A large deviations principle for infinite-server queues in a random environment

H.M. Jansen, Michel Mandjes, Koen de Turck, S Wittevrongel

A large deviations principle for infinite-server queues in a random environment

H. M. Jansen[1,2], M. R. H. Mandjes[1], K. De Turck[2], S. Wittevrongel[2]

December 22, 2015

Abstract

This paper studies an infinite-server queue in a random environment, meaning that the arrival rate, the service requirements and the server work rate are modulated by a general càdlàg stochastic background process. To prove a large deviations principle, the concept of attainable parameters is introduced. Scaling both the arrival rates and the background process, a large deviations principle for the number of jobs in the system is derived using attainable parameters. Finally, some known results about Markov-modulated infinite-server queues are generalized and new results for several background processes and scalings are established in examples.

Keywords. infinite-server queue ⋆ random environment ⋆ modulation ⋆ large deviations principle

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

[2] TELIN, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium.

E-mail. {h.m.jansen|m.r.h.mandjes}@uva.nl, {kdeturck|sw}@telin.ugent.be

1 Introduction

The infinite-server queue is one of the fundamental models in queueing theory. Its distinguishing feature is the presence of an infinite number of servers, so that jobs are served independently and there are no waiting times. This leads to explicit formulas for many quantities of interest, especially for M/M/∞ queues, where jobs arrive according to a Poisson process and the service requirements have an exponential distribution. In practice, however, one often observes time-varying arrival intensities, service requirement distributions and server work rates. This calls for adequate modeling.

A natural way to incorporate time-dependence is to consider an infinite-server queue in a random environment. In this case there is an independent

background process that modulates the arrival rate, the service requirement distribution and the work rate of the servers.

Model. In this paper, we study the case where the background process is a general stochastic process J whose paths are right-continuous and have finite left limits, i.e., J has càdlàg paths. The process J modulates the arrival rate, the service requirement distribution and the server work rate in the following way. When J is in state x, jobs arrive according to a Poisson process with intensity $\lambda(x)$. Upon arrival, a job draws an independent service requirement with distribution F_x if J is in state x when the job arrives. Then the service requirement of the job is processed by a server, whose work rate is $\mu(x)$ while J is in state x. Immediately after its service requirement has been processed, a job leaves the system.

Main result. The main result of this paper is a full large deviations principle (LDP) for the transient number of jobs in the system, under a scaling of the arrival rate and the background process. To arrive at this result, we first show that the number of jobs in the system at time $t \geq 0$ has a Poisson distribution with random parameter $\phi_t(J)$. Then we scale $\lambda \mapsto n\lambda$ and we scale $J \mapsto J_n$ such that the normalized random parameter $\phi_t(J_n)$ satisfies an LDP. Under this scaling, we derive the LDP for the transient number of jobs in the system.

Literature. The amount of literature related to our main result is quite small. Moreover, almost all papers on infinite-server queues in a random environment (with notable exception [5]) study Model I or Model II (cf. [3]). In both models, jobs arrive according to a Poisson process with intensity $\lambda(x)$ when the background process is in state x. In Model I, service requirements have a standard exponential distribution and servers work at rate $\mu(x)$ when the background process is in state x. This is equivalent to the jobs being subject to a modulated hazard rate. In Model II, service requirements have an exponential distribution with parameter $\kappa(x)$ when the background process is in state x and servers work at constant rate 1.

An early reference is [17], which analyzes Model I when the background process is a continuous-time Markov chain. Important results in [17] are a recursion for the factorial moments of the number of jobs and the observation that the steady-state distribution is not of some 'matrix-Poisson' type.

Other important results can be found in [8], which studies Model I when the background process is a semi-Markov process with finite state space. The crucial observation in [8] is that the stationary number of jobs has a Poisson distribution with a random parameter that is determined by the background process. Moreover, the factorial moments of the number of jobs are computed via a recursion. These results are generalized in [14].

The observation in [8] is used to obtain time-scaling results in both the central limit regime and the large deviations regime. In the central limit regime, [2] and [4] derive central limit theorems for Markov-modulated infinite-server queues for several models and scalings. In this regime, the so-called deviation matrix (cf. [7]) plays an important role. In the large deviations regime, [3] and [5] compute optimal paths to obtain rate functions under a linear scaling of the arrival rates, given that the background process is an irreducible continuous-time

Markov chain. The former studies Model I, whereas the latter studies Model II for a class of service requirement distributions that includes the exponential distribution.

As mentioned, we show that the number of jobs in the system has a Poisson distribution with a random parameter. This means that the probability distribution of the number of jobs in the system is a mixture of Poisson distributions. This significantly complicates a large deviations analysis: even in elementary cases a mixture may not satisfy an LDP. Nevertheless, papers such as [1], [6] and [11] have studied large deviations of mixtures and identified conditions under which a mixture does satisfy an LDP. However, our model does not fit into the framework of these publications. We will elaborate on this in the next section.

Contributions. In more detail, the contributions of this paper are the following. We generalize known models by considering a general càdlàg background process instead of a semi-Markov background process with finite state space. We also generalize known models by including general service requirement distributions. Moreover, in our model the background process modulates both the service requirement distributions and the server work rate, whereas previous papers considered models in which either the service requirement distributions or the server work rate was modulated. In particular, our model generalizes Model I and Model II.

Using elementary arguments, we show that in this model the transient number of jobs has a Poisson distribution with random parameter. We scale the arrival rate linearly and we scale the background process such that the normalized random parameter satisfies an LDP. Under this scaling, we obtain a full LDP for the number of jobs in the system. To the best of our knowledge, this is the first time that a full LDP is presented for modulated infinite-server queues. To prove the LDP, we exploit properties of the queueing system, introduce the concept of attainable parameters and use a generalization of Varadhan's Lemma. These tools enable us to avoid the assumptions in [1], [6] and [11].

The theory is illustrated by examples that show rate functions that cannot be obtained via background processes with finite state space. Additionally, we show that completely different background processes may lead to the same LDP, even in highly nontrivial cases. We also show examples that do not fit into the framework of [1], [6] and [11].

Organization. The rest of this paper is organized as follows. In Section 2, we give a more practical motivation to study our model en discuss some of the technical subtleties. In Section 3, we describe the model and provide some of its basic properties. Additionally, we fix some notation. In Section 4, we introduce the concept of attainable parameters and prove an LDP for the number of jobs in the system. In Section 5, we show that the rate function corresponding to this LDP has a simple description when we do not scale the background process. As an illustration, we work out some examples. In Section 6, we give examples in which we do scale the background process. In Section 7, we briefly discuss the results and point out some topics for future research. The appendices provide some technical details about the number of jobs in the system (Section A),

continuity and convergence in Skorokhod space (Section B) and properties of Poisson random variables (Section C).

2 Motivation

Modulated infinite-server queues are used to model various phenomena in communications systems, road traffic and hospital capacity planning, for example. For us, the main practical motivation to study this model stems from biology. It is well known that the production of molecules in a cell may be 'bursty', meaning that periods of high production activity are followed by periods of low production activity. In [12], this phenomenon is modelled using an interrupted Poisson process, i.e., a Poisson process that is modulated by a stochastic ON/OFF switch. After a molecule is produced, it degrades during a random time interval. The resulting model for the number of molecules (mRNA in the case of [12]) still present in the cell is, essentially, a simple modulated infinite-server queue. However, other publications (cf. [18, p. 22]) indicate that for more complicated production processes one may need a more general background process with a larger state space.

Another important observation (cf. [19, pp. 605-606]) is that the production process and the switching process may be on different time scales. This gives rise to different regimes (cf. [19, Fig. 28.4]). Mathematically, this phenomenon is captured by the modulated infinite-server queue via a linear scaling of the arrival rate and a scaling of the switching process.

Motivated by these observations, we study the modulated infinite-server queue using very few assumptions on the background process and the service requirement distributions. In particular, we are interested in the large deviations behavior of the number of jobs (molecules) in the system under a linear scaling of the arrival rate and a very general scaling of the background process. Our goal is to prove a full large deviations principle for the number of jobs in the system.

From a more mathematical point of view, this leads to some interesting problems. First of all, the model that we consider includes general service requirements and generalizes two well-known models. Next to that, we impose very few assumptions on the background process and its scaling, whereas other studies use semi-Markov processes with finite state space and specific scalings. The scaled background processes considered here induce rate functions that do not necessarily have compact level sets. This leads to a large deviations problem that seems not to have been discussed before. We will explain this in more detail.

In the next section, we introduce a linear scaling of the arrival rate $\lambda \mapsto n\lambda$ and a general scaling of the background process $J \mapsto J_n$. We denote the number of jobs in the system at time t by $M_n(t)$ and we would like to prove an LDP for $\frac{1}{n}M_n(t)$. It turns out that $\frac{1}{n}M_n(t)$ is a mixture of Poisson distributions with a certain mixing measure ν_n. Large deviations of mixtures have been investigated in papers such as [1], [6] and [11]. However, our model does not fit into the framework of these publications. We will indicate why.

A probability measure $\mathbb{Q}_n(\cdot)$ is called a mixture if there exist a family of probability measures $\{\mathbb{Q}_n(\theta;\cdot) : \theta \in \Theta\}$ (where the $\mathbb{Q}_n(\theta;\cdot)$ are the 'conditional' probabilities) and a probability measure ν_n on Θ (the mixing measure) such that $\mathbb{Q}_n(F) = \int_\Theta \mathbb{Q}_n(\theta;F)\,\mathrm{d}\nu_n(\theta)$, where $n \in \mathbb{N}$. To prove an LDP for a mixture, one clearly needs to assume that each conditional probability satisfies an LDP and that the mixing measure also satisfies an LDP (ignoring some trivial cases). In general, however, these assumptions are not sufficient for a mixture to satisfy an LDP. Indeed, in [11, Ex. 4.2] it is shown that a mixture may fail to satisfy an LDP under these assumptions, even when $\nu_n = \nu$ and thus ν_n automatically satisfies an LDP.

These problems may be circumvented by imposing additional assumptions. One way is to take $\Theta = \mathbb{R}$ and $\nu_n = \nu$ (cf. [11, Th. 4.2]). Another way is to assume that the sequence of probability measures ν_n is exponentially tight (cf. [1, Th. 1] and [6, Th. 2.3]). This assumption implies that the rate function corresponding to ν_n is good, meaning that it has compact sublevel sets (cf. [10, Lem. 1.2.18] and [6, Lem. 2.1]).

Although we do have $\Theta = \mathbb{R}$ in our case, we do not assume that $\nu_n = \nu$ nor that the rate function corresponding to ν_n has compact sublevel sets. Consequently, we cannot use the known results about LDPs for mixtures. As mentioned in Section 1, we solve this relying on special properties of the queueing model and a generalization of Varadhan's Lemma as presented in [15]. This approach allows us to work with a much larger class of background processes and also shows that common assumptions about good rate functions in large deviations theory may sometimes be unnecessarily restrictive when dealing with queueing systems.

3 Model and problem description

We study an infinite-server queue with modulated arrival rates, service requirements and server work rates. The precise mathematical setup of the model and some of its basic properties are provided in Section A. Heuristically, the model may be described as follows.

Let $(J(t))_{t\geq 0}$ be a càdlàg stochastic process with state space $\mathcal{E}$, which is assumed to be a metric space. We will refer to the process J as the background process or modulating process. For each $j \in \mathcal{E}$, let $Z(1,j), Z(2,j), \ldots$ be a sequence of independent, nonnegative, identically distributed random variables with cumulative distribution function F_j. We assume that the map $(\omega, j) \mapsto Z(k,j)(\omega)$ is measurable. Let λ and μ be continuous functions defined on $\mathcal{E}$ and taking values in $[0,\infty)$.

While the background process is in state $x \in \mathcal{E}$, jobs enter the system following a Poisson process with intensity $\lambda(x) \geq 0$. When job k enters the system, its service requirement is given by $Z(k,y)$ if the background process is in state $y \in \mathcal{E}$ upon its arrival. Server k processes this service requirement at rate $\mu(z)$ while the background process is in state $z \in \mathcal{E}$. Job k leaves the system when its service requirement has been processed.

We denote a modulated infinite-server queue (under the conditions detailed in Section A) by the quadruple (J, Z, λ, μ). Additionally, we denote the number of jobs in this system at time t by $M(t)$.

Given a modulated infinite-server queue (J, Z, λ, μ), we associate the map ϕ_t with it, where $\phi_t \colon D([0,\infty); \mathcal{E}) \to [0,\infty)$ is given by

$$\phi_t(f) = \int_0^t \left(1 - F_{f(s)}\left(\int_s^t \mu(f(r))\,\mathrm{d}r\right)\right)\lambda(f(s))\,\mathrm{d}s. \tag{1}$$

The map ϕ_t will be called the parameter map of (J, Z, λ, μ). In Section A it is shown that $M(t)$ has a Poisson distribution with random parameter $\phi_t(J)$. This will turn out to be a crucial property in this paper.

We are interested in events with an unusual number of jobs in the system. More precisely, we would like to prove an LDP for the number of jobs in the system. A sequence of probability measures $\{\tau_n\}_{n\in\mathbb{N}}$ is said to satisfy an LDP with rate function ρ if there exists a lower semi-continuous function $\rho \colon \mathcal{X} \to [0,\infty]$ such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \tau_n(F) \leq -\inf_{a\in F} \rho(a)$$

for all closed sets F and

$$\liminf_{n\to\infty} \frac{1}{n} \log \tau_n(G) \geq -\inf_{a\in G} \rho(a)$$

for all open sets G, where each τ_n is defined on the Borel σ-algebra of the topological space $\mathcal{X}$. A sequence of random variables is said to satisfy an LDP with rate function ρ if the sequence of measures induced by the random variables satisfies an LDP with rate function ρ. Importantly, we do not assume that ρ is a good rate function, i.e., we do not assume that ρ has compact level sets.

As mentioned, we would like to prove an LDP for the number of jobs in the system. To analyze this problem, we will scale the arrival rates via $\lambda(x) \mapsto n\lambda(x)$, i.e., we linearly speed up the arrivals. In addition, we will scale the background process via $J \mapsto J_n$. Formally, scaling $\lambda(x) \mapsto n\lambda(x)$ and $J \mapsto J_n$ means that we start with an infinite-server queue (J, Z, λ, μ) and then consider the sequence of infinite-server queues $\{(J_n, Z, n\lambda, \mu)\}_{n\in\mathbb{N}}$.

Given the scalings $\lambda(x) \mapsto n\lambda(x)$ and $J \mapsto J_n$, we denote the corresponding number of jobs in the system by $M_n(t)$. It follows immediately from equation (1) that $M_n(t)$ has a Poisson distribution with random parameter

$$n\phi_t(J_n) = \int_0^t \left(1 - F_{J_n(s)}\left(\int_s^t \mu(J_n(r))\,\mathrm{d}r\right)\right)n\lambda(J_n(s))\,\mathrm{d}s,$$

where ϕ_t is the parameter map associated with (J, Z, λ, μ). The normalized random parameter $\phi_t(J_n)$ induces a sequence of probability measures $\{\nu_n\}_{n\in\mathbb{N}}$ on $\mathbb{R}$ via $\nu_n(B) = \mathbb{P}(\phi_t(J_n) \in B)$ for Borel sets $B \subset \mathbb{R}$.

