Methodol Comput Appl Probab (2019) 21:125–153
https://doi.org/10.1007/s11009-018-9644-1

Linear Stochastic Fluid Networks: Rare-Event Simulation and Markov Modulation

O. J. Boxma[1] · E. J. Cahen[2] · D. Koops[3] · M. Mandjes[3]

Received: 7 March 2017 / Revised: 16 May 2018 /
Accepted: 22 May 2018 / Published online: 4 June 2018

Abstract We consider a linear stochastic fluid network under Markov modulation, with a focus on the probability that the joint storage level attains a value in a rare set at a given point in time. The main objective is to develop efficient importance sampling algorithms with provable performance guarantees. For linear stochastic fluid networks without modulation, we prove that the number of runs needed (so as to obtain an estimate with a given precision) increases polynomially (whereas the probability under consideration decays essentially exponentially); for networks operating in the slow modulation regime, our algorithm is asymptotically efficient. Our techniques are in the tradition of the rare-event simulation procedures that were developed for the sample-mean of i.i.d. one-dimensional light-tailed random variables, and intensively use the idea of exponential twisting. In passing, we also point out how to set up a recursion to evaluate the (transient and stationary) moments of the joint storage level in Markov-modulated linear stochastic fluid networks.

Keywords Linear networks · Stochastic processes · Queues · Rare events · Importance sampling

Mathematics Subject Classification (2010) 60K25 · 60F10 · 65C05

✉ M. Mandjes
m.r.h.mandjes@uva.nl

[1] EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands

[2] CWI, Science Park 123, 1098 XG Amsterdam, The Netherlands

[3] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 105-107, 1098 XG Amsterdam, The Netherlands

1 Introduction

Linear stochastic fluid networks, as introduced in Kella and Whitt (1999), can be informally described as follows. Consider a network consisting of L stations. Jobs, whose sizes are i.i.d. samples from some general L-dimensional distribution, arrive at the stations according to a Poisson process. At each of the nodes, in between arrivals the storage level decreases exponentially. Processed traffic is either transferred to the other nodes or leaves the network (according to a given routing matrix). In addition to this basic version of the linear stochastic fluid network, there is also its *Markov modulated* counterpart (Kella and Stadje 2002), in which the arrival rate, the distribution of the job sizes, and the routing matrix depend on the state of an external, autonomously evolving finite-state continuous-time Markov chain (usually referred to as the *background process*).

Linear stochastic fluid networks can be seen as natural fluid counterparts of corresponding infinite-server queues. As such, they inherit several nice properties of those infinite-server queues. In particular, separate infinitesimally small fluid particles, moving through the network, do not interfere, and are therefore mutually independent. Essentially due to this property, linear stochastic fluid networks allow explicit analysis; in particular, the joint Laplace transform of the storage levels at a given point in time can be expressed in closed form as a function of the arrival rate, the Laplace transform of the job sizes and the routing matrix (Kella and Whitt 1999, Thm. 5.1).

When Markov modulation is imposed, the analysis becomes substantially harder. Conditional on the path of the background process, again explicit expressions can be derived, cf. (Kella and Stadje 2002, Thm. 1). Unconditioning, however, cannot be done in a straightforward manner. As a consequence the results found are substantially less explicit than for the non-modulated linear stochastic fluid network. In Kella and Stadje (2002) also a system of ordinary differential equations has been set up that provides the transform of the stationary storage level; in addition, conditions are identified that guarantee the existence of such a stationary distribution.

In this paper we focus on rare events for Markov-modulated linear stochastic fluid networks. More specifically, in a particular scaling regime (parameterized by n) we analyze the probability p_n that at a given point in time the network storage vector is in a given rare set. By scaling the arrival rate as well as the rare set (which amounts to multiplying them by a scaling parameter n), the event of interest becomes increasingly rare. More specifically, under a Cramér-type assumption on the job-size distribution, application of large-deviations theory yields that p_n decays (roughly) exponentially. As p_n can be characterized only asymptotically, one could consider the option of using simulation to obtain precise estimates. The effectiveness, however, of such an approach is limited due to the rarity of the event under consideration: in order to get a reliable estimate, one needs sufficiently many runs in which the event occurs. This is the reason why one often resorts to simulation using *importance sampling* (or: *change of measure*). This is a variance reduction technique in which one replaces the actual probability measure by an alternative measure under which the event under consideration is *not* rare; correcting the simulation output with appropriate likelihood ratios yields an unbiased estimate.

The crucial issue when setting up an importance sampling procedure concerns the choice of the alternative measure: one would like to select one that provides a substantial variance reduction, or is even (in some sense) optimal. The objective of this paper is to develop a change of measure which performs provably optimally.

Our ultimate goal is to obtain an efficient simulation procedure for Markov-modulated linear stochastic fluid networks. We do so by (i) first considering a single node without

modulation, (ii) then multi-node systems, still without modulation, and (iii) finally modulated multi-node systems. There are two reasons for this step-by-step setup:

- For the non-modulated models we have more refined results than for the modulated models. More specifically, for the non-modulated models we have developed estimates for the number of runs Σ_n required to obtain an estimate with predefined precision (showing that Σ_n grows polynomially in the rarity parameter n), whereas for modulated models we can just prove that Σ_n grows subexponentially.
- In addition, this approach allows the reader to get gradually familiar with the concepts used in this paper.

The construction and analysis of our importance sampling methodology is based on the ideas developed in Blom and Mandjes (2013); there the focus was on addressing similar issues for a single-node Markov modulated infinite-server system. In line with Blom and Mandjes (2013), we consider the regime in which the background process is 'slow': while we (linearly) speed up the driving Poisson process, we leave the rates of the Markovian background process unalterned.

A traditional, thoroughly examined, importance sampling problem concerns the sample mean S_n of n i.i.d. light-tailed random variables $X_1, \ldots, X_n$; the objective there is to estimate $\mathbb{P}(S_n \geqslant a)$ for $a > \mathbb{E}X_1$ and n large. As described in (Asmussen and Glynn 2007, Section VI.2), in this situation importance sampling (i.e., sampling under an alternative measure, and translating the simulation output back by applying appropriate likelihood ratios) works extremely well. To this end, the distribution of the X_i s should be *exponentially twisted*. As it turns out, in our setup, the probability of our interest can be cast in terms of this problem. Compared to the standard setup of sample means of one-dimensional random variables, however, there are a few complications: (i) in our case it is not a priori clear how to sample from the exponentially twisted distributions, (ii) we consider multi-dimensional distributions (i.e., rare-event probabilities that concern the storage levels of all individual buffers in the network), (iii) we impose Markov modulation. We refer to e.g. Glasserman and Juneja (2008) and Kuhn et al. (2017) for earlier work on similar problems.

In passing, we also point out how to set up a recursion to evaluate the (transient and stationary) moments of the joint storage level in Markov-modulated linear stochastic fluid networks (where the results in Kella and Stadje (2002) are restricted to just the first two stationary moments at epochs that the background process jumps).

The single-node model without modulation falls in the class of (one-dimensional) *shot-noise* models, for which efficient rare-event simulation techniques have been developed over the past, say, two decades. Asmussen and Nielsen (1995) and Ganesh et al. (2007) consider the probability that a shot-noise process decreased by a linear drift ever exceeds some given level. Relying on sample-path large deviations results, an asymptotically efficient importance sampling algorithm is developed, under the same scaling as the one we consider in our paper. The major difference with our model (apart from the fact that we deal with considerably more general models, as we focus on networks and allow modulation) is that we focus on a rare-event probability that relates to the position of the process at a fixed point in time; in this setting we succeed in finding accurate estimates of the number of runs needed to get an estimate of given precision.

There is a vast body of literature related to the broader area of rare-event simulation for queueing systems. We refer to the literature overviews (Blanchet and Mandjes 2009; Juneja et al. 2006); interesting recent papers include (Asmussen and Kortschak 2015; Cahen et al. 2017; Sezer 2009).

This paper is organized as follows. In Section 2 the focus is on a single-node network, without Markov modulation (addressing complication (i) above), Section 3 addresses the extension to multi-node systems (addressing complication (ii)), and in Section 4 the feature of modulation is added (addressing complication (iii)). In each of these three sections, we propose a change of measure, quantify its performance, and demonstrate its efficiency through simulation experiments. In Section 4.1 we include the explicit expressions for the moments in Markov-modulated linear stochastic fluid networks. A discussion and concluding remarks are found in Section 5.

2 Single Resource, No Modulation

To introduce the concepts we work with in this paper, we analyze in this section a linear stochastic fluid network consisting of a single node, in which the input is just compound Poisson (so no Markov modulation is imposed). More precisely, in the model considered, jobs arrive according to a Poisson process with rate λ, bring along i.i.d. amounts of work (represented by the sequence of i.i.d. random variables $(B_1, B_2, \ldots)$), and the workload level decays exponentially at a rate $r > 0$. This model belongs to the class of *shot-noise processes*. As mentioned in the introduction, we gradually extend the model in the next sections.

2.1 Preliminaries

We first present a compact representation for the amount of work in the system at time t, which we denote by $X(t)$, through its moment generating function. To this end, let $N(t)$ denote a Poisson random variable with mean λt, and $(U_1, U_2, \ldots)$ i.i.d. uniformly distributed random variables (on the interval $[0, t]$). Assume in addition that the random objects $(B_1, B_2, \ldots)$, $N(t)$, and $(U_1, U_2, \ldots)$ are independent. Then it is well-known that the value of our shot-noise process at time t can be expressed as

$$X(t) = \sum_{j=1}^{N(t)} B_j \mathrm{e}^{-r(t-U_j)} \stackrel{\mathrm{d}}{=} \sum_{j=1}^{N(t)} B_j \mathrm{e}^{-rU_j}, \tag{1}$$

where the distributional equality is a consequence of the fact that the distribution of U is symmetric on the interval $[0, t]$. It is easy to compute the moment generating function (MGF) of $X(t)$, by conditioning on the value of $N(t)$:

$$\begin{aligned} M(\vartheta) := \mathbb{E}\,\mathrm{e}^{\vartheta X(t)} &= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \left(\mathbb{E}\exp(\vartheta B\,\mathrm{e}^{-rU})\right)^k \\ &= \exp\left(\lambda \int_0^t \left(\beta(\mathrm{e}^{-ru}\,\vartheta) - 1\right) \mathrm{d}u\right), \end{aligned} \tag{2}$$

where $\beta(\cdot)$ is the MGF corresponding to B (throughout assumed to exist). By differentiating and inserting $\vartheta = 0$, it follows immediately that

$$\mathbb{E}\,X(t) = \frac{\lambda}{r}(1 - \mathrm{e}^{-rt})\,\mathbb{E}\,B =: m(t).$$

Higher moments can be found by repeated differentiation. We note that, as t is held fixed throughout the document, we often write N rather than $N(t)$.

2.2 Tail Probabilities, Change of Measure

The next objective is to consider the asymptotics of the random variable $X(t)$ under a particular scaling. In this scaling we let the arrival rate be $n\lambda$ rather than just λ, for $n \in \mathbb{N}$. The value of the shot-noise process is now given by

$$Y_n(t) := \sum_{i=1}^{n} X_i(t),$$

with the vector $(X_1(t), \ldots, X_n(t))$ consisting of i.i.d. copies of the random variable $X(t)$ introduced above; here the infinite divisibility of a Compound Poisson distribution is used.

Our goal is to devise techniques to analyze the tail distribution of $Y_n(t)$. Standard theory now provides us with the asymptotics of

$$p_n(a) = \mathbb{P}(Y_n(t) \geqslant na)$$

for some $a > m(t)$; we are in the classical 'Cramér setting' (Dembo and Zeitouni 1998, Section 2.2) if it is assumed that $M(\vartheta)$ is finite in a neighborhood around the origin (which requires that the same property is satisfied by $\beta(\cdot)$). Let $I(a)$ and $\vartheta^\star \equiv \vartheta^\star(a)$, respectively, be defined as

$$I(a) := \sup_{\vartheta} \big(\vartheta a - \log M(\vartheta)\big), \quad \vartheta^\star := \arg\sup_{\vartheta} \big(\vartheta a - \log M(\vartheta)\big),$$

with $M(\cdot)$ as above. Using 'Cramér', we obtain that, under mild conditions,

$$\lim_{n\to\infty} \frac{1}{n} \log p_n(a) = -I(a) = -\vartheta^\star a + \log M(\vartheta^\star).$$

More refined asymptotics are available as well; we get back to this issue in Section 2.3.

As these results apply in the regime that n is large, a relevant issue concerns the development of efficient techniques to estimate $p_n(a)$ through simulation. An important rare-event simulation technique is importance sampling, relying on the commonly used exponential twisting technique. We now investigate how to construct the exponentially twisted version $\mathbb{Q}$ (with twist $\vartheta^\star$) of the original probability measure $\mathbb{P}$. The main idea is that under $\mathbb{Q}$ the $X_i(t)$ have mean a, such that under the new measure the event under study is not rare anymore.

