Limit theorems for Markov-modulated queues

Halldora Thorsdottir

Limit theorems for Markov-modulated queues

UNIVERSITY OF AMSTERDAM

The research described in this thesis was conducted at and financially supported by Centrum Wiskunde & Informatica (CWI), the Dutch national research institute for mathematics and computer science, and the University of Amsterdam (UvA).

Cover design by Bouwe van der Molen
Printed by Proefschriftmaken.nl | Uitgeverij BOXPress

Limit theorems for Markov-modulated queues

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het College voor Promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel
op vrijdag 13 mei 2016, te 14:00 uur

door

Halldóra Þórsdóttir

geboren te Urbana, Illinois, Verenigde Staten.

Promoticommissie:

Promotor:	Prof. dr. M.R.H. Mandjes	Universiteit van Amsterdam
Copromotor:	Prof. dr. U. Ayesta	University of the Basque Country
Overige leden:	Prof. dr. I.J.B.F. Adan	Technische Universiteit Eindhoven
	Prof. dr. D.T. Crommelin	Universiteit van Amsterdam
	Prof. dr. M.C.M. de Gunst	Vrije Universiteit Amsterdam
	Prof. dr. R. Núñez-Queija	Universiteit van Amsterdam
	Dr. W.R.W. Scheinhardt	Universiteit Twente
	Dr. P.J.C. Spreij	Universiteit van Amsterdam
	Prof. dr. J.H. van Zanten	Universiteit van Amsterdam

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

Dedicated to Ása,
born in the same week as this thesis.

Contents

Chapter 1

Introduction

We all live in a random environment. This can lead to delightful surprises but also to a complicated existence, due to the fact that the environment controls other processes in our daily lives. Examples include the weather, on which we may base our decision to go to the park or stay at home, it can be a traffic jam due to an accident, causing road delay or it can be high body temperature, which can lead to an increased protein production. The list of situations and systems affected by such unpredictable external circumstances can contain anything from biology to large scale computing.

In this thesis, which falls under the field of probability theory called queueing theory, we focus on queueing systems embedded in a random environment. In a nutshell, queueing theory describes probabilistic service systems where the main players are usually discrete entities called customers; here they will also be referred to as jobs, particles or molecules. A natural way to incorporate a general, external random environment into such mathematical models, is to let the model parameters governing the queue be affected by this environment.

The best known queueing system is the *single server* queue, a typical example being a hot dog stand. Customers line up to get service, namely the preparation of their hot dog. Two key features of the queueing system include (i) how much time passes between subsequent customers and (ii) how long it takes to prepare each hot dog; (i) is referred to as *inter-arrival time*, (ii) as *service time*, both are random quantities that are assumed to behave according to some probability distribution. It is also important to define how much service a queueing system can provide. If the hot dog server gets too busy on his own, he or she might consider adding servers to the hot dog stand, leading to a *many server* queue.

Some service systems have the special feature that the customers *bring their own server*, so that all customers can obtain the service that they request simultaneously, independently of each other. This is called an *infinite server* system. In the example of the hot dog stand this would

mean that each time a new customer enters, a server on standby jumps to attend to the customer immediately, regardless of how many are already being served. A more realistic example of an infinite server system is that of new protein molecules being generated in our bodies. Then their life-span can be seen as their service time. Importantly, the molecules do not need to wait for others to obtain or complete their service. In this model there is no waiting line, but techniques from queueing theory can still be helpful to analyze e.g. how many molecules are in the system at any given time. The assumption of infinite servers is also used to *approximate* many servers in systems where customers rarely need to wait.

A crucial specification of a queueing system with a limited number of servers is who gets served first. At the hot dog stand, the common consensus is that of first come, first serve (FCFS). This is probably the most studied queueing discipline, but plenty of others exist and are widely used. Another well-known service discipline is the processor sharing (PS) policy, where the service resources are equally divided between all the customers present. The name is derived from computer processors, which can serve multiple jobs *simultaneously*, just as the PS discipline dictates. Yet other service disciplines employ priorities. These are frequently applied by airlines, that differentiate between *customer classes* by e.g. inviting their business class passengers to board ahead of other passengers.

Possible probability distributions of the inter-arrival and service times come in all shapes and sizes. Most commonly, the probability of the time between two arrivals being more than t is assumed to be exponential, $e^{-\lambda t}$, for some parameter $\lambda > 0$. The exponential distribution is also a common assumption for service times. An important implication of this assumption is that the probability that a given number of customers will arrive in the future does not depend on how many have arrived in the past. Similarly, remaining service times do not depend on the elapsed service times, which makes for easier computations. This memorylessness is called the *Markov* property and a queue of this type with s servers is called an $\mathrm{M/M}/s$ queue, the M's standing for Markovian arrivals and Markovian services. Markovian queues have been thoroughly studied and their properties are listed in numerous textbooks.

Queueing is traditionally a field of operations research and performance is of special interest and importance. Typical performance measures for queues include queue length, waiting times and workload that the service system has yet to process. The probability distributions to which these performance measures adhere are well known for many queueing systems. It is for example known that in *steady-state*, the number of cus-

tomers in the M/M/1 queue follows the geometric distribution, and in an M/M/∞ queue it is Poisson distributed.

In what follows in this thesis, we frequently work with Markovian arrivals and services, but let some of the parameters be affected by an external, random environment. Even though the environment itself may be Markovian, the fluctuations that it dictates partially remove the convenient Markov property from the arrival and service processes, making analysis more difficult.

In the chapters to come we focus on two classes of queueing systems in a random environment, starting with an infinite server system and followed by a single server system. The first three chapters contain analysis of the steady-state and transient behavior of infinite server systems. The systems' most convenient feature is their unlimited service capacity and lack of waiting, resulting in various independence properties.

In the last chapter we analyze the steady-state of a single server queue under the so-called discriminatory processor sharing (DPS) service discipline, which is an extension of the PS discipline mentioned above. DPS is applied to a system with multiple customer classes, where all customers are served simultaneously but with a weight according to their class. Contrary to the infinite server system, this leads to full interdependency between the customers present. As a consequence, the way in which service is rendered is crucial for the performance. The DPS discipline is particularly suited for applications where jobs do not require the complete attention of the server. An example of this is internet data traffic, where transmission can occur piecewise, e.g. per packet of groups thereof. Moreover, the more jobs that are present in a DPS system, the less service each one gets, just as when dividing bandwidth between different tasks.

One of the overarching themes of the thesis is that of *Markov-modulation*, which is how the random environment is formalized, see Section 1.2. Although the queues analyzed here are to some extent Markovian, adding modulation typically complicates any exact analysis by adding a second probabilistic layer to the problem.

The limiting behavior of the aforementioned queueing systems, see Section 1.1, is the other common denominator behind the main results of the thesis. For this purpose some of the parameters, such as the rate of arrivals, may be taken to be extremely large so that it threatens the stability of the system. Another approach that we take is to study the random environment on a *time scale* such that it moves much faster or slower than the main queueing process. In such cases the performance metrics of the

otherwise complicated queues turn out to follow well known and well behaving distributions, allowing for direct conclusions about the properties of the limiting queue.

The remainder of this introduction is organized as follows. We start with background information, literature and preliminaries for limit theorems and Markov-modulation, the two main themes that have guided the research presented in this thesis. The concept of limit theorems is broadly introduced in Section 1.1 and some preliminaries about e.g. convergence of random variables are provided. We explain the two limiting regimes applied in Chapters 2 to 5, the central limit theorem (Section 1.1.1) and heavy traffic (Section 1.1.2) and give simple examples thereof. A comparison between the two limiting regimes is given in Section 1.1.3. Section 1.2 contains a brief literature review of Markov-modulated queues, followed by an overview of the more technical aspects and assumptions.

We continue with the contents and contributions of the thesis, including short explanations of the models given in Section 2.1, in particular the discriminatory processor sharing service discipline which is studied in Chapter 5. A discussion about the different chapters and the methods used is in Section 2.2. Finally, Section 3 contains a description of the results derived for the infinite server models in Chapters 2, 3 and 4 and the single server model in Chapter 5.

1 Background and literature

In this section we present selected background information and literature about the two main branches underlying this thesis, limit theorems and Markov-modulation.

1.1 Limit theorems

When faced with a highly complex system, studying limit regimes can serve the purpose of obtaining simpler approximations of complicated processes, whose properties we cannot explicitly compute. These limits should offer the researcher a way to zoom out and see the big picture and, to some extent, separate the important parts from the details. In the field of queues, the limiting approximation processes should ideally be known and nondegenerate *stochastic processes*.

A stochastic process is a collection of random variables, often describing the evolution of these variables over time. Therefore any stochastic limit is based on the convergence of random variables. Such convergence can

be obtained in different ways, here we define convergence in distribution. A sequence (M_N) of random variables is said to converge *in distribution* to a random variable M if $\lim_{N\to\infty} \mathbb{P}(M_N \leq m) = \mathbb{P}(M \leq m)$ at all continuity points of the distribution function. This type of convergence is also known as *weak convergence* and is applied throughout the thesis. In addition to weak convergence for random variables, one can also obtain weak convergence at the *process level*, yielding a so-called functional limit. Proving a functional result is considerably more involved, we return to this concept in Section 1.1.1.

For random variables and stochastic processes, limits can be obtained by scaling the model in question appropriately in time and space. Typically, system parameters are multiplied with scalars, which may then be taken to, say, zero or infinity. How the scaling is imposed on the temporal and spatial scales highlights one of the key features of many complex systems, namely the different behavior they may exhibit at different time scales. Often only one or a few of the different levels are of interest to the observer. This can justify putting the focus on one main process, which in turn can greatly simplify the analysis. In some cases it may be relevant and sufficient to consider some average of the 'secondary' processes, this is addressed in e.g. Chapter 2 of this thesis. Then the limit is not only a goal in itself but also a way to alleviate the original complexity. For example, in biological systems or chemical reaction networks, the difference in concentration of various molecule types can be so large that exploiting the scale separation is a natural and even necessary way to reduce the complexity of such a model, see e.g. the discussion in [13].

Furthermore, by deriving asymptotic limits one can establish boundary conditions for the main process in question. Such boundary conditions can provide a useful benchmark when testing the validity of computer programs and other experiments, or for statistical tests [93].

The next two subsections are dedicated to the two classical scaling regimes used in this thesis: central limit theorem type of scaling and heavy-traffic type of scaling. We conclude with a discussion on how the two scaling regimes differ and overlap.

1.1.1 CLT scalings

The classical central limit theorem (CLT) concerns the centered and scaled sum of N independent and identically distributed (i.i.d.) random variables, converging in distribution to the normal distribution when N is large enough. It describes the stochastic fluctuations around the mean (or center) as seen at a given time scale, showing that these centered values

adhere to the statistical regularity of the well-known bell curve.

Before discussing more involved forms of the CLT we demonstrate how such a limit can be derived for performance measures of queues using a moment generating function (MGF). Let M denote the length of the M/M/∞ queue in steady-state, which is known to follow the Poisson distribution. We define its MGF as $\mathbb{E}[e^{\theta M}]$, with $\mathbb{E}$ denoting the expected value and $\theta > 0$ being a parameter. Let λ denote the rate of arrivals to the queue, which we scale with N to speed up the arrival stream. Then the *scaled* queue length, denoted $M^{(N)}$, is known to be Poisson distributed with parameter $N\rho$, where $\rho = \lambda/\mu$ is the ratio between the arrival rate and service rate μ. The corresponding MGF is $e^{N\rho(e^{\theta}-1)}$. We compute the limit

$$\begin{aligned}
&\lim_{N\to\infty} \mathbb{E}\left[\exp\left(\frac{\theta(M^{(N)} - N\rho)}{\sqrt{N}}\right)\right] \\
&\quad = \lim_{N\to\infty} e^{-\rho\theta\sqrt{N}} \cdot \mathbb{E}[e^{\theta M^{(N)}/\sqrt{N}}] \\
&\quad = \lim_{N\to\infty} e^{-\rho\theta\sqrt{N}} \cdot e^{N\rho(e^{\theta/\sqrt{N}}-1)} \\
&\quad = \lim_{N\to\infty} \exp\left(-\rho\theta\sqrt{N} + N\rho[\theta/\sqrt{N} + \theta^2/2N + \mathcal{O}(N^{-3/2})]\right) \\
&\quad = \exp\left(\theta^2\rho/2\right),
\end{aligned}$$

where in the second step we have inserted the known MGF of the Poisson distribution and applied Taylor expansion in the third step. The computations yield the MGF of the normal distribution. In other words,

$$\frac{M^{(N)} - N\rho}{\sqrt{N}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, \rho), \tag{1.1}$$

where $\xrightarrow{\mathrm{d}}$ denotes convergence in distribution and $\mathcal{N}(a, b)$ denotes the normal distribution with mean a and variance b.

The classical CLT was generalized in Donsker's theorem (see e.g. [93]) to a *functional* CLT (FCLT). Where the basic form of the CLT applies to the sum of N i.i.d. random variables, for N large, the name functional CLT stems from the fact that the convergence in distribution holds for many functionals, such as the maximum of the first N sums. In fact, an FCLT a is weak convergence of a stochastic process, and since a stochastic process is an infinite dimensional collection of random variables, this is a stronger result than the weak convergence of single random variables defined above. To clarify the difference, suppose that a given FCLT

shows convergence to a Gaussian stochastic process, describing the evolution of random variables over time. Then any time point of the Gaussian process corresponds to a normally distributed random variable, and any finite collection of such time points has a multivariate normal distribution. Finally, the joint distribution of all those infinitely many random variables, i.e. all time points, is the distribution of the Gaussian process. These steps also partly describe how an FCLT can be derived. First, one establishes convergence for a single time point, then for a finite set of time points (see e.g. Chapter 3). The final step is typically a so-called tightness argument (see e.g. [85]). An alternative approach to proving an FCLT can be seen in Chapter 4, where the *martingale* CLT (see [39]) is used.

When the limiting process resulting from an FCLT is a *diffusion process*, it is called a diffusion limit. A diffusion process is a continuous time Markov process with almost surely continuous sample paths; the process is completely determined by its drift and diffusion terms. Two examples are the well-known Brownian motion and the Ornstein-Uhlenbeck process.

Brownian motion describes the diffusion of a particle in fluid. In the field of stochastic processes, Brownian motion is often described by the Wiener process, a Gaussian process which is characterized by its independent and normally distributed stationary increments. The Ornstein-Uhlenbeck (OU) process describes a Brownian particle under friction. The difference in the mathematical description of the two processes lies in the drift term, which attracts the process back to a central location. The OU is a Gaussian process, but unlike the Wiener process, it is mean-reverting.

In [52], Iglehart showed how the Markovian many server queue can be approximated with an OU process as the number of servers and the arrival rate both go to infinity. From that functional result it is easily derived that at any point in time, the properly scaled queue length follows the normal distribution. The steady-state derivation resulting in Eqn. (1.1) above can be seen as a special case of such a result.

In Chapters 2 and 3, CLT scalings are applied to the length of an infinite server queue with randomly varying parameters and it is shown that the result is normally distributed. This is done for both steady-state and transient queues. The first step there is to establish a law of large numbers (LLN) for the process of interest to obtain its mean. Then we proceed to a CLT, where the scaling looks as follows. Consider a sequence of time-dependent stochastic processes $M(t)$. Inserting the scaling, e.g. speeding up the arrival rates, yields a mapping $M(t) \mapsto M^{(N)}(t)$, for $N \geq 1$, where the latter quantity denotes the scaled process. The LLN scaling we denote

with $\bar{M}^{(N)}(t)$, and the CLT scaling by $\hat{M}^{(N)}(t)$. Respectively, they are defined as

$$\bar{M}^{(N)}(t) \equiv \frac{M^{(N)}(t)}{N} \Rightarrow \rho(t), \quad \hat{M}^{(N)}(t) \equiv \frac{M^{(N)}(t) - N\rho(t)}{a_N},$$

where $\rho(t)$ is continuous and deterministic, a_N is a deterministic sequence chosen carefully such that the fraction goes neither to zero nor to infinity, but to a nondegenerate limit. These are also referred to as fluid scaling and diffusion scaling; in fact fluid queues are also popular approximations for many complex queueing systems. For both scalings the arrival stream is accelerated by N, however for the fluid scaling space is scaled down by N whereas the diffusion-scaled processes are centered and then scaled down by a_N, which is usually $\sqrt{N}$. The CLT tells us that the range of values for the centered $M^{(N)}(t)$ should be of order a_N [93]. As an approximation, the CLT is a refinement of the LLN approximation, since it gives information about the variance, in addition to the mean. Fluid and diffusion approximations are discussed in e.g. [29], with a special focus on queues in a slowly changing random environment.

1.1.2 Heavy traffic

Chapter 5 employs the so-called *heavy-traffic* scaling to analyze a multi-class queue in a random environment. The heavy-traffic limiting regime is commonly studied for systems which become unstable due to their finite serving capacity. This is of course never the case for infinite server systems, that is, they never become unstable. A queueing system is said to be in heavy traffic when the total input load approaches that of the service capacity. Typically for heavy-traffic analysis, this causes classic performance metrics such as queue length, waiting time and workload to grow out of bounds, but by imposing a carefully selected scaling, a non-degenerate stochastic process limit may be found. This usually involves multiplying the process of interest with $(1 - \rho)$, with ρ being the traffic intensity, and then letting ρ go to 1. The resulting limiting quantity is frequently exponentially distributed. That the M/M/1 queue length is exponentially distributed in heavy traffic can be derived using the well-known fact that the queue length in steady-state is geometrically distributed. The corresponding MGF is $\frac{1-\rho}{1-\rho e^{\theta}}$, with $\rho = \lambda/\mu$ as before. After

scaling we compute the limit

$$\begin{aligned}\lim_{\rho\to 1}\mathbb{E}[e^{(1-\rho)\theta M}] &= \lim_{\rho\to 1}\frac{1-\rho}{1-\rho e^{(1-\rho)\theta}}\\ &= \lim_{\rho\to 1}\frac{-1}{-e^{(1-\rho)\theta}(1-\rho\theta)} \quad \text{(by l'Hôpital's rule)}\\ &= \frac{1}{1-\theta},\end{aligned}$$

which is the MGF of the exponential distribution with mean 1.

The paper of Kingman [63] is often cited as one of the first studies of the later well to be known heavy traffic limiting regime. There he shows that the scaled workload of an FCFS GI/GI/1 queue, i.e. with generally distributed and independent arrivals and services, is exponentially distributed when the rate of arrivals approaches that of the service. Later this was extended in [53] to show that the queue length converges weakly to reflected Brownian motion (RBM). As the name implies, RBM is constructed from Brownian motion and a reflecting boundary, e.g. the positive axis; this results in non-negativity which is clearly an important feature of the queue length and many other performance metrics in queueing theory. RBM is the typical functional counterpart of the exponential limiting distribution that applies in heavy-traffic scaling.

The convergence of the steady-state distribution to an exponential distribution, or RBM if the process level is considered, holds for several classes of single server queues, including the Markov-modulated M/M/1 queue under the FCFS discipline, see [7]. Remarkably, in particular cases queues with non-Markovian input can be approximated by queues with Markovian input in the heavy-traffic regime. An example is the heavy-traffic approximation for the queue length process of the non-Markovian GI/GI/s queue; also in this case RBM is the limit process [77].

1.1.3 Choice of scaling and terminology

Choosing an appropriate scaling and limiting regime for the model in question can come naturally. In the terminology of this thesis, the heavy-traffic scaling refers to pushing a queue towards instability, then finding the limiting distribution of a single, scaled performance metric, whereas results of the CLT kind involve the limit of a centered and scaled quantity, derived after e.g. speeding up arrivals. Somewhat confusingly, there is a body of literature where CLT type of results are presented under the name of heavy traffic, see e.g. [77, 82]. These limiting results, which are what here is referred to as CLT results, follow from Donsker's FCLT and

the continuous mapping theorem, an approach which is thouroughly discussed in [93]. See also [76] for a survey of CLT results for many and infinite server queues using martingale theory. Although any scaling regime which pushes the traffic intensity to its critical point can, intuitively speaking, earn the name heavy traffic, we will stick to the dichotomy described at the beginning of this paragraph.

As mentioned earlier, CLT scaling typically yields the normal distribution or a Gaussian limit, whereas heavy-traffic scaling for queues results in the exponential distribution or reflected Brownian motion. An example of a crossover between the two limiting regimes is given by Ward and Glynn in [92], which is also a convenient overview of diffusion approximations. In their studies of diffusion limits for queues with reneging, the authors adapt the natural point of view of seeing infinite server systems as queues with *only* reneging. Their objective is to combine heavy traffic with reneging, so the focus is on choosing a scaling which yields both types of limits. The authors of [92] establish diffusion limits, namely an OU process, and note that "the reflected OU process plays the same role in the reneging context as does reflected Brownian motion in the setting of conventional queues." This is consistent with findings for the infinite server queue, as discussed above and in Chapter 4.

1.2 Markov-modulation

Some customer service systems are very sensitive to their surroundings. For an ice cream parlor, a sunny day gives a great boost to the rate of arrivals, whereas rain has the opposite effect. The weather is thus the quintessential example of a *modulating*, external random environment, which can greatly affect the behavior of a queueing system. The simplest way to include modulation is to let a given parameter switch between on and off according to a two state environment. Breakdowns in a factory line can be represented with on-off modulation, for arrivals, service or both. A slightly more complicated example from biology concerns so-called mRNA molecules. Lab experiments have shown that there is great variability in the generation rate of new mRNA molecules in cells, even if they are grown under the same circumstances and possess identical genetic materical. This is used as a basis for the models in [87, 86]. The molecules' lifetime can be viewed as service time in an infinite server system and since production of molecules can in queueing terms be called arrival of new customers, the randomly fluctuating rate is a prime example to model as a queueing system in a random environment. The external process may then represent the state of the cell in which the molecule

production is taking place, which in turn may affect the production rate. Yet another motivation is the modeling of modulated service capacity, see [74]. There, part of the service capacity is dedicated to priority customers, while the remaining, fluctuating capacity is divided between any other customers present.

When the random environment is formalized by a Markov process, we refer to the queue as being *Markov-modulated*. Usually, the rate of the arrival or service process fluctuates according to this external Markov process, adding flexibility to the model.

To clarify the idea we introduce the model more formally, as well as the notation that will be used later in this section: Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_D)$ denote the equilibrium distribution of the environment, which is an irreducible continuous time Markov chain with state space $\{1, \ldots, D\}$. Further, let λ_i denote the arrival rate while in state $i = 1, \ldots, D$ and its time-average with $\lambda_\infty := \sum_i \pi_i \lambda_i$. Let F_i be the service distribution associated with state i, $1/\mu_i$ be its mean and the corresponding time-average $\mu_\infty := \sum_i \pi_i \mu_i$. Finally, an important quantity is the steady-state probability that the queue is empty while the Markov chain is in state i, which we denote by $p_{0,i}$.

Markov-modulation is one way to relax the widespread assumption of independent and identically distributed arrival and service times. In particular, a fluctuating rate causing, say, high and low tides of arrivals, results in correlation in the stream of customers, see discussion on Markov-modulated arrivals in Section 1.2.2. Observe that if, say, the arrival distribution affected by the external environment is exponential, albeit with a fluctuating rate, then its evolution *conditioned on the environment* is memoryless. In this thesis we apply modulation to queues in which both the inter-arrival and service distribution are exponential (Chapter 2 and 4) as well as only the inter-arrival distribution (Chapter 3 and 5).

The inter-arrival or service distribution rate can also be assumed to fluctuate in a deterministic fashion. An example of that is heterogeneous Poisson arrivals or departures, where the rate fluctuates as a function of time, see e.g. [37, 36, 95] for non-homogeneous, bulk and compound Poisson arrivals, respectively. The non-homogeneity already greatly complicates the analysis and may explain why there is less literature available than on queues with constant rates [69].

The remaining discussion on Markov-modulation is split into two parts. Section 1.2.1 contains a short review of the literature on modulated queues in terms of number of servers, various service disciplines, customer classes and types of environment. The aim is not to give a fully comprehensive list but rather to include the results that guided the author in writing

this thesis. Section 1.2.2 describes the more technical aspects of Markov-modulated queues, challenges and important assumptions, which serve as preliminaries for the results in the thesis.

1.2.1 Literature review

Single server queues
The single server queue with Markov-modulated arrival or service times, or both, has been studied extensively since the 1970s. Early papers include Yechiali and Naor [97] and Eisen and Tainiter [38], in which a two state background process modulates the arrival and service rates of an M/M/1 queue. Later this was generalized to a finite-state background process by Purdue, who in [79] derived results for the busy period, equilibrium conditions and emptiness probability. In [71] Neuts presented matrix-geometric methods to analyze single and many server queues in a random environment. A key result on Markov-modulated single server queues is that the stationary distribution of the number of customers is of matrix-geometric form, it is thus a 'matrix-generalization' of the normal M/M/1 queue. Asmussen studies the waiting time distribution for the Markov-modulated M/G/1 queue in [8]. In [78], Prabhu and Zhu survey various types of modulated single server queues, including ones where the modulating background process moves in a countable but not necessarily finite set.

For results on other service disciplines, see e.g. [74] about the processor sharing (PS) queue. Sengupta characterizes in [88] the Laplace-Stieltjes transform (LST) of the sojourn times of the modulated M/M/1 under four different service disciplines, FCFS, LCFS, PS and round-robin. In [49] the authors use a simple modulated system to study correlations in subsequently arriving jobs and compare performance between seven different service disciplines.

The area of multiclass queues has not been left untouched by modulation. Using a time-changing argument for a fairly general class of service disciplines, Takine [89] obtains exact results for a system with modulated arrivals and service times. Arrivals can only occur at transition epochs of the modulating process and customers require generally distributed amounts of work. For another type of multiclass system, a form of the $c\mu$-rule (a scheduling rule dictating which class should be served first) is shown to be optimal in [27]. Here the system is studied in heavy traffic from a scheduling point of view, by formulating a Brownian control problem. Finally, a recent paper by Dorsman et al. [35] contains heavy-traffic results for a *network* of queues with Markov-modulated service speeds.

Infinite servers
In comparison to the much studied single server model, the queueing literature for infinite servers under modulation is not as rich. Some results for the system's steady-state behavior have already been available for some time; see e.g. [32, 75] for the factorial moments of the number of customers in the system. Furthermore, in [75] it is shown that contrary to the M/M/1 queue, the 'matrix-generalization' does not hold for the infinite server case when moving from the classic to the modulated queue. The stationary distribution of the number of customers in an $M/M/\infty$ queue with Markov-modulated service times has been derived in [14].

For the sake of completeness, we point out that after the scaling results on Markov-modulated infinite server queues in Chapters 3 and 4 were published, research in the same direction has moved further, see [19, 20] for CLT type of results where the environment can be relatively faster or slower than the arrival process, and FCLT results for modulated service rates, respectively.

Non-Markovian modulation
There are more ways to model random rate fluctuation than by an independent, Markovian environment. Related results have been established for semi-Markov-modulated queues, where the sojourn times of the external modulating environment are not exponential. See e.g. [44] for queue length results in such an infinite server system; in addition, in this paper the service requirements are Erlang or hyperexponentially distributed. Another semi-Markov-modulated result is [32], where a stochastic decomposition formula for the $M/M/\infty$ queue is derived. The results of Chapter 2 are also for a semi-Markov-modulated queue. In [26] the modulation is governed by a general random variable. There it is not only used to change the parameters of otherwise Markovian queues, but rather to switch between completely different service distributions. The results concern the workload of an M/G/1 queue with two service speeds: high speed negative exponential periods and generally distributed low speed service periods.

Queues with modulated service depending on the state of the system, such as its workload, are easy to motivate with applications. An overview can be found in [15].

1.2.2 Busy period, arrivals and services

A fundamental challenge in getting explicit solutions for modulated systems already starts when trying to find an explicit expression to probabilistically describe the busy period; in line with this, characterizing the emptiness probability $p_{0,i}$ (per state i of the environment) is problematic. In e.g. heavy-traffic analysis, it is important to prove that the probability of a queue being empty vanishes in the appropriate limit. For many non-modulated queues this is immediate, but not necessarily so in the modulated case. In [71] the queue length probability distribution, and in particular $p_{0,i}$, is expressed by an implicit vector-matrix multiplication. In [79] it is shown that the busy period matrix, with entries based on the environment's state at the beginning and the end of a busy period, satisfies an integral equation.

Modulated arrivals
Modulating the rate of the arrival stream can be seen as a first step in introducing a stochastic, non-homogeneous Poisson distribution. This is suitable for applications where the generation of new jobs exhibits a certain burstiness or is time-varying in a non-deterministic way. Even when service rates are kept fixed, modulated arrivals affect the queue length distribution, and can in the case of limited service capacity move the system between overload and underload, while overall stability is maintained. Its practical importance is explained in e.g. [28], where Burman and Smith study delay and queue length using light and heavy-traffic approximations. The authors also explain how the independence of inter-arrival times is removed by the time-varying arrival rate. Due to the variation, the inter-arrival times give additional information about the number of customers. This is particularly interesting in the case of multiclass systems, where the modulation induces correlations between subsequent jobs which may have different service requirements based on their class, see e.g. [49]. Correlated inter-arrival times and service times are also studied in [1].

Modulated service
While modulating arrival rates leaves little ambiguity as to how it should be done, the modulation of service rates is another story and can be done in primarily two different ways. The two models, now discussed briefly, can both be reasonable for different scenarios, but they often require separate analysis. In some cases, e.g. [19], both are studied with a unified approach.

First we discuss the model where a job's residual service time may change when the background process jumps to another state; we call this a model with *continuous modulation*. Then the random variable F_i can be used to denote the service time precisely while the background process is in state i. In the case of exponential service times, this entails that a job can take on many service rates during its stay.

Some classical results for this variant of the modulated M/M/1 queue with the FCFS discipline are e.g. [97, 71]. The stability condition for this model is that the average arrival rate should be less than the average service rate, that is the traffic intensity $\lambda_\infty/\mu_\infty$ should be less than 1. An important observation is that contrary to the non-modulated single server queue, the busy fraction $1 - \sum_i p_{0,i}$ does *not* equal the traffic intensity in general. The main challenge with this model is that certain performance measures are also continuously modulated. One example is the workload, which under this assumption can increase and decrease with the changes of the environment. This makes for a more difficult analysis and is addressed in Chapter 5. A common assumption in the literature of modulated queues is to let only the service *capacity* be continuously modulated, see [26, 35, 68, 89, 88]. The service *requirement* is then assumed to be independent of the environment.

An alternative form of service modulation is when a job's service distribution is based on the state of the background process *upon its arrival*. That is, a job's service time is denoted by F_i, $i = 1, \ldots, D$, if the environment is in state i when it arrives. The service distribution does not change during its presence in the system, regardless of changes in the environment. Regterschot and De Smit analyze waiting times and queue lengths at arrival epochs and in continuous time in [84], using Wiener-Hopf techniques. Their steady-state results are based on the stability condition $\sum_i \pi_i \lambda_i/\mu_i < 1$, notably different to the continuous modulation variant. This modulation is also used in [8, 34] on the Markov-modulated M/G/1 queue and [78] with modulated compound Poisson arrivals.

This service model effectively divides the jobs into D different classes, determined by the state of the environment upon their arrival. This is addressed for the infinite server system in [19], where a CLT for this variant (referred to as Model II) yields a D-dimensional normal distribution. In Chapter 5, Remark 2.1, we show how in case of a multiclass system, this type of modulation can be rewritten to simply extend the number of job classes, leaving out many cumbersome details of the modulation.

A model which sits somewhat in between these two was proposed by Neuts [72], where the service time of a job depends only on the environment at the beginning of its service.

2 Models and methods

In this section we review the contents of Chapters 2-5 of the thesis. Most importantly, Chapters 2, 3 and 4 contain results on infinite server systems in a CLT regime, whereas Chapter 5 is about a modulated M/G/1 type queue in heavy traffic, with a special focus on the DPS service discipline. First, we briefly explain the models, in particular the service discipline used in Chapter 5. This is followed by an overview of the characterizing features of the different chapters in relation with each other. We conclude with a section on the contributions of the thesis.

2.1 Models

In Chapters 2-4 we study a few variants of modulated infinite server systems under a particular scaling. The infinite server queue can be used to model a system where the service that a job receives is not affected by the other jobs present. This assumption can either reflect actually infinite service capacity or be an approximation for a system where the service capacity is deemed to be large enough. Since there is no waiting, a queue is never formed in an infinite server system, so one may also refer to it simply as a birth and death process. Importantly, the notion of a service discipline becomes irrelevant. Chapters 3 and 4 make use of an environment, referred to as the *background process*, which is formalized by an irreducible continuous time Markov chain, whereas in Chapter 2 the environment has deterministic transition times.

Chapter 5 contains the study of an M/G/1 type queue under Markov-modulation in heavy traffic. Whereas the arrivals follow a modulated Poisson process, the service distribution is general in the first half of the chapter, and in the second half a special case is studied under exponential service times. In particular, that part of the chapter employs the DPS service discipline which will now be explained. We also include a brief literature overview.

The DPS model extends the processor sharing (PS) service discipline to a multiclass system with class-dependent weights. The egalitarian PS is a model introduced by Kleinrock in [65], where service resources are simultaneously shared equally among all customers present. In the discriminatory variant with, say, K customer classes, each class k is assigned a weight g_k, $k = 1, \ldots, K$. If M_k denotes the number of class-k customers present, the server dedicates the fraction

$$\frac{g_k}{\sum_{j=1}^{K} g_j M_j},$$

to each customer of class k. When all the weights are equal, DPS and PS are the same. DPS has turned out to be suitable to model the simultaneous parsing of diverse tasks, such as processing network data. Most available results are in terms of limit theorems and moments, which is somewhat telling for the challenging nature of the DPS discipline.

Fayolle et al. [42] established the mean sojourn time conditioned on the service requirement, as well as the mean queue lengths of the different classes, which were shown to depend on the entire service requirement distributions of all classes. The DPS model has queue lengths with a finite mean, irrespective of any higher-order characteristics of the service distribution, see Avrachenkov et al. [11]. This is an extension of a result for the PS system, which holds while the queue is stable. In [83] Rege and Sengupta prove a heavy-traffic limit theorem for the joint queue length distribution in a DPS system where service times are exponentially distributed. They also develop a recursive formula to compute all moments of said distribution. A thourough overview of DPS results can be found in [2]. Later research includes e.g. [91], which extends the heavy-traffic results of [83] to phase-type distributed service requirements, and [54] with approximations for the mean sojourn times, based on combining light and heavy-traffic limits.

2.2 Methods

Overall, the main results of this thesis are in terms of particular scalings, i.e. it is shown that in the limit of an appropriate scaling, key performance measures, primarily the queue length distribution, converge to well-known distributions, the normal and the exponential distribution. In Chapters 3-5 we assume that the external environment is represented by an irreducible continuous time Markov chain on a finite state space, but in Chapter 2 it is a semi-Markov process.

In Chapter 2 and 4, we apply modulation to a Poisson arrival process, with the service times being non-modulated and exponential. Chapters 3 and 5 allow for more general service distribution and there, both arrivals and services are affected by the modulating process.

Whereas the results of Chapter 5 are in steady-state, those of Chapters 2-4 are at a transient level, furthermore the results of Chapter 4 are at the process level. The work presented in Chapter 4 is derived under a different framework and methodology from the other chapters. It is primarily based on unit-rate Poisson processes, as will be briefly explained in the following section, as well as martingale theory and the martingale CLT. This toolkit immediately yields a functional limiting result.

The methodology of Chapters 2, 3 and 5 will be more familiar to the reader of queueing literature. There the starting point is fixed-point and conditional equations, used to describe the infinite server queue in Chapters 2 and 3, and balance equations used for the M/G/1 queue in Chapter 5. Due to the modulating background process, however, those equations become particularly uninviting. By applying the right scaling and Taylor expansion, the equations simplify considerably, which helps in deriving the limits.

Chapters 2-4 all exploit the concept of time-scale separation in a very explicit way. The scaling of choice is then applied to both the environment and the arrival process to the queue, both are pushed to infinity albeit at different speeds. On the one hand, when the environment is sped up faster than the arrival process, it will only be perceived as an average from the perspective of the main process, the queue length. More precisely, instead of having multiple arrival rates due to the modulation, one only observes an average arrival rate. In this case, the averaging obtained after taking the limit greatly simplifies the analysis. On the other hand, slowing the environment down relative to the arrivals, as seen in Chapter 4, results in a sequence of temporary steady-states. In that case the actual deviation between the transient and the equilibrium distribution becomes more apparent.