We will assume that the sequence of probability measures $\{\nu_n\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ. The sequence $\{\nu_n\}_{n\in\mathbb{N}}$ trivially satisfies an LDP when $\nu_n = \nu_{n+1}$ for all $n \in \mathbb{N}$, so this assumption covers the case in which the background process is not scaled.

Given the scaling, we denote the number of jobs in the system at time t by $M_n(t)$ and consider the normalized random variable $\frac{1}{n}M_n(t)$. Our goal is to prove an LDP for $\frac{1}{n}M_n(t)$ and to describe the corresponding rate function.

Throughout this paper, we will also use the following notation. We denote the closure of a set A by cl$A$. We write $B(x,\epsilon)$ for the open ball with center $x \in \mathbb{R}^d$ and radius $\epsilon > 0$ and $B[x,\epsilon]$ for its closure. The Borel σ-algebra of a topological space $\mathcal{E}$ will be denoted by $\mathcal{B}(\mathcal{E})$. For notational convenience, we will sometimes write $\mathbb{R}_+$ for $[0,\infty)$, $B_+(x,\epsilon)$ for $B(x,\epsilon)\cap\mathbb{R}_+$ and $B_+[x,\epsilon]$ for $B[x,\epsilon]\cap\mathbb{R}_+$. As is customary, we define $\exp(-\infty) = 0$ and $\log(0) = -\infty$.

4 A large deviations principle

Let (J, Z, λ, μ) be a modulated infinite-server queue with associated parameter map ϕ_t. In this section we will prove an LDP for the number of jobs in the system under a scaling of the arrival rates and the background process, i.e., we will prove an LDP for $\frac{1}{n}M_n(t)$. It will turn out that so-called attainable parameters determine the rate function corresponding to the LDP.

Definition 4.1. Given a scaling $J \mapsto J_n$, a real number $\gamma \in [0,\infty)$ is called an *attainable parameter* at time $t \geq 0$ if for all $\epsilon > 0$ there exists $N_\epsilon \in \mathbb{N}$ such that $\mathbb{P}(\phi_t(J_n) \in B(\gamma,\epsilon)) = \nu_n(B(\gamma,\epsilon)) > 0$ for all $n \geq N_\epsilon$. The set of all attainable parameters at time t is denoted by $\mathcal{R}(t)$.

The intuition behind attainable parameters is as follows. The number of jobs in the system has a Poisson distribution with a random parameter that is completely determined by the background process. Basically, the background process samples the Poisson parameter. A real number γ is an attainable parameter if, for all n large enough, the scaled background process samples parameters close to γ with positive probability.

As mentioned before, we will prove an LDP for $\frac{1}{n}M_n(t)$ by scaling $\lambda(x) \mapsto n\lambda(x)$ and $J \mapsto J_n$ such that the sequence of probability measures $\{\nu_n\}_{n\in\mathbb{N}}$ induced by the sequence of random parameters $\{\phi_t(J_n)\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ. The rate function $I\colon \mathbb{R} \to [0,\infty]$ governing the LDP for $\frac{1}{n}M_n(t)$ is given by

$$I(a) = \inf_{\gamma\in\mathcal{R}(t)} [\ell(\gamma; a) + \psi(\gamma)], \tag{2}$$

where $\ell(\gamma;\cdot)$ is the Fenchel-Legendre transform of the Poisson cumulant generating function with parameter γ. It will turn out (cf. Lemma 4.2) that

$$I(a) = \inf_{\gamma\in\mathcal{R}(t)} [\ell(\gamma; a) + \psi(\gamma)] = \inf_{\gamma\in\{\psi<\infty\}} [\ell(\gamma; a) + \psi(\gamma)]. \tag{3}$$

However, we will take the infimum over $\mathcal{R}(t)$ rather than over $\{\psi < \infty\}$ to stress that attainability of parameters is the crucial property for proving the LDP.

Before we can give the proof, we have to settle some technical details. First, it is not immediately clear whether the function I is indeed a rate function or even whether I is well defined. In particular, it is not clear whether $\mathcal{R}(t)$ is a non-empty set. However, the assumption that the sequence $\{\nu_n\}_{n\in\mathbb{N}}$ satisfies an LDP implies that $\mathcal{R}(t)$ is non-empty, as the following lemma shows.

Lemma 4.2. *Let the scaling $J \mapsto J_n$ be such that $\{\nu_n\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ. Then $\mathcal{R}(t)$ is a non-empty closed subset of $[0,\infty)$ and $\{\psi < \infty\} \subset \mathcal{R}(t)$.*

Proof. Suppose that $\gamma \in \mathbb{R} \setminus \mathcal{R}(t)$. Then there exists $\epsilon > 0$ such that for all $n \in \mathbb{N}$ there exists $k_n \in \mathbb{N}$ such that $k_n \geq n$ and $\nu_{k_n}(B(\gamma,\epsilon)) = 0$. This implies that $B(\gamma,\epsilon) \subset \mathbb{R} \setminus \mathcal{R}(t)$, so $\mathcal{R}(t)$ is closed. Moreover, we must have

$$\liminf_{n\to\infty} \frac{1}{n} \log \nu_n(B(\gamma,\epsilon)) = -\infty = -\inf_{a\in B(\gamma,\epsilon)} \psi(a),$$

so $\psi(a) = \infty$ for all $a \in B(\gamma,\epsilon)$. Then $\mathbb{R} \setminus \mathcal{R}(t) \subset \{\psi = \infty\}$ and $\{\psi < \infty\} \subset \mathcal{R}(t)$. The fact that ψ is a rate function implies that $\{\psi < \infty\}$ is non-empty. The statement of the lemma follows immediately. □

From the previous lemma it follows that I is a well defined function. The fact that I is a rate function is implied by Proposition C.5 and the functions ℓ and ψ being rate functions.

The next lemma is a generalization of Varadhan's Lemma. Contrary to Varadhan's Lemma, it does not require that a given function f is continuous. Instead, it requires that a weaker condition is fulfilled. We will use this lemma to obtain the large deviations upper bound, by applying it to functions f of the form described in Proposition C.4.

Lemma 4.3. *Let $\mathcal{X}$ be a topological space and let $\{\xi_n\}_{n\in\mathbb{N}}$ be a sequence of measures defined on its Borel σ-algebra. Suppose that $\{\xi_n\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ϱ. Let $f\colon \mathcal{X} \to [-\infty, 0]$ be a Borel measurable function such that $f^{-1}([a,b])$ is a closed set for all $a, b \in (-\infty, 0]$ satisfying*

$$\sup_{x\in\mathcal{X}} [f(x) - \varrho(x)] \leq a \leq b \leq 0.$$

Then it holds that

$$\limsup_{n\to\infty} \frac{1}{n} \log \int_{\mathcal{X}} e^{nf(x)} \xi_n(\mathrm{d}x) \leq \sup_{x\in\mathcal{X}} [f(x) - \varrho(x)].$$

Proof. This follows immediately from [15, Cor. 2.3]. □

With these technical details settled, we can prove the following LDP for the number of jobs in the system.

Theorem 4.4. *Consider a modulated infinite-server queue (J, Z, λ, μ) as described in Section 3. Scale $\lambda(x) \mapsto n\lambda(x)$ and $J \mapsto J_n$ such that $\{\nu_n\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ. Then the rescaled number of jobs in the system $\frac{1}{n}M_n(t)$ satisfies an LDP with rate function I as defined in equation (2), so*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}M_n(t) \in F\right) \leq -\inf_{a\in F} I(a) \tag{4}$$

for any closed set $F \subset \mathbb{R}$ and

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}M_n(t) \in G\right) \geq -\inf_{a\in G} I(a) \tag{5}$$

for any open set $G \subset \mathbb{R}$.

Proof. For $\gamma \geq 0$, let $P_0(\gamma), P_1(\gamma), P_2(\gamma), \ldots$ denote a sequence of i.i.d. random variables that have a Poisson distribution with parameter γ. Let $F \subset \mathbb{R}$ be a closed set and let $G \subset \mathbb{R}$ be an open set.

To prove the upper bound (4), recall that $M_n(t)$ has a Poisson distribution with random parameter $n\phi_t(J_n)$. Then we may write

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}M_n(t) \in F\right) &= \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} \mathbb{P}\left(\frac{1}{n}P_0(n\gamma) \in F\right) \nu_n(\mathrm{d}\gamma) \\
&= \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} P_i(\gamma) \in F\right) \nu_n(\mathrm{d}\gamma) \\
&\leq \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} 2e^{n[-\inf_{a\in F} \ell(\gamma;a)]}\, \nu_n(\mathrm{d}\gamma) \\
&= \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} e^{n[-\inf_{a\in F} \ell(\gamma;a)]}\, \nu_n(\mathrm{d}\gamma).
\end{aligned}$$

The inequality above follows from [11, Lem. 4.1].

According to Proposition C.4, the function $\gamma \mapsto -\inf_{a\in F} \ell(\gamma; a)$ satisfies the assumptions of Lemma 4.3. Moreover, $\{\nu_n\}_{n\in\mathbb{N}}$ satisfies an LDP both in $\mathbb{R}$ and in $[0, \infty)$ with rate function ψ (cf. [10, Lem. 4.1.5]). Hence, we may apply Lemma 4.3 to obtain

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}M_n(t) \in F\right) &\leq \limsup_{n\to\infty} \frac{1}{n} \log \int_{[0,\infty)} e^{n[-\inf_{a\in F} \ell(\gamma;a)]}\, \nu_n(\mathrm{d}\gamma) \\
&\leq \sup_{\gamma\in[0,\infty)} \left[-\inf_{a\in F} \ell(\gamma; a) - \psi(\gamma)\right] \\
&= -\inf_{a\in F} \inf_{\gamma\in[0,\infty)} [\ell(\gamma; a) + \psi(\gamma)] \\
&= -\inf_{a\in F} \inf_{\gamma\in\mathcal{R}(t)} [\ell(\gamma; a) + \psi(\gamma)] \\
&= -\inf_{a\in F} I(a).
\end{aligned}$$

The fact that we only have to consider the infimum over $\mathcal{R}(t)$ follows from Lemma 4.2. This proves the upper bound.

To prove the lower bound (5), let $\gamma \in \mathcal{R}(t)$ and $\epsilon > 0$. Define $\gamma_\epsilon^- = \max\{0, \gamma - \epsilon\}$ and $\gamma_\epsilon^+ = \gamma + \epsilon$. By definition of the set $\mathcal{R}(t)$ there exists N_ϵ such that $\mathbb{P}(\phi_t(J_n) \in B(\gamma, \epsilon)) > 0$ for all $n \geq N_\epsilon$.

Fix $x \in G$. Because G is open, there exists $\delta > 0$ such that $B(x, \delta) \subset G$. Observe that

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n} M_n(t) \in G\right) &\geq \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x, \delta)\right) \\
&\geq \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x, \delta)\,;\, \phi_t(J_n) \in B(\gamma, \epsilon)\right) \\
&= \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x, \delta) \,\middle|\, \phi_t(J_n) \in B(\gamma, \epsilon)\right) \mathbb{P}(\phi_t(J_n) \in B(\gamma, \epsilon))
\end{aligned}$$

for all $n \geq N_\epsilon$, where the equality follows from the fact that $\mathbb{P}(\phi_t(J_n) \in B(\gamma, \epsilon)) > 0$ for all $n \geq N_\epsilon$. Then we get

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in G\right) \geq \\
\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x, \delta) \,\middle|\, \phi_t(J_n) \in B(\gamma, \epsilon)\right) \\
+ \liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\phi_t(J_n) \in B(\gamma, \epsilon)).
\end{aligned}$$

Recall that $\phi_t(J_n)$ satisfies an LDP with rate function ψ, so

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\phi_t(J_n) \in B(\gamma, \epsilon)) \geq - \inf_{a \in B(\gamma, \epsilon)} \psi(a)$$

by assumption. Moreover, it holds that

$$\begin{aligned}
\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x, \delta) \,\middle|\, \phi_t(J_n) \in B(\gamma, \epsilon)\right) = \\
\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} M_n(t) \in B(x, \delta) \,\middle|\, \phi_t(J_n) \in B(\gamma, \epsilon) \cap \mathbb{R}_+\right) \geq \\
\liminf_{n\to\infty} \inf_{\xi \in B(\gamma, \epsilon) \cap \mathbb{R}_+} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n P_i(\xi) \in B(x, \delta)\right) = \\
\min_{\xi \in \{\gamma_\epsilon^-, \gamma_\epsilon^+\}} \left[- \inf_{a \in B(x, \delta)} \ell(\xi; a) \right].
\end{aligned}$$

In the display above, the inequality follows from Lemma A.2 and the second equality is established in Proposition C.3. Combining the results, we obtain that

$$\mathbb{P}\left(\frac{1}{n} M_n(t) \in G\right) \geq \min_{\xi \in \{\gamma_\epsilon^-, \gamma_\epsilon^+\}} \left[- \inf_{a \in B(x, \delta)} \ell(\xi; a) \right] - \inf_{a \in B(\gamma, \epsilon)} \psi(a).$$

This holds for all $\epsilon > 0$ and small enough $\delta > 0$. Taking limits, we get

$$\lim_{\epsilon \downarrow 0} \min_{\xi \in \{\gamma_\epsilon^-, \gamma_\epsilon^+\}} \left[- \inf_{a \in B(x,\delta)} \ell(\xi; a) \right] = - \inf_{a \in B(x,\delta)} \ell(\gamma; a)$$

thanks to Proposition C.4 and

$$\lim_{\epsilon \downarrow 0} \inf_{a \in B(\gamma,\epsilon)} \psi(a) = \psi(\gamma),$$

because ψ is lower semi-continuous. Similarly, we get $\lim_{\delta \downarrow 0} \inf_{a \in B(x,\delta)} \ell(\gamma; a) = \ell(\gamma; x)$. Hence, it follows that

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} M_n(t) \in G \right) &\geq \lim_{\delta \downarrow 0} \lim_{\epsilon \downarrow 0} \left[\min_{\xi \in \{\gamma_\epsilon^-, \gamma_\epsilon^+\}} \left[- \inf_{a \in B(x,\delta)} \ell(\xi; a) \right] - \inf_{a \in B(\gamma,\epsilon)} \psi(a) \right] \\ &= -[\ell(\gamma; x) + \psi(\gamma)]. \end{aligned}$$

Since $x \in G$ and $\gamma \in \mathcal{R}(t)$ were arbitrary, we obtain

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n} M_n(t) \in G \right) &\geq \sup_{x \in G} \sup_{\gamma \in \mathcal{R}(t)} [-[\ell(\gamma; x) + \psi(\gamma)]] \\ &= - \inf_{a \in G} I(a), \end{aligned}$$

which completes the proof. □

The proof of Theorem 4.4 contains familiar elements. First, the upper bound is proved using a Chernoff bound combined with a generalization of Varadhan's Lemma. Second, the lower bound is proved by considering 'the most likely of all unlikely scenarios', which is similar to the method used in [3] and [5]. However, the proofs there relied on properties of irreducible continuous-time Markov chains and the computation of optimal paths, whereas we consider general càdlàg background processes via attainable parameters.