More concretely, exponential twisting with parameter $\vartheta^\star$ means that under the new measure $\mathbb{Q}$, the $X_i(t)$ should have the MGF

$$\mathbb{E}_{\mathbb{Q}}\, \mathrm{e}^{\vartheta X(t)} = \frac{\mathbb{E}\, \mathrm{e}^{(\vartheta+\vartheta^\star)X(t)}}{\mathbb{E}\, \mathrm{e}^{\vartheta^\star X(t)}} = \frac{M(\vartheta + \vartheta^\star)}{M(\vartheta^\star)}; \tag{3}$$

under this choice the random variable has the desired mean:

$$\mathbb{E}_{\mathbb{Q}}\, X(t) = \frac{M'(\vartheta^\star)}{M(\vartheta^\star)} = a.$$

The question is now: how to sample a random variable that has this MGF? To this end, notice that $M(\vartheta) = \exp(-\lambda t + \lambda t\, \mathbb{E} \exp(\vartheta B \mathrm{e}^{-rU}))$ and

$$M(\vartheta + \vartheta^\star) = \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t\, \mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU}))^k}{k!} \left(\frac{\mathbb{E} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU})}{\mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU})} \right)^k,$$

such that Eq. 3 equals

$$\sum_{k=0}^{\infty} \exp(-\lambda t\, \mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU})) \frac{(\lambda t\, \mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU}))^k}{k!} \left(\frac{\mathbb{E} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU})}{\mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU})} \right)^k.$$

From this expression we can see how to sample the $X_i(t)$ under $\mathbb{Q}$, as follows. In the first place we conclude that under $\mathbb{Q}$ the number of arrivals becomes Poisson with mean

$$\lambda t\, \mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU}) = \lambda \int_0^t \beta(\mathrm{e}^{-ru}\, \vartheta^\star) \mathrm{d}u, \tag{4}$$

rather than λt (which is an increase). Likewise, it entails that under $\mathbb{Q}$ the distribution of the $B_j \mathrm{e}^{-rU_j}$ should be twisted by $\vartheta^\star$, in the sense that these random variables should have under $\mathbb{Q}$ the MGF

$$\mathbb{E}_{\mathbb{Q}} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU}) = \frac{\mathbb{E} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU})}{\mathbb{E} \exp(\vartheta^\star B\, \mathrm{e}^{-rU})}.$$

We now point out how such a random variable should be sampled. To this end, observe that

$$\mathbb{E} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU}) = \int_0^t \frac{\beta(\mathrm{e}^{-ru}(\vartheta + \vartheta^\star))}{\beta(\mathrm{e}^{-ru}\, \vartheta^\star)} \frac{1}{t}\, \beta(\mathrm{e}^{-ru}\, \vartheta^\star) \mathrm{d}u,$$

so that

$$\mathbb{E}_{\mathbb{Q}} \exp((\vartheta + \vartheta^\star) B\, \mathrm{e}^{-rU}) = \int_0^t \frac{\beta(\mathrm{e}^{-ru}(\vartheta + \vartheta^\star))}{\beta(\mathrm{e}^{-ru}\, \vartheta^\star)} \frac{\beta(\mathrm{e}^{-ru}\, \vartheta^\star)}{\int_0^t \beta(\mathrm{e}^{-rv}\, \vartheta^\star) \mathrm{d}v} \mathrm{d}u.$$

From this representation two conclusions can be drawn. In the first place, supposing there are k arrivals, then the arrival epochs $U_1, \ldots, U_k$ are i.i.d. under $\mathbb{Q}$, with the density given by

$$f_U^{\mathbb{Q}}(u) = \frac{\beta(\mathrm{e}^{-ru}\, \vartheta^\star)}{\int_0^t \beta(\mathrm{e}^{-rv}\, \vartheta^\star)\, \mathrm{d}v}.$$

In the second place, given that the k-th arrival occurs at time u, the density of the corresponding job size B_k should be exponentially twisted by $\mathrm{e}^{-ru}\, \vartheta^\star$ (where each of the job sizes is sampled independently of everything else).

Now that we know how to sample from $\mathbb{Q}$ it is straightforward to implement the importance sampling. Before we describe its complexity (in terms of the number of runs required to obtain an estimate with given precision), we first provide an example in which we demonstrate how the change of measure can be performed.

Example 1 In this example we consider the case that the B_i are exponentially distributed with mean μ^{-1}. Applying the transformation $w := \mathrm{e}^{-ru}\, \vartheta/\mu$, it is first seen that

$$\begin{aligned} \int_0^s \beta(\mathrm{e}^{-ru}\, \vartheta) \mathrm{d}u &= \int_0^s \frac{\mu}{\mu - \mathrm{e}^{-ru}\, \vartheta} \mathrm{d}u = \frac{1}{r} \int_{\mathrm{e}^{-rs}\vartheta/\mu}^{\vartheta/\mu} \frac{1}{1-w} \frac{1}{w} \mathrm{d}w \\ &= \frac{1}{r} \left[\log \frac{w}{1-w} \right]_{\mathrm{e}^{-rs}\vartheta/\mu}^{\vartheta/\mu} = \frac{1}{r} \log \left(\frac{\mu \mathrm{e}^{rs} - \vartheta}{\mu - \vartheta} \right). \end{aligned}$$

As $\vartheta^\star$ solves the equation $M'(\vartheta^\star)/M(\vartheta^\star) = a$, we obtain the quadratic equation

$$m(t) = a \left(1 - \frac{\vartheta}{\mu} \right) \left(1 - \frac{\vartheta}{\mu} \mathrm{e}^{-rt} \right),$$

leading to

$$\vartheta^\star = \frac{\mu e^{rt}}{2}\left((1+e^{-rt}) - \sqrt{(1-e^{-rt})^2 + 4e^{-rt}\frac{m(t)}{a}}\right)$$

(where it is readily checked that $\vartheta^\star \in (0,\mu)$).

Now we compute what the alternative measure $\mathbb{Q}$ amounts to. In the first place, the number of arrivals should become Poisson with parameter

$$\frac{\lambda}{r}\log\left(\frac{\mu e^{rt}-\vartheta^\star}{\mu-\vartheta^\star}\right)$$

(which is larger than λt). In addition, we can check that

$$F_U^{\mathbb{Q}}(u) := \mathbb{Q}(U \leqslant u) = \log\left(\frac{\mu e^{ru}-\vartheta^\star}{\mu-\vartheta^\star}\right) \Big/ \log\left(\frac{\mu e^{rt}-\vartheta^\star}{\mu-\vartheta^\star}\right)$$

(rather than u/t). The function $F_U^{\mathbb{Q}}(u)$ has the value 0 for $u=0$ and the value 1 for $u=t$, and is concave. This concavity reflects that the arrival epochs of the shots tend to be closer to 0 under $\mathbb{Q}$ than under $\mathbb{P}$. This is because we identified each of the U_i with t minus the actual corresponding arrival epoch in Eq. 1; along the most likely path of $Y_n(t)$ itself the shots will be typically closer to t under $\mathbb{Q}$. Observe that one can sample U under $\mathbb{Q}$ using the classical inverse distribution function method (Asmussen and Glynn 2007, Section II.2a): with H denoting a uniform number on [0, 1), we obtain such a sample by

$$\frac{1}{r}\log\left(\left(e^{rt}-\frac{\vartheta^\star}{\mu}\right)^H\left(1-\frac{\vartheta^\star}{\mu}\right)^{1-H}+\frac{\vartheta^\star}{\mu}\right).$$

Also, conditional on a U_i having attained the value u, the jobs B_i should be sampled from an exponential distribution with mean $(\mu - e^{-ru}\,\vartheta^\star)^{-1}$.

Remark 1 In the model we study in this section, the input of the linear stochastic fluid network is a compound Poisson process. As pointed out in Kella and Whitt (1999) the class of inputs can be extended to the more general class of increasing Lévy processes in a straightforward manner.

2.3 Efficiency Properties of Importance Sampling Procedure

In this subsection we analyze the performance of the procedure introduced in the previous section. The focus is on a characterization of the number of runs needed to obtain an estimate with a given precision (at a given confidence level).

In every run $Y_n(t)$ is sampled under $\mathbb{Q}$, as pointed out above. As $\mathbb{Q}$ is an implementation of an exponential twist (with twist $\vartheta^\star$), the likelihood ratio (of sampling $Y_n(t)$ under $\mathbb{P}$ relative to $\mathbb{Q}$) is given by

$$L = \frac{d\mathbb{P}}{d\mathbb{Q}} = e^{-\vartheta^\star Y_n(t)} e^{n\,\log M(\vartheta^\star)}.$$

In addition, define I as the indicator function of the event $\{Y_n(t) \geqslant na\}$. Clearly, $\mathbb{E}_{\mathbb{Q}}(LI) = p_n(a)$. We keep generating samples LI (under $\mathbb{Q}$), and estimate $p_n(a)$ by the corresponding sample mean, until the ratio of the half-width of the confidence interval (with critical value T) and the estimator drops below some predefined ε (say, 10%). Under $\mathbb{P}$ the number of runs needed is effectively inversely proportional to $p_n(a)$, hence exponentially increasing in n. We now focus on quantifying the reduction of the number of runs when using the importance sampling procedure we described above, i.e., the one based on the measure $\mathbb{Q}$.

Using a Normal approximation, it is a standard reasoning that when performing N runs the ratio of the half-width of the confidence interval and the estimator is approximately

$$\frac{1}{p_n(a)} \cdot \frac{T}{\sqrt{N}} \sqrt{\mathbb{V}\mathrm{ar}_{\mathbb{Q}}(L^2 I)},$$

and hence the number of runs needed is roughly

$$\Sigma_n := \frac{T^2}{\varepsilon^2} \frac{\mathbb{V}\mathrm{ar}_{\mathbb{Q}}(L^2 I)}{(p_n(a))^2}.$$

We now analyze how Σ_n behaves as a function of the 'rarity parameter' n. Due to the Bahadur-Rao result (Bahadur and Rao 1960), with $f_n \sim g_n$ denoting $f_n/g_n \to 1$ as $n \to \infty$,

$$p_n(a) = \mathbb{E}_{\mathbb{Q}}(LI) \sim \frac{1}{\sqrt{n}} \frac{1}{\vartheta^\star \sqrt{2\pi\tau}} \mathrm{e}^{-nI(a)}, \quad \tau := \frac{\mathrm{d}^2}{\mathrm{d}\vartheta^2} \log M(\vartheta) \Big|_{\vartheta=\vartheta^\star}. \tag{5}$$

Using the same proof technique as in Bahadur and Rao (1960), it can be shown that

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \frac{1}{\sqrt{n}} \frac{1}{2\vartheta^\star \sqrt{2\pi\tau}} \mathrm{e}^{-2nI(a)}; \tag{6}$$

see Appendix A for the underlying computation. It also follows that $\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \mathbb{V}\mathrm{ar}_{\mathbb{Q}}(L^2 I)$.

We can use these asymptotics, to conclude that under $\mathbb{Q}$ the number of runs required grows slowly in n. More specifically, Σ_n is essentially proportional to $\sqrt{n}$ for n large. This leads to the following result; cf. (Blanchet et al. 2008, Section 2) for related findings in a more general context.

Proposition 1 *As $n \to \infty$,*

$$\Sigma_n \sim \alpha \sqrt{n}, \quad \alpha = \frac{T^2}{\varepsilon^2} \vartheta^\star \cdot \frac{1}{2} \sqrt{2\pi\tau}. \tag{7}$$

2.4 Simulation Experiments

In this subsection we present numerical results for the single-node model without Markov modulation. We focus on the case of exponential jobs, as in Example 1. We simulate until the estimate has reached the precision $\varepsilon = 0.1$, with confidence level 0.95 (such that the critical value is $T = 1.96$). The parameters chosen are: $t = 1$, $r = 1$, $\lambda = 1$, and $\mu = 1$. We set $a = 1$ (which is larger than $m(t) = 1 - e^{-1}$). As it turns out, $\vartheta^\star = 0.2918$ and

$$\tau = \frac{\lambda}{r} \left(\frac{1}{(\mu - \vartheta^\star)^2} - \frac{1}{(\mu \mathrm{e}^{rt} - \vartheta^\star)^2} \right) = 1.8240.$$

The top-left panel of Fig. 1 confirms the exponential decay of the probability of interest, as a function of n. In the top-right panel we verify that the number of runs indeed grows proportionally to $\sqrt{n}$; the value of α, as defined in Eq. 7, is 198.7, which is depicted by the horizontal line. The bottom-left panel shows the density of the arrival epochs, which confirms that the arrival epochs tend to be closer to 0 under $\mathbb{Q}$ than under $\mathbb{P}$; recall that under $\mathbb{P}$ these epochs are uniformly distributed on $[0, t]$. Recall that we reversed time in Eq. 1: for the actual shot-noise system that we are considering, it means that in order to reach the

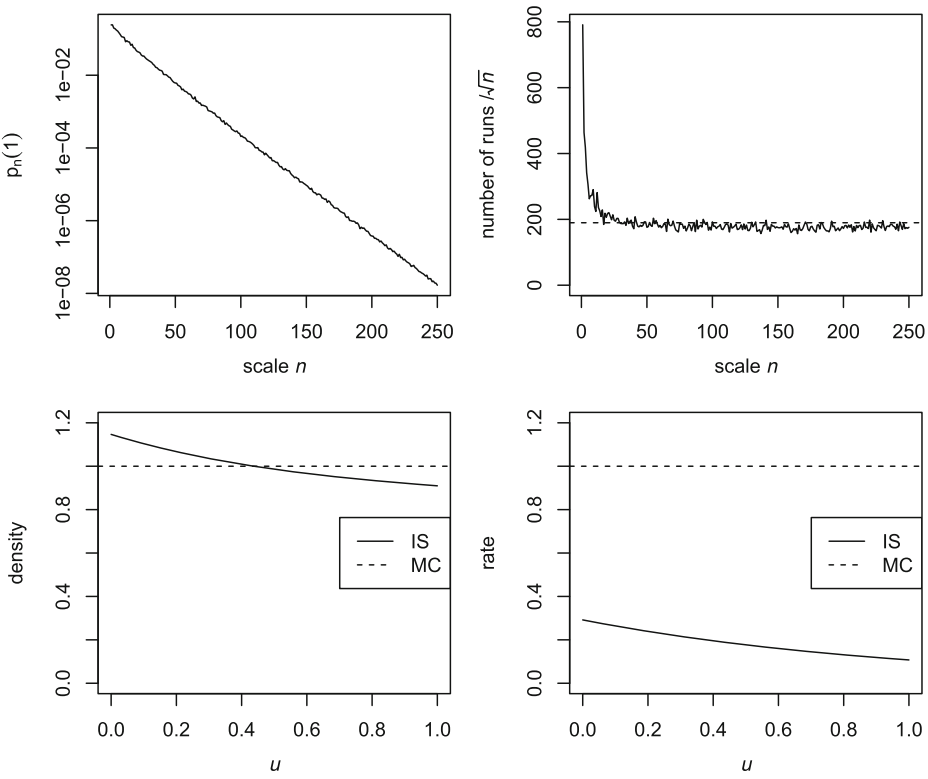

Fig. 1 Numerical results for Section 2.4

desired level at time t, the arrival epochs tend to be closer to t under $\mathbb{Q}$ than under $\mathbb{P}$. The bottom-right panel presents the rate of the exponential job sizes as a function of u. Using (4), the arrival rate under $\mathbb{Q}$ turns out to be 1.2315.