The heavy-traffic scaling in Chapter 5 lets the arrival rate be increased in such a way that the traffic intensity reaches its critical point. In the limit of this scaling, the distribution of the environment becomes independent of the queue length process.

Non-scaled results are also presented in Chapters 3 and 5, primarily recursions for moments and differential equations that describe properties of the distribution of the number of customers or workload. These results are obtained using transforms, and the same holds for the majority of the results in the whole thesis, with the exception of Chapter 4. Such transforms, i.e. Laplace transforms, moment generating functions or probability generating functions, fully and uniquely characterize a probability distibution and facilitate certain computations. In Chapters 2, 3 and 5 the idea is to evaluate transforms under the particular scaling imposed. After taking the limit, we obtain the transform of a known distribution, which implies that the sequence of random quantities under consideration converges to one following this particular distribution. In this way, the goal of a known, nondegenerate limit is reached.

3 Contribution

As mentioned in Section 1.2.1, there is much less literature available on modulated infinite server systems than on its finite server counterparts. This partly motivates the research presented in Chapters 2-4.

In Chapter 2 we study an infinite server queue fed by a modulated Poisson arrival stream, both steady-state and transient behavior. The exponential service rate is not affected by the background process, which has deterministic transition times. This is called a *semi-Markov-modulated* system. Two scaling regimes are analyzed, the *Poisson regime*, where the background process is sped up by a scaling parameter N, and the *CLT regime*, where the arrival rates λ_i are sped up by N but the transition rates of the background process are scaled by $N^{1+\varepsilon}$, for some $\varepsilon > 0$. In the first regime, the arrival process becomes asymptotically Poisson with a uniform rate λ_∞. The second regime makes use of this result, showing that the scaled and centered queue length variable is normally distributed. In both cases the road map is to set up a system of equations for the appropriate generating function, send N to infinity to obtain a differential equation and then solve it to obtain a known generating function. Especially the second step requires some delicate handling.

Chapter 3 contains results for the $M/G/\infty$ queue where both arrival rates and service times are modulated. The service distribution of a job is based on the state of the environment at its arrival. The CLT scaling regime from Chapter 2 is also analyzed to obtain a multi-dimensional CLT, i.e. for a vector of time points. Similarly to Chapter 2, this is achieved by setting up a system of differential equations for the MGF of the number-in-system. To obtain a meaningful, solvable form we use properties of the deviation matrix (see [31]) for Markov chains, which, roughly speaking, is a measure of the deviation of the time-dependent transition probabilities of the Markov chain from its invariant distribution. The chapter also includes a number of non-scaled results, such as the transient mean and stationary variance of the queue length, and as a special case, recursive moments in the case of exponentially distributed service times.

In Chapter 4 we make use of a Markovian framework developed by Kurtz and collaborators, see e.g. [5, 13], in studying the $M/M/\infty$ queue with modulated arrivals and a uniform service rate. This framework has been used extensively for studying chemical reaction networks, i.e. where a continuous time Markov chain represents a molecule count. It is well suited to analyze birth and death processes, such as the infinite server system, but has also been used for single server queueing models, such as in [27]. Since we are concerned with the queue length, the

essence is to count molecules. Namely, counting arriving molecules and subtracting departing (or decaying) ones. This is done using unit-rate Poisson processes based on the time-change representation argument: If $Y_\lambda(\cdot)$ is a Poisson process with rate λ, then the distribution of $Y_1(\lambda t)$, a unit-rate Poisson process *evaluated* at λt, is the same as that of a Poisson process with rate λ evaluated at t,

$$\mathbb{P}[Y_1(\lambda t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \mathbb{P}[Y_\lambda(t) = k].$$

This setup can easily incorporate fluctuating rates and even many types of molecules, i.e. a multiclass system. After appropriately centering and scaling the queue length process, we are able to rewrite the problem to find a suitable sequence of martingales. Subsequently we use the martingale CLT (see [39]) to prove weak convergence to an Ornstein-Uhlenbeck process. In addition to the scaling applied in Chapters 2 and 3, here we also speed up the background and arrival process in such a way that the environment can also move slower than the arrival process, based on a scaling parameter α. When faster, the parameters of the OU process are in accordance with the results obtained in Chapters 2 and 3, whereas when the environment moves slower than the arrival process, the correct scaling is shown to depend on α. Furthermore, in this case the variance parameters of the OU process contain terms from the deviation matrix also used in the previous chapter.

The main novelty of Chapter 5 is the combination of a multiclass queue under the discriminatory processor sharing (DPS) service discipline, with Markov-modulation, i.e. modulated arrivals, service requirements and service capacity. We study two variants of an M/G/1 type queue affected by a random environment, formalized by Markov-modulation. The main results are derived under heavy-traffic scaling. First, we derive the mean of the total workload assuming generally distributed service requirements and any service discipline which does not depend on the modulating environment. There we assume that the service requirements are based on the state of the environment upon a customer's arrival, whereas the capacity of the server is continuously modulated. It is then shown that the workload is exponentially distributed under heavy-traffic scaling, which is an extension of [34]. This result can be applied to the analysis of a modulated multiclass system, which leads to the second part of the chapter, where the focus is on a multiclass system under DPS. For this second variant, we assume exponential, class-dependent service requirements. It is shown that the joint queue length distribution undergoes a state-space collapse when subject to heavy-traffic scaling. This re-

sult is inspired by [83, 91]. Using the workload result, the limiting queue length distribution is shown to be exponential, times a deterministic vector representing the different customer classes.

The Chapters 2 to 5 are all based on papers which have been published. In this thesis, each chapter's content has been kept as close as possible to the published version to maintain accuracy. Therefore, each chapter contains its own introduction, in some cases acknowledgements and up to a point, notation. This also implies that each chapter is self-contained and can thus be read independently of the others.

Chapter 2

Two time-scalings for a semi-Markov-modulated queue

In this chapter we study semi-Markov-modulated $M/M/\infty$ queues. The Poisson input rate is modulated by a Markovian background process where the times spent in each of its states are assumed deterministic, and the service times are exponential. Two specific scalings are considered, both in terms of transient and steady-state behavior. In the former the transition times of the background process are divided by N, and then N is sent to ∞; a Poisson limit is obtained. In the latter both the transition times and the Poissonian input rates are scaled, but the background process is sped up more than the arrival process; here a central limit type regime applies. The accuracy and convergence rate of the limiting results are demonstrated with numerical experiments. The remainder of this chapter has appeared as article [22].

1 Introduction

Adding the effect of a random environment to the classical $M/M/\infty$ queue provides us with a natural framework to model various real-life phenomena. In this model an infinite server queue is fed by a Poisson arrival stream whose rate is modulated by a Markovian background process, and the service times are exponential. The resulting model, usually called a *semi-Markov-modulated infinite server queue*, is a suitable candidate for several applications, for instance in telecommunication network engineering, where the arrival rates of customers vary between different modes [94]. Another example is the synthesis and later degradation of mRNA strings in cells, after transcription of the DNA which tends to occur in a clustered fashion [87].

In our work we focus on the following variant of the semi-Markov-modulated infinite server queue. The Poisson arrivals to the queue have

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

Other variants of the semi-Markov-modulated queues, such as the single server counterpart, have been widely studied (see for example the discussion in the introduction of [51]); considerably less attention has been paid to the infinite server case. Without aiming at giving a full account of the existing literature, we mention that results for the system's steady-state behavior (mostly in terms of factorial moments) have already been available for some time; see e.g. [32, 75]. Also Markov-modulated infinite server queues with Erlang or hyperexponentially distributed service times have been addressed [44]. Finally we mention that the stationary distribution of the number of customers in an $M/M/\infty$ queue with Markov-modulated service times has been derived in [14].

A new line of research started in [51], where time-scalings are imposed so as to obtain explicit expressions for the resulting limiting distribution. The main result of that paper relates to speeding up the transition times by a factor N; the resulting arrival process turns out to be a Poisson process (with a rate λ_∞ that can be given explicitly in terms of the λ_i and the invariant measure associated with the generator of the background process).

In the present chapter we apply two scalings, to which we refer to (for obvious reasons) as the *Poisson regime* and the *CLT regime*. The first one amounts to dividing the t_i by N, as described above. In the second scaling the rate of arrivals to the system is scaled linearly with N (that is, the arrival rates become $N\lambda_i$), but the transition times are scaled *superlinearly* (that is, they become $t_i/N^{1+\varepsilon}$, for some $\varepsilon > 0$).

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

- Whereas in [51] it was proved that under the Poisson scaling the input process converges to a Poisson process with rate λ_∞, we show in Section 3 that the steady-state number of customers in the system converges to a Poisson distribution with mean $\varrho := \lambda_\infty/\mu$. The transient variant of this result is established in Section 4. At the method-

ological level, the argument used is (i) set up a system of equations for the probability generating function (PGF) of the quantity of interest, (ii) then send N to ∞, and obtain a differential equation for the PGF, (iii) and finally conclude the stated by solving the differential equation. Note that the differential equation is in terms of the argument of the PGF in the steady-state case, and in time in the transient case.

- Under the CLT scaling, we obtain results of the diffusion-type. Essentially two effects are combined. By scaling the transition times by $N^{1+\varepsilon}$, by virtue of the findings in [51], the input process converges to a Poisson process with rate λ_∞. The effect of scaling the λ_is as well is that a central-limit type of regime kicks in, as described in e.g. [85, Section 6.6]. As a consequence, the number of customers minus its expected value, divided by $\sqrt{N}$ converges to a zero-mean random variable. Sections 5 and 6 establish the transient and steady-state version of this result, respectively. The underlying argumentation, although considerably more delicate, resembles the one developed for the Poisson regime.

- We have extensively tested the resulting approximation in a set of numerical experiments, to confirm the speed of convergence to the limiting distribution. In Section 7 we present results, showing that the asymptotics lead to quite accurate approximations, already for relatively low N.

- In a discussion section, we comment on a number of extensions: (i) state-dependent service rate, (ii) general transition times, and (iii) large deviations results.

2 Model description

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

Consider an irreducible semi-Markov process $X(t)$ on a finite state space $\{1, \ldots, d\}$, with $d \in \mathbb{N}$. Its transition matrix is given by $P = (p_{ij})_{i,j=1}^{d}$, where p_{ii} need not necessarily be zero. The time spent in state i is distributed as a non-negative random variable T_i (to be referred to as a *transition time*). The subsequent transition times in state i, say $(T_{i,j})_{j\in\mathbb{N}}$, constitute a sequence of i.i.d. random variables; in addition the sequences

$(T_{i,j})_{j\in\mathbb{N}}$, for various $i \in \{1,\ldots,d\}$, are assumed independent. There is also independence between the jumps of the semi-Markov process and the transition times. While the process $X(t)$, often referred to as the *background process*, is in state i, customers arrive according to a Poisson process with rate $\lambda_i \geq 0$. The service times are assumed to be exponentially distributed with mean $1/\mu$, irrespective of the state of the background process.

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

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

3 Steady-state, Poisson regime

Let, following [51], M_i denote the steady-state number of customers in the system when the background process enters state i. We denote by $\gamma_i(\cdot)$ the probability generating function of M_i: $\gamma_i(z) := \mathbb{E}z^{M_i}$. The probabilities of the time-reversed process, that is of coming from state j, given that the process just jumped to state i, are denoted by $\tilde{p}_{ij} = p_{ji}\pi_j/\pi_i$. Then, from [51, Thm. 2], for the case of deterministic transition times,

$$\gamma_i(z) = \sum_{j=1}^{d} \tilde{p}_{ij} g_j(z) \gamma_j(h_j(z)), \tag{2.1}$$

with

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

We now scale, as in [51, Section 4.2], $t_i \mapsto t_i/N$ (in self-evident notation), and study the solution for $\gamma_i(z)$ in the above fixed point relation (2.1). Intuitively, this scaling means that the background process moves fast between the states in the state space, so that it is conceivable that the particle arrival process tends to a Poisson process as N grows large. It was shown in [51, Section 4.2] that this is indeed the case, with associated arrival rate

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

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

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

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

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

$$\gamma_i^{(N)}(z) = \sum_{j=1}^{d} \tilde{p}_{ij} \left[\left(1 - \frac{\lambda_j t_j}{N}(1-z)\right) \times \left(\gamma_j^{(N)}(z) + \left(\gamma_j^{(N)}\right)'(z) \cdot (1-z)\frac{t_j\mu}{N}\right)\right] + O\left(\frac{1}{N^2}\right). \tag{2.2}$$

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

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

We conclude that $\lim_{N\to\infty} \gamma_i^{(N)}(z) = \gamma(z)$ for a PGF $\gamma(\cdot)$ that does not depend on i.
Now multiply Eqn. (2.2) by N, multiply by π_i and sum over i:

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

Note that by definition of transition probabilities, it holds for any vector ζ that

$$\sum_{i=1}^{d}\sum_{j=1}^{d}\pi_i\tilde{p}_{ij}\zeta_j = \sum_{j=1}^{d}\pi_j\zeta_j\sum_{i=1}^{d}p_{ji} = \sum_{j=1}^{d}\pi_j\zeta_j. \tag{2.3}$$

We will refer to this equality several times throughout the chapter.
Combining the previous three displays and letting $N \to \infty$ we obtain the

differential equation

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

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

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

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

Theorem 3.1. *Under the scaling $t_i \mapsto t_i/N$,*

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

4 Transient, Poisson regime

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

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

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

Define $\varrho_t := (\lambda_\infty/\mu)\cdot(1-e^{-\mu t})$. An elementary conditioning argument yields that

$$\bar{\gamma}_i^{(N)}(z,t) = \sum_{k=0}^{\infty} e^{-\lambda_i t_i/N} \frac{(\lambda_i t_i/N)^k}{k!} (p_i^{(N)}(z,t))^k \sum_{j=1}^{d} p_{ij} \bar{\gamma}_j^{(N)}\left(z, t - \frac{t_i}{N}\right). \tag{2.4}$$

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

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

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

$$\begin{aligned} \bar{\gamma}_i^{(N)}(z,t) &= \exp\left(\frac{\lambda_i t_i}{N} e^{-\mu t}(z-1)\right) \sum_{j=1}^{d} p_{ij} \left(\bar{\gamma}_j^{(N)}(z,t) - \frac{t_i}{N} \frac{\mathrm{d}}{\mathrm{d}t} \bar{\gamma}_j^{(N)}(z,t)\right) \\ &= \left(1 + \frac{\lambda_i t_i}{N} e^{-\mu t}(z-1)\right) \sum_{j=1}^{d} p_{ij} \left(\bar{\gamma}_j^{(N)}(z,t) - \frac{t_i}{N} \frac{\mathrm{d}}{\mathrm{d}t} \bar{\gamma}_j^{(N)}(z,t)\right). \end{aligned} \tag{2.5}$$

Letting $N \to \infty$, we conclude again that $\lim_{N\to\infty} \gamma_i^{(N)}(z,t) = \gamma(z,t)$ for a PGF $\gamma(\cdot,t)$ that does not depend on i. Now multiply Eqn. (2.5) by N, multiply by π_i and sum over i. Due to the 'p_{ij} analogue' of (2.3) the $O(N)$-terms cancel. By virtue of the state independence of $\gamma(\cdot,t)$, we obtain when sending $N \to \infty$,

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

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

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

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

Theorem 4.1. *Under the scaling $t_i \mapsto t_i/N$,*

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

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

5 Steady-state, CLT regime

In, e.g., [85, Section 6.6] an M/M/∞ queue is observed under a linear scaling of the arrival rate λ, that is, one scales $\lambda \mapsto \lambda N$. With Nx_0 customers being present at time 0, a central-limit-theorem (CLT) type of result is proven. More specifically, it is shown that the number of customers in this M/M/∞ system at time t, minus its expected value $N\, m(\lambda, \mu)$, and divided by $\sqrt{N}$, tends to a zero-mean Normal random variable, with variance $v(\mu, \lambda)$; here

$$\begin{aligned} m(\lambda, \mu) &:= x_0 e^{-\mu t} + N\lambda/\mu \cdot (1 - e^{-\mu t}), \\ v(\lambda, \mu) &:= x_0 e^{-\mu t}(1 - e^{-\mu t}) + \lambda/\mu \cdot (1 - e^{-\mu t}). \end{aligned}$$

(In fact, [85, Thm. 6.14] provides us with a considerably more refined result: a functional central limit theorem. More precisely, there is weak convergence of the queueing process to a specific Gaussian process, viz. an Ornstein-Uhlenbeck process.)

The idea of this section (as well as the next section) is that we scale the arrival rate linearly ($\lambda_i \mapsto \lambda_i N$, that is), but we scale the transition times *superlinearly* ($t_i \mapsto t_i/N^{1+\varepsilon}$, for some $\varepsilon > 0$). The effect of this scaling is that we combine the convergence of the particle arrival process to a Poisson process of rate λ_∞ (essentially as in [51, Section 4.2]) with the CLT regime kicking in (as in [85, Section 6.6]). As a consequence, one would expect convergence of the steady-state number of customers, minus $N\, m(\lambda_\infty, \mu)$, divided by $\sqrt{N}$, to a zero-mean Normal random variable with variance $v(\lambda_\infty, \mu)$. The objective of this section is to verify the steady-state counterpart of this claim (in which the variance is λ_∞/μ), where the transient version is established in the next section.

As mentioned above, we now let the arrival rates be $\lambda_i N$, and the sojourn times $t_i/N^{1+\varepsilon}$. Define, with $\varrho := \lambda_\infty/\mu$, the moment generating function (MGF) of $(M_i^{(N)} - N\varrho)/\sqrt{N}$,

$$\delta_i^{(N)}(\vartheta) := \mathbb{E}\left(\exp\left(\frac{\vartheta M_i^{(N)} - \vartheta N\varrho}{\sqrt{N}}\right)\right) = \gamma_i^{(N)}\left(e^{\vartheta/\sqrt{N}}\right) \cdot e^{-\vartheta\sqrt{N}\varrho}, \quad (2.6)$$

where $\gamma_i^{(N)}(\cdot)$ is the probability generating function of $M_i^{(N)}$. From [51, Thm. 2], for deterministic transition times,

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

Expanding the exponential terms yields:

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

Likewise,

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

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

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

with an error term of $O\left(N^{-\frac{5}{2}-2\varepsilon}\right)$. Differentiating Eqn. (2.6), we get

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

and thus,

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

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

$$\begin{aligned} \delta_i{}^{(N)}(\vartheta)=\sum_{j=1}^{d}\Bigg[\tilde{p}_{ij}\left(1+\frac{\lambda_j t_j\vartheta}{N^{\frac{1}{2}+\varepsilon}}+\frac{\lambda_j t_j\vartheta^2}{2N^{1+\varepsilon}}\right)\left(\frac{\mu t_j}{N^{1+\varepsilon}}\left(1-e^{\vartheta/\sqrt{N}}\right)\times\right. \\ \left.\left(\sqrt{N}e^{-\vartheta/\sqrt{N}}(\delta_j^{(N)})'(\vartheta)+N\varrho e^{-\vartheta/\sqrt{N}}\right)+\delta_j^{(N)}(\vartheta)\right)\Bigg] \end{aligned} \tag{2.7}$$

Now multiply (2.7) with $N^{1+\varepsilon}\pi_i$ and sum over i to obtain, relying on Eqn. (2.3) and some Tayloring,

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

Simplifying we arrive at

$$\sum_{j=1}^{d} \pi_j \mu t_j \vartheta (\delta_j^{(N)})'(\vartheta) \tag{2.8}$$
$$= \sum_{j=1}^{d} \pi_j \left(\mu t_j \vartheta^2 \frac{\varrho}{2} + \frac{\lambda_j t_j \vartheta^2}{2} - \sqrt{N}(\mu t_j \vartheta \varrho - \lambda_j t_j \vartheta) \right) \delta_j^{(N)}(\vartheta)$$
$$+ O\left(\frac{1}{N^{\varepsilon}}\right) + O\left(\frac{1}{\sqrt{N}}\right).$$

Note that, on multiplying Eqn. (2.7) with $\sqrt{N}$, we obtain that

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

we thus conclude that also $\sqrt{N}\delta_i^{(N)}(\vartheta)$ is independent of i in the limit $N \to \infty$. In addition, due to the very definition of ϱ, the $\sqrt{N}$ terms of Eqn. (2.8) cancel when sending N to ∞. We consequently obtain the differential equation

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

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

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

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

Theorem 5.1. *Under the scaling $t_i \mapsto t_i/N^{1+\varepsilon}$ and $\lambda_i \mapsto \lambda_i N$,*

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

6 Transient, CLT regime

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

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

In the light of the above remark, we can focus on the number of customers arriving in $(0,t]$ that are still present at time t. Let, as before, in case the modulating process is in state i at time 0, this number be denoted by $\bar{M}_i^{(N)}(t)$; as mentioned earlier, $\bar{M}_i^{(N)}(t)$ is independent of $\check{M}^{(N)}(t)$. The objective of the present section is to analyze $\bar{M}_i^{(N)}(t)$ in the CLT regime.

Recall that

$$\varrho_t := (\lambda_\infty/\mu)\cdot(1-e^{-\mu t}).$$

Then, with $\bar{\gamma}_i^{(N)}(\cdot,t)$ now denoting the moment generating function of $\bar{M}_i^{(N)}(t)$, we define the MGF

$$\bar{\delta}_i^{(N)}(\vartheta,t) := \mathbb{E}\left(\exp\left(\frac{\vartheta\bar{M}_i^{(N)}(t) - \vartheta N\varrho_t}{\sqrt{N}}\right)\right) = \bar{\gamma}_i^{(N)}\left(\frac{\vartheta}{\sqrt{N}},t\right)\cdot e^{-\vartheta\sqrt{N}\varrho_t}. \tag{2.9}$$

Similar to Eqn. (2.4) we note that

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

Here the $p_i^{(N)}(\vartheta,t)$ are MGFs of random variables that are alternatively distributed on 0 and 1, that is $p_i^{(N)}(\vartheta,t) = 1-p(t)+p(t)e^{\vartheta}$, with $p(t)$, the probability of equalling 1,

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

Consequently,

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

and

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

In addition, neglecting higher-order terms as before,

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

Upon combining the above,

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

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

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

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

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

In line with the preceding sections, we multiply the equation by $N^{1+\varepsilon}\pi_i$ and sum over i. Due to the 'p_{ij}-analogue' of Eqn. (2.3) we obtain

$$\begin{aligned} &\sum_{j=1}^{d} \sum_{i=1}^{d} \pi_i p_{ij} t_i \frac{\mathrm{d}}{\mathrm{d}t} \bar{\delta}_j^{(N)}(\vartheta, t) \qquad (2.10) \\ &= \vartheta \sum_{j=1}^{d} \sum_{i=1}^{d} \pi_i p_{ij} t_i \left(\lambda_i e^{-\mu t} \frac{\vartheta}{2} + \lambda_i e^{-\mu t} \sqrt{N} - \varrho_t' \sqrt{N} \right) \bar{\delta}_j^{(N)}(\vartheta, t), \end{aligned}$$

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

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

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

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

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

converges to a Normally distributed random variable with mean 0 and variance ϱ_t.

Taking into account the contribution of the Nx_0 customers that were already present at time 0, our findings can be summarized in the following statement.

Theorem 6.1. *Under the scaling $t_i \mapsto t_i/N^{1+\varepsilon}$ and $\lambda_i \mapsto \lambda_i N$,*

$$\frac{M_i^{(N)}(t) - Nx_0 e^{-\mu t} - N\varrho_t}{\sqrt{N}} \xrightarrow{\mathrm{d}} \mathbb{N}\mathrm{orm}\left(0, x_0 e^{-\mu t}(1 - e^{-\mu t}) + \varrho_t\right).$$

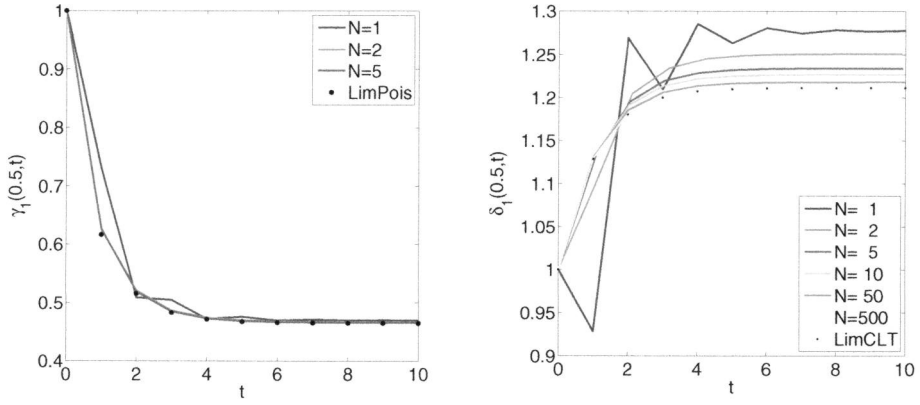


Figure 2.1: (*Left panel*) The transient Poisson regime converging to its limiting curve, $z = 0.5$. (*Right panel*) Convergence of the transient CLT regime to its limiting curve, $\vartheta = 0.5, \varepsilon = 0.5$.

7 Computational results

This section contains numerical results corresponding to the limiting regimes studied in Sections 4 and 6. We compare the resulting approximations with the explicit solutions to Eqns. (2.4) and (2.9), for a range of values of the scaling parameter N. To enable easy numerical evaluation of (2.4) and (2.9), we assume that the deterministic transition times are equal: $t_i = 1$ for all $i \in \{1, \ldots, d\}$; as a consequence, computing the PGF of the transient number of customers present reduces to elementary matrix multiplications.

We consider a two state system, with arrival rates $\lambda_1 = 1$ and $\lambda_2 = 2$, and service rate $\mu = 1$. The probability transition matrix is

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

In Fig. 2.1 we present the results for the Poisson regime (left) and the CLT regime (right). The two graphs demonstrate the convergence behavior to the limiting curve. In the left panel we see that the Poisson regime converges very quickly (at $N = 5$ the maximum error is just $O(10^{-3})$), whereas from the right panel it is observed that in the CLT regime a substantially larger value of N is required to reach the same accuracy level.

For the CLT regime we note that the solution curve, $\bar{\delta}_1^{(N)}(\vartheta, t)$, corre-

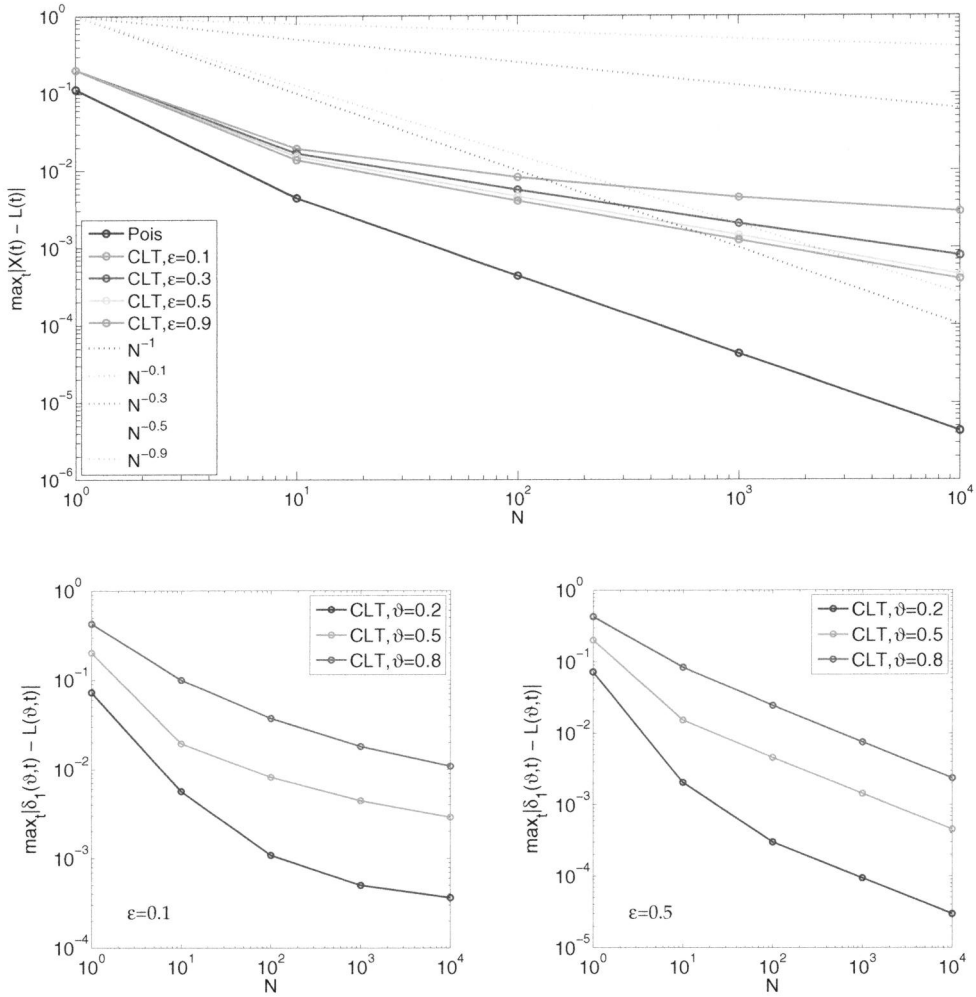


Figure 2.2: (*Top panel*) Maximum error for both regimes with varying ε as a function of N. $X(t)$ represents the solution curve δ_1 or γ_1 for the CLT and the Poisson regime, respectively. The superimposed dotted lines demonstrate the convergence rate. (*Bottom panels*) Maximum error for varying ϑ, (*left panel*) $\varepsilon = 0.1$, (*right panel*) $\varepsilon = 0.5$ in the CLT regime.

sponding to $N = 1$ exhibits large jumps. This can be explained by the specific choice of the matrix P, for which jumping between the two states is highly probable. In fact, the complementing solution curve for $\bar{\delta}_2^{(N)}(\vartheta, t)$ (not depicted here), exhibits jumps in the opposite direction at the early stages.

To get insight into the rate of convergence we look at the maximum difference between the transient PGF and MGF on the one hand, and the limiting curve on the other hand, over the computed time periods for the two limiting regimes, that is $\max_t |X_i(t) - L(t)|$, where $X_i(t)$ is $\bar{\gamma}_i^{(N)}(z, t)$ in the case of the Poisson regime, and $\bar{\delta}_i^{(N)}(\vartheta, t)$ in the case of the CLT regime, and where $L(t)$ denotes the limiting curve. The maximum difference is depicted as a function of N in the top panel of Fig. 2.2. For the CLT regime in particular we note how the convergence rate grows with ε until $\varepsilon = 0.5$; from that point on the $\sqrt{N}$ error term takes over, as noted in Section 6).

In the bottom panels of Fig. 2.2 we see the effect of the choice of ϑ in the MGF for the CLT regime. It is seen that the accuracy improves as ϑ gets smaller, but the convergence rate remains the same for all ϑ.

8 Discussion and concluding remarks

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

It is not hard to verify that state-dependent service rates can be incorporated. Some care needs to be taken, though. As mentioned in [51], in the steady-state regime we can let the service time depend on the state the background process is currently in. The analysis remains essentially the same; the μ in the definition of $h_j(z)$ is to be replaced by μ_j. Inspection of the proofs, however, reveals that it is not straightforward to incorporate this type of state-dependence in the transient cases; it is *not* hard, though, to let in these transient cases the service times depend on the state of the background process upon arrival of the particle (it essentially means that the μ in the definition of $p_i^{(N)}(z, t)$ should be replaced by μ_i).

In [51] it is indicated how the case of general transition times can be addressed in the Poisson regime. The intuition behind the argument is that the probability that a next transition occurs in a small time interval is essentially proportional to the reciprocal of the mean transition time. As a consequence, the same limiting random variables apply, but with the deterministic transition times t_i replaced by the mean transition times

$\mathbb{E}T_i$.

The proof method applied in this chapter can be used to obtain large deviation properties of the customers in the system. By establishing the existence of the limit of the appropriate moment generating function, the Gärtner-Ellis theorem [33, Thm. 2.3.6] can be applied to the random variable under consideration. A simpler variant of the model studied in this chapter, is the one in which the arrivals result from N i.i.d. Markov-modulated input streams with transition times scaled with $1/N^{\varepsilon}$ for $\varepsilon > 0$; for ease we let the system start empty. Then the crucial observation is that the number of customers present in this system at time t, say $\bar{M}^{(N)}(t)$, can be written as the sum of N i.i.d. contributions. The corresponding limiting cumulant function can easily be derived following the method of the previous sections; we eventually find for $a \geq \varrho_t$,

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

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

Chapter 3

Analysis and CLT scaling of a modulated M/G/∞ queue

This chapter analyzes several aspects of the Markov-modulated infinite server queue. In particular, service times adhere to a distribution function $F_i(\cdot)$ when the state of the background process (as seen at arrival) is i. We start by setting up explicit formulas for the mean and variance of the number of particles in the system at time $t \geq 0$, given the system started empty. The special case of exponential service times is studied in detail, resulting in a recursive scheme to compute the moments of the number of particles at an exponentially distributed time, as well as their steady-state counterparts. Then we consider an asymptotic regime in which the arrival rates are sped up by a factor N, and the transition times by a factor $N^{1+\varepsilon}$ (for some $\varepsilon > 0$). Under this scaling it turns out that the number of particles at time $t \geq 0$ obeys a central limit theorem; the convergence of the finite-dimensional distributions is proven. By assuming generally distributed service times and establishing multi-dimensional convergence, the results of this chapter are an extension of Chapter 2. The remainder of this chapter has appeared as article [21].

1 Introduction

Owing to its wide applicability and its attractive mathematical features, the infinite server queue has been intensively studied. Such a system describes units of work, e.g., particles or customers, arriving at a resource, that stay present for some random duration that is independent of other customers (in that there is no waiting). In the special case that these customers arrive according to a Poisson process with rate λ, and the sojourn times are i.i.d. random variables with finite mean $1/\mu$ — a system commonly referred to as the M/G/∞ queue — it is known that the stationary number of particles in the system has a Poisson distribution with mean

λ/μ. Also the transient behavior of such an M/G/∞ queue is well understood; see e.g. [93, p. 355].

When relaxing the model assumptions mentioned above, several interesting variants arise. In one branch of the literature, for instance, attention has been paid to the case of renewal (rather than Poisson) arrivals [46, 47]. In the present chapter, however, we consider a variant in which we introduce some sort of 'burstiness' in the arrivals and service times, using the concept of *Markov modulation*. This means that both the arrival process and the service-time distributions are driven by an external Markov process ('background process'), in the following manner. Let $X(t)$ denote an irreducible continuous-time Markov process defined on a finite state space $\{1, \ldots, d\}$. When $X(t) = i$, then the (Poissonian) arrival rate at time t equals λ_i, where $\boldsymbol{\lambda} \equiv (\lambda_1, \ldots, \lambda_d)$ is a vector with nonnegative entries. In addition, it is assumed that the time a particle remains in the system, the service time, has some general distribution with distribution function $F_i(\cdot)$ that depends on the state of the background process as seen upon arrival by the particle.

The resulting model could be called a *Markov-modulated* M/G/∞ *queue*, or an infinite server queue in a Markov-modulated environment. This type of system is relevant in a broad variety of application domains, ranging from telecommunication networks to biology. The rationale behind using this model in a communication networks setting is that the arrival rate and service times of customers may vary during the day, or on shorter timescales. In the biological context, one could think of mRNA strings being transcribed and degraded in a cell, where these transcriptions typically tend to occur in a clustered fashion; the proposed model captures the key characteristics of this mechanism well, as argued in [87].

A variety of results exist on Markov-modulated single and many server queues, whereas the literature on their infinite server counterpart is, surprisingly, considerably scarcer. In the case of a single server, the key result is that the stationary distribution of the number of customers is of matrix-geometric form [71], so this system can be viewed as a 'matrix generalization' of the normal M/M/1 queue where the stationary distribution is scalar-geometric. In [75] the stationary distribution for the case of infinitely many servers is considered; the results are in terms of the factorial moments of the number of customers. More particularly, it is shown that the corresponding distribution is *not* of matrix-Poisson type; in other words: this system is not the 'matrix generalization' of the M/M/∞, which has a scalar-Poisson distribution. A somewhat more general model that includes retrials has been studied in [59].