5 Examples: unscaled background processes

Given a modulated infinite-server queue (J, Z, λ, μ) and a scaling $\lambda \mapsto n\lambda$ and $J \mapsto J_n$, Theorem 4.4 provides a full LDP for $\frac{1}{n} M_n(t)$ and describes the corresponding rate function. In the upcoming examples we will consider cases in which the background process is not scaled and we will use Theorem 4.4 to verify or extend known results and to obtain new results.

Throughout this section we will assume that the background process is not scaled, i.e., $J_n = J$ for all $n \in \mathbb{N}$ for some càdlàg stochastic process J. This is similar to the situation shown in the bottom figures in [19, Fig. 28.4], where the arrival process is very fast relative to the background process. We will show how to obtain an LDP in this case.

The following lemma is trivial, but plays a central role in this section.

Lemma 5.1. *If $J_n = J$ for all $n \in \mathbb{N}$, then the sequence $\{\phi_t(J_n)\}_{n\in\mathbb{N}}$ satisfies an LDP with some rate function ψ. In this case $\mathcal{R}(t)$ coincides with the support of $\phi_t(J)$ and $\mathcal{R}(t) = \{\psi < \infty\} = \{\psi = 0\}$.*

Proof. Suppose that $J_n = J$ for all $n \in \mathbb{N}$. Let ν denote the law of $\phi_t(J)$. Clearly, for each $x \in \mathbb{R}$ either $\nu(B(x,\epsilon)) > 0$ for all $\epsilon > 0$ or there exists $\delta > 0$ such that $\nu(B(x,\delta)) = 0$. The set of all $x \in \mathbb{R}$ with $\nu(B(x,\epsilon)) > 0$ for all $\epsilon > 0$ is the support of $\phi_t(J)$. Hence, if $J_n = J$, then $\mathcal{R}(t)$ equals the support of $\phi_t(J)$.

The rate function $\psi\colon \mathbb{R} \to [0,\infty]$ is defined by taking $\psi(a) = 0$ for $a \in \mathcal{R}(t)$ and $\psi(a) = \infty$ if $a \notin \mathcal{R}(t)$.

Suppose that $G \subset \mathbb{R}$ is open. Then $\nu(G) > 0$ if and only if $G \cap \mathcal{R}(t) \neq \emptyset$. Hence, $\liminf_{n\to\infty} \frac{1}{n} \log \nu(G) = -\inf_{a\in G} \psi(a)$.

Suppose that $F \subset \mathbb{R}$ is closed. If $F \cap \mathcal{R}(t) \neq \emptyset$, then trivially $\limsup_{n\to\infty} \frac{1}{n} \log \nu(F) \leq 0 = -\inf_{a\in F} \psi(a)$. If $F \cap \mathcal{R}(t) = \emptyset$, then there exists an open set $F^* \supset F$ such that $F^* \cap \mathcal{R}(t) = \emptyset$, because $\mathcal{R}(t)$ is closed (cf. Lemma 4.2). Then $\nu(F^*) = 0$ (see the argument for open sets G) and we have $\limsup_{n\to\infty} \frac{1}{n} \log \nu(F) \leq \limsup_{n\to\infty} \frac{1}{n} \log \nu(F^*) = -\infty = -\inf_{a\in F} \psi(a)$. □

Hence, when the background process is not scaled, we have the special property that $\mathcal{R}(t) = \{\psi = 0\}$. This will enable us to compute explicit rate functions in the examples. In these computations, we will extensively use the following properties of the rate function I and properties of step functions in Skorokhod space.

Recall that the rate function I is given by

$$I(a) = \inf_{\gamma\in\mathcal{R}(t)} [\ell(\gamma; a) + \psi(\gamma)],$$

and that $\mathcal{R}(t) = \{\psi = 0\}$ (cf. Lemma 5.1). Hence, we get

$$I(a) = \inf_{\gamma\in\mathcal{R}(t)} \ell(\gamma; a). \tag{6}$$

In this case, we can give a simpler and more explicit description of I, using the following properties of the function ℓ.

For $\gamma \geq 0$, the function $\ell(\gamma;\cdot)$ is the Fenchel-Legendre transform of the Poisson cumulant generating function with parameter γ and is given by

$$\ell(\gamma; a) = \begin{cases} \infty & a < 0; \\ \gamma & a = 0; \\ \gamma - a + a\log(a/\gamma) & a > 0. \end{cases} \tag{7}$$

For $\gamma = 0$ and $a > 0$, we understand that $\gamma - a + a\log(a/\gamma) = \infty$. An important observation is that the following inequalities hold for $0 \leq \gamma_1 \leq \gamma_2 < \infty$:

$$\ell(\gamma_1; a) \leq \ell(\gamma_2; a) \qquad \forall a \in [0, \gamma_1]; \tag{8}$$

$$\ell(\gamma_1; a) \geq \ell(\gamma_2; a) \qquad \forall a \in [\gamma_2, \infty). \tag{9}$$

See Figure 1 for an illustration.

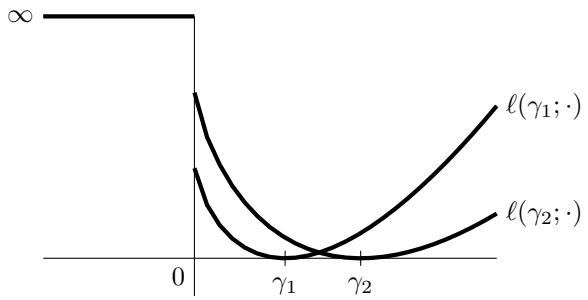

Figure 1: Graphs of the functions $\ell(\gamma_1;\cdot)$ and $\ell(\gamma_2;\cdot)$ for $0 < \gamma_1 < \gamma_2 < \infty$

Because in the present case I is just an infimum of Poisson rate functions, these inequalities imply that I has some special properties. They are described in the following proposition.

Proposition 5.2. *In the present case, $I(a) = 0$ if and only if $a \in \mathcal{R}(t)$. If $I(a) > 0$ for some $a \in \mathbb{R}$, then exactly one of the following three scenarios is true:*

1. *$a < c_- = \inf \mathcal{R}(t)$ and $I(b) = \ell(c_-; b)$ for all $b \in (-\infty, c_-]$;*
2. *$a > c_+ = \sup \mathcal{R}(t)$ and $I(b) = \ell(c_+; b)$ for all $b \in [c_+, \infty)$;*
3. *the previous two cases do not hold and $I(b) = \min\{\ell(c_-; b), \ell(c_+; b)\}$ for all $b \in [c_-, c_+]$, where $c_- = \sup(\mathcal{R}(t) \cap (-\infty, a))$ and $c_+ = \inf(\mathcal{R}(t) \cap (a, \infty))$.*

Proof. It follows immediately from equations (6) and (7) that $I(a) = 0$ if and only if $a \in \mathcal{R}(t)$. Hence, $I(a) > 0$ implies that the distance of a to $\mathcal{R}(t)$ is strictly positive, since $\mathcal{R}(t)$ is closed. The three scenarios now follow from the inequalities (8) and (9). □

The previous proposition may seem rather abstract. To get some intuition, the following example describes a typical rate function.

Example 5.3. Suppose that $\mathcal{R}(t) = [\alpha, \beta] \cup [\gamma, \delta]$ for some $0 < \alpha < \beta < \gamma < \delta < \infty$. Then the function I looks like the graph shown in Figure 2: it equals 0 on the intervals $[\alpha, \beta]$ and $[\gamma, \delta]$, whereas it equals the minimum of $\ell(\beta;\cdot)$ and $\ell(\gamma;\cdot)$ on the interval (β, γ) in between. On the interval $(-\infty, \alpha]$ the function I equals $\ell(\alpha;\cdot)$ and on the interval $[\delta, \infty)$ the function I equals $\ell(\delta;\cdot)$.

In the remainder of this section, we focus on the modulated M/M/∞ queue (J, Z, λ, μ) as described in Example A.3 under a linear scaling of the arrival rates. The associated parameter map ϕ_t is given by equation (16) (cf. Example A.3), which is continuous (cf. Lemma B.4).

To compute $\mathcal{R}(t)$ in this case, it is often convenient to use the following properties of step functions in $D([0, \infty); \mathcal{E})$. (For the definition of a step function, see Section B.) Recall that the set of all step functions in $D([0, \infty); \mathcal{E})$ is denoted by $\mathcal{S}([0, \infty); \mathcal{E})$.

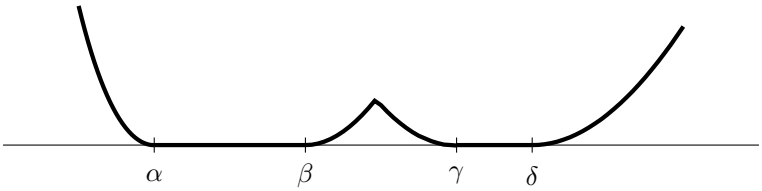

Figure 2: Visualization of the function I in Example 5.3

Lemma 5.4. *If* $\{\phi_t(f)|f \in \mathcal{S}([0,\infty);\mathcal{E})\} \subset \mathcal{R}(t)$*, then*

$$\mathcal{R}(t) = \mathrm{cl}\{\phi_t(f)|f \in \mathcal{S}([0,\infty);\mathcal{E})\} = \{\phi_t(f)|f \in \mathcal{D}([0,\infty);\mathcal{E})\}.$$

Proof. This follows from Lemma 4.2 and Corollary B.3 and the fact that ϕ_t is continuous under the present assumptions. □

Lemma 5.5. *If* $\{\phi_t(f)|f \in \mathcal{S}([0,\infty);\mathcal{E})\} \subset \mathcal{R}(t)$*, then* $\mathcal{R}(t)$ *is a closed interval.*

Proof. It suffices to show that $\mathcal{R}(t)$ is convex. Let $f_{\mathrm{c}}^1, f_{\mathrm{c}}^2 \in \mathcal{S}([0,\infty);\mathcal{E})$. We may assume that $\phi_t\big(f_{\mathrm{c}}^1\big) \le \phi_t\big(f_{\mathrm{c}}^2\big)$. For $x \in [0,t]$ we define the function g_x via

$$g_x(s) = \mathbb{1}_{\{s<x\}} f_{\mathrm{c}}^1(s) + \mathbb{1}_{\{s\ge x\}} f_{\mathrm{c}}^2(s)$$

for $s \in [0,\infty)$. Clearly, $g_x \in \mathcal{S}([0,\infty);\mathcal{E})$ for each $x \in [0,t]$.

Since f_{c}^1 and f_{c}^2 are step functions, there exists a finite set $\mathcal{E}^* \subset \mathcal{E}$ such that $g_x \in \mathcal{S}([0,\infty);\mathcal{E}^*)$ for each $x \in [0,t]$. Suppose that $x_1, x_2 \in [0,t]$ with $x_1 < x_2$ and $x_2 - x_1 = \epsilon$. Then $g_{x_1}(s) = g_{x_2}(s)$ for all $s \in [0,t] \setminus [x_1, x_2)$. Since the interval $[x_1, x_2)$ has length ϵ, Lemma B.6 implies that

$$|\phi_t(g_{x_1}) - \phi_t(g_{x_2})| \le \lambda_+(1+t)\big(1 - e^{-\epsilon\kappa_+ + \mu_+} + \epsilon\big),$$

where $\lambda_+ = \max_{x\in\mathcal{E}^*} \lambda(x)$, $\mu_+ = \max_{x\in\mathcal{E}^*} \mu(x)$ and $\kappa_+ = \max_{x\in\mathcal{E}^*} \kappa(x)$. This shows that the function $x \mapsto \phi_t(g_x)$ is a continuous function from $[0,t]$ to $\mathbb{R}$.

Observe that $\phi_t(g_0) = \phi_t\big(f_{\mathrm{c}}^2\big)$ and $\phi_t(g_t) = \phi_t\big(f_{\mathrm{c}}^1\big)$. Now applying the Intermediate Value Theorem to the continuous function $x \mapsto \phi_t(g_x)$, it follows that

$$\big[\phi_t\big(f_{\mathrm{c}}^1\big), \phi_t\big(f_{\mathrm{c}}^2\big)\big] = [\phi_t(g_t), \phi_t(g_0)] \subset \{\phi_t(g_x)\,|\,x \in [0,t]\} \subset \mathcal{R}(t).$$

Combined with Lemma 5.4, this implies the statement of the lemma. □

Let $f_{\mathrm{c}} \in \mathcal{S}([0,\infty);\mathcal{E})$ be a step function. Clearly, f_{c} has a unique minimal representation $\{(t_i,\alpha_i)\}_{i=0}^k$, where $k \in \mathbb{N}$, $0 = t_0 < t_1 < \ldots < t_k < \infty$ and $\alpha_0, \ldots, \alpha_k \in \mathcal{E}$ are such that $f_{\mathrm{c}}(t) = \alpha_i$ for $t \in [t_i, t_{i+1})$ and $i = 0, \ldots, k-1$ and $f_{\mathrm{c}}(t) = \alpha_k$ for $t \in [t_k, \infty)$. Given this minimal representation, we define its truncated minimal step size by

$$\Delta_{f_{\mathrm{c}}} = 1 \wedge \min_{i=1,\ldots,k} \{t_i - t_{i-1}\}.$$

Additionally, we define $t_{k+1} = t_k \vee t$. The truncated minimal step size and t_{k+1} will be used for computing attainable parameters.

In the upcoming examples, we would like to compute rate functions via attainable parameters. To compute attainable parameters, we use the following strategy. We fix a certain path f, often a step function. This gives us a parameter value $\phi_t(f)$. Then we would like to show that, with positive probability, the background process stays 'close' to f, which will imply that $\phi_t(f)$ is an attainable parameter.

Staying 'close' to f depends on properties of $\mathcal{E}$ and the background process. In most cases, the background process needs a little bit of room (both in time and in space) to jump near a discontinuity of f. This is where the truncated minimal step size comes in: it is an upper bound on the time we give the background process for jumping near a discontinuity of a step function. The precise meaning of this will become clearer in the examples.