3 Multi-node Systems, No Modulation

In this section we consider multi-node stochastic fluid linear stochastic fluid networks, of the type analyzed in the work by Kella and Whitt (1999). It is instructive to first consider the simplest multi-node system: a tandem network without external input in the downstream node and no traffic leaving after having been served by the upstream node (and rate r_ℓ for node ℓ, $\ell = 1, 2$); later we extend the ideas developed to general linear stochastic fluid networks.

3.1 Preliminaries

As mentioned above, we first consider the two-node tandem. The content of the first node is, as before,

$$X^{(1)}(t) = \sum_{j=1}^{N} B_j \mathrm{e}^{-r_1(t-U_j)}$$

(with N having a Poisson distribution with mean λt), but it can be argued that the content of the second node satisfies a similar representation. More specifically, using the machinery developed in Kella and Whitt (1999), it turns out that

$$X^{(2)}(t) = \sum_{j=1}^{N} B_j \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2(t-U_j)} - \mathrm{e}^{-r_1(t-U_j)} \right) \stackrel{\mathrm{d}}{=} \sum_{j=1}^{N} B_j \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 U_j} - \mathrm{e}^{-r_1 U_j} \right). \tag{8}$$

As before, perform the scaling by n, meaning that the arrival rate λ is inflated by a factor n. It leads to the random vectors $(X_1^{(1)}(t), \ldots, X_n^{(1)}(t))$ and $(X_1^{(2)}(t), \ldots, X_n^{(2)}(t))$. With these vectors we can define $Y_n^{(1)}(t)$ and $Y_n^{(2)}(t)$, analogously to how this was done in the single-node case; these two random quantities represent the contents of the upstream resource and the downstream resource, respectively.

The state of this tandem system can be uniquely characterized in terms of its (bivariate) moment generating function. The technique to derive an explicit expression is by relying on the above distributional equality (8). Again, the key step is to condition on the number of shots that have arrived in the interval $[0, t]$: with $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2)$,

$$\begin{aligned}
M(\boldsymbol{\vartheta}) &:= \mathbb{E}\, \mathrm{e}^{\vartheta_1 X^{(1)}(t) + \vartheta_2 X^{(2)}(t)} \\
&= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \left(\mathbb{E} \exp\left(\vartheta_1 B \mathrm{e}^{-r_1 U} + \vartheta_2 B \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 U} - \mathrm{e}^{-r_1 U} \right) \right) \right)^k \\
&= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \left(\int_0^t \frac{1}{t} \mathbb{E} \exp\left(\vartheta_1 B \mathrm{e}^{-r_1 u} + \vartheta_2 B \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 u} - \mathrm{e}^{-r_1 u} \right) \right) \mathrm{d}u \right)^k \\
&= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \left(\int_0^t \frac{1}{t} \beta \left(\mathrm{e}^{-r_1 u} \vartheta_1 + \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 u} - \mathrm{e}^{-r_1 u} \right) \vartheta_2 \right) \mathrm{d}u \right)^k \\
&= \exp\left(\lambda \int_0^t \left(\beta \left(\mathrm{e}^{-r_1 u} \vartheta_1 + \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 u} - \mathrm{e}^{-r_1 u} \right) \vartheta_2 \right) - 1 \right) \mathrm{d}u \right).
\end{aligned} \tag{9}$$

The above computation is for the two-node tandem system, but the underlying procedure can be extended to the case of networks with more than 2 nodes, and external input in each of the nodes. To this end, we consider the following network consisting of L nodes. Jobs are generated according to a Poisson process. At an arrival epoch, an amount is added to the content of each of the resources $\ell \in \{1, \ldots, L\}$, where the amount added to resource ℓ is distributed as the (non-negative) random variable $B^{(\ell)}$; $\beta(\boldsymbol{\vartheta})$, with $\boldsymbol{\vartheta} \in \mathbb{R}^L$, is the joint MGF of $B^{(1)}$ up to $B^{(L)}$ (note that the components are not assumed independent). In addition, let the traffic level at node ℓ decay exponentially with rate r_ℓ (i.e., the value of the output rate is linear in the current level, with proportionality constant r_ℓ). A deterministic fraction $p_{\ell\ell'} \geqslant 0$ ($\ell \neq \ell'$) is then fed into node ℓ', whereas a fraction $p_{\ell\ell} \geqslant 0$ leaves the network (with $\sum_{\ell'=1}^{L} p_{\ell\ell'} = 1$). We denote $r_{\ell\ell'} := r_\ell p_{\ell\ell'}$. As an aside we mention that this general model covers models in which some arrivals (of the Poisson process with parameter λ) actually lead to arrivals at only a subset of the L queues (since the job sizes $B^{(1)}, \ldots, B^{(L)}$ are allowed to equal 0).

We now point out how the joint buffer content process can be analyzed. Again our objective is to evaluate the moment generating function. Define the matrix R as follows: its (ℓ, ℓ)-th entry is $r_{\ell\ell} + \sum_{\ell' \neq \ell} r_{\ell\ell'}$, whereas its (ℓ, ℓ')-th entry (with $\ell \neq \ell'$) is $-r_{\ell\ell'}$. We have,

according to Kella and Whitt (1999), with N again Poisson with mean λt, the following distributional equality: for any $\ell \in \{1, \ldots, L\}$,

$$X^{(\ell)}(t) = \sum_{\ell'=1}^{L} \sum_{j=1}^{N} B_j^{(\ell')} (\mathrm{e}^{-R(t-U_j)})_{\ell'\ell}.$$

It means we can compute the joint MGF of $X^{(1)}(t)$ up to $X^{(L)}(t)$ as follows, cf. (Kella and Whitt 1999, Thm. 5.1):

$$\begin{aligned}
M(\boldsymbol{\vartheta}) &:= \mathbb{E} \exp\left(\sum_{\ell=1}^{L} \vartheta_\ell X^{(\ell)}(t) \right) \\
&= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \left(\mathbb{E} \exp\left(\sum_{\ell=1}^{L} \vartheta_\ell \sum_{\ell'=1}^{L} B^{(\ell')} (\mathrm{e}^{-R(t-U)})_{\ell'\ell} \right) \right)^k \\
&= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \left(\int_0^t \frac{1}{t} \mathbb{E} \exp\left(\sum_{\ell=1}^{L} \vartheta_\ell \sum_{\ell'=1}^{L} B^{(\ell')} (\mathrm{e}^{-Ru})_{\ell'\ell} \right) \mathrm{d}u \right)^k \\
&= \sum_{k=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^k}{k!} \left(\int_0^t \frac{1}{t} \beta \left(\sum_{\ell=1}^{L} (\mathrm{e}^{-Ru})_{1\ell} \vartheta_\ell, \ldots, \sum_{\ell=1}^{L} (\mathrm{e}^{-Ru})_{L\ell} \vartheta_\ell \right) \mathrm{d}u \right)^k \\
&= \exp\left(-\lambda t + \lambda \int_0^t \beta \left(\sum_{\ell=1}^{L} (\mathrm{e}^{-Ru})_{1\ell} \vartheta_\ell, \ldots, \sum_{\ell=1}^{L} (\mathrm{e}^{-Ru})_{L\ell} \vartheta_\ell \right) \mathrm{d}u \right) \\
&= \exp\left(\lambda \int_0^t \left(\beta \left(\mathrm{e}^{-Ru}\, \boldsymbol{\vartheta} \right) - 1 \right) \mathrm{d}u \right),
\end{aligned}$$

which is the matrix/vector-counterpart of the expression (2) that we found in the single-node case; for the two-node case the special form (9) applies.

3.2 Tail Probabilities, Change of Measure

In this subsection we introduce the change of measure that we use in our importance sampling approach. Many of the concepts are analogous to concepts used for the single-node case in Section 2.

Define (in self-evident notation)

$$p_n(\boldsymbol{a}) := \mathbb{P}\left(Y_n^{(1)}(t) \geqslant n a_1, \ldots, Y_n^{(L)}(t) \geqslant n a_L \right).$$

Due to the multivariate version of Cramér's theorem, with $A := [a_1, \infty) \times \cdots \times [a_L, \infty)$,

$$\lim_{n\to\infty} \frac{1}{n} \log p_n(\boldsymbol{a}) = -\inf_{\boldsymbol{b}\in A} I(\boldsymbol{b}), \quad \text{where} \quad I(\boldsymbol{b}) := \sup_{\boldsymbol{\vartheta}} \left(\langle \boldsymbol{\vartheta}, \boldsymbol{b} \rangle - \log M(\boldsymbol{\vartheta}) \right). \tag{10}$$

More refined asymptotics than the logarithmic asymptotics of Eq. 10 are available as well, but these are not yet relevant in the context of the present subsection; we return to these 'exact asymptotics' in Section 3.3.

We assume that the set A is 'rare', in the sense that

$$\boldsymbol{m}(t) \notin A, \quad \text{with} \quad m_i(t) := \left. \frac{\partial M(\boldsymbol{\vartheta})}{\partial \vartheta_i} \right|_{\boldsymbol{\vartheta}=\boldsymbol{0}}.$$

Let us now construct the importance sampling measure. Let $\boldsymbol{\vartheta}^\star$ be the optimizing $\boldsymbol{\vartheta}$ in the decay rate of $p_n(\boldsymbol{a})$. Mimicking the reasoning we used in the single-node case, the number of arrivals becomes Poisson with mean

$$\lambda \int_0^t \beta \left(\mathrm{e}^{-r_1 u}\, \vartheta_1^\star + \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 u} - \mathrm{e}^{-r_1 u} \right) \vartheta_2^\star \right) \mathrm{d}u$$

(rather than λt). The density of U under $\mathbb{Q}$ becomes

$$f_U^{\mathbb{Q}}(u) = \frac{\beta \left(\mathrm{e}^{-r_1 u} \vartheta_1^\star + \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 u} - \mathrm{e}^{-r_1 u} \right) \vartheta_2^\star \right)}{\int_0^t \beta \left(\mathrm{e}^{-r_1 v}\, \vartheta_1^\star + \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 v} - \mathrm{e}^{-r_1 v} \right) \vartheta_2^\star \right) \mathrm{d}v}.$$

Given a sample from this distribution attains the value u, the distribution of the corresponding random variable B should be twisted by

$$\mathrm{e}^{-r_1 u} \vartheta_1^\star + \frac{r_1}{r_1 - r_2} \left(\mathrm{e}^{-r_2 u} - \mathrm{e}^{-r_1 u} \right) \vartheta_2^\star.$$

Analogously to what we found above in the two-node tandem, we can identify $\mathbb{Q}$ for general linear stochastic fluid networks. We find that under $\mathbb{Q}$ the number of arrivals becomes Poisson with parameter

$$\lambda \int_0^t \beta \left(\mathrm{e}^{-Ru}\, \boldsymbol{\vartheta}^\star \right) \mathrm{d}u.$$

The arrival epochs should be drawn using the density

$$f_U^{\mathbb{Q}}(u) = \frac{\beta \left(\mathrm{e}^{-Ru}\, \boldsymbol{\vartheta}^\star \right)}{\int_0^t \beta \left(\mathrm{e}^{-Rv}\, \boldsymbol{\vartheta}^\star \right) \mathrm{d}v}.$$

Given an arrival at time u, $(B^{(1)}, \ldots, B^{(L)})$ should be exponentially twisted by

$$\left((\mathrm{e}^{-Ru}\, \boldsymbol{\vartheta}^\star)_1, \ldots, (\mathrm{e}^{-Ru}\, \boldsymbol{\vartheta}^\star)_L \right).$$

3.3 Efficiency Properties of Importance Sampling Procedure

We now consider the efficiency properties of the change of measure proposed in the previous subsection. To this end, we first argue that the vector $\boldsymbol{\vartheta}$ generally has some (at least one) strictly positive entries, whereas the other entries equal 0; i.e., there are *no* negative entries. To this end, we first denote by $\boldsymbol{b}^\star$ the 'most likely point' in A:

$$\boldsymbol{b}^\star := \arg \inf_{\boldsymbol{b} \in A} I(\boldsymbol{b}),$$

so that $\boldsymbol{\vartheta}^\star = \boldsymbol{\vartheta}(\boldsymbol{b}^\star)$. It is a standard result from convex optimization that

$$\frac{\partial I(\boldsymbol{b})}{\partial b_i} = \vartheta_i(\boldsymbol{b}). \tag{11}$$

Suppose now that $\vartheta_i(\boldsymbol{b}^\star) < 0$. Increasing the i-th component of the $\boldsymbol{b}^\star$ (while leaving all other components unchanged) would lead to a vector that is still in A, but that by virtue of Eq. 11 corresponds to a lower value of the objective function $I(\cdot)$, thus yielding that $\boldsymbol{b}^\star$ was not the optimizer; we have thus found a contradiction. Similarly, when $\vartheta_i(\boldsymbol{b}^\star) = 0$ we have that $b_i^\star > a_i$ (as otherwise a reduction of the objective function value would be possible, which contradicts $\boldsymbol{b}^\star$ being minimizer).