The case of Markov-modulated *renewal* (rather than Poisson) arrivals,

but exponential service times, is covered in [70]. Related results can be found in [67] as well, where special attention is paid to the autocorrelations in infinite server systems of various types. Steady-state results for the infinite server queue with modulated service rates have been derived in [14]. Falin [40] furthermore considers the simultaneous modulation of arrival and service rates and finds the mean number of customers in steady state.

It should also be noted that introducing burstiness using a Markovian background process is by no means the only way to incorporate a nonhomogeneous arrival rate. Willmot and Drekic [95] apply bulk arrivals with a random bulk size, whereas Economou and Fakinos [36] study arrivals generated by a compound Poisson process, both to find the transient distribution of the number of customers in the system using a generating functions based approach.

D'Auria [32] studies the same model as we do in the present chapter. Among several other results, he finds a recursion for the factorial moments of the stationary number of particles in the system. A key observation in his analysis is that the number of particles present has, in stationarity, a Poisson distribution with random parameter. Fralix and Adan [44] focus on the situation that the service times have specific phase-type distributions. In Hellings *et al.* [51] it was shown that if the transition times of the background process are sped up by a factor N, then the arrival process tends (as $N \to \infty$) to a Poisson process; the queue under consideration then essentially behaves as an M/G/∞ system.

While the above results focus on Markov-modulated infinite server queues in stationarity, there are considerably fewer results on the associated transient behavior. In [22], we studied both the transient and stationary behavior of a model similar to the one studied in the present chapter, viz. the one with exponential service times and a Markovian background process with deterministic transition times. The main focus of [22] lies on specific time scalings. In the first scaling, just the background process' transition times are sped up by a factor N; then it turns out that the distribution of the resulting queueing system converges to that of an appropriate M/M/∞ queue (which has, in steady-state, a Poisson distribution). In the second scaling, the background process jumps at a faster rate than the arrival process: the arrival rates are scaled by a factor N and the transition times by a factor $N^{1+\varepsilon}$ for some $\varepsilon > 0$. Under this scaling a central limit result was proven, for both the transient and stationary distribution.

The main contributions of this chapter are the following. In the first place we develop in Section 2 expressions for the transient mean and

variance for the number of particles in the system at time $t \geq 0$. For exponential service times the resulting expressions simplify considerably. In Section 3 we exclusively consider the special case of exponential service times: we develop a differential equation that describes the moment generating function of the number of particles in the system, and show how this differential equation facilitates the computation of moments (at an exponentially distributed time epoch, as well as in steady-state). This section also includes a recursive scheme to compute the higher moments. Section 4 considers one of the scalings studied in [22]: the arrival rates λ_i are replaced by $N\lambda_i$, while the transition times of the background Markov process are sped up by a factor $N^{1+\varepsilon}$, for some $\varepsilon > 0$, where N grows large. The objective is to prove a central limit theorem for the number of particles in the system in a finite-dimensional setting, that is, at multiple points in time. The result is established by first setting up a system of differential equations for the number of particles in the system at multiple points in time in the *non*-scaled system, then applying the scaling, and then deriving (by using Taylor approximations) a limiting differential equation (as $N \to \infty$) which eventually provides us with the claimed multivariate central limit theorem. Finally, Section 5 contains examples demonstrating analytically and numerically the results from Sections 3 and 4.

2 General results

In full detail, the model can be described as follows. Consider an irreducible continuous-time Markov process $X(t)$ on a finite state space $\{1, \ldots, d\}$, with $d \in \mathbb{N}$. $X(t)$, often referred to as the *background process*, has a transition rate matrix given by $Q = (q_{ij})_{i,j=1}^{d}$. The steady-state distribution of $X(t)$ is given by $\boldsymbol{\pi}$, a d-dimensional vector with non-negative entries summing to 1, solving $\boldsymbol{\pi}Q = \mathbf{0}$. Denote $q_i := -q_{ii} = \sum_{j \neq i} q_{ij}$.

Now consider the embedded discrete-time Markov chain that corresponds to the jump epochs of $X(t)$. It has a probability transition matrix $P = (p_{ij})_{i,j=1}^{d}$, with diagonal elements equalling 0 and $p_{ij} := q_{ij}/q_i$. Let $\hat{\pi}_i$ be the stationary probability vector at the jump epochs of $X(t)$; it solves (after normalization to 1) the linear system $\hat{\boldsymbol{\pi}} D_Q^{-1} Q = 0$, with $D_Q := \mathrm{diag}\{\boldsymbol{q}\}$. The time spent by $X(t)$ in state i, denoted T_i, is referred to as *transition time*. T_i has an exponential distribution with mean $1/q_i$. There is the following obvious relation between $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$:

$$\pi_i := \frac{\hat{\pi}_i \mathbb{E}T_i}{\sum_{j=1}^{d} \hat{\pi}_j \mathbb{E}T_j} = \frac{\hat{\pi}_i / q_i}{\sum_{j=1}^{d} \hat{\pi}_j / q_j}.$$

While the process $X(t)$ is in state i, particles arrive according to a Poisson process with rate $\lambda_i \geq 0$, for $i = 1, \ldots, d$. The service times are assumed to be i.i.d. samples distributed as a random variable F_i with mean $1/\mu_i$ if the particle was generated when the background process was in state i; the corresponding distribution function is $F_i(x) := \mathbb{P}(F_i \leq x)$, with $x \geq 0$. The service times are independent of the background process $X(t)$ and the arrival process. The system we consider is an *infinite server queue*, meaning that each particle stays in the system just for the duration of its service time (that is, there is no waiting). In the rest of this section we focus on analyzing the probabilistic properties of the number of particles in the system at given points in time, starting empty. It is assumed that the background process is in stationarity at time 0.

We start by considering a somewhat different model than the one introduced above, where the relation with our model becomes clear soon. Consider an M/G/∞ queue with (i) a *nonhomogeneous* arrival process with rate function $\lambda(\cdot)$ (such that the Poissonian arrival rate is $\lambda(s)$ at time s), and (ii) a *time dependent* distribution function $F(s, \cdot)$ (to be interpreted as the probability that a customer that arrives at time s leaves before time $t + s$ is $F(s, t)$). Observe that, conditional on the event that there are n arrivals by time t, the joint distribution of the arrival times is that of the order statistics taken from independent random variables with density

$$\frac{\lambda(s)}{\Lambda(t)} 1_{[0,t]}(s),$$

where $\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$. It now follows that if $M(t)$ is the number of particles in the system at time t, starting with an empty system, then with $\bar{F}(\cdot) := 1 - F(\cdot)$ we have that $M(t)$ has a Poisson distribution:

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

and we note for later that

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

After this general observation, we return to the initial context. Whereas we so far assumed that the input rate function and service-time distribution function were deterministic, we now assume that they are stochastic. More specifically, we use λ_i for the arrival rate when the background process $X(\cdot)$ is in state i, and $F_i(\cdot)$ for the distribution function of particles arriving while the background process is in the state i.

By conditioning on the sample path of the background process, say $X(s) = f(s)$, we find that $M(t)$ is Poisson distributed with parameter

$$\int_0^t \bar{F}_{f(t-s)}(s)\lambda_{f(t-s)}\mathrm{d}s.$$

Then by unconditioning, i.e., averaging over all paths $f(\cdot)$ of the background process, its probability generating function (PGF) equals the moment generating function (MGF) of its random parameter, evaluated at $(z-1)$:

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

see e.g. [32, p. 226]. Recalling that $X(\cdot)$ is assumed to be stationary, we have the distributional equality $\{X(t+u)|\ u \in \mathbb{R}\} \overset{\mathrm{d}}{=} \{X(u)|\ u \in \mathbb{R}\}$, so that

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

or, denoting by $\hat{X}(\cdot)$ the time-reversed version of $X(\cdot)$,

$$\begin{aligned}\mathbb{E}\, z^{M(t)} &= \mathbb{E}\exp\left(-(1-z)\int_0^t \bar{F}_{\hat{X}(s)}(s)\lambda_{\hat{X}(s)}\mathrm{d}s\right)\\ &= \mathbb{E}\exp\left(-(1-z)\int_0^t a_{\hat{X}(s)}(s)\mathrm{d}s\right),\end{aligned}$$

with $a_i(s) := \lambda_i \bar{F}_i(s)$. This probability generating function allows us to analyze the mean and variance of $M(t)$. It is immediate that the mean of $M(t)$ equals, cf. [80, Thm. 2.1],

$$\mathbb{E}M(t) = \mathbb{E}\int_0^t a_{\hat{X}(s)}(s)\mathrm{d}s = \int_0^t \mathbb{E}a_{\hat{X}(s)}(s)\mathrm{d}s = \sum_{i=1}^d \pi_i\lambda_i \int_0^t \bar{F}_i(s)\mathrm{d}s. \quad (3.1)$$

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

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

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

First notice that

$$\mathbb{V}\mathrm{ar}(M(t)|\hat{X}) = \mathbb{E}(M(t)|\hat{X}) = \int_0^t a_{\hat{X}(s)}(s)\mathrm{d}s,$$

because $(M(t) \mid \hat{X})$ has a Poisson distribution (as was noted above). Hence,

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

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

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

where for $u < s$

$$\begin{aligned}
\mathbb{C}\mathrm{ov}\left(1\{\hat{X}(u)=i\}, 1\{\hat{X}(s)=j\}\right) &= \pi_i \left(e^{\hat{Q}(s-u)}\right)_{ij} - \pi_i\pi_j \\
&= \pi_j \left(e^{Q(s-u)}\right)_{ji} - \pi_i\pi_j. \qquad (3.2)
\end{aligned}$$

We now make the expressions more explicit for the case that t tends to ∞. With $D_\pi = \mathrm{diag}\{\boldsymbol{\pi}\}$, Q and $\hat{Q} = D_\pi^{-1} Q^{\mathrm{T}} D_\pi^{-1}$ are the transition rate matrices of X and $\hat{X}$, respectively. Let us denote the matrix $\Sigma(s) = (\sigma_{ij}(s))_{i,j=1}^d$ through

$$\sigma_{ij}(s) := \pi_j \left(e^{Qs}\right)_{ji} - \pi_i\pi_j.$$

Letting $t \to \infty$, we obtain

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

When the service-time distributions are exponential, that is, $\bar{F}_i(t) = e^{-\mu_i t}$, so that $a_i(t) = \lambda_i e^{-\mu_i t}$, we have that

$$\mathbb{V}\mathrm{ar}\left(\int_0^\infty a_{\hat{X}(s)}(s)\mathrm{d}s\right) = 2\sum_{i,j}\frac{\lambda_i\lambda_j}{\mu_i+\mu_j}\int_0^\infty e^{-\mu_j s}\sigma_{ij}(s)\mathrm{d}s. \tag{3.3}$$

We summarize (some of) our findings.

Proposition 2.1. *The transient mean of the number of particles is*

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

whereas the stationary variance is

$$\mathbb{V}\mathrm{ar}M(\infty) = \sum_{i=1}^d \pi_i\frac{\lambda_i}{\mu_i} + 2\sum_{i,j=1}^d\int_0^\infty\int_0^\infty a_i(u)a_j(u+s)\sigma_{ij}(s)\mathrm{d}u\,\mathrm{d}s,$$

provided that the system started empty.

We finish this section by performing some explicit calculations for the case that X is reversible and exponential service times; later on we further focus on the situation of $d = 2$. Due to the reversibility, $\pi_i q_{ij} = \pi_j q_{ji}$ for all $i, j \in \{1, \ldots, d\}$. As a consequence $D_\pi Q = Q^{\mathrm{T}} D_\pi$, so that the matrix

$$D_\pi^{1/2} Q D_\pi^{-1/2}$$

is symmetric, and can be written as $G(-\Delta)G^{\mathrm{T}}$, where G is a (real-valued) orthogonal matrix, and $\Delta = \mathrm{diag}\{\boldsymbol{\delta}\}$ is a (real-valued) diagonal matrix (where it is noted that, owing to the background process' irreducibility all but one entries of $\boldsymbol{\delta}$ are strictly positive). It follows that

$$Q = (D_\pi^{-1/2}G)(-\Delta)(D_\pi^{-1/2}G)^{-1},$$

and therefore

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

It now follows that

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

is symmetric, and hence for each $i, j \in \{1, \ldots, d\}$ we can write

$$\sigma_{ij}(s) = \sum_{k=1}^{d} c_{ijk} e^{-\delta_k s} - \pi_i \pi_j.$$

Consequently,

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

In the case of $d = 2$, we have that $\pi_1 = q_2/\bar{q} = 1 - \pi_2$, with $\bar{q} := q_1 + q_2$. It is readily verified that $\delta_1 = 0$ and $\delta_2 = \bar{q}$. It requires a standard computation to verify that

$$e^{Qs} = \begin{pmatrix} \pi_1 + \pi_2 e^{-\bar{q}s} & \pi_2 - \pi_2 e^{-\bar{q}s} \\ \pi_1 - \pi_1 e^{-\bar{q}s} & \pi_2 + \pi_1 e^{-\bar{q}s} \end{pmatrix},$$

and also

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

Elementary calculus now yields that (3.3) equals

$$\frac{q_1 q_2}{\bar{q}^2} \left(\frac{\lambda_1^2}{\mu_1} \cdot \frac{1}{\bar{q} + \mu_1} + \frac{\lambda_2^2}{\mu_2} \cdot \frac{1}{\bar{q} + \mu_2} - 2\frac{\lambda_1 \lambda_2}{\mu_1 + \mu_2} \left(\frac{1}{\bar{q} + \mu_1} + \frac{1}{\bar{q} + \mu_2}\right)\right).$$

3 Exponential service times

In this section we analyze the special case of exponential service times in greater detail. We set up a system of differential equations for the moment generating function of the transient number of particles in the system. Then this is used to determine the mean and higher moments after an exponential amount of time.

We start this section with some preliminaries and additional notation. Here and in the remaining sections we denote by $M_i(t)$ the number of particles in the system at time t, conditional on the background process being in state i at time 0. It is evident that $M_i(t)$ can be written as the sum of two independent components: the number of particles still present at time t out of the original population of size x_0 (in the sequel denoted by

$\check{M}(t)$), increased by the number of particles that arrived in $(0,t]$ that is still present at time t (in the sequel denoted by $\bar{M}_i(t)$ in case the background process is in state i at time 0).

Due to the assumption that the service times are exponentially distributed, there are positive numbers μ_i, for $i = 1, \ldots, d$, such that $\bar{F}_i(t) = e^{-\mu_i t}$. In the case that the μ_i are identical (say equal to $\mu > 0$), $\check{M}(t)$ follows a binomial distribution with parameters x_0 and $e^{-\mu t}$. In the case the μ_i are not identical, we need to know the number $x_{0,i}$ particles present at time 0 that were generated while the background process was in state i. The resulting (independent) random variables $\check{M}_i(t)$ follow binomial distributions with parameters $x_{0,i}$ and $e^{-\mu_i t}$; indeed, $\check{M}(t) = \sum_i \check{M}_i(t)$.

Given these observations we concentrate in the remainder of this section on the more complicated component of $M(t)$, that is $\bar{M}_i(t)$.

3.1 Differential equation

Recall that we write, for ease of notation, $q_i := 1/\mathbb{E}T_i$, and $q_{ij} := p_{ij}q_i$ (where $i \neq j$), with $q_{ii} = -q_i$. The main quantity in this subsection is the moment generating function of $\bar{M}_i(t)$:

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

Consider a small time period Δt, and focus on all terms of magnitude $O(\Delta t)$ or larger. In our continuous-time Markov setting, the background process has either zero jumps (with probability $1 - q_i\Delta t + o(\Delta t)$), or a jump to state $j \neq i$ (with probability $q_{ij}\Delta t + o(\Delta t)$); the probability of more than one transition is $o(\Delta t)$ (see for instance [73, Thm. 2.8.2]).

Note that

$$\begin{aligned}\Lambda_i(\vartheta, t) = \sum_{k=0}^{\infty} e^{-\lambda_i \Delta t} \frac{(\lambda_i \Delta t)^k}{k!} \left(p_i(\vartheta, t)\right)^k \times \\ \left(\sum_{j \neq i} q_{ij}\Delta t\, \Lambda_j(\vartheta, t - \Delta t) + \left(1 - \sum_{j \neq i} q_{ij}\Delta t\right) \Lambda_i(\vartheta, t - \Delta t)\right) \\ + O\left((\Delta t)^2\right);\end{aligned} \tag{3.4}$$

here $p_i(\vartheta, t)$ is the MGF of a random variable distributed on $\{0, 1\}$, indicating whether a particle arriving in the time period $(0, \Delta t)$ is still present

at t. It is seen that the value 1 occurs with probability

$$\int_0^{\Delta t} \frac{1}{\Delta t} \left(\int_{t-u}^{\infty} \mu_i e^{-\mu_i v} \mathrm{d}v \right) \mathrm{d}u$$
$$= \frac{e^{-\mu_i t}}{\Delta t} \int_0^{\Delta t} e^{\mu_i u} \mathrm{d}u = \frac{e^{-\mu_i t}}{\mu_i \Delta t} (e^{\mu_i \Delta t} - 1) = e^{-\mu_i t} + O(\Delta t).$$

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

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

Now $q_i = \sum_{j \neq i} q_{ij}$ yields

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

with an additional $O\left((\Delta t)^2\right)$ term, and where the derivative is with respect to t. We have derived the following system of differential equations.

Proposition 3.1. *The MGF s $\Lambda_i(\vartheta, t)$ satisfy*

$$\lambda_i \, (e^{\vartheta} - 1) e^{-\mu_i t} \Lambda_i(\vartheta, t) = \Lambda_i'(\vartheta, t) - \sum_{j=1}^{d} q_{ij} \Lambda_j(\vartheta, t). \tag{3.5}$$

Now define $\psi_i(\alpha, \vartheta) := \int_0^{\infty} \alpha e^{-\alpha t} \Lambda_i(\vartheta, t) \mathrm{d}t$. Then, by integrating,

$$\int_0^{\infty} \alpha e^{-\alpha t} \Lambda_i'(\vartheta, t) \mathrm{d}t = \alpha(\psi_i(\alpha, \vartheta) - 1).$$

We thus obtain

$$\lambda_i(e^{\vartheta}-1)\frac{\alpha}{\alpha+\mu_i}\psi_i(\alpha+\mu_i,\vartheta)=\alpha(\psi_i(\alpha,\vartheta)-1)-\sum_{j=1}^{d}q_{ij}\psi_j(\alpha,\vartheta); \quad (3.6)$$

cf. [10, Eqn. (4.6), Cor. 1] in the one-dimensional case and [60, Thm. 3] in the network case for equations that resemble (3.5) for Markov-modulated shot-noise models. These may be viewed as continuous state-space analogues or weak limits of the infinite server queue (see [61] regarding a general framework that includes both for the network version in the non-modulated case).

3.2 Mean

To compute $\mathbb{E}\bar{M}_i(\tau_\alpha)$, with $\tau_\alpha \sim \exp(\alpha)$, we differentiate Eqn. (3.6) with respect to ϑ and let $\vartheta \downarrow 0$, thus obtaining

$$\lambda_i\frac{\alpha}{\alpha+\mu_i}\psi_i(\alpha+\mu_i,0)=\alpha\cdot\lim_{\vartheta\downarrow 0}\frac{\mathrm{d}}{\mathrm{d}\vartheta}\psi_i(\alpha,\vartheta)-\sum_{j=1}^{d}q_{ij}\cdot\lim_{\vartheta\downarrow 0}\frac{\mathrm{d}}{\mathrm{d}\vartheta}\psi_j(\alpha,\vartheta),$$

or

$$\begin{aligned}\lambda_i\frac{\alpha}{\alpha+\mu_i} &= \alpha\int_0^\infty \alpha e^{-\alpha t}\mathbb{E}\bar{M}_i(t)\mathrm{d}t-\sum_{j=1}^{d}q_{ij}\int_0^\infty \alpha e^{-\alpha t}\mathbb{E}\bar{M}_j(t)\mathrm{d}t\\ &= \alpha\mathbb{E}\bar{M}_i(\tau_\alpha)-\sum_{j=1}^{d}q_{ij}\mathbb{E}\bar{M}_j(\tau_\alpha). \quad (3.7)\end{aligned}$$

Now consider the special case that the background process is in equilibrium at time 0. It turns out that the expressions simplify significantly. We have, due to (3.7), using that $\sum_i \pi_i q_{ij}=0$,

$$\sum_{i=1}^{d}\pi_i\mathbb{E}\bar{M}_i(\tau_\alpha)=\sum_{i=1}^{d}\pi_i\lambda_i\frac{1}{\alpha+\mu_i}.$$

Laplace inversion yields that

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

in line with (3.1). Now consider steady-state behavior, that is, we let $\alpha \downarrow 0$. From the above, we obtain an expression that could as well have

been found by applying Little's law:

$$\sum_{i=1}^{d} \pi_i \mathbb{E}\bar{M}_i(\infty) = \sum_{i=1}^{d} \pi_i \frac{\lambda_i}{\mu_i}.$$

3.3 Higher moments

A second differentiation of (3.6) yields

$$2\lambda_i \frac{\alpha}{\alpha+\mu_i} \mathbb{E}\bar{M}_i(\tau_{\alpha+\mu_i}) + \lambda_i \frac{\alpha}{\alpha+\mu_i} = \alpha \mathbb{E}\bar{M}_i^2(\tau_\alpha) - \sum_{j=1}^{d} q_{ij} \mathbb{E}\bar{M}_j^2(\tau_\alpha).$$

In other words, once we know the $\mathbb{E}\bar{M}_i(\tau_\alpha)$ for all $\alpha > 0$, we can compute the associated second moment as well. Along the same lines,

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

As a consequence, these higher moments (at exponentially distributed epochs) can be recursively determined. Again there is a simplification if the background process is in equilibrium at time 0. Then we have the equation

$$\sum_{i=1}^{d} \pi_i \mathbb{E}\bar{M}_i^n(\tau_\alpha) = \sum_{i=1}^{d} \pi_i \lambda_i \frac{1}{\alpha+\mu_i} \sum_{k=0}^{n-1} \binom{n}{k} \mathbb{E}\bar{M}_i^k(\tau_{\alpha+\mu_i}).$$

For the steady-state we obtain, cf. [10],

$$\sum_{i=1}^{d} \pi_i \mathbb{E}\bar{M}_i^n(\infty) = \sum_{i=1}^{d} \pi_i \frac{\lambda_i}{\mu_i} \sum_{k=0}^{n-1} \binom{n}{k} \mathbb{E}\bar{M}_i^k(\tau_{\mu_i}).$$

4 Asymptotic normality for general service times

In this section we consider our Markov-modulated infinite server system, but, as opposed to the setting discussed in the previous section, now with *generally* distributed service times. The main result is a central limit theorem (for $N \to \infty$) under the scaling $q_{ij} \mapsto N^{1+\varepsilon} q_{ij}$ and $\lambda_i \mapsto N\lambda_i$; here

$\varepsilon > 0$. The intuitive idea behind this scaling is that the state of the background process moves at a faster time scale than the arrival processes (so that the arrival process is effectively a single Poisson process as $N \to \infty$), while this arrival process is sped up by a factor N (so that a central limit regime kicks in).

Remark 4.1. We already observed that the number $\check{M}_i^{(N)}(t)$ of particles still present at time t, out of the initial population of size Nx_0 and that arrived while the background process was in state i, is not affected by the evolution of the background process, as the departure rate has been determined upon arrival. Specializing to the case of exponential service times (with mean μ_i^{-1} if the particle under consideration had entered while the background process was in state i), the corresponding random variables have independent binomial distributions with parameters $Nx_{0,i}$ and $e^{-\mu_i t}$. $Nx_{0,i}$ denotes the number of particles present at time 0 that arrived while the background was in state i. Therefore, as $N \to \infty$,

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

For other service-time distributions the quantities $e^{-\mu_i t}$ have to be replaced by the appropriate survival probability associated with the residual lifetime of a particle that is present at time 0 and that had arrived while the background process was in i.

In light of the above, it suffices to focus on establishing a central limit theorem for the number of particles that arrived in $(0, t]$ that are still present at time t. Let, in line with earlier definitions, this number be denoted by $\bar{M}_i^{(N)}(t)$ in case the modulating process is in state i at time 0. $\Diamond$

One of the leading intuitions of this section is that, due to the fact that the timescale of the background process is faster than that of the arrival process, we can essentially replace our Markov-modulated infinite server system, as $N \to \infty$, by an M/G/∞ queue. This effectively means that, irrespective of the initial state i, $\bar{M}_i^{(N)}(t)$ can be approximated by a Poisson distribution with parameter $N\varrho_t$. The candidate for ϱ_t can be easily identified using the theory of Section 2, namely Eqn. (3.1):

$$\varrho_t := \sum_{i=1}^{d} \pi_i \lambda_i \int_0^t \bar{F}_i(s)\,\mathrm{d}s. \tag{3.8}$$

Let us now focus on identifying a candidate for the limiting covariance between $\bar{M}_i^{(N)}(t)$ and $\bar{M}_i^{(N)}(t+u)$; this is a rather elementary computation

that we include for the sake of completeness. Let $N(t)$ be the number present in an M/G/∞ queue that started off empty at time 0; the arrival rate is λ and the distribution function of the service times is denoted by $F(\cdot)$. In this system it is possible to compute the covariance between $N(t)$ and $N(t+u)$ explicitly in terms of the arrival rate and the distribution function $F(\cdot)$ of the service times. Realize that $N(t+u)$ can be written as the sum of the particles that were already present at time t and that are still present at time $t+u$ (which we denote by $N_t(t+u)$), and the ones that have arrived in $(t, t+u]$ and that are still present at time $t+u$. The latter quantity being independent of $N(t)$, we have

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

It thus suffices to compute the quantity $\mathbb{C}\mathrm{ov}(N(t), N_t(t+u))$. To this end, define

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

the first of these quantities can be interpreted as the probability that an arbitrary particle that has arrived in $[0, t)$ has already left the system at time t, the second as the probability that it is still present at time t but not at $t+u$ anymore, and the third as the probability that it is still present at time $t+u$. It now follows that

$$\begin{aligned}
\mathbb{E}N(t)\, N_t(t+u) &= \sum_{k=0}^{\infty} \sum_{\ell=0}^{k} k\ell\, \mathbb{P}(N(t) = k, N_t(t+u) = \ell) \\
&= \sum_{m=0}^{\infty} e^{-\lambda t} \frac{(\lambda t)^m}{m!} \sum_{k=0}^{m} \sum_{\ell=0}^{k} k\ell \binom{m}{k, \ell} (q^{\mathrm{A}})^{m-k} (q^{\mathrm{B}})^{k-\ell} (q^{\mathrm{C}})^{\ell},
\end{aligned}$$

which turns out to equal (after some elementary computations) $q^{\mathrm{C}}\, \lambda t + q^{\mathrm{C}}(1 - q^{\mathrm{A}})\lambda^2 t^2$. As $\mathbb{E}N(t) = (1 - q^{\mathrm{A}})\lambda t$ and $\mathbb{E}N_t(t+u) = q^{\mathrm{C}}\, \lambda t$, it follows that

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

This computation provides us with the candidate for the central limit result in the case of general service times. Define in this context, for $t_1 \leq t_2$,

$$c_{t_1,t_2} := \sum_{i=1}^{d} \pi_i \lambda_i \int_0^{t_1} \bar{F}_i(v + t_2 - t_1) \mathrm{d}v$$

(while if $t_2 < t_1$ we put $c_{t_1,t_2} = c_{t_2,t_1}$).

The following result covers the asymptotic multivariate normality. In the proof we consider the bivariate case (time epochs t and $t+u$), but the extension to a general dimension (time epochs t_1 up to t_K with, without loss of generality, $t_1 \leq \ldots \leq t_K$) is straightforward and essentially a matter of careful bookkeeping.

Theorem 4.1. *For any $\boldsymbol{\alpha} \in \mathbb{R}^K$ and $\boldsymbol{t} \in \mathbb{R}^K$ (with $t_1 \leq \ldots \leq t_K$), and general service times, as $N \to \infty$,*

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

with

$$\sigma^2 := \sum_{k=1}^{K} \alpha_k^2 \varrho_{t_k} + 2 \sum_{k=1}^{K-1} \sum_{\ell=1}^{k-1} \alpha_k \alpha_\ell c_{t_k,t_\ell}.$$

This theorem shows convergence of the finite-dimensional distributions to a multivariate Normal distribution. A next step would be to prove convergence *at the process level*, viz. convergence of

$$\left(\frac{\bar{M}_i^{(N)}(t) - N\varrho_t}{\sqrt{N}} \right)_{t \geq 0}$$

to a Gaussian process with a specific correlation structure. Such a result has been proven for the regular (that is, non-modulated) M/M/∞ queue in which the Poisson arrival rate is scaled by N; the limiting process is then an Ornstein-Uhlenbeck process — see e.g. [85]. The proofs of such weak convergence results typically consist of three steps: single-dimensional convergence, finite-dimensional convergence, and a tightness argument, where the tightness step tends to be relatively complicated. In our setup (that is, the Markov-modulated M/G/∞ queue) we have proven the first two steps; the third step (tightness) is beyond the scope of this chapter.

We prove Thm. 4.1 for the case of $K = 2$, with $t_1 = t$ and $t_2 = t + u$ (for $t, u \geq 0$); higher dimensions can be dealt with fully analogously but

these require substantially more administration. Our starting point is to set up a system of differential equations for the non-scaled process. This system is derived in the very same way as the differential equations for the univariate exponential case (see Prop. 3.1). Define, for fixed scalars α_1, α_2, and for $u \geq 0$ given,

$$\Lambda_i(\vartheta, t) := \mathbb{E} \exp(\vartheta \alpha_1 \bar{M}_i(t) + \vartheta \alpha_2 \bar{M}_i(t+u)).$$

In addition, let

$$\begin{aligned} p_i(\vartheta, t) &:= F_i(t) + e^{\vartheta \alpha_1}(F_i(t+u) - F_i(t)) + e^{\vartheta(\alpha_1+\alpha_2)}(1 - F_i(t+u)) \\ &= e^{\vartheta(\alpha_1+\alpha_2)} - (e^{\vartheta\alpha_1} - 1)F_i(t) - e^{\vartheta\alpha_1}(e^{\vartheta\alpha_2} - 1)F_i(t+u). \end{aligned}$$

Proposition 4.2. *The MGF s $\Lambda_i(\vartheta, t)$ satisfy*

$$\bar{p}_i(\vartheta, t)\Lambda_i(\vartheta, t) = \Lambda_i'(\vartheta, t) - \sum_{j=1}^{d} q_{ij}\Lambda_j(\vartheta, t),$$

where $\bar{p}_i(\vartheta, t) := \lambda_i(p_i(\vartheta, t) - 1)$.

Proof. Let $I_i(t)$ be the indicator function of the event that a particle arriving in $(0, \Delta t]$ (while the background process was in state i) is still in the system at time t, and consider the random variable $\alpha_1 I_i(t) + \alpha_2 I_i(t+u)$. Similarly to what we did earlier in this section, $\alpha_1 I_i(t) + \alpha_2 I_i(t+u)$ can be split into three contributions; one corresponding to the event that a particle that arrived in $(0, \Delta t]$ has already left the system at time t, one corresponding to the event that it is still present at time t but not anymore at time $t+u$, and finally one corresponding to the event that it is still present at time $t+u$. With some standard calculus it is readily obtained that

$$\mathbb{E} \exp\left(\vartheta \alpha_1 I_i(t) + \vartheta \alpha_2 I_i(t+u)\right) = p_i(\vartheta, t) + O(\Delta t).$$

This means that we obtain

$$\begin{aligned} \Lambda_i(\vartheta, t) = {} & \lambda_i \Delta t \cdot p_i(\vartheta, t)\Lambda_i(\vartheta, t - \Delta t) \\ & + \sum_{j \neq i} q_{ij}\Delta t \cdot \Lambda_j(\vartheta, t - \Delta t) + (1 - \lambda_i \Delta t - q_i \Delta t)\, \Lambda_i(\vartheta, t - \Delta t) + o(\Delta t). \end{aligned}$$

Now subtracting $\Lambda_i(\vartheta, t - \Delta t)$ from both sides, dividing by Δt, and letting $\Delta t \downarrow 0$ leads to the desired system of differential equations. $\square$

Proof of Thm. 4.1. Now we are ready to prove the bivariate asymptotic normality for the case of general service times. The idea behind the proof is to (i) start off with the differential equations for the non-scaled system as derived in Prop. 4.2; (ii) incorporate the scaling in the differential equations, and apply the centering and normalization corresponding to the central limit regime; (iii) use Taylor expansions (for large N); (iv) obtain a limiting differential equation (as $N \to \infty$). This limiting differential equation finally yields the claimed central limit theorem.

