CENTRAL LIMIT THEOREMS FOR MARKOV-MODULATED INFINITE-SERVER QUEUES

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

Abstract. This paper studies an infinite-server queue in a Markov environment, that is, an infinite-server queue with arrival rates and service times depending on the state of an independently evolving Markovian background process. Scaling the arrival rates λ_i by a factor N and the rates q_{ij} of the background process by a factor N^α, with $\alpha \in \mathbb{R}^+$, we establish a central limit theorem as N tends to ∞. We find different scaling regimes, which depend on the specific value of α. Remarkably, for $\alpha < 1$, we find a central limit theorem in which the centered process has to be normalized by $N^{1-\alpha/2}$ rather than $\sqrt{N}$; in the expression for the variance deviation matrices appear.

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.*

[•] 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: July 2, 2013.

1. Introduction

Infinite-server queues have found widespread use in various application domains, often as an approximation for models with many servers. In these systems jobs arrive, are served in parallel, and leave when their service is completed; the jobs do not interfere with each other. The infinite-server queue was originally developed (over a century ago) to describe the dynamics of the number of calls in progress in a communication network. More recently, however, applications in various other domains have been explored, such as road traffic [14] and biology [12].
In the standard infinite-server model, jobs arrive according to a Poisson process with rate λ, where their service times form a sequence of independent and identically distributed (i.i.d.) random variables (distributed as a random variable B with finite first moment), independent of the call arrival process. A key result states that in this M/G/∞ queue the stationary number of jobs in the system obeys a Poisson distribution with mean $\lambda\,\mathbb{E}B$; i.e. there is insensitivity, in that the stationary distribution depends on B only through its mean.

In many practical situations, however, the assumptions underlying the standard infinite-server model are not realistic: there is no constant arrival rate, and the jobs do not stem from a single distribution. A model that allows the input process to exhibit some sort of 'burstiness' is the *Markov-modulated* infinite-server queue. In this system, a finite-state irreducible continuous-time Markov process (often referred to as the *background process*) modulates the input process: if the background process is in state i, the arrival process is a Poisson process with rate, say, λ_i, while the service times are distributed as a random variable, say, B_i (while the obvious independence conditions are imposed). Often the B_i s are assumed exponential with mean μ_i^{-1}.
The Markov-modulated infinite-server queue has attracted (relatively limited) attention over the past decades. The main focus in the literature so far has been on characterizing (through the derivation of moments, or even the full probability generating function) the steady-state number of jobs in the system; see e.g. [4, 6, 9, 11] and references therein. Interestingly, under an appropriate time-scaling [1, 7] in which the transitions of the background process occur at a faster rate than the Poisson arrivals, we retrieve the Poisson distribution for the steady-state number of jobs in the system. Recently, transient results have been obtained as well, under specific scalings of the arrival rates and transition times of the modulating Markov chain [1, 2].

Contribution. The present paper considers one of the scalings studied in [1, 2]: the arrival rates λ_i are scaled by a factor N and the transition rates q_{ij} of the background process by a factor N^α, for some $\alpha \in (0,\infty)$. However, where in [1, 2] only the situation of $\alpha > 1$ was considered, we now allow α to be any positive number. We focus on the number of jobs in the scaled system at time t, denoted by $M^{(N)}(t)$, aiming 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 α: for $\alpha > 1$ the variance of $M^{(N)}(t)$ scales essentially linearly in N, while for $\alpha < 1$ it behaves as $N^{2-\alpha}$.

The approach is as follows. We first derive differential equations for the probability generating functions (pgf s) of $M^{(N)}(t)$ and $M^{(N)}$. Then we establish laws of large numbers for both random quantities, so that we know how these quantities should be centered. Finally, the resulting centered random variables are scaled, so as to obtain a CLT; as could be expected from the properties of the variance of $M^{(N)}(t)$ and $M^{(N)}$, as we mentioned above, the appropriate scaling is $\sqrt{N}$ for $\alpha > 1$, and $N^{1-\alpha/2}$ for $\alpha < 1$. The proofs rely on (non-trivial) manipulations of the differential equations that describe the pgf s; interestingly *deviation matrices* [5] play a crucial role here.

There are two variants of the model: in the first (referred to as *Model* I) jobs present at time t are subject to a hazard rate determined by the state of the background process at time t, whereas in the second (referred to as *Model* II) the service times are determined by the state of the modulating process at the job's arrival epoch. Our analysis covers both cases (whereas in [1] just Model II is analyzed for $\alpha > 1$).

Organization. The organization of the rest of this paper is as follows. In Section 2, we explain the model in detail and introduce some notation. In Section 3, systems of differential equations are derived that describe the stationary and transient distribution of the number of jobs in the system for Model I. Then in Sections 4–5, we state and prove for Model I the CLT s mentioned above, for the stationary and transient distribution, respectively. Section 6 covers the corresponding results for Model II. The single-dimensional convergence can be extended to convergence of the finite-dimensional distributions (viz. at different points in time); see Section 7. The final section of the paper, Section 8, contains a discussion and concluding remarks.

2. Model description, preliminaries, and motivation

As described in the introduction, this paper studies an infinite-server queue with Markov-modulated Poisson arrivals and general service times. In full detail, the model is described as follows.

Consider an irreducible continuous-time Markov process $(J(t))_{t\in\mathbb{R}}$ on a finite state space $\{1,\dots,d\}$, with $d \in \mathbb{N}$. Its transition rate matrix is given by $Q := (q_{ij})_{i,j=1}^{d}$; the q_{ij} are nonnegative if $i \neq j$, whereas $q_{ii} = -\sum_{j\neq i} q_{ij}$. Let π_i the stationary probability that the background process is in state i, for $i = 1,\dots,d$. The time spent in state i (often referred to as the *transition time*) has an exponential distribution with mean $1/q_i$, where $q_i := -q_{ii}$. While the process $(J(t))_{t\in\mathbb{R}}$, often referred to as the *background process* or *modulating process*, is in state i, jobs arrive according to a Poisson process with rate $\lambda_i \geq 0$. The service times are assumed to be exponentially distributed with rate μ_i, but, importantly this statement can be interpreted in two ways:

- ▷ In the first variant all jobs present at a certain time instant t are subject to a hazard rate determined by the state of background chain at time t, regardless of when they arrived;
- ▷ In the second variant the service rate is determined by the background state as seen by the job upon its arrival.

The first part of this paper (Sections 3–5) focuses on the former variant, whereas in Section 6 we analyze the latter variant.

For notational convenience, we introduce the diagonal matrices Λ 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 $\boldsymbol{\pi}^{\rm T}Q = \mathbf{0}^{\rm T}$ and $\boldsymbol{\pi}^{\rm T}\mathbf{1} = 1$, where $\mathbf{0}$ and $\mathbf{1}$ denote vectors of zeros and ones, respectively. In the sequel we frequently use the 'time-average arrival rate' $\lambda_\infty := \sum_{i=1}^d \pi_i\lambda_i = \boldsymbol{\pi}^{\rm T}\Lambda\mathbf{1}$ and 'time average departure rate' $\mu_\infty := \sum_{i=1}^d \pi_i\mu_i = \boldsymbol{\pi}^{\rm T}\mathcal{M}\mathbf{1}$.