Now define Θ as the subset of $i \in \{1, \ldots, L\}$ such that $\vartheta_i > 0$, and let $D \in \{1, \ldots, L\}$ the number of elements of Θ. We now argue that the number of runs needed to obtain an estimate of predefined precision scales as $n^{D/2}$. Relying on the results from Chaganthy and Sethuraman (1996) (in particular their Thm. 3.4), it follows that $p_n(\boldsymbol{a})$ behaves (for n large) proportionally to $n^{-D/2} \exp(-nI(\boldsymbol{b}^\star))$; using the same machinery, $\mathbb{E}_{\mathbb{Q}}(L^2 I)$ behaves proportionally to $n^{-D/2} \exp(-2nI(\boldsymbol{b}^\star))$. Mimicking the line of reasoning of Section 2.3, we conclude that the number of runs needed is essentially proportional to $n^{D/2}$. The formal statement is as follows; in Appendix A we comment on the underlying computations.

Proposition 2 *As $n \to \infty$,*

$$\Sigma_n \sim \alpha\, n^{D/2}, \quad \alpha = \frac{T^2}{\varepsilon^2} \left(\prod_{i \in D} \vartheta_i^\star \right) \cdot \frac{1}{2^D} \left(\sqrt{2\pi} \right)^D \sqrt{\tau}, \tag{12}$$

where τ is the determinant of the Hessian of $\log M(\boldsymbol{\vartheta})$ *in* $\boldsymbol{\vartheta}^\star$.

We further illustrate the ideas and intuition behind the qualitative result described in the above proposition by considering the case $L = 2$. It is noted that three cases may arise: (i) $\Theta = \{1, 2\}$, (ii) $\Theta = \{1\}$, (iii) $\Theta = \{2\}$; as case (iii) can be dealt with in the same way as case (ii), we concentrate on the cases (i) and (ii) only. In case (i), where $D = 2$, the necessary condition (Chaganthy and Sethuraman 1996, Eqn. (3.4)) is fulfilled as $\boldsymbol{\vartheta} > 0$ componentwise. As in addition the conditions A–C of (Chaganthy and Sethuraman 1996) are in place, it is concluded that (Chaganthy and Sethuraman 1996, Thm. 3.4) can be applied, leading to $\boldsymbol{b}^\star = \boldsymbol{a}$, and

$$p_n(\boldsymbol{a}) \sim \frac{1}{n} \frac{1}{\vartheta_1^\star \vartheta_2^\star \cdot 2\pi\sqrt{\tau}} \mathrm{e}^{-nI(\boldsymbol{a})},$$

where τ is the determinant of the Hessian of $\log M(\boldsymbol{\vartheta})$ in $\boldsymbol{\vartheta}^\star$. Along the same lines, it can be shown that

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \frac{1}{n} \frac{1}{4\vartheta_1^\star \vartheta_2^\star \cdot 2\pi\sqrt{\tau}} \mathrm{e}^{-2nI(\boldsymbol{a})}.$$

It now follows that Σ_n is roughly linear in n: with ε and T as introduced in Section 2.3,

$$\Sigma_n = \alpha\, n, \quad \alpha := \frac{T^2}{\varepsilon^2} \vartheta_1^\star \vartheta_2^\star \cdot \frac{\pi\sqrt{\tau}}{2}. \tag{13}$$

In case (ii), we do not have that $\boldsymbol{\vartheta} > 0$ componentwise, and hence (Chaganthy and Sethuraman 1996, Thm. 3.4) does not apply; in the above terminology, $D = 1 < 2 = L$. Observe that in this case the exponential decay rate of the event $\{Y_n^{(1)}(t) \geqslant na_1, Y_n^{(2)}(t) < na_2\}$ strictly majorizes that of $\{Y_n^{(1)}(t) \geqslant na_1\}$ (informally: the former event is substantially less likely than the latter). It thus follows that $b_1^\star = a_1$ and $b_2^\star > a_2$, and

$$\begin{aligned} p_n(\boldsymbol{a}) &= \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1\right) - \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1, Y_n^{(2)}(t) < na_2\right) \\ &\sim \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1\right) \sim \frac{1}{\sqrt{n}} \frac{1}{\vartheta_1^\star \sqrt{2\pi\tau}} \mathrm{e}^{-2nI(\boldsymbol{b}^\star)}, \quad \tau := \frac{\mathrm{d}}{\mathrm{d}\vartheta^2} \log M(\vartheta, 0)\Big|_{\vartheta = \vartheta_1^\star}, \end{aligned}$$

and in addition

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \frac{1}{\sqrt{n}} \frac{1}{2\vartheta_1^\star \sqrt{2\pi\tau}} \mathrm{e}^{-2nI(\boldsymbol{b}^\star)}.$$

As a consequence in this regime Σ_n grows essentially proportional to $\sqrt{n}$ for n large:

$$\Sigma_n \sim \alpha \sqrt{n}, \quad \alpha := \frac{T^2}{\varepsilon^2} \vartheta_1^\star \cdot \frac{1}{2} \sqrt{2\pi\tau}.$$

In case (iii) Σ_n behaves proportionally to $\sqrt{n}$ as well.

3.4 Simulation Experiments

We conclude this section by providing a few numerical illustrations. In the first set we focus on the downstream queue only (i.e., we analyze $p_n(0, a_2)$), whereas in the second set we consider the joint exceedance probability $p_n(\boldsymbol{a})$. The precision and confidence have been chosen as in Example 1. Throughout we take $t = 1$, $r_1 = 2$, $r_2 = 1$, $\lambda = 1$, and $\mu = 1$.

In the first set of experiments we take $a_1 = 0$ and $a_2 = 1$. Elementary numerical analysis yields that $\vartheta^\star = 0.8104$ and $\tau = 1.4774$, leading to α, as defined in Eq. 13, equalling 474.3. For graphs on the behavior of $p_n(1)$ as a function of n and the number of runs needed, we refer to (Boxma et al. 2018, Fig. 2). The two panels of Fig. 2 should be interpreted as the bottom panels of Fig. 1. Interestingly, the left panel indicates that in the tandem system it does not pay off to let jobs arrive right before t (as they first have to go through the first resource to end up in the second resource), as reflected by the shape of the density of the arrival epochs under $\mathbb{Q}$; to this end, recall that we reversed time in Eq. 8, so that a low density at $u = 0$ in the graph corresponds to a high density at $u = t$ in the actual system. The arrival rate under $\mathbb{Q}$ is 1.5103.

In the second set of experiments we take $a_1 = 1.2$ and $a_2 = 1.1$; all other parameters are the same as in the first set. As mentioned above, we now consider the joint exceedance probability. As it turns out, $\vartheta_1^\star = 0.1367$ and $\vartheta_2^\star = 0.2225$. For graphs describing the behavior of $p_n(1.2, 1.1)$ as a function of n and the number of runs needed, we refer to (Boxma et al. 2018, Fig. 3); the latter graph reveals that for this specific parameter setting Σ_n/n converges to the limiting constant rather slowly. Concerning the left panel of Fig. 3, note that in Section 2 we saw that to make sure the first queue gets large it helps to have arrivals at the end of the interval, whereas above we observed that to make the second queue large arrivals should occur relatively early. We now focus on the event that *both* queues are large, and consequently the arrival distribution becomes relatively uniform again, as shown in the left panel of Fig. 3. The arrival rate under $\mathbb{Q}$ is 2.3478.

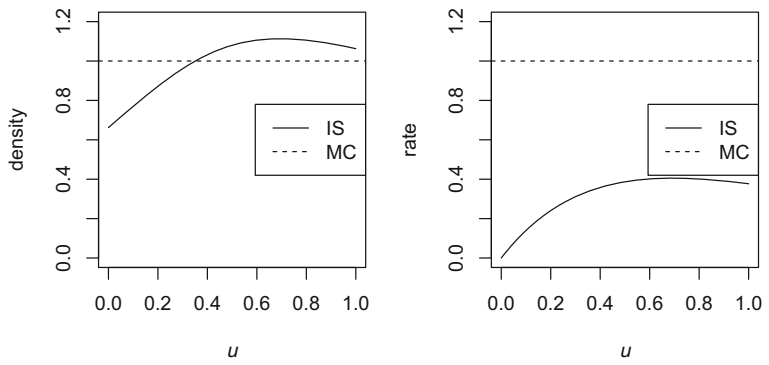

Fig. 2 Numerical results for Section 3.4: downstream queue only

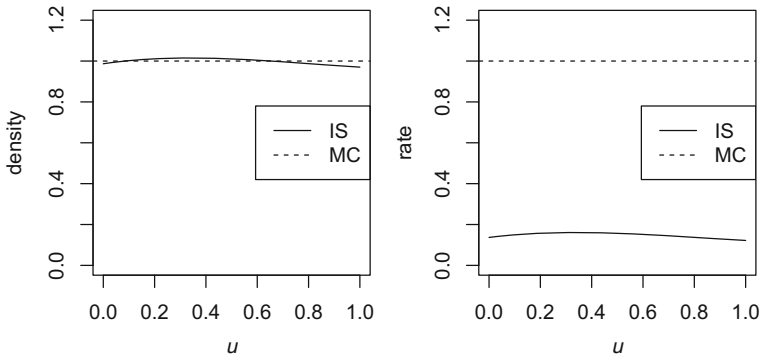

Fig. 3 Numerical results for Section 3.4: both queues

4 Multi-node Systems Under Markov Modulation

In this section consider the networks analyzed in the previous section, but now in a random environment. More specifically, the type of random environment we focus on here is known as *Markov modulation*: the system dynamics are affected by the state of an external finite-state irreducible Markov process $J(\cdot)$ with generator matrix $Q = (q_{jj'})_{j,j'=1}^{d}$. When this Markov process (usually referred to as the *background process*) is in state j, arrivals occur according to a Poisson process with rate λ_j, the MGF of the job size is $\beta_j(\boldsymbol{\vartheta})$, and the routing matrix is R_j. Analogously to the definitions used in the case without Markov modulation, this routing matrix' (i,i)-th entry is

$$(R_j)_{ii} := r_{ii}^{(j)} + \sum_{i' \neq i} r_{ii'}^{(j)},$$

which can be interpreted as the rate at which fluid leaves server i when the background process is in j. Likewise, for $i \neq i'$,

$$(R_j)_{ii'} := -r_{ii'}^{(j)},$$

which is the rate at which fluid flows from server i to i' when the background process is in j.

Below we assume that $J(0) = j_0$ for a fixed state $j_0 \in \{1, \ldots, d\}$; it is seen that all results generalize to an arbitrary initial distribution in a straightforward manner.

The structure of the section is as follows: we consecutively describe general results for the model under consideration (extending earlier stationary results from Kella and Stadje (2002) to their transient counterpart), propose an importance sampling measure, establish efficiency properties of the corresponding estimator, and present a number of numerical experiments.

Note that the setup of this section slightly differs from that of the previous sections. For the models covered in Sections 2 and 3, already detailed explicit analysis is available; see e.g. the results in terms of transforms and moments in Kella and Whitt (1999). Such a complete analysis is lacking for the model featuring in the present section. With the results of our paper added to the literature, the situation has become 'uniform': for all three setups (i.e., Sections 2, 3, and 4), one has results on transient transforms, transient moments, as well as recipes for efficient rare-event simulation.

4.1 Exact Expressions for Moments

Before focusing on simulation-based techniques, this subsection (which can be read independently of the rest of the section) shows that various moment-related quantities can be computed in closed form.