We first 'center' the random variable $\alpha_1 \bar{M}_i^{(N)}(t) + \alpha_2 \bar{M}_i^{(N)}(t+u)$; to this end we subtract $N\varrho(t,u)$ from this random variable, with

$$\varrho(t,u) := \alpha_1 \varrho_t + \alpha_2 \varrho_{t+u},$$

and ϱ_t defined as in Eqn. (3.8). At this point we impose the scaling, that is, we replace q_{ij} by $N^{1+\varepsilon} q_{ij}$, and λ_i by $N\lambda_i$. With these parameters, we now study the appropriately centered and scaled random variable

$$\frac{\vartheta\alpha_1 \bar{M}_i^{(N)}(t) + \vartheta\alpha_2 \bar{M}_i^{(N)}(t+u) - N\vartheta\varrho(t)}{\sqrt{N}},$$

where we suppress the argument u in $\varrho(t,u)$ (as u is held fixed throughout the proof). It means that we study the 'centered and scaled MGF'

$$\tilde{\Lambda}_i^{(N)}(\vartheta,t) := \Lambda_i\left(\frac{\vartheta}{\sqrt{N}},t\right)\exp\left(-\sqrt{N}\vartheta\varrho(t)\right), \tag{3.9}$$

where, due to Prop. 4.2, $\Lambda_i(\vartheta/\sqrt{N},t)$ satisfies

$$N\bar{p}_i\left(\frac{\vartheta}{\sqrt{N}},t\right)\Lambda_i\left(\frac{\vartheta}{\sqrt{N}},t\right) = \Lambda_i'\left(\frac{\vartheta}{\sqrt{N}},t\right) - N^{1+\varepsilon}\sum_{j=1}^{d} q_{ij}\Lambda_j\left(\frac{\vartheta}{\sqrt{N}},t\right).$$

Realize that, as a straightforward application of the chain rule,

$$\left(\tilde{\Lambda}_i^{(N)}\right)'(\vartheta,t) = \Lambda_i'\left(\frac{\vartheta}{\sqrt{N}},t\right)\exp\left(-\sqrt{N}\vartheta\varrho(t)\right) - \sqrt{N}\vartheta\varrho'(t)\tilde{\Lambda}_i^{(N)}(\vartheta,t).$$

Upon combining the above, we find a relation which is completely in terms of the centered/scaled MGF $\tilde{\Lambda}_i^{(N)}(\vartheta,t)$:

$$\begin{aligned} N\bar{p}_i\left(\frac{\vartheta}{\sqrt{N}},t\right)\tilde{\Lambda}_i^{(N)}(\vartheta,t) &= \left(\tilde{\Lambda}_i^{(N)}\right)'(\vartheta,t) \\ &\quad + \sqrt{N}\vartheta\varrho'(t)\tilde{\Lambda}_i^{(N)}(\vartheta,t) - N^{1+\varepsilon}\sum_{j=1}^{d} q_{ij}\tilde{\Lambda}_j^{(N)}(\vartheta,t). \end{aligned} \tag{3.10}$$

We now study the solution of this system of differential equations for N large by 'Tayloring' the function $\bar{p}_i(\vartheta/\sqrt{N}, t)$ with respect to N. It is an elementary exercise to check that

$$\bar{p}_i\left(\frac{\vartheta}{\sqrt{N}}, t\right) = \frac{h_{1,i}(\vartheta, t)}{\sqrt{N}} + \frac{h_{2,i}(\vartheta, t)}{N} + O(N^{-\frac{3}{2}}),$$

with

$$\begin{aligned} h_{1,i}(\vartheta, t) &:= \lambda_i \left(\vartheta\alpha_1 \bar{F}_i(t) + \vartheta\alpha_2 \bar{F}_i(t+u)\right), \\ h_{2,i}(\vartheta, t) &:= \frac{\lambda_i}{2}\left(\vartheta^2\alpha_1^2 \bar{F}_i(t) + \vartheta^2(\alpha_2(2\alpha_1+\alpha_2))\bar{F}_i(t+u)\right). \end{aligned}$$

We thus obtain the differential equation

$$\begin{aligned} &\left(\sqrt{N}\ \left(h_{1,i}(\vartheta, t) - \vartheta\varrho'(t)\right) + h_{2,i}(\vartheta, t) + O(N^{-\frac{1}{2}})\right) \tilde{\Lambda}_i^{(N)}(\vartheta, t) \\ &\quad = \left(\tilde{\Lambda}_i^{(N)}\right)'(\vartheta, t) - N^{1+\varepsilon} \sum_{j=1}^{d} q_{ij} \tilde{\Lambda}_j^{(N)}(\vartheta, t), \end{aligned}$$

or in self-evident matrix/vector notation,

$$\begin{aligned} N^{1+\varepsilon} Q \tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) &= \left(\tilde{\mathbf{\Lambda}}^{(N)}\right)'(\vartheta, t) - \sqrt{N}\left(H_1(\vartheta, t) - \vartheta\varrho'(t)\right) \tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) \\ &\quad - H_2(\vartheta, t)\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) + O(N^{-\frac{1}{2}}). \end{aligned}$$

Now premultiply this equation by $\mathcal{F} := (\Pi - Q)^{-1}$, the so-called *fundamental matrix*, where $\Pi := \mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$. It holds that $\Pi^2 = \Pi$, $\mathcal{F}\Pi = \Pi\mathcal{F} = \Pi$, and $Q\mathcal{F} = \mathcal{F}Q = \Pi - I$; see for these properties and more background on fundamental matrices and deviations matrices e.g. [31]. We then obtain

$$\begin{aligned} N^{1+\varepsilon} \tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) = {} & N^{1+\varepsilon} \Pi \tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) - \mathcal{F}\left(\tilde{\mathbf{\Lambda}}^{(N)}\right)'(\vartheta, t) \\ & + \sqrt{N}\mathcal{F}\left(H_1(\vartheta, t) - \vartheta\varrho'(t)\right) \tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) \\ & + \mathcal{F}H_2(\vartheta, t)\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) + O(N^{-\frac{1}{2}}). \end{aligned}$$

Iterating this identity once, we obtain

$$\begin{aligned} N^{1+\varepsilon} \tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) = {} & N^{1+\varepsilon} \Pi \tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) - \Pi\mathcal{F}\left(\tilde{\mathbf{\Lambda}}^{(N)}\right)'(\vartheta, t) \\ & + \sqrt{N}\mathcal{F}\left(H_1(\vartheta, t) - \vartheta\varrho'(t)\right) \Pi\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) \\ & + \mathcal{F}H_2(\vartheta, t)\Pi\tilde{\mathbf{\Lambda}}^{(N)}(\vartheta, t) + O(N^{-\frac{1}{2}}) + O(N^{-\varepsilon}). \end{aligned}$$

Now premultiply the equation by $d\boldsymbol{\pi}^{\mathrm{T}} = \mathbf{1}^{\mathrm{T}}\Pi$. Recalling the identity $\Pi\mathcal{F} = \Pi$ and noting that it follows from the definition of $\varrho(t)$ that

$$\mathbf{1}^{\mathrm{T}}\Pi\left(H_1(\vartheta,t) - \vartheta\varrho'(t)\right)\mathbf{1} = 0,$$

all $O(N^\alpha)$ terms with $\alpha > 0$ cancel. For $\lim_{N\to\infty} \boldsymbol{\pi}^{\mathrm{T}}\tilde{\boldsymbol{\Lambda}}^{(N)}(\vartheta,t) =: \tilde{\Lambda}(\vartheta,t)$ we thus obtain the following differential equation:

$$\tilde{\Lambda}'(\vartheta,t) = \left(\sum_{i=1}^{d}\pi_i h_{2,i}(\vartheta,t)\right)\tilde{\Lambda}(\vartheta,t).$$

Using the technique of separation of variables, it follows that

$$\tilde{\Lambda}(\vartheta,t) = \exp\left(\int_0^t \sum_{i=1}^{d}\pi_i h_{2,i}(\vartheta,s)\mathrm{d}s\right)\kappa(\vartheta,u),$$

or

$$\tilde{\Lambda}(\vartheta,t) = \exp\left(\frac{\vartheta^2}{2}\sum_{i=1}^{d}\pi_i\int_0^t\left(\lambda_i\left[\alpha_1^2\bar{F}_i(s) + (2\alpha_1+\alpha_2)\alpha_2\bar{F}_i(s+u)\right]\right)\mathrm{d}s\right)\kappa(\vartheta,u),$$

for some function $\kappa(\vartheta,u)$ that is independent of t. Now note that this expression should not depend on α_1 if $t = 0$. In addition, if we insert $u = 0$, then α_1 and α_2 should appear in the expression as $\alpha_1 + \alpha_2$. This enables us to identify $\kappa(\vartheta,u)$. We eventually obtain

$$\tilde{\Lambda}(\vartheta,t) \;=\; \exp\left(\frac{\vartheta^2}{2}\left(\alpha_1^2\varrho_t + 2\alpha_1\alpha_2 c_{t,t+u} + \alpha_2^2\varrho_{t+u}\right)\right), \qquad (3.11)$$

as desired. We have proven the claimed convergence.

Remark 4.2. It is remarked that the central limit theorem does not carry over to the case $\varepsilon \in (-1,0]$, as then the term of order $N^{1-2\varepsilon}$ cannot be neglected relative to the term of order $N^{1-\varepsilon}$. As a result, in that situation the variance featuring in the central limit theorem will contain the fundamental matrix $\mathcal{F}$ for these values of ε. ◊

5 Examples

5.1 Two-state model

In this example we consider the case $d = 2$, and exponential sojourn times of the background process, that is, the time spent in state i is exponential

with mean $1/q_i \in (0,\infty)$. From $\mathbb{E}\bar{\boldsymbol{M}}(\tau_\alpha) = (A(\alpha))^{-1}\boldsymbol{\varphi}(\alpha)$ we obtain for the mean number in the system after an exponential time with mean $1/\alpha$ (ignoring the effect of an initial population)

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

When sending α to ∞, we indeed obtain that $\mathbb{E}\bar{M}_i(\tau_\infty) = 0$; when sending α to 0, the resulting formula is consistent with the long-term mean number in the system, as found earlier. Replacing q_i by Nq_i (for $i = 1, 2$), we obtain that both components of $\mathbb{E}\bar{\boldsymbol{M}}(\tau_\alpha)$ converge (as $N \to \infty$) to

$$\pi_1\frac{\lambda_1}{\alpha+\mu_1} + \pi_2\frac{\lambda_2}{\alpha+\mu_2},$$

which is for $\mu_1 = \mu_2$ in line with the findings in [51].

We now focus on computing the second moment; for ease we consider the stationary case. From Section 3.3, we have

$$\sum_{i=1}^{d}\frac{2\pi_i\lambda_i}{\alpha+\mu_i}\mathbb{E}\bar{M}_i(\tau_{\alpha+\mu_i}) + \sum_{i=1}^{d}\frac{\pi_i\lambda_i}{\alpha+\mu_i} = \sum_{i=1}^{d}\pi_i\mathbb{E}\bar{M}_i^2(\tau_\alpha),$$

which becomes after sending α to 0,

$$\mathbb{E}\bar{M}^2(\infty) := \sum_{i=1}^{d}\pi_i\mathbb{E}\bar{M}_i^2(\infty) = \sum_{i=1}^{d}2\pi_i\frac{\lambda_i}{\mu_i}\mathbb{E}\bar{M}_i(\tau_{\mu_i}) + \sum_{i=1}^{d}\pi_i\frac{\lambda_i}{\mu_i};$$

obviously, $\pi_1 = 1-\pi_2 = q_2/(q_1+q_2)$.

We now find a lower bound on the variance of the stationary number of particles in the system. Restricting ourselves to the case $\mu_i \equiv \mu$ for $i = 1,\ldots,d$, elementary computations yield

$$\begin{aligned}\mathbb{E}\bar{M}^2(\infty) = {}& \frac{\pi_1 r_1}{\mu-q}\left((\mu-q_2)r_1 - q_1 r_2\right) + \pi_1 r_1 \\ &+ \frac{\pi_2 r_2}{\mu-q}\left((\mu-q_1)r_2 - q_2 r_1\right) + \pi_2 r_2,\end{aligned}$$

with $r_i := \lambda_i/\mu$ and $q := q_1 + q_2$. We now claim that, with R denoting the stationary mean $\pi_1 r_1 + \pi_2 r_2$, the stationary variance is larger than this R, or equivalently

$$\mathbb{E}\bar{M}^2(\infty) \geq R^2 + R, \tag{3.12}$$

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

$$a^2(f_1 - \pi_1)\pi_1 + a(f_2 - \pi_2)\pi_1 + a(g_1 - \pi_1)\pi_2 + (g_2 - \pi_2)\pi_2 \geq 0, \tag{3.13}$$

with equality only if $a = 1$; here

$$f_1 = 1 - f_2 := \frac{\mu - q_2}{\mu - q}, \quad g_2 := 1 - g_1 := \frac{\mu - q_1}{\mu - q}.$$

Observe that $f_1 > \pi_1$, so that the left-hand side of (3.13) has a minimum. Now realize that $f_1 - \pi_1 = -(f_2 - \pi_2)$ and $g_2 - \pi_2 = -(g_1 - \pi_1)$. As a result, (3.13) reduces to

$$(a-1)(a(f_1 - \pi_1)\pi_1 - (g_2 - \pi_2)\pi_2) \geq 0,$$

which, due to $(f_1 - \pi_1)\pi_1 = (g_2 - \pi_2)\pi_2$, can be rewritten as

$$(f_1 - \pi_1)\pi_1(a-1)^2 \geq 0.$$

Claim (3.12) thus follows. We conclude that $\mathbb{V}\mathrm{ar}\bar{M}(\infty) \geq \mathbb{E}\bar{M}(\infty)$, with equality if and only if $\lambda_1 = \lambda_2$.

This result can be intuitively understood. As argued before, $\bar{M}(\infty)$ is distributed as a Poisson random variable with a *random* parameter. In the introduction of [51], it was shown with an elementary argument that this entails that $\mathbb{V}\mathrm{ar}\bar{M}(\infty) \geq \mathbb{E}\bar{M}(\infty)$; informally, this says that Markov modulation increases the variability of the stationary distribution. We have now shown that for $d = 2$ this inequality is in fact strict, unless the λ_i match (and equal, say λ). In fact, then the queue is just an M/M/∞ system which has the Poisson(λ/μ) distribution as the equilibrium distribution, for which mean and variance coincide (and have the value λ/μ). In other words, for $d = 2$ there are no other ways to obtain a Poisson stationary distribution than letting all λ_i be equal.

5.2 Computational results

We include computational results demonstrating the converging behavior of the two-state scaled process in one dimension (i.e., $K = 1$ in Thm. 4.1). Unscaled, the parameters are $\boldsymbol{\lambda} = (1, 2)$, $\boldsymbol{\mu} = (1, 1)$, and $\boldsymbol{q} = (1, 3)$.

Depicted in Figure 3.1 is the limiting behavior of Eqn. (3.9) assuming exponential service times, obtained by solving the scaled version of the differential equation (3.5) with the MGF parameter $\vartheta = 0.5$ and $\varepsilon = 0.5$. The corresponding limiting curve from Eqn. (3.11) is plotted as well. As in the case with deterministic transition times [22], we observe loglinear convergence, with the solution curve closely following the limiting curve for $N = 1000$. Tweaking the parameters results in the same convergence behavior.

Acknowledgments

The authors like to thank Koen de Turck (Ghent University) for helpful discussions.

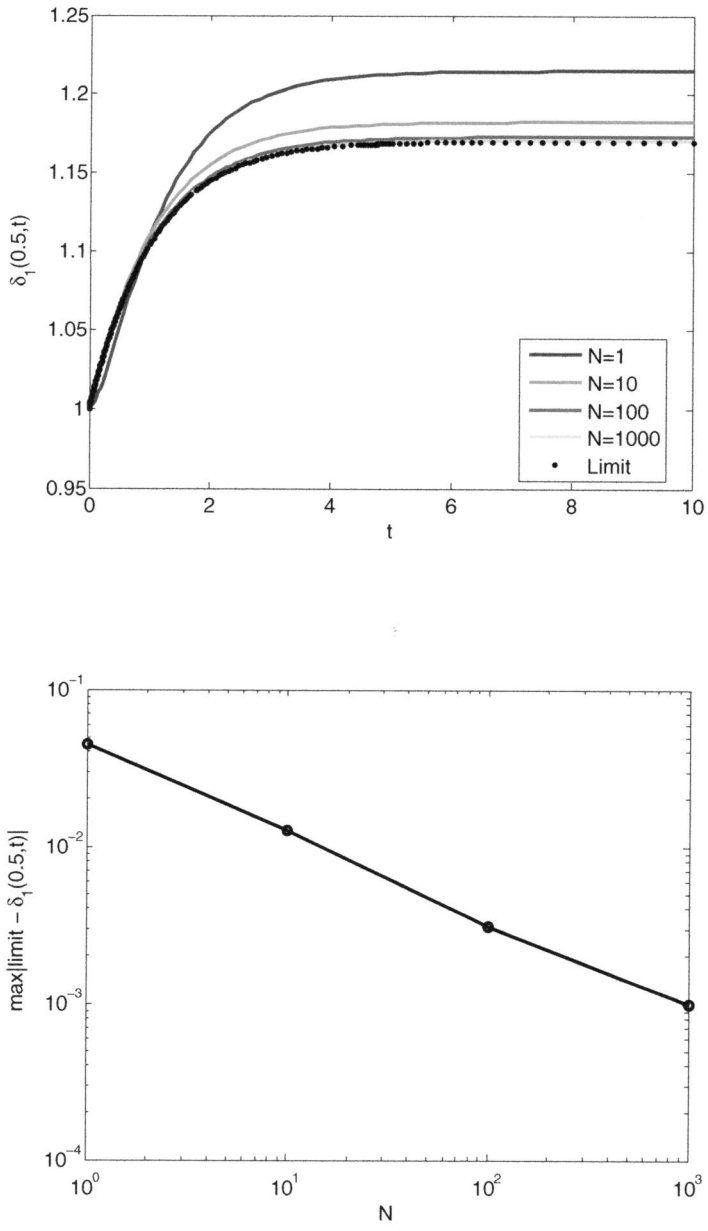


Figure 3.1: (above) The scaled process, $\tilde{\Lambda}^{(N)}(0.5, t)$, approaches the limiting curve as N grows larger. (below) Maximum error as a function of N shows loglinear convergence.

Chapter 4

An FCLT using martingale theory: fast and slow environment

We consider a model in which the production of new molecules in a chemical reaction network occurs in a stochastic fashion, modeled as a Poisson process with a varying rate. It is assumed that molecules decay after an exponential time with a uniform rate. The quantity of interest is distribution of the number of molecules in the system, under a time-scaling similar to that seen in Chapters 2 and 3; namely the background process is sped up by a factor N^α, for some $\alpha > 0$, whereas the arrival rates are multiplied by N, for N large. Note however, that whereas in Chapters 2 and 3 the condition was that $\alpha > 1$, the results in this chapter also include the case when $0 < \alpha \leq 1$. The main result is a functional central limit theorem (FCLT), in that the number of molecules, after centering and scaling, converges to an Ornstein-Uhlenbeck process. An interesting dichotomy is observed: (i) if $\alpha > 1$ the background process jumps faster than the arrival process, and consequently the arrival process behaves essentially as a (homogeneous) Poisson process, so that the scaling in the FCLT is the usual $\sqrt{N}$, whereas (ii) for $\alpha \leq 1$ the background process is relatively slow, and the scaling in the FCLT is $N^{1-\alpha/2}$. In the latter regime, the parameters of the limiting Ornstein-Uhlenbeck process contain the deviation matrix associated with the background process $J(\cdot)$. The remainder of this chapter has appeared as article [4].

1 Introduction

When modeling chemical reaction networks within cells, the dynamics of the numbers of molecules of the various types are often described by deterministic differential equations. These models ignore the inherent

stochasticity that may play a role, particularly when the number of molecules is relatively small. To remedy this, the use of stochastic representations of chemical networks has been advocated, see e.g. [6, 30, 45].

In this chapter we use the formulation as in [5, 13] where the numbers of molecules evolve as a continuous-time Markov chain. A concise description of this formulation is the following, with our specific model more formally developed in Section 2. Consider a model consisting of a finite number, ℓ, of species and a finite number, K, of reaction channels. We let $\boldsymbol{M}(t)$ be the ℓ-dimensional vector whose ith component gives the number of molecules of the ith species present at time t. For the kth reaction channel we denote by $\boldsymbol{\nu}_k \in \mathbb{Z}^{\ell}_{\geq 0}$ the number of molecules of each species needed for the reaction to occur, and by $\boldsymbol{\nu}'_k \in \mathbb{Z}^{\ell}_{\geq 0}$ the number produced. We let $\mu_k(\boldsymbol{x})$ denote the rate, or intensity (termed *propensity* in the biology literature), at which the kth reaction occurs when the numbers of molecules present equals the vector $\boldsymbol{x}$. Then, $\boldsymbol{M}(t)$ may be represented as the solution to the (vector-valued) equation

$$\boldsymbol{M}(t) = \boldsymbol{M}(0) + \sum_{k=1}^{K} (\boldsymbol{\nu}'_k - \boldsymbol{\nu}_k)\, Y_k \left(\int_0^t \mu_k(\boldsymbol{M}(s)) \mathrm{d}s \right), \tag{4.1}$$

where the stochastic processes $Y_k(\cdot)$ are independent unit-rate Poisson processes [5]. Note that if, for some $k^\star \in \{1, \ldots, K\}$, $\boldsymbol{\nu}'_{k^\star} - \boldsymbol{\nu}_{k^\star}$ equals the ith unit vector $\boldsymbol{e}_i$, then the $k^\star$th reaction channel corresponds to the external arrival of molecules of species i. For the specific situation that subnetworks operate at disparate timescales, these can be analyzed separately by means of lower dimensional approximations, as pointed out in e.g. [57].

In this chapter we study a model of the type described above, for the special case that there is just one type of molecular species (i.e., $\ell = 1$), and that there are external arrivals. The distinguishing feature is that the rate of the external input is determined by an independent continuous-time Markov chain $J(\cdot)$ (commonly referred to as the *background process*) defined on the finite state space $\{1, \ldots, d\}$. More concretely, we study a reaction system that obeys the stochastic representation

$$M(t) = M(0) + Y_1 \left(\int_0^t \lambda_{J(s)} \mathrm{d}s \right) - Y_2 \left(\mu \int_0^t M(s) \mathrm{d}s \right),$$

where $Y_1(\cdot)$ and $Y_2(\cdot)$ are independent unit-rate Poisson processes, and $\lambda_{J(s)}$ takes the value $\lambda_i \geq 0$ when the background process is in state i. Hence, in this model external molecules flow into the system according

to a Poisson process with rate λ_i when the background process $J(\cdot)$ is in state i, while each molecule decays after an exponentially distributed time with mean μ^{-1} (independently of other molecules present).

The main result of the chapter is a functional central limit theorem (FCLT) for the process $M(t)$, where we impose a specific scaling on the transition rates $Q = (q_{ij})_{i,j=1}^{d}$ of the background process $J(t)$, as well as on the external arrival rates $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^{\mathrm{T}}$ (note that all vectors are to be understood as column vectors). More precisely, the transition rates of the background process are sped up by a factor N^{α}, with $\alpha > 0$, while the arrival rates are sped up linearly, that is, they become $N\lambda_i$. Then we consider the process $U^N(t)$ (with a superscript N to stress the dependence on N), obtained from $M^N(t)$ by centering, that is, subtracting the mean $\mathbb{E}M^N(t)$, and normalizing, that is, dividing by an appropriate polynomial in N. It is proven that $U^N(t)$ converges (as $N \to \infty$) weakly to a specific Gauss-Markov process, viz. an Ornstein-Uhlenbeck (OU) process with certain parameters (which are given explicitly in terms of $\boldsymbol{\lambda}$, μ, and the matrix Q). Our proofs are based on martingale techniques; more specifically, an important role is played by the martingale central limit theorem.

Interestingly, if $\alpha > 1$ the normalizing polynomial in the FCLT is the usual $\sqrt{N}$, but for $\alpha \leq 1$ it turns out that we have to divide by $N^{1-\alpha/2}$. The main intuition behind this dichotomy is that for $\alpha > 1$ the timescale of the background process is faster than that of the arrival process, and hence the arrival process is effectively a (homogeneous) Poisson process. As a result the corresponding FCLT is in terms of the corresponding Poisson rate (which we denote by $\lambda_\infty := \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{\lambda}$, where $\boldsymbol{\pi}$ is the stationary distribution of the background process) and μ only. For $\alpha \leq 1$, on the contrary, the background process jumps relatively slowly; the limiting OU process is in terms of $\boldsymbol{\lambda}$ and μ, but features the deviation matrix [31] associated to the background process $J(\cdot)$ as well.

In earlier works [18, 21] a similar setting was studied. However, where we use a martingale-based approach in the present chapter, in [18, 21] another technique was applied: (i) we set up a system of differential equations for the Laplace transform of $M^N(t)$ jointly with the state of the background process $J^N(t)$, (ii) modified these into a system of differential equations for the transform of the (centered and normalized) process $U^N(t)$ jointly with $J^N(t)$, (iii) approximated these by using Taylor expansions, and (iv) then derived an ordinary differential equation for the limit of the transform of $U^N(t)$ (as $N \to \infty$). Since this differential equation yielded a Normal distribution, the CLT was established. Importantly, the

results derived in [18, 21] crucially differ from the ones in the present chapter. The most significant difference is that those results are *no* FCLT: just the *finite-dimensional convergence* to the OU process was established, rather than convergence at the process level (i.e., to prove weak convergence an additional 'tightness argument' would be needed). For the sake of completeness, we mention that [21] covers just the case $\alpha > 1$, that is, the regime in which the arrival process is effectively Poissonian, while [18] allows all $\alpha > 0$.

The previous line of work in [18, 21, 22] was presented in the language of *queueing theory*; the model described above can be seen as an infinite server queue with Markov-modulated input. In comparison to Markov-modulated single server queues (and to a lesser extent Markov-modulated many server queues), this infinite server model has been much less intensively studied. This is potentially due to the fact that the presence of infinitely many servers may be considered less realistic, perhaps rightfully so in the context of operational research, the major application field of queueing theory. In the context of chemical reactions, however, it can be argued that the concept of infinitely many servers is quite natural: each molecule brings its own 'decay'-server.

We conclude this introduction with a few short remarks on the relation of our work with existing literature. Incorporating Markov modulation in the external arrival rate the infinite server queue becomes, from a biological perspective, a more realistic model [87]. It is noted that deterministic modulation has been studied in [37] for various types of non-homogeneous arrival rate functions ($\mathrm{M}_t/\mathrm{G}/\infty$ queue). For earlier results on the stationary distribution of Markov-modulated infinite server queues, for instance in terms of a recursive scheme that determines the moments, we refer to e.g. [32, 44, 59, 75].

As mentioned above, at the methodological level, our work heavily relies on the so-called martingale central limit theorem (MCLT), see for instance [39, 94]. It is noted that convergence to OU has been established in the non-modulated setting before: an appropriately scaled $\mathrm{M}/\mathrm{M}/\infty$ queue weakly converges to an OU process. For a proof, see e.g. [85, Section 6.6]; cf. also [23, 52].

The rest of this chapter is organized as follows. In Section 2 we set up the model, its properties and quantities of interest, and present the essential mathematical tools. The N^α-scaled background process is thoroughly investigated in Section 3; most notably we derive its FCLT relying on the MCLT. This takes us to Section 4, where we first show that

$\bar{M}^N(t) := N^{-1}M^N(t)$ converges to a deterministic solution, denoted by $\varrho(t)$, to finally establish the FCLT for $M^N(t)$ by proving asymptotic normality of the process $N^\beta\,(\bar{M}^N(t) - \varrho(t))$, with $\beta \in (0, 1/2)$ appropriately chosen. As indicated earlier, the parameters specifying the limiting OU process depend on which speedup is 'faster': the one corresponding to the background process (i.e., $\alpha > 1$) or that of the arrival rates (i.e., $\alpha < 1$). We conclude this chapter by a set of numerical experiments, that illustrate the impact of the value of α.

2 The model and mathematical tools

In this section we first describe our model in detail, and then present preliminaries (viz. a version of the law of large numbers for Poisson processes and the MCLT).

Model. This chapter considers the following Markovian model. Let $J(t)$ be an irreducible continuous-time Markov process on the finite state space $\{1, \ldots, d\}$. Define its generator matrix by $Q = (q_{ij})_{i,j=1}^d$ and the (necessarily unique) invariant distribution by $\boldsymbol{\pi}$; as a consequence, $\boldsymbol{\pi}^{\mathrm{T}}Q = \mathbf{0}^{\mathrm{T}}$. Let $X_i(t)$ be the indicator function of the event $\{J(t) = i\}$, for $i = 1, \ldots, d$; in other words: $X_i(t) = 1$ if $J(t) = i$ and 0 otherwise. It is assumed that $J(\cdot)$ is in stationarity at time 0 and hence at any t; we thus have $\mathbb{P}(J(t) = i) = \pi_i$. As commonly done in the literature, the transient distribution $\mathbb{P}(J(s) = j \mid J(0) = i)$ is denoted by $p_{ij}(s)$ and is computed as $(e^{Qs})_{i,j}$.

The model considered in this chapter is a so-called *Markov-modulated infinite server queue*. Its dynamics can be described as follows. For any time $t \geq 0$, molecules arrive according to a Poisson process with rate λ_i if $X_i(t) = 1$. We let the service/decay rate of each molecule be μ irrespective of the state of the background process. There are infinitely many servers so that the molecules' sojourn times are their service times; the molecules go in service immediately upon arrival. Throughout this chapter, $M(t)$ denotes the number of molecules present at time t.

Scaling. In this chapter an FCLT under the following scaling is investigated. The background process as well as the arrival process are sped up, while the service-time distribution remains unaffected. More specifically, the transition matrix of the background process becomes $N^\alpha Q$ for some $\alpha > 0$, while the arrival rates, λ_i for $i = 1, \ldots, d$, are scaled linearly (i.e., become $N\lambda_i$); then N is sent to ∞. To indicate the fact that they depend on the scaling parameter N, we write in the sequel $J^N(t)$ for the back-

ground process, $X_i^N(t)$ for the indicator function associated with state i of the background process at time t, and $M^N(t)$ for the number of molecules in the system at time t. Later in the chapter we let the transitions of the background process go from being sublinear (i.e., $\alpha < 1$) to superlinear (i.e., $\alpha > 1$); one of our main findings is that there is a *dichotomy*, in the sense that there is crucially different behavior in these two regimes, with a special situation at the boundary, i.e., $\alpha = 1$.

The above model can be put in terms of a chemical reaction network, as formulated in the introduction. It turns out to be convenient to do so by interpreting the background process as a model for a single molecule transitioning between d different states, with $X_i^N(t)$ denoting the number of molecules in state i at time t. Since there is at most one such molecule, we see $X_i^N(t) \in \{0, 1\}$. The following table informally summarizes the relevant reactions and corresponding intensity functions for the model of interest:

Reaction	Intensity function	Description
$X_i \to X_j$ (for $i \neq j$)	$N^\alpha q_{ij} X_i^N(t)$	$J(\cdot)$ jumps from i to j
$\emptyset \to M$	$\sum_{i=1}^d N\lambda_i X_i^N(t)$	Arrival
$M \to \emptyset$	$\mu M^N(t)$	Departure

As mentioned above, it is assumed that $\alpha > 0$; in addition, $q_{ij} \geq 0$ for $i \neq j$ (while $Q\mathbf{1} = \mathbf{0}$), $\lambda_i \geq 0$, and $\mu > 0$. The dynamics can be phrased in terms of the stochastic representation framework, as described in the introduction. In the first place, the evolution of the indicator functions can be represented as

$$\begin{aligned} X_i^N(t) = X_i^N(0) &- \sum_{\substack{j=1 \\ j\neq i}}^d Y_{i,j}\left(N^\alpha q_{ij}\int_0^t X_i^N(s)\mathrm{d}s\right) \\ &+ \sum_{\substack{j=1 \\ j\neq i}}^d Y_{j,i}\left(N^\alpha q_{ji}\int_0^t X_j^N(s)\mathrm{d}s\right) \end{aligned} \tag{4.2}$$

where the $Y_{i,j}$ ($i, j = 1, \ldots, d$ with $i \neq j$) are independent unit-rate Poisson processes. It is readily verified that if the $X_i^N(0)$, with $i = 1, \ldots, d$, are indicator functions summing up to 1, then so are the $X_i^N(t)$ for any $t \geq 0$. The second (third, respectively) term in the right-hand side represents the number of times that $J^N(\cdot)$ leaves (enters) state i in $[0, t]$.

In the second place, the number of molecules in the system evolves as

$$M^N(t) = M^N(0) + Y_1\left(N\int_0^t \sum_{i=1}^d \lambda_i X_i^N(s)\mathrm{d}s\right) - Y_2\left(\mu\int_0^t M^N(s)\mathrm{d}s\right), \tag{4.3}$$

where Y_1, and Y_2 are independent unit-rate Poisson processes (also independent of the $Y_{i,j}$).

The objective of this chapter is to describe the limiting behavior of the system as $N \to \infty$, for different values of α. Our main result is an FCLT for the process $M^N(\cdot)$; to establish this, we also need an FCLT for the *state frequencies* of $J^N(\cdot)$ on $[0,t]$, defined as

$$\boldsymbol{Z}^N(t) = (Z_1^N(t), \ldots, Z_d^N(t))^{\mathrm{T}}, \quad \text{with} \quad Z_i^N(t) := \int_0^t X_i^N(s)\mathrm{d}s.$$

This chapter essentially makes use of two more or less standard 'tools' from probability theory: the law of large numbers applied to Poisson processes and the martingale central limit theorem (MCLT). For the sake of completeness, we state the versions used here.

Lemma 2.1. [5, Thm. 2.2] *Let Y be a unit rate Poisson process. Then for any $U > 0$,*

$$\lim_{N\to\infty} \sup_{0\le u\le U} \left|\frac{Y(Nu)}{N} - u\right| = 0,$$

almost surely.

The following is known as (a version of) the MCLT, and is a corollary to Thm. 7.1.4 and the proof of Thm. 7.1.1 in [39]. Here and in the sequel, '$\Rightarrow$' denotes weak convergence; in addition, $[\cdot,\cdot]_t$ is the quadratic covariation process.

Theorem 2.2. *Let $\{\boldsymbol{\mathcal{M}}^N\}$, for $N \in \mathbb{N}$, be a sequence of $\mathbb{R}^d$-valued martingales with $\boldsymbol{\mathcal{M}}^N(0) = \mathbf{0}$ for any $N \in \mathbb{N}$. Suppose*

$$\lim_{N\to\infty} \mathbb{E}\left[\sup_{s\le t} |\boldsymbol{\mathcal{M}}^N(s) - \boldsymbol{\mathcal{M}}^N(s-)|\right] = \mathbf{0}, \text{ with } \boldsymbol{\mathcal{M}}^N(s-) := \lim_{u\uparrow s} \boldsymbol{\mathcal{M}}^N(u),$$

and, as $N \to \infty$,

$$[\mathcal{M}_i^N, \mathcal{M}_j^N]_t \to C_{ij}(t)$$

for a deterministic matrix $C_{ij}(t)$ that is continuous in t, for $i,j = 1,\ldots,d$ and for all $t > 0$. Then $\boldsymbol{\mathcal{M}}^N \Rightarrow \boldsymbol{W}$, where $\boldsymbol{W}$ denotes a Gaussian process with independent increments and $\mathbb{E}\left[\boldsymbol{W}(t)\boldsymbol{W}(t)^{\mathrm{T}}\right] = C(t)$ (such that $\mathbb{E}[W_i(t)\,W_j(t)] = C_{ij}(t)$).

There is an extensive body of literature on the MCLT; for more background, see e.g. [56, 81, 94].

3 A functional CLT for the state frequencies

In this section we establish the FCLT for the integrated background processes $\boldsymbol{Z}^N(t)$, that is, the state frequencies of the Markov process $J^N(\cdot)$ on $[0,t]$. This FCLT is a crucial element in the proof of the FCLT of $M^N(t)$, as will be given in the next section. It is noted that there are several ways to establish this FCLT; we refer for instance to the related weak convergence results in [16, 66], as well as the nice, compact proof for the single-dimensional convergence in [9, Ch. II, Thm. 4.11]. We chose to include our own derivation, as it is straightforward, insightful and self-contained, while at the same time it also introduces a number of concepts and techniques that are used in the MCLT-based proof of the FCLT for $M^N(t)$ in the next section.

We first identify the corresponding law of large numbers. To this end, we consider the process $\boldsymbol{X}^N(t)$ by dividing both sides of (4.2) by N^α and letting $N \to \infty$. Since $X_i^N(t) \in \{0,1\}$ for all t, the $X_i^N(t)$ and $X_i^N(0)$ terms both go to zero as $N \to \infty$. Thus we may apply Lemma 2.1 to see that almost surely, as $N \to \infty$,

$$-\sum_{j \neq i} q_{ij} Z_i^N(t) + \sum_{j \neq i} q_{ji} Z_j^N(t) \to 0,$$

or $\lim_{N\to\infty} \boldsymbol{Z}^N(t)^{\mathrm{T}} Q = \mathbf{0}^{\mathrm{T}}$. Bearing in mind that $\mathbf{1}^{\mathrm{T}} \boldsymbol{Z}^N(t) = t$, the limit of $\boldsymbol{Z}^N(t)$ solves the global balance equations, entailing that

$$Z_j^N(t) \to \pi_j t \tag{4.4}$$

almost surely as $N \to \infty$, where we recall that $\boldsymbol{\pi}$ is the stationary distribution associated with the background process $J(\cdot)$.

As mentioned above, the primary objective of this section is to establish an FCLT for $\boldsymbol{Z}^N(\cdot)$ as $N \to \infty$. More specifically, we wish to identify a covariance matrix C such that, as $N \to \infty$,

$$N^{\alpha/2} \left(\boldsymbol{Z}^N(t) - \boldsymbol{\pi} t \right) \Rightarrow \boldsymbol{W}(t), \tag{4.5}$$

with $\boldsymbol{W}(\cdot)$ representing a (d-dimensional) Gaussian process with independent increments such that $\mathbb{E}\left[\boldsymbol{W}(t)\boldsymbol{W}(t)^{\mathrm{T}}\right] = C\,t$. In other words: our

goal is to show weak convergence to a d-dimensional Brownian motion (with dependent components).