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 α. We call the resulting background process $(J^{(N)}(t))_{t\in\mathbb{R}}$, to stress the dependence on the scaling parameter N.
The main objective of this paper to derive CLTs for the number of jobs in the system, as N grows large; it turns out to matter whether α 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': with $\varrho(t) := \lim_{N\to\infty} \mathbb{E}M^{(N)}(t)/N$,

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

converges in distribution to a zero-mean Normal distribution with a certain variance, say, $\sigma^2(t)$. Here, importantly, for $\alpha > 1$ we have that the scaling parameter γ equals the usual $1/2$, while for $\alpha \leq 1$ it has the uncommon value $1 - \alpha/2$. A similar dichotomy holds for the stationary counterpart $M^{(N)}$.
In Fig. 1 we see typical sample paths of the number of jobs in the system. In the left panel the background process evolves on a substantially slower time scale than the arrival process (α close to 0), so that the number of jobs converges to a local equilibrium during each transition time. In the right panel the background process is faster than the arrival process (α substantially larger than 1), so that the process essentially behaves as a (non-modulated) M/M/∞ system.
In addition, we prove that for $\alpha > 1$ the variance $\sigma^2(t)$ equals $\varrho(t)$. The intuition here is that in this regime the background process jumps essentially faster than the arrival process, so that the arrival stream is nearly Poisson with parameter λ_∞. The resulting system is therefore, as $N \to \infty$, close to an M/M/∞, in which the transient distribution is Poissonian, thus explaining the fact that both the normalized mean and the normalized variance equal $\varrho(t)$. If $\alpha < 1$ the background process is essentially slower than the arrival process. Here the computations are substantially more complex: $\sigma^2(t)$ turns out to be a linear combination of the entries of the so-called *deviation matrix* D of the transition rate matrix Q. For a number of fundamental properties of deviation matrices we refer to e.g. the standard texts [8, 10, 13]; for a compact survey, see [5].

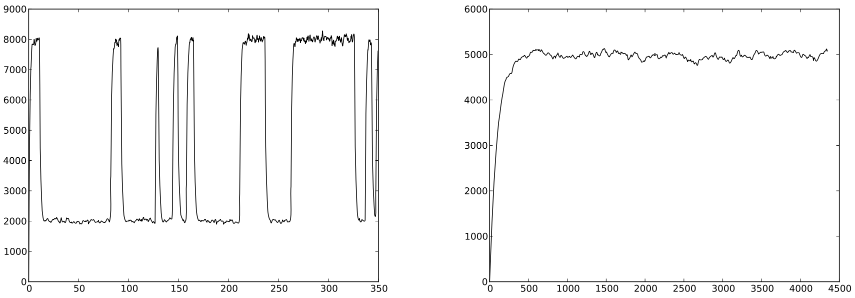

FIGURE 1. Evolution of number of jobs in the system (with $d = 2$). Left: background process is *slow* relative to arrival process; right: background process is *fast* relative to arrival process.

We illustrate the above dichotomy by determining, through an elementary computation, the asymptotic variance $\sigma^2 := \lim_{t\to\infty}\sigma^2(t)$, which reveals some of the key steps of the (considerably more elaborate) derivations later in this paper. Suppose we wish to compute $\mathbb{V}\mathrm{ar}\, M^{(N)}$. To this end, we recall the formula in [11] for the n-th factorial moment in Model I:

$$\mathbb{E}\left[M^{(N)}(M^{(N)}-1)\cdot\ldots\cdot(M^{(N)}-n+1)\right] = n!N^n\boldsymbol{\pi}^{\mathrm{T}}\Lambda X_1\Lambda X_2\Lambda\cdot\ldots\cdot X_{n-1}\Lambda X_n\mathbf{1},$$

where $X_n := (n\mathcal{M} - N^\alpha Q)^{-1}$. To keep this introductory derivation as focused as possible, we consider 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).

As a first elementary computation, we find an expression for $\mathbb{E}[M^{(N)}]$. According to the above formula, this mean equals $N\boldsymbol{\pi}^{\mathrm{T}}\Lambda X_1\mathbf{1}$. Realize that, for any $n \in \mathbb{N}$, by virtue of $Q^i\mathbf{1} = \mathbf{0}$ for $i \in \{1, 2, \ldots\}$,

$$X_n\mathbf{1} = \frac{1}{n\mu}\left(I - N^\alpha Q\frac{1}{n\mu}\right)^{-1}\mathbf{1} = \frac{1}{n\mu}\sum_{i=0}^{\infty}\left(N^\alpha Q\frac{1}{n\mu}\right)^i\mathbf{1} = \frac{1}{n\mu}\mathbf{1}.$$

It now follows that $\mathbb{E}[M^{(N)}] = N\varrho$, with $\varrho := \lambda_\infty/\mu$.

Now concentrate on the variance. By virtue of the above relation for the factorial moments,

$$\mathbb{E}\left[M^{(N)}(M^{(N)}-1)\right] = \frac{1}{\mu}N^2\boldsymbol{\pi}^{\mathrm{T}}\Lambda X_1\Lambda\mathbf{1}.$$

To evaluate this expression, we recall some concepts pertaining to the theory of deviation matrices of Markov processes; see e.g. [5]. In particular, we let $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$ denote the *ergodic matrix.* We also define the *fundamental matrix* $F := (\Pi - Q)^{-1}$ and the *deviation matrix* $D := F - \Pi$. 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 $F\mathbf{1} = \mathbf{1}$. The (i, j)-th entry of the deviation matrix can be alternatively computed as

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

with $p_{ij}(t) := \mathbb{P}(J(t) = j \mid J(0) = i)$.

Lemma 1. *We have that (i)* $X_n\Pi = (n\mu)^{-1}\Pi$ *and (ii)* $X_n = (n\mu)^{-1}\Pi + N^{-\alpha}D + O(N^{-2\alpha})$.

Proof. First note that, $n\mu X_n - N^\alpha X_n Q = I$. By postmultiplying both sides by Π, claim (i) follows immediately. Also, noting that $QD = \Pi - I$, we find

$$X_n = X_n\Pi + N^{-\alpha}(I - n\mu X_n)D = \frac{1}{n\mu}\Pi + N^{-\alpha}(I - n\mu X_n)D.$$

Iterating this recursion, we obtain

$$X_n = \frac{1}{n\mu}\Pi + N^{-\alpha}(I - n\mu X_n\Pi)D + O(N^{-2\alpha}),$$

which yields claim (ii) because of $\Pi D = 0$. □

With this lemma in place, the variance can be asymptotically analyzed in a fairly straightforward manner. Note that

$$\mathbb{E}\left[M^{(N)}(M^{(N)} - 1)\right] = \frac{1}{\mu}N^2\boldsymbol{\pi}^{\mathrm{T}}\Lambda X_1\Lambda\mathbf{1} = N^2\varrho^2 + N^{2-\alpha}\frac{1}{\mu}\boldsymbol{\pi}^{\mathrm{T}}\Lambda D\Lambda\mathbf{1} + O(N^{2-2\alpha}),$$

which leads to

$$\begin{aligned}\mathbb{V}\mathrm{ar}[M^{(N)}] &= \mathbb{E}\left[M^{(N)}(M^{(N)} - 1)\right] - \mathbb{E}\left[M^{(N)}\right]^2 + \mathbb{E}\left[M^{(N)}\right]\\ &= N^{2-\alpha}\sigma_m^2 + N\varrho + O(N^{2-2\alpha}),\end{aligned}$$

where

$$\sigma_m^2 := \frac{1}{\mu}\boldsymbol{\pi}^{\mathrm{T}}\Lambda D\Lambda\mathbf{1} = \frac{1}{\mu}\sum_{i=1}^{d}\sum_{j=1}^{d}\pi_i\lambda_i\lambda_j[D]_{ij}.$$