Multi-node systems under Markov modulation have been studied in detail by Kella and Stadje (2002). We start this subsection by providing a compact derivation of a PDE characterizing the system's transient behavior, which was not included in that paper. To this end, we define, for $j \in \{1, \ldots, d\}$,

$$\Xi_j(\boldsymbol{\vartheta}, t) := \mathbb{E}\left(\exp\left(\sum_{\ell=1}^{L} \vartheta_\ell X^{(\ell)}(t)\right) 1_j(t)\right),$$

with $1_j(t)$ the indicator function of the event that $J(t) = j$. Using the standard 'Markov machinery', $\Xi_j(\boldsymbol{\vartheta}, t + \Delta t)$ equals (up to $o(\Delta t)$ terms) the sum of a contribution

$$\lambda_j \, \Delta t \, \Xi_j(\boldsymbol{\vartheta}, t)\beta_j(\boldsymbol{\vartheta})$$

due to the scenario that an arrival occurs between t and $t + \Delta t$, a contribution

$$\sum_{j' \neq j} q_{j'j} \, \Delta t \, \Xi_{j'}(\boldsymbol{\vartheta}, t)$$

due to the scenario that the modulating Markov process jumps between t and $t + \Delta t$, and a contribution

$$\big(1 - \lambda_j \, \Delta t - q_j \, \Delta t\big) \, \mathbb{E}\left(\exp\left(\sum_{\ell=1}^{L}\left(\vartheta_\ell - \sum_{\ell'=1}^{L} \vartheta_{\ell'} (R_j)_{\ell\ell'} \, \Delta t\right) X^{(\ell)}(t)\right) 1_j(t)\right),$$

with $q_j := -q_{jj}$; regarding the last term, observe that when the background process is in state j, and no new job arrives between t and $t + \Delta t$,

$$X^{(\ell)}(t + \Delta t) = X^{(\ell)}(t) - (R_j)_{\ell\ell} \, \Delta t \, X^{(\ell)}(t) - \sum_{\ell' \neq \ell} (R_j)_{\ell'\ell} \, \Delta t \, X^{(\ell')}(t).$$

We thus find that

$$\begin{aligned}\Xi_j(\boldsymbol{\vartheta}, t + \Delta t) \; = \; & \lambda_j \, \Delta t \, \beta_j(\boldsymbol{\vartheta}) \, \Xi_j(\boldsymbol{\vartheta}, t) + \sum_{j' \neq j} q_{j'j} \, \Delta t \, \Xi_{j'}(\boldsymbol{\vartheta}, t) + \\ & \big(1 - \lambda_j \, \Delta t - q_j \, \Delta t\big) \; \Xi_j\left(\boldsymbol{\vartheta} - R_j \boldsymbol{\vartheta} \, \Delta t, t\right) + o(\Delta t).\end{aligned}$$

This immediately leads to (by subsequently subtracting $\Xi_j(\boldsymbol{\vartheta}, t)$ from both sides, dividing by Δt, and letting $\Delta t \downarrow 0$)

$$\frac{\partial}{\partial t} \Xi_j(\boldsymbol{\vartheta}, t) = \lambda_j\big(\beta_j(\boldsymbol{\vartheta}) - 1\big)\Xi_j(\boldsymbol{\vartheta}, t) + \sum_{j'=1}^{d} q_{j'j} \, \Xi_{j'}(\boldsymbol{\vartheta}, t) - \sum_{\ell'=1}^{L} \big(R_j \, \boldsymbol{\vartheta}\big)_{\ell'} \frac{\partial}{\partial \vartheta_{\ell'}} \Xi_j(\boldsymbol{\vartheta}, t). \tag{14}$$

Let us now compactly summarize the relation (14), in vector/matrix notation. This notation will prove practical when computing higher moments; in other (but related) contexts, similar procedures have been proposed in e.g. Huang et al. (2016) and Rabehasaina (2006). Let $\mathscr{M}^{n_1 \times n_2}$ be the set of $\mathbb{R}$-valued matrices of dimension $n_1 \times n_2$ (for generic $n_1, n_2 \in \mathbb{N}$).

In addition, I_n is the identity matrix of dimension $n \in \mathbb{N}$. We introduce the following three matrices in $\mathscr{M}^{d\times d}$, $\mathscr{M}^{d\times d}$, and $\mathscr{M}^{Ld\times Ld}$, respectively:

$$\Lambda := \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_d \end{pmatrix}, \quad B(\boldsymbol{\vartheta}) := \begin{pmatrix} \beta_1(\boldsymbol{\vartheta}) & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \beta_d(\boldsymbol{\vartheta}) \end{pmatrix}, \quad R := \begin{pmatrix} R_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & R_d \end{pmatrix}.$$

We use the conventional notation $\otimes$ for the Kronecker product. Recall that the Kronecker product is bilinear, associative and distributive with respect to addition; these properties we will use in the sequel without mentioning. It also satisfies the mixed product property $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. Furthermore, note that $I_{n_1} \otimes I_{n_2} = I_{n_1 n_2}$.

We now consider some differentiation rules for matrix-valued functions which will allow us to iteratively evaluate moments. In the first place we define the operator $\nabla_{\boldsymbol{\vartheta}}$ for $\boldsymbol{\vartheta} \in \mathbb{R}^L$; to keep notation compact, we often suppress the subscript $\boldsymbol{\vartheta}$, and write just ∇. Let $f \equiv f(\boldsymbol{\vartheta})$ be a mapping of $\mathbb{R}^L$ to $\mathscr{M}^{n_1\times n_2}$. Then $\nabla f \equiv \nabla f(\boldsymbol{\vartheta}) \in \mathscr{M}^{n_1 L\times n_2}$ is defined by

$$\nabla f = \begin{pmatrix} \nabla f_{11} & \nabla f_{12} & \cdots & \nabla f_{1n_2} \\ \nabla f_{21} & \nabla f_{22} & \cdots & \nabla f_{2n_2} \\ \vdots & \vdots & \ddots & \vdots \\ \nabla f_{n_1 1} & \nabla f_{n_1 2} & \cdots & \nabla f_{n_1 n_2} \end{pmatrix}, \quad \text{where} \quad \nabla f_{ij} := \begin{pmatrix} \partial_1 f_{ij} \\ \partial_2 f_{ij} \\ \vdots \\ \partial_L f_{ij} \end{pmatrix}.$$

In the above definition $\nabla f_{ij} \equiv \nabla f_{ij}(\boldsymbol{\vartheta})$ is to be understood as the usual gradient; the symbol ∂_i is used to denote the partial derivative with respect to the i-th variable, in the sense of

$$\partial_i f_{ij} := \frac{\partial}{\partial \vartheta_i} f_{ij}(\boldsymbol{\vartheta}).$$

Furthermore, we define inductively $\nabla^k f \equiv \nabla^k f(\boldsymbol{\vartheta}) := \nabla(\nabla^{k-1} f)$, $k \in \mathbb{N}$, with $\nabla^0 f := f$. It is checked that $\nabla^k f(\boldsymbol{\vartheta})$ is a mapping of $\mathbb{R}^L$ to $\mathscr{M}^{L^k n_1\times n_2}$.

In the sequel we use a couple of differentiation rules, that we have listed below. Let $A(\cdot)$ be a matrix-valued function from $\mathbb{R}^L$ to $\mathscr{M}^{n_1\times n_2}$, and $B(\cdot)$ a matrix-valued function from $\mathbb{R}^L$ to $\mathscr{M}^{n_2\times n_3}$, and let I_q be a $q \times q$ identity matrix (for some $q \in \mathbb{N}$). Then,

– *Product rule*:

$$\nabla_{\boldsymbol{\vartheta}} \big(A(\boldsymbol{\vartheta})\, B(\boldsymbol{\vartheta})\big) = (\nabla_{\boldsymbol{\vartheta}} A(\boldsymbol{\vartheta}))\, B(\boldsymbol{\vartheta}) + (A(\boldsymbol{\vartheta}) \otimes I_L)\, \nabla_{\boldsymbol{\vartheta}} B(\boldsymbol{\vartheta});$$

being an element of $\mathscr{M}^{Ln_1\times n_3}$.

– *Differentiation of Kronecker product (1)*:

$$\nabla_{\boldsymbol{\vartheta}} (I_q \otimes A(\boldsymbol{\vartheta})) = I_q \otimes (\nabla_{\boldsymbol{\vartheta}} A(\boldsymbol{\vartheta})).$$

– *Differentiation of Kronecker product (2)*:

$$\begin{aligned} \nabla_{\boldsymbol{\vartheta}} (A(\boldsymbol{\vartheta}) \otimes I_q) &= (K_{n_1,q} \otimes I_L)(I_q \otimes (\nabla_{\boldsymbol{\vartheta}} A(\boldsymbol{\vartheta})))K_{q,n_2} \\ &= (K_{n_1,q} \otimes I_L)K_{q,n_2}(\nabla_{\boldsymbol{\vartheta}} A(\boldsymbol{\vartheta}) \otimes I_q), \end{aligned}$$

where $K_{m,n}$ is the commutation matrix defined by

$$K_{m,n} := \sum_{i=1}^{m} \sum_{j=1}^{n} (H_{ij} \otimes H_{ij}^{\mathrm{T}}),$$

and $H_{ij} \in \mathscr{M}^{m\times n}$ denotes a matrix with a 1 at its (i, j)-th position and zeros elsewhere, cf. Magnus and Neudecker (1979).

The first rule can be checked componentwise and the second rule is trivial. The third rule follows from the first and second rule in combination with the fact that the Kronecker product commutes after a correction with the commutation matrices. Moreover, we use the property $K_{m,n}^{-1} = K_{n,m}$. An overview of the properties of commutation matrices can be found in Magnus and Neudecker (1979).

In the introduced terminology, it follows that Eq. 14 can be written as

$$\frac{\partial}{\partial t}\,\Xi(\vartheta,t) = \Lambda\big(B(\vartheta) - I_d\big)\,\Xi(\vartheta,t) + Q^{\mathrm{T}}\,\Xi(\vartheta,t) - \big(I_d \otimes \vartheta^{\mathrm{T}}\big)R^{\mathrm{T}}\,\nabla_\vartheta\,\Xi(\vartheta,t). \tag{15}$$

We now point out how (transient and stationary) moments can be evaluated; note that Kella and Stadje (2002) focuses on the first two stationary moments at epochs that the background process jumps. We throughout use the notation $z_i(t)$ for the i-th derivative of $\Xi(\vartheta,t)$ in $(\mathbf{0},t)$, for $t \geqslant 0$:

$$z_i(t) := \nabla_\vartheta^i\,\Xi(\vartheta,t)\Big|_{\vartheta=\mathbf{0}} \in \mathscr{M}^{L^i d\times d},$$

for $i \in \mathbb{N}$. Note that, with $\pi_j(t) = (\exp(Qt))_{j_0,j}$,

$$\Xi(\vartheta,0) = \boldsymbol{e}_{j_0}, \quad \Xi(\mathbf{0},t) = \boldsymbol{\pi}(t)^{\mathrm{T}} \equiv (\pi_1(t),\ldots,\pi_d(t)).$$

◦ We start by characterizing the first moments. Applying the operator $\nabla \equiv \nabla_\vartheta$ to the differential equation (15) yields

$$\begin{aligned}\nabla_\vartheta\left(\frac{\partial}{\partial t}\,\Xi(\vartheta,t)\right) = {}& (\Lambda\otimes I_L)(\nabla_\vartheta B(\vartheta))\,\Xi(\vartheta,t) + \\ & \Big(Q^{\mathrm{T}}\otimes I_L + \Lambda(B(\vartheta)-I_d)\otimes I_L - R^{\mathrm{T}}\Big)\,\nabla_\vartheta\,\Xi(\vartheta,t) - \\ & \Big(((I_d\otimes\vartheta^{\mathrm{T}})R^{\mathrm{T}})\otimes I_L\Big)\,\nabla_\vartheta^2\,\Xi(\vartheta,t),\end{aligned} \tag{16}$$

using standard properties of the Kronecker product in combination with

$$\nabla_\vartheta(I_d\otimes\vartheta^{\mathrm{T}}) = I_d\otimes(\nabla_\vartheta\vartheta^{\mathrm{T}}) = I_d\otimes(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_L) = I_d\otimes I_L = I_{dL},$$

where $\boldsymbol{e}_i$ denotes the L-dimensional column vector in which component i equals 1 and all other components are 0. Then, inserting $\vartheta = \mathbf{0}$ into Eq. 16 yields the system of (non-homogeneous) linear differential equations

$$z_1'(t) = (\Lambda\otimes I_L)\,\nabla B(\mathbf{0})\,\boldsymbol{\pi}(t) + \big((Q^{\mathrm{T}}\otimes I_L) - R^{\mathrm{T}}\big)\,z_1(t). \tag{17}$$

In the stationary case, we obtain

$$z_1(\infty) = \big(R^{\mathrm{T}} - (Q^{\mathrm{T}}\otimes I_L)\big)^{-1}(\Lambda\otimes I_L)\,\nabla B(\mathbf{0})\,\boldsymbol{\pi}, \tag{18}$$

with $\boldsymbol{\pi} := \lim_{t\to\infty}\boldsymbol{\pi}(t)$ (being the unique non-negative solution of $\boldsymbol{\pi}^{\mathrm{T}}Q = \mathbf{0}^{\mathrm{T}}$ such that the entries of $\boldsymbol{\pi}$ sum to 1).

◦ We now move to second moments. Applying the ∇_ϑ-operator to Eq. 16,

$$\begin{aligned}\nabla_\vartheta^2\left(\frac{\partial}{\partial t}\,\Xi(\vartheta,t)\right) = {}& (\Lambda\otimes I_{L^2})(\nabla_\vartheta^2 B(\vartheta))\,\Xi(\vartheta,t) + \\ & \big(((\Lambda\otimes I_L)\nabla_\vartheta B(\vartheta))\otimes I_L\big)\nabla_\vartheta\,\Xi(\vartheta,t) + \\ & \nabla_\vartheta(\Lambda B(\vartheta)\otimes I_L)\nabla_\vartheta\,\Xi(\vartheta,t) + \\ & \big(Q^{\mathrm{T}}\otimes I_{L^2} + \Lambda(B(\vartheta)-I_d)\otimes I_{L^2} - R^{\mathrm{T}}\otimes I_L\big)\,\nabla_\vartheta^2\,\Xi(\vartheta,t) - \\ & (((I_d\otimes\vartheta^{\mathrm{T}})R^{\mathrm{T}})\otimes I_{L^2})\,\nabla_\vartheta^3\,\Xi(\vartheta,t) - \\ & \nabla_\vartheta(((I_d\otimes\vartheta^{\mathrm{T}})R^{\mathrm{T}})\otimes I_L)\,\nabla_\vartheta^2\,\Xi(\vartheta,t),\end{aligned}$$

in which the factor $\nabla_{\boldsymbol{\vartheta}}(\Lambda B(\boldsymbol{\vartheta}) \otimes I_L)$ can be expressed more explicitly as

$$(K_{d,L} \otimes I_L) K_{L,dL}(((\Lambda \otimes I_L)\nabla_{\boldsymbol{\vartheta}} B(\boldsymbol{\vartheta})) \otimes I_L),$$

and the factor $\nabla_{\boldsymbol{\vartheta}}(((I_d \otimes \boldsymbol{\vartheta}^{\mathrm{T}})R^{\mathrm{T}}) \otimes I_L)$ simplifies to $(K_{d,L} \otimes I_L)K_{L,dL}(R^{\mathrm{T}} \otimes I_L)$. Inserting $\boldsymbol{\vartheta} = 0$ yields the system of linear differential equations

$$\begin{aligned} z_2'(t) \;=\; & (\Lambda \otimes I_{L^2})\,(\nabla^2 B(\mathbf{0}))\,\boldsymbol{\pi}(t) + \\ & (Q^{\mathrm{T}} \otimes I_{L^2} - ((K_{d,L} \otimes I_L)K_{L,dL} + I_{dL^2})(R^{\mathrm{T}} \otimes I_L))\, z_2(t) + \\ & \big(((\Lambda \otimes I_L)(\nabla B(\mathbf{0}))) \otimes I_L\big) z_1(t) + \\ & (K_{d,L} \otimes I_L)K_{L,dL}(((\Lambda \otimes I_L)\nabla B(\mathbf{0})) \otimes I_L) z_1(t) \end{aligned}$$

where $z_1(t)$ solves (17). As before, the stationary quantities can be easily derived (by equating $z_2'(t)$ to 0). One has to keep in mind, however, that some of the mixed partial derivatives occur multiple times in z_k, for $k \in \{2, 3, \ldots\}$, and therefore the solution will only be unique after removing the corresponding redundant rows. Alternatively, the system can be completed by including equations which state that these mixed partial derivatives are equal.