We start our exposition by identifying a candidate covariance matrix C, by studying the asymptotic behavior (that is, as $N \to \infty$) of

$$\mathbb{C}\mathrm{ov}(Z_i^N(t), Z_j^N(t))$$

for fixed t. Bearing in mind that $Z_i^N(t)$ is the integral over s of $X_i^N(s)$, and using standard properties of the covariance, this covariance can be rewritten as

$$\int_0^t \int_0^s \mathbb{C}\mathrm{ov}\left(X_i^N(r), X_j^N(s)\right) \, \mathrm{d}r\mathrm{d}s + \int_0^t \int_s^t \mathbb{C}\mathrm{ov}\left(X_i^N(r), X_j^N(s)\right) \, \mathrm{d}r\mathrm{d}s.$$

Recalling that the process $J^N(\cdot)$ starts off in equilibrium at time 0, and that $X_i(s)$ is the indicator function of the event $\{J^N(s) = i\}$, this expression can be rewritten as

$$\int_0^t \int_0^s \left(\pi_i p_{ij}^N(s-r) - \pi_i \pi_j\right) \, \mathrm{d}r\mathrm{d}s + \int_0^t \int_s^t \left(\pi_j p_{ji}^N(r-s) - \pi_i \pi_j\right) \, \mathrm{d}r\mathrm{d}s,$$

where we use the notation $p_{ij}^N(s) := \mathbb{P}(J^N(s) = j \mid J^N(0) = i)$. Performing the change of variable $u := rN^\alpha$ we thus find that

$$\begin{aligned} N^\alpha \mathbb{C}\mathrm{ov}(Z_i^N(t), Z_j^N(t)) = \pi_i \int_0^t \int_0^{sN^\alpha} (p_{ij}(u) - \pi_j) \, \mathrm{d}u\mathrm{d}s \\ + \pi_j \int_0^t \int_0^{(t-s)N^\alpha} (p_{ji}(u) - \pi_i) \, \mathrm{d}u\mathrm{d}s. \end{aligned}$$

A crucial role in the analysis is played by the *deviation matrix* $D = (D_{ij})_{i,j=1}^d$ associated with the finite-state Markov process $J(\cdot)$; it is defined by

$$D_{ij} := \int_0^\infty (p_{ij}(t) - \pi_j) \mathrm{d}t; \tag{4.6}$$

see e.g. [31] for background and a survey of the main results on deviation matrices. Combining the above, we conclude that, as $N \to \infty$, with $C_{ij} := \pi_i D_{ij} + \pi_j D_{ji}$ we have identified that candidate covariance matrix, in the sense that we have shown that, for given t,

$$N^\alpha \, \mathbb{C}\mathrm{ov}(Z_i^N(t), Z_j^N(t)) \to C_{ij} t$$

as $N \to \infty$. The objective of the remainder of this section is to establish the weak convergence (4.5) with the covariance matrix $C = (C_{ij})_{i,j=1}^d$.

We now prove this weak convergence relying on the MCLT. We start by considering linear combinations of the $X_i^N(t)$ processes based on Eqn. (4.2) and introduce $\tilde{X}_i^N(t) := X_i^N(t) - X_i^N(0)$ for notational convenience. For any real constants f_i, $i = 1, \ldots, d$, we have that

$$\begin{aligned}\sum_{i=1}^{d} f_i \tilde{X}_i^N(t) &= \sum_{i=1}^{d} \sum_{j \neq i} f_i \Big(Y_{ji} \left(N^\alpha Z_j^N(t) q_{ji} \right) - Y_{ij} \left(N^\alpha Z_i^N(t) q_{ij} \right) \Big) \\ &= \sum_{i=1}^{d} \sum_{j=1}^{d} (f_j - f_i) Y_{ij} \left(N^\alpha Z_i^N(t) q_{ij} \right), \end{aligned} \tag{4.7}$$

where we do not need to define the processes Y_{ii} as the terms containing them are zero anyway. Note that the quadratic variation of this linear combination is equal to

$$\left[\sum_{i=1}^{d} f_i \tilde{X}_i^N \right]_t = \sum_{i=1}^{d} \sum_{j=1}^{d} (f_j - f_i)^2 Y_{ij}(N^\alpha Z_i^N(t) q_{ij}), \tag{4.8}$$

as the quadratic variation of a Poisson process is equal to itself. Due to Lemma 2.1 and Eqn. (4.4), we have for $N \to \infty$,

$$\left[N^{-\alpha/2} \sum_{i=1}^{d} f_i \tilde{X}_i^N \right]_t \to t \sum_{i=1}^{d} \sum_{j=1}^{d} (f_j - f_i)^2 \pi_i q_{ij}. \tag{4.9}$$

The crucial step in proving the weak convergence and consequently applying the MCLT is the identification of a suitable martingale. We prove the following lemma.

Lemma 3.1. *Let D denote the deviation matrix of the background Markov chain $J(t)$. $\boldsymbol{V}^N(t) := N^{-\alpha/2} \tilde{\boldsymbol{X}}^N(t)^T D + N^{\alpha/2} \left(\boldsymbol{Z}^N(t) - \boldsymbol{\pi} t \right)$ is an $\mathbb{R}^d$-valued martingale.*

Proof. We center our unit-rate Poisson processes by introducing

$$\tilde{Y}_{i,j}(u) := Y_{i,j}(u) - u.$$

The following algebraic manipulations are easily verified:

$$
\begin{aligned}
&N^{-\alpha/2}\sum_{i=1}^{d}\sum_{j=1}^{d}(D_{jk}-D_{ik})\tilde{Y}_{ij}\left(N^{\alpha}Z_i^N(t)q_{ij}\right)\\
&=N^{-\alpha/2}\sum_{i=1}^{d}\tilde{X}_i^N(t)D_{ik}-N^{\alpha/2}\sum_{i=1}^{d}\sum_{j=1}^{d}(D_{jk}-D_{ik})Z_i^N(t)q_{ij}\\
&=N^{-\alpha/2}\left(\tilde{\boldsymbol{X}}^N(t)^TD\right)_k-N^{\alpha/2}\sum_{i=1}^{d}\sum_{j=1}^{d}Z_i^N(t)q_{ij}D_{jk}\\
&=N^{-\alpha/2}\left(\tilde{\boldsymbol{X}}^N(t)^TD\right)_k+N^{\alpha/2}(Z_k^N(t)-\pi_kt),
\end{aligned}
$$

where we used Eqn. (4.7), the fact that $\sum_{i=1}^{d}Z_i^N(t)=t$ and the property $QD=\Pi-I$, with $\Pi=\mathbf{1}\boldsymbol{\pi}^{\mathrm{T}}$. As centered Poisson processes are martingales, and linear combinations preserve the martingale property, this concludes the proof. □

We now wish to apply the MCLT to $\boldsymbol{V}^N(t)$ as $N\to\infty$. As the second term is absolutely continuous, and the first term is a jump process with jump sizes $N^{-\alpha/2}$, we have indeed vanishing jump sizes as required by the MCLT. We now compute the covariations of $\boldsymbol{V}^N(t)$, and note that as the second term is absolutely continuous and thus does not contribute to the covariation, we have that

$$
\begin{aligned}
[V_i^N,V_j^N]_t&=N^{-\alpha}[((\tilde{\boldsymbol{X}}^N)^TD)_i,((\tilde{\boldsymbol{X}}^N)^TD)_j]_t\\
&=\frac{1}{2}N^{-\alpha}\Big([((\tilde{\boldsymbol{X}}^N)^TD)_i+((\tilde{\boldsymbol{X}}^N)^TD)_j]_t\\
&\qquad-[((\tilde{\boldsymbol{X}}^N)^TD)_i]_t-[((\tilde{\boldsymbol{X}}^N)^TD)_j]_t\Big)\\
&=\frac{1}{2}N^{-\alpha}\left(\left[\sum_{k=1}^{d}\tilde{X}_k^N(D_{ki}+D_{kj})\right]_t\right.\\
&\qquad\left.-\left[\sum_{k=1}^{d}\tilde{X}_k^ND_{ki}\right]_t-\left[\sum_{k=1}^{d}\tilde{X}_k^ND_{kj}\right]_t\right),
\end{aligned}
\tag{4.10}
$$

where we have used the polarization identity $2[X,Y]_t=[X+Y]_t-[X]_t-[Y]_t$.

Using Eqn. (4.9), this converges to

$$\begin{aligned}[V_i^N, V_j^N]_t \to \frac{1}{2}t\sum_{k=1}^{d}\sum_{\ell=1}^{d}\pi_k q_{k\ell}\times \\ \left((D_{ki}+D_{kj}-D_{\ell i}-D_{\ell j})^2-(D_{ki}-D_{\ell i})^2-(D_{kj}-D_{\ell j})^2\right) \\ = t\sum_{k=1}^{d}\sum_{\ell=1}^{d}\pi_k q_{k\ell}(D_{ki}-D_{\ell i})(D_{kj}-D_{\ell j}) \\ = t(\pi_j D_{ji}+\pi_i D_{ij}) \qquad (4.11)\end{aligned}$$

where we have used the properties $Q\mathbf{1}=0$, $\Pi Q=0$, $QD=\Pi-I$ and $\Pi D=0$. Thus, from the MCLT we have that $\boldsymbol{V}^N(t)$ converges weakly to d-dimensional Brownian motion with covariance matrix

$$C := D^{\mathrm{T}}\mathrm{diag}\{\boldsymbol{\pi}\}+\mathrm{diag}\{\boldsymbol{\pi}\}D.$$

As the first term of $\boldsymbol{V}^N(t)$ vanishes for $N\to\infty$, we have established the desired FCLT:

Proposition 3.2. *As $N\to\infty$,*

$$N^{\alpha/2}\left(\boldsymbol{Z}^N(t)-\boldsymbol{\pi}t\right)\Rightarrow \boldsymbol{W}_C(t),$$

where $\boldsymbol{W}_C(\cdot)$ is a zero-mean Gaussian process with independent increments and covariance structure $\mathbb{E}\left[\boldsymbol{W}_C(t)\boldsymbol{W}_C(t)^{\mathrm{T}}\right]=C\,t$.

4 A functional CLT for the process $M^N(t)$

Using the FCLT for the process $\boldsymbol{Z}^N(t)$, as established in the previous section, we are now in a position to understand the limiting behavior of the main process of interest, $M^N(t)$, as N grows large. As before, we begin by considering the average behavior of the quantity of interest. Dividing both sides of (4.3) by N, and denoting $\bar{M}^N(t):=N^{-1}M^N(t)$, we have

$$\bar{M}^N(t)=\bar{M}^N(0)+N^{-1}Y_1\left(N\sum_{i=1}^{d}\lambda_i Z_i^N(t)\right)-N^{-1}Y_2\left(N\mu\int_0^t\bar{M}^N(s)ds\right).$$

Assuming that $\bar{M}^N(0)$ converges almost surely to some value ϱ_0, the use of Lemma 2.1 in conjunction with Eqn. (4.4) yields that $\bar{M}^N(t)$ converges almost surely to the solution of the deterministic integral equation

$$\varrho(t)=\varrho_0+\left(\sum_{i=1}^{d}\lambda_i\pi_i\right)t-\mu\int_0^t\varrho(s)\mathrm{d}s=\varrho_0+\lambda_\infty t-\mu\int_0^t\varrho(s)\mathrm{d}s, \quad (4.12)$$

with $\lambda_\infty := \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{\lambda}$. It is readily verified that this solution is given by

$$\varrho(t) = \varrho_0 e^{-\mu t} + \frac{\lambda_\infty}{\mu}(1 - e^{-\mu t}). \tag{4.13}$$

As our goal is to derive an FCLT, we center and scale the process $M^N(\cdot)$; this we do by subtracting $N\varrho(\cdot)$, and dividing by $N^{1-\beta}$ for some $\beta > 0$ to be specified later. More concretely, we introduce the process

$$U_\beta^N(t) := N^\beta\big(\bar{M}^N(t) - \varrho(t)\big).$$

Letting $\beta > 0$ be arbitrary (for the moment), we have that due to Eqn. (4.12),

$$\begin{aligned}
&N^\beta\left(\bar{M}^N(t) - \varrho(t)\right) \\
&\quad = N^\beta\left(\bar{M}^N(0) - \varrho_0\right) - N^\beta(\varrho(t) - \varrho_0) \\
&\quad + N^\beta\left(N^{-1}Y_1\left(N\sum_{i=1}^d \lambda_i Z_i^N(t)\right) - N^{-1}Y_2\left(N\mu\int_0^t \bar{M}^N(s)\mathrm{d}s\right)\right) \\
&\quad = N^\beta(\bar{M}^N(0) - \varrho_0) - N^\beta\left(\lambda_\infty t - \mu\int_0^t \varrho(s)\mathrm{d}s\right) \\
&\quad + N^\beta\left(N^{-1}\tilde{Y}_1\left(N\sum_{i=1}^d \lambda_i Z_i^N(t)\right) - N^{-1}\tilde{Y}_2\left(N\mu\int_0^t \bar{M}^N(s)\mathrm{d}s\right)\right) \\
&\quad + N^\beta\sum_{i=1}^d \lambda_i Z_i^N(t) - N^\beta\mu\int_0^t \bar{M}^N(s)\mathrm{d}s.
\end{aligned}$$

This identity can be written in a more convenient form by defining the process

$$R_\beta(t) := \tilde{Y}_1\left(N\sum_{i=1}^d \lambda_i Z_i^N(t)\right) - \tilde{Y}_2\left(N\mu\int_0^t \bar{M}^N(s)\mathrm{d}s\right),$$

which is a martingale [5]. The resulting equation for $U_\beta^N(t)$ is

$$\begin{aligned}
U_\beta^N(t) &= U_\beta^N(0) + N^{\beta-1}R_\beta(t) \\
&\quad + N^\beta\left(\sum_{i=1}^d \lambda_i Z_i^N(t) - \lambda_\infty t\right) - \mu\int_0^t U_\beta^N(s)\mathrm{d}s.
\end{aligned} \tag{4.14}$$

We wish to establish the weak convergence of the process $U_\beta^N(t)$, as $N \to \infty$. We must simultaneously consider how to choose the parameter

β. To do so we separately inspect the terms involving $R_\beta(t)$ and $Z_i^N(t)$ in Eqn. (4.14).

First note that the sequence of martingales $\{N^{\beta-1}R_\beta(t)\}$, for $N \in \mathbb{N}$, clearly satisfies the first condition of Thm. 2.2, that of vanishing jump sizes, under the condition that $\beta < 1$, which we impose from now on. To obtain its weak limit, we compute its quadratic variation

$$\left[N^{\beta-1}R_\beta\right]_t = N^{2\beta-2}\left(Y_1\left(N\sum_{i=1}^{d}\lambda_i Z_i^N(t)\right) + Y_2\left(N\mu\int_0^t \bar{M}^N(s)\mathrm{d}s\right)\right). \tag{4.15}$$

For this term to converge in accordance with Thm. 2.2, we need $\beta \leq \frac{1}{2}$, which we impose from now on. With $\beta = \frac{1}{2}$, the term (4.15) will converge to $\lambda_\infty t + \mu\int_0^t \varrho(s)\mathrm{d}s$. Choosing $\beta < \frac{1}{2}$ will take it to zero. Turning to the Z_i^N terms of (4.14), by Prop. 3.2, and recalling that $\lambda_\infty = \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{\lambda}$, we have that for $\beta = \alpha/2$,

$$N^\beta\left(\sum_{i=1}^{d}\lambda_i Z_i^N(t) - \lambda_\infty t\right) \Rightarrow \boldsymbol{\lambda}\cdot\boldsymbol{W}_C(t), \tag{4.16}$$

which is distributionally equivalent to $W(\text{Þ}t)$, where W is a standard Brownian motion and

$$\text{Þ} := \sum_{i=1}^{d}\sum_{j=1}^{d}\lambda_i\lambda_j C_{ij}. \tag{4.17}$$

If $\beta < \alpha/2$ the term on the left-hand side of (4.16) converges to zero. Combining the above leads us to select $\beta = \min\{\alpha/2, 1/2\}$. In the sequel we distinguish between $\alpha > 1$, $\alpha < 1$, and $\alpha = 1$.

Before we treat the three cases, we first recapitulate the class of Ornstein-Uhlenbeck (OU) processes. We say that $S(t)$ is OU(a, b, c) if it satisfies the stochastic differential equation (SDE)

$$\mathrm{d}S(t) = (a - b\,S(t))\mathrm{d}t + \sqrt{c}\,\mathrm{d}W(t),$$

with $W(t)$ a standard Brownian motion. This SDE is solved by

$$S(t) = S(0)e^{-bt} + a\int_0^t e^{-b(t-s)}\mathrm{d}s + \sqrt{c}\int_0^t e^{-b(t-s)}\mathrm{d}W(s).$$

By using standard stochastic calculus it can be verified that (taking $u \leq t$)

$$\begin{aligned} \mathbb{E}S(t) &= S(0)e^{-bt} + \frac{a}{b}(1 - e^{-bt}), \\ \mathbb{V}\mathrm{ar}\, S(t) &= \frac{c}{2b}(1 - e^{-2bt}), \\ \mathbb{C}\mathrm{ov}(S(t), S(u)) &= \frac{ce^{-bu}}{2b}\left(e^{bt} - e^{-bt}\right). \end{aligned} \tag{4.18}$$

For t large, we see that

$$\mathbb{E}S(\infty) = \frac{a}{b}, \quad \mathbb{V}\mathrm{ar}\, S(\infty) = \frac{c}{2b}, \quad \lim_{t\to\infty} \mathbb{C}\mathrm{ov}(S(t), S(t+u)) = \frac{c}{2b}e^{-bu}.$$

After this intermezzo, we now treat the three cases separately.

Case 1: $\alpha > 1$

In this case we pick $\beta = 1/2$. The term (4.15) converges to

$$\sum_{i=1}^{d} \lambda_i \pi_i t + \mu \int_0^t \varrho(s)\mathrm{d}s = \lambda_\infty t + \mu \int_0^t \varrho(s)\mathrm{d}s,$$

while the term (4.16) converges to zero and is therefore neglected. Hence, $U^N_{1/2}(t)$ converges in distribution to the solution of

$$U_{1/2}(t) = U_{1/2}(0) + W\left(\lambda_\infty t + \mu \int_0^t \varrho(s)\mathrm{d}s\right) - \mu \int_0^t U_{1/2}(s)\mathrm{d}s,$$

where W is a standard Brownian motion. The above solution is distributionally equivalent to the solution of the Itô formulation of the SDE

$$U_{1/2}(t) = U_{1/2}(0) + \int_0^t \sqrt{\lambda_\infty + \mu\varrho(s)}\mathrm{d}W(s) - \mu \int_0^t U_{1/2}(s)\mathrm{d}s.$$

This SDE can be solved using standard techniques to obtain

$$U_{1/2}(t) = e^{-\mu t}\left(U_{1/2}(0) + \int_0^t \sqrt{\lambda_\infty + \mu\varrho(s)}e^{\mu s}\mathrm{d}W(s)\right).$$

We now demonstrate how to compute the variance of $U_{1/2}(t)$. To this end, first recall that by virtue of (4.13),

$$\begin{aligned} \mathbb{V}\mathrm{ar}\, U_{1/2}(t) &= \int_0^t (\lambda_\infty + \mu\varrho(s))e^{-2\mu(t-s)}\mathrm{d}s \\ &= \int_0^t \left(\lambda_\infty + \mu\left(\varrho_0 e^{-\mu s} + \frac{\lambda_\infty}{\mu}(1 - e^{-\mu s})\right)\right)e^{-2\mu(t-s)}\mathrm{d}s. \end{aligned}$$

After routine calculations, this yields

$$\mathbb{V}\mathrm{ar}\, U_{1/2}(t) = \left(\varrho_0 e^{-\mu t} + \frac{\lambda_\infty}{\mu}\right)(1 - e^{-\mu t}),$$

cf. the expressions in [21, Section 4]. In a similar fashion, we can derive that

$$\mathbb{C}\mathrm{ov}(U_{1/2}(t), U_{1/2}(t+u)) = e^{-\mu u}\left(\varrho_0 e^{-\mu t} + \frac{\lambda_\infty}{\mu}\right)(1 - e^{-\mu t}).$$

It is seen that for $t \to \infty$ the limiting process is $\mathrm{OU}(0, \mu, 2\lambda_\infty)$.

Case 2: $\alpha < 1$

In this case we pick $\beta = \alpha/2$ and the term (4.15), and therefore the term $N^{\beta-1}R_\beta(t)$, converges to zero, whereas the term (4.16) converges to $W(\text{Þ}t)$, where W is a standard Brownian motion and Þ as defined by Eqn. (4.17). Hence, $U^N_{\alpha/2}(t)$ converges weakly to the solution of

$$U_{\alpha/2}(t) = U_{\alpha/2}(0) + W(\text{Þ}t) - \mu\int_0^t U_{\alpha/2}(s)\mathrm{d}s,$$

It is straightforward to solve this equation:

$$U_{\alpha/2}(t) = e^{-\mu t}\left(U_{\alpha/2}(0) + \int_0^t \sqrt{\text{Þ}}\, e^{\mu s}\mathrm{d}W(s)\right).$$

This process has variance

$$\mathbb{V}\mathrm{ar}\, U_{\alpha/2}(t) = \int_0^t \text{Þ}e^{-2\mu(t-s)}\mathrm{d}s = \text{Þ}\frac{1 - e^{-2\mu t}}{2\mu}. \tag{4.19}$$

It is readily checked that this process is $\mathrm{OU}(0, \mu, \text{Þ})$; this is due to

$$\mathbb{C}\mathrm{ov}(U_{\alpha/2}(t), U_{\alpha/2}(t+u)) = \text{Þ}e^{-\mu u}\frac{1 - e^{-2\mu t}}{2\mu}.$$

Case 3: $\alpha = 1$

In this case we put $\beta = 1/2$, and the terms $N^{\beta-1}R_\beta(t)$ and (4.16) are of the same order. Hence, their sum converges weakly to

$$W\left(\lambda_\infty t + \mu\int_0^t \varrho(s)ds + \text{Þ}t\right),$$

where W is a standard Browian motion. In this case $U^N_{1/2}(t)$ converges weakly to the solution of

$$U_{1/2}(t) = U_{1/2}(0) + W\left(\int_0^t (\lambda_\infty + Þ + \mu\varrho(s))\,\mathrm{d}s\right) - \mu\int_0^t U_{1/2}(s)\mathrm{d}s.$$

Solving the above in a similar fashion to cases 1 and 2 yields

$$U_{1/2}(t) = e^{-\mu t}\left(U_{1/2}(0) + \int_0^t \sqrt{\lambda_\infty + Þ + \mu\varrho(s)}\, e^{\mu s}\mathrm{d}W(s)\right).$$

with the corresponding variance

$$\begin{aligned}\mathbb{V}\mathrm{ar}\, U_{1/2}(t) &= \int_0^t (\lambda_\infty + Þ + \mu\varrho(s))\, e^{-2\mu(t-s)}\mathrm{d}s \\ &= \left(\varrho_0 e^{-\mu t} + \frac{\lambda_\infty}{\mu}\right)(1 - e^{-\mu t}) + \frac{Þ}{2\mu}(1 - e^{-2\mu t})\end{aligned}$$

and covariance

$$\begin{aligned}&\mathbb{C}\mathrm{ov}(U_{1/2}(t), U_{1/2}(t+u)) \\ &\quad = e^{-\mu u}\left(\left(\varrho_0 e^{-\mu t} + \frac{\lambda_\infty}{\mu}\right)(1 - e^{-\mu t}) + \frac{Þ}{2\mu}(1 - e^{-2\mu t})\right).\end{aligned}$$

For t large this process is $\mathrm{OU}(0, \mu, 2\lambda_\infty + Þ)$.

We summarize the above results in the following theorem; it is the FCLT for $M^N(t)$ that we wished to establish. It identifies the Gauss-Markov process to which $M^N(\cdot)$ weakly converges, after centering and scaling; this limiting process behaves, modulo the effect of the initial value ϱ_0, as an OU process. More specifically, the theorem describes the limiting behavior of the centered and normalized version $U^N_\beta(\cdot)$ of $M^N(\cdot)$: the focus is on the process

$$U^N_\beta(t) = N^\beta\left(\bar{M}^N(t) - \varrho(t)\right) = \frac{M^N(t) - N\varrho(t)}{N^{1-\beta}}. \tag{4.20}$$

It is observed that for $\alpha \geq 1$, we have the usual $\sqrt{N}$ CLT-scaling; for $\alpha < 1$, however, the normalizing polynomial is $N^{1-\alpha/2}$, that is $\beta = \min\{\alpha/2, 1/2\}$.

Theorem 4.1. *As $N \to \infty$, the process $U^N_\beta(t)$ converges in distribution to the solution of*

$$U_\beta(t) = e^{-\mu t}\left(U_\beta(0) + \int_0^t \sigma(s)e^{\mu s}\,\mathrm{d}W(s)\right)$$

where

$$\sigma(s) := \begin{cases} \sqrt{\lambda_\infty + \mu\varrho(s)}, & \alpha > 1,\ \beta = 1/2; \\ \sqrt{\text{Þ}}, & \alpha < 1,\ \beta = \alpha/2; \\ \sqrt{\lambda_\infty + \text{Þ} + \mu\varrho(s)}, & \alpha = 1,\ \beta = 1/2, \end{cases}$$

W is standard Brownian motion and $\text{Þ} = \sum_{i=1}^{d}\sum_{j=1}^{d}\lambda_i\lambda_j C_{ij}$.

5 Discussion and an Example

Above we identified two crucially different scaling regimes: $\alpha > 1$ and $\alpha < 1$ (where the boundary case of $\alpha = 1$ had to be dealt with separately). In case $\alpha > 1$, the background process evolves fast relative to the arrival process, and as a consequence the arrival stream is effectively Poisson with rate $N\lambda_\infty$. When the arrival process is simplified in such a way, the system essentially behaves as an M/M/∞ queue. This regime was discussed in greater detail in e.g. [21], focusing on convergence of the finite-dimensional distributions. On the other hand, for $\alpha < 1$ the arrival rate is sped up more than the background process. Intuitively, then the system settles in a temporary (or local) equilibrium.

5.1 A two-state example

In this example we numerically study the limiting behavior of $U_\beta^N(t) := N^\beta(\bar{M}^N(t) - \varrho(t))$ with $\beta = \min\{\alpha/2, 1/2\}$ (as $N \to \infty$) in a two-state system for different α. For various values of N, we compute the moment generating function (MGF) of $U_\beta^N(t)$ by numerically evaluating the system of differential equations derived in [21, Section 3.1]. We have shown that the limiting distribution of $U_\beta^N(t)$ is Gaussian with specific parameters. We now explain how the MGF of the limiting random variable can be computed. Introducing the notation $\Lambda(t,\theta) := \mathbb{E}e^{\theta U_\beta(t)}$ and $\Lambda^N(t,\theta) := \mathbb{E}e^{\theta U_\beta^N(t)}$, it is immediate from Thm. 4.1 that we have

$$\Lambda(t,\theta) = \exp\left(\frac{\theta^2}{2}\left[\left(\varrho_0 e^{-\mu t} + \frac{\lambda_\infty}{\mu}\right)(1-e^{-\mu t})\,1_{\{\alpha\geq 1\}} + \text{Þ}\frac{(1-e^{-2\mu t})}{2\mu}\,1_{\{\alpha\leq 1\}}\right]\right), \quad (4.21)$$

In the regime that $\alpha \leq 1$, we need to evaluate the parameter Þ, which can be easily computed for $d = 2$. For the generator matrix $Q = (q_{ij})_{i,j=1}^2$, let

$q_i := -q_{ii}$ and note that $q_i > 0$. With $\bar{q} := q_1 + q_2$, the matrix exponential is given by

$$e^{Qt} = \frac{1}{\bar{q}} \begin{bmatrix} q_2 + q_1 e^{-\bar{q}t} & q_1 - q_1 e^{-\bar{q}t} \\ q_2 - q_2 e^{-\bar{q}t} & q_1 + q_2 e^{-\bar{q}t} \end{bmatrix}.$$

Since $\pi_1 = q_2/\bar{q}$ and $\pi_2 = q_1/\bar{q}$, we can now compute the components of the deviation matrix D (see Eqn. (4.6)) and covariance matrix C:

$$D = \frac{1}{\bar{q}^2} \begin{bmatrix} q_1 & q_1 \\ -q_2 & q_2 \end{bmatrix}, \quad C = \frac{2q_1q_2}{\bar{q}^3} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

From the above we find the value of Þ:

$$Þ = \frac{2q_1q_2}{\bar{q}^3}(\lambda_1 - \lambda_2)^2.$$

Fig. 4.2 illustrates the convergence of $U_\beta^N(t)$ to $U_\beta(t)$ when $\boldsymbol{q} = (1, 3)$, $\boldsymbol{\lambda} = (1, 4)$ and $\mu = 1$. We assume $\varrho_0 = 0$ and let $\theta = 0.5$. In Figs. 4.2.(a)–(c) we see the effect of α. Fig. 4.1 depicts the convergence rate, computed as

$$\max_{t \geq 0} \left| \Lambda^N(t, \theta) - \Lambda(t, \theta) \right|.$$

We observe a roughly loglinear convergence speed for $\alpha \geq 1$, whereas for $\alpha < 1$ the convergence turns out to be substantially slower.

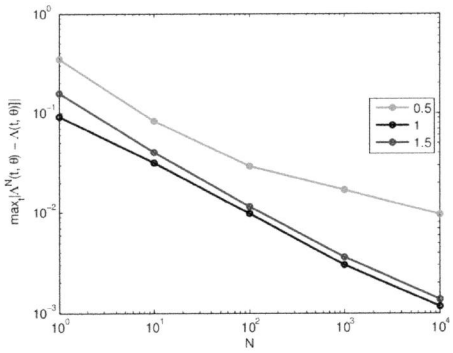
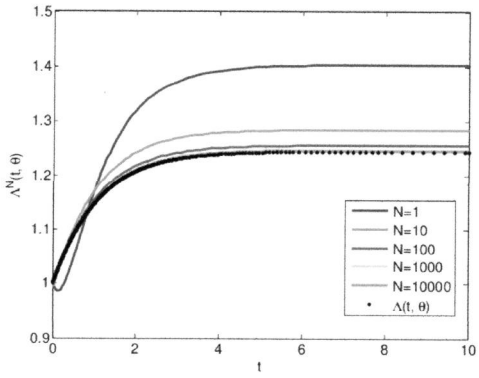


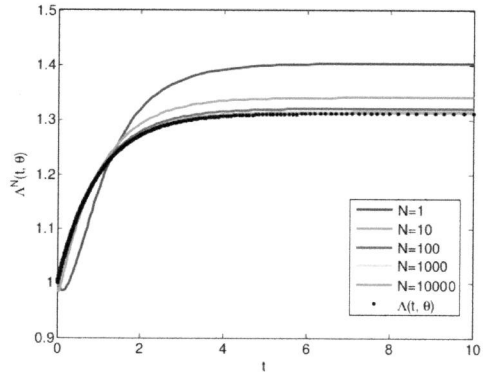
Figure 4.1: Convergence rate as a function of N; error depending on $\alpha \in \{0.5, 1, 1.5\}$.

Acknowledgments

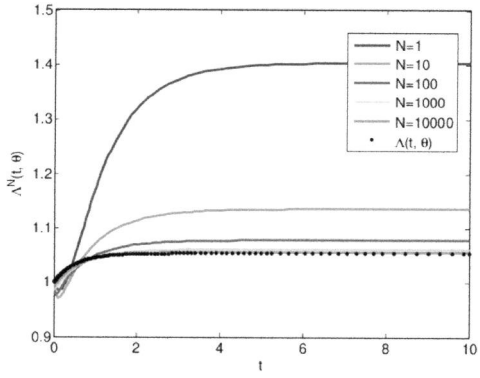
We wish to thank T. Kurtz (University of Wisconsin – Madison) for pointing us at several useful references.



(a) $\alpha = 1.5$



(b) $\alpha = 1$



(c) $\alpha = 0.5$

Figure 4.2: Convergence of the MGF of $U_\beta^N(t)$ as a function of $t \geq 0$. Panels (a)-(c) correspond to three values of α.

Chapter 5

Heavy traffic analysis: DPS and modulated workload

This chapter deals with a single-server queue with modulated arrivals, service requirements and service capacity. In the first result, we derive the mean of the total workload assuming generally distributed service requirements and any service discipline which does not depend on the modulating environment. We then show that the workload is exponentially distributed under heavy-traffic scaling. In our second result, we focus on the discriminatory processor sharing (DPS) discipline. Assuming exponential, class-dependent service requirements, we show that the joint distribution of the queue lengths of different customer classes under DPS undergoes a state-space collapse when subject to heavy-traffic scaling. That is, the limiting distribution of the queue length vector is shown to be exponential, times a deterministic vector. The distribution of the scaled workload, as derived for general service disciplines, is a key quantity in the proof of the state-space collapse. This chapter has been published as article [90].

1 Introduction

Markov-modulation is a way to formalize the embedding of queues in a random environment. The parameters of the queue in question, typically arrival rates, service requirements or both, are governed by an external Markov chain, thereby creating an extra layer of randomness around the stochastic queueing process. For classical results on Markov-modulated single-server queues with the first-come-first-serve (FCFS) discipline see e.g. [71, 79, 84]. Recent work on systems in a Markov-modulated environment can be found in for example [43, 62, 98].

In this chapter we will analyse a modulated queue under a heavy-traffic scaling, that is, evaluate the system at its critical load. It is a well-

known result from the literature of single-server queues, that under fairly general conditions [63], the steady-state distributions of the appropriately scaled queue length and workload become exponential when the critical load is approached. This property has been seen to carry over to certain systems where arrivals and service times are modulated by an external Markov process, see [7, 34, 41]. In fact, [7] establishes an even stronger result: convergence of the queue length *process* to a reflected Brownian motion. *Multi-class single-server queues* under a heavy-traffic scaling have been studied in e.g. [58] for FCFS with feedback routing, [12] for the discriminatory random order of service discipline, and [48, 91] for the discriminatory processor sharing (DPS) policy. In particular, [12, 91] show that the steady-state queue length vector undergoes a so-called *state-space collapse* and converges to an exponentially distributed variable, times a deterministic vector. The cited multi-class results under heavy-traffic scaling are all for non-modulated systems. In light of this, we will in this chapter put special emphasis on a modulated multi-class single-server queue, and the limiting steady-state queue length distribution is derived.

While there is little ambiguity in how arrival rates are modulated, there are in the literature typically two ways in which to modulate the service rates. One can (i) let the departure rate be *continuously* modulated throughout a customer's service, the other approach is to (ii) let a customer's service requirement distribution be based on the state of the environment when it arrives and remain the same until the customer departs. We note that by adapting the number of different customer classes, the fixed service requirements of case (ii) can be seen as a special case of the continuously modulated requirements (i); we further elaborate on this later in the chapter in Remark 2.1 in Section 2.

The way the load or traffic intensity for modulated queues is defined goes hand in hand with the way the service rates are affected by the environment. In case (i), the load is typically the average of arrival rates (where the averaging is with respect to the equilibrium distribution of the environment), say λ_∞, divided by the average of service rates, say μ_∞ (see e.g. [71]). In case (ii), the load is taken as the average over the arrival rate times the mean service requirement, say, λ_d/μ_d per state d of the environment (see e.g. [34, 84]). The two load definitions represent different scenarios. In particular, when load (ii) is equal to 1 (the critical load) it means that in at least one state of the environment, the total load over all classes must exceed 1, i.e. for at least one state we must have overload. This is true only for special cases of definition (i).

In this chapter, special focus will be given to a multi-class single-server

queue under the DPS discipline. The DPS discipline was first introduced by Kleinrock in [64] as an extension of the well-known egalitarian processor sharing (PS) discipline and has turned out to be very suitable to model the simultaneous parsing of diverse tasks, such as processing network data. Under this service discipline, the service capacity is divided between all present customers in proportion to their prescribed weights. Due to the challenging nature of DPS systems, most available results are in terms of limit theorems and moments. Fayolle et al. [42] established the mean sojourn time conditioned on the service requirement. That analysis also yielded the mean queue lengths of the different classes, which were shown to depend on the entire service requirement distributions of all classes. The DPS model has finite mean queue lengths irrespective of any higher-order characteristics of the service distribution, see Avrachenkov et al. [11]. This is an extension of a result for the Processor Sharing (PS) system, which holds while the queue is stable. DPS under a heavy-traffic regime was analysed in [48] assuming finite second moments of the service requirement distributions. Assuming exponential service requirement distributions, a direct approach to establish a heavy-traffic limit for the joint queue length distribution was described in [83] and extended to phase-type distributions in [91]. Combining light and heavy-traffic limits, in [55] an interpolation approximation is derived for the steady-state distribution of the queue length and waiting time of DPS. The performance of DPS in overload is considered in [3]. Asymptotics of the sojourn time have received attention in [25, 24]. Game-theoretic aspects of DPS have been studied in [96, 50]. A thorough overview of DPS results can be found in [2].