The first example treats the familiar case of a Markov-modulated M/M/∞ queue, i.e., the case in which the background process is an irreducible Markov chain. This case is partly studied in [3] (Model I) and [5] (Model II). In the example, we recover [3, Th. 2] and [5, Th. 1]. Additionally, we generalize these results to our model and extend them to a full LDP.

Example 5.6. Let J be an irreducible, continuous-time Markov process with finite state space $\mathcal{E} = \{1, \dots, d\}$. We consider the modulated infinite-server queue (J, Z, λ, μ) under the scaling $\lambda \mapsto n\lambda$. Theorem 4.4 (combined with Lemma 5.1) shows that $\frac{1}{n}M_n(t)$ satisfies an LDP with rate function I. This rate function may be computed as follows.

Fix any function $g \in \mathcal{S}([0,\infty); \mathcal{E})$ with minimal representation $\{(t_i, \alpha_i)\}_{i=0}^k$ and take any $\epsilon \in (0,1)$. Define $\mathcal{W}(g;\epsilon)$ as the set of all $f \in D([0,\infty);\mathcal{E})$ such that

$$\begin{aligned} f(t) &= \alpha_{i-1} \qquad && \forall t \in \left[t_{i-1} + \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_i - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\right) \quad \forall i \in \{1,\dots,k\}, \\ f(t) &= \alpha_k && \forall t \in [t_k, t_{k+1}]. \end{aligned}$$

Intuitively speaking, the set $\mathcal{W}(g;\epsilon)$ consists of all paths $f \in D([0,\infty);\mathcal{E})$ that coincide with g on the intervals described above. These intervals cover $[0,t]$, except around 0 and around time points at which g jumps.

Observe that the set $\mathcal{W}(g;\epsilon)$ is constructed such that each $f \in \mathcal{W}(g;\epsilon)$ coincides with g on $[0,t]$, except possibly on a subset with Lebesgue measure at most ϵ. Since $\mathcal{E}$ is finite and the parameter map ϕ_t is given by equation (16) under the present assumptions, Lemma B.6 then implies that

$$\sup_{f \in \mathcal{W}(g;\epsilon)} |\phi_t(f) - \phi_t(g)| \to 0$$

as $\epsilon \to 0$.

Also observe that $\mathbb{P}(J \in \mathcal{W}(g;\epsilon)) > 0$, thanks to the irreducibility of J. Consequently, $\{\phi_t(g) \,|\, g \in \mathcal{S}([0,\infty);\mathcal{E})\} \subset \mathcal{R}(t)$. Then Lemma 5.5 implies that $\mathcal{R}(t) = \{\phi_t(g) \,|\, g \in D([0,\infty);\mathcal{E})\}$ and that $\mathcal{R}(t)$ is a closed interval. Because $\mathcal{E}$

is finite, we immediately get

$$\mathcal{R}(t) = [a_-, a_+],$$

where $0 \leq a_- \leq a_+ < \infty$ with $a_- = \inf_{g \in D([0,\infty);\mathcal{E})} \phi_t(g)$ and $a_+ = \sup_{g \in D([0,\infty);\mathcal{E})} \phi_t(g)$. Now applying Proposition 5.2, it follows that the rate function I is given by

$$I(a) = \begin{cases} \infty & a \in (-\infty, 0); \\ \ell(a_-; a) & a \in [0, a_-]; \\ 0 & a \in [a_-, a_+]; \\ \ell(a_+; a) & a \in [a_+, \infty). \end{cases} \tag{10}$$

The result of the previous example depends neither on the initial distribution nor on the transition rate matrix of the irreducible Markov chain. Moreover, the analysis in the previous example implies the following lemma. It shows that we always obtain a good rate function when the background process has a finite state space.

Lemma 5.7. *Let $J^{(1)}$ be a background process with finite state space $\mathcal{E}$ and let $J^{(2)}$ be an irreducible Markov chain with the same state space. Consider the two modulated M/M/∞ queues $\left(J^{(1)}, Z, \lambda, \mu\right)$ and $\left(J^{(2)}, Z, \lambda, \mu\right)$. Scaling $\lambda \mapsto n\lambda$, we obtain in both cases an LDP for the number of jobs in the system with corresponding rate functions $I^{(1)}$ and $I^{(2)}$. Then it holds that $I^{(1)}(a) \geq I^{(2)}(a)$ for all $a \in \mathbb{R}$. In particular, both $I^{(1)}$ and $I^{(2)}$ are good rate functions.*

In the next example we will modulate an M/M/∞ queue by another Markov-modulated infinite-server queue. This setup differs from the setup considered in [3] and [5]. In particular, the state space of the background process is countably infinite, so that we may obtain a rate function that is not good.

Example 5.8. Consider a Markov-modulated infinite-server queue as described in [17], i.e., a Markov-modulated infinite-server queue under the assumptions of Model I. Assume that neither the arrival rates nor the server work rates are identically equal to 0 and that the system starts empty. Let $J(t)$ be the number of jobs in this Markov-modulated infinite-server queue at time $t \geq 0$. Then J is a càdlàg stochastic process and its state space is $\mathcal{E} = \mathbb{Z}_{\geqslant 0}$.

Consider the modulated M/M/∞ queue (J, Z, λ, μ) and impose the scaling $\lambda \mapsto n\lambda$. Then $\frac{1}{n} M_n(t)$ satisfies an LDP with rate function I, according to Theorem 4.4 and Lemma 5.1. This rate function may be computed as follows.

Recall that J stays in state $m \in \mathcal{E}$ during $[t, t + \Delta t]$ with positive probability for arbitrarily large Δt. Moreover, because neither the arrival rates nor the server work rates are identically equal to 0, the process J also has the following property. If $J(t) = m_1$ at time $t \geq 0$, then it jumps to state $m_2 \in \mathcal{E}$ during $[t, t + \Delta t]$ with positive probability for arbitrarily small Δt.

Roughly speaking, these two properties mean that the background process is irreducible, in the sense that it can jump to or stay in any state during any time interval we would like. Of course, this is very similar to the Markov

chain being irreducible in the previous example. Consequently, our strategy for determining the attainable parameters will be very similar, although there are some subtleties related to the state space being infinite.

Fix any $g \in \mathcal{S}([0,\infty);\mathcal{E})$ with minimal representation $\{(t_i,\alpha_i)\}_{i=0}^k$ and take any $\epsilon \in (0,1)$. Let $\mathcal{W}(g;\epsilon)$ denote the set of all $f \in D([0,\infty);\mathcal{E})$ with

$$\begin{aligned} f(t) &= \alpha_{i-1} \qquad && \forall t \in \left[t_{i-1} + \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_i - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\right) \quad \forall i \in \{1,\dots,k\}, \\ f(t) &= \alpha_k && \forall t \in [t_k, t_{k+1}], \end{aligned}$$

and

$$\begin{aligned} 0 \le f(t) &\le \alpha_0 && \forall t \in \left[0, \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\right), \\ \alpha_{i-1} \wedge \alpha_i \le f(t) &\le \alpha_{i-1} \vee \alpha_i && \forall t \in \left[t_i - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_i + \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\right) \\ & && \forall i \in \{1,\dots,k-1\}, \\ \alpha_{k-1} \wedge \alpha_k \le f(t) &\le \alpha_{k-1} \vee \alpha_k && \forall t \in \left[t_k - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_k\right]. \end{aligned}$$

Observe that each $f \in \mathcal{W}(g;\epsilon)$ coincides with g, except possibly on a subset with Lebesgue measure at most ϵ. Moreover, each $f \in \mathcal{W}(g;\epsilon)$ takes values in the finite set $\mathcal{E}^* = \{0,\dots,\alpha_+\}$, where $\alpha_+ = \max\{\alpha_i \,|\, i \in \{0,\dots,k\}\}$. Since the parameter map ϕ_t is given by equation (16) under the present assumptions, Lemma B.6 implies that

$$\sup_{f \in \mathcal{W}(g;\epsilon)} |\phi_t(f) - \phi_t(g)| \to 0$$

as $\epsilon \to 0$.

The two properties of the background process described above imply that $\mathbb{P}(J \in \mathcal{W}(g;\epsilon)) > 0$. It follows that $\{\phi_t(g) \,|\, g \in \mathcal{S}([0,\infty);\mathcal{E})\} \subset \mathcal{R}(t)$. Write $a_- = \inf_{g \in D([0,\infty);\mathcal{E})} \phi_t(g)$ and $a_+ = \sup_{g \in D([0,\infty);\mathcal{E})} \phi_t(g)$. Lemma 5.4 and Lemma 5.5 imply that $\mathcal{R}(t) = [a_-, a_+]$ if $a_+ < \infty$ and $\mathcal{R}(t) = [a_-, \infty)$ if $a_+ = \infty$. Hence,

$$I(a) = \begin{cases} \infty & a \in (-\infty, 0); \\ \ell(a_-; a) & a \in [0, a_-]; \\ 0 & a \in [a_-, a_+]; \\ \ell(a_+; a) & a \in [a_+, \infty) \end{cases} \tag{11}$$

if $a_+ < \infty$ and

$$I(a) = \begin{cases} \infty & a \in (-\infty, 0); \\ \ell(a_-; a) & a \in [0, a_-]; \\ 0 & a \in [a_-, \infty) \end{cases} \tag{12}$$

if $a_+ = \infty$. Note that I is not a good rate function if $a_+ = \infty$.

The previous example only depends on the state space being countable and discrete and on the background process being irreducible in the sense described

above. Consequently, the same result holds if the background process is an irreducible Markov process with a countable, discrete state space.

In the last example of this section we compare rate functions that are obtained using two different background processes. One background process is a Markov chain, whereas the other background process is a reflected Brownian motion, which has an uncountable state space. It turns out that both background processes lead to the same LDP, even though the background processes are completely different. Apparently, two very different modulating processes may lead to the same rate function for the LDP, even if the arrival rates, service requirements and server work rates are nontrivial.

Example 5.9. Let $\mathcal{E} = [0,1]$ be equipped with the Euclidean metric. Recall that, under the present assumptions, the $Z(k,j)$ depend on the function κ. Assume that $\lambda\colon [0,1] \to [0,1]$ is given by $\lambda(x) = x$, $\kappa\colon [0,1] \to [0,1]$ is given by $\kappa(x) = 1$ and $\mu\colon [0,1] \to [0,1]$ is given by $\mu(x) = 1 - x$.

Let J^{MC} be an irreducible, continuous-time Markov chain with state space $\{0,1\}$. Let J^{rBM} be a reflected Brownian motion with reflecting barriers 0 and 1. For simplicity, assume that J^{rBM} starts in $x_0 \in (0,1)$, so

$$J^{\mathrm{rBM}}(t) = x_0 + W(t) + L(t) - U(t)$$

for some standard Brownian motion W, lower-regulator process L and upper-regulator process U (cf. [9]).

Consider the two modulated M/M/∞ queues $\big(J^{\mathrm{MC}}, Z, \lambda, \mu\big)$ and $\big(J^{\mathrm{rBM}}, Z, \lambda, \mu\big)$. Under the scaling $\lambda \mapsto n\lambda$, both $\frac{1}{n}M_n^{\mathrm{rBM}}(t)$ and $\frac{1}{n}M_n^{\mathrm{MC}}(t)$ satisfy an LDP with the same good rate function I, which is given by

$$I(a) = \begin{cases} \infty & a \in (-\infty, 0); \\ 0 & a \in [0,t]; \\ \ell(t;a) & a \in [t,\infty). \end{cases} \tag{13}$$

The rate function for the LDP corresponding to $\frac{1}{n}M_n^{\mathrm{MC}}(t)$ is derived in Example 5.6. It is easy to see that the rate function has the form claimed above.

We will show that $\frac{1}{n}M_n^{\mathrm{rBM}}(t)$ satisfies an LDP with the same rate function. Fix $g \in \mathcal{S}([0,\infty);\mathcal{E})$ with minimal representation $\{(t_i,\alpha_i)\}_{i=0}^k$ and take any $\epsilon > 0$. Define $\mathcal{W}(g;\epsilon)$ as the set of all $f \in D([0,\infty);\mathcal{E})$ such that

$$\begin{aligned} |f(t) - \alpha_{i-1}| &\le \epsilon \qquad \forall t \in \big[t_{i-1} + \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g, t_i - \tfrac{\epsilon}{2}\tfrac{1}{k}\Delta_g\big) \quad \forall i \in \{1,\dots,k\}, \\ |f(t) - \alpha_k| &\le \epsilon \qquad \forall t \in [t_k, t_{k+1}]. \end{aligned}$$

Then we get

$$\sup_{f \in \mathcal{W}(g;\epsilon)} \phi_t(f) \le (\phi_t(g) + \epsilon t + \epsilon)e^{\epsilon t + \epsilon}$$

and

$$\inf_{f \in \mathcal{W}(g;\epsilon)} \phi_t(f) \ge (\phi_t(g) - \epsilon t - \epsilon)e^{-\epsilon t - \epsilon}.$$

Now observe that

$$\mathbb{P}\big(J^{\mathrm{rBM}} \in \mathcal{W}(g;\epsilon)\big) \geq \mathbb{P}(x_0 + W \in \mathcal{W}(g;\epsilon)) > 0,$$

due to the definition of J^{rBM} and W being a Brownian motion.

It follows that $\{\phi_t(g) \,|\, g \in \mathcal{S}([0,\infty);\mathcal{E})\} \subset \mathcal{R}^{\mathrm{rBM}}(t)$, so $\mathcal{R}^{\mathrm{rBM}}(t) = [0,t]$ and the corresponding rate function is given by the function I above.

In this section we considered examples in which the background process was not scaled. As shown, this implies some special properties, which we can use to explicitly compute rate functions. In the next section, we will scale the background process, too. Although explicit computations are not possible in general, there are still cases for which we may derive rate functions.

6 Examples: scaled background processes

In this section we will give two examples in which the background process is scaled. In the first example, we will consider the Markov-modulated M/M/∞ queue and derive an explicit rate function under a superlinear time-scaling. This scaling corresponds to the top figures in [19, Fig. 28.4], where the arrival process is very slow relative to the background process.

In the second example, we will consider a new model: we take the service requirements from Example A.4 and let the background process be a Brownian motion. Besides being useful for modelling purposes, Brownian motion also induces mixing measures that are not exponentially tight. We will show this and derive an LDP using Theorem 4.4. In this case, the rate function will be given as the solution of a variational problem.

Example 6.1. Consider the modulated M/M/∞ queue (J, Z, λ, μ) with parameter map ϕ_t, as described in Example A.3. Assume that J is an irreducible continuous-time Markov chain with finite state space $\{1,\dots,d\}$ and generator matrix Q.