∘ It now follows that higher moments can be found recursively, using the three differentiation rules that we stated above.

Remark 2 Various variants of our model can be dealt with similarly. In this remark we consider the slightly adapted model in which shots only occur simultaneously with a jump in the modulating Markov chain. Then (up to $o(\Delta t)$ terms) $\Xi_j(\boldsymbol{\vartheta}, t + \Delta t)$ is the sum of a contribution

$$\sum_{j' \neq j} q_{j'j} \Delta t \; \Xi_{j'}(\boldsymbol{\vartheta}, t)\beta_j(\boldsymbol{\vartheta})$$

due to the scenario that there is a jump in the modulating chain in the interval $[t, t + \Delta t]$ (which also induces a shot), and a contribution of

$$(1 - q_j \Delta t)\; \mathbb{E}\left(\exp\Big(\sum_{\ell=1}^{L} \Big(\vartheta_\ell - \sum_{\ell'=1}^{d} \vartheta_{\ell'} (R_j)_{\ell\ell'} \Delta t \Big) X^{(\ell)}(t) \Big) 1_j(t) \right),$$

with $q_j := -q_{jj}$, in the scenario that there is no jump. Performing the same steps as above, we obtain

$$\frac{\partial}{\partial t}\Xi_j(\boldsymbol{\vartheta}, t) = q_j(\beta_j(\boldsymbol{\vartheta}) - 1)\Xi_j(\boldsymbol{\vartheta}, t) + \sum_{j'=1}^{d} q_{j'j}\Xi_{j'}(\boldsymbol{\vartheta}, t)\beta_j(\boldsymbol{\vartheta}) - \sum_{j'=1}^{L} (R_j \boldsymbol{\vartheta})_{j'} \frac{\partial}{\partial \vartheta_{j'}} \Xi_j(\boldsymbol{\vartheta}, t),$$

which has a similar structure as Eq. 14. It follows that the moments can be found as before. With $\widetilde{Q} := \mathrm{diag}\{q_1, \ldots, q_d\}$, it turns out that the transient means are given by

$$z_1'(t) = \nabla B(\mathbf{0})(Q^{\mathrm{T}} + \widetilde{Q})\boldsymbol{\pi}(t) + \big((Q^{\mathrm{T}} \otimes I_L) - R^{\mathrm{T}}\big) z_1(t).$$

In particular, the stationary first moment equals

$$z_1(\infty) = \big(R^{\mathrm{T}} - (Q^{\mathrm{T}} \otimes I_L)\big)^{-1} \nabla B(\mathbf{0})(Q^{\mathrm{T}} + \widetilde{Q})\boldsymbol{\pi}.$$

Consider the following numerical example for the computation of the expected values and variances, in which the technique described above is illustrated.

Example 2 In this example, we choose the parameters in such a way that we see non-monotonic behavior. Our example corresponds to a case in which the system does not start

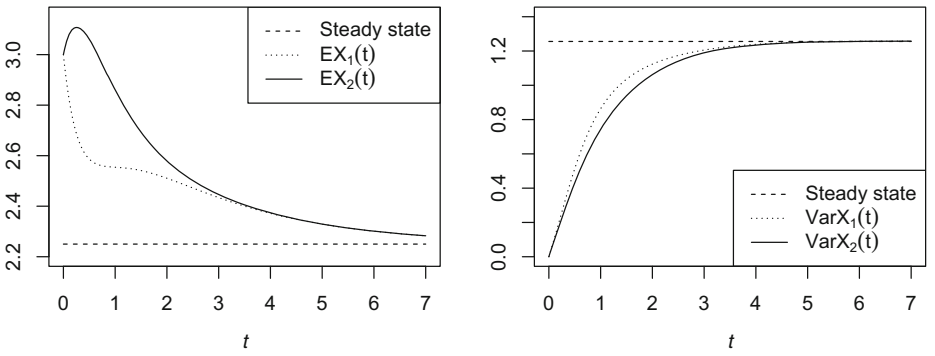

Fig. 4 Transient expected values and variances of Example 2

empty, which is dealt with by imposing suitable starting conditions. We consider a two-dimensional ($L = 2$) queueing system, with a two-dimensional state space of the Markov modulating process ($d = 2$). We pick $q_{12} = q_{21} = 1$, $\lambda_1 = \lambda_2 = 1$, $\mathbb{E}\, B_1 = \mathbb{E}\, B_2 = \mathbb{E}\, B_1^2 = \mathbb{E}\, B_2^2 = 1$, $J(0) = 1$, $(X_1(0), X_2(0)) = (3, 3)$, and the rate matrices

$$R_1 = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}, \qquad R_2 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}.$$

Due to the symmetry in the choice of the parameters, one can expect that for both states of the background process expected value tends (as t grows large) to the same steady-state value; the same is anticipated for the stationary variance. This is confirmed by Fig. 4. For t small, the two queues behave differently due to $J(0) = 1$, which implies that queue 1 drains faster. Note that $\mathbb{E}\, X_2(t)$ even increases for t small, due to the fact that the flow from node 1 to 2 equals the flow from 2 to 1, constituting a net flow of zero, so that the additional contribution of external output to node 2 leads to a net increase of $\mathbb{E}\, X_2(t)$. The transient correlation is plotted in Fig. 5. At time $t = 0$ the queues are perfectly correlated, since the starting state is deterministic. Then the correlation decreases due to the asymmetric flow rates until around $t = 1$, which is when the Markov chain J is expected to switch, after which the correlation monotonously tends to the steady state.

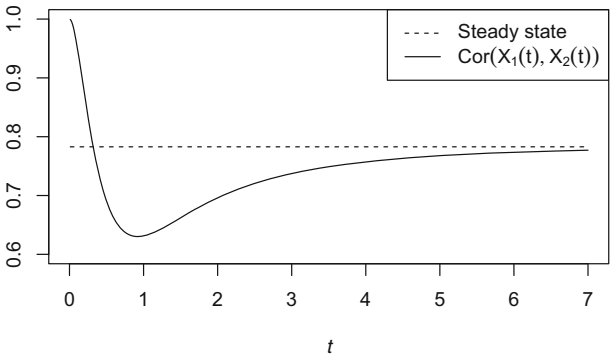

Fig. 5 Transient correlation between $X_1(t)$, $X_2(t)$ of Example 2

4.2 Tail Probabilities, Change of Measure

We now characterize the decay rate of the rare-event probability under study, and we propose a change of measure to efficiently estimate it. In the notation we have been using so far, we again focus on

$$p_n(\boldsymbol{a}) := \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1, \ldots, Y_n^{(L)}(t) \geqslant na_L\right) = \mathbb{P}\left(\boldsymbol{Y}_n(t) \in A\right),$$

where $\boldsymbol{Y}_n(t) = (Y_n^{(1)}(t), \ldots, Y_n^{(L)}(t))$. It is stressed that, following (Blom and Mandjes 2013), we consider the regime in which the background process is 'slow'. In concrete terms, this means that we linearly speed up the driving Poisson process (i.e., we replace the arrival rates λ_j by $n\lambda_j$), but leave the rates of the Markovian background process unaltered.

First we find an alternative characterization of the state of the system at time t. Let $\mathscr{F}_t$ denote the set of all functions from $[0, t]$ onto the states $\{1, \ldots, d\}$. Consider a path $f \in \mathscr{F}_t$. Let f have $K(f)$ jumps between 0 and t, whose epochs we denote by $t_1(f)$ up to $t_{K(f)}(f)$ (and in addition $t_0(f) := 0$ and $t_{K(f)+1}(f) := t$). Let

$$j_i(f) := \lim_{t \downarrow t_i(f)} f(t)$$

(i.e., the state of f immediately after the i-th jump). We also introduce

$$D_i(u, f) := \exp\left(-(t_{i+1}(f) - u)\, R_{j_i(f)}\right), \quad D_i(f) := \exp\left(-(t_{i+1}(f) - t_i(f))\, R_{j_i(f)}\right).$$

Suppose now that the Markov process $J(\cdot)$ follows the path $f \in \mathscr{F}_t$. Then the contribution to the MGF of $\boldsymbol{X}(t)$ due to shots that arrived between $t_i(f)$ and $t_{i+1}(f)$ is, mimicking the arguments that we used in Section 3.2 for non-modulated networks,

$$\psi_i(f, \boldsymbol{\vartheta}) := \exp\left(\lambda_{j_i(f)} \int_{t_i(f)}^{t_{i+1}(f)} \left(\beta_{j_i(f)}\left(D_i(u, f) D_{i+1}(f) \cdots D_{K(f)}(f)\, \boldsymbol{\vartheta}\right) - 1\right) \mathrm{d}u\right).$$

As a consequence, the MGF of $X(t)$ given the path f is

$$M_f(\boldsymbol{\vartheta}) := \prod_{i=0}^{K(f)} \psi_i(f, \boldsymbol{\vartheta}).$$

First conditioning on the path of $J(\cdot) \in \mathscr{F}_t$ between 0 and t and then unconditioning, it then immediately follows that the MGF of $\boldsymbol{X}(t)$ is given by

$$M(\boldsymbol{\vartheta}) = \mathbb{E}\, M_J(\boldsymbol{\vartheta}).$$

Then, precisely as is shown in Blom and Mandjes (2013) for a related stochastic system, the decay rate can be characterized as follows:

$$\lim_{n \to \infty} \frac{1}{n} \log p_n(\boldsymbol{a}) = -\inf_{f \in \mathscr{F}_t} \mathbb{I}_f(\boldsymbol{a}), \quad \mathbb{I}_f(\boldsymbol{a}) := \inf_{\boldsymbol{b} \in A} \sup_{\boldsymbol{\vartheta}} \left(\langle \boldsymbol{\vartheta}, \boldsymbol{b} \rangle - \log M_f(\boldsymbol{\vartheta})\right). \tag{19}$$

The argumentation to show this is analogous to the one in (Blom and Mandjes 2013, Thm. 1), and can be summarized as follows. In the first place, let $f^\star$ be the optimizing path in Eq. 19. Then, as $J(\cdot)$ does not depend on n, we can choose a 'ball' $\mathscr{B}_t(f^\star)$ around $f^\star$ such that the decay rate of the probability of $J(\cdot)$ being in that ball is 0. The lower bound follows by only taking into account the contribution due to paths in $\mathscr{B}_t(f^\star)$. The upper bound follows by showing that the contribution of all $f \in \mathscr{F}_t \setminus \mathscr{B}_t(f^\star)$ is negligible.

Informally, the path $f^\star$ has the interpretation of the most likely path of $J(\cdot)$ given that the rare event under consideration happens. To make sure that the event under consideration is rare, we assume that for all $f \in \mathscr{F}_t$

$$\left(\frac{\partial}{\partial \vartheta_1} M_f(\boldsymbol{\vartheta}) \Big|_{\boldsymbol{\vartheta}=\mathbf{0}}, \ldots, \frac{\partial}{\partial \vartheta_L} M_f(\boldsymbol{\vartheta}) \Big|_{\boldsymbol{\vartheta}=\mathbf{0}} \right) \notin A.$$

The change of measure we propose is the following. In every run we first sample the path $J(s)$ for $s \in [0, t]$ *under the original measure* $\mathbb{P}$ (i.e., with $J(0) = j_0$, and then using the generator matrix Q). We call the resulting path $f \in \mathscr{F}_t$. For this path, define $\boldsymbol{\vartheta}^\star_f \geqslant \mathbf{0}$ as the optimizing $\boldsymbol{\vartheta}$ in the definition of $\mathbb{I}(f)$ in Eq. 19; $\boldsymbol{b}^\star_f \in A$ is the optimizing $\boldsymbol{b}$.