We are not aware of work analysing a DPS system under modulation. We refer to [74] where the Processor Sharing discipline (DPS discipline with unit weights) was analysed in a Markovian random environment. Multi-class queues in a random environment have been studied for different models in [27, 89]. In [89], a modulated system is studied where arrivals can only occur at transition epochs of the modulating process but service requirements are class-dependent and generally distributed. Using a time-changing argument, the waiting time distribution is derived under the FCFS discipline. In [27], the authors derive a Brownian control problem to establish a form of the $c\mu$ scheduling rule in heavy traffic under continuously modulated service requirements. By using a particular scaling, the time-scale separation of the external environment and the queue length process is exploited. Similar scaling of a modulated queue can also be seen in results on the Markov-modulated infinite server queue in e.g. [21].

The system we analyse in this chapter is a single-server queue where the arrival rates, service requirements and service capacity are modulated. We focus on the setting where a customer's service requirement distribution is based on the state of the environment when it arrives and does not change throughout its service. The service capacity is however continuously modulated. This assumption is in line with the literature for various types of modulated queues, see [26, 35, 68, 89]. In Remark 6.1 in Section 6, we discuss how part of our results can be extended to a more general model with continuously modulated service requirements. We derive the distribution of the workload under a heavy-traffic scaling for generally distributed service requirements and any service discipline which does not depend on the environment. We then turn our attention to the DPS discipline in a multi-class queue, which is a particular case of the general modulated system as described above. The weights of the DPS system determining the service proportion, depend on a customer's class and do not change with the environment, which means that the workload result remains valid. An important finding in the present chapter is that the queue length vector under DPS becomes independent of the modulating process in the heavy-traffic limit, which is consistent with the modulated M/G/1 queue in e.g. [7, 34]. This, together with the obtained result on the workload, allows us to derive the distribution of the queue length vector under DPS and to show that it undergoes a state-space collapse.

The remainder of the chapter is organized as follows. In Section 2 we describe the model. In Section 3 we study the workload of a single-server queue with modulated arrivals, service capacity and service requirement distribution, establishing its distribution in heavy traffic. From Section 4 onwards, the focus is on DPS. In Section 4 we derive some basic properties of the queue length distribution, obtain a rate conservation law and derive an equation for the moments of the queue lengths weighted with the modulated service capacity. Section 5 is devoted to the heavy-traffic scaling; there we show that the distribution of the environment becomes independent of that of the queue length vector, in addition to deriving two technical lemmas. The exponential limiting distribution of the joint queue length in heavy traffic follows in Section 6. The result is shown in two steps in the subsections 6.1 about the state-space collapse and 6.2 about the exact limiting distribution, where we rely on the workload result of Section 3. We conclude with a summary and some open questions in Section 7.

2 Model

We analyse a single-server queue modulated by an independent external environment, which is formalized by an irreducible continuous time Markov chain on a finite state-space $\{1, \ldots, D\}$. The modulating process is denoted with Z and is governed by an infinitesimal generator matrix $Q = (q_{d\ell})_{d,\ell=1}^{D}$ with an invariant distribution $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_D)$. In what follows, vectors are generally denoted in bold. New customers arrive according to a Poisson distribution with rate λ_d when the environment is in state d. A customer arriving in state d has a service requirement distribution given by a function $H_d(\cdot)$ with Laplace-Stieltjes transform (LST) $h_d(\cdot)$. The service requirement does not change further with the environment. The first and second moment are given by h_{d1} and h_{d2}, respectively. In addition we let the service capacity be scaled by a factor c_d during the environment's stay in state d, this can thus change during the service of the customers. The traffic intensity will be measured as

$$\rho_\infty = \sum_d \pi_d \lambda_d h_{d1} / c_\infty, \tag{5.1}$$

with $c_\infty := \sum_d \pi_d c_d$ being the service capacity averaged over the environment. The workload is defined as the time it takes to empty the system at an arbitrary moment in time given the observed environment and is denoted by W. In Section 3 we study the workload and the environment as a two-dimensional process (W, Z), under any service discipline that is independent of the environment.

The first main result of this chapter concerns the distribution of the workload when the traffic intensity approaches its critical point. The system is said to be in heavy traffic when ρ_∞ approaches 1. Let $N > 0$ and define the following parametrization

$$\lambda_d^{(N)} := \frac{\lambda_d}{\rho_\infty}(1 - 1/N) \to \frac{\lambda_d}{\rho_\infty} =: \hat{\lambda}_d, \quad \text{as } N \to \infty, \tag{5.2}$$

where ρ_∞ is based on the unscaled parameters. Prelimit quantities will be denoted with a superscript (N); the prelimit traffic intensity is thus $\rho_\infty^{(N)}$ and is equal to $1 - 1/N$. Limiting quantities will have a $\hat{}$; in heavy traffic the traffic intensity is denoted $\hat{\rho}_\infty$ and is equal to 1.

In the remainder of the chapter, starting in Section 4, we analyse a single-server queue with K customer classes under the discriminatory processor sharing policy; the queue is again embedded in a random environment. Let $\alpha_{k,d}$ be the probability that a customer, that arrives while

the environment is in state d, is of class k; note that $\sum_k \alpha_{k,d} = 1$ for a given d. The Poisson arrival rate of a class-k customer is denoted with $\lambda_{k,d} := \alpha_{k,d}\lambda_d$ and for each class k it is assumed that $\lambda_{k,d} > 0$ for at least one state d. In the multi-class setting we assume that a class-k customer has an exponentially distributed service requirement with mean $1/\mu_k$. We believe that the results obtained in this chapter can be extended to phase-type distributed service requirements, the latter being dense in the space of all distributions on $[0, \infty)$. For the non-modulated DPS queue, the phase-type analysis was performed in [91] using similar proof techniques. For ease of exposition, however, we focus here on the exponential case.

We no longer let the service requirement of a particular customer be environment-dependent (although the distribution of an arbitrary customer is, as explained in Section 6.2). One can however retrieve the environment-dependent service requirements by introducing additional classes for each environment, see Remark 2.1. By referring to a class-k customer's service rate while in state d as $\mu_{k,d} := \mu_k c_d$, we take the modulated service capacity into account. Most of the results for the queue length can in fact be shown without assuming this product form, representing a system where the service requirements are continuously modulated, see Remark 6.1, Section 6.1, for further details.

By $\lambda_{k,\infty} := \sum_d \lambda_{k,d}\pi_d$ we denote the average arrival rate of class-k customers, similarly we denote the average service rate for class k by $\mu_{k,\infty} := \sum_d \mu_{k,d}\pi_d = \mu_k c_\infty$ and $\rho_{k,\infty} := \lambda_{k,\infty}/\mu_{k,\infty}$. The aggregate traffic intensity for the multi-class model is defined as

$$\rho_\infty := \sum_{k=1}^{K} \rho_{k,\infty},$$

which is consistent with the definition in Eqn. (5.1).

Let the state of the multi-class system be described by the vector of random variables $(M_1, \ldots, M_K, Z) =: (\boldsymbol{M}, Z)$, where M_k is the number of class-k customers, for $k = 1, \ldots, K$. As before, Z represents the state of the background process. In a DPS system, the random fraction of service given to a class-k customer is

$$\frac{g_k}{\sum_j g_j M_j},$$

where g_k are weight parameters associated with each class k.

When in heavy traffic, we denote $\hat{\rho}_{k,\infty} := \hat{\lambda}_{k,\infty}/\mu_{k,\infty}$ and thus have for the multi-class system

$$\sum_k \hat{\rho}_{k,\infty} = \sum_k \frac{\hat{\lambda}_{k,\infty}}{\mu_{k,\infty}} = \frac{1}{\rho_\infty} \sum_k \frac{\lambda_{k,\infty}}{\mu_{k,\infty}} = \frac{1}{\rho_\infty} \sum_k \rho_{k,\infty} = 1.$$

Remark 2.1 (Modulated service requirements rewritten to classes). Any multi-class system where the service requirement distribution is determined by the modulating process at a customer's arrival, can be written as a multi-class model with non-modulated, only class-dependent, service rates μ_k, as illustrated below: While in state d of the environment, class-k customers arrive with rate $\lambda_{k,d}$ and have exponential service requirement with mean $1/\mu_{k,d}$ and weight g_k. Such customers we refer to as being of class (k, d) and count with $M_{k,d}$, hence we need to keep track of $K \cdot D$ different customer "classes". Arrivals to class (k, d) are inactive while not in state d.

classes	arrival rates		serv. rate	weight
	$d = 1$	$d = 2$	$d \in \{1, 2\}$	
(1,1)	$\lambda_{1,1}$	0	$\mu_{1,1}$	g_1
(1,2)	0	$\lambda_{1,2}$	$\mu_{1,2}$	g_1
(2,1)	$\lambda_{2,1}$	0	$\mu_{2,1}$	g_2
(2,2)	0	$\lambda_{2,2}$	$\mu_{2,2}$	g_2

Table 5.1: A multi-class system where the service requirement distribution is fixed upon arrival can be translated into one with non-modulated service requirements.

From Table 5.1 one can easily see how a $K = D = 2$ system can be written into one with $K = 4$ and $D = 2$. The arrival rates are still modulated, but in an on-off way. The service rates $\mu_{k,d}$ are now non-modulated.

3 Workload

In this section we consider the workload in a modulated queue and extend the results for an M/G/1 type queue from Falin and Falin [41] and Dimitrov [34] to include modulated service capacity. We derive the mean of the workload and then its distribution in the heavy-traffic regime.

Let $p_{0,d} = P(W = 0, Z = d)$ and let $\boldsymbol{a} = (a_1, \ldots, a_D)^T$ be a vector solving

$$[Q \cdot \boldsymbol{a}]_d = c_d - \lambda_d h_{d1} - c_\infty(1 - \rho_\infty), \tag{5.3}$$

for $d = 1, \ldots, D$. Note that such a solution always exists since the right hand side vector of Eqn. (5.3) is orthogonal to $\boldsymbol{\pi}$. We obtain the following result for the mean workload.

Proposition 3.1. *For any service requirement distribution $H_d(\cdot)$ and any service discipline that does not depend on the state of the environment, the mean of the workload satisfies*

$$\mathbb{E}W = \frac{\sum_d [\pi_d \lambda_d h_{d2}/2 + a_d \pi_d (\lambda_d h_{d1} - c_d) + p_{0,d} c_d a_d]}{c_\infty (1 - \rho_\infty)}, \tag{5.4}$$

where $\boldsymbol{a}$ is a solution of Eqn. (5.3). *Furthermore, $1 - \rho_\infty = \sum_d p_{0,d} c_d / c_\infty$.*

Remark 3.1. Although the solution vector $\boldsymbol{a}$ is not unique, the term

$$\sum_{d=1}^{D} a_d [\pi_d (\lambda_d h_{d1} - c_d) + p_{0,d} c_d],$$

as appearing in Eqn. (5.4), is. This is due to the following argument: Suppose $\boldsymbol{a}$ and $\boldsymbol{a}^*$ are two solutions to Eqn. (5.3). Then $0 = Q\boldsymbol{a} - Q\boldsymbol{a}^* = Q(\boldsymbol{a} - \boldsymbol{a}^*)$, so $(\boldsymbol{a} - \boldsymbol{a}^*)$ is in the nullspace of Q. But Q is a generator matrix so it can easily be seen that $Q\boldsymbol{r} = 0$ for any vector $r \cdot \mathbf{1}^T = (r, r, \ldots, r)^T$, $r \in \mathbb{R}$. Also, since the environment is an irreducible Markov chain, the nullspace of Q has dimension 1, and therefore, $(\boldsymbol{a} - \boldsymbol{a}^*) = r_a \cdot \mathbf{1}^T$, for some $r_a \in \mathbb{R}$. Thus,

$$\sum_{d=1}^{D} (a_d - a_d^*)[\pi_d (\lambda_d h_{d1} - c_d) + p_{0,d} c_d] = r_a c_\infty (\rho_\infty - 1) + r_a c_\infty (1 - \rho_\infty) = 0,$$

where the first term follows by definition of ρ_∞ and c_∞ and the second term comes from Eqn. (5.7) in the proof below.

Proof of Proposition 3.1. Define $F_d(x, t) = P(W(t) < x, Z(t) = d)$, for some time $t > 0$. In an infinitesimal time $\mathrm{d}t$, a new arrival requiring service x changes the workload with probability $\lambda_d H_d(x)\mathrm{d}t$. The service capacity is scaled by c_d when the environment is in state d, meaning that in $\mathrm{d}t$ time, the workload is reduced by $c_d \mathrm{d}t$, yielding by a classic birth-and-death argument for the M/G/1 queue,

$$\begin{aligned} F_d(x, t + \mathrm{d}t) = {} & (1 - \lambda_d \mathrm{d}t + q_{dd}\mathrm{d}t) F_d(x + c_d \mathrm{d}t, t) \\ & + \sum_{\ell \neq d} q_{\ell d} F_\ell(x + c_\ell \mathrm{d}t, t)\mathrm{d}t + \lambda_d \mathrm{d}t \int_0^x F_d(x + c_d \mathrm{d}t - y, t)\mathrm{d}H_d(y). \end{aligned}$$

We let $t \to \infty$ to go to steady-state and since

$$\frac{F_d(x + c_d \mathrm{d}t) - F_d(x)}{\mathrm{d}t} = c_d \frac{F_d(x + c_d \mathrm{d}t) - F_d(x)}{c_d \mathrm{d}t} \underset{\mathrm{d}t \downarrow 0}{\longrightarrow} c_d F_d'(x),$$

we obtain

$$c_d F_d'(x) = (\lambda_d - q_{dd}) F_d(x) - \sum_{\ell \neq d} q_{\ell d} F_\ell(x) - \lambda_d \int_0^x F_d(x-y) \mathrm{d}H_d(y).$$

Denote the LST of $F_d(\cdot)$ by $\varphi_d(s) = \mathbb{E}[\mathrm{e}^{-sW(t)}, Z(t) = d] = p_{0,d} + \int_{0+}^{\infty} \mathrm{e}^{-sx} \mathrm{d}F_d(x)$. The corresponding transform equation becomes

$$\sum_{\ell=1}^{D} q_{\ell d} \varphi_\ell(s) = [\lambda_d (1 - h_d(s)) - s c_d] \varphi_d(s) + s p_{0,d} c_d. \tag{5.5}$$

It is now convenient to sum over d and divide through Eqn. (5.5) with s to get zero on the left hand side, leading to

$$\sum_d p_{0,d} c_d = \sum_d \varphi_d(s) \left[c_d - \lambda_d \frac{(1 - h_d(s))}{s} \right]. \tag{5.6}$$

Using that $\varphi_d(0) = \pi_d$ and by l'Hôpital

$$\lim_{s \downarrow 0} \frac{(1 - h_d(s))}{s} = -\lim_{s \downarrow 0} h_d'(s) = h_{d1},$$

we get by taking the limit $s \to 0$ of Eqn. (5.6) that

$$\sum_d \frac{p_{0,d} c_d}{c_\infty} = 1 - \rho_\infty. \tag{5.7}$$

We differentiate Eqn. (5.6) w.r.t. s:

$$\sum_d \varphi_d(s) \lambda_d \left[\frac{h_d'(s)}{s} + \frac{1 - h_d(s)}{s^2} \right] = \sum_d \varphi_d'(s) \left[c_d - \lambda_d \frac{1 - h_d(s)}{s} \right],$$

which in the limit of $s \to 0$ results in

$$\sum_d \frac{\pi_d \lambda_d h_{d2}}{2} = \sum_d W_d \left[c_d - \lambda_d h_{d1} \right], \tag{5.8}$$

with the first moment of the workload while in state d being

$$-\lim_{s \downarrow 0} \varphi_d'(s) = \mathbb{E}[W, Z = d] =: W_d.$$

Now multiply Eqn. (5.5) with a_d, sum over d, take the derivative w.r.t. s and let $s \to 0$ to obtain

$$-\sum_d W_d[c_\infty(\rho_\infty - 1) + c_d - \lambda_d h_{d1}] = \sum_d \left[a_d \pi_d (\lambda_d h_{d1} - c_d) + p_{0,d} c_d a_d\right],$$

by using Eqn. (5.3). Adding this equation to Eqn.(5.8) yields

$$c_\infty(1-\rho_\infty)\sum_d W_d = \sum_d \left[\frac{\pi_d \lambda_d h_{d2}}{2} + a_d \pi_d(\lambda_d h_{d1} - c_d) + p_{0,d} c_d a_d\right], \quad (5.9)$$

giving the desired expression for the mean workload, $\mathbb{E}W = \sum_d W_d$. □

Eqn. (5.7) makes it clear that in heavy traffic, that is when $\rho_\infty^{(N)} \to 1$, all the probabilities $p_{0,d}^{(N)}$ go to zero. Also, in heavy traffic, the right hand side of Eqn. (5.3) reduces to $c_d - \hat{\lambda}_d h_{d1}$. Recalling the parametrization $1 - \rho_\infty^{(N)} = 1/N$ in Section 2, we obtain the following result.

Proposition 3.2. *For any service distribution $H_d(\cdot)$ and any service discipline that does not depend on the state of the environment, the mean of the workload in heavy traffic satisfies*

$$\lim_{N\to\infty} \frac{1}{N}\mathbb{E}W^{(N)} = \frac{1}{c_\infty}\sum_d \pi_d \left[\hat{\lambda}_d h_{d2}/2 + a_d(\hat{\lambda}_d h_{d1} - c_d)\right], \quad (5.10)$$

where $\boldsymbol{a}$ is a solution of $[Q \cdot \boldsymbol{a}]_d = c_d - \hat{\lambda}_d h_{d1}$.

Proof. Under the heavy traffic scaling, the empty probabilities $p_{0,d}^{(N)}$ go to zero for $d = 1, \ldots, D$. The result then follows immediately from Eqn. (5.4) and $1 - \rho_\infty^{(N)} = 1/N$. □

This leads to the main result of this section.

Theorem 3.3. *In heavy traffic, the scaled workload $\frac{1}{N}W^{(N)}$, converges in distribution to $\hat{W}$, where $\hat{W}$ is exponentially distributed with mean given in Eqn.* (5.10).

Proof. This follows from combining Proposition 3.2 with Theorem 4 in [34]. The full proof is in the Appendix. □

The above results yields that $\hat{W}$ is relatively compact, which together with the metric space being separable and complete implies that the scaled workload $\frac{1}{N}W^{(N)}$ is tight, by Prohorov's theorem [17].

4 Queue length vector under DPS

In the remainder of the chapter we focus on the multi-class model under DPS, where we assume that the service requirements of class-k customers are exponentially distributed with rate μ_k. In this section we establish some properties of the joint queue length distribution. We start with the flow equations, followed by a rate conservation law and an equation for the moments of the queue lengths conditioned on the environment.

Equating the flow in and out of state $(\boldsymbol{M}, Z) = (\boldsymbol{m}, d)$ yields (noting that $-q_{dd} = \sum_{\ell \neq d} q_{d\ell}$)

$$\begin{aligned}
&\left(\sum_{k=1}^{K} (\lambda_{k,d} + \frac{g_k m_k}{\sum_i g_i m_i} \mu_{k,d} \cdot \mathbf{1}_{\{m_k > 0\}}) - q_{dd} \right) p_{\boldsymbol{m},d} \\
&= \sum_{k=1}^{K} (\lambda_{k,d} p_{\boldsymbol{m}-\boldsymbol{e}_k,d} \cdot \mathbf{1}_{\{m_k>0\}} + \frac{g_k (m_k+1)}{\sum_i g_i m_i + g_k} \mu_{k,d} p_{\boldsymbol{m}+\boldsymbol{e}_k,d}) + \sum_{\ell \neq d} q_{d\ell} p_{\boldsymbol{m},\ell},
\end{aligned} \tag{5.11}$$

where $p_{\boldsymbol{m},d} := \mathbb{P}((\boldsymbol{M}, Z) = (\boldsymbol{m}, d))$ and $\boldsymbol{e}_k$ is the vector with 1 in the k-th place and zeros elsewhere. We now define the partial probability generating function (PGF) for when the background process is in state d:

$$\begin{aligned}
P_d(\boldsymbol{z}) &:= \mathbb{E}[z_1^{M_1} \cdots z_K^{M_K} \cdot \mathbf{1}_{\{Z=d\}}] \\
&:= \sum_{m_1=0}^{\infty} \cdots \sum_{m_K=0}^{\infty} \mathbb{P}(M_1 = m_1, \ldots, M_K = m_K, Z = d) \cdot z_1^{m_1} \cdots z_K^{m_K} \\
&= \sum_{\boldsymbol{m} \geq \boldsymbol{0}} p_{\boldsymbol{m},d} \boldsymbol{z}^{\boldsymbol{m}},
\end{aligned}$$

where $z_1^{m_1} \cdots z_K^{m_K} =: \boldsymbol{z}^{\boldsymbol{m}}$ and $(m_1, \ldots, m_K) \geq (0, \ldots, 0)$, i.e. $\boldsymbol{m} \geq \boldsymbol{0}$. Then the overall generating function for the queue length is given by $P(\boldsymbol{z}) := \mathbb{E}[z_1^{M_1} \cdots z_K^{M_K}] = \sum_{d=1}^{D} P_d(\boldsymbol{z})$. We also define

$$\begin{aligned}
R_d(\boldsymbol{z}) &:= \sum_{\boldsymbol{m} \geq \boldsymbol{0}} \frac{p_{\boldsymbol{m},d} \boldsymbol{z}^{\boldsymbol{m}}}{\sum_j g_j m_j} \cdot \mathbf{1}_{\{\sum_{j=1}^{K} m_j > 0\}}, \quad \text{hence} \\
\frac{\partial R_d(\boldsymbol{z})}{\partial z_k} &= z_k^{-1} \sum_{\boldsymbol{m} \geq \boldsymbol{e}_k} \frac{m_k}{\sum_j g_j m_j} p_{\boldsymbol{m},d} \boldsymbol{z}^{\boldsymbol{m}} \cdot \mathbf{1}_{\{\sum_{j=1}^{K} m_j > 0\}}.
\end{aligned}$$

By multiplying the flow equation (5.11) with $\boldsymbol{z}^{\boldsymbol{m}}$, summing over all vectors $\boldsymbol{m} \geq \boldsymbol{0}$ and rearranging terms, we can eventually write it in terms of

the PGF $P_d(\boldsymbol{z})$ and the partial derivative $\partial R_d(\boldsymbol{z})/\partial z_k$, that is,

$$\sum_{k=1}^{K}\left[\lambda_{k,d}(1-z_k)P_d(\boldsymbol{z})+\mu_{k,d}g_k(z_k-1)\frac{\partial R_d(\boldsymbol{z})}{\partial z_k}\right]=\sum_{\ell=1}^{D}P_\ell(\boldsymbol{z})q_{\ell d}. \quad (5.12)$$

It will be convenient to write the equation fully in terms of $\partial R_d/\partial z_k$, so we note the relation

$$P_d(\boldsymbol{z})=\sum_{k=1}^{K}g_k z_k\frac{\partial R_d(\boldsymbol{z})}{\partial z_k}+p_{\mathbf{0},d} \quad (5.13)$$

where $p_{\mathbf{0},d}=P((\boldsymbol{M},Z)=(\mathbf{0},d))$ is the probability of an empty queue in state d, which is equivalent to the probability of no workload defined in Section 3. We incorporate this into Eqn. (5.12) to obtain

$$\begin{aligned}\sum_{k=1}^{K}\lambda_{k,d}(1-z_k)\left[\sum_{j=1}^{K}g_j z_j\frac{\partial R_d(\boldsymbol{z})}{\partial z_j}+p_{\mathbf{0},d}\right]\\ +\sum_{k=1}^{K}\mu_{k,d}g_k(z_k-1)\frac{\partial R_d(\boldsymbol{z})}{\partial z_k}=\sum_{\ell=1}^{D}\left[\sum_{k=1}^{K}g_k z_k\frac{\partial R_\ell(\boldsymbol{z})}{\partial z_k}+p_{\mathbf{0},\ell}\right]q_{\ell d}.\end{aligned} \quad (5.14)$$

This equation we will use later when deriving the heavy-traffic limit.

For the M/M/1 queue with modulated arrivals and service times, moments of the queue length and the sojourn times can be found for the FCFS service discipline, in [97] and [68], respectively. In [83] the authors establish a recursive formula to calculate moments of the queue length in a non-modulated DPS system. In a similar fashion, we obtain an expression for the sum of the state-dependent moments weighted with the capacity of the server. We also derive a *rate conservation law*, which shows how the average arrival rates per class are proportional to the resources allocated to that same class, and the service they receive. Both results can be found in the following proposition.

Proposition 4.1. *When the queue is stable, the average number of class-k arrivals is proportional to the service resources allocated to class-k customers, i.e.*

$$\lambda_{k,\infty}=\sum_d \mu_{k,d}\mathbb{E}\left[\frac{g_k M_k}{\sum_j g_j M_j}\cdot\mathbf{1}_{\{\sum_j M_j>0\}}\cdot\mathbf{1}_{\{Z=d\}}\right],$$

for $k = 1, \ldots, K$. Furthermore, the state-dependent expectations of M_k satisfy

$$\sum_d c_d \mathbb{E}[M_k \cdot \mathbf{1}_{\{Z=d\}}] = \frac{\lambda_{k,\infty}}{\mu_k} + \sum_{d,j} g_j \frac{\lambda_{k,d}\mathbb{E}[M_j \cdot \mathbf{1}_{\{Z=d\}}] + \lambda_{j,d}\mathbb{E}[M_k \cdot \mathbf{1}_{\{Z=d\}}]}{\mu_k g_k + \mu_j g_j}.$$

Proof. We sum Eqn. (5.12) over d and take the derivative w.r.t. z_i to obtain

$$\sum_d \left[-\lambda_{i,d} P_d(\boldsymbol{z}) + \sum_j \lambda_{j,d}(1 - z_j) \frac{\partial P_d(\boldsymbol{z})}{\partial z_j} + \mu_{i,d} g_i \frac{\partial R_d(\boldsymbol{z})}{\partial z_i} + \sum_j \mu_{j,d} g_j (z_j - 1) \frac{\partial^2 R_d(\boldsymbol{z})}{\partial z_i \partial z_j} \right] = 0.$$

Letting $\boldsymbol{z} \to 1$ yields

$$\sum_d \left[\mu_{i,d} g_i \frac{\partial R_d(\boldsymbol{z})}{\partial z_i}\bigg|_{\boldsymbol{z}\to 1} - \lambda_{i,d} P_d(\boldsymbol{z})\bigg|_{\boldsymbol{z}\to 1} \right] = 0. \tag{5.15}$$

Since

$$\lim_{\boldsymbol{z}\to 1} P_d(\boldsymbol{z}) = \pi_d,$$

the following conservation law results from Eqn. (5.15), for $k = 1, \ldots, K$:

$$\begin{aligned} \lambda_{k,\infty} &= \sum_d \mu_{k,d} g_k \frac{\partial R_d(\boldsymbol{z})}{\partial z_k}\bigg|_{\boldsymbol{z}\to 1} \\ &= \sum_d \mu_{k,d} \mathbb{E}\left[\frac{g_k M_k}{\sum_j g_j M_j} \cdot \mathbf{1}_{\{\sum_j M_j > 0\}} \cdot \mathbf{1}_{\{Z=d\}} \right]. \end{aligned} \tag{5.16}$$

By taking partial derivatives of Eqn. (5.13), we obtain after standard calculations the recursive relation

$$\frac{\partial^j P_d(\boldsymbol{z})}{\partial z_{i_1} \cdots \partial z_{i_j}}\bigg|_{\boldsymbol{z}\to 1} = \sum_{k=1}^{K} g_k \frac{\partial^{j+1} R_d(\boldsymbol{z})}{\partial z_{i_1} \cdots \partial z_{i_j} \partial z_k} + \sum_{\ell=1}^{j} g_{i_\ell} \frac{\partial^j R_d(\boldsymbol{z})}{\partial z_{i_1} \cdots \partial z_{i_j}}. \tag{5.17}$$

In particular, this yields the explicit form

$$\mathbb{E}[M_k \cdot \mathbf{1}_{\{Z=d\}}] = \frac{\partial P_d}{\partial z_k}\bigg|_{\boldsymbol{z}\to 1} = \sum_j g_j \frac{\partial^2 R_d}{\partial z_k \partial z_j}\bigg|_{\boldsymbol{z}\to 1} + g_k \frac{\partial R_d}{\partial z_k}\bigg|_{\boldsymbol{z}\to 1}.$$

Proceeding from the rate conservation law, Eqn. (5.16), and by using $\mu_{k,d} = \mu_k c_d$, we have

$$\sum_d c_d g_k \frac{\partial R_d}{\partial z_k}\bigg|_{\boldsymbol{z}\to 1} = \frac{\lambda_{k,\infty}}{\mu_k}.$$

By taking two partial derivatives of the balance equation Eqn. (5.12) we can solve for a second mixed derivative of R_d, namely

$$\begin{aligned}\frac{\partial^2 R_d}{\partial z_k \partial z_j}\bigg|_{\boldsymbol{z}\to 1} &= \frac{\sum_\ell \mathbb{E}[M_k M_j \cdot \mathbf{1}_{\{Z=\ell\}}] q_{\ell d}}{\mu_{k,d} g_k + \mu_{j,d} g_j} \\ &\quad + \frac{\lambda_{k,d}\mathbb{E}[M_j \cdot \mathbf{1}_{\{Z=d\}}] + \lambda_{j,d}\mathbb{E}[M_k \cdot \mathbf{1}_{\{Z=d\}}]}{\mu_{k,d} g_k + \mu_{j,d} g_j},\end{aligned}$$

thus also yielding a mixed moment. Summing over the weighted moments of the number of class-k customers while in state d, we obtain a linear equation resembling Eqn. (16) in [83] for the non-modulated DPS queue:

$$\begin{aligned}\sum_d c_d \mathbb{E}[M_k \cdot \mathbf{1}_{\{Z=d\}}] &= \frac{\lambda_{k,\infty}}{\mu_k} + \sum_{d,j} c_d g_j \frac{\partial^2 R_d}{\partial z_k \partial z_j}\bigg|_{\boldsymbol{z}\to 1} \quad \text{(by Eqn. (5.17))} \\ &= \frac{\lambda_{k,\infty}}{\mu_k} + \sum_{d,j} g_j \frac{\lambda_{k,d}\mathbb{E}[M_j \cdot \mathbf{1}_{\{Z=d\}}] + \lambda_{j,d}\mathbb{E}[M_k \cdot \mathbf{1}_{\{Z=d\}}]}{\mu_k g_k + \mu_j g_j},\end{aligned}$$

where the last equality comes from $\sum_d q_{\ell d} = 0$. □

5 Preliminary results for the queue length in heavy traffic

We proceed to show that in heavy traffic, the distribution of the environment and the joint queue length become independent. This result, along with two technical lemmas that we derive in this section, will later help to establish the main result about the limiting queue length under DPS, presented in Section 6. Here we consider the queue length vector $(M_1, \ldots, M_K)$ scaled with $1/N$ and evaluate the PGF in $z^{1/N}$. The objective is to determine the distribution of $\frac{1}{N}(M_1^{(N)}, \ldots, M_K^{(N)}) \cdot \mathbf{1}_{\{Z=d\}}$ as N goes to infinity. We will state the existence of the limiting vector, and thus also the limit of the generating function $P_d^{(N)}(\boldsymbol{z}^{1/N})$, as an assumption. This assumption will be proven in Section 6.2. The superscript N denotes dependency on the prelimit parameter $\lambda_d^{(N)}$.

We make use of the change of variables $e^{-s_k} = z_k$, for $s_k > 0$, and denote $e^{-\boldsymbol{s}/N} := (e^{-s_1/N}, \ldots, e^{-s_K/N})$. Assuming that the limit exists, we use the new variables to define the heavy-traffic quantities: We let $\hat{p}_{\mathbf{0},d} := \lim_{N\to\infty} p^{(N)}_{\mathbf{0},d}$, $\hat{P}_d(\boldsymbol{s}) := \lim_{N\to\infty} P^{(N)}_d(e^{-\boldsymbol{s}/N}) = \lim_{N\to\infty} \mathbb{E}[e^{-\sum_j s_j M_j/N} \cdot \mathbf{1}_{\{Z=d\}}]$ and $\hat{P}(\boldsymbol{s}) := \sum_d \hat{P}_d(\boldsymbol{s}) = \lim_{N\to\infty} \mathbb{E}[e^{-\sum_j s_j M_j/N}]$. We denote by $(\hat{M}_1, \ldots, \hat{M}_K)$ the random vector corresponding to the LST $\hat{P}(\boldsymbol{s})$. Finally, let

$$\hat{R}_d(\boldsymbol{s}) := \mathbb{E}\left[\frac{1 - e^{-\sum_j s_j \hat{M}_j}}{\sum_j g_j \hat{M}_j} \cdot \mathbf{1}_{\{\sum_j \hat{M}_j > 0\}} \cdot \mathbf{1}_{\{Z=d\}}\right],$$

where the 1 in the numerator is to ensure that the bracketed expression remains bounded when the queue length quantities $\hat{M}_k$ are all near zero. We can now proceed to the following lemma.