From this expression, we observe that for $\alpha > 1$ the variance essentially behaves as $N\varrho$ (so that we have 'Poisson-like' behavior), while for $\alpha < 1$ it grows like $N^{2-\alpha}$, with a proportionality constant that features the deviation matrix D. The objective of this paper is now to verify whether this observation for the variance (that we derived for the special case in which we assumed identical service rates μ_i) translates into fully-fledged CLT s, both under stationarity and in the transient case, for Model I as well as Model II. The above computation suggests that in case $\alpha > 1$ we have to impose the 'ordinary' $\sqrt{N}$ scaling to the centered process, while for $\alpha < 1$ it is anticipated that we have to scale by $N^{1-\alpha/2}$. We use a fairly classical approach to proving these CLT s: we show that under the appropriate scaling, the moment generating function converges to that of the Normal distribution. The general outline of the proofs in this paper is as follows. We use the following three vector-valued generating functions throughout the paper: $\boldsymbol{p}$ denotes the unscaled probability generating function (pgf); $\tilde{\boldsymbol{p}}$ denotes the corresponding moment generating function (mgf), scaled and centered appropriately for the central limit theorem at hand; and $\bar{\boldsymbol{p}}$ denotes the mgf under the law-of-large-numbers scaling. For the transient cases, these generating functions involve an extra argument t to incorporate time. All three generating functions are vector-valued (of dimension d as we consider distributions jointly with the background process $J^{(N)}(\cdot)$). Lastly, ϕ denotes the (scalar) mgf under the scaling.
Our approach consists of the following steps:

▷ We derive a differential equation for the pgf $\boldsymbol{p}$.

- ▷ We establish the weak law of large numbers under the scaling by making use of the mgf $\bar{\boldsymbol{p}}$, so as to establish the mean behavior of the underlying random variable.
- ▷ We scale and center the pgf so as to obtain a differential equation in terms of the mgf $\tilde{\boldsymbol{p}}$, which depends on the scaling parameter N. This differential equation is further manipulated (leading to expressions involving deviation matrices, which are then iterated and approximated by suitable Taylor expansions).
- ▷ By discarding asymptotically vanishing terms ($N \to \infty$), we show that the differential equation has in the limit a unique solution, viz. $\phi(\vartheta) = \exp(\vartheta^2\sigma^2)$, for some σ^2 that we explicitly identify; this corresponds to a zero-mean Normal distribution with variance σ^2. Due to Lévy's continuity theorem, pointwise convergence of characteristic functions implies convergence in distribution to a 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.

3. Model I: Stationary and transient distribution

In this section we derive systems of differential equations for the pgf of the number of jobs in the system in Model I, both for the stationary and time-dependent behavior. There is a direct relation with the results on stationary factorial moments, as presented in [11]; our results distinguish themselves from those in [11] in the sense that we uniquely characterize the pgf, and in addition our analysis also covers the transient case. For ease we consider the unscaled model (that is, $N = 1$); the differential equations can be translated easily into those for the N-scaled process introduced in Section 2.

We consider the process $(J^{(1)}(t), M^{(1)}(t))_{t\in\mathbb{R}}$ which is an ergodic Markov process on the state space $\{1, \dots, d\} \times \mathbb{N}$. With the state 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 invariant distribution $(\boldsymbol{p}_k)_{k=0}^{\infty}$, where $\boldsymbol{p}_k$ is a d-dimensional row-vector whose entries are defined by $[\boldsymbol{p}_k]_j := \mathbb{P}(M^{(1)} = k, J^{(1)} = j)$. The (row-vector-)pgf $\boldsymbol{p}(z)$ is then given by

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

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

$$\boldsymbol{p}(z)Q = (z-1)[\boldsymbol{p}'(z)\mathcal{M} - \boldsymbol{p}(z)\Lambda].$$

Proof. We immediately have that

$$\boldsymbol{p}_{k-1}\Lambda + \boldsymbol{p}_k(Q - \Lambda - k\mathcal{M}) + (k+1)\boldsymbol{p}_{k+1}\mathcal{M} = 0, \tag{1}$$

for all $k \in \mathbb{N}$, if we conveniently set $\boldsymbol{p}_{-1} = 0$. From the standard relations

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

we obtain by multiplying both sides of (1) by z^k and summing over $k \in \mathbb{N}$,

$$z\boldsymbol{p}(z)\Lambda + \boldsymbol{p}(z)(Q-\Lambda) - z\boldsymbol{p}'(z)\mathcal{M} + \boldsymbol{p}'(z)\mathcal{M} = 0.$$

The claim follows directly. □

Substituting $z = 1$ gives us: $\boldsymbol{p}(1)Q = 0$, so that $\boldsymbol{p}(1) = \boldsymbol{\pi}^{\mathrm{T}}$, as desired: the stationary distribution of the background chain in the entire Markov chain must be the same as the stationary distribution of the background chain in isolation. We can find the factorial moments of the queue content by repeated differentations and subsequently substituting $z = 1$. Our results agree with those in [11], in particular the formula for the factorial moments as mentioned in Section 2.

We now present the analogous differential equation for the transient case; the system of ordinary differential equations becomes a system of partial differential equations, as expected.

Proposition 2. *Consider Model* I. *The generating function $\boldsymbol{p}(t,z)$ 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. Let $\boldsymbol{p}_k(t) \in \mathbb{R}$ be a row-vector with entries $[\boldsymbol{p}_k(t)]_j := \mathbb{P}(M^{(1)}(t) = k, J^{(1)}(t) = j)$. By virtue of the Chapman-Kolgomorov equation, we have that

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

for all $k \in \mathbb{N}$, if we put $\boldsymbol{p}_{-1}(t) = 0$ for all $t \geq 0$. From this point on, we can follow the lines of the proof of Prop. 1. □

4. Model I: Limit results for stationary distribution

The main goal of this section is to establish the CLT for Model I in the stationary regime. The starting point of the analysis is a system of ordinary differential equations for the scaled model. Under the scaling $\Lambda \mapsto N\Lambda$ and $Q \mapsto N^{\alpha}Q$, appealing to Prop. 1, it is immediate that we have the following modified differential equation for the row-vector $\boldsymbol{p}^{(N)}(z)$:

$$\boldsymbol{p}^{(N)}(z)Q = N^{-\alpha}(z-1)[(\boldsymbol{p}^{(N)})'(z)\mathcal{M} - N\boldsymbol{p}^{(N)}(z)\Lambda]. \tag{2}$$

We postmultiply the equation in the previous display with the fundamental matrix F to obtain:

$$\boldsymbol{p}^{(N)}(z) = \boldsymbol{p}^{(N)}(z)\Pi + N^{-\alpha}(z-1)\left[N\boldsymbol{p}^{(N)}(z)\Lambda - (\boldsymbol{p}^{(N)})'(z)\mathcal{M}\right]F. \tag{3}$$

We first establish the mean number of jobs in the system in stationarity. More specifically, we prove the following claim.

Lemma 2. *Consider Model* I. $N^{-1}M^{(N)}$ *converges in probability to* $\varrho = \lambda_\infty/\mu_\infty$ *as* $N \to \infty$.