Conditional on the path f of the background process, under the new measure $\mathbb{Q}$ the number of external arrivals between $t_i(f)$ and $t_{i+1}(f)$ is Poisson with parameter

$$\int_{t_i(f)}^{t_{i+1}(f)} \lambda_{j_i(f)} \beta_{j_i(f)} \left(P_i(u, f)\, \boldsymbol{\vartheta}^\star_f \right) \mathrm{d}u,$$

where $P_i(u, f) := D_i(u, f) D_{i+1}(f) \cdots D_{K(f)}(f)$. The arrival epochs between $t_i(f)$ and $t_{i+1}(f)$ should be drawn using the density

$$f_U^{\mathbb{Q}}(u) = \frac{\beta_{j_i(f)} \left(P_i(u, f)\, \boldsymbol{\vartheta}^\star_f \right)}{\displaystyle\int_{t_i(f)}^{t_{i+1}(f)} \beta_{j_i(f)} \left(P_i(v, f)\, \boldsymbol{\vartheta}^\star_f \right) \mathrm{d}v}.$$

Given an arrival at time u between $t_i(f)$ and $t_{i+1}(f)$, the job sizes $(B^{(1)}, \ldots, B^{(L)})$ should be sampled from a distribution with MGF $\beta_{j_i(f)}(\boldsymbol{\vartheta})$, but then exponentially twisted by

$$\left(\left(P_i(u, f)\, \boldsymbol{\vartheta}^\star_f \right)_1, \ldots, \left(P_i(u, f)\, \boldsymbol{\vartheta}^\star_f \right)_L \right).$$

Remark 3 As mentioned above, the background process is sampled under the original measure, whereas an alternative measure is used for the number of arrivals, the arrival epochs, and the job sizes. The intuition behind this, is that the rare event under consideration is caused by two effects:

- In the first place, samples of the background process J should be close to $f^\star$. Under $\mathbb{P}$ a reasonable fraction ends up close to $f^\star$ — more precisely, the event of J being close to $f^\star$ does not become increasingly rare when n grows. As a consequence, no change of measure is needed here.
- In the second place, given the path of the background process, the $Y_n^{(\ell)}(t)$ should exceed the values na_ℓ, for $\ell = 1, \ldots, L$. This event *does* become exponentially rare as n grows, so importance sampling is to be applied here.

4.3 Efficiency Properties of Importance Sampling Procedure

We now analyze the speed up realized by the change of measure introduced in the previous subsection. Unlike our results for the non-modulated systems, now we cannot find the precise rate of growth of Σ_n. What *is* possible though, is proving *asymptotic efficiency* (also sometimes referred to as *logarithmic efficiency*), in the sense that we can show that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{Q}}(L^2 I) = \lim_{n\to\infty} \frac{2}{n} \log p_n(\boldsymbol{a}) = -2 \inf_{f\in\mathscr{F}_t} \inf_{\boldsymbol{b}\in A} \sup_{\boldsymbol{\vartheta}} \left(\langle \boldsymbol{\vartheta}, \boldsymbol{b} \rangle - \log M_f(\boldsymbol{\vartheta}) \right)$$

(where the second equality is a consequence of Eq. 19). This equality is proven as follows. As by Jensen's inequality $\mathbb{E}_{\mathbb{Q}}(L^2 I) \geqslant (\mathbb{E}_{\mathbb{Q}}(LI))^2 = (p_n(\boldsymbol{a}))^2$, we are left to prove the upper bound:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}_{\mathbb{Q}}(L^2 I) \leqslant \lim_{n\to\infty} \frac{2}{n} \log p_n(\boldsymbol{a}).$$

If the path of $J(\cdot)$ equals $f \in \mathscr{F}_t$, it follows by an elementary computation that we have constructed the measure $\mathbb{Q}$ such that

$$L = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \prod_{\ell=1}^{L} \exp\left(-\langle \boldsymbol{\vartheta}^\star_f, \boldsymbol{Y}_n(t)\rangle + n \log M_f(\boldsymbol{\vartheta}^\star_f)\right).$$

The fact that $\boldsymbol{\vartheta}^\star_f$ is componentwise non-negative, in combination with the fact that $\boldsymbol{Y}_n(t) \geqslant \boldsymbol{a}$ when $I = 1$, entails that

$$LI \leqslant \exp\left(-n\,\langle \boldsymbol{\vartheta}^\star_f, \boldsymbol{a}\rangle + n \log M_f(\boldsymbol{\vartheta}^\star_f)\right) = \exp\left(-n\,\langle \boldsymbol{\vartheta}^\star_f, \boldsymbol{b}^\star_f\rangle + n \log M_f(\boldsymbol{\vartheta}^\star_f)\right) = \mathrm{e}^{-n\,\mathbb{I}_f(\boldsymbol{a})},$$

noting that $\boldsymbol{a}$ and $\boldsymbol{b}^\star_f$ may only differ if the corresponding entry of $\boldsymbol{\vartheta}^\star_f$ equals 0 (that is, $\langle \boldsymbol{a} - \boldsymbol{b}^\star_f, \boldsymbol{\vartheta}^\star_f\rangle = 0$). The upper bound thus follows: with $f^\star$ the minimizing path in Eq. 19, recalling that $J(\cdot)$ is sampled under $\mathbb{P}$,

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \leqslant \mathbb{E}\,\mathrm{e}^{-2n\,\mathbb{I}_J(\boldsymbol{a})} \leqslant \mathrm{e}^{-2n\,\mathbb{I}_{f^\star}(\boldsymbol{a})}.$$

We have established the following result.

Proposition 3 *As $n \to \infty$, the proposed importance sampling procedure is asymptotically efficient. This means that the number of runs needed grows subexponentially:*

$$\lim_{n\to\infty} \frac{1}{n} \log \Sigma_n = 0.$$

Remark 4 In the scaling considered, for both the logarithmic asymptotics of $p_n(\boldsymbol{a})$ and our importance sampling algorithm, the precise transition rates q_{ij} do not matter; the only crucial element is that the background process is irreducible. Observe that, even though the logarithmic asymptotics of $p_n(a)$ do not depend on the actual values of the transition rates q_{ij}, the probability $p_n(a)$ itself and its exact asymptotics *do* depend on those rates. We refer to Blom et al. (2017) for the exact asymptotics of a related infinite-server model; it is noted that the derivation of such precise asymptotics is typically highly involved.

The above reasoning indicates that the proposed procedure remains valid under more general conditions: the ideas carry over to any situation in which the rates are piecewise constant along the most likely path.

4.4 Simulation Experiments

We performed experiments featuring a single-node system under Markov modulation. In our example the job sizes stem from an exponential distribution. When the background process is in state i, the arrival rate is λ_i, the job-size distribution is exponential with parameter μ_i, and the rate at which the storage level decays is r_i, for $i \in \{1, \ldots, d\}$.

The change of measure is then implemented as follows. As pointed out in Section 4.2, per run a path f of the background process is sampled under the original measure $\mathbb{P}$. Suppose along this path there are K transitions (remarking that, for compactness, we leave out the argument f here), say at times t_1 up to t_K; with $t_0 = 0$ and $t_{K+1} = t$, the state between t_i

and t_{i+1} is denoted by j_i, for $i = 0, \ldots, K$. Per run a specific change of measure is to be computed, parametrized by the t_i and j_i, as follows.

We define

$$P_i(u) := \bar{P}_i e^{r_{j_i} u}, \quad \bar{P}_i := e^{-r_{j_i} t_{i+1}} \prod_{i'=i+1}^{K} e^{-r_{j_{i'}} (t_{i'+1} - t_{i'})};$$

the product in this expression should be interpreted as 1 if $i + 1 > K$. It is readily checked that

$$M(\vartheta) = \prod_{i=0}^{K} \exp\left(\lambda_{j_i} \int_{t_i}^{t_{i+1}} \frac{P_i(u)\,\vartheta}{\mu_{j_i} - P_i(u)\,\vartheta} \mathrm{d}u \right).$$

Let $\vartheta^\star$ be the maximizing argument of $\vartheta a - \log M(\vartheta)$.

We can now provide the alternative measure $\mathbb{Q}$ for this path of the background process. The number of arrivals between t_i and t_{i+1} (for $i = 0, \ldots, K$) becomes Poisson with parameter

$$\begin{aligned} \int_{t_i}^{t_{i+1}} \lambda_{j_i} \frac{\mu_{j_i}}{\mu_{j_i} - P_i(u)\,\vartheta^\star} \mathrm{d}u &= \frac{\lambda_{j_i}}{r_{j_i}} \log\left(\frac{\mu_{j_i} - \bar{P}_i e^{r_{j_i} t_i} \vartheta^\star}{\mu_{j_i} e^{-r_{j_i}(t_{i+1} - t_i)} - \bar{P}_i e^{r_{j_i} t_i} \vartheta^\star} \right) \\ &= \frac{\lambda_{j_i}}{r_{j_i}} \log\left(\frac{\mu_{j_i} - \bar{P}_i e^{r_{j_i} t_i} \vartheta^\star}{\mu_{j_i} - \bar{P}_i e^{r_{j_i} t_{i+1}} \vartheta^\star} \right) + \lambda_{j_i} (t_{i+1} - t_i). \end{aligned}$$

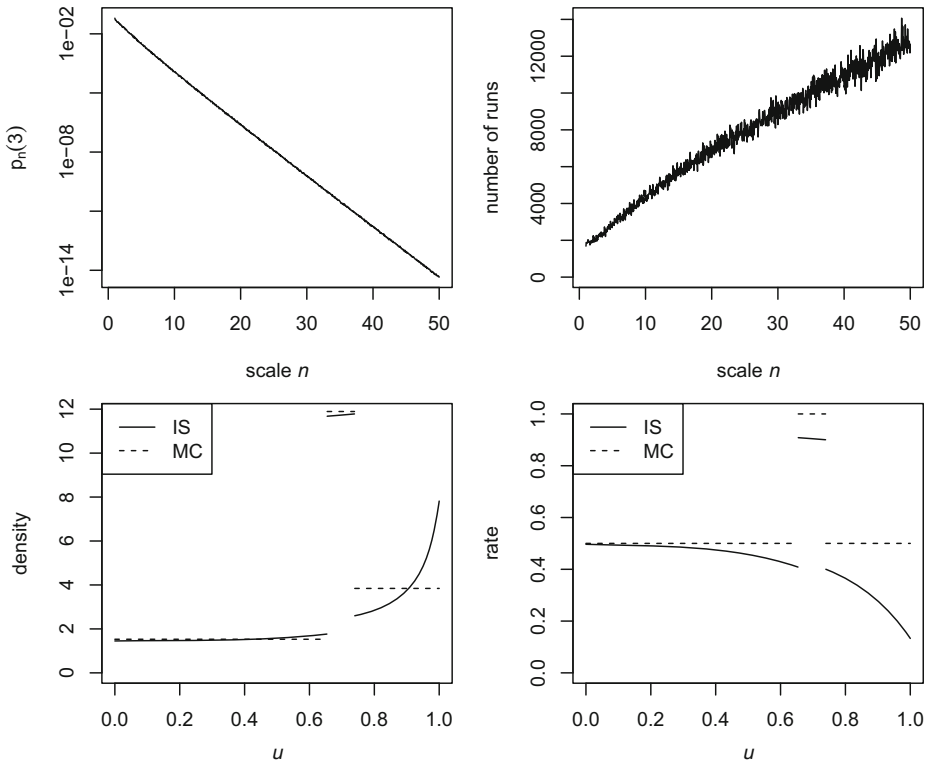

Fig. 6 Numerical results for Section 4.4: first example

(where it is noted that this expression is larger than $\lambda_{j_i}(t_{i+1}-t_i)$, which was the parameter under $\mathbb{P}$). The density of each of the arrivals between t_i and t_{i+1} becomes

$$\left(\frac{1}{\mu_{j_i}-P_i(u)\,\vartheta^\star}\right)\Bigg/\int_{t_i}^{t_{i+1}}\left(\frac{1}{\mu_{j_i}-P_i(v)\,\vartheta^\star}\right)\mathrm{d}v$$
$$=\left(\frac{\mu_{j_i}}{\mu_{j_i}-P_i(u)\,\vartheta^\star}\right)\Bigg/\frac{1}{r_{j_i}}\log\left(\frac{\mu_{j_i}-\bar{P}_i e^{r_{j_i}t_i}\vartheta^\star}{\mu_{j_i}e^{-r_{j_i}(t_{i+1}-t_i)}-\bar{P}_i e^{r_{j_i}t_i}\vartheta^\star}\right)$$

(rather than a uniform distribution, as was the case under $\mathbb{P}$); sampling from this distribution is easy, since the inverse distribution function can be determined in closed form. Given an arrival that takes place at time u between t_i and t_{i+1}, the job size is exponential with parameter $\mu_{j_i}-P_i(u)\,\vartheta^\star$ (rather than exponential with parameter μ_{j_i}).

We now describe two examples in which the dimension of the background process is $d=2$, $q_{12}=q_{21}=2$, and $t=1$. In the first example we fix $a=3$, $\boldsymbol{\lambda}=(2,1)$, $\boldsymbol{\mu}=(\frac{1}{2},1)$, and $\boldsymbol{r}=(5,1)$, in the second example $a=0.8$, $\boldsymbol{\lambda}=(0.9,1)$, $\boldsymbol{\mu}=(0.9^{-1},1)$, and $\boldsymbol{r}=(0.3,0.6)$. As before, we simulate until the precision of the estimate has reached $\varepsilon=0.1$. The top two panels in Figs. 6–7 should be read as those in Figs. 1–3; the bottom two panels correspond to the density of the arrival epochs and the rate of the exponential job sizes, respectively, for f the 'empirical maximizer' of $\mathbb{I}_f(a)$ (i.e., the maximizer of $\mathbb{I}_f(a)$ over all paths f of the background process that were sampled in the simulation experiment).