Lemma 5.1. *If* $\lim_{N\to\infty} P^{(N)}_d(e^{-\boldsymbol{s}/N})$ *exists, then it satisfies*

$$\hat{P}_d(\boldsymbol{s}) = \sum_{k=1}^{K} g_k \frac{\partial \hat{R}_d(\boldsymbol{s})}{\partial s_k} + \hat{p}_{\mathbf{0},d}. \tag{5.18}$$

Proof. From Eqn. (5.13),

$$\lim_{N\to\infty} P^{(N)}_d(e^{-\boldsymbol{s}/N}) = \lim_{N\to\infty} \sum_{k=1}^{K} g_k z_k \frac{\partial R^{(N)}_d(\boldsymbol{z})}{\partial z_k}\bigg|_{\boldsymbol{z}=e^{-\boldsymbol{s}/N}} + \hat{p}_{\mathbf{0},d}. \tag{5.19}$$

Note that

$$\begin{aligned}
&\lim_{N\to\infty} z_k \frac{\partial R^{(N)}_d(\boldsymbol{z})}{\partial z_k}\bigg|_{\boldsymbol{z}=e^{-\boldsymbol{s}/N}} \\
&\quad= \lim_{N\to\infty} \sum_{\boldsymbol{m}\geq \boldsymbol{e}_k} \frac{m_k}{\sum_j g_j m_j} p^{(N)}_{\boldsymbol{m},d} \boldsymbol{z}^{\boldsymbol{m}} \cdot \mathbf{1}_{\{\sum_j m_j > 0\}}\bigg|_{\boldsymbol{z}=e^{-\boldsymbol{s}/N}} \\
&\quad= \lim_{N\to\infty} \sum_{\boldsymbol{m}\geq \boldsymbol{e}_k} \frac{m_k}{\sum_j g_j m_j} p^{(N)}_{\boldsymbol{m},d} e^{-s_1 m_1/N} \cdot \ldots \cdot e^{-s_K m_K/N} \cdot \mathbf{1}_{\{\sum_j m_j > 0\}} \\
&\quad= \lim_{N\to\infty} \mathbb{E}\left[\frac{M^{(N)}_k}{\sum_j g_j M^{(N)}_j} e^{-\sum_j s_j M^{(N)}_j/N} \cdot \mathbf{1}_{\{\sum_j M^{(N)}_j/N > 0\}} \cdot \mathbf{1}_{\{Z=d\}}\right] \\
&\quad= \mathbb{E}\left[\frac{\hat{M}_k}{\sum_j g_j \hat{M}_j} e^{-\sum_j s_j \hat{M}_j} \cdot \mathbf{1}_{\{\sum_j \hat{M}_j > 0\}} \cdot \mathbf{1}_{\{Z=d\}}\right] \\
&\quad= \frac{\partial \hat{R}_d(\boldsymbol{s})}{\partial s_k}.
\end{aligned} \tag{5.20}$$

The second-to-last step follows from the fact that $\frac{M_k^{(N)}}{\sum_j g_j M_j^{(N)}} \cdot \mathrm{e}^{-\sum_j s_j M_j^{(N)}} \cdot \mathbf{1}_{\{\sum_j M_j^{(N)}>0\}}$ is upper bounded by $1/\min_j(g_j)$. By the continuous mapping theorem (see Billingsley's [17]), it converges in distribution to $\frac{\hat{M}_k}{\sum_j g_j \hat{M}_j} \cdot \mathrm{e}^{-\sum_j s_j \hat{M}_j} \cdot \mathbf{1}_{\{\sum_j \hat{M}_j>0\}}$. The environment is not affected by the heavy-traffic scaling. Eqns. (5.19) and (5.20) now conclude the proof. □

With the help of Lemma 5.1 we obtain:

Proposition 5.2. *If* $\lim_{N\to\infty} P_d^{(N)}(\mathrm{e}^{-\boldsymbol{s}/N})$ *exists, the joint queue length distribution is independent of the environment in heavy traffic, that is*

$$\hat{P}_d(\boldsymbol{s}) = \pi_d \hat{P}(\boldsymbol{s}).$$

Proof. We use the change of variables $z_k = \mathrm{e}^{-s_k}$. Since

$$z_k \frac{\partial R_d}{\partial z_k}(\boldsymbol{z})\bigg|_{\boldsymbol{z}=\mathrm{e}^{-\boldsymbol{s}}} = -\frac{\partial R_d}{\partial s_k}(\mathrm{e}^{-\boldsymbol{s}}),$$

we obtain, by applying the heavy traffic scaling to Eqn. (5.14),

$$\begin{aligned}\sum_k \Bigg[\lambda_{k,d}^{(N)}(1-\mathrm{e}^{-s_k/N}) \Bigg[\sum_{j=1}^K g_j \frac{-\partial R_d^{(N)}}{\partial s_j}(\mathrm{e}^{-\boldsymbol{s}/N}) + p_{\mathbf{0},d}^{(N)}\Bigg] \\ - \mu_{k,d} g_k (\mathrm{e}^{-s_k/N}-1)\mathrm{e}^{s_k/N} \frac{\partial R_d^{(N)}}{\partial s_k}(\mathrm{e}^{-\boldsymbol{s}/N})\Bigg] \\ = \sum_{\ell=1}^D \Bigg[\sum_{k=1}^K g_k \frac{-\partial R_\ell^{(N)}}{\partial s_k}(\mathrm{e}^{-\boldsymbol{s}/N}) + p_{\mathbf{0},\ell}^{(N)}\Bigg] q_{\ell d}.\end{aligned}$$

With Taylor expansion we obtain

$$\begin{aligned}\sum_k \Bigg[\lambda_{k,d}^{(N)}\left(\frac{s_k}{N} - \frac{s_k^2}{N^2}\right) \Bigg[\sum_{j=1}^K g_j \frac{-\partial R_d^{(N)}}{\partial s_j}(\mathrm{e}^{-\boldsymbol{s}/N}) + p_{\mathbf{0},d}^{(N)}\Bigg] \\ - \mu_{k,d} g_k \left(\frac{s_k}{N} + \frac{s_k^2}{N^2}\right) \frac{-\partial R_d^{(N)}}{\partial s_k}(\mathrm{e}^{-\boldsymbol{s}/N})\Bigg] \\ = \sum_{\ell=1}^D \Bigg[\sum_{k=1}^K g_k \frac{-\partial R_\ell^{(N)}}{\partial s_k}(\mathrm{e}^{-\boldsymbol{s}/N}) + p_{\mathbf{0},\ell}^{(N)}\Bigg] q_{\ell d} + \mathcal{O}(N^{-3}).\end{aligned} \tag{5.21}$$

Since $\frac{-\partial R_d^{(N)}}{\partial s_j}$ is bounded (see proof of Lemma 5.1) and converges to $\frac{\partial \hat{R}_d}{\partial s_j}$, we obtain as $N \to \infty$,

$$\boldsymbol{\nu} \cdot [Q]_d = \sum_{\ell=1}^{D} \left[\sum_{k=1}^{K} g_k \frac{\partial \hat{R}_\ell(\boldsymbol{s})}{\partial s_k} + \hat{p}_{\mathbf{0},\ell} \right] q_{\ell d} = 0, \quad \boldsymbol{s} \geq \mathbf{0}, \quad \forall d, \tag{5.22}$$

where $[Q]_d$ denotes the d^{th} column of Q and $\boldsymbol{\nu}$ is a row vector such that $\nu_\ell = \sum_{k=1}^{K} g_k \frac{\partial \hat{R}_\ell(\boldsymbol{s})}{\partial s_k} + \hat{p}_{\mathbf{0},\ell}$. This implies that $\boldsymbol{\nu} Q = 0$, and since Q is a generator we conclude that

$$\nu_d = \sum_{k=1}^{K} g_k \frac{\partial \hat{R}_d(\boldsymbol{s})}{\partial s_k} + \hat{p}_{\mathbf{0},d} = \pi_d x,$$

where x does not depend on d. Observe now that by Lemma 5.1, we have

$$\begin{aligned} \hat{P}_d(\boldsymbol{s}) - \hat{p}_{\mathbf{0},d} &= \sum_{k=1}^{K} g_k \frac{\partial \hat{R}_d(\boldsymbol{s})}{\partial s_k} \\ &= \pi_d x - \hat{p}_{\mathbf{0},d} \\ &= \mathbb{E}[\mathbf{1}_{\{Z=d\}}] x - \hat{p}_{\mathbf{0},d}. \end{aligned}$$

Since $\hat{P}_d(\boldsymbol{s}) = \mathbb{E}\left[\mathrm{e}^{-\sum_j s_j \hat{M}_j} \cdot \mathbf{1}_{\{Z=d\}} \right]$ this implies that

$$x = \mathbb{E}\left[\mathrm{e}^{-\sum_j s_j \hat{M}_j} \right] = \hat{P}(\boldsymbol{s}).$$

This shows that the environment becomes independent from the joint queue-length process in the heavy-traffic limit. □

The flow equation, Eqn. (5.14), simplifies considerably in heavy traffic, as shown in the following lemma.

Lemma 5.3. *If* $\lim_{N\to\infty} P_d^{(N)}(\mathrm{e}^{-\boldsymbol{s}/N})$ *exists, then* $\hat{R}_d(\boldsymbol{s})$ *satisfies the following equation:*

$$0 = \sum_{k,d} F_{k,d}(\boldsymbol{s}) \frac{\partial \hat{R}_d(\boldsymbol{s})}{\partial s_k}, \quad \forall \boldsymbol{s} \geq \mathbf{0},$$

with $F_{k,d}(\boldsymbol{s})$ *defined as*

$$F_{k,d}(\boldsymbol{s}) := g_k \left(\sum_j \hat{\lambda}_{j,d} s_j - \mu_{k,d} s_k \right).$$

Proof. We start by multiplying through Eqn. (5.21) with N, followed by summing over d. Due to Q being a generator, this eliminates the right hand side with the transition rates $q_{\ell d}$:

$$\sum_{k,d}\left[\lambda_{k,d}^{(N)}\left(s_k-\frac{s_k^2}{N}\right)\left[\sum_{j=1}^{K}g_j\frac{-\partial R_d^{(N)}}{\partial s_j}(\mathrm{e}^{-\boldsymbol{s}/N})+p_{\mathbf{0},d}^{(N)}\right]\right]$$
$$-\sum_{k,d}\left[\mu_{k,d}g_k\left(s_k+\frac{s_k^2}{N}\right)\frac{-\partial R_d^{(N)}}{\partial s_k}(\mathrm{e}^{-\boldsymbol{s}/N})\right]+\mathcal{O}(N^{-2})=0.$$

Taking the limit $N\to\infty$ yields

$$\begin{aligned}0&=\sum_{k,d}\hat{\lambda}_{k,d}s_k\sum_j g_j\frac{\partial\hat{R}_d(\boldsymbol{s})}{\partial s_j}-\sum_{k,d}\mu_{k,d}g_k s_k\frac{\partial\hat{R}_d(\boldsymbol{s})}{\partial s_k}\\&=\sum_{k,d}g_k\frac{\partial\hat{R}_d(\boldsymbol{s})}{\partial s_k}\sum_j\hat{\lambda}_{j,d}s_j-\sum_{k,d}\mu_{k,d}g_k s_k\frac{\partial\hat{R}_d(\boldsymbol{s})}{\partial s_k}\\&=\sum_{k,d}g_k\left(\sum_j\hat{\lambda}_{j,d}s_j-\mu_{k,d}s_k\right)\frac{\partial\hat{R}_d(\boldsymbol{s})}{\partial s_k}\\&=\sum_{k,d}F_{k,d}(\boldsymbol{s})\frac{\partial\hat{R}_d(\boldsymbol{s})}{\partial s_k}.\end{aligned}\tag{5.23}$$

□

In what follows we focus on

$$F_{k,\infty}(\boldsymbol{s}):=\sum_d F_{k,d}(\boldsymbol{s})\pi_d=g_k\left(\sum_j\hat{\lambda}_{j,\infty}s_j-\mu_{k,\infty}s_k\right),$$

and denote its vector counterpart by $\boldsymbol{F}_\infty(\boldsymbol{s})=(F_{1,\infty}(\boldsymbol{s}),\ldots,F_{K,\infty}(\boldsymbol{s}))$.

6 Queue length distribution in heavy traffic

We now state and consequently prove our main result about the queue length distribution.

Theorem 6.1. *When scaled by* $1/N = (1 - \rho_\infty^{(N)})$*, the queue-length vector converges in distribution as* $(\rho_{1,\infty}^{(N)}, \ldots, \rho_{K,\infty}^{(N)}) \to (\hat{\rho}_{1,\infty}, \ldots, \hat{\rho}_{K,\infty})$ *i.e.,* $\rho_\infty^{(N)} \to 1$*, namely*

$$\begin{aligned}\frac{1}{N}(M_1^{(N)}, \ldots, M_K^{(N)}) \cdot \mathbf{1}_{\{Z=d\}} &\xrightarrow{\mathrm{d}} \left(\hat{M}_1, \ldots, \hat{M}_K\right) \cdot \mathbf{1}_{\{Z=d\}} \\ &\stackrel{\mathrm{d}}{=} \pi_d \left(\frac{\hat{\rho}_{1,\infty}}{g_1}, \ldots, \frac{\hat{\rho}_{K,\infty}}{g_K}\right) \cdot X, \end{aligned} \tag{5.24}$$

where $\xrightarrow{\mathrm{d}}$ *denotes convergence in distribution and* X *is exponentially distributed with mean*

$$\mathbb{E}X = \frac{\sum_k \hat{\rho}_{k,\infty}/\mu_k - \sum_d c_d \pi_d a_d (1 - \hat{\rho}_d)}{c_\infty \sum_k \hat{\rho}_{k,\infty}/(g_k \mu_k)}, \tag{5.25}$$

with $\hat{\rho}_d := c_d^{-1} \sum_k \hat{\lambda}_{k,d}/\mu_k$ *and* $\boldsymbol{a} = (a_1, \ldots, a_D)^T$ *being a solution of*

$$[Q \cdot \boldsymbol{a}]_d = c_d(1 - \hat{\rho}_d).$$

We will prove this theorem in the two following subsections, first showing in Section 6.1 the state-space collapse observed in Eqn. (5.24) and then in Section 6.2 we will show that X is exponentially distributed with the mean given by Eqn. (5.25).

6.1 State-space collapse

The first part of the proof of Theorem 6.1 is the state-space collapse. In this section, we assume $\lim_{N\to\infty} P_d^{(N)}(\mathrm{e}^{-\boldsymbol{s}/N})$ exists.

Observe that due to Proposition 5.2,

$$\begin{aligned}\hat{R}_d(\boldsymbol{s}) &= \mathbb{E}\left[\frac{1 - \mathrm{e}^{-\sum_k s_k \hat{M}_k}}{\sum_k g_k \hat{M}_k} \cdot \mathbf{1}_{\{\sum_k \hat{M}_k > 0\}} \cdot \mathbf{1}_{\{Z=d\}}\right] \\ &= \mathbb{E}\left[\frac{1 - \mathrm{e}^{-\sum_k s_k \hat{M}_k}}{\sum_k g_k \hat{M}_k} \cdot \mathbf{1}_{\{\sum_k \hat{M}_k > 0\}}\right] \cdot \pi_d \\ &=: \hat{R}(\boldsymbol{s})\pi_d, \end{aligned} \tag{5.26}$$

where the last equation defines $\hat{R}(\boldsymbol{s})$, which is independent of d. We now derive some properties of $\hat{R}(\boldsymbol{s})$.

Lemma 6.2. *$\hat{R}(\boldsymbol{s})$ is constant on a $(K-1)$-dimensional hyperplane $\mathcal{H}_c$, where*

$$\mathcal{H}_c := \left\{\boldsymbol{s} \geq \mathbf{0} : \sum_k \frac{\hat{\rho}_{k,\infty}}{g_k} s_k = c\right\}, \quad c > 0.$$

Proof. We follow closely the steps of the proof of Lemma 3 in [91]. The proof has 3 steps: (i) Show that $F_{k,\infty}(\boldsymbol{s})$ is parallel to the hyperplane. Hence, any flow corresponding to $F_{k,\infty}$ that starts in the plane, stays in the plane. (ii) Show that $\hat{R}(\boldsymbol{s})$ is constant along each flow in the hyperplane and (iii) show that each flow in the hyperplane converges to a unique point. This implies that $\hat{R}(\boldsymbol{s})$ is constant on the hyperplane.

(i) $\boldsymbol{F}_\infty(\boldsymbol{s})$ is parallel to $\mathcal{H}_c$
Observe that with $1 = \sum_k \hat{\rho}_{k,\infty}$ and $\hat{\rho}_{k,\infty} = \hat{\lambda}_{k,\infty}/\mu_{k,\infty}$,

$$\begin{aligned}\sum_k \frac{\hat{\rho}_{k,\infty}}{g_k} F_{k,\infty}(\boldsymbol{s}) &= \sum_k \hat{\rho}_{k,\infty} \left(\sum_j \hat{\lambda}_{j,\infty} s_j - \mu_{k,\infty} s_k \right) \\ &= \sum_j \hat{\lambda}_{j,\infty} s_j - \sum_k \hat{\rho}_{k,\infty} \mu_{k,\infty} s_k \\ &= \sum_k \hat{\lambda}_{k,\infty} s_k - \sum_k \hat{\lambda}_{k,\infty} s_k \\ &= 0.\end{aligned}$$

This indicates that the K-dimensional vector $F_\infty(\boldsymbol{s})$ is parallel to the hyperplane.

(ii) $\hat{R}(\boldsymbol{s})$ is constant along flows in $\mathcal{H}_c$
For each $\boldsymbol{s} \geq \boldsymbol{0}$, there exists a unique flow $\boldsymbol{f}(u) = (f_1(u), \ldots, f_K(u))^T$ parametrized by $u \geq 0$, such that

$$\boldsymbol{f}(0) = \boldsymbol{s} \quad \text{and} \quad \frac{\mathrm{d}f_k(u)}{\mathrm{d}u} = F_{k,\infty}(\boldsymbol{f}(u)). \tag{5.27}$$

Due to (i), any flow that starts in $\mathcal{H}_c$, stays in $\mathcal{H}_c$. Now,

$$\begin{aligned}\frac{\mathrm{d}\hat{R}(\boldsymbol{f}(u))}{\mathrm{d}u} &= \sum_{k=1}^K \frac{\mathrm{d}f_k(u)}{\mathrm{d}u} \cdot \frac{\partial \hat{R}(\boldsymbol{s})}{\partial s_k}\bigg|_{\boldsymbol{s}=\boldsymbol{f}(u)} \\ &= \sum_{k=1}^K F_{k,\infty}(\boldsymbol{f}(u)) \cdot \frac{\partial \hat{R}(\boldsymbol{s})}{\partial s_k}\bigg|_{\boldsymbol{s}=\boldsymbol{f}(u)} \\ &= \sum_{k=1}^K \sum_{d=1}^D F_{k,d}(\boldsymbol{f}(u))\pi_d \cdot \frac{\partial \hat{R}(\boldsymbol{s})}{\partial s_k}\bigg|_{\boldsymbol{s}=\boldsymbol{f}(u)} \\ &= \sum_{k=1}^K \sum_{d=1}^D F_{k,d}(\boldsymbol{f}(u)) \cdot \frac{\partial \hat{R}_d(\boldsymbol{s})}{\partial s_k}\bigg|_{\boldsymbol{s}=\boldsymbol{f}(u)} \\ &= 0, \qquad \text{by Eqn. (5.23),}\end{aligned}$$

implying that $\hat{R}(\boldsymbol{f}(u))$ is constant along each flow $\boldsymbol{f}(u)$ which lies in $\mathcal{H}_c$.

(iii) Each flow in $\mathcal{H}_c$ converges to a unique point

Here we first write the flow specifications in a vector-matrix form, then show that one eigenvalue of that matrix is zero with eigenvector $\boldsymbol{s}^* \in \mathcal{H}_1$, and the other eigenvalues are negative, and thus we can write $\boldsymbol{f}(u) = c \cdot \boldsymbol{s}^* + \boldsymbol{g}(u)$ where $\lim_{u\to\infty} \boldsymbol{g}(u) = 0$.
Eqn. (5.27) can be written in matrix-vector form as

$$\boldsymbol{f}'(u) = A\boldsymbol{f}(u),$$

with

$$A = \begin{pmatrix} g_1(\hat{\lambda}_{1,\infty} - \mu_{1,\infty}) & g_1\hat{\lambda}_{2,\infty} & \cdots & g_1\hat{\lambda}_{K,\infty} \\ g_2\hat{\lambda}_{1,\infty} & g_2(\hat{\lambda}_{2,\infty} - \mu_{2,\infty}) & \cdots & g_2\hat{\lambda}_{K,\infty} \\ \vdots & & \ddots & \vdots \\ g_K\hat{\lambda}_{1,\infty} & \cdots & & g_K(\hat{\lambda}_{K,\infty} - \mu_{K,\infty}) \end{pmatrix}.$$

Let D be the diagonal matrix with $d_i = \hat{\rho}_{i,\infty}/g_i$ on the diagonal. Then with

$$\begin{aligned} S &:= DAD^{-1} \\ &= \begin{pmatrix} g_1(\hat{\lambda}_{1,\infty} - \mu_{1,\infty}) & g_2\frac{\hat{\rho}_{1,\infty}}{\hat{\rho}_{2,\infty}}\hat{\lambda}_{2,\infty} & \cdots & g_K\frac{\hat{\rho}_{1,\infty}}{\hat{\rho}_{K,\infty}}\hat{\lambda}_{K,\infty} \\ g_1\frac{\hat{\rho}_{2,\infty}}{\hat{\rho}_{1,\infty}}\hat{\lambda}_{1,\infty} & g_2(\hat{\lambda}_{2,\infty} - \mu_{2,\infty}) & \cdots & g_K\frac{\hat{\rho}_{2,\infty}}{\hat{\rho}_{K,\infty}}\hat{\lambda}_{K,\infty} \\ \vdots & & \ddots & \vdots \\ g_1\frac{\hat{\rho}_{K,\infty}}{\hat{\rho}_{1,\infty}}\hat{\lambda}_{1,\infty} & \cdots & & g_K(\hat{\lambda}_{K,\infty} - \mu_{K,\infty}) \end{pmatrix}, \end{aligned}$$

S^T is a generator corresponding to a finite-state Markov chain.

From the proof of Lemma 4 in [91], it is easily seen that the Markov chain corresponding to S^T is irreducible (since we assume that all $\lambda_{k,d} > 0$ for all k and at least one d). Retracing the arguments stated there, for completeness, it follows that this Markov chain has a unique equilibrium distribution (column) vector, $\boldsymbol{\eta}$, such that $\boldsymbol{\eta}^T S^T = 0$. In particular, 0 is an eigenvalue with multiplicity one and all other eigenvalues have a strictly negative real part, see [9]. Since the eigenvalues of S^T and A are the same, 0 is also an eigenvalue of A with corresponding right eigenvector $\boldsymbol{s}^* = D^{-1}\boldsymbol{\eta}$, $\boldsymbol{s}^* \geq \mathbf{0}, \boldsymbol{s}^* \in \mathcal{H}_1$. The solution of the linear system $\boldsymbol{f}'(u) = A\boldsymbol{f}(u)$, $\boldsymbol{f}(0) \in \mathcal{H}_c$ can now be written as the sum of the homogeneous and the particular solution, i.e. $\boldsymbol{f}(u) = c \cdot \boldsymbol{s}^* + \boldsymbol{g}(u)$, where $\lim_{u\to\infty} \boldsymbol{g}(u) = \mathbf{0}$. This implies that all the flows in $\mathcal{H}_c$ converge to one common point $c \cdot \boldsymbol{s}^*$. Combining (i), (ii) and (iii), we conclude that the function $\hat{R}(\boldsymbol{s})$ is constant on $\mathcal{H}_c$. □

As a consequence of Lemma 6.2, the function $\hat{R}(\boldsymbol{s})$ depends on $\boldsymbol{s}$ only through the sum $\sum_{k=1}^{K}(\hat{\rho}_{k,\infty}/g_k)s_k$. Therefore, there exists a function $\hat{R}^*$: $\mathbb{R} \to \mathbb{R}$ such that $\hat{R}(\boldsymbol{s}) = \hat{R}^*(\sum_{k=1}^{K}(\hat{\rho}_{k,\infty}/g_k)s_k)$. Then

$$\frac{\partial}{\partial s_k}\hat{R}(\boldsymbol{s}) = \frac{\hat{\rho}_{k,\infty}}{g_k}\frac{\mathrm{d}\hat{R}^*(v)}{\mathrm{d}v}\bigg|_{v=\sum_{k=1}^{K}(\hat{\rho}_{k,\infty}/g_k)s_k},$$

so we obtain

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{-\sum_{k=1}^{K}s_k\hat{M}_k}] &= \lim_{N\to\infty}\sum_{d=1}^{D}P_d^{(N)}(\mathrm{e}^{-\boldsymbol{s}/N}) \\
&= \sum_{d=1}^{D}\sum_{k=1}^{K}g_k\frac{\partial\hat{R}_d(\boldsymbol{s})}{\partial s_k} \\
&= \sum_{d=1}^{D}\sum_{k=1}^{K}g_k\frac{\partial\hat{R}(\boldsymbol{s})}{\partial s_k}\pi_d \\
&= \sum_{k=1}^{K}\hat{\rho}_{k,\infty}\frac{\mathrm{d}\hat{R}^*(v)}{\mathrm{d}v}\bigg|_{v=\sum_{k=1}^{K}(\hat{\rho}_{k,\infty}/g_k)s_k} \\
&= \frac{\mathrm{d}\hat{R}^*(v)}{\mathrm{d}v}\bigg|_{v=\sum_{k=1}^{K}(\hat{\rho}_{k,\infty}/g_k)s_k}, \qquad (5.28)
\end{aligned}$$

which only depends on $v = \sum_{k=1}^{K}(\hat{\rho}_{k,\infty}/g_k)s_k$. Since we also have

$$\begin{aligned}
\mathbb{E}[\mathrm{e}^{-\sum_{k=1}^{K}s_k\hat{M}_k}] = \mathbb{E}\Bigg[&\mathrm{e}^{-\frac{g_1}{\hat{\rho}_{1,\infty}}v\hat{M}_1}\cdot\mathrm{e}^{-\frac{\hat{\rho}_{2,\infty}}{g_2}s_2\left(\frac{g_2}{\hat{\rho}_{2,\infty}}\hat{M}_2-\frac{g_1}{\hat{\rho}_{1,\infty}}\hat{M}_1\right)} \\
&\cdot\cdot\cdot\cdot\cdot\mathrm{e}^{-\frac{\hat{\rho}_{K,\infty}}{g_K}s_K\left(\frac{g_K}{\hat{\rho}_{K,\infty}}\hat{M}_K-\frac{g_1}{\hat{\rho}_{1,\infty}}\hat{M}_1\right)}\Bigg],
\end{aligned}$$

this together with Eqn. (5.28) implies that $\left(\frac{g_j}{\hat{\rho}_{j,\infty}}\hat{M}_j - \frac{g_1}{\hat{\rho}_{1,\infty}}\hat{M}_1\right) = 0$, for all $j = 1, \ldots, K$. Thus $(g_k/\hat{\rho}_{k,\infty})\hat{M}_k \overset{\mathrm{d}}{=} (g_j/\hat{\rho}_{j,\infty})\hat{M}_j$, for all k, j, almost surely. Combining this finding with that of Eqn. (5.22), we obtain Eqn. (5.24) with X distributed as $(g_1/\hat{\rho}_{1,\infty})\hat{M}_1$.

Remark 6.1 (Continuously modulated service requirements). In Section 3 we saw that the critical load is indeed reached when $\rho_\infty \to 1$, since then $p_{0,d} \to 0$. This indicates that $(1 - \rho_\infty)$ is the right heavy-traffic scaling when $\mu_{k,d} = \mu_k c_d$. This is less clear for a general $\mu_{k,d}$, that is, for continuously modulated service requirements, where the environment can

influence the departure rate of customers present in the system. For that setting, the workload process is no longer independent of the employed scheduling discipline, since the decision on which class to serve impacts the rate at which customers leave. We are not aware of any results on workload and waiting time distributions where the service distribution is a general function of both class and environment.

The majority of the preceding queue length results in this chapter can however be proven without the restriction of the product form, i.e., for continuously modulated service requirements. The traffic intensity for this variant is defined as for the multi-class model above, only this time one cannot split the average class-k service rate into $\mu_{k,\infty} = \mu_k c_\infty$. The traffic intensity per class k, $\rho_{k,\infty} = \lambda_{k,\infty}/\mu_{k,\infty}$, is in line with the Markov-modulated single-server queues. Assuming there exists a scaling $f(N)$ such that $f(N)(M_1, \ldots, M_K)\mathbf{1}_{\{Z=d\}}$ converges in distribution, it can be shown that the empty probabilities $p_{\mathbf{0},d}$ vanish in heavy traffic as $N \to \infty$, for a general $\mu_{k,d}$. This property then follows from Proposition 5.2 without relying on the workload results from Section 3 and the product form assumed there. Furthermore, under this assumption, all results in Section 6.1 hold, implying that a state-space collapse will appear. In other words, we can prove the first half of Theorem 6.1. However, we do not know what the distribution of the common factor X will be.

6.2 Distribution of the common factor

In order to prove that the limiting queue length distribution exists and to find the common factor of the queue length distribution in heavy traffic, the random variable X, we make use of the results on the workload of the total system. From [91] and Eqn. (5.24) we have that

$$\hat{W} \stackrel{\mathrm{d}}{=} \sum_{k=1}^{K} \frac{\hat{M}_k}{\mu_k} = X \cdot \sum_{k=1}^{K} \frac{\hat{\rho}_{k,\infty}}{g_k \mu_k}. \tag{5.29}$$

In order to apply the workload result of Section 3, we first derive the service requirement of an arbitrary customer while being in state d. If $H_k(\cdot)$ is the distribution function of a class-k customer's service requirement, then the probability of a class-k customer arriving and requiring service not exceeding x is $\alpha_{k,d} H_k(x)$. Summing over k now yields the desired distribution,

$$H_d(x) := \sum_{k=1}^{K} \alpha_{k,d} H_k(x).$$

The overall service requirement distribution thus depends on the state of the environment at its arrival. With exponential service requirements, the corresponding LST is given by

$$h_d(s) = \sum_{k=1}^{K} \frac{\alpha_{k,d}\mu_k}{\mu_k + s}, \quad s \geq 0, \tag{5.30}$$

and the first and second moment are given by

$$h_{d1} = \sum_k \frac{\alpha_{k,d}}{\mu_k}, \quad h_{d2} = 2\sum_k \frac{\alpha_{k,d}}{\mu_k^2}. \tag{5.31}$$

We can now apply the result of Theorem 3.3 with moments as given in Eqn. (5.31). Hence, we have that $\hat{W}$ is exponentially distributed with mean

$$\mathbb{E}\hat{W} = c_\infty^{-1}\left(\sum_k \hat{\rho}_{k,\infty}/\mu_k - \sum_d c_d\pi_d a_d(1-\hat{\rho}_d)\right),$$

where $\boldsymbol{a}$ is a solution of $[Q\cdot\boldsymbol{a}]_d = c_d - \hat{\lambda}_d\sum_k \frac{\alpha_{k,d}}{\mu_k} = c_d(1-\hat{\rho}_d)$. Along with Eqn. (5.29) this yields the mean of the exponential random variable X:

$$\begin{aligned}\mathbb{E}X &= \frac{\mathbb{E}\hat{W}}{\sum_k \hat{\rho}_{k,\infty}/(g_k\mu_k)} \\ &= \frac{\sum_k \hat{\rho}_{k,\infty}/\mu_k - \sum_d c_d\pi_d a_d(1-\hat{\rho}_d)}{c_\infty\sum_k \hat{\rho}_{k,\infty}/(g_k\mu_k)}.\end{aligned} \tag{5.32}$$

The first term of the numerator is in accordance with the results of [83] and [91], the second term is a result of the random environment.

The results in Sections 5 and 6.1 are based on the assumption that $\lim_{N\to\infty}\frac{1}{N}\boldsymbol{M}\cdot\mathbf{1}_{\{Z=d\}}$ exists. Since the scaled workload is tight, see Section 3, so is the scaled queue length. Then, by Prohorov's theorem ([17]) there exists a subsequence of N such that $\frac{1}{N}M_k$ converges in distribution, and hence for this subsequence $\lim_{N\to\infty} P_d^{(N)}(\mathrm{e}^{-\boldsymbol{s}/N})$ exists. Since each converging subsequence yields the same limit, the limit itself exists (see corollary page 59 in [17]), i.e. $\frac{1}{N}(\boldsymbol{M}, Z=d) \xrightarrow{\mathrm{d}} \hat{\boldsymbol{M}}\cdot\mathbf{1}_{\{Z=d\}}$, as $N\to\infty$, with the limiting vector as in Eqn. (5.24).

This concludes the proof of Theorem 6.1.

7 Conclusion and future work

We first studied the workload for a queue with modulated arrivals, service requirements and service capacity, and derived that the scaled work-

load converges to an exponentially distributed random variable in heavy traffic. The workload results obtained are valid for any service distribution and for any service discipline which does not depend on the environment. We then focussed on the special setting of a multi-class queue under the DPS policy and showed that the joint queue length distribution for such a system undergoes a state-space collapse in heavy traffic. Under the scaling of $(1-\rho_\infty)$, the vector-valued limiting distribution is independent of the modulating environment and converges in distribution to a one-dimensional random variable times a deterministic vector. In this derivation, the distribution of the scaled workload is a key quantity. With this we extend known results about the DPS queue to a Markov-modulated setting.

Clearly an interesting question for future consideration is whether the state-space collapse for the DPS policy carries over to continuously modulated service requirements, as discussed in Remark 6.1. Another open question concerns the characterization of the moments of the queue lengths for the modulated DPS queue, outside of heavy traffic. Last but not least, modulating the weights of the DPS would open the possibility of dynamical scheduling based on the environment. The latter would be a study on its own, as already the stability conditions will no longer be independent of the weights of the DPS policy.

Acknowledgements

We would like to thank Urtzi Ayesta, Joke Blom and Michel Mandjes for helpful discussions. This research was partially supported by the SMI program of INP Toulouse.

8 Appendix: Proof of Theorem 3.3

The proof Theorem 3.3 is based on Theorem 4 in [34], which can be adapted to our model as follows:

We start with notation and preliminaries. Let $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_D)$, $\bar{H}(s) = \text{diag}(1 - h_1(s), \ldots, 1 - h_D(s))$, $C = \text{diag}(c_1, \ldots, c_D)$ and $\boldsymbol{p}_0 = (p_{0,1}, \ldots, p_{0,D})$. Furthermore $\bar{H}_1$ and $\bar{H}_2$ are the diagonal matrices corresponding to the moments h_{d1} and h_{d2}, respectively, for $d = 1, \ldots, D$. Recall Eqn. (5.3), $[Q \cdot \boldsymbol{a}]_d = c_d - \lambda_d h_{d1} - c_\infty(1 - \rho_\infty)$. We will now construct a partial inverse of Q to make it easier to find a vector $\boldsymbol{a}$ which

solves this equation. Let Q_1 and R be matrices such that

$$Q_1 = \begin{pmatrix} q_{22} & q_{23} & \cdots & q_{2D} \\ \vdots & & & \\ q_{D2} & q_{D3} & \cdots & q_{DD} \end{pmatrix}, \quad R = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & Q_1^{-1} & & \\ 0 & & & \end{pmatrix}.$$

Then $\det Q_1 \neq 0$ and due to Q being a generator (for more details see [34]), we have

$$QR = \begin{pmatrix} 0 & \frac{-\pi_2}{\pi_1} & \frac{-\pi_3}{\pi_1} & \cdots & \frac{-\pi_D}{\pi_1} \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \ddots & \ddots \\ 0 & 0 & \cdots & \cdots & 1 \end{pmatrix}.$$

It follows that for any vector $\boldsymbol{x}$, it holds that

$$\boldsymbol{x}QR = \boldsymbol{x} - \frac{x_1}{\pi_1}\boldsymbol{\pi}. \tag{5.33}$$

Then it can be verified with straightforward calculations that

$$\begin{aligned} \boldsymbol{a} &= (a_1, \ldots, a_D) \\ &= R(c_1 - \lambda_1 h_{11} - c_\infty(1-\rho_\infty), \ldots, c_D - \lambda_D h_{D1} - c_\infty(1-\rho_\infty))^T \\ &= R[C - \Lambda \bar{H}_1]\boldsymbol{e} - c_\infty(1-\rho_\infty)R\boldsymbol{e} \\ &= R[C - \Lambda \bar{H}_1]\boldsymbol{e} - \boldsymbol{r}, \end{aligned} \tag{5.34}$$

is a possible solution vector, with $\boldsymbol{r} := c_\infty(1-\rho_\infty)R\boldsymbol{e}$.

Define the vector $\boldsymbol{\varphi} = (\varphi_1(s), \ldots, \varphi_D(s))$ and write Eqn. (5.5) in matrix-vector terms,

$$\boldsymbol{\varphi}(s)Q = \boldsymbol{\varphi}(s)[\Lambda \bar{H}(s) - sC] + s\boldsymbol{p}_0 C. \tag{5.35}$$

Observe that, according to Eqn. (5.7)

$$\boldsymbol{p}_0 C\boldsymbol{e} = c_\infty(1-\rho_\infty).$$

Now multiply from the right both sides of the new vector-matrix equation, Eqn. (5.35), with a D-dimensional vector of 1's, $\boldsymbol{e}$, to obtain

$$\boldsymbol{\varphi}(s)[\Lambda \bar{H}(s) - sC]\boldsymbol{e} + sc_\infty(1-\rho_\infty) = 0. \tag{5.36}$$

Multiply from the right both sides of Eqn. (5.35) with the matrix R to get

$$\boldsymbol{\varphi}(s)QR = \boldsymbol{\varphi}(s)[\Lambda \bar{H}(s) - sC]R + s\boldsymbol{p}_0 CR.$$

Rewrite this equation by using the property of Eqn. (5.33) to obtain

$$\boldsymbol{\varphi}(s) = \frac{\varphi_1(s)}{\pi_1}\boldsymbol{\pi} + \boldsymbol{\varphi}(s)[\Lambda\bar{H}(s) - sC]R + s\boldsymbol{p}_0 CR. \tag{5.37}$$

Iterate Eqn. (5.37) with itself by inserting $\boldsymbol{\varphi}(s)$ into the right hand side of the equation to obtain, after some algebraic transformations,

$$\boldsymbol{\varphi}(s) = \frac{\varphi_1(s)}{\pi_1}\boldsymbol{\pi}[I + G(s)R] + \boldsymbol{y}(s), \tag{5.38}$$

with $G(s) := \Lambda\bar{H}(s) - sC$ and

$$\boldsymbol{y}(s) := \boldsymbol{\varphi}(s)[G(s)R]^2 + s\boldsymbol{p}_0 CR[G(s)R + I]. \tag{5.39}$$

Substitute Eqn. (5.38) into Eqn. (5.36) to obtain an expression for $\varphi_1(s)$,

$$\begin{aligned} 0 &= \left[\frac{\varphi_1(s)}{\pi_1}\boldsymbol{\pi}[I + G(s)R] + \boldsymbol{y}(s)\right] \cdot G(s)\boldsymbol{e} + sc_\infty(1 - \rho_\infty) \\ &= \frac{\varphi_1(s)}{\pi_1}\left[\boldsymbol{\pi}G(s)\boldsymbol{e} + \boldsymbol{\pi}G(s)RG(s)\boldsymbol{e}\right] + \boldsymbol{y}(s)G(s)\boldsymbol{e} + sc_\infty(1 - \rho_\infty) \\ &= \frac{\varphi_1(s)}{\pi_1}\left[B_2(s) + B_3(s)\right] + B_1(s), \end{aligned} \tag{5.40}$$

with $B_1(s) = \boldsymbol{y}(s)G(s)\boldsymbol{e} + sc_\infty(1 - \rho_\infty)$, $B_2(s) = \boldsymbol{\pi}G(s)\boldsymbol{e}$ and $B_3(s) = \boldsymbol{\pi}G(s)RG(s)\boldsymbol{e}$.

