \alpha/2, 1/2\}$, *and* $\sigma^2(t) := \sigma_m^2(t)1_{\{\alpha \leq 1\}} + \varrho(t)1_{\{\alpha \geq 1\}}$, *with* $\sigma_m^2(t) := \mu^{-1}(1 - e^{-2\mu t})U$.

**Corollary 1.** *Consider Model* I *or* II *with* $\mu_i = \mu$ *for all* $i \in \{1, \ldots, d\}$. *The random variable*

$$\frac{M^{(N)} - N\varrho}{N^\gamma}$$

*converges to a Normal distribution with zero mean and variance* $\sigma^2$ *as* $N \to \infty$*; here the parameter* $\gamma$ *equals* $\max\{1 - \alpha/2, 1/2\}$, *and* $\sigma^2 := \sigma_m^2 1_{\{\alpha \leq 1\}} + \varrho 1_{\{\alpha \geq 1\}}$, *with* $\sigma_m^2 := \mu^{-1}U$.

## 4. Model I: Stationary and transient distribution

In this section we briefly recall the results of the steps for Model I, both for the stationary and time-dependent behavior. The proofs are analogous to those in the previous section. Comparing the results of Lemma 2 and Thm. 2 with those of Lemma 1 and Thm. 1, respectively, the effect of heterogeneous service rates becomes visible.

**Proposition 2.** *Consider Model* I. *In the stationary case the pgf* $\boldsymbol{p}(z)$ *satisfies the following differential equation:*

$$\boldsymbol{p}(z)Q = (z-1)\left(\frac{\mathrm{d}\boldsymbol{p}(z)}{\mathrm{d}z}\mathcal{M} - \boldsymbol{p}(z)\Lambda\right).$$

*In the transient case the pgf* $\boldsymbol{p}(t, z)$ *satisfies the following differential equation:*

$$\frac{\partial \boldsymbol{p}(t,z)}{\partial t} = \boldsymbol{p}(t,z)\,Q + (z-1)\left(\boldsymbol{p}(t,z)\,\Lambda - \frac{\partial \boldsymbol{p}(t,z)}{\partial z}\mathcal{M}\right).$$

Define $\varrho^{(\mathrm{I})} := \lambda_\infty/\mu_\infty$, and $\varrho^{(\mathrm{I})}(t) = \varrho^{(\mathrm{I})}\,(1 - e^{-\mu_\infty t})$.

**Lemma 2.** *Consider Model* I. *As* $N \to \infty$,

(1) $N^{-1}M^{(N)}(t)$ *converges in probability to* $\varrho^{(\mathrm{I})}(t)$.
(2) $N^{-1}M^{(N)}$ *converges in probability to* $\varrho^{(\mathrm{I})}$.

**Theorem 2.** *Consider Model* I. *The random variable*

$$\frac{M^{(N)}(t) - N\varrho^{(\mathrm{I})}(t)}{N^\gamma}$$

*converges to a Normal distribution with zero mean and variance* $\sigma^2(t)$ *as* $N \to \infty$*; here* $\sigma^2(t) := \sigma_m^2(t)1_{\{\alpha \leq 1\}} + \varrho^{(\mathrm{I})}(t)1_{\{\alpha \geq 1\}}$, *with*

$$\sigma_m^2(t) := 2e^{-2\mu_\infty t}\int_0^t e^{2\mu_\infty s}\boldsymbol{\pi}^{\mathrm{T}}(\Lambda - \varrho^{(\mathrm{I})}(s)\mathcal{M})D(\Lambda - \varrho^{(\mathrm{I})}(s)\mathcal{M})\mathbf{1}\,\mathrm{d}s.$$

*The random variable*

$$\frac{M^{(N)} - N\varrho^{(\mathrm{I})}}{N^{\gamma}}$$

*converges to a Normal distribution with zero mean and variance* $\sigma^2$ *as* $N \to \infty$*; here* $\sigma^2 := \sigma_m^2 1_{\{\alpha \leq 1\}} + \varrho^{(\mathrm{I})} 1_{\{\alpha \geq 1\}}$*, with*

$$\sigma_m^2 := \mu_\infty^{-1} \boldsymbol{\pi}^{\mathrm{T}} (\Lambda - \varrho^{(\mathrm{I})} \mathcal{M}) D (\Lambda - \varrho^{(\mathrm{I})} \mathcal{M}) \mathbf{1}.$$

*In both cases the parameter* $\gamma$ *equals* $\max\{1 - \alpha/2, 1/2\}$.

The formula for $\sigma_m^2(t)$ can be evaluated more explicitly. Define $G_{m,n}(t) := e^{-m\mu_\infty t} - e^{-n\mu_\infty t}$, for $m, n \in \mathbb{N}$. Direct computations yield that $\sigma_m^2(t)$ equals

$$U \frac{1}{\mu_\infty} G_{0,2}(t) + \hat{U} \frac{\varrho^{(\mathrm{I})}}{\mu_\infty} \left(2G_{1,2}(t) - G_{0,2}(t)\right) + \check{U} \frac{(\varrho^{(\mathrm{I})})^2}{\mu_\infty} \left(G_{0,2}(t) - 4G_{1,2}(t) + 2\mu_\infty t e^{-2\mu_\infty t}\right),$$

with $\hat{U} := \boldsymbol{\pi}^{\mathrm{T}} \mathcal{M} D \Lambda \mathbf{1} + \boldsymbol{\pi}^{\mathrm{T}} \Lambda D \mathcal{M} \mathbf{1}$ and $\check{U} := \boldsymbol{\pi}^{\mathrm{T}} \mathcal{M} D \mathcal{M} \mathbf{1}$. It is readily verified that $\sigma_m^2(t) \to \sigma_m^2$ as $t \to \infty$, as expected.

## 5. Results for Model II

In this section we study Model II: the service times are now determined by the background state as seen by the job upon arrival. The approach is as before: we first derive a system of differential equations (Section 5.1), then establish the mean behavior by means of laws of large numbers (Section 5.2), and finally derive the CLTs (Section 5.3).

5.1. **Differential equations for the pgf $\boldsymbol{p}$.** For the transient distribution, a system of differential equations was previously derived in [2]. It is based on the observation that $M(t)$ has a Poisson distribution with (random) parameter $\varphi(J)$, see (1). The intuition behind this formula is that a job arriving at time $s$ survives in the system until time $t$ with probability $e^{-\mu_i (t-s)}$ (assuming that the background process is in state $i$), which is distributionally equivalent with 'thinning' the Poisson parameter with exactly this fraction. This description yields, after some manipulations, the following differential equation for the pgf, the row vector $\boldsymbol{p}(t,z)$:

$$\frac{\partial \boldsymbol{p}(t,z)}{\partial t} = \boldsymbol{p}(t,z)\tilde{Q} + (z-1)\boldsymbol{p}(t,z)\Delta(t), \tag{12}$$

where $\tilde{Q} = (\tilde{q}_{ij})_{i,j=1}^d$ is the transition rate matrix of the time-reversed version of $J(\cdot)$ (i.e., $\tilde{q}_{ij} := q_{ji}\pi_j/\pi_i$), and $\Delta(t)$ denotes a diagonal matrix with entries $[\Delta(t)]_{ii} := \lambda_i \exp(-\mu_i t)$.