Proof. We introduce the scaled moment generating function $\bar{\boldsymbol{p}}^{(N)}(\vartheta) := \boldsymbol{p}^{(N)}(z(\vartheta))$, with $z \equiv z^{(N)}(\vartheta) = \exp(\vartheta/N)$. Evidently,

$$\frac{\mathrm{d}\bar{\boldsymbol{p}}^{(N)}(\vartheta)}{\mathrm{d}\vartheta} = \frac{\mathrm{d}\boldsymbol{p}^{(N)}(z)}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}\vartheta} = \frac{1}{N}\cdot z\,\frac{\mathrm{d}\boldsymbol{p}^{(N)}(z)}{\mathrm{d}z}.$$

Substituting these expressions in Eqn. (3), and noting that $z = 1 + \vartheta N^{-1} + O(N^{-2})$, we obtain

$$\bar{\boldsymbol{p}}^{(N)}(\vartheta) = \bar{\boldsymbol{p}}^{(N)}(\vartheta)\Pi + N^{-\alpha}\vartheta\left[\bar{\boldsymbol{p}}^{(N)}(\vartheta)\Lambda - (\bar{\boldsymbol{p}}^{(N)})'(\vartheta)\mathcal{M}\right]F + o(N^{-\alpha}).$$

Note that $\bar{\boldsymbol{p}}^{(N)}(\vartheta) = \bar{\boldsymbol{p}}^{(N)}(\vartheta)\Pi + O(N^{-\alpha})$ and $(\bar{\boldsymbol{p}}^{(N)})'(\vartheta) = (\bar{\boldsymbol{p}}^{(N)})'(\vartheta)\Pi + O(N^{-\alpha})$, so that by postmultiplying the previous display by $\mathbf{1}$ we obtain

$$0 = N^{-\alpha}\vartheta\left[\bar{\boldsymbol{p}}^{(N)}(\vartheta)\mathbf{1}\lambda_\infty - (\bar{\boldsymbol{p}}^{(N)})'(\vartheta)\mathbf{1}\mu_\infty\right] + o(N^{-\alpha}),$$

recalling that $F\mathbf{1} = \mathbf{1}$. We thus find a differential equation in $\phi^{(N)}(\vartheta) := \bar{\boldsymbol{p}}^{(N)}(\vartheta)\mathbf{1}$. Multiplying by N^α and sending N to ∞, we obtain a limiting differential equation with the solution

$$\phi(\vartheta) = K\exp\left(\frac{\lambda_\infty}{\mu_\infty}\vartheta\right);$$

here K is an integration constant, which equals 1 due to the fact that $\phi(\vartheta)$ is a mgf. We have thus found the mgf of the constant ϱ. By Lévy's continuity theorem, we have convergence in distribution of $N^{-1}M^{(N)}$ to ϱ, but convergence in probability to a constant is implied by convergence in distribution to the same constant. This completes the proof. □

Now we know that $N^{-1}M^{(N)}$ can be centered by subtracting ϱ. In the next result we identify the right scaling such that the centered random variable obeys a CLT. We explicitly identify the corresponding variance.

Theorem 1. *Consider Model* I. *The random variable*

$$\frac{M^{(N)} - N\varrho}{N^{1-\beta/2}}$$

converges to a Normal distribution with zero mean and variance σ^2 *as* $N \to \infty$*; here the scaling parameter* β *equals* $\min\{\alpha, 1\}$, *and* $\sigma^2 := \sigma_m^2 1_{\{\alpha\le 1\}} + \varrho 1_{\{\alpha\ge 1\}}$, *with* $\sigma_m^2 := \mu_\infty^{-1}\boldsymbol{\pi}^{\mathrm{T}}(\Lambda - \varrho\mathcal{M})D(\Lambda - \varrho\mathcal{M})\mathbf{1}$.

Proof. The proof strategy is as outlined at the end of Section 2. As a first step, we introduce the centered and scaled mgf $\tilde{\boldsymbol{p}}^{(N)}(\vartheta)$. We perform a change of variables in Eqn. (3) so as to obtain a differential equation in $\tilde{\boldsymbol{p}}^{(N)}(\vartheta)$. Note that

$$\tilde{\boldsymbol{p}}^{(N)}(\vartheta) = \exp(-\varrho\vartheta N^{\beta/2})\cdot\boldsymbol{p}^{(N)}\left(\exp(\vartheta N^{-1+\beta/2})\right),$$

which can be written as

$$\boldsymbol{p}^{(N)}(z) = \exp(\varrho\vartheta N^{\beta/2})\cdot\tilde{\boldsymbol{p}}^{(N)}(\vartheta),$$

where $z \equiv z^{(N)}(\vartheta) = \exp(\vartheta N^{-1+\beta/2})$. Consider the Taylor expansions of z and z^{-1}:

$$z^{\pm 1} \quad = \quad 1 \pm \vartheta N^{-1+\beta/2} + \frac{1}{2}\vartheta^2 N^{-2+\beta} + O(N^{-3+3\beta/2}).$$

It is readily verified that

$$\frac{\mathrm{d}\boldsymbol{p}^{(N)}(z)}{\mathrm{d}z}\frac{\mathrm{d}z}{\mathrm{d}\vartheta} = \exp(\varrho\vartheta N^{\beta/2})\left(\varrho N^{\beta/2}\tilde{\boldsymbol{p}}^{(N)}(\vartheta) + \frac{\mathrm{d}\tilde{\boldsymbol{p}}^{(N)}(\vartheta)}{\mathrm{d}\vartheta}\right),$$

$$\frac{\mathrm{d}z}{\mathrm{d}\vartheta} = N^{-1+\beta/2}\exp(\vartheta N^{1-\beta/2}) = N^{-1+\beta/2}z.$$

Upon combining the above, we conclude that

$$\frac{\mathrm{d}\boldsymbol{p}^{(N)}(z)}{\mathrm{d}z} = \frac{1}{z}\cdot\exp(\varrho\vartheta N^{\beta/2})\left(N\varrho\,\tilde{\boldsymbol{p}}^{(N)}(\vartheta) + N^{1-\beta/2}\frac{\mathrm{d}\tilde{\boldsymbol{p}}^{(N)}(\vartheta)}{\mathrm{d}\vartheta}\right).$$

Now perform the change of variables, and substitute the expressions for $\boldsymbol{p}^{(N)}(z)$ and $(\boldsymbol{p}^{(N)})'(z)$ into Eqn. (3), yielding

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(\vartheta) \quad &= \quad \tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Pi + N^{1-\alpha}\left(z^{(N)}(\vartheta) - 1\right)\tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Lambda F \\ &\quad - N^{1-\alpha}\left(1 - \frac{1}{z^{(N)}(\vartheta)}\right)\varrho\,\tilde{\boldsymbol{p}}^{(N)}(\vartheta)\mathcal{M}F \\ &\quad - N^{1-\alpha-\beta/2}\left(1 - \frac{1}{z^{(N)}(\vartheta)}\right)(\tilde{\boldsymbol{p}}^{(N)})'(\vartheta)\mathcal{M}F. \end{aligned} \tag{4}$$

The next step is to apply the Taylor expansions for z and z^{-1}, as given above. We assume that $\beta \leq 1$ and $\beta \leq \alpha$, as is consistent with the proof statement. By deleting every term that has a provably smaller order than $N^{-\alpha}$, we obtain

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(\vartheta) \quad &= \quad \tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Pi + \vartheta N^{\beta/2-\alpha}\tilde{\boldsymbol{p}}^{(N)}(\vartheta)(\Lambda - \varrho\mathcal{M})F \\ &\quad + \frac{\vartheta^2 N^{\beta-1-\alpha}}{2}\tilde{\boldsymbol{p}}^{(N)}(\vartheta)(\Lambda + \varrho\mathcal{M})F - \vartheta N^{-\alpha}(\tilde{\boldsymbol{p}}^{(N)})'(\vartheta)\mathcal{M}F, \end{aligned} \tag{5}$$

with an error term that is $o(N^{-\alpha})$. It takes some careful but elementary steps to verify that this is indeed justified, where the restrictions imposed on β are intensively used:

- ▷ The third-order Taylor term for the second and third term of the right-hand side of Eqn. (4) has order $N^{1-\alpha-3+3\beta/2}$, which is indeed smaller than $N^{-\alpha}$.
- ▷ The fourth term has as a second order Taylor term with degree $N^{-1-\alpha+\beta/2}$, being smaller than $N^{-\alpha}$ as well.

