European Journal of Operational Research 000 (2016) 1–9

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Stochastics and Statistics

Refined large deviations asymptotics for Markov-modulated infinite-server systems[☆]

Joke Blom[a], Koen De Turck[b,*], Michel Mandjes[a,c,d,e]

[a] CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands
[b] Laboratoire Signaux et Systèmes (L2S, CNRS UMR8506), École CentraleSupélec, Université Paris Saclay, 3 Rue Joliot Curie, Plateau de Moulon, Gif-sur-Yvette 91190, France
[c] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, Amsterdam 1098 XH, the Netherlands
[d] Eurandom, Eindhoven University of Technology, Eindhoven, the Netherlands
[e] IBIS, Faculty of Economics and Business, University of Amsterdam, Amsterdam, the Netherlands

ARTICLE INFO

Article history:
Received 1 April 2016
Accepted 26 October 2016
Available online xxx

Keywords:
Queueing
Communication networks
Markov-modulation
Rare events
Large deviations

ABSTRACT

Many networking-related settings can be modeled by Markov-modulated infinite-server systems. In such models, the customers' arrival rates and service rates are modulated by a Markovian background process; additionally, there are infinitely many servers (and consequently the resulting model is often used as a proxy for the corresponding many-server model). The Markov-modulated infinite-server model hardly allows any explicit analysis, apart from results in terms of systems of (ordinary or partial) differential equations for the underlying probability generating functions, and recursions to obtain all moments. As a consequence, recent research efforts have pursued an asymptotic analysis in various limiting regimes, notably the central-limit regime (describing fluctuations around the average behavior) and the large-deviations regime (focusing on rare events). Many of these results use the property that the number of customers in the system obeys a Poisson distribution with a random parameter. The objective of this paper is to develop techniques to accurately approximate tail probabilities in the large-deviations regime. We consider the scaling in which the arrival rates are inflated by a factor N, and we are interested in the probability that the number of customers exceeds a given level Na. Where earlier contributions focused on so-called *logarithmic asymptotics* of this exceedance probability (which are inherently imprecise), the present paper improves upon those results in that *exact asymptotics* are established. These are found in two steps: first the distribution of the random parameter of the Poisson distribution is characterized, and then this knowledge is used to identify the exact asymptotics. The paper is concluded by a set of numerical experiments, in which the accuracy of the asymptotic results is assessed.

1. Introduction, notation, and preliminaries

Consider an infinite-server queue modulated by a finite-state irreducible continuous-time Markov chain J: when the so-called *background process* J is in state $i \in \{1, \dots, d\}$, jobs arrive according to a Poisson process with rate λ_i, while the departure rate is μ_i. These Markov-modulated infinite-server queues have attracted some attention during the past decades; see e.g. the early contributions of D'Auria (2008), Keilson and Servi (1993), O'Cinneide and Purdue (1986) and later Fralix and Adan (2009). Importantly, considerably fewer results are available for this model than for the corresponding *single*-server queue. This is primarily due to the fact that, despite the system's simple structure, the Markov-modulated infinite-server queue hardly allows any explicit analysis: whereas the Markov-modulated single-server queue has a matrix-geometric stationary distribution, no such result applies to its infinite-server counterpart. The results obtained so far are implicit, in that they are in terms of partial differential equations characterizing the probability generating functions related to the system's transient behavior, and recursions for the corresponding moments (where in each step of the recursion a system of non-homogeneous ordinary differential equations needs to be solved).

[☆] M. Mandjes' research is partly funded by the NWO Gravitation project NETWORKS, grant no. 024.002.003.

[*] Corresponding author.

E-mail addresses: Joke.Blom@cwi.nl (J. Blom), koen.deturck@centralesupelec.fr (K. De Turck), m.r.h.mandjes@uva.nl (M. Mandjes).

http://dx.doi.org/10.1016/j.ejor.2016.10.050
0377-2217/© 2016 Published by Elsevier B.V.

The Markov-modulated infinite-server queue can be applied in various domains, ranging from biology to the performance analysis of particular communication networks. In the present paper the focus lies on the latter application, where the model with an infinite number of servers typically serves as a proxy for its counterpart with a large but finite number of servers. The Markov modulation of the arrival rates and service rates facilitates the modeling of some sort of 'burstiness'; although the concept of Markov modulation has been around for a few decades, it still spurs a considerable amount of research effort (Horváth, 2015; O'Reilly, 2014). For instance, the model can be used to describe the fluctuations in the users' activity level (where each user alternates between transmitting data or being silent). Also, e.g. in a wireless setting, the modulation of the service rate can represent channel conditions that vary over time. In the context of communication networks, a particularly relevant feature concerns *rare events*. More specifically, a high activity level corresponds to congestion, and therefore the system should be designed such that such high activity levels occur relatively infrequently.

Given that, as argued above, explicit analysis is hardly possible, recent research efforts have focused on the exploration of various limiting regimes. In the first place, significant progress has been made in terms of the derivation of (functional) central limit theorems under specific parameter scalings. When inflating the arrival rates by a factor N, and speeding up the background process by a factor N^{α} (for some $\alpha > 0$), in e.g. (Anderson, Blom, Mandjes, Thorsdottir, and De Turck, 2014; Blom, De Turck, and Mandjes, 2015; Blom, De Turck, and Mandjes, 2016) it has been proven that the (transient as well as stationary) number of jobs present in the system is, after centering and normalizing, asymptotically Normally distributed. An interesting dichotomy was identified, in that the regimes $\alpha < 1$ and $\alpha > 1$ lead to qualitatively different asymptotics.