**Remark 1.** It is noted that the definition of $\boldsymbol{p}$ is slightly different from the one used in [2]. In the present paper we consider the generating function of the number of jobs present at time $t$ *jointly with the state of the background process at time* $t$, whereas [2, Prop. 2] considers the generating function of the number of jobs present at time $t$ *conditioned on the background state at time* 0. As

a consequence, we obtain a slightly different equation, but it is easy to translate them into each other. ♢

Our objective is to set up our proof such that it facilitates proving both the transient and stationary CLT. Naïvely, one could try to obtain a differential equation for the stationary behavior by sending $t \to \infty$ in (12), but it is readily checked that this yields a trivial relation only: $\mathbf{0} = \mathbf{0}$. A second naïve approach would be to establish the CLT for $M^{(N)}(t)$, and to send then $t$ to $\infty$; it is clear, however, that this procedure relies on interchanging two limits ($N \to \infty$ and $t \to \infty$), of which a formal justification is lacking.

We therefore resort to an alternative approach. It relies on a description based on a more general state space: we do not only keep track of the number of jobs present, but we rather record the numbers of jobs present *of each type*, where 'type' refers to the state of the background process upon arrival. To this end, we introduce the $d$-dimensional stochastic process

$$\boldsymbol{M}(t) = (M_1(t), \ldots, M_d(t))_{t\in\mathbb{R}}\,,$$

where the $k$-th entry denotes the number of particles of type $k$ in the system at time $t$. The transient and stationary total numbers of jobs present are denoted by

$$M(t) := \sum_{k=1}^{d} M_k(t), \quad M := \sum_{k=1}^{d} M_k,$$

respectively. As usual, we add a superscript $^{(N)}$ when working with the model in which imposed our scaling on the arrival rates and the transition rates of the background process.

As before, we first derive a differential equation for the unscaled model. The generating function $\boldsymbol{p}(t, \boldsymbol{z})$ is defined as follows:

$$[\boldsymbol{p}(t, \boldsymbol{z})]_j = \mathbb{E}\left(\prod_{k=1}^{d} z_k^{M_k(t)} 1_{\{J(t)=j\}}\right).$$

In addition, $E_k$ is a matrix for which $[E_k]_{kk} = 1$, and whose other entries are zero. For a row vector $\boldsymbol{q}$, the multiplication $\boldsymbol{q}\, E_k$ thus results in a (row) vector which leaves the $k$-th entry of $\boldsymbol{q}$ unchanged while the other entries become zero. The following result covers the transient case.

**Proposition 3.** *Consider Model* II. *The pgf* $\boldsymbol{p}(t, \boldsymbol{z})$ *satisfies the following differential equation:*

$$\frac{\partial \boldsymbol{p}(t, \boldsymbol{z})}{\partial t} = \boldsymbol{p}(t, \boldsymbol{z})Q + \sum_{k=1}^{d}(z_k - 1)\left(\lambda_k\, \boldsymbol{p}(t, \boldsymbol{z})\, E_k - \mu_k \frac{\partial \boldsymbol{p}(t, \boldsymbol{z})}{\partial z_k}\right).$$

With the pgf $\boldsymbol{p}(z_1, \ldots, z_d)$ defined in the obvious way, the differential equation for the stationary case is the following.

**Proposition 4.** *Consider Model* II. *The pgf* $\boldsymbol{p}(\boldsymbol{z})$ *satisfies the following differential equation:*

$$0 = \boldsymbol{p}(\boldsymbol{z})Q + \sum_{k=1}^{d}(z_k - 1)\left(\lambda_k\, \boldsymbol{p}(\boldsymbol{z})\, E_k - \mu_k \frac{\partial \boldsymbol{p}(\boldsymbol{z})}{\partial z_k}\right).$$

The proofs of these propositions are straightforward, and follow the same lines as before: we consider the generator of the Markov process, and transform the Kolmogorov equation (for the transient case) and the invariance equation (for the stationary case).

The partial differential equation for the transient scaled model follows directly from Prop. 3, by replacing $\lambda_k$ by $N\lambda_k$, and $Q$ by $N^\alpha Q$. It results in

$$\frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial t} = N^\alpha \boldsymbol{p}^{(N)}(t,\boldsymbol{z})\, Q + \sum_{k=1}^{d}(z_k - 1)\left(N\lambda_k\, \boldsymbol{p}^{(N)}(t,\boldsymbol{z})\, E_k - \mu_k \frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial z_k}\right). \tag{13}$$

The stationary case can be dealt with analogously, relying on Prop. 4.

Our objective is to derive the CLT for both the transient and stationary case. We do so by presenting the full analysis for the transient case; in the stationary case we can leave out one term. Importantly, this approach does not have the problem of illegitimately interchanging two limits.

5.2. **Mean behavior.** As before, we first derive the law of large numbers. Again we rewrite the differential equations (13) as a recurrence relation for $\boldsymbol{p}^{(N)}$ that involves the fundamental matrix $F$:

$$\begin{aligned}\boldsymbol{p}^{(N)}(t,\boldsymbol{z}) \;&=\; \boldsymbol{p}^{(N)}(t,\boldsymbol{z})\Pi + N^{-\alpha}\sum_{k=1}^{d}(z_k - 1)\left(N\lambda_k\, \boldsymbol{p}^{(N)}(t,\boldsymbol{z})\, E_k - \mu_k \frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial z_k}\right)F \\ &\quad - N^{-\alpha}\frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial t}F \end{aligned} \tag{14}$$

for the transient case, and likewise for the stationary case.

The following lemma establishes weak laws of large numbers for $\boldsymbol{M}^{(N)}(t)$ and $M^{(N)}(t)$, as well as their steady-state counterparts $\boldsymbol{M}^{(N)}$ and $M^{(N)}$. We first define

$$\varrho_k^{(\mathrm{II})}(t) := \pi_k \frac{\lambda_k}{\mu_k}\,(1 - e^{-\mu_k t}), \quad \varrho_k^{(\mathrm{II})} := \pi_k \frac{\lambda_k}{\mu_k}.$$

Also, $\varrho^{(\mathrm{II})}(t) := \sum_k \varrho_k^{(\mathrm{II})}(t)$ and $\varrho^{(\mathrm{II})} := \sum_k \varrho_k^{(\mathrm{II})}$.