Denote the stationary distribution corresponding to Q by $\pi = (\pi_1,\dots,\pi_d)$ and define $\mu_\infty = \sum_{j=1}^d \pi_j\mu_j$ and

$$\varrho_t = \sum_{j=1}^d \pi_j\lambda_j \int_0^t e^{-\kappa_j\mu_\infty(t-s)}\,\mathrm{d}s = \sum_{j=1}^d \pi_j \frac{\lambda_j}{\kappa_j\mu_\infty}\big(1 - e^{-\kappa_j\mu_\infty t}\big).$$

Scale $\lambda \mapsto n\lambda$ and $J \mapsto J_n$, where $J_n(t) = J\big(n^{1+\epsilon}t\big)$. It is easy to see that scaling $J \mapsto J_n$ is equivalent to scaling $Q \mapsto n^{1+\epsilon}Q$.

The sequence of random parameters $\{\phi_t(J_n)\}_{n\in\mathbb{N}}$ satisfies an LDP with rate function ψ, where

$$\psi(a) = \begin{cases} 0 & a = \varrho_t; \\ \infty & a \neq \varrho_t. \end{cases}$$

Indeed, this follows from the fact that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\phi_t(J_n) \in B(\rho_t, \eta)) = 0$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\phi_t(J_n) \notin B(\rho_t, \eta)) = -\infty$$

for all $\eta > 0$. These equalities are an immediate result from the proof of [3, Th. 3].

Given this LDP for $\{\phi_t(J_n)\}_{n\in\mathbb{N}}$, Theorem 4.4 implies that $\frac{1}{n}M_n(t)$ satisfies an LDP with rate function I, where

$$I(a) = \ell(\varrho_t; a).$$

Hence, under this superlinear time-scaling of the background Markov chain, the LDP for $\frac{1}{n}M_n(t)$ is governed by a Poisson rate function with parameter ϱ_t.

Example 6.2. Consider the (nonexponential) modulated infinite-server queue (J, Z, λ, μ) with parameter map ϕ_t, as described in Example A.4. Assume that the background process J is a standard Brownian motion W on $[0, \infty)$. By $\overline{W}$ we denote its restriction to the interval $[0, t]$. The sample paths of $\overline{W}$ are elements of $C_0[0, t]$, the space of continuous functions $f\colon [0, t] \to \mathbb{R}$ with $f(0) = 0$.

Equip $C_0[0, t]$ with the supremum metric. Of course, we may view the function ϕ_t as a map from $C_0[0, t]$ to $[0, \infty)$ and this map is continuous under the supremum metric.

Scale $\lambda \mapsto n\lambda$ and $J \mapsto J_n^\alpha$ for some fixed $\alpha \in [0, \infty)$, where J_n^α is given by a time-scaling: $J_n^\alpha(s) = W(n^{-\alpha}s)$ for $s \geq 0$. Under this scaling, the arrivals are sped up linearly, whereas the time scale of the Brownian motion is slowed down sublinearly, linearly or superlinearly.

Since W is a Brownian motion, we have $\phi_t\big(J_n^1\big) \stackrel{\mathrm{d}}{=} \phi_t\Big(\frac{1}{\sqrt{n}}W\Big) = \phi_t\Big(\frac{1}{\sqrt{n}}\overline{W}\Big)$. Schilder's Theorem (cf. [10, Th. 5.2.3]) states that $\frac{1}{\sqrt{n}}\overline{W}$ satisfies an LDP in $C_0[0, t]$ with good rate function

$$\xi(f) = \begin{cases} \frac{1}{2}\int_0^t |\dot{f}(s)|^2 \,\mathrm{d}s & f \in H_1([0, t]); \\ \infty & \text{else.} \end{cases}$$

Here, $H_1([0, t])$ denotes the set of all absolutely continuous functions $f \in C_0[0, t]$ that have square integrable derivative $\dot{f}$.

Recall that the parameter map ϕ_t is given by equation (17) and that ϕ_t is continuous under the supremum metric on $C_0[0, t]$. The contraction principle (cf. [10, Th. 4.2.1]) now implies that $\phi_t\big(J_n^1\big)$ satisfies an LDP with good rate function ψ, where ψ is given by

$$\psi(a) = \inf\{\xi(f) \,|\, f \in H_1([0, t]), \phi_t(f) = a\}.$$

Clearly, $\psi(a) = 0$ if and only if $a = \phi_t(f_0)$, where $f_0(s) = 0$ for all $s \in [0,t]$. Now writing

$$\frac{1}{n}\log \mathbb{P}(\phi_t(J_n^\alpha) \in B) = n^{\alpha-1}\frac{1}{n^\alpha}\log \mathbb{P}\Big(\phi_t\Big(n^{-\alpha/2}\overline{W}\Big) \in B\Big)$$

for Borel sets B, it is straightforward to verify that for each $\alpha \in [0,\infty)$ the random variable $\phi_t(J_n^\alpha)$ satisfies an LDP with rate function ψ^α, which takes the following form. For $\alpha > 1$, we have $\psi^\alpha(a) = 0$ if $a = \phi_t(f_0)$ and $\psi^\alpha(a) = \infty$ if $a \neq \phi_t(f_0)$. For $\alpha = 1$, we have $\psi^\alpha = \psi$. For $\alpha \in [0,1)$, we have $\psi^\alpha(a) = 0$ if $a \in \{\psi < \infty\}$ and $\psi^\alpha(a) = \infty$ if $a \in \{\psi = \infty\}$.

Observe that for $\alpha \in [0,1)$ the set $\{\psi < \infty\}$ is not necessarily compact, for instance when $\lambda(x) = 1 + x^2$ and $\mu(x) = \kappa(x) = 1$. Hence, the sequence of probability measures induced by $\phi_t(J_n^\alpha)$ may not be exponentially tight for $\alpha \in [0,1)$. For $\alpha \in (0,1)$, this scaling is not covered by the results in [1], [6] and [11].

Nevertheless, it follows from Theorem 4.4 that $\frac{1}{n}M_n^\alpha(t)$ satisfies an LDP with rate function I^α, where $\frac{1}{n}M_n^\alpha(t)$ is the number of jobs in the system when the background process is J_n^α and I^α is given by

$$I^\alpha(a) = \inf_{\gamma \in \mathcal{R}^\alpha(t)}[\ell(\gamma; a) + \psi^\alpha(\gamma)].$$

Now recall that $\{\psi^\alpha < \infty\} \subset \mathcal{R}^\alpha(t)$. Also observe that $\{\xi < \infty\} = H_1([0,t])$ and that $\{\psi < \infty\} = \{\phi_t(f) | f \in H_1([0,t])\}$. Then we may rewrite I^α as $I^\alpha(a) = \ell(\phi_t(f_0); a)$ if $\alpha > 1$, $I^\alpha(a) = \inf_{f \in H_1([0,t])} \ell(\phi_t(f); a)$ if $\alpha \in [0,1)$ and

$$I^\alpha(a) = \inf_{f \in H_1([0,t])}[\ell(\phi_t(f); a) + \psi(\phi_t(f))]$$

if $\alpha = 1$.

7 Discussion and concluding remarks

In this paper, we studied an infinite-server queue in a random environment and proved a full LDP for the transient number of jobs in the system. The proof of this LDP has two essential ingredients, namely the result that the transient number of jobs in the system has a Poisson distribution with a random parameter and the assumption that the random parameter satisfies an LDP. Hence, the large deviations behavior of the random parameter seems to be the crucial factor that determines the large deviations behavior of the number of jobs in this specific queueing system.

The rate function corresponding to the LDP for the number of jobs is rather abstract. Nevertheless, we showed in the examples how to compute the rate function in certain specific cases. In particular, we recovered earlier obtained results for Markov-modulated infinite-server queues and strengthened these to a full LDP. Additionally, we proved LDPs when the background process has an

uncountable state space. In all examples, knowledge about the behavior of the background process could be exploited to describe the rate function.

The results in this paper also show that we do not have to restrict ourselves to background processes with finite state space or service requirements with an exponential distribution. Moreover, the proof of the LDP shows that assumptions about good rate functions used to study large deviations of mixtures may be unnecessarily restrictive when dealing with queueing systems.

There are several interesting topics for future research on the modulated infinite-server queue presented here. In this paper, we only looked at large deviations of the number of jobs at a fixed time $t \geq 0$. However, for certain applications it may be desirable to know the deviations over the whole time interval $[0, t]$. Therefore, it would be interesting to consider sample path large deviations. Also moderate deviations could be worth investigating, so as to bridge the gap between the central limit theorems and the large deviations results for modulated infinite-server queues. It is unlikely, though, that we may obtain such results under as few assumptions as in this paper.

Furthermore, it would be interesting to see whether the large deviations results for modulated infinite-server queues carry over to modulated Ornstein-Uhlenbeck processes. To the best of our knowledge, this has not been investigated so far.

Acknowledgement. This research has been partly funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

A Transient number of jobs in the system

In this section, we provide the precise mathematical description of the model and determine the distribution of the number of jobs in the system at time $t \geq 0$, which is denoted by $M(t)$. We mentioned in Section 1 that the steady-state distribution of the number of jobs in the system has already been determined for specific background processes and service requirements in Model I and Model II. However, in this case we would like to determine the transient distribution given a general càdlàg background process for the model described below, which generalizes Model I and Model II and includes general service times.

Throughout this section, we denote by $D([0,\infty);\mathcal{E})$ the space of càdlàg functions from $[0,\infty)$ to $\mathcal{E}$, where $\mathcal{E}$ is a metric space with metric ρ. Throughout, we assume that $\mathcal{E}$ is equipped with the Borel σ-algebra $\mathcal{B}(\mathcal{E})$ induced by ρ. We define, in the usual way, a metric d° on $D([0,\infty);\mathcal{E})$ that generates the Skorokhod J_1 topology. (For more details, see Section B and references there.)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which we have defined a standard Poisson process $\overline{Y}$ and a càdlàg stochastic process J with state space $\mathcal{E}$. Assume that $\mathcal{F}_\infty^{\overline{Y}}$ and $\mathcal{F}_\infty^{J}$ are independent.

Also assume that, for each $j \in \mathcal{E}$, we have defined on $(\Omega, \mathcal{F}, \mathbb{P})$ a sequence of independent, identically distributed, nonnegative random variables $Z(1,j)$, $Z(2,j), \ldots$ such that the map $(\omega, j) \mapsto Z(k,j)(\omega)$ is $\mathcal{Z} \otimes \mathcal{B}(\mathcal{E})/\mathcal{B}([0,\infty])$ mea-

surable, where $\mathcal{Z} = \sigma(Z(k,j) \,|\, k \in \mathbb{N}, j \in \mathcal{E})$. We denote the cumulative distribution function of $Z(1,j)$ by F_j. Note that $\mathcal{F}_\infty^{\overline{Y}}$, $\mathcal{F}_\infty^J$ and $\mathcal{Z}$ are independent and that the maps $(\omega, j, t) \mapsto \mathbb{1}_{\{Z(k,j)(\omega) \le t\}}$ and $(j,t) \mapsto F_j(t)$ are measurable with respect to the obvious σ-algebras.

Intuitively, $Z(k,j)$ describes the service requirement of job k if the background process J is in state j upon arrival of job k. The measurability assumption means that the background process should select the particular service requirement of job k 'in a measurable and independent way'. This is, of course, a very reasonable assumption and easily verifiable in many cases.

Now we know what the background process looks like and how the service requirements are modulated. To modulate the arrival rate and the server work rate, we take continuous functions $\lambda\colon \mathcal{E} \to [0,\infty)$ and $\mu\colon \mathcal{E} \to [0,\infty)$. Then the arrival rate at time $s \ge 0$ will be given by $\lambda(J(s))$ and the server work rate at time $s \ge 0$ will be given by $\mu(J(s))$.

Given a background process J, service requirements $Z(k,j)$ and functions λ and μ under the conditions described above, we will denote a modulated infinite-server queue by the quadruple (J, Z, λ, μ).

The modulated infinite-server queue (J, Z, λ, μ) is constructed as follows. We define the modulated Poisson process Y via

$$Y(t) = \overline{Y}\bigg(\int_0^t \lambda(J(s))\,\mathrm{d}s\bigg).$$

The process Y will be the arrival process. We denote the jump times of Y by $\tau_1, \tau_2, \ldots$ and the jump times of $\overline{Y}$ by $\overline{\tau}_1, \overline{\tau}_2, \ldots$. For convenience, we set $\tau_0 = \overline{\tau}_0 = 0$. The jump times τ_k and $\overline{\tau}_k$ are related via $\tau_k = \Lambda^-(\overline{\tau}_k)$ and $\overline{\tau}_k = \Lambda(\tau_k)$, where $\Lambda(t) = \int_0^t \lambda(J(s))\,\mathrm{d}s$ and $\Lambda^-(r) = \inf\{t \ge 0 \,|\, \Lambda(t) \ge r\}$.

Define the interarrival times $\sigma_k = \tau_k - \tau_{k-1}$ and $\overline{\sigma}_k = \overline{\tau}_k - \overline{\tau}_{k-1}$ for $k \in \mathbb{N}$. For later use, we note that $\overline{\sigma}_1, \overline{\sigma}_2, \ldots$ is a sequence of i.i.d. random variables with a standard exponential distribution.

At time $t = 0$ there are no jobs in the system. At each jump time of Y exactly one job arrives. Hence, the number of jobs that have entered the system during the time interval $[0,t]$ is given by the (a.s. finite) random variable $\sum_{k=1}^\infty \mathbb{1}_{\{\tau_k \le t\}}$.

When job k enters the system at time τ_k, its service requirement is given by $Z(k, J(\tau_k))$. In other words, job k draws an independent service requirement with cumulative distribution function $F_{J(\tau_k)}$. Job k leaves the system when its service requirement has been processed by the server, whose work rate is modulated by the background process J and is equal to $\mu(J(s))$ for $s \ge 0$.

Hence, job k has both entered and left the system before time $t \ge 0$ if and only if $\tau_k \le t$ and $Z(k, J(\tau_k)) \le \int_{[\tau_k, t)} \mu(J(r))\,\mathrm{d}r$. We get

$$M(t) = \sum_{k=1}^\infty \bigg(\mathbb{1}_{\{\tau_k \le t\}} - \mathbb{1}_{\{\tau_k \le t\}} \mathbb{1}_{\left\{ Z(k, J(\tau_k)) \le \int_{[\tau_k, t)} \mu(J(r))\,\mathrm{d}r \right\}} \bigg).$$

Note that $M(t)$ is a càdlàg stochastic process.

If J is deterministic, then it is relatively easy to determine the distribution of $M(t)$. For instance, one may compute the characteristic function of $M(t)$ via the following steps.