Also the large-deviations regime has been explored, resulting in so-called *logarithmic asymptotics* (Blom, Kella, Mandjes, & De Turck, 2014; Blom & Mandjes, 2013; Blom, De Turck, & Mandjes, 2013). In these papers the arrival rates are scaled by a factor N and the background process is either left unchanged or accelerated by a factor $N^{1+\varepsilon},\ \varepsilon > 0$. With $M^{(N)}(t)$ the number of jobs present at time t in the resulting system, these papers determine the limit

$$\lim_{N\to\infty}\frac{1}{N}\log p_t^{(N)}(a) =: -I(a),\quad \text{with}\quad p_t^{(N)}(a) := \mathbb{P}\Big(M^{(N)}(t) \geq Na\Big), \tag{1}$$

as well as the corresponding limit for $M^{(N)}(t)$'s steady-state counterpart $M^{(N)}$. It is observed that these asymptotics are inherently imprecise, as they essentially just entail that

$$p_t^{(N)}(a) = e^{-NI(a)}\Psi(N),$$

for some *unknown* subexponential function $\Psi(N)$; we only know that $\Psi(N)$ has the property that, as $N \to \infty$,

$$\frac{1}{N}\log\Psi(N) \to 0. \tag{2}$$

Observe that (2) still leaves a substantial amount of freedom: $\Psi(N)$ could be for instance a constant, but also any polynomial function of N, or even 'big functions' of the type $10^6 \cdot \exp(N^{0.99})$. We conclude that logarithmic asymptotics of the type (1) typically provide valuable insight into the system's rare-event behavior, but that they may be too inaccurate to be used for performance evaluation purposes. This shows that there is a clear need for more precise asymptotic results.

The main contribution of the present paper is to improve the logarithmic asymptotics (1) to so-called *exact* asymptotics: we identify an explicit function $\zeta(\cdot)$ such that, as $N \to \infty$,

$$\frac{p_t^{(N)}(a)}{\zeta(N)} \to 1.$$

As it turns out, this $\zeta(N)$ is the product of the exponential term identified above ($e^{-NI(a)}$), a polynomial term (which is typically of the form N^{-C}, for some $C > 0$), and a constant. The proof of this property consists of two steps, and relies on the property that $M^{(N)}(t)$ obeys a Poisson distribution with random parameter (as was observed in e.g. Blom et al., 2014; D'Auria, 2008).

- In the first step a system of partial differential equations is set up for the distribution of this Poisson parameter.
- In the second step, this is combined with (a uniform version) of the classical result by Bahadur and Rao (1960), Höglund (1979) on the exact tail asymptotics of sample means of i.i.d. random variables, so as to obtain the exact asymptotics of the tail probability of our interest.

Model and notation. As mentioned above, λ_i is the (Poissonian) arrival rate when the background process is in state i. We let

$$Q = (q_{ij})_{i,j=1}^{d}$$

be the ($d \times d$) transition rate matrix of the (irreducible) background process J, with $\boldsymbol{\pi}$ denoting the corresponding invariant probability measure (which is a d-dimensional vector $\boldsymbol{\pi}$). The entries of Q are non-negative, except for those on the diagonal; the row-sums are assumed to be 0, where we define $q_i := -q_{ii} \geq 0$.

Concerning the departure process, two models are considered. In the first, referred to as Model I, each job present is experiencing a departure rate μ_i when J is in state i; as a consequence, this hazard rate may change during the job's sojourn time (that is, when the background process makes a transition). In the second, Model II, the crucial difference is that the job's sojourn time is sampled upon arrival: when the background process is then in state i, it has an exponential distribution with mean $1/\mu_i$. The evident independence assumptions are imposed.

Preliminaries. In Models I and II, we have that $M^{(N)}(t)$ has a mixed Poisson distribution, i.e., a Poisson distribution with random parameter (Blom et al., 2014; D'Auria, 2008). More specifically, with $P(b)$ denoting a Poisson random variable with mean $b > 0$, our target probability $p_t^{(N)}(a)$ equals the probability $\mathbb{P}(P(N\phi_t(J)) \geq Na)$ in Model I and $\mathbb{P}(P(N\psi_t(J)) \geq Na)$ in Model II, where the functionals $\phi_t(J)$ and $\psi_t(J)$ of the path $J \equiv \{J(s)\colon s \in [0, t]\}$ are given by, respectively,

$$\phi_t(J) := \int_0^t \lambda_{J(s)} e^{-\int_s^t \mu_{J(r)}\mathrm{d}r}\mathrm{d}s \quad \text{and} \quad \psi_t(J) := \int_0^t \lambda_{J(s)} e^{-(t-s)\mu_{J(s)}}\mathrm{d}s.$$

An intuitive explanation for this property is the following. In Model II the probability of a job that has arrived at time s is still present at time $t \in (s, \infty)$ is

$$e^{-(t-s)\mu_{J(s)}},$$

as $\mu_{J(s)}$ is its hazard rate during its entire lifetime. In Model I this hazard rate may change over time, in the sense that when the background process is in state i it is μ_i; therefore, the probability of a job that has arrived at time s is still present at t is

$$e^{-\int_s^t \mu_{J(r)}\mathrm{d}r}.$$

In an earlier paper (Blom et al., 2014) we have developed a technique to determine for Model I numbers $a_t^{(-,\mathrm{I})}$ and $a_t^{(+,\mathrm{I})}$ (such that $0 \leq a_t^{(-,\mathrm{I})} \leq a_t^{(+,\mathrm{I})}$) being the smallest, resp. largest numbers that $\phi_t(J)$ can attain. The analogous result for $\psi_t(J)$ (featuring in Model II) has been presented in Blom and Mandjes (2013), resulting in numbers $a_t^{(-,\mathrm{II})}$ and $a_t^{(+,\mathrm{II})}$.