**Lemma 3.** *Consider Model* II. *As* $N \to \infty$,

(1) $N^{-1}\boldsymbol{M}^{(N)}(t)$ *converges in probability to* $\boldsymbol{\varrho}^{(\mathrm{II})}(t)$.
(2) $N^{-1}\boldsymbol{M}^{(N)}$ *converges in probability to* $\boldsymbol{\varrho}^{(\mathrm{II})}$.
(3) $N^{-1}M^{(N)}(t)$ *converges in probability to* $\varrho^{(\mathrm{II})}(t)$, *and* $N^{-1}M^{(N)}$ *to* $\varrho^{(\mathrm{II})}$.

*Proof.* Similarly to the proof of Lemma 1, we first introduce the scaled moment generating function $\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) := \boldsymbol{p}^{(N)}(t,\boldsymbol{z})$, with $z_k \equiv z_k^{(N)}(\vartheta_k) = \exp(\vartheta_k/N)$, for $k = 1,\ldots,d$. We see immediately that

$$\frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t} = \frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial t}, \qquad \frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial \vartheta_k} = \frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial z_k}\frac{\mathrm{d}z_k}{\mathrm{d}\vartheta_k} = \frac{z_k}{N}\frac{\partial \boldsymbol{p}(t,\boldsymbol{z})}{\partial z_k}.$$

Now we substitute these expressions in Eqn. (14), and note that $z_k^{\pm 1} = 1 \pm \vartheta_k N^{-1} + O(N^{-2})$. As a consequence,

$$\begin{aligned}\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) &= \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\Pi + N^{-\alpha}\sum_{k=1}^{d}\vartheta_k\left(\lambda_k\,\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\,E_k - \mu_k\frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial \vartheta_k}\right)F \\ &\quad - N^{-\alpha}\frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t}F + o(N^{-\alpha}).\end{aligned}$$

It directly follows that $\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) = \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\Pi + O(N^{-\alpha})$, and hence also

$$\frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t} = \frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t}\,\Pi + O(N^{-\alpha}), \qquad \frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial \vartheta_k} = \frac{\partial \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial \vartheta_k}\,\Pi + O(N^{-\alpha}).$$

The next step is to postmultiply the previous display by $\mathbf{1}\,N^{\alpha}$, and after some elementary steps we obtain the following scalar partial differential equation in $\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\mathbf{1}$:

$$\frac{\partial(\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\mathbf{1})}{\partial t} = \sum_{k=1}^{d}\vartheta_k\left(\pi_k\lambda_k(\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\mathbf{1}) - \mu_k\frac{\partial(\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\mathbf{1})}{\partial \vartheta_k}\right) + o(1).$$

Now let $N \to \infty$; define $\bar{\boldsymbol{p}}(t,\boldsymbol{\vartheta})\mathbf{1} := \lim_{N\to\infty}\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\mathbf{1}$. We propose the following form for the limiting function $\bar{\boldsymbol{p}}(t,\boldsymbol{\vartheta})\mathbf{1}$:

$$\bar{\boldsymbol{p}}(t,\boldsymbol{\vartheta})\mathbf{1} = \exp\left(\sum_{k=1}^{d}\vartheta_k\bar{\varrho}_k(t)\right),$$

for specific functions $\bar{\varrho}_k(\cdot)$ (to be determined later). Plugging this form into the differential equation, it means that the following equation must be fulfilled by the $\bar{\varrho}_k(\cdot)$:

$$\sum_{k=1}^{d}\vartheta_k\left(\bar{\varrho}_k'(t) - \pi_k\lambda_k + \mu_k\bar{\varrho}_k(t)\right) = 0.$$

As this must hold for any $\vartheta_k$, this equation leads to a separate differential equation for every $\bar{\varrho}_k(t)$, which moreover agrees with the one in the first part of the claim ($\bar{\varrho}_k(t) = \varrho_k^{(\mathrm{II})}(t)$, that is). We conclude that we have established the claim for the transient case: $N^{-1}\boldsymbol{M}^{(N)}(t)$ converges in probability to $\boldsymbol{\varrho}^{(\mathrm{II})}(t)$ as $N \to \infty$.

For the stationary case, we can follow precisely the same procedure, but without the partial derivative with respect to time, so that we now end up with a differential equation in $\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})\mathbf{1}$ as follows:

$$0 = \sum_{k=1}^{d}\vartheta_k\left(\pi_k\lambda_k(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})\mathbf{1}) - \mu_k\frac{\partial(\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})\mathbf{1})}{\partial \vartheta_k}\right),$$

for which $\bar{\boldsymbol{p}}(\boldsymbol{\vartheta})\mathbf{1} = \exp(\sum_{k=1}^{d} \vartheta_k \varrho_k^{(\mathrm{II})})$ forms a solution. This completes the proof of the second claim. The third claim follows trivially. □

5.3. **Central limit theorems.** Next, we state and prove the CLT result for Model II. To this end, we first define the (symmetric) matrices $V(t)$ and $V := \lim_{t\to\infty} V(t)$ with entries

$$[V(t)]_{jk} := \frac{\lambda_j \lambda_k [\bar{D}]_{jk}}{\mu_j + \mu_k}(1 - e^{-(\mu_j+\mu_k)t}), \quad [V]_{jk} = \frac{\lambda_j \lambda_k [\bar{D}]_{jk}}{\mu_j + \mu_k};$$

here $\bar{D}$ denotes the (symmetric) matrix defined by $[\bar{D}]_{jk} = (\pi_j [D]_{jk} + \pi_k [D]_{kj})$. Also, $C := \lim_{t\to\infty} C(t)$, where

$$[C(t)]_{jk} := [V(t)]_{jk} 1_{\{\alpha\le 1\}} + \varrho_j^{(\mathrm{II})}(t) 1_{\{\alpha\ge 1\}} 1_{\{j=k\}}.$$

The following theorem is the main result of this section.

**Theorem 3.** *Consider Model* II. *The random vector*

$$\frac{\boldsymbol{M}^{(N)}(t) - N\boldsymbol{\varrho}^{(\mathrm{II})}(t)}{N^\gamma}$$

*converges to a d-dimensional Normal distribution with zero mean and covariance matrix $C(t)$ as $N \to \infty$. In both cases the parameter $\gamma$ equals* $\max\{1 - \alpha/2, 1/2\}$. *The random vector*

$$\frac{\boldsymbol{M}^{(N)} - N\boldsymbol{\varrho}^{(\mathrm{II})}}{N^\gamma}$$

*converges to a d-dimensional Normal distribution with zero mean and covariance matrix $C$ as $N \to \infty$.*

*Proof.* Mimicking the proof of the CLT in Section 3, we start again with setting up a recurrence relation for the centered and normalized mgf $\tilde{\boldsymbol{p}}^{(N)}$. Define $\boldsymbol{z}$ by $z_k \equiv z_k^{(N)}(\vartheta_k) := \exp(\vartheta_k N^{-\gamma})$, for $k = 1, \ldots, d$, with the value of $\gamma$ to be determined later on. We first concentrate on the transient case and introduce the centered and normalized mgf $\tilde{\boldsymbol{p}}(t, \boldsymbol{\vartheta})$:

$$\tilde{\boldsymbol{p}}^{(N)}(t, \boldsymbol{\vartheta}) = \exp\left(-N^{1-\gamma} \sum_{k=1}^{d} \vartheta_k \varrho_k^{(\mathrm{II})}(t)\right) \boldsymbol{p}^{(N)}(t, \boldsymbol{z}).$$