Suppose that $J(\omega,t) = f(t)$ for all $\omega \in \Omega$ and $t \geq 0$ for some function $f \in D([0,\infty);\mathcal{E})$. We may write the characteristic function of $M(t)$ as

$$\mathbb{E}\exp(\mathrm{i}\theta M(t)) = \mathbb{E}\exp\left(\mathrm{i}\theta\sum_{k=1}^{\infty}\left(\mathbb{1}_{\{\tau_k\leq t\}} - \mathbb{1}_{\{\tau_k\leq t\}}\mathbb{1}_{\left\{Z(k,J(\tau_k))\leq\int_{t\wedge\tau_k}^{t}\mu(J(r))\,\mathrm{d}r\right\}}\right)\right) =$$

$$= \mathbb{E}\mathbb{1}_{\{\tau_1>t\}} + \sum_{n=1}^{\infty}\mathbb{E}\mathbb{1}_{\{\tau_n\leq t;\tau_{n+1}>t\}}\exp\left(\mathrm{i}\theta\left(n - \sum_{k=1}^{n}\mathbb{1}_{\left\{Z(k,f(\tau_k))\leq\int_{t\wedge\tau_k}^{t}\mu(f(r))\,\mathrm{d}r\right\}}\right)\right).$$

Clearly, $\mathbb{E}\mathbb{1}_{\{\tau_1>t\}} = e^{-\int_0^t\lambda(f(s))\,\mathrm{d}s} = e^{-\Lambda(t)}$. We are left with computing the infinite sum above. Fix $n\in\mathbb{N}$ and note that

$$\mathbb{E}\mathbb{1}_{\{\tau_n\leq t;\tau_{n+1}>t\}}\exp\left(\mathrm{i}\theta\left(n - \sum_{k=1}^{n}\mathbb{1}_{\left\{Z(k,f(\tau_k))\leq\int_{t\wedge\tau_k}^{t}\mu(f(r))\,\mathrm{d}r\right\}}\right)\right) =$$

$$= \mathbb{E}\left(\mathbb{1}_{\{\tau_n\leq t;\tau_{n+1}>t\}}\exp(\mathrm{i}\theta n)\mathbb{E}\left[\exp\left(-\mathrm{i}\theta\sum_{k=1}^{n}\mathbb{1}_{\left\{Z(k,f(\tau_k))\leq\int_{t\wedge\tau_k}^{t}\mu(f(r))\,\mathrm{d}r\right\}}\right)\,\middle|\,\tau_1,\tau_2,\ldots\right]\right)$$

$$= \mathbb{E}\mathbb{1}_{\{\tau_n\leq t;\tau_{n+1}>t\}}\prod_{k=1}^{n}\left(\exp(\mathrm{i}\theta) + (1+\exp(\mathrm{i}\theta))F_{f(\tau_k)}\left(\int_{t\wedge\tau_k}^{t}\mu(f(r))\,\mathrm{d}r\right)\right),$$

because Y and the collection of service requirements are independent. For convenience, we write

$$h(\tau_k) = \exp(\mathrm{i}\theta) + (1+\exp(\mathrm{i}\theta))F_{f(\tau_k)}\left(\int_{t\wedge\tau_k}^{t}\mu(f(r))\,\mathrm{d}r\right).$$

Summarizing, we get

$$\mathbb{E}\exp(\mathrm{i}\theta M(t)) = \mathbb{E}\mathbb{1}_{\{\tau_1>t\}} + \sum_{n=1}^{\infty}\mathbb{E}\mathbb{1}_{\{\tau_n\leq t;\tau_{n+1}>t\}}\prod_{k=1}^{n}h(\tau_k).$$

Next, observe that

$$\begin{aligned}\mathbb{E}\mathbb{1}_{\{\tau_n\leq t;\tau_{n+1}>t\}}\prod_{k=1}^{n}h(\tau_k) &= \mathbb{E}\left(\mathbb{1}_{\{\tau_n\leq t\}}\left(\prod_{k=1}^{n}h(\tau_k)\right)\mathbb{E}\left[\mathbb{1}_{\{\sigma_{n+1}>t-\tau_n\}}\,\middle|\,\tau_1,\ldots,\tau_n\right]\right)\\ &= \mathbb{E}\left(\mathbb{1}_{\{\tau_n\leq t\}}\left(\prod_{k=1}^{n}h(\tau_k)\right)e^{-(\Lambda(t)-\Lambda(\tau_n))}\right).\end{aligned}$$

We define $x_k^+ = x_1 + \cdots + x_k$. Straightforward calculations give

$$\begin{aligned}
&\mathbb{E}\left(\mathbb{1}_{\{\tau_n \le t\}} \left(\prod_{k=1}^{n} h(\tau_k)\right) e^{\Lambda(\tau_n)}\right) = \\
&= \mathbb{E}\mathbb{1}_{\{\overline{\tau}_n \le \Lambda(t)\}} \left(\prod_{k=1}^{n} h\big(\Lambda^-(\overline{\tau}_k)\big)\right) e^{\overline{\tau}_n} \\
&= \int_{x_1=0}^{\Lambda(t)} \int_{x_2=0}^{\Lambda(t)-x_1^+} \cdots \int_{x_n=0}^{\Lambda(t)-x_{n-1}^+} \prod_{k=1}^{n} h\big(\Lambda^-\big(x_k^+\big)\big)\, \mathrm{d}x_n \ldots \mathrm{d}x_1 \\
&= \int_{y_1=0}^{\Lambda(t)} \int_{y_2=y_1}^{\Lambda(t)} \cdots \int_{y_n=y_{n-1}}^{\Lambda(t)} \prod_{k=1}^{n} h\big(\Lambda^-(y_k)\big)\, \mathrm{d}y_n \ldots \mathrm{d}y_1 \\
&= \int_{z_1=0}^{t} \int_{z_2=z_1}^{t} \cdots \int_{z_n=z_{n-1}}^{t} \prod_{k=1}^{n} [h(z_k)\lambda(f(z_k))]\, \mathrm{d}z_n \ldots \mathrm{d}z_1.
\end{aligned}$$

Now note that for an integrable function g we have

$$\left[\int_0^t g(s)\, \mathrm{d}s\right]^n = n! \int_{z_1=0}^{t} \int_{z_2=z_1}^{t} \cdots \int_{z_n=z_{n-1}}^{t} \prod_{k=1}^{n} g(z_k)\, \mathrm{d}z_n \ldots \mathrm{d}z_1.$$

As a result, it holds that

$$\begin{aligned}
&\int_{z_1=0}^{t} \int_{z_2=z_1}^{t} \cdots \int_{z_n=z_{n-1}}^{t} \prod_{k=1}^{n} [h(z_k)\lambda(f(z_k))]\, \mathrm{d}z_n \ldots \mathrm{d}z_1 = \\
&= \frac{1}{n!} \left[\int_0^t h(s)\lambda(f(s))\, \mathrm{d}s\right]^n \\
&= \sum_{k=0}^{n} \frac{1}{k!} [\exp(\mathrm{i}\theta)\Lambda(t)]^k \frac{1}{(n-k)!} \left[(1-\exp(\mathrm{i}\theta)) \int_0^t F_{f(s)}\left(\int_s^t \mu(f(r))\, \mathrm{d}r\right) \lambda(f(s))\, \mathrm{d}s\right]^{n-k}.
\end{aligned}$$

Now we may write

$$
\begin{aligned}
&\mathbb{E}\exp(\mathrm{i}\theta M(t)) = \\
&= \mathbb{E}\mathbb{1}_{\{\tau_1 > t\}} + \sum_{n=1}^{\infty} \mathbb{E}\mathbb{1}_{\{\tau_n \le t; \tau_{n+1} > t\}} h(\tau_k) \\
&= e^{-\Lambda(t)} + \sum_{n=1}^{\infty} e^{-\Lambda(t)} \sum_{k=0}^{n} \frac{1}{k!} [\exp(\mathrm{i}\theta)\Lambda(t)]^k \frac{1}{(n-k)!} \left[(1-\exp(\mathrm{i}\theta)) \int_0^t F_{f(s)}\left(\int_s^t \mu(f(r))\,\mathrm{d}r \right) \lambda(f(s))\,\mathrm{d}s \right]^{n-k} \\
&= e^{-\Lambda(t)} \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{1}{k!} [\exp(\mathrm{i}\theta)\Lambda(t)]^k \frac{1}{(n-k)!} \left[(1-\exp(\mathrm{i}\theta)) \int_0^t F_{f(s)}\left(\int_s^t \mu(f(r))\,\mathrm{d}r \right) \lambda(f(s))\,\mathrm{d}s \right]^{n-k} \\
&= e^{-\Lambda(t)} \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \frac{1}{k!} [\exp(\mathrm{i}\theta)\Lambda(t)]^k \frac{1}{n!} \left[(1-\exp(\mathrm{i}\theta)) \int_0^t F_{f(s)}\left(\int_s^t \mu(f(r))\,\mathrm{d}r \right) \lambda(f(s))\,\mathrm{d}s \right]^{n} \\
&= \exp\left(-\Lambda(t) + \exp(\mathrm{i}\theta)\Lambda(t) + (1-\exp(\mathrm{i}\theta)) \int_0^t F_{f(s)}\left(\int_s^t \mu(f(r))\,\mathrm{d}r \right) \lambda(f(s))\,\mathrm{d}s \right) \\
&= \exp\left((\exp(\mathrm{i}\theta)-1) \int_0^t \left(1 - F_{f(s)}\left(\int_s^t \mu(f(r))\,\mathrm{d}r \right) \right) \lambda(f(s))\,\mathrm{d}s \right).
\end{aligned}
$$

Hence, in this case $M(t)$ has a Poisson distribution with parameter $\phi_t(f)$, where

$$
\phi_t(f) = \int_0^t \left(1 - F_{f(s)}\left(\int_s^t \mu(f(r))\,\mathrm{d}r \right) \right) \lambda(f(s))\,\mathrm{d}s. \tag{14}
$$

Given our modulated infinite-server queue, we may view ϕ_t as a map from $D([0,\infty);\mathcal{E})$ to $[0,\infty)$ and we will call ϕ_t the parameter map associated with the modulated infinite-server queue.

Note that $s \mapsto F_{f(s)}\left(\int_s^t \mu(f(r))\,\mathrm{d}r \right)$ is $\mathcal{B}([0,\infty))/\mathcal{B}([0,1])$ measurable, as it is a composition of measurable maps. Also note that

$$
0 \le \left(1 - F_{f(s)}\left(\int_s^t \mu(f(r))\,\mathrm{d}r \right) \right) \lambda(f(s)) \le \lambda(f(s))
$$

and that $s \mapsto \lambda(f(s))$ is càdlàg. Hence, the integral in equation (14) is actually well defined and finite.

Now suppose that J is not deterministic. In this case, we may use the independence of J and standard arguments to obtain that

$$
\begin{aligned}
\mathbb{E}\exp(\mathrm{i}\theta M(t)) &= \mathbb{E}\mathbb{E}[\exp(\mathrm{i}\theta M(t)) \mid J] \\
&= \mathbb{E}\exp\left([\exp(\mathrm{i}\theta)-1] \int_0^t \left(1 - F_{J(s)}\left(\int_s^t \mu(J(r))\,\mathrm{d}r \right) \right) \lambda(J(s))\,\mathrm{d}s \right).
\end{aligned}
$$

We summarize our findings in the following lemma.

Lemma A.1. *Under the stated conditions, $M(t)$ has a Poisson distribution with random parameter $\phi_t(J)$, where ϕ_t is the parameter map associated with*

the modulated infinite-server queue as defined via equation (14) *and thus*

$$\phi_t(J) = \int_0^t \left(1 - F_{J(s)}\left(\int_s^t \mu(J(r))\,\mathrm{d}r\right)\right)\lambda(J(s))\,\mathrm{d}s.$$

If we scale $\lambda(x) \mapsto n\lambda(x)$ *and* $J \mapsto J_n$*, then the number of jobs in the system* $M_n(t)$ *has a Poisson distribution with random parameter* $n\phi_t(J_n)$*.*

Lemma A.1 states that $M(t)$ has a Poisson distribution with random parameter $\phi_t(J)$, meaning that

$$\mathbb{E}\exp(\mathrm{i}\theta M(t)) = \mathbb{E}u(\phi_t(J),\theta) = \int_{[0,\infty)} u(\gamma,\theta)\,\mathrm{d}\nu(\gamma),$$

where $u(\gamma,\cdot)\colon \mathbb{R} \to \mathbb{C}$ is the characteristic function of a Poisson distribution with parameter $\gamma \in [0,\infty)$ and ν is the law of $\phi_t(J)$. We may also describe this as $M(t)$ having a mixed Poisson distribution with mixing measure ν, meaning that the law of $M(t)$ may be represented as

$$\mathbb{Q}(A) = \int_{[0,\infty)} \mathbb{Q}_\gamma(A)\,\mathrm{d}\nu(\gamma), \tag{15}$$

where $\mathbb{Q}_\gamma$ is the law of a random variable with a Poisson distribution with parameter $\gamma \in [0,\infty)$. Indeed, suppose that $\mathbb{Q}$ is defined by (15). Then standard measure-theoretic arguments (cf. [16, Pr III.2.1]) show that

$$\int_{\mathbb{R}} \exp(\mathrm{i}\theta x)\,\mathrm{d}\mathbb{Q}(x) = \int_{[0,\infty)}\int_{\mathbb{R}} \exp(\mathrm{i}\theta x)\,\mathrm{d}\mathbb{Q}_\gamma(x)\,\mathrm{d}\nu(\gamma) = \int_{[0,\infty)} u(\gamma,\theta)\,\mathrm{d}\nu(\gamma),$$

so $\mathbb{Q}$ is actually the law of $M(t)$.

We may use these observations about modulated infinite-server queues and mixed Poisson distributions to prove the following intuitive lemma.

Lemma A.2. *Let* A *and* B *be measurable subsets of* $\mathbb{R}$*. If* $\mathbb{P}(\phi_t(J) \in B) > 0$*, then*

$$\inf_{\gamma\in B} \mathbb{Q}_\gamma(A) \le \mathbb{P}(M(t) \in A \,|\, \phi_t(J) \in B) \le \sup_{\gamma\in B} \mathbb{Q}_\gamma(A),$$

where $\mathbb{Q}_\gamma$ *is the law of a random variable with a Poisson distribution with parameter* $\gamma \in [0,\infty)$*.*

Proof. Denote $\Omega^* = \{\phi_t(J) \in B\}$. If $\mathbb{P}(\Omega^*) > 0$, we define a new probability space $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$ by taking $\mathcal{F}^* = \{F \cap \Omega^* \,|\, F \in \mathcal{F}\}$ and $\mathbb{P}^*(F) = \mathbb{P}(F)/\mathbb{P}(\Omega^*)$ for $F \in \mathcal{F}^*$.

For each random function X defined on $(\Omega, \mathcal{F}, \mathbb{P})$ we denote its restriction to Ω^* by X^*, which is a random function on $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$. Using the independence assumptions (stated at the beginning of this section), it is easy to verify that J^* is an independent càdlàg stochastic process, $\overline{Y}^*$ is an independent standard Poisson process and $Z^*(1,j), Z^*(2,j), \ldots$ is a sequence of independent,

identically distributed random variables for each $j \in \mathcal{E}$. Moreover, the map $(\omega, j) \mapsto Z^*(k,j)(\omega)$ is $\mathcal{Z}^* \otimes \mathcal{B}(\mathcal{E})/\mathcal{B}([0,\infty])$ measurable and $Z^*(k,j)$ has cumulative distribution function F_j.