Our goal is to transform the coupled system of ordinary differential equations in $\tilde{\boldsymbol{p}}^{(N)}$ into a single-dimensional ordinary differential equation in terms of $\phi^{(N)}(\vartheta) := \tilde{\boldsymbol{p}}^{(N)}(\vartheta)\mathbf{1}$. This can be done as follows. First iterate Eqn. (5) until all terms in the right-hand side either contain $\tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Pi$ or are of $O(N^{-\alpha})$ using the restrictions on β. In the latter terms $\tilde{\boldsymbol{p}}^{(N)}$ or $(\tilde{\boldsymbol{p}}^{(N)})'$ can be replaced by $\tilde{\boldsymbol{p}}^{(N)}\Pi$ or $(\tilde{\boldsymbol{p}}^{(N)})'\Pi$, respectively, since $\tilde{\boldsymbol{p}}^{(N)} = \tilde{\boldsymbol{p}}^{(N)}\Pi + o(1)$.

This yields

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(\vartheta) &= \tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Pi + \vartheta N^{\beta/2-\alpha}\tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Pi(\Lambda - \varrho\mathcal{M})F \\ &\quad + \vartheta^2 N^{\beta-2\alpha}\tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Pi(\Lambda - \varrho\mathcal{M})F(\Lambda - \varrho\mathcal{M})F \\ &\quad + \frac{\vartheta^2 N^{\beta-1-\alpha}}{2}\tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Pi(\Lambda + \varrho\mathcal{M})F - \vartheta N^{-\alpha}(\tilde{\boldsymbol{p}}^{(N)})'(\vartheta)\Pi\mathcal{M}F,\end{aligned}$$

with an error term that is $o(N^{-\alpha})$. Now postmultiply the resulting identity by $\mathbf{1}\,N^{\alpha}/\vartheta$; realize that $\Pi\mathbf{1} = \mathbf{1}$ and $F\mathbf{1} = \mathbf{1}$. Observe that, from the definition of ϱ,

$$\tilde{\boldsymbol{p}}^{(N)}(\vartheta)\Pi(\Lambda - \varrho\mathcal{M})F\mathbf{1} = \phi^{(N)}(\vartheta)\,\boldsymbol{\pi}^{\mathrm{T}}(\Lambda - \varrho\mathcal{M})\mathbf{1} = 0.$$

We thus obtain

$$(\phi^{(N)})'(\vartheta) = \vartheta N^{\beta-\alpha}\phi^{(N)}(\vartheta)\frac{\boldsymbol{\pi}^{\mathrm{T}}(\Lambda - \varrho\mathcal{M})F(\Lambda - \varrho\mathcal{M})\mathbf{1}}{\mu_\infty} + \vartheta N^{\beta-1}\varrho\phi^{(N)}(\vartheta) + o(1),$$

using $(\tilde{\boldsymbol{p}}^{(N)})'(\vartheta)\Pi\mathcal{M}F\mathbf{1} = (\phi^{(N)})'(\vartheta)\mu_\infty$ and $\boldsymbol{\pi}^{\mathrm{T}}(\Lambda + \varrho\mathcal{M})\mathbf{1} = 2\lambda_\infty$. First, note that if we choose β smaller than both α and 1, we do not obtain a CLT, but rather that the random variable under study converges in distribution to the constant 0. Hence, we take $\beta = \min\{\alpha, 1\}$, in which case the largest term dominates, with both terms contributing if $\alpha = 1$. We find that $\phi^{(N)}(\vartheta)$ converges, as $N \to \infty$, to

$$\phi(\vartheta) = \frac{1}{2}\sigma^2\vartheta^2,$$

where $\sigma^2 := \sigma_m^2 1_{\{\alpha\leq 1\}} + \varrho 1_{\{\alpha\geq 1\}}$, with

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

(where the rightmost equality in the previous display follows from $F = D + \Pi = D + \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$, in conjunction with the definition of ϱ). We recognize the mgf of a centered Normally distributed random variable. We have thus established the claim. □

From Thm. 1 we conclude that the variance σ^2 equals ϱ for $\alpha > 1$, in agreement with the intuition presented earlier: the system essentially behaves as a normal M/M/∞ system, with mean and variance roughly equalling $N\varrho$. If $\alpha < 1$ the timescale of the background process is relatively slow, so that the variance of $M^{(N)}$ is more than linear. In addition we note that if $\alpha = 1$ *both* terms appear in σ^2: we then have that $\sigma^2 = \sigma_m^2 + \varrho$.
In case $\mu_i = \mu$ for all $i \in \{1, \ldots, d\}$, we find, using $D\mathbf{1} = \mathbf{0}$ and $\boldsymbol{\pi}^{\mathrm{T}}D = \mathbf{0}$,

$$\boldsymbol{\pi}^{\mathrm{T}}(\Lambda - \varrho\mathcal{M})D(\Lambda - \varrho\mathcal{M})\mathbf{1} = \boldsymbol{\pi}^{\mathrm{T}}\Lambda D\Lambda\mathbf{1},$$

so that $\sigma_m^2 = \mu^{-1}\boldsymbol{\pi}^{\mathrm{T}}\Lambda D\Lambda\mathbf{1}$, in agreement with the findings presented in Section 2.

5. Model I: Limit results for transient distribution

We consider the transient CLT for Model I, which turns out harder to prove than its stationary counterpart. However, as the proof follows essentially the same lines as the stationary CLT, we provide an appropriately abridged derivation in this section. 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. 2 implies that we have the following system of partial differential equations:

$$\frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial t} = N^\alpha \boldsymbol{p}^{(N)}(t,z)\, Q + (z-1)\left(N\boldsymbol{p}^{(N)}(t,z)\Lambda - \frac{\partial \boldsymbol{p}^{(N)}(t,z)}{\partial z}\mathcal{M}\right). \tag{6}$$

As before, this equation is postmultiplied with the fundamental matrix F and divided by N^α:

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

Notice that this differential equation is basically the same as in the previous section, except for the last term (containing the partial derivative with respect to t). To make the proofs compact, we primarily concentrate on this new term. We start by analyzing the mean number in the system at time t; recall that $\varrho := \lambda_\infty/\mu_\infty$.

Lemma 3. *Consider Model* I. *$N^{-1}M^{(N)}(t)$ converges in probability to $\varrho(t) = \varrho\,(1-e^{-\mu_\infty t})$ as $N\to\infty$.*

Proof. We follow the lines of the proof of Lemma 2 closely, and introduce the transient scaled moment generating function $\bar{\boldsymbol{p}}^{(N)}(t,\vartheta)$:

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

The expressions for z and the derivative with respect to ϑ do not change, except for the fact that the ordinary derivative becomes a partial derivative. The partial derivative with respect to t is simply $\partial\bar{\boldsymbol{p}}^{(N)}/\partial t = \partial\boldsymbol{p}^{(N)}/\partial t$. In the same way as in the stationary case, we obtain

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

Analogously to Lemma 2, postmultiplying the previous display by $\mathbf{1}$ yields

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

We thus obtain a partial differential equation in $\bar{\boldsymbol{p}}^{(N)}(t,\vartheta)\mathbf{1}$; 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^α and let $N\to\infty$. It is straightforward to check that $\bar{\boldsymbol{p}}(t,\vartheta)\mathbf{1} = \exp(\vartheta\varrho\,(1-\exp(-\mu_\infty t)))$ 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. □

Now that we have derived the weak law of large numbers, we proceed with stating and proving the corresponding CLT.