We wish to perform a change of variables in Eqn. (14) to obtain a recurrence relation in $\tilde{\boldsymbol{p}}^{(N)}(t, \boldsymbol{\vartheta})$. To this end, note that

$$\frac{\partial \boldsymbol{p}^{(N)}(t, \boldsymbol{z})}{\partial z_k} \frac{\mathrm{d}z_k}{\mathrm{d}\vartheta_k} = \exp\left(N^{1-\gamma} \sum_{k=1}^{d} \vartheta_k \varrho_k^{(\mathrm{II})}(t)\right) \left(\varrho_k^{(\mathrm{II})}(t) N^{1-\gamma} \tilde{\boldsymbol{p}}^{(N)}(t, \boldsymbol{\vartheta}) + \frac{\partial \tilde{\boldsymbol{p}}^{(N)}(t, \boldsymbol{\vartheta})}{\partial \vartheta_k}\right),$$

where

$$\frac{\mathrm{d}z_k}{\mathrm{d}\vartheta_k} = N^{-\gamma} \exp(\vartheta_k N^{-\gamma}) = N^{-\gamma} z_k.$$

Also,

$$\frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial t} = \exp\left(N^{1-\gamma}\sum_{k=1}^{d}\vartheta_k \varrho_k^{(\mathrm{II})}(t)\right)\left(\sum_k \vartheta_k \frac{\mathrm{d}\varrho_k^{(\mathrm{II})}(t)}{\mathrm{d}t} N^{1-\gamma}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) + \frac{\partial \tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t}\right).$$

Now perform the change of variables, and substitute the expressions for the partial derivatives of $\boldsymbol{p}^{(N)}(t,\boldsymbol{z})$ into Eqn. (14). Dividing the equation by $\exp(N^{1-\gamma}\sum_{k=1}^{d}\vartheta_k\varrho_k^{(\mathrm{II})}(t))$ gives the following recurrence relation for $\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{z})$:

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) &= \tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\Pi + N^{1-\alpha}\sum_{k=1}^{d}(z_k-1)\lambda_k\,\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\,E_k\,F\\ &\quad - N^{-\alpha}\sum_{k=1}^{d}\left(1-\frac{1}{z_k}\right)N^{\gamma}\mu_k\left(N^{1-\gamma}\varrho_k^{(\mathrm{II})}(t)\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) + \frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial\vartheta_k}\right)F\\ &\quad - N^{1-\alpha-\gamma}\sum_{k=1}^{d}\vartheta_k\frac{\mathrm{d}\varrho_k^{(\mathrm{II})}(t)}{\mathrm{d}t}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})F - N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t}F.\end{aligned}$$

The next step is to introduce the second order Taylor expansions for $z_k$ and $z_k^{-1}$:

$$z_k^{\pm 1} = 1 \pm \vartheta_k N^{-\gamma} + \frac{1}{2}\vartheta_k^2 N^{-2\gamma} + O(N^{-3\gamma}).$$

Ignoring all terms that are provably smaller than $N^{-\alpha}$ under the assumption that $\gamma > 1/3$ (justified later), and combining terms of the same order, we obtain

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) &= \tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\,\Pi + N^{1-\alpha-\gamma}\sum_{k=1}^{d}\vartheta_k\,\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\left(\lambda_k E_k - \mu_k\varrho_k^{(\mathrm{II})}(t)I - \frac{\mathrm{d}\varrho_k^{(\mathrm{II})}(t)}{\mathrm{d}t}I\right)F\\ &\quad + N^{1-\alpha-2\gamma}\sum_{k=1}^{d}\frac{\vartheta_k^2}{2}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\left(\lambda_k E_k + \mu_k\varrho_k^{(\mathrm{II})}(t)I\right)F\\ &\quad - N^{-\alpha}\sum_{k=1}^{d}\vartheta_k\mu_k\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial\vartheta_k}F - N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t}F,\end{aligned}$$

up to an error term that is $o(N^{-\alpha})$. As we did in the proof of the CLT in Section 3 with Eqn. (8), we iterate and manipulate this relation, under the assumption that $\gamma \geq 1-\alpha/2$ (justified later), until all terms in the right-hand side contain $\tilde{\boldsymbol{p}}^{(N)}\Pi$. Then we postmultiply with $\mathbf{1}\,N^{\alpha}$, and develop a differential equation in terms of $\phi^{(N)}(t,\boldsymbol{\vartheta}) := \tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\,\mathbf{1}$. After some (by now quite familiar) manipulations, we obtain the following partial differential equation in $\phi^{(N)}(t,\boldsymbol{\vartheta})$:

$$\begin{aligned}\frac{\partial\phi^{(N)}(t,\boldsymbol{\vartheta})}{\partial t} + \sum_{k=1}^{d}\vartheta_k\mu_k\frac{\partial\phi^{(N)}(t,\boldsymbol{\vartheta})}{\partial\vartheta_k} &= \frac{1}{2}\phi^{(N)}(t,\boldsymbol{\vartheta})\left(N^{2-\alpha-2\gamma}\sum_{j=1}^{d}\sum_{k=1}^{d}\vartheta_j\vartheta_k\lambda_j\lambda_k[\bar{D}]_{jk}\right.\\ &\qquad\left. + N^{1-2\gamma}\sum_{k=1}^{d}\vartheta_k^2\pi_k(\lambda_k+\mu_k\varrho_k^{(\mathrm{II})}(t))\right) + o(1),\end{aligned}$$

where we have used that

$$\boldsymbol{\pi}^{\mathrm{T}}\left(\sum_{j=1}^{d}\vartheta_j(\lambda_j E_j-\mu_j\varrho_j^{(\mathrm{II})}(t)I-\frac{\mathrm{d}\varrho_k^{(\mathrm{II})}(t)}{\mathrm{d}t}I)\right)F\left(\sum_{k=1}^{d}\vartheta_k(\lambda_k E_k-\mu_k\varrho_k^{(\mathrm{II})}(t)I-\frac{\mathrm{d}\varrho_k^{(\mathrm{II})}(t)}{\mathrm{d}t}I)\right)\mathbf{1}$$
$$=\sum_{j=1}^{d}\sum_{k=1}^{d}\vartheta_j\vartheta_k\lambda_j\lambda_k\left(\boldsymbol{\pi}^{\mathrm{T}}E_jDE_k\mathbf{1}\right)=\frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d}\vartheta_j\vartheta_k\lambda_j\lambda_k[\bar{D}]_{jk}.$$