Following the procedure described in this section, we construct the modulated infinite-server queue (J^*, Z^*, λ, μ) on $(\Omega^*, \mathcal{F}^*, \mathbb{P}^*)$ and we denote the number of jobs in the system at time $t \geq 0$ by $K(t)$. Lemma A.1 implies that $K(t)$ has a Poisson distribution with random parameter $\phi_t(J^*)$.

It is easy to verify that

$$\begin{aligned}
\mathbb{P}^*(K(t) \in A) &= \mathbb{P}^*\left(\sum_{k=1}^{\infty}\left(\mathbb{1}_{\{\tau_k^* \leq t\}} - \mathbb{1}_{\{\tau_k^* \leq t\}}\mathbb{1}_{\{Z^*(k,J^*(\tau_k^*)) \leq \int_{[\tau_k^*,t)} \mu(J^*(r))\,\mathrm{d}r\}}\right) \in A\right) \\
&= \mathbb{P}^*\left(\left\{\sum_{k=1}^{\infty}\left(\mathbb{1}_{\{\tau_k \leq t\}} - \mathbb{1}_{\{\tau_k \leq t\}}\mathbb{1}_{\{Z(k,J(\tau_k)) \leq \int_{[\tau_k,t)} \mu(J(r))\,\mathrm{d}r\}}\right) \in A\right\} \cap \Omega^*\right) \\
&= \mathbb{P}(\{M(t) \in A\} \cap \Omega^*)/\mathbb{P}(\Omega^*) \\
&= \mathbb{P}(M(t) \in A \,|\, \phi_t(J) \in B).
\end{aligned}$$

Since $K(t)$ has a Poisson distribution with random parameter $\phi_t(J^*)$, it follows that $\mathbb{P}(M(t) \in A \,|\, \phi_t(J) \in B) = \int_{[0,\infty)} \mathbb{Q}_\gamma(A)\,\mathrm{d}\nu^*(\gamma)$, where ν^* is the law of $\phi_t(J^*)$. But we have $\mathbb{P}^*(\phi_t(J^*) \in \mathbb{R} \setminus B) = 0$, so $\mathbb{P}(M(t) \in A \,|\, \phi_t(J) \in B) = \int_B \mathbb{Q}_\gamma(A)\,\mathrm{d}\nu^*(\gamma)$. Now observe that

$$\inf_{\gamma \in B} \mathbb{Q}_\gamma(A) = \int_B \inf_{\beta \in B} \mathbb{Q}_\beta(A)\,\mathrm{d}\nu^*(\gamma) \leq \int_B \mathbb{Q}_\gamma(A)\,\mathrm{d}\nu^*(\gamma) \leq \int_B \sup_{\beta \in B} \mathbb{Q}_\beta(A)\,\mathrm{d}\nu^*(\gamma) = \inf_{\gamma \in B} \mathbb{Q}_\gamma(A),$$

which completes the proof. □

In the following example we will describe the main example of a modulated infinite-server queue in this paper, which is essentially the modulated M/M/∞ queue. It is constructed such that job k has a service requirement with an exponential distribution with parameter $\kappa(j)$ if the background process is in state j upon its arrival, with κ some continuous function. This example includes Model I and Model II as mentioned in Section 1.

Example A.3. Let J be a background process with state space $\mathcal{E}$, as described above. Let λ, μ and κ be continuous maps from $\mathcal{E}$ to $[0,\infty)$ and let $\overline{Z}_1, \overline{Z}_2, \ldots$ be a sequence of independent standard exponential random variables. Define

$$Z(k,j)(\omega) = \begin{cases} \overline{Z}_k(\omega)/\kappa(j) & \text{if } \kappa(j) > 0; \\ \infty & \text{if } \kappa(j) = 0. \end{cases}$$

Clearly, $\mathcal{F}_\infty^{\overline{Y}}$, $\mathcal{F}_\infty^J$ and $\mathcal{Z}$ are independent in this example. Using that

$$\begin{aligned}
\left\{(\omega, j) \in \Omega \times \mathcal{E} \,\middle|\, \overline{Z}_k(\omega)/\kappa(j) \in (a,\infty)\right\} = \\
\bigcup_{q \in (0,\infty) \cap \mathbb{Q}} \left\{\omega \in \Omega \,\middle|\, \overline{Z}_k(\omega) \in (qa,\infty)\right\} \times \{j \in \mathcal{E} \,|\, \kappa(j) \in (0,q)\}
\end{aligned}$$

for $a \in (0,\infty)$, it is readily verified that the map $(\omega, j) \mapsto Z(k,j)(\omega)$ is $\mathcal{Z} \otimes \mathcal{B}(\mathcal{E})/\mathcal{B}([0,\infty])$ measurable.

In this case, the number of jobs $M(t)$ in the infinite-server queue (J, λ, μ, Z) has a Poisson distribution with random parameter $\phi_t(J)$, where the parameter map ϕ_t is given by

$$\phi_t(f) = \int_0^t \lambda(f(s)) e^{-\kappa(f(s)) \int_s^t \mu(f(r)) \, \mathrm{d}r} \, \mathrm{d}s. \tag{16}$$

This follows immediately from Lemma A.1 and the construction of Z.

The next example describes a nonexponential queue: the service requirement distributions will be Pareto distributions. The background process will determine the shape parameter of the Pareto distributions.

Example A.4. Let J be a background process with state space $\mathcal{E} = \mathbb{R}$. Let λ and μ be continuous maps from $\mathcal{E}$ to $[0,\infty)$. Let κ be a continuous map from $\mathcal{E}$ to $(0,\infty)$ and let $\overline{Z}_1, \overline{Z}_2, \ldots$ be a sequence of independent random variables having a Pareto distribution with scale parameter 1 and shape parameter 1, so that $\mathbb{P}\big(\overline{Z}_1 > x\big) = 1/(1 \vee x)$ for $x \in \mathbb{R}$. Define

$$Z(k,j)(\omega) = \big(\overline{Z}_k(\omega)\big)^{1/\kappa(j)}.$$

In a similar way as in the previous example one checks that $\mathcal{F}_\infty^{\overline{Y}}$, $\mathcal{F}_\infty^J$ and $\mathcal{Z}$ are independent and that the map $(\omega, j) \mapsto Z(k,j)(\omega)$ is $\mathcal{Z} \otimes \mathcal{B}(\mathcal{E})/\mathcal{B}([0,\infty])$ measurable.

Given these service requirements, the number of jobs $M(t)$ in the infinite-server queue (J, λ, μ, Z) has a Poisson distribution with random parameter $\phi_t(J)$, where the parameter map ϕ_t is given by

$$\phi_t(f) = \int_0^t \lambda(f(s)) \bigg(1 \vee \int_s^t \mu(f(r)) \, \mathrm{d}r\bigg)^{-\kappa(f(s))} \, \mathrm{d}s. \tag{17}$$

As before, this follows from Lemma A.1 and the construction of Z.

B Continuity and convergence in Skorokhod space

In the previous section we showed that $M(t)$ has a Poisson distribution with a random parameter $\phi_t(J)$, where ϕ_t is the parameter map associated with the modulated infinite-server queue (J, Z, λ, μ). For specific choices of the service requirements $Z(k,j)$, the map ϕ_t enjoys several continuity and convergence properties. We explore some of these properties in this section, mainly for the setup of Example A.3.

Let $\mathcal{E}$ be a metric space with metric ρ. Let $D([0,\infty); \mathcal{E})$ denote the space of càdlàg functions $f\colon [0,\infty) \to \mathcal{E}$, i.e., $\lim_{s \downarrow t} f(s) = f(t)$ and $\lim_{s \uparrow t} f(s)$ exists in $\mathcal{E}$ for every $t \geq 0$, where $\lim_{s \uparrow 0} f(s) := f(0)$ by convention.

Define a metric d° on $D([0,\infty);\mathcal{E})$ via

$$d^\circ(f,g) = \inf_{\lambda\in\Lambda}\left[\gamma(\lambda) \vee \int_0^\infty e^{-u} d(f,g,\lambda,u)\,\mathrm{d}u\right].$$

Here, Λ denotes the space of increasing homeomorphisms of $[0,\infty)$,

$$\gamma(\lambda) = \sup_{t>s\geq 0} |\log(\lambda(t)-\lambda(s)) - \log(t-s)|$$

and

$$d(f,g,\lambda,u) = \sup_{t\in[0,\infty)} [1 \wedge \rho(f(t\wedge u), g(\lambda(t)\wedge u))].$$

The metric d° induces the Skorokhod J_1 topology. For more details, see [13] or [21].

Definition B.1. A function $f_\mathrm{c} \in D([0,\infty);\mathcal{E})$ is called a *piecewise constant function* or a *step function* if there exist $n \in \mathbb{N}$, finitely many time points $0 = t_0 < t_1 < \ldots < t_n < \infty$ and $\alpha_0,\ldots,\alpha_n \in \mathcal{E}$ such that $f_\mathrm{c}(t) = \alpha_i$ for $t \in [t_i, t_{i+1})$ and $i = 0,\ldots,n-1$ and $f_\mathrm{c}(t) = \alpha_n$ for $t \in [t_n,\infty)$.

The set of step functions in $D([0,\infty);\mathcal{E})$ is denoted by $\mathcal{S}([0,\infty);\mathcal{E})$.

Proposition B.2. *Let $f \in D([0,\infty);\mathcal{E})$. For all $T > 0$ and $\epsilon > 0$ there exists a step function $f_\mathrm{c} \in \mathcal{S}([0,\infty);\mathcal{E})$ such that*

$$\sup_{t\in[0,T]} \rho(f(t), f_\mathrm{c}(t)) < \epsilon.$$

Proof. This is derived in the same way as [21, Th. 12.2.2]. □

Corollary B.3. *The set $\mathcal{S}([0,\infty);\mathcal{E})$ is dense in $D([0,\infty);\mathcal{E})$.*

Consequently, every continuous function on $D([0,\infty);\mathcal{E})$ is completely determined by its behavior on the set of step functions.

Now we will investigate properties of the parameter map under the assumptions of Example A.3. Let $\lambda\colon \mathcal{E} \to [0,\infty)$, $\kappa\colon \mathcal{E} \to [0,\infty)$ and $\mu\colon \mathcal{E} \to [0,\infty)$ be continuous. For $t \geq 0$, we would like to show that the function $\phi_t\colon D([0,\infty);\mathcal{E}) \to [0,\infty)$ defined by equation (16) is continuous. Note that ϕ_t has the form

$$\phi_t(f) = \int_0^t \lambda(f(s)) e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r}\,\mathrm{d}s$$

in this case and that it is the parameter map obtained in Example A.3.

First, we observe that the map $c_\lambda\colon D([0,\infty);\mathcal{E}) \to D([0,\infty);\mathbb{R})$ defined via $c_\lambda(f)(t) = \lambda(f(t))$ is continuous, because λ is continuous. Similarly, the functions c_κ and c_μ are continuous.

Next, let $f,g \in D([0,\infty);\mathbb{R})$. Then pointwise multiplication of f and g is defined via $(fg)(t) = f(t)g(t)$. This is a measurable map which is continuous at (f,g) if f or g is continuous (cf. [20, Th. 4.2]).

Finally, let $f \in D([0,\infty);\mathbb{R})$. Then the map $\psi\colon D([0,\infty);\mathbb{R}) \to D([0,\infty);\mathbb{R})$ defined via $\psi(t) = \int_0^t f(s)\,\mathrm{d}s$ is continuous. This follows almost immediately from the definition of ψ and the characterization in [13, Pr. 3.5.3].

Now note that the sequence of functions $\{\lambda(f_n)\}_{n\in\mathbb{N}}$ is bounded in the sup norm over $[0,t]$ if $f_n \to f$ in $D([0,\infty);\mathcal{E})$. Then it suffices to show that

$$\int_0^t e^{-\kappa(f_n(s))\int_s^t \mu(f_n(r))\,\mathrm{d}r}\,\mathrm{d}s \to \int_0^t e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r}\,\mathrm{d}s$$

as $f_n \to f$ in $D([0,\infty);\mathcal{E})$. But this follows from repeated applications of the first three observations.

Hence, the map ϕ_t must be continuous. Observe that continuity of λ, κ and μ is crucial to obtain this result. We summarize these findings in the following lemma.

Lemma B.4. *Let $\lambda\colon \mathcal{E} \to [0,\infty)$, $\kappa\colon \mathcal{E} \to [0,\infty)$ and $\mu\colon \mathcal{E} \to [0,\infty)$ be continuous. Then the function $\phi_t\colon D([0,\infty);\mathcal{E}) \to [0,\infty)$ as defined in equation (16) is continuous. Consequently, the parameter map obtained in Example A.3 is continuous.*

Another property of the map ϕ_t as defined in equation (16) is described in Lemma B.6. We will use the following easy lemma in the proof of Lemma B.6.

Lemma B.5. *Let $x, y \in (-\infty, 0]$. If $0 \le \alpha \le \beta < \infty$, then $|(e^x)^\alpha - (e^y)^\alpha| \le 1 - e^{-\beta|x-y|}$.*

Lemma B.6. *Let $\lambda\colon \mathcal{E} \to [0,\infty)$, $\kappa\colon \mathcal{E} \to [0,\infty)$ and $\mu\colon \mathcal{E} \to [0,\infty)$ be continuous and let $\phi_t\colon D([0,\infty);\mathcal{E}) \to [0,\infty)$ be defined by equation (16).*

Let $f, g \in D([0,\infty);\mathcal{E})$ and assume that there exists a finite set $\mathcal{E}^* \subset \mathcal{E}$ such that $f(s) \in \mathcal{E}^*$ and $g(s) \in \mathcal{E}^*$ for all $s \in [0,t]$ and that the set $A = \{s \in [0,t] \,|\, f(s) \neq g(s)\}$ has Lebesgue measure ϵ. Then

$$|\phi_t(f) - \phi_t(g)| \le \lambda_+(1+t)\big(1 - e^{-\epsilon\kappa_+\mu_+} + \epsilon\big), \tag{18}$$

where $\lambda_+ = \max_{x\in\mathcal{E}^*} \lambda(x)$, $\mu_+ = \max_{x\in\mathcal{E}^*} \mu(x)$ and $\kappa_+ = \max_{x\in\mathcal{E}^*} \kappa(x)$.

Proof. Clearly, we have

$$\begin{aligned}|\phi_t(f) - \phi_t(g)| \le \int_A \Big|\lambda(f(s))e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r} - \lambda(g(s))e^{-\kappa(g(s))\int_s^t \mu(g(r))\,\mathrm{d}r}\Big|\,\mathrm{d}s \\ + \int_{[0,t]\setminus A} \Big|\lambda(f(s))e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r} - \lambda(g(s))e^{-\kappa(g(s))\int_s^t \mu(g(r))\,\mathrm{d}r}\Big|\,\mathrm{d}s.\end{aligned}$$

Denote the first integral on the right-hand side by I_1 and the second integral on the right-hand side by I_2. It is easy to see that I_1 is bounded above by $\epsilon\lambda_+$.