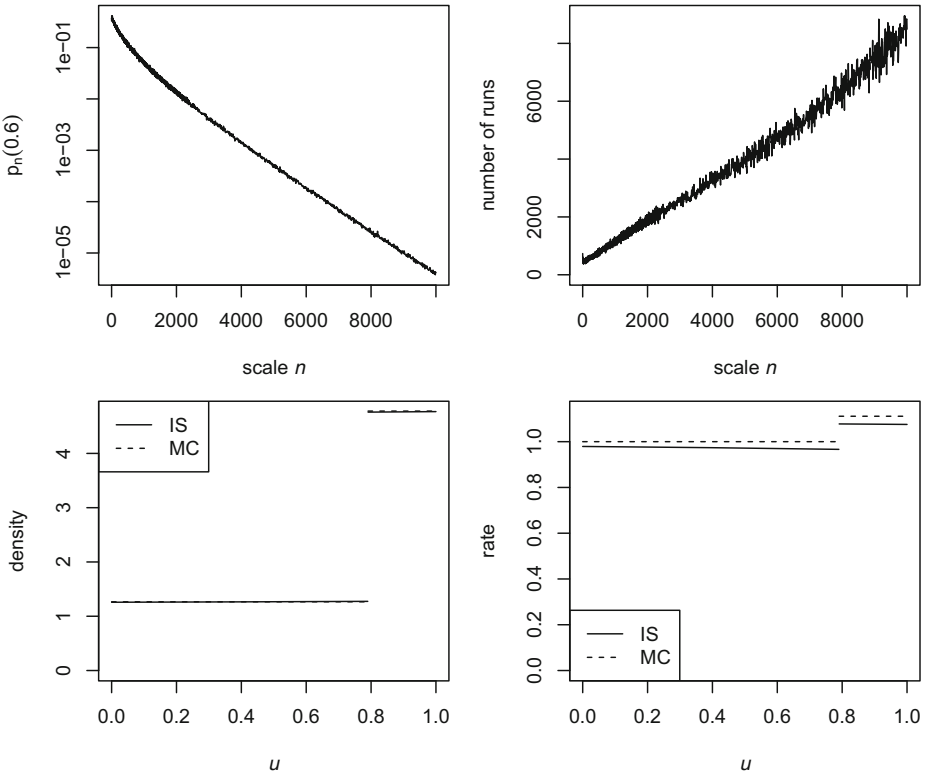

Fig. 7 Numerical results for Section 4.4: second example

In the first example the thus obtained 'optimal path' subsequently visits states 1, 2, and 1, where the corresponding jump times are $t_1^\star = 0.654$ and $t_2^\star = 0.739$, and the decay rate is 0.573. The mean numbers of arrivals in the three parts of the optimal path are 1.392, 0.090 and 0.963 respectively, whereas for Monte Carlo sampling these are 1.308, 0.085 and 0.522 respectively.

In the second example the optimal path subsequently visits states 2 and 1, where the corresponding jump time is $t_1^\star = 0.790$. In this case the decay rate has the value 0.000806. The mean numbers of arrivals in the two parts of the optimal path are 0.812 and 0.195 respectively, which are slightly higher than the corresponding values under Monte Carlo sampling (0.790 and 0.189 respectively). Observe that in this example the difference between the two measures is relative small, also reflected by the small value of the decay rate; the event under consideration technically qualifies as 'rare' in that $p_n(0.8) \to 0$ as $n \to \infty$, but has a relatively high likelihood (e.g. as compared to the first example). As a consequence of the fact that both measures almost coincide, the two densities in the bottom-left panel can hardly be distinguished.

We observe that the top panels confirm that in both examples (i) $p_n(a)$ decays roughly exponentially in n, (ii) the number of runs needed grows roughly linearly in n (in the first example slightly sublinearly).

5 Discussion and Concluding Remarks

In this paper we have considered the probability of attaining a value in a rare set A at a fixed point in time t: with $A = [a_1, \infty) \times \cdots \times [a_L, \infty)$,

$$p_n(a) = \mathbb{P}\left(Y_n^{(1)}(t) \geqslant na_1, \ldots, Y_n^{(L)}(t) \geqslant na_L\right).$$

A relevant related quantity is the probability of having reached the set A *before* t:

$$\mathbb{P}\left(\exists s \leqslant t : Y_n^{(1)}(s) \geqslant na_1, \ldots, Y_n^{(L)}(s) \geqslant na_L\right); \tag{20}$$

observe that this probability increases to 1 as $t \to \infty$. Alternatively, one could study the probability that all a_ℓ (for $\ell = 1, \ldots, L$) are exceeded before t, *but not necessarily at the same time*:

$$\mathbb{P}\left(\exists s_1 \leqslant t : Y_n^{(1)}(s_1) \geqslant na_1, \ldots, \exists s_L \leqslant t : Y_n^{(L)}(s_L) \geqslant na_L\right). \tag{21}$$

Powerful novel sample-path large deviations results by Budhiraja and Nyquist (2015), which deal with a general class of multi-dimensional shot-noise processes, may facilitate the development of efficient importance sampling algorithms for non-modulated linear stochastic fluid networks. The results in (Budhiraja and Nyquist 2015) do not cover Markov modulation, though.

In the current setup of Section 4 the speed of the background process is kept fixed, i.e., not scaled by n. For modulated diffusions a sample-path large deviation principle has been recently established in Huang et al. (2016) for the case that the background process is sped up by a factor n (which amounts to multiplying the generator matrix Q by n); the rate function decouples into (i) a part concerning the rare-event behavior of the background process and (ii) a part concerning the rare-event behavior of the diffusion (conditional on

the path of the background process). With a similar result for the Markov-modulated linear stochastic fluid networks that we have studied in this paper, one could potentially set up an efficient importance sampling procedure for the probabilities (20) and (21) under this scaling.

Acknowledgments The research of O. Boxma, D. Koops and M. Mandjes was partly funded by the NWO Gravitation Project NETWORKS, Grant Number 024.002.003. The research of O. Boxma was also partly funded by the Belgian Government, via the IAP Bestcom project. The research of E. Cahen was funded by an NWO grant, Grant Number 613.001.352.

Appendix A

We here point out how (6) can be established; the line of reasoning is precisely the same as in the derivation of (5) in (Dembo and Zeitouni 1998, Thm. 3.7.4). First write

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) = \mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\vartheta^\star Y_n(t)}\mathrm{e}^{2n\log M(\vartheta^\star)}1_{\{Y_n(t)\geqslant na\}}) = \mathrm{e}^{-2nI(a)}\,\mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\vartheta^\star(Y_n(t)-na)}1_{\{Y_n(t)\geqslant na\}}),$$

which, with $Z_n := (Y_n(t) - na)/\sqrt{n}$, equals

$$\mathrm{e}^{-2nI(a)}\,\mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\vartheta^\star Z_n\sqrt{n}}1_{\{Z_n\geqslant 0\}}).$$

Observe that $\mathbb{E}_{\mathbb{Q}}\, Y_n = na$, due to the very choice of $\mathbb{Q}$. This entails that Z_n converges in distribution to a centered Normal random variable; as can be verified, the corresponding variance is τ (where τ is defined in Eq. 5). Using the Berry-Esseen-based justification presented in (Dembo and Zeitouni 1998, page 111), we conclude that, as $n \to \infty$,

$$\mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\vartheta^\star Z_n\sqrt{n}}1_{\{Z_n\geqslant 0\}}) \sim \int_0^\infty \mathrm{e}^{-2\vartheta^\star\sqrt{n}\,x}\,\frac{1}{\sqrt{2\pi\tau}}\mathrm{e}^{-x^2/(2\tau)}\mathrm{d}x.$$

Completing the square, the right-hand side of the previous display equals, with $\mathcal{N}(\mathrm{M}, \mathrm{V})$ a normal random variable with mean M and variance V,

$$\mathrm{e}^{2(\vartheta^\star)^2 n\tau}\,\mathbb{P}\left(\mathcal{N}(-2\vartheta^\star\sqrt{n}\,\tau, \tau) > 0\right) = \mathrm{e}^{2(\vartheta^\star)^2 n\tau}\,\mathbb{P}\left(\mathcal{N}(0,1) > 2\vartheta^\star\sqrt{n\tau}\right).$$

Now we use the standard equivalence (as $x \to \infty$)

$$\mathbb{P}(\mathcal{N}(0,1) > x) \sim \frac{1}{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2},$$

to obtain

$$\int_0^\infty \mathrm{e}^{-2\vartheta^\star\sqrt{n}\,x}\,\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/(2\tau)}\mathrm{d}x \sim \frac{1}{\sqrt{n}}\,\frac{1}{2\vartheta^\star\sqrt{2\pi\tau}}.$$

Combining the above, we derive the claim:

$$\mathbb{E}_{\mathbb{Q}}(L^2 I) \sim \frac{1}{\sqrt{n}}\,\frac{1}{2\vartheta^\star\sqrt{2\pi\tau}}\mathrm{e}^{-2nI(a)}.$$

We now proceed with the computations underlying (12). To this end, first observe that

$$L = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = \mathrm{e}^{-\langle \boldsymbol{\vartheta}^\star, \boldsymbol{Y}_n(t)\rangle \mathrm{e}^{n \log M(\boldsymbol{\vartheta}^\star)}}.$$

As a consequence, in line with the above computation for the one-dimensional case,

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}(LI) &= \mathrm{e}^{-nI(\boldsymbol{b}^\star)}\mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-\langle \boldsymbol{\vartheta}^\star, \boldsymbol{Y}_n(t)-n\boldsymbol{a}\rangle}1_{\{\boldsymbol{Y}_n(t)\in A\}}),\\ \mathbb{E}_{\mathbb{Q}}(L^2I) &= \mathrm{e}^{-2nI(\boldsymbol{b}^\star)}\mathbb{E}_{\mathbb{Q}}(\mathrm{e}^{-2\langle \boldsymbol{\vartheta}^\star, \boldsymbol{Y}_n(t)-n\boldsymbol{a}\rangle}1_{\{\boldsymbol{Y}_n(t)\in A\}}).\end{aligned}$$

It was proven in (Chaganthy and Sethuraman 1996, Thm. 3.4) that

$$p_n(\boldsymbol{a}) = \mathbb{E}_{\mathbb{Q}}(LI) \sim \frac{1}{\sqrt{\tau}}\left(\prod_{i\in D}\vartheta_i^\star\right)^{-1}(2\pi n)^{-D/2}\,\mathrm{e}^{-nI(\boldsymbol{b}^\star)},$$

while at the same time

$$\mathbb{E}_{\mathbb{Q}}(L^2I) \sim \frac{1}{\sqrt{\tau}}\left(\prod_{i\in D}(2\vartheta_i^\star)\right)^{-1}(2\pi n)^{-D/2}\,\mathrm{e}^{-2nI(\boldsymbol{b}^\star)}.$$

This immediately leads to Eq. 12.

References

Asmussen S, Glynn P (2007) Stochastic simulation. Springer, New York

Asmussen S, Kortschak D (2015) Error rates and improved algorithms for rare event simulation with heavy Weibull tails. Methodol Comput Appl Probab 17:441–461

Asmussen S, Nielsen H (1995) Ruin probabilities via local adjustment coefficients. J Appl Probab 33:736–755

Bahadur R, Rao RR (1960) On deviations of the sample mean. Ann Math Stat 31:1015–1027

Blanchet J, Leder K, Glynn P (2008) Strongly efficient algorithms for light-tailed random walks: an old folk song sung to a faster new tune. In: L'Ecuyer P, Owen A (eds) Proceedings of the Eighth International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing (MCQMC 2008). Springer, Berlin, pp 227-248

Blanchet J, Mandjes M (2009) Rare event simulation for queues. In: Rubino G, Tuffin B (eds) Rare event simulation using Monte Carlo methods. Wiley, Chichester, pp 87–124

Blom J, De Turck K, Mandjes M (2017) Refined large deviations asymptotics for Markov-modulated infinite-server systems. Euro J Oper Res 259:1036–1044

Blom J, Mandjes M (2013) A large-deviations analysis of Markov-modulated infinite-server queues. Oper Res Lett 41:220–225

Boxma O, Cahen E, Koops D, Mandjes M (2018) Linear networks: rare-event simulation and Markov modulation. arXiv:1705.10273

Budhiraja A, Nyquist P (2015) Large deviations for multidimensional state-dependent shot-noise processes. J Appl Probab 52:1097–1114

Cahen E, Mandjes M, Zwart B (2017) Rare event analysis and efficient simulation for a multi-dimensional ruin problem. Probab Eng Inform Sci 31:265–283

Chaganthy N, Sethuraman J (1996) Multidimensional strong large deviation theorems. J Stat Plan Inference 55:265–280

Dembo A, Zeitouni O (1998) Large deviations techniques and applications, 2nd ed. Springer, New York

Ganesh A, Macci C, Torrisi G (2007) A class of risk processes with reserve-dependent premium rate: sample path large deviations and importance sampling. Queueing Syst 55:83–94

Glasserman P, Juneja S (2008) Uniformly efficient importance sampling for the tail distribution of sums of random variables. Math Oper Res 33:36–50

Huang G, Jansen HM, Mandjes M, Spreij P, De Turck K (2016) Markov-modulated Ornstein-Uhlenbeck processes. Adv Appl Probab 48:235–254

Huang G, Mandjes M, Spreij P (2016) Large deviations for Markov-modulated diffusion processes with rapid switching. Stoch Process Appl 126:1785–1818

Juneja S, Shahabuddin P, Nelson B (2006) Rare event simulation techniques: an introduction and recent advances. In: Henderson S (ed) Handbook in operations research and management sciences, volume 13: Simulation, pp 291–350

Kella O, Stadje W (2002) Markov modulated linear fluid networks with Markov additive input. J Appl Probab 39:413–420

Kella O, Whitt W (1999) Linear stochastic fluid networks. J Appl Probab 36:244–260

Kuhn J, Mandjes M, Taimre T (2017) Exact asymptotics of sample-mean related rare-event probabilities. Probability in the Engineering and Informational Sciences, to appear

Magnus J, Neudecker H (1979) The commutation matrix: some properties and applications. Ann Stat 7:381–394

Rabehasaina L (2006) Moments of a Markov-modulated irreducible network of fluid queues. J Appl Probab 43:510–522

Sezer A (2009) Importance sampling for a Markov modulated queuing network. Stoch Process Appl 119:491–517