The last part of the proof concerns the limiting behavior as $N\to\infty$. Pick, as before, $\gamma=\max\{1-\alpha/2,1/2\}$, to obtain the following partial differential equation:

$$\frac{\partial\phi(t,\boldsymbol{\vartheta})}{\partial t}+\sum_{k=1}^{d}\vartheta_k\mu_k\frac{\partial\phi(t,\boldsymbol{\vartheta})}{\partial\vartheta_k}$$
$$=\frac{1}{2}\phi(t,\boldsymbol{\vartheta})\left(\sum_{j=1}^{d}\sum_{k=1}^{d}\vartheta_j\vartheta_k\lambda_j\lambda_k[\bar{D}]_{jk}1_{\{\alpha\leq 1\}}+\sum_{k=1}^{d}\vartheta_k^2(\pi_k\lambda_k+\mu_k\varrho_k^{(\mathrm{II})}(t))1_{\{\alpha\geq 1\}}\right).$$

It is straightforward to verify that the following expression constitutes a solution for this differential equation:

$$\phi(t,\boldsymbol{\vartheta})=\exp\left(\frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d}\vartheta_j\vartheta_k[V(t)]_{jk}1_{\{\alpha\leq 1\}}+\frac{1}{2}\sum_{k=1}^{d}\vartheta_k^2\varrho_k^{(\mathrm{II})}(t)1_{\{\alpha\geq 1\}}\right).$$

If we redo the derivation for the stationary case (i.e., we now discard the terms originating from the derivative with respect to $t$ in the original partial differential equation), we end up with

$$\phi(\boldsymbol{\vartheta})=\exp\left(\frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d}\vartheta_j\vartheta_k[V]_{jk}1_{\{\alpha\leq 1\}}+\frac{1}{2}\sum_{k=1}^{d}\vartheta_k^2\varrho_k^{(\mathrm{II})}1_{\{\alpha\geq 1\}}\right).$$

This completes the proof. □

**Corollary 2.** *Consider Model* II. *An immediate consequence of Thm.* 3 *is that, with* $\gamma$ *as defined before, the random variables*

$$\frac{M^{(N)}-N\varrho^{(\mathrm{II})}}{N^\gamma}\quad\textit{and}\quad\frac{M^{(N)}(t)-N\varrho^{(\mathrm{II})}(t)}{N^\gamma}$$

*converge to Normal distributions with zero mean and variances*

$$\sum_{j=1}^{d}\sum_{k=1}^{d}[V]_{jk}1_{\{\alpha\leq 1\}}+\varrho^{(\mathrm{II})}1_{\{\alpha\geq 1\}}\quad\textit{and}\quad\sum_{j=1}^{d}\sum_{k=1}^{d}[V(t)]_{jk}1_{\{\alpha\leq 1\}}+\varrho^{(\mathrm{II})}(t)1_{\{\alpha\geq 1\}},$$

*respectively, as* $N\to\infty$.

## 6. Correlation across time

Above we analyzed the joint distribution of the two queues at a given point in time. A related question, to be covered in this section, concerns the joint distribution at distinct time epochs. For

ease we assume that the service rates are identical (and equal to $\mu$), so that Model I and Model II coincide.

6.1. **Differential equation.** We follow the line of reasoning of [2, Prop. 2]; we consider again the non-scaled model, but, as before, these results can be trivially translated in terms of the $N$-scaled model. Fix time epochs $0 \equiv s_1 \leq s_2 \leq \cdots \leq s_K$ for some $K \in \mathbb{N}$. The goal of this subsection is to characterize the joint transform, for $j = 1, \ldots, d$,

$$\Psi_j(t, \boldsymbol{z}) := \mathbb{E}\left( \prod_{k=1}^{K} z_k^{M(t+s_k)} \,\middle|\, J(0) = j \right).$$

Assume a job arrives between 0 and $\Delta t$, for an infinitesimally small $\Delta t$. Then it is still in the system at time $t+s_k$, but not anymore at $t+s_{k+1}$ with probability $f_k(t) - f_{k+1}(t)$, where $f_k(t) := e^{-\mu(t+s_k)}$. As a consequence, we obtain the following relation:

$$\begin{aligned}\Psi_j(t, \boldsymbol{z}) &= \lambda_j \Delta t \, b(t, \boldsymbol{z}) \, \Psi_j(t - \Delta t, \boldsymbol{z}) \\ &\quad + \sum_{i \neq j} q_{ji} \Delta t \, \Psi_i(t - \Delta t, \boldsymbol{z}) + \left( 1 - \lambda_j \Delta t - \sum_{i \neq j} q_{ji} \Delta t \right) \Psi_j(t - \Delta t, \boldsymbol{z}) + o(\Delta t),\end{aligned}$$

where

$$\begin{aligned} b(t, \boldsymbol{z}) \;&:=\; (1 - f_1(t)) + z_1(f_1(t) - f_2(t)) + \cdots \\ &\qquad + (z_1 \cdots z_{K-1})(f_{K-1}(t) - f_K(t)) + (z_1 \cdots z_K) f_K(t). \end{aligned}$$

With elementary manipulations, we obtain

$$\frac{\Psi_j(t, \boldsymbol{z}) - \Psi_j(t - \Delta t, \boldsymbol{z})}{\Delta t} = \sum_{i=1}^{d} q_{ji} \Psi_i(t - \Delta t, \boldsymbol{z}) + a_j(t, \boldsymbol{z}) \Psi_j(t - \Delta t, \boldsymbol{z}) + o(1),$$

where $a_j(t, \boldsymbol{z}) := \lambda_j \left( b(t, \boldsymbol{z}) - 1 \right)$. Now letting $\Delta t \downarrow 0$, and defining $A(t, \boldsymbol{z}) := \operatorname{diag}\{\boldsymbol{a}(t, \boldsymbol{z})\}$, we obtain the differential equation, in vector notation,

$$\frac{\partial}{\partial t} \boldsymbol{\Psi}(t, \boldsymbol{z}) = (Q + A(t, \boldsymbol{z})) \boldsymbol{\Psi}(t, \boldsymbol{z}).$$

6.2. **Covariance.** We now explicitly compute $\mathbb{C}\mathrm{ov}(M(s), M(t))$, assuming, without loss of generality, that $s \leq t$; the computations are similar to the ones in Section 2.2 (and therefore some steps are left out). The 'law of total covariance', with $J \equiv (J(r))_{r=0}^{t}$, entails that

$$\mathbb{C}\mathrm{ov}(M(s), M(t)) = \mathbb{E}\, \mathbb{C}\mathrm{ov}(M(s), M(t) \,|\, J) + \mathbb{C}\mathrm{ov}(\mathbb{E}(M(s) \,|\, J), \mathbb{E}(M(t) \,|\, J)). \tag{15}$$