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

$$\frac{M^{(N)}(t) - N\varrho(t)}{N^{1-\beta/2}}$$

converges to a Normal distribution with zero mean and variance $\sigma^2(t)$ *as* $N \to \infty$*; here the scaling parameter* β *equals* $\min\{\alpha, 1\}$, *and* $\sigma^2(t) := \sigma^2_m(t)1_{\{\alpha\leq 1\}} + \varrho(t)1_{\{\alpha\geq 1\}}$, *with*

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

Proof. The structure of the proof follows that of the stationary case, but we finally obtain a *partial* (rather than an ordinary) differential equation. As in the proof of Lemma 3, we concentrate on the new term $-N^{-\alpha}\,\partial\boldsymbol{p}^{(N)}(t,z)/\partial t\, F$. We introduce the mgf $\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)$, scaled and centered as in the claim. Note that

$$\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta) = \exp(-\varrho(t)\vartheta N^{\beta/2})\,\boldsymbol{p}^{(N)}\left(t, \exp(\vartheta N^{-1+\beta/2})\right),$$

The expressions for z and the derivative with respect to ϑ are the same as in the stationary case, except for the type of derivative (ordinary versus partial), while ϱ should obviously be replaced by $\varrho(t)$. Also,

$$\frac{\partial\boldsymbol{p}^{(N)}(t,\exp(\vartheta N^{-1+\beta/2}))}{\partial t} = \exp(\varrho(t)\vartheta N^{\beta/2})\left(\varrho'(t)\vartheta N^{\beta/2}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta) + \frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t}\right).$$

As there is no 'Tayloring' required in this term, we can fastforward to the equivalent of differential equation (5):

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta) &= \tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\Pi + \vartheta N^{\beta/2-\alpha}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)(\Lambda - \varrho(t)\mathcal{M} - \varrho'(t))F\\ &\quad + \frac{\vartheta^2 N^{\beta-1-\alpha}}{2}\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)(\Lambda + \varrho(t)\mathcal{M})F\\ &\quad - \vartheta N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial\vartheta}\mathcal{M}F - N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)}{\partial t}F + o(N^{-\alpha}).\end{aligned}\tag{7}$$

Iterating (7), we find

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

The next step is to postmultiply this equation by $\mathbf{1}\,N^{\alpha}$, so as to obtain a differential equation in terms of $\phi^{(N)}(t,\vartheta) = \tilde{\boldsymbol{p}}^{(N)}(t,\vartheta)\mathbf{1}$; as in the stationary case, various terms cancel. We eventually find,

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

up to an $o(1)$ error term. It is immediately seen from the definition of $\varrho(t)$ that the first term on the right-hand side vanishes. In addition, it takes some elementary algebra to check that

$$\boldsymbol{\pi}^{\mathrm{T}}(\Lambda - \varrho(t)\mathcal{M} - \varrho'(t))F(\Lambda - \varrho(t)\mathcal{M} - \varrho'(t))\mathbf{1} = \boldsymbol{\pi}^{\mathrm{T}}(\Lambda - \varrho(t)\mathcal{M})D(\Lambda - \varrho(t)\mathcal{M})\mathbf{1}$$

and

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

Let β be $\min\{\alpha, 1\}$, we obtain by sending $N \to \infty$, with $\phi(t, \vartheta) := \lim_{N\to\infty} \phi^{(N)}(t, \vartheta)$,

$$\frac{\partial \phi(t, \vartheta)}{\partial t} + \vartheta\mu_\infty \frac{\partial \phi(t, \vartheta)}{\partial \vartheta} = \vartheta^2 \phi(t, \vartheta)\, g(t).$$

with $g(t) = (\boldsymbol{\pi}^{\mathrm{T}}(\Lambda - \varrho(t)\mathcal{M})D(\Lambda - \varrho(t)\mathcal{M})\mathbf{1})\, 1_{\{\alpha\leq 1\}} + (\lambda_\infty(1 - e^{-\mu_\infty t}/2))\, 1_{\{\alpha \geq 1\}}$.
We propose the *ansatz*

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

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

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

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 as given in the statement of the theorem. □

As an aside, we mention that we can explicitly compute the integral in the definition of $\sigma^2(t)$. After some elementary manipulations we find that

$$\sigma^2(t) = e^{-2\mu_\infty t} f(t) = \left(\sigma_m^2 + ve^{-\mu_\infty t} + (-\sigma_m^2 - v + wt)e^{-2\mu_\infty t}\right) 1_{\{\alpha\leq 1\}} + \varrho(t)1_{\{\alpha\geq 1\}},$$

where v is given by $2\mu_\infty^{-1}\varrho(\boldsymbol{\pi}^{\mathrm{T}}\mathcal{M}D(\Lambda - \rho\mathcal{M})\mathbf{1} + \boldsymbol{\pi}^{\mathrm{T}}(\Lambda - \rho\mathcal{M})D\mathcal{M}\mathbf{1})$, whereas w denotes $2\varrho^2\boldsymbol{\pi}^{\mathrm{T}}\mathcal{M}D\mathcal{M}\mathbf{1}$. As $t \to \infty$, we have that $\sigma_m^2(t) \to \sigma_m^2$, as expected.

6. Results for Model II

In this section we study Model II: service times are determined by the background state as seen by the jobs upon arrival. The approach is as before: we first derive a system of differential equations, and then we manipulate these under the centering and scaling considered. This results in a single differential equation, that immediately yields the desired CLT.

For the transient distribution, a system of differential equations was previously derived in [1]. It is based on the distributional identity, cf. [1, 4],

$$M^{(N)}(t) \overset{d}{=} P^{(N)}\left(\varphi\left(J^{(N)}\right)\right),$$

where $P^{(N)}(\lambda)$ is a Poisson random variable with mean $N\lambda$, and

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

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 $\boldsymbol{p}(t,z)$:

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

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 $[\Delta(t)]_{ii} := \lambda_i \exp(-\mu_i t)$. (It is noted that the definition of $\boldsymbol{p}$ is slightly different from the one used in [1]. 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 [1, 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 different equation, but it is easy to translate them into each other.)

Naïvely, one could try to obtain a differential equation for the stationary behavior by sending $t \to \infty$, 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 ∞; 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 a description with a slightly more general state space: we keep track of the number 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

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

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 equal to

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

respectively.

As before, we first derive a differential equation for the unscaled model (that is, $N = 1$). The generating function $\boldsymbol{p}(t,\boldsymbol{z})$ is defined as follows:

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

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 generating function $\boldsymbol{p}(t,\boldsymbol{z})$ 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 $\boldsymbol{p}(z_1,\dots,z_d)$ defined in the obvious way, the differential equation for the stationary case is the following.

Proposition 4. *Consider Model* II. *The generating function $\boldsymbol{p}(\boldsymbol{z})$ 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 differential equations for the scaled model follow directly from the above propositions, by replacing λ_k by $N\lambda_k$, and Q by $N^\alpha Q$. Later on, it turns out to be convenient to rewrite the resulting differential equation in terms of the fundamental matrix F:

$$\begin{aligned}\boldsymbol{p}^{(N)}(t,\boldsymbol{z}) &= \boldsymbol{p}^{(N)}(t,\boldsymbol{z})\Pi \\ &+N^{-\alpha}\sum_{k=1}^{d}(z_k-1)\left(N\lambda_k\,\boldsymbol{p}^{(N)}(t,\boldsymbol{z})\,E_k - \mu_k\frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial z_k}\right)F - N^{-\alpha}\frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial t}F \end{aligned} \tag{8}$$

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

We now present 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.