The next step is to insert the scaling $s \mapsto s/N$ for each term. Recall that using the heavy traffic parametrization introduced in Section 2, we have $(1 - \rho_\infty) = 1/N$. Now observe that, as $N \to \infty$,

$$\frac{\bar{H}(s/N)}{s/N} \to \bar{H}_1, \quad \frac{\bar{H}_1 s/N - \bar{H}\,(s/N)}{(s/N)^2} \to \frac{\bar{H}_2}{2}.$$

Therefore the limit

$$\frac{G(s/N)}{s/N} \to \Lambda\bar{H}_1 - C,$$

is a constant and since $|\boldsymbol{\varphi}(s/N)| \leq 1$ and $\boldsymbol{p}_0^{(N)} \to 0$ (see Eqn. (5.7)), we have

$$\begin{aligned} \frac{\boldsymbol{y}(s/N)}{s/N} &= \boldsymbol{\varphi}(s/N)\frac{[G(s/N)R]^2}{s/N} + \boldsymbol{p}_0^{(N)} CR[G(s/N)R + I] \\ &= \boldsymbol{\varphi}(s/N)\left(\frac{s}{N}\right)\left[\frac{G(s/N)}{s/N}R\right]^2 + \boldsymbol{p}_0^{(N)} CR[G(s/N)R + I] \\ &\to 0, \end{aligned}$$

as $N \to \infty$. Combining the above we obtain

$$\begin{aligned} B_1(s/N) &= \boldsymbol{y}(s/N)G(s/N)\boldsymbol{e} + sc_\infty(1-\rho_\infty)/N \\ &= sc_\infty(1-\rho_\infty)/N + o(N^{-2}) = sc_\infty/N^2 + o(N^{-2}). \end{aligned}$$

Then

$$\begin{aligned} B_2(s/N) &= \boldsymbol{\pi}\left[\Lambda\bar{H}\left(s/N\right) - sC/N\right]\boldsymbol{e} \\ &= \boldsymbol{\pi}\Lambda\left[\bar{H}\left(s/N\right) - \bar{H}_1 s/N\right]\boldsymbol{e} + \boldsymbol{\pi}\left[\Lambda\bar{H}_1 s/N - sC/N\right]\boldsymbol{e} \\ &= \boldsymbol{\pi}\Lambda\frac{\bar{H}\left(s/N\right) - \bar{H}_1 s/N}{(s/N)^2}(s/N)^2\boldsymbol{e} - sc_\infty(1-\rho_\infty)/N \\ &= -\boldsymbol{\pi}\Lambda\frac{\bar{H}_2}{2}(s/N)^2\boldsymbol{e} + o(N^{-2}) - sc_\infty/N^2 \\ &= -(s/N)^2\sum_{d=1}^{D}\pi_d\lambda_d h_{d2}/2 - sc_\infty/N^2 + o(N^{-2}), \end{aligned}$$

since $\boldsymbol{\pi}\Lambda\bar{H}_1\boldsymbol{e} = \sum_{d=1}^{D}\pi_d\lambda_d h_{d1} = \rho_\infty c_\infty$. Furthermore,

$$\begin{aligned} B_3(s/N) &= \boldsymbol{\pi}\left[\Lambda\bar{H}\left(s/N\right) - Cs/N\right]R\left[\Lambda\bar{H}\left(s/N\right) - Cs/N\right]\boldsymbol{e} \\ &= \boldsymbol{\pi}(s/N)^2\left[\Lambda\frac{\bar{H}\left(s/N\right)}{(s/N)} - C\right]R\cdot\left[\Lambda\frac{\bar{H}\left(s/N\right)}{(s/N)} - C\right]\boldsymbol{e} \\ &= \boldsymbol{\pi}(s/N)^2\left[\Lambda\bar{H}_1 - C\right]R\left[\Lambda\bar{H}_1 - C\right]\boldsymbol{e} + o(N^{-2}) \\ &= -\boldsymbol{\pi}(s/N)^2\left[\Lambda\bar{H}_1 - C\right](\boldsymbol{a}+\boldsymbol{r}) + o(N^{-2}) \\ &= -\sum_d \pi_d\left[(a_d+r_d)(\lambda_d h_{d1} - c_d)\right](s/N)^2 + o(N^{-2}), \end{aligned}$$

due to $R\left[C - \Lambda\bar{H}_1\right]\boldsymbol{e} = \boldsymbol{a} + \boldsymbol{r}$, see Eqn. (5.34). Under the heavy-traffic scaling,

$$\boldsymbol{r} = c_\infty(1-\rho_\infty)R\boldsymbol{e} = c_\infty N^{-1}R\boldsymbol{e},$$

is an $o(1)$ term. Observe that

$$\begin{aligned} &-(B_2(s/N) + B_3(s/N)) \\ &\quad = (s/N)^2\sum_{d=1}^{D}\pi_d[\lambda_d h_{d2}/2 + (a_d + o(1))(\lambda_d h_{d1} - c_d)] \\ &\qquad + sc_\infty/N^2 + o(N^{-2}). \end{aligned}$$

Rearranging Eqn. (5.40) yields

$$\begin{aligned}\varphi_1(s/N) &= \pi_1 \frac{B_1(s/N)}{-(B_2(s/N)+B_3(s/N))} \\ &= \pi_1 \frac{c_\infty s/N^2 + o(N^{-2})}{(s/N)^2 \sum_{d=1}^{D} \pi_d[\lambda_d h_{d2}/2 + a_d(\lambda_d h_{d1} - c_d)] + c_\infty s/N^2 + o(N^{-2})} \\ &= \pi_1 \frac{1+o(1)}{1 + c_\infty^{-1} \sum_d \pi_d \left[\lambda_d h_{d2}/2 + a_d(\lambda_d h_{d1} - c_d)\right] s + o(1)}.\end{aligned}$$

Let M be the desired mean stated in Theorem 3.3, that is

$$M := c_\infty^{-1} \sum_d \pi_d \left[\hat{\lambda}_d h_{d2}/2 + a_d(\hat{\lambda}_d h_{d1} - c_d)\right].$$

Then, taking the heavy-traffic limit,

$$\lim_{N\to\infty} \boldsymbol{\varphi}(s/N) = \lim_{N\to\infty} \frac{\varphi_1(s/N)}{\pi_1} \boldsymbol{\pi} = \frac{\boldsymbol{\pi}}{1+Ms},$$

i.e. the LST $\boldsymbol{\varphi}(s)$ converges in distribution to the LST of an exponentially distributed random variable with mean M.

Bibliography

[1] I. Adan and V. Kulkarni. Single-server queue with Markov-dependent inter-arrival and service times. *Queueing Syst.*, 45(2):113–134, 2003.

[2] E. Altman, K. Avrachenkov, and U. Ayesta. A survey on discriminatory processor sharing. *Queueing Syst.*, 53(1-2):53–63, 2006.

[3] E. Altman, T. Jimenez, and D. Kofman. DPS queues with stationary ergodic service times and the performance of TCP in overload. *Proceedings of IEEE Infocom*, 2:975–983, 2004.

[4] D. Anderson, J. Blom, M. Mandjes, H. Thorsdottir, and K. de Turck. A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodol. Comput. Appl.*, 18(1):153–168, 2014.

[5] D. Anderson and T. Kurtz. Continuous-time Markov chain models for chemical reaction networks. In H. Koeppl et al., editor, *Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology*, pages 3–42. Springer, 2011.

[6] A. Arazia, E. Ben-Jacob, and U. Yechiali. Bridging genetic networks and queueing theory. *Physica A*, 332:585–616, 2004.

[7] S. Asmussen. The heavy traffic limit of a class of Markovian queueing models. *Oper. Res. Letters*, 6:301–306, 1987.

[8] S. Asmussen. Ladder heights and the Markov-modulated M/G/1 queue. *Stoch. Proc. Appl.*, 37(2):313–326, 1991.

[9] S. Asmussen. *Applied Probability and Queues*. Springer Science & Business Media, 2nd edition, 2003.

[10] S. Asmussen and O. Kella. Rate modulation in dams and ruin problems. *J. Appl. Probab.*, 33:523–535, 1996.

[11] K. Avrachenkov, U. Ayesta, P. Brown, and R. Núñez-Queija. Discriminatory processor sharing revisited. *Proceedings of IEEE Infocom*, 2:784–795, 2005.

[12] U. Ayesta, A. Izagirre, and I. M. Verloop. Heavy-traffic analysis of a multi-class queue with relative priorities. *Probab. Engrg. Inform. Sci.*, 29(2):153–180, 2015.

[13] K. Ball, T. Kurtz, L. Popovic, and G. Rempala. Asymptotic analysis of multi-scale approximations to reaction networks. *Ann. Appl. Probab.*, 16:1925–1961, 2006.

[14] M. Baykal-Gürsoy and W. Xiao. Stochastic decomposition in $M/M/\infty$ queues with Markov-modulated service rates. *Queueing Syst.*, 48(1-2):75–88, 2004.

[15] R. Bekker. *Queues with State-Dependent Rates*. PhD thesis, Eindhoven University of Technology, 2005.

[16] R. N. Bhattacharya. On the functional central limit theorem and the law of the iterated logarithm for Markov processes. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 60:185–201, 1982.

[17] P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2nd edition, 1999.

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

[19] J. Blom, K. De Turck, and M. Mandjes. Analysis of Markov-modulated infinite-server queues in the central-limit regime. *Probab. Engrg. Inform. Sci.*, 29(3):433–459, 2015.

[20] J. Blom, K. De Turck, and M. Mandjes. Functional central limit theorems for Markov-modulated infinite-server systems. *Math. Method. Oper. Res.*, 2015.

[21] J. Blom, O. Kella, M. Mandjes, and H. Thorsdottir. Markov-modulated infinite-server queues with general service times. *Queueing Syst.*, 76(4):403–424, 2014.

[22] J. Blom, M. Mandjes, and H. Thorsdottir. Time-scaling limits for Markov-modulated infinite-server queues. *Stoch. Models*, 29(1):112–127, 2013.

[23] A. Borovkov. On limit laws for service processes in multi-channel systems. *Siberian Math. J.*, 8:746–763, 1967.

[24] S. Borst, R. Núñez-Queija, and B. Zwart. Sojourn time asymptotics in processor-sharing queues. *Queueing Syst.*, 53(1-2):31–51, 2006.

[25] S. Borst, D. van Ooteghem, and B. Zwart. Tail asymptotics for discriminatory processor-sharing queues with heavy-tailed service requirements. *Perform. Eval.*, 61(2):281–298, 2005.

[26] O. J. Boxma and I. A. Kurkova. The M/G/1 queue with two service speeds. *Adv. Appl. Probab.*, 33(2):520–540, 2001.

[27] A. Budhiraja, A. Ghosh, and X. Liu. Scheduling control for Markov-modulated single-server multiclass queueing systems in heavy traffic. *Queueing Syst.*, 78(1):57–97, 2014.

[28] D. Y. Burman and D. R Smith. An asymptotic analysis of a queueing system with Markov-modulated arrivals. *Oper. Res.*, 34(1):105–119, 1986.

[29] G. L. Choudhury, A. Mandelbaum, M. I. Reiman, and W. Whitt. Fluid and diffusion limits for queues in slowly changing environments. *Stoch. Models*, 13(1):121–146, 1997.

[30] N. Cookson, W. Mather, T. Danino, O. Mondragón-Palomino, R. Williams, L. Tsimring, and J. Hasty. Queueing up for enzymatic processing: Correlated signaling through coupled degradation. *Mol. Syst. Biol.*, 7(1):561, 2011.

[31] P. Coolen-Schrijner and E. van Doorn. The deviation matrix of a continuous-time Markov chain. *Probab. Engrg. Inform. Sci.*, 16:351–366, 2002.

[32] B. D'Auria. $M/M/\infty$ queues in semi-Markovian random environment. *Queueing Syst.*, 58(3):221–237, 2008.

[33] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2nd edition, 1998.

[34] M. Dimitrov. Single-server queueing system with Markov-modulated arrivals and service times. *Pliska Stud. Math. Bulgar.*, 20:53–62, 2011.

[35] J. L. Dorsman, M. Vlasiou, and B. Zwart. Heavy-traffic asymptotics for networks of parallel queues with Markov-modulated service speeds. *Queueing Syst.*, 79(3-4):293–319, 2015.

[36] A. Economou and D. Fakinos. The infinite server queue with arrivals generated by a non-homogeneous compound Poisson process and heterogeneous customers. *Stoch. Models*, 15(5):993–1002, 1999.

[37] S. Eick, W. Massey, and W. Whitt. The physics of the $M_t/G/\infty$ queue. *Oper. Res.*, 41:731–742, 1993.

[38] M. Eisen and M. Tainiter. Stochastic variations in queuing processes. *Oper. Res.*, 11(6):922–927, 1963.

[39] S. Ethier and T. Kurtz. *Markov Processes: Characterization and Convergence*. John Wiley & Sons, 1986.

[40] G. Falin. The M/M/ ∞ queue in a random environment. *Queueing Syst.*, 58(1):65–76, 2008.

[41] G. I. Falin and A. I. Falin. Heavy traffic analysis of M/G/1 type queueing systems with Markov-modulated arrivals. *TOP*, 7(2):279–291, 1999.

[42] G. Fayolle, I. Mitrani, and R. Iasnogorodski. Sharing a processor among many job classes. *J. ACM*, 27(3):519–532, 1980.

[43] D. Fiems and E. Altman. Markov-modulated stochastic recursive equations with applications to delay-tolerant networks. *Perform. Eval.*, 70(11):965–980, 2013.

[44] B. Fralix and I. Adan. An infinite-server queue influenced by a semi-Markovian environment. *Queueing Syst.*, 61(1):65–84, 2009.

[45] D. Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:35–55, 2007.

[46] P. Glynn. Large deviations for the infinite server queue in heavy traffic. In F. P. Kelly and R. J. Williams, editors, *Stochastic Networks*, volume 71 of *Mathematics and its Applications*, pages 387–394. Institute for Mathematics and Its Applications, 1995.

[47] P. Glynn and W. Whitt. A new view of the heavy-traffic limit theorem for infinite-server queues. *Adv. Appl. Probab.*, 23:188–209, 1991.

[48] S. Grishechkin. On a relationship between processor-sharing queues and Crump-Mode-Jagers branching processes. *Adv. Appl. Probab.*, 24(3):653–698, 1992.

[49] V. Gupta, M. Burroughs, and M. Harchol-Balter. Analysis of scheduling policies under correlated job sizes. *Perform. Eval.*, 67(11):996–1013, 2010.

[50] R. Hassin and M. Haviv. *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59. Springer Science & Business Media, 2003.

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

[52] D. Iglehart. Limiting diffusion approximations for the many server queue and the repairman problem. *J. Appl. Probab*, 2:429–441, 1965.

[53] D. Iglehart and W. Whitt. Multiple channel queues in heavy traffic. I. *Adv. Appl. Probab*, 2(1):150–177, 1970.

[54] A. Izagirre, U. Ayesta, and I. M. Verloop. Sojourn time approximations in a multi-class time-sharing server. *Proceedings of IEEE Infocom*, pages 2786–2794, 2014.

[55] A. Izagirre, U. Ayesta, and I.M. Verloop. Interpolation approximations for the steady-state distribution in multi-class resource-sharing systems. *Perform. Eval.*, 91:56–79, 2015.

[56] J. Jacod and A. Shiryayev. *Limit Theorems for Stochastic Processes.* Springer, 1987.

[57] H. Kang and T. Kurtz. Separation of time-scales and model reduction for stochastic reaction networks. *Ann. Appl. Probab.*, 23:529–583, 2013.

[58] T. Katsuda. State-space collapse in stationarity and its application to a multiclass single-server queue in heavy traffic. *Queueing Syst.*, 65(3):237–273, 2010.

[59] J. Keilson and L. Servi. The matrix M/M/∞ system: retrial models and Markov-modulated sources. *Adv. Appl. Probab.*, 25:453–471, 1993.

[60] O. Kella and W. Stadje. Markov-modulated linear fluid networks with Markov additive input. *J. Appl. Probab.*, 39:413–420, 2002.

[61] O. Kella and W. Whitt. Linear stochastic fluid networks. *J. Appl. Probab.*, 36:244–260, 1999.

[62] B. Kim and J. Kim. A single server queue with Markov modulated service rates and impatient customers. *Perform. Eval.*, 83:1–15, 2015.

[63] J.F.C. Kingman. The single server queue in heavy traffic. *Proc. Cambridge Philos.*, 57(4):902–904, 1961.

[64] L. Kleinrock. Time-shared systems: A theoretical treatment. *J. ACM*, 14(2):242–261, 1967.

[65] L. Kleinrock. *Queueing Systems, vol. 2.* John Wiley & Sons, 1976.

[66] T. Kurtz and P. Protter. Wong-Zakai corrections, random evolutions, and simulation schemes for SDEs. In E. Mayer-Wolf, E. Merzbach, and A. Schwartz, editors, *Stochastic Analysis*, pages 331–346. Academic Press, 1991.

[67] L. Liu and J. Templeton. Autocorrelations in infinite server batch arrival queues. *Queueing Syst.*, 14(3-4):313–337, 1993.

[68] S. Mahabhashyam and N. Gautam. On queues with Markov modulated service rates. *Queueing Syst.*, 51(1-2):89–113, 2005.

[69] W.A. Massey. The analysis of queues with time-varying rates for telecommunication models. *Telecommun. Syst.*, 21(2-4):173–204, 2002.

[70] M. Neuts and S. Chen. The infinite server queue with semi-Markovian arrivals and negative exponential services. *J. Appl. Probab.*, 9:178–184, 1972.

[71] M. F. Neuts. *Matrix-geometric solutions in stochastic models: an algorithmic approach.* Johns Hopkins University Press, 1981.

[72] M.F. Neuts. A queue subject to extraneous phase changes. *Adv. Appl. Probab.*, 3:78–119, 1971.

[73] J. Norris. *Markov Chains.* Cambridge University Press, 1997.

[74] R. Núñez-Queija. *Processor-Sharing Models for Integrated-Services Networks.* PhD thesis, Eindhoven University of Technology, 2000.

[75] C. O'Cinneide and P. Purdue. The M/M/∞ queue in a random environment. *J. Appl. Probab.*, 23:175–184, 1986.

[76] G. Pang, R. Talreja, and W. Whitt. Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surveys*, 4:193–267, 2007.

[77] G. Pang and W. Whitt. Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Syst.*, 65(4):325–364, 2010.

[78] N.U. Prabhu and Y. Zhu. Markov-modulated queueing system. *Queueing Syst.*, 5(1):215–246, 1989.

[79] P. Purdue. The M/M/1 queue in a Markovian environment. *Oper. Res.*, 22:562–569, 1974.

[80] P. Purdue and D. Linton. An infinite-server queue subject to an extraneous phase process and related models. *J. Appl. Probab.*, 18:236–244, 1981.

[81] R. Rebolledo. Central limit theorems for local martingales. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 51:269–286, 1980.

[82] J. Reed and R. Talreja. Distribution-valued heavy-traffic limits for the G/GI/∞ queue. *Ann. Appl. Probab.*, 25(3):1420–1474, 2015.

[83] K. M. Rege and B. Sengupta. Queue-length distribution for the discriminatory processor-sharing queue. *Oper. Res.*, 44(4):653–657, 1996.

[84] G. J. K. Regterschot and J. H. A. de Smit. The queue M/G/1 with Markov-modulated arrivals and service. *Math. Opns. Res.*, 11(3):465–483, 1986.

[85] Ph. Robert. *Stochastic Networks and Queues*. Springer, 2003.

[86] A. Schwabe. *Non-genetic cell-to-cell variability: theory and experiments*. PhD thesis, Vrije Universiteit Amsterdam, 2014.

[87] A. Schwabe, K. Rybakova, and F. Bruggeman. Transcription stochasticity of complex gene regulation models. *Biophys. J.*, 103:1152–1161, 2012.

[88] B. Sengupta. Sojourn time distributions for the M/M/1 queue in a Markovian environment. *Eur. J. Oper. Res.*, 32:140–149, 1987.

[89] T. Takine. Single-server queues with Markov-modulated arrivals and service speed. *Queueing Syst.*, 49(1):7–22, 2005.

[90] H. Thorsdottir and I. M. Verloop. Markov-modulated M/G/1-type queue in heavy traffic and its application to time-sharing disciplines. *Queueing Syst.*, 82(1):1–27, 2016.

[91] I. M. Verloop, U. Ayesta, and R. Núñez-Queija. Heavy-traffic analysis of a multiple-phase network with discriminatory processor sharing. *Oper. Res.*, 3(59):648–660, 2011.

[92] A.R. Ward and P.W. Glynn. A diffusion approximation for a Markovian queue with reneging. *Queueing Syst.*, 43(1-2):103–128, 2003.

[93] W. Whitt. *Stochastic-process Limits*. Springer, New York, 2001.

[94] W. Whitt. Proofs of the martingale FCLT. *Probab. Surveys*, 4:268–302, 2007.

[95] G.E. Willmot and S. Drekic. Time-dependent analysis of some infinite server queues with bulk Poisson arrivals. *INFOR*, 47(4):297–303, 2009.

[96] Y. Wu, L. Bui, and R. Johari. Heavy traffic approximation of equilibria in resource sharing games. *IEEE J. Sel. Area. Comm.*, 30(11):2200–2209, 2012.

[97] U. Yechiali and P. Naor. Queueing problems with heterogeneous arrival and service. *Oper. Res.*, 19:722–734, 1971.

[98] A.V. Zorine. On ergodicity conditions in a polling model with Markov modulated input and state-dependent routing. *Queueing Syst.*, 76(2):223–241, 2014.

Publications

The work in this thesis is based on the articles mentioned below. Chapters 2, 3, 4 and 5 contain the material of [22], [21], [4] and [90], respectively.

Halldora Thorsdottir carried out the research and writing of [22, 21, 4] in collaboration with Michel Mandjes (promotor) and Joke Blom (daily supervisor). After getting acquainted with the subject and techniques in [22], gradually increasing contributions followed in [21, 4]. Offer Kella (Hebrew University of Jerusalem) contributed in particular to Section 2 in [21]. David Anderson (University fo Wisconsin, Madison) introduced the framework and main tools of [4] to his co-authors, while also sketching out the rough result. Among other important contributions, Koen de Turck provided a useful martingale presented in Lemma 3.1.

The work in [90] is the result of a collaboration with Maaike Verloop (CNRS IRIT, Toulouse), as well as discussions with Urtzi Ayesta (co-promotor, CNRS LAAS, Toulouse). Halldora took the lead in setting up this collaboration and formulating the research question.

List of publications

[22] J. Blom, M. Mandjes, and H. Thorsdottir. Time-scaling limits for Markov-modulated infinite-server queues. *Stoch. Models*, 29(1):112–127, 2013.

[21] J. Blom, O. Kella, M. Mandjes, and H. Thorsdottir. Markov-modulated infinite-server queues with general service times. *Queueing Syst.*, 76(4):403–424, 2014.

[4] D. Anderson, J. Blom, M. Mandjes, H. Thorsdottir, and K. de Turck. A functional central limit theorem for a Markov-modulated infinite-server queue. *Methodol. Comput. Appl.*, 18(1):153–168, 2014.

[90] H. Thorsdottir and I. M. Verloop. Markov-modulated M/G/1-type queue in heavy traffic and its application to time-sharing disciplines. *Queueing Syst.*, 82(1):1–27, 2016.

Summary

This thesis considers queueing systems affected by a random environment, with the primary focus being that of evaluating the performance of these queues in specific asymptotic regimes. Embedding a queueing system in a random environment is a way to add flexibility to a model. This flexibility comes at the cost of increased complexity, in that the queue becomes a doubly random system, i.e. the already stochastic arrival and service processes are also assumed to have randomly fluctuating parameters governed by the external environment. In the scope of this thesis the environment can in theory represent anything that can be modelled as a random process on a discrete state space, such as weather dynamics or the state of the economy. We frequently refer to the environment as the background process or modulating process. In addition to an introduction, the thesis consists of four chapters; Chapters 2 to 5 are based on journal papers that have been published. Chapters 2, 3 and 4 all study infinite server systems, whereas in Chapter 5 a single server queue is studied.

Chapters 2, 3 and 4 all exploit the concept of time-scale separation. The scaling of choice is applied to both the environment and the arrival process. Both are pushed to infinity albeit at different speeds, imposing a central limit theorem (CLT) type of scaling. On the one hand, when the environment is sped up more than the arrival process, it will only be perceived as an average from the perspective of the main process, the queue length. Instead of having multiple arrival and service rates due to the modulation, in the limit one effectively only observes an average arrival and service rate, which greatly simplifies the analysis. On the other hand, slowing down the environment relative to the arrivals, as in Chapter 4, yields a sequence of temporary steady-states. In that case the deviation between the transient and the equilibrium distribution of the environment, expressed in terms of the so-called deviation matrix for Markov chains, plays a crucial role.

Under the CLT scaling, Gaussian limits are derived. In Chapters 2 and 3, which contain results at the transient level, the limiting distribution is identified as the normal distribution, whereas the functional CLT of Chapter 4 results in a limiting Gaussian process of the Ornstein-

Uhlenbeck (OU) type. The heavy-traffic scaling in Chapter 5 lets the arrival rate be increased such that the traffic intensity reaches its critical point. Typically the scaling yields reflected Brownian motion limits; here its stationary counterpart, the exponential distribution, is obtained as the limiting distribution of the steady-state workload and queue length.

The methodology of Chapters 2, 3 and 5 is likely familiar to the reader of queueing literature. We set off with fixed-point equations to describe the infinite server queue in Chapters 2 and 3, and balance equations for the M/G/1 queue in Chapter 5. Due to the modulating background process the equations become particularly uninviting. By applying the right scaling and Taylor expansion, the equations simplify considerably, which helps in deriving the limits. The work in Chapter 4 is derived under a different framework and methodology, primarily based on unit-rate Poisson processes and the martingale CLT. The main advantage of this toolkit is that it yields a functional limiting result, whereas the methods of the other chapters yield finite-dimensional convergence.

While the main topic of this thesis is the behaviour of modulated queues in particular scaling regimes, it contains specific results for non-scaled processes as well, presented in Chapters 3 and 5. These results, which are obtained using transforms, are primarily in terms of recursions for moments and differential equations that describe properties of the distribution of the number of customers and workload.

We summarize the main results of the thesis. After speeding up the environment that modulates the Poisson arrivals to an infinite server queue, the arrival process is shown to be asymptotically Poisson with a uniform rate, see Chapter 2. By also speeding up the arrival rates, the scaled and centered queue length converges to a normally distributed random variable. Here the background process has deterministic transition times, yielding a semi-Markov-modulated system. These results are extended in Chapter 3 to a multi-dimensional CLT for an $M/G/\infty$ queue under Markov-modulation. Chapter 4 contains a functional CLT for the queue length process with Markov-modulated arrivals and nonmodulated, exponential service times. The martingale CLT is applied to prove weak convergence to an OU process, where the environment moves either faster or slower than the arrival process. In Chapter 5, assuming generally distributed service requirements and a fairly general class of service disciplines, it is shown that the workload of an M/G/1 queue with Markov-modulated service capacity converges to an exponentially distributed random variable in heavy traffic. The discriminatory processor sharing service discipline is applied to the case of exponentially distributed service requirements and multiple customer classes. The queue

length vector for the various classes undergoes a state-space collapse in the limit of heavy-traffic scaling, also giving the exponential distribution.

Samenvatting

Dit proefschrift onderzoekt wachtrijsystemen die worden beïnvloed door een aan toeval onderhevige omgeving. De primaire focus ligt op de karakterisering van het gedrag van deze wachtrijen in bepaalde asymptotische schalingsregimes. Door een wachtrijsysteem te beschouwen in een stochastisch-fluctuerende (in plaats van statische) omgeving, wint het model aan flexibiliteit en realisme. Hier staat echter tegenover dat de complexiteit toeneemt, in de zin dat de wachtrij een 'dubbel stochastisch' systeem wordt; de aankomst- en verwerkingsprocessen die in standaard wachtrijsystemen al stochastisch zijn, krijgen nu ook stochastisch variërende, door de omgeving bepaalde, parameters.

In dit proefschrift correspondeert het omgevingsproces, ook wel *modulerend proces* of *achtergrondproces*, met een stochastisch proces op een discrete toestandsruimte; in toepassingen kan men bijvoorbeeld denken aan een proces dat de dynamiek van weersomstandigheden beschrijft of van de toestand van de economie. Naast een inleiding, bestaat het proefschrift uit vier hoofdstukken, die reeds gepubliceerd zijn. In hoofdstukken 2, 3 en 4 worden systemen bestudeerd met een oneindig aantal verwerkingseenheden (zogenaamde *infinite-server systemen)*, in hoofdstuk 5 een wachtrij met een enkele server (een *single-server* systeem).

In de hoofdstukken 2, 3 en 4 wordt gebruik gemaakt van het principe van scheiding van tijdschalen: één bepaalde schaling wordt toegepast op het omgevingsproces, en een ander op het aankomstproces. Onder een schaling waar beide processen 'oneindig worden versneld', maar met ongelijke factoren, geldt een Centrale Limietstelling (*central limit theorem*, CLT). Wanneer het omgevingsproces meer versneld wordt dan het aankomstproces, dan wordt alleen het *gemiddelde* effect van de omgeving ervaren door het wachtrijproces: vanwege de modulatie zijn er meerdere aankomst- en verwerkingssnelheden, maar in de limiet ervaart de wachtrij slechts één aankomst- en één verwerkingssnelheid. Wordt, daarentegen, het omgevingsproces minder versneld dan het aankomstproces (zoals in hoofdstuk 4), dan krijgen we een aaneenschakeling van 'tijdelijke' *steady-states*. In dit geval speelt de afwijking tussen de tijdsafhankelijke en de evenwichtsverdeling van het omgevingsproces een belangrijke rol; deze wordt uitgedrukt in termen van de zogenaamde *deviatie-matrix*.

Onder de CLT schaling die we hierboven bespraken worden Gaussische limieten afgeleid. In de hoofdstukken 2 en 3 zijn dit resultaten die betrekking hebben op een enkel punt in de tijd, waarin de wachtrij asymptotisch normaal verdeeld is. De CLT van hoofdstuk 4 is daarentegen *in functionaalvorm*: hij beschrijft een Gaussisch limietproces van het Ornstein-Uhlenbeck-type. De *heavy-traffic* schaling van hoofdstuk 5 laat de aankomstfrequenties toenemen zodat de verkeersintensiteit een kritisch niveau benadert. Normaal gesproken leidt deze schaling tot een *reflected Brownian motion* (RBM) limiet; hier wordt de stationaire variant van de RBM, de exponentiële verdeling, gevonden als de limietverdeling van de stationaire *workload* én de wachtrijlengte.

De lezer van de wachtrijliteratuur zal waarschijnlijk bekend zijn met de methodologie die wordt gebruikt in de hoofdstukken 2, 3 en 5. In de hoofdstukken 2 en 3 beginnen we met vastpuntsvergelijkingen om de *infinite-server* systemen te beschrijven; in hoofdstuk 5 is het startpunt het opstellen van de balansvergelijkingen voor de M/G/1 wachtrij. Vanwege de modulatie door het achtergrondproces, zijn de resulterende vergelijkingen echter zeer gecompliceerd. Door nu gebruik te maken van onder onze schaling geldende benaderingen (onder andere gebaseerd op Taylor-ontwikkelingen) worden deze vereenvoudigd, zodat limietresultaten afgeleid kunnen worden. In hoofdstuk 4 wordt gebruik gemaakt van een ander kader en een andere methodologie, voornamelijk gebaseerd op de *martingale CLT*. Het voordeel hiervan is dat het leidt tot een limiet in functionaalvorm, terwijl de methoden uit de andere hoofdstukken leiden tot convergentieresultaten voor eindig-dimensionale stochasten.

Het hoofdonderwerp van dit proefschrift is hoe gemoduleerde wachtrijen zich gedragen onder bepaalde schalingen, maar het bevat ook resultaten voor ongeschaalde processen (zie hoofdstuk 3 en 5). Deze resultaten, gevonden met gebruik van transformaties, zijn voornamelijk recursies voor momenten en differentiaalvergelijkingen die de eigenschappen beschrijven van de verdeling van de *workload* en de hoeveelheid klanten in het systeem.

Nu volgen de belangrijkste resultaten van dit proefschrift. In hoofdstuk 2 wordt bewezen dat door het versnellen van het omgevingsproces, dat de Poisson-aankomsten van een *infinite-server* wachtrij moduleert, het aankomstproces zich gedraagt als een uniform (d.w.z. niet-gemoduleerd) Poissonproces. Als de aankomstfrequenties zelf ook nog worden versneld, convergeert de geschaalde en gecentreerde wachtrijlengte naar een normaal-verdeelde stochast. In dit hoofdstuk heeft het achtergrondproces deterministische transitietijden, waardoor het een *semi-Markov-*

gemoduleerd proces is. Het resultaat wordt in hoofdstuk 3 uitgebreid naar een multi-dimensionale CLT voor een $M/G/\infty$ wachtrij met Markov-modulatie. Hoofdstuk 4 bevat een CLT in functionaalvorm voor het wachtrijproces met Markov-gemoduleerde aankomsten en exponentiële (niet-gemoduleerde) verwerkingstijden. De *martingale CLT* wordt toegepast om zwakke convergentie (d.w.z. convergentie in verdeling) naar een OU proces te bewijzen, daarbij onderscheid makend of het omgevingsproces sneller dan wel langzamer is dan het aankomstproces. In hoofdstuk 5 wordt, voor willekeurige verdeelde bedieningseisen en voor een ruime klasse van bedieningsdisciplines, bewezen dat de *workload* van een $M/G/1$ queue met Markov-gemoduleerde bedieningscapaciteit in *heavy traffic* convergeert naar een exponentiëel verdeelde stochast. De *discriminatory processor sharing* bedieningsdiscipline is toegepast voor het geval van exponentiëel verdeelde bedieningseisen en verschillende soorten klanten. De vector van de wachtrijlengte voor de verschillende soorten ondergaat een *state-space collapse* in de limiet van *heavy-traffic* schaling, en wordt ook exponentiëel verdeeld.

Acknowledgements

Over the course of more than four years, many things have influenced the outcome that is this thesis, both people and circumstances. During my masters project I got a taste of the joy and challenge of doing research in the great workplace that is CWI. This is also thanks to my wonderful colleagues there, and my supervisor Jason Frank, who suggested I apply for what came to be my PhD project.

What a great choice of project; I got to enter the intriguing world of stochastics with the help of my two excellent advisors, Michel Mandjes and Joke Blom. Their enthusiasm for mathematics, that requires one to work hard with great precision, along with their continuous support, was exactly the framework I needed to finish this thesis. Thank you for the many discussions and for pushing me to get everything possible out of being an *OiO*, a researcher in training. I don't think I could have asked for better supervisors.

Special thanks to Maaike Verloop and Urtzi Ayesta for receiving me so well in Toulouse and for the enjoyable cooperation. Thanks as well to my coauthors, Koen de Turck, David Anderson and Offer Kella.

I have made many great friends during my time at CWI, inspiring people willing to discuss anything, no matter how crazy it might sound. To my two office mates, Keith and Jaldert, thanks to you both for being such intellectual explorers, I had so much fun. Thanks also to my dear paranymphs, Thibault and Paula, always up for talking about life and love and even science. I mention also my always social life-science colleagues, fellow Praethuys attenders and CWI supportive staff.

Finally, thanks to my family for giving me the spirit to seek adventure abroad and especially my grandfather Tómas, for showing his enthusiasm for me conducting research. And to Bouwe, who has made my life in Amsterdam such a happy one.