Due to the fact that $M(s)$ obeys a Poisson distribution with the random parameter $\varphi(J)$, the second term in the right hand side of (15) can be written as $I_1 + I_2$, where

$$I_1 := \sum_{i=1}^{d}\sum_{j=1}^{d}\lambda_i\lambda_j K_{ij}, \text{ where } K_{ij} := \int_0^s\int_0^v e^{-\mu(s-u)}e^{-\mu(t-v)}\pi_i\left(p_{ij}(v-u)-\pi_j\right)\mathrm{d}u\mathrm{d}v,$$

$$I_2 := \sum_{i=1}^{d}\sum_{j=1}^{d}\lambda_i\lambda_j L_{ij}, \text{ where } L_{ij} := \int_0^s\int_v^t e^{-\mu(s-u)}e^{-\mu(t-v)}\pi_j\left(p_{ji}(u-v)-\pi_i\right)\mathrm{d}u\mathrm{d}v.$$

It takes some standard algebra to obtain

$$\begin{aligned} K_{ij} &= e^{-\mu t}\pi_i\int_0^s\left(\int_w^s e^{2\mu v}\mathrm{d}v\right)e^{-\mu(s+w)}\left(p_{ij}(w)-\pi_j\right)\mathrm{d}w \\ &= \frac{1}{2\mu}e^{-\mu t}\pi_i\int_0^s\left(e^{\mu(s-w)}-e^{-\mu(s-w)}\right)\left(p_{ij}(w)-\pi_j\right)\mathrm{d}w. \end{aligned}$$

Similarly, $L_{ij} = L_{ij}^{(1)} + L_{ij}^{(2)}$, where

$$\begin{aligned} L_{ij}^{(1)} &:= \frac{1}{2\mu}e^{-\mu t}\pi_j\left(e^{\mu s}-e^{-\mu s}\right)\int_0^{t-s}e^{\mu w}\left(p_{ji}(w)-\pi_i\right)\mathrm{d}w, \\ L_{ij}^{(2)} &:= \frac{1}{2\mu}e^{-\mu s}\pi_j\int_{t-s}^{t}\left(e^{\mu(t-w)}-e^{-\mu(t-w)}\right)\left(p_{ji}(w)-\pi_i\right)\mathrm{d}w. \end{aligned}$$

Now concentrate on the first term in the right hand side of (15). To this end, consider the following decomposition:

$$M(s) := M^{(1)}(s,t) + M^{(2)}(s,t), \quad M(t) := M^{(2)}(s,t) + M^{(3)}(s,t),$$

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

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

Mimicking the arguments used in [6], it is immediate that $M^{(2)}(s,t)$ has a Poisson distribution with random parameter $\xi(J)$, where

$$\xi(f) := \int_0^s \lambda_{f(r)}e^{-\mu_{f(r)}(t-r)}\mathrm{d}r.$$

We conclude that

$$\mathbb{E}\,\mathbb{C}\mathrm{ov}(M(s),M(t)\,|\,J) = \mathbb{E}\xi(J) = \sum_{i=1}^{d}\pi_i\lambda_i\int_0^s e^{-\mu(t-r)}\mathrm{d}r = \varrho(s)\,e^{-\mu(t-s)}.$$

When scaling $\boldsymbol{\lambda} \mapsto N\boldsymbol{\lambda}$ and $Q \mapsto N^\alpha Q$, for $\alpha > 0$, it is readily verified that for $N$ large,

$$\mathbb{C}\mathrm{ov}(M^{(N)}(s), M^{(N)}(t)) \sim N\varrho(s)\, e^{-\mu(t-s)} + N^{2-\alpha}\frac{e^{-\mu t}}{\mu}\left(e^{\mu s} - e^{-\mu s}\right) U,$$

recalling that $U := \boldsymbol{\pi}^{\mathrm{T}}\Lambda D\Lambda\mathbf{1}$. When taking $s = t$, we obtain formulae for the variance that are in line with our findings of Section 2.2.

6.3. **Limit results.** We again consider the situation in which the modulating Markov chain $J(\cdot)$ is sped up by a factor $N^\alpha$ (for some positive $\alpha$), while the arrival rates $\lambda_i$ are sped up by $N$. In this subsection we consider the (multivariate) distribution of the number of jobs in the system at different points in time. While in [2] we just covered the case of $\alpha > 1$, we now establish a CLT for general $\alpha$.

As the techniques used are precisely the same as before, we just state the result. We first introduce some notation. Define $[\check{C}(t)]_{k\ell} = [\check{C}(t)]_{\ell k}$, where for $k \geq \ell$

$$[\check{C}(t)]_{k\ell} := \frac{U}{\mu}\left(1 - e^{-2\mu(t+s_\ell)}\right) e^{-\mu(s_k - s_\ell)} 1_{\{\alpha \leq 1\}} + \frac{\lambda_\infty}{\mu}\left(1 - e^{-\mu(t+s_\ell)}\right) e^{-\mu(s_k - s_\ell)} 1_{\{\alpha \geq 1\}}.$$

**Theorem 4.** *The random vector*

$$\left(\frac{M^{(N)}(t+s_1) - N\varrho(t+s_1)}{N^\gamma}, \ldots, \frac{M^{(N)}(t+s_K) - N\varrho(t+s_K)}{N^\gamma}\right)$$

*converges to a $K$-dimensional Normal distribution with zero mean and covariance matrix $\check{C}(t)$ as $N \to \infty$. The parameter $\gamma$ equals* $\max\{1 - \alpha/2, 1/2\}$.

As $t \to \infty$, $\check{C}(t) \to \check{C}$, where

$$[\check{C}]_{k\ell} = \frac{u_{k\ell}}{2\mu}, \quad \text{with} \quad u_{k\ell} := 2\left(U 1_{\{\alpha \leq 1\}} + \lambda_\infty 1_{\{\alpha \geq 1\}}\right) e^{-\mu(s_k - s_\ell)}.$$

We observe that the limiting centered and scaled process, as $t \to \infty$, has the correlation structure of an Ornstein-Uhlenbeck process $S(t)$ (at the level of finite-dimensional distributions), that is, the solution to the stochastic differential equation

$$\mathrm{d}S(t) = -\mu S(t)\mathrm{d}t + \left(2U 1_{\{\alpha \leq 1\}} + \sqrt{\lambda_\infty + \mu\varrho(t)}\, 1_{\{\alpha \geq 1\}}\right)\mathrm{d}W(t),$$

with $W(\cdot)$ standard Brownian motion.