As before, we first derive the law of large numbers. The following lemma covers both the transient and stationary cases. Define $\varrho_k(t) := \pi_k\lambda_k/\mu_k\,(1-e^{-\mu_k t})$ and $\varrho_k := \pi_k\lambda_k/\mu_k$. We (re-)define $\varrho(t) := \sum_k \varrho_k(t)$, and $\varrho := \sum_k \varrho_k$.

Lemma 4. *Consider Model* II. *$N^{-1}\boldsymbol{M}^{(N)}(t)$ converges in probability to $\boldsymbol{\varrho}(t)$ as $N\to\infty$. Moreover, $N^{-1}\boldsymbol{M}^{(N)}$ converges in probability to $\boldsymbol{\varrho}$ as $N\to\infty$. Lastly, $N^{-1}M^{(N)}(t)$ converges in probability to $\varrho(t)$, and $N^{-1}M^{(N)}$ to ϱ as $N\to\infty$.*

Proof. Similarly to previous proofs, we first introduce the scaled moment generating function $\bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) := \boldsymbol{p}^{(N)}(t,\boldsymbol{z})$, with $z_k \equiv z_k^{(N)}(\vartheta_k) = \exp(\vartheta_k/N)$. 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. (8), and note that $z_k = 1 + \vartheta_k N^{-1} + O(N^{-2})$. As a consequence,

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

It now 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 by introducing $\phi^{(N)}(t,\boldsymbol{\vartheta}) := \bar{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\mathbf{1}$, we derive after some elementary steps

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

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

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

for specific functions $\bar{\varrho}_k(\cdot)$. Plugging in 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 ϑ_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 ($\bar{\varrho}_k(t) = \varrho_k(t)$, that is). We conclude that we have established the claim for the transient case: $N^{-1}\boldsymbol{M}^{(N)}(t)$ converges in probability to $\boldsymbol{\varrho}(t)$ 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 $\phi(\boldsymbol{\vartheta})$ as follows:

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

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

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

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

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

$$[C(t)]_{jk} := [V(t)]_{jk} 1_{\{\alpha\leq 1\}} + \varrho_j(t) 1_{\{\alpha\geq 1\}} 1_{\{j=k\}}.$$

It is noted that the matrix $\bar{D}$ is invariant under time-reversal of the background Markov chain, and hence the following CLT is also invariant under such time-reversal.

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

$$\frac{\boldsymbol{M}^{(N)} - N\boldsymbol{\varrho}}{N^{1-\beta/2}}$$

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

$$\frac{\boldsymbol{M}^{(N)}(t) - N\boldsymbol{\varrho}(t)}{N^{1-\beta/2}}$$

converges to a d-dimensional Normal distribution with zero mean and covariance matrix $C(t)$ as $N \to \infty$. In both cases the scaling parameter β equals $\min\{\alpha, 1\}$.

Proof. Define $\boldsymbol{z}$ by $z_k \equiv z_k^{(N)}(\vartheta_k) := \exp(\vartheta_k N^{-1+\beta/2})$. We first concentrate on the transient case and introduce the centered and scaled mgf $\tilde{\boldsymbol{p}}(t,\boldsymbol{\vartheta})$:

$$\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) = \exp\left(-N^{\beta/2}\sum_{k=1}^{d}\vartheta_k\varrho_k(t)\right)\boldsymbol{p}^{(N)}(t,\boldsymbol{z}).$$

We wish to perform a change of variables in Eqn. (8) to obtain a differential equation in $\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})$. The second order Taylor expansions of z_k and z_k^{-1} are given by

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

Mimicking the proof of the CLTs in Model I, we note that

$$\frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial z_k}\frac{\mathrm{d}z_k}{\mathrm{d}\vartheta_k} = \exp\left(N^{\beta/2}\sum_{k=1}^{d}\vartheta_k\varrho_k(t)\right)\left(\varrho_k(t)N^{\beta/2}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) + \frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial\vartheta_k}\right),$$

where

$$\frac{\mathrm{d}z_k}{\mathrm{d}\vartheta_k} = N^{-1+\beta/2}\exp(\vartheta_k N^{-1+\beta/2}) = N^{-1+\beta/2}z_k.$$

Also,

$$\frac{\partial \boldsymbol{p}^{(N)}(t,\boldsymbol{z})}{\partial t} = \exp\left(N^{\beta/2}\sum_{k=1}^{d}\vartheta_k\varrho_k(t)\right)\left(\sum_k \vartheta_k\varrho_k'(t)N^{\beta/2}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) + \frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t}\right).$$

Now perform the change of variables, and substitute the expressions for the partial derivatives of $\boldsymbol{p}^{(N)}(t,z)$ into Eqn. (8):

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) &= \tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\Pi + N^{-\alpha}\sum_{k=1}^{d}(z_k-1)N\lambda_k\,\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\,E_k\,F\\ &\quad -N^{-\alpha}\sum_{k=1}^{d}\left(1-\frac{1}{z_k}\right)N^{1-\beta/2}\mu_k\left(N^{\beta/2}\varrho_k(t)\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})+\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial\vartheta_k}\right)F\\ &\quad -N^{-\alpha+\beta/2}\sum_{k=1}^{d}\vartheta_k\varrho_k'(t)\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})F - N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t}F.\end{aligned}$$

The next step is to introduce the Taylor expansions for z_k and z_k^{-1}, assuming that $\beta\leq 1$ and $\beta\leq\alpha$, in line with the proof statement. Ignoring all terms that are provably smaller than $N^{-\alpha}$, and combining terms of the same order, we obtain

$$\begin{aligned}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta}) &= \tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\,\Pi + N^{\beta/2-\alpha}\sum_{k=1}^{d}\vartheta_k\,\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\,\big(\lambda_kE_k-\mu_k\varrho_k(t)I-\varrho_k'(t)I\big)\,F\\ &\quad +N^{\beta-1-\alpha}\sum_{k=1}^{d}\frac{\vartheta_k^2}{2}\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})\,(\lambda_kE_k+\mu_k\varrho_k(t)I)\,F\\ &\quad -N^{-\alpha}\sum_{k=1}^{d}\vartheta_k\mu_k\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial\vartheta_k}F - N^{-\alpha}\frac{\partial\tilde{\boldsymbol{p}}^{(N)}(t,\boldsymbol{\vartheta})}{\partial t}F,\end{aligned}$$

up to an error term that is $o(N^{-\alpha})$. As we did in the proofs of the CLTs corresponding to Model I, we iterate this relation, 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 $\phi^{(N)}(t,\boldsymbol{\vartheta})$:

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

as we note that

$$\begin{aligned}&\boldsymbol{\pi}^{\mathrm{T}}\left(\sum_{j=1}^{d}\vartheta_j(\lambda_jE_j-\varrho_j(t)\mu_jI-\varrho_j'(t)I)\right)F\left(\sum_{k=1}^{d}\vartheta_k(\lambda_kE_k-\varrho_k(t)\mu_kI-\varrho_k'(t)I)\right)\mathbf{1}\\ &\qquad= \sum_{j=1}^{d}\sum_{k=1}^{d}\vartheta_j\vartheta_k\lambda_j\lambda_k\left(\boldsymbol{\pi}^{\mathrm{T}}E_jDE_k\mathbf{1}\right)=\frac{1}{2}\sum_{j=1}^{d}\sum_{k=1}^{d}\vartheta_j\vartheta_k\lambda_j\lambda_k[\bar{D}]_{jk}.\end{aligned}$$

Pick, as before, $\beta = \min\{1, \alpha\}$, to obtain the following partial differential equation, by sending $N \to \infty$,

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

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

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

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

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

This completes the proof. □

Corollary 1. *Consider Model* II. *An immediate consequence of Thm. 3 is that, with β as defined before, the random variables*

$$\frac{M^{(N)} - N\varrho}{N^{1-\beta/2}} \quad \textit{and} \quad \frac{M^{(N)}(t) - N\varrho(t)}{N^{1-\beta/2}}$$

converge to Normal distributions with zero mean and variances

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

respectively, as $N \to \infty$.