We may find an upper bound for I_2 as follows. For $s \in [0,t] \setminus A$ we have

$$\begin{aligned}\Big|\lambda(f(s))e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r} - \lambda(g(s))e^{-\kappa(g(s))\int_s^t \mu(g(r))\,\mathrm{d}r}\Big| = \\ \Big|\lambda(f(s))e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r} - \lambda(f(s))e^{-\kappa(f(s))\int_s^t \mu(g(r))\,\mathrm{d}r}\Big| \leq \\ \lambda_+\Big|\Big(e^{-\int_s^t \mu(f(r))\,\mathrm{d}r}\Big)^{\kappa(f(s))} - \Big(e^{-\int_s^t \mu(g(r))\,\mathrm{d}r}\Big)^{\kappa(f(s))}\Big| \leq \\ \lambda_+\big(1 - e^{-\kappa_+\epsilon\mu_+}\big).\end{aligned}$$

To obtain the last inequality, we apply Lemma B.5 and use that

$$\sup_{s\in[0,t]}\Big|\int_s^t \mu(f(r))\,\mathrm{d}r - \int_s^t \mu(g(r))\,\mathrm{d}r\Big| \leq \epsilon\mu_+.$$

It follows that

$$\begin{aligned}\int_{[0,t]\setminus A}\Big|\lambda(f(s))e^{-\kappa(f(s))\int_s^t \mu(f(r))\,\mathrm{d}r} - \lambda(g(s))e^{-\kappa(g(s))\int_s^t \mu(g(r))\,\mathrm{d}r}\Big|\,\mathrm{d}s \leq \\ \lambda_+ t\big(1 - e^{-\kappa_+\epsilon\mu_+}\big).\end{aligned}$$

Combining the upper bounds for I_1 and I_2 proves the lemma. □

To conclude this section, we provide a lemma asserting the continuity of the map ϕ_t as defined by equation (17). The continuity is established using the same arguments as above.

Lemma B.7. *Let $\lambda\colon \mathcal{E} \to [0,\infty)$ and $\mu\colon \mathcal{E} \to [0,\infty)$ be continuous. Then the map $\phi_t\colon D([0,\infty);\mathcal{E}) \to [0,\infty)$ defined by*

$$\phi_t(f) = \int_0^t \lambda(f(s))\Big(1 \vee \int_s^t \mu(f(r))\,\mathrm{d}r\Big)^{-f(s)}\,\mathrm{d}s \tag{19}$$

is continuous. Consequently, the parameter map obtained in Example A.4 is continuous.

C Properties of Poisson random variables

For $\gamma \geq 0$, let $P_0(\gamma), P_1(\gamma), P_2(\gamma), \ldots$ denote a sequence of i.i.d. random variables that have a Poisson distribution with parameter γ. In this section, we will fix an arbitrary $x \in \mathbb{R}$, $\delta > 0$, $\lambda \geq 0$ and $\epsilon > 0$ and define $\lambda_\epsilon^- = \max\{0, \lambda - \epsilon\}$ and $\lambda_\epsilon^+ = \lambda + \epsilon$. Recall that $B_+(\lambda,\epsilon) = B(\lambda,\epsilon) \cap \mathbb{R}_+$.

We would like to prove a large deviations lower bound for

$$\liminf_{n\to\infty} \inf_{\gamma\in B_+(\lambda,\epsilon)} \frac{1}{n}\log \mathbb{P}\bigg(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in B(x,\delta)\bigg).$$

Of course, the difficulty here is the presence of the infimum over a range of parameters. We will show in Proposition C.1 that this infimum may be taken over certain restricted subsets of $B_+(\lambda,\epsilon)$. For each of these subsets we will provide a large deviations lower bound, from which we will derive a lower bound when the infimum is taken over $B_+(\lambda,\epsilon)$. This is the content of Proposition C.3.

Proposition C.1. *For all $x \in \mathbb{R}$, $\delta > 0$, $\lambda \geq 0$ and $\epsilon > 0$ it holds that*

$$\inf_{\gamma \in B_+(\lambda,\epsilon)} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in B(x,\delta)\right) = \inf_{\gamma \in (B(\lambda,\epsilon)\cap B[x,\delta]) \cup \{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in B(x,\delta)\right).$$

Proof. Let $0 \leq \gamma_- \leq \gamma_+ < \infty$. For $y \in \mathbb{R}$ it holds that

$$\mathbb{P}(P_0(\gamma_+) = y) \geq \mathbb{P}(P_0(\gamma_-) = y) \qquad \text{if} \qquad y \geq \gamma_+ \geq \gamma_- \tag{20}$$

and

$$\mathbb{P}(P_0(\gamma_+) = y) \leq \mathbb{P}(P_0(\gamma_-) = y) \qquad \text{if} \qquad \gamma_+ \geq \gamma_- \geq y. \tag{21}$$

Because we are working with i.i.d. Poisson random variables, we may write

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in B(x,\delta)\right) = \mathbb{P}(P_0(n\gamma) \in (n(x-\delta), n(x+\delta))). \tag{22}$$

Now the statement of the proposition is an easy consequence of the equations (20), (21) and (22) combined. □

Proposition C.2. *Let $x \in \mathbb{R}$ and $\delta > 0$. If $B_+(x,\delta) \neq \emptyset$, then*

$$\lim_{n\to\infty} \inf_{\gamma \in B_+[x,\delta]} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in B(x,\delta)\right) = 0.$$

Proof. For a Borel set $A \subset \mathbb{R}$, define $p_n(A|\gamma) = \mathbb{P}\big(\frac{1}{n}\sum_{i=1}^n P_i(\gamma) \in A\big)$. Now suppose that $B_+(x,\delta) \neq \emptyset$. Then the diameter of $B_+(x,\delta)$ is strictly positive and bounded above by $r = \min\{2\delta, x+\delta\}$.

Let $N_r \in \mathbb{N}$ be such that $\frac{1}{N_r} < \frac{r}{2}$. Then for all $n \geq N_r$ and $\gamma \in B_+[x,\delta]$ we define $\gamma_n^- = \frac{1}{n}\lfloor n\gamma \rfloor$, $\gamma_n^+ = \frac{1}{n}\lceil n\gamma \rceil$ and

$$\gamma_n^* = \min\big\{\{\gamma_n^-, \gamma_n^+\} \cap B(x,\delta)\big\}.$$

Then $\max\{|\gamma - \gamma_n^-|, |\gamma - \gamma_n^+|\} \leq \frac{1}{n} < \frac{r}{2}$ and $p_n(B(x,\delta)|\gamma) \geq p_n(\{\gamma_n^*\}|\gamma)$ for each $n \in \mathbb{N}$ and each $\gamma \in B_+[x,\delta]$. Using that $n! \leq n^{n+1/2}e^{-n+1}$, we get

$$\begin{aligned} p_n(\{\gamma_n^*\}|\gamma) &\geq \left(\frac{n\gamma}{n\gamma+1}\right)^{n\gamma_n^*} e^{n(\gamma_n^*-\gamma)} e^{-1} (n\gamma_n^*)^{-1/2} \\ &\geq \left(1 - \frac{1}{n(x+\delta)+1}\right)^{n(x+\delta)} e^{-2} (n(x+\delta))^{-1/2} \end{aligned}$$

for each $n \in \mathbb{N}$ and each $\gamma \in B_+[x,\delta]$. This implies the statement. □

Combined with Cramér's Theorem in $\mathbb{R}$, the two previous propositions enable us to prove the following large deviations bound. Note that we prove an equality rather than an inequality and that the limit exists.

Proposition C.3. *For all $x \in \mathbb{R}$, $\delta > 0$, $\lambda \geq 0$ and $\epsilon > 0$ it holds that*

$$\lim_{n\to\infty} \inf_{\gamma \in B_+(\lambda,\epsilon)} \frac{1}{n} \log \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} P_i(\gamma) \in B(x,\delta) \right) = \min_{\gamma \in \{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \left[- \inf_{a \in B(x,\delta)} \ell(\gamma; a) \right]. \tag{23}$$

Proof. Define $p_n(A \,|\, \gamma) = \mathbb{P}\big(\frac{1}{n}\sum_{i=1}^{n} P_i(\gamma) \in A\big)$ for Borel sets $A \subset \mathbb{R}$ and $C = (B(\lambda,\epsilon) \cap B[x,\delta]) \cup \{\lambda_\epsilon^-, \lambda_\epsilon^+\}$. Thanks to Proposition C.1 we may write

$$\lim_{n\to\infty} \inf_{\gamma \in B_+(\lambda,\epsilon)} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma) = \lim_{n\to\infty} \inf_{\gamma \in C} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma).$$

It follows from Proposition C.2 that we may restrict the infimum to the set $\{\lambda_\epsilon^-, \lambda_\epsilon^+\}$, so

$$\begin{aligned}
\lim_{n\to\infty} \inf_{\gamma \in C} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma) &= \lim_{n\to\infty} \min_{\gamma \in \{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma) \\
&= \min_{\gamma \in \{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \lim_{n\to\infty} \frac{1}{n} \log p_n(B(x,\delta) \,|\, \gamma) \\
&= \min_{\gamma \in \{\lambda_\epsilon^-, \lambda_\epsilon^+\}} \left[- \inf_{a \in B(x,\delta)} \ell(\gamma; a) \right].
\end{aligned}$$

The last equality is an application of Cramér's Theorem for i.i.d. Poisson random variables; the limit exists because $B(x,\delta)$ is a continuity set for the Fenchel-Legendre transform corresponding to a Poisson distribution. □

As shown in the inequalities (8) and (9), the Fenchel-Legendre transforms corresponding to Poisson distributions are nicely ordered in some sense. This property leads to the following propositions. Their proofs are elementary but tedious and are therefore omitted.

Proposition C.4. *Let $F \subset \mathbb{R}$ be closed and define $f\colon [0,\infty) \to [-\infty, 0]$ via*

$$f(\gamma) = - \inf_{a \in F} \ell(\gamma; a).$$

If $F \subset (-\infty, 0)$, then $f \equiv -\infty$. If $F \cap [0,\infty) \neq \emptyset$, then f is real-valued and continuous on $(0,\infty)$. Additionally, $\lim_{\gamma \downarrow 0} f(\gamma) = f(0)$, where $f(0) = 0$ if $0 \in F$ and $f(0) = \infty$ if $0 \notin F$. In any case, $f^{-1}([a,b])$ is closed for all $a, b \in (-\infty, 0]$ with $a \leq b$.

Proposition C.5. *Let $\mathcal{R} \subset [0,\infty)$ be a non-empty, closed set. Let $\psi\colon \mathbb{R} \to [0,\infty]$ be a lower semi-continuous function. Then the function $I\colon \mathbb{R} \to [0,\infty]$ defined via*

$$I(a) = \inf_{\gamma \in \mathcal{R}} [\ell(\gamma; a) + \psi(\gamma)]$$

is a lower semi-continuous function.

References

[1] J. D. Biggins. Large deviations for mixtures. *Electronic Communications in Probability*, 9:60–71, 2004.

[2] J. Blom, O. Kella, M. Mandjes, and H. Thorsdottir. Markov-modulated infinite-server queues with general service times. *Queueing Systems*, 76(4):403–424, 2014.

[3] Joke Blom, Koen De Turck, Offer Kella, and Michel Mandjes. Tail asymptotics of a Markov-modulated infinite-server queue. *Queueing Systems*, 78(4):337–357, 2014.

[4] Joke Blom, Koen De Turck, and Michel Mandjes. Analysis of Markov-modulated infinite-server queues in the central-limit regime. *Probability in the Engineering and Informational Sciences*, 29(3):433–459, 2015.

[5] Joke Blom and Michel Mandjes. A large-deviations analysis of Markov-modulated infinite-server queues. *Operations Research Letters*, 41(3):220–225, 2013.

[6] Narasinga R. Chaganty. Large deviations for joint distributions and statistical applications. *Sankhyā A*, 59:147–166, 1997.

[7] Pauline Coolen-Schrijner and Erik A. van Doorn. The deviation matrix of a continuous-time Markov chain. *Probability in the Engineering and Informational Sciences*, 16(3):351–366, 2002.

[8] B. D'Auria. M/M/∞ queues in semi-Markovian random environment. *Queueing Systems*, 58(3):221–237, 2008.

[9] B. D'Auria, J. Ivanovs, O. Kella, and M. Mandjes. Two-sided reflection of Markov-modulated Brownian motion. *Stochastic Models*, 28(2):316–332, 2012.

[10] Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, second edition, 1998.

[11] I. H. Dinwoodie and S. L. Zabell. Large deviations for exchangeable random vectors. *The Annals of Probability*, 20(3):1147–1166, 1992.

[12] Maciej Dobrzyński and Frank J. Bruggeman. Elongation dynamics shape bursty transcription and translation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(8):2583–2588, 2009.

[13] Stewart N. Ethier and Thomas G. Kurtz. *Markov Processes: Characterization and Convergence*. Wiley, New York, 1986.

[14] Brian H. Fralix and Ivo J. B. F. Adan. An infinite-server queue influenced by a semi-Markovian environment. *Queueing Systems*, 61(1):65–84, 2009.

[15] H. M. Jansen, M. R. H. Mandjes, K. De Turck, and S. Wittevrongel. On the upper bound in Varadhan's Lemma. *Statistics and Probability Letters*, 103(1):24–29, 2015.

[16] Jacques Neveu. *Mathematical Foundations of the Calculus of Probability*. Holden-Day, San Francisco, 1965.

[17] C. A. O'Cinneide and P. Purdue. The M/M/∞ queue in a random environment. *Journal of Applied Probability*, 23(1):175–184, 1986.

[18] Alvaro Sanchez, Sandeep Choubey, and Jane Kondev. Stochastic models of transcription: From single molecules to single cells. *Methods*, 62:13–25, 2013.

[19] A. Schwabe, M. Dobrzyński, K. Rybakova, P. Verschure, and F. J. Bruggeman. Origins of stochastic intracellular processes and consequences for cell-to-cell variability and cellular survival strategies. In Daniel Jameson, Malkhey Verma, and Hans V. Westerhoff, editors, *Methods in Systems Biology*, volume 500 of *Methods in Enzymology*, pages 597–625. Academic Press, Burlington, 2011.

[20] Ward Whitt. Some useful functions for functional limit theorems. *Mathematics of Operations Research*, 5(1):67–85, 1980.

[21] Ward Whitt. *Stochastic-Process Limits: an Introduction to Stochastic-Process Limits and Their Application to Queues*. Springer, New York, 2002.