## 7. Numerical illustration

In this section, we briefly illustrate the accuracy of the approximations that are suggested by the limit theorems of this paper. In particular, we consider the variance of the queue content $M^{(N)}$ under stationarity for Model I. In this case, there is an exact expression for the variance [14]:

$$\mathbb{V}\mathrm{ar}[M^{(N)}] = 2N^2\boldsymbol{\pi}^{\mathrm{T}}\Lambda(\mathcal{M} - N^\alpha Q)^{-1}\Lambda(2\mathcal{M} - N^\alpha Q)^{-1}\mathbf{1} + N\varrho^{(\mathrm{I})} - N^2(\varrho^{(\mathrm{I})})^2.$$

On the other hand, Theorem 2 suggests the following asymptotic expression for the variance of $M^{(N)}$:

$$V_1(N) := N\varrho^{(\mathrm{I})}1_{\{\alpha\geq 1\}} + N^{2-\alpha}\sigma_m^2 1_{\{\alpha\leq 1\}}.$$

This expression discards one of the two contributions to the variance, and may therefore be less accurate when both terms are of comparable size. To remedy this effect, we propose the following simple alternative

$$V_2(N) := N\varrho^{(\mathrm{I})} + N^{2-\alpha}\sigma_m^2,$$

which is asymptotically equivalent with $V_1(N)$ as $N$ grows large.

In Fig. 1, we illustrate these approximations in the three different regimes $\varrho \gg \sigma_m^2$, $\varrho \approx \sigma_m^2$, and $\varrho \ll \sigma_m^2$, for a two-state Markov process with generator $Q$ and varying values of $\alpha$. The parameter values for the three cases are

$$Q = \begin{pmatrix} -1 & 1 \\ 3 & -3 \end{pmatrix}, \quad \begin{pmatrix} -2 & 2 \\ 1 & -1 \end{pmatrix}, \quad \begin{pmatrix} -1 & 1 \\ 3 & -3 \end{pmatrix},$$

and $\boldsymbol{\lambda} = [1,2],\ [1,2],\ [1,50]$, $\boldsymbol{\mu} = [2,1],\ [100,1],\ [2,1]$. We observe that in all cases both approximations tend to the exact values as $N$ gets larger, but the errors are dependent on the specific choices of the parameters of the Markov process. As to be expected, $V_2(N)$ is the more accurate one. The contourplots in the middle row give the relative error in the approximation $V_1(N)$. They nicely show the effect of the absence of one of the terms in the approximation: for $\varrho \gg \sigma_m^2$ the relative error is almost one if $\alpha = 1-\varepsilon$, whereas for $\varrho \ll \sigma_m^2$ this is the case for $\alpha = 1+\varepsilon$. If the two terms are in balance ($\varrho \approx \sigma_m^2$), we see an increase of the relative error around $\alpha \approx 1$, which is absent in approximation $V_2(N)$, plotted in the bottom row.

## 8. Discussion and conclusion

In this paper we derived central limit theorems (CLT s) for infinite-server queues with Markov-modulated input. In our approach the modulating Markov chain is sped up by a factor $N^\alpha$ (for some positive $\alpha$), while the arrival process is sped up by $N$. Interestingly, there is a *phase transition* in the sense that the normalization to be used in the CLT depends on the value of $\alpha$: rather than the standard normalization by $\sqrt{N}$, it turned out that the centered process should be divided by $N^\gamma$, with $\gamma$ equal to $\max\{1-\alpha/2, 1/2\}$. We have proved this by first establishing systems of differential equations for the (transient and stationary) distribution of the number of jobs in the system, and then studying their behavior under the scaling described above.

We have also derived a CLT for the *multivariate* distribution of the number of jobs present at different time instants, complementing the analysis for just $\alpha > 1$ in [2]. We anticipate weak

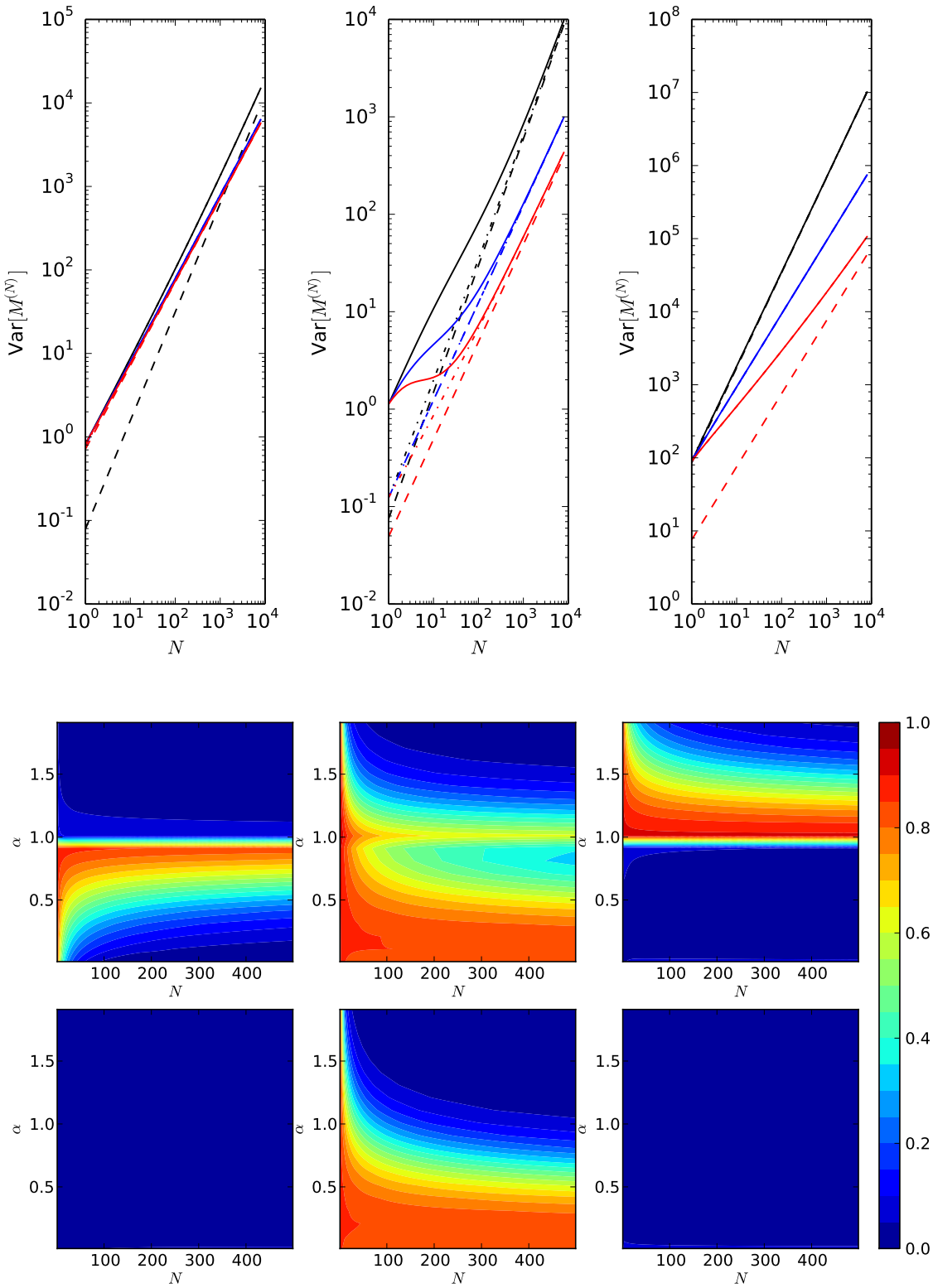


FIGURE 1. Illustration of the behaviour of the approximation in the three different regimes: (left) $\varrho \gg \sigma_m^2$, (middle) $\varrho \approx \sigma_m^2$, (right) $\varrho \ll \sigma_m^2$. Top row: Plots of the variance of $M^{(N)}$ along with two approximations; black: $\alpha = 0.7$; blue: $\alpha = 1.0$; red: $\alpha = 1.3$. Full lines represent the exact values, dashed lines represent the first approximation $V_1(N)$ and dash-dotted lines represent $V_2(N)$. Middle row: Contourplots of the relative error in the approximation $V_1(N)$ for varying $\alpha$ and $N$. Bottom row: Same for approximation $V_2(N)$.

convergence to an Ornstein-Uhlenbeck process with appropriate parameters, but establishing such a claim will require different techniques.