7. 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 μ), so that Model I and Model II coincide.

7.1. **Differential equation.** We follow the line of reasoning of [1, 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 $0 \equiv s_1 \le s_2 \le \cdots \le s_K$ for some $K \in \mathbb{N}$. Our goal is to characterize the joint transform, for $j = 1, \ldots, d$,

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

Assume a job arrives between 0 and Δt, for an infinitesimally small Δt. 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 $a_j(t,\boldsymbol{z}):=\lambda_j\,(b(t,\boldsymbol{z})-1)$. Now letting $\Delta t\downarrow 0$, and defining $A(t,\boldsymbol{z}):=\text{diag}\{\boldsymbol{a}(t,\boldsymbol{z})\}$, we obtain the differential equation, in vector notation,

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

7.2. **Covariance.** Let us now analyze $\mathbb{C}\text{ov}(M^{(1)}(t),M^{(1)}(t+s))$; for ease we assume here that the background process is in equilibrium at time 0. An explicit formula for $m_j(t):=\mathbb{E}_jM^{(1)}(t)$, where the subscript indicates that we condition on $\{J^{(1)}=j\}$, is already known [1]; with $\boldsymbol{\Lambda}(t):=e^{-\mu t}\Lambda\boldsymbol{1}$, it solves the nonhomogeneous system of linear differential equations $\boldsymbol{m}'(t)=\Lambda(t)+Q\,\boldsymbol{m}(t)$. It is immediate that we have, bearing in mind that $\boldsymbol{m}(0)=\boldsymbol{0}$,

$$\boldsymbol{m}(t)=e^{Qt}\int_0^t e^{-Qs}\boldsymbol{\Lambda}(s)\mathrm{d}s.$$

It is readily seen that this entails (use $\boldsymbol{\pi}^{\mathrm{T}}Q=\boldsymbol{0}^{\mathrm{T}}$)

$$\mathbb{E}M^{(1)}(t)=\boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{m}(t)=\sum_{i=1}^d\frac{\pi_i\lambda_i}{\mu}(1-e^{-\mu t}).$$

We now concentrate on computing $C_j(t):=\mathbb{E}(M^{(1)}(t)M^{(1)}(t+s)\,|\,J^{(1)}(0)=j)$ for given $s\geq 0$; realize that $C_j(0)=0$. As before, we set up a system of differential equations. We obtain, up to $O(\Delta t)$-terms, with $t_\Delta:=t-\Delta t$,

$$\begin{aligned}C_j(t)= \quad &\lambda_j\Delta t\left(\begin{array}{l}(1-e^{-\mu t})\,C_j(t_\Delta)+(e^{-\mu t}-e^{-\mu(t+s)})\,(m_j(t_\Delta+s)+C_j(t_\Delta))\\ +\,e^{-\mu(t+s)}\,(1+m_j(t_\Delta)+m_j(t_\Delta+s)+C_j(t_\Delta))\end{array}\right)\\ &\qquad+\sum_{i\neq j}q_{ji}\Delta t\;\;C_i(t_\Delta)+\left(1-\lambda_j\Delta t-\sum_{i\neq j}q_{ji}\Delta t\right)C_j(t_\Delta).\end{aligned}$$

Subtracting $C_j(t_\Delta)$ from both sides, dividing by Δt and letting $\Delta t \downarrow 0$, this directly leads to the nonhomogeneous system of linear differential equations

$$\boldsymbol{C}'(t) = e^{-\mu(t+s)}\Lambda(\boldsymbol{1}+\boldsymbol{m}(t)) + e^{-\mu t}\Lambda\boldsymbol{m}(t+s) + Q\boldsymbol{C}(t),$$

which is solved by

$$\boldsymbol{C}(t) = e^{Qt}\int_0^t e^{-Qu}\left(e^{-\mu(u+s)}\Lambda(\boldsymbol{1}+\boldsymbol{m}(u)) + e^{-\mu u}\Lambda\boldsymbol{m}(u+s)\right)\mathrm{d}u.$$

7.3. **Limit results.** We again consider the situation in which the modulating Markov chain $J(\cdot)$ is sped up by a factor N^α (for some positive α), while the arrival rates λ_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 [1] we just covered the case of $\alpha > 1$, we now establish a CLT for general α.

As the techniques used are precisely the same as before, we just state the result. We first introduce some notation. Define $U := \sum_{i=1}^d \sum_{j=1}^d \lambda_i\lambda_j[\bar{D}]_{ij}$. In addition, $[\check{C}(t)]_{k\ell} = [\check{C}(t)]_{\ell k}$, where for $k \geq \ell$

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

Theorem 4. *The random vector*

$$\left(\frac{M^{(N)}(t+s_1)-N\varrho(t+s_1)}{N^{1-\beta/2}},\ldots,\frac{M^{(N)}(t+s_K)-N\varrho(t+s_K)}{N^{1-\beta/2}}\right)$$

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

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

$$[\check{C}]_{k\ell} = \frac{u_{k\ell}}{2\mu},\quad \text{with}\quad u_{k\ell} := \left(U1_{\{\alpha\leq 1\}} + 2\lambda_\infty 1_{\{\alpha\geq 1\}}\right)e^{-\mu(s_k-s_\ell)}.$$

We observe that the limiting 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

$$\mathrm{d}S(t) = (\lambda_\infty-\mu)S(t)\mathrm{d}t + \left(U1_{\{\alpha\leq 1\}} + 2\lambda_\infty 1_{\{\alpha\geq 1\}}\right)\mathrm{d}W(t),$$

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

8. Discussion and conclusion

In this paper we derived central limit theorems (CLTs) for infinite-server queues with Markov-modulated input. In our approach the modulating Markov chain is sped up by a factor N^α (for some positive α), while the arrival process is sped up by N. Interestingly, there is a *phase transition* in the sense that the scaling to be used in the CLT depends on the value of α: rather than the standard normalization by $\sqrt{N}$, it turned out that the centered process should be divided by $N^{1-\beta/2}$, with β equal to $\min\{\alpha,1\}$. 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 [1]. We anticipate weak 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 [3]. 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 PDE s are of a similar form and moreover quasi-linear, we can suffice by establishing uniqueness for the two types of PDE s, the first of which is as follows:

$$\sum_k \mu_k \vartheta_k \frac{\partial \phi}{\partial \vartheta_k} = g(\boldsymbol{\vartheta})\, \phi(\vartheta_1, \dots, \vartheta_d),$$

for some function $g(\cdot)$ with boundary condition $\phi(0, \dots, 0) = 1$. This pertains to differential equations in the proofs of Lemma 4 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 (ODE s):

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

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

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

while the ODE for ϕ 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 $(\vartheta_1, \dots, \vartheta_d)$, 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:

$$\frac{\partial \phi}{\partial t} + \sum_k \mu_k \vartheta_k \frac{\partial \phi}{\partial \vartheta_k} = g(t, \vartheta)\, \phi(t, \vartheta_1, \dots, \vartheta_d),$$

with the boundary condition $\phi(0, \vartheta_1, \dots, \vartheta_d) = 1$ (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. 2 and 3 (as well as in the proofs Lemma 3 and 4 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] J. Blom, O. Kella, M. Mandjes, and H. Thorsdottir (2012). Markov-modulated infinite server queues with general service times. *Queueing Systems,* to appear.

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

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

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

[5] 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.

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

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

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

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

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

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

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

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

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