## Appendix A. Uniqueness of solutions of the PDEs

In the various proofs of this article, we have 'solved' the differential equations by guessing a solution and establishing that it satisfies both the differential equation itself and the boundary conditions. We now show that the solutions are indeed unique by relying on the method of characteristics [5]. The method consists of rewriting the partial differential equation (PDE) as a system of ordinary differential equations along so-called characteristic curves, for which the theory of existence and uniqueness is well-developed.

As all occurring PDEs are of a similar form and moreover quasi-linear, we can suffice by establishing uniqueness for the two types of PDEs, the first of which is as follows:

$$\sum_{k=1}^{d} \mu_k \vartheta_k \frac{\partial \phi}{\partial \vartheta_k} = g(\boldsymbol{\vartheta})\, \phi(\vartheta_1, \ldots, \vartheta_d),$$

for some function $g(\cdot)$ with boundary condition $\phi(0, \ldots, 0) = 1$. This pertains to differential equations in the proofs of Lemma 3 and Thm. 3. Let us consider a parametric curve

$$(\vartheta_1(t), \cdots, \vartheta_d(t), \phi(t)),$$

where $\phi(t) := \phi(\vartheta_1(t), \cdots, \vartheta_d(t))$ (with a slight but customary abuse of notation), subject to the following system of ordinary differential equations (ODEs):

$$\frac{\mathrm{d}\vartheta_k(t)}{\mathrm{d}t} = \mu_k \vartheta_k(t) \qquad \text{and} \qquad \frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = g(\vartheta_1(t), \ldots, \vartheta_d(t))\phi(t).$$

The ODEs in $\vartheta_k(t)$ have the following solution:

$$\vartheta_k(t) = \vartheta_k(0) \exp(\mu_k t),$$

while the ODE for $\phi$ is also quasi-linear with a continuous function $g(\cdot)$, such that a general solution can be found with one undetermined constant. In order to construct the solution at an arbitrary point $(\vartheta_1, \ldots, \vartheta_d)$, one puts $\vartheta_k(0) = \vartheta_k$ and then combines this with the boundary condition $1 = \phi(0, \cdots, 0)$, which indeed gives us the condition to make the solution of the ODE in $\phi(t)$ unique.

Next, we consider the PDE:

$$\frac{\partial \phi}{\partial t} + \sum_{k=1}^{d} \mu_k \vartheta_k \frac{\partial \phi}{\partial \vartheta_k} = g(t, \vartheta)\, \phi(t, \vartheta_1, \ldots, \vartheta_d),$$

with the boundary condition $\phi(0, \vartheta_1, \ldots, \vartheta_d) = 1$ (i.e., an empty system at $t = 0$) for which the uniqueness question can be tackled in a similar but slightly different fashion (as $t$ is now an explicit variable of the problem). This form occurs in the proofs of Thms. 1 and 3 (as well as in the proofs Lemmas 1 and 3 with the slight difference that there is a negative sign in the $\partial/\partial t$-term, which hardly changes our argument). Indeed, we consider the parametric curve:

$$(t, \vartheta_1(t), \cdots, \vartheta_d(t), \phi(t)),$$

with the same ODEs imposed on $\vartheta_k(t)$ (and hence having the same solution as well), while

$$\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} = g(t, \vartheta_1(t), \ldots, \vartheta_d(t))\, \phi(t)$$

has again a solution with one undetermined constant. In order to find the solution at $(t, \vartheta_1, \ldots, \vartheta_d)$, we put $\vartheta_k(t) = \vartheta_k$, from which we find $\vartheta_k(0) = \vartheta_k \exp(-\mu_k t)$. These relations together with $\phi(0) = 1$ ensure that each ODE has a unique solution, and hence the original PDE has a unique solution as well.

## References


[1] D. Anderson, J. Blom, M. Mandjes, H. Thorsdottir, and K. de Turck (2014). A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodology and Computing in Applied Probability*, DOI 10.1007/s11009-014-9405-8.

[2] J. Blom, O. Kella, M. Mandjes, and H. Thorsdottir (2014). Markov-modulated infinite server queues with general service times. *Queueing Systems,* **76**, 403–424.

[3] J. Blom, K. de Turck, and M. Mandjes (2013). A central limit theorem for Markov-modulated infinite-server queues. In: *Proceedings ASMTA 2013*, Ghent, Belgium. Lecture Notes in Computer Science (LNCS) Series, **7984**, pp. 81-95.

[4] J. Blom, M. Mandjes, and H. Thorsdottir (2013). Time-scaling limits for Markov-modulated infinite-server queues. *Stochastic Models*, **29**, 112–127.

[5] D. Hilbert and R. Courant (1924). *Methoden der mathematischen Physik, Vol II.* Springer, Berlin.

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

[7] P. Coolen-Schrijner and E. van Doorn (2002). The deviation matrix of a continuous-time Markov chain. *Probability in the Engineering and Informational Sciences*, **16**, 351–366.

[8] G. Falin (2008). The $M/M/\infty$ queue in a random environment. *Queueing Systems*, **58**, 65–76.

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

[10] T. Hellings, M. Mandjes, and J. Blom (2012). Semi-Markov-modulated infinite-server queues: approximations by time-scaling. *Stochastic Models*, **28**, 452–477.

[11] J. Keilson (1979). *Markov Chain Models: Rarity and Exponentiality.* Springer, New York.

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

[13] J. Kemeny and J. Snell (1961). *Finite Markov chains.* Van Nostrand, New York.

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

[15] A. Schwabe, M. Dobrzyński, and F. Bruggeman (2012). Transcription stochasticity of complex gene regulation models. *Biophysical Journal*, **103**, 1152-1161.

[16] R. Syski (1978). Ergodic potential. *Stochastic Processes and their Applications*, **7**, 311-336.

[17] T. van Woensel and N. Vandaele (2007). Modeling traffic flows with queueing models: a review. *Asia-Pacific Journal of Operational Research*, **24**, 235–261.