LARGE DEVIATIONS OF AN INFINITE-SERVER SYSTEM WITH A LINEARLY SCALED BACKGROUND PROCESS

K.E.E.S. DE TURCK [†], M.R.H. MANDJES [•,⋆]

ABSTRACT. This paper studies an infinite-server queue in a Markov environment, that is, an infinite-server queue with arrival rates and service times depending on the state of a Markovian background process. We focus on the probability that the number of jobs in the system attains an unusually high value. Scaling the arrival rates λ_i by a factor N and the transition rates ν_{ij} of the background process as well, a large-deviations based approach is used to examine such tail probabilities (where N tends to ∞). The paper also presents qualitative properties of the system's behavior conditional on the rare event under consideration happening.

KEYWORDS. Queues ⋆ infinite-server systems ⋆ Markov modulation ⋆ large deviations

Work done while K. de Turck was visiting Korteweg-de Vries Institute for Mathematics, University of Amsterdam, the Netherlands, with greatly appreciated financial support of *Fonds Wetenschappelijk Onderzoek / Research Foundation – Flanders*.

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

† TELIN, Ghent University, St.-Pietersnieuwstraat 41, B9000 Gent, Belgium.

M. Mandjes is also with EURANDOM, Eindhoven University of Technology, Eindhoven, the Netherlands, and IBIS, Faculty of Economics and Business, University of Amsterdam, Amsterdam, the Netherlands.

1. INTRODUCTION

Queues with infinitely many servers have found widespread use in various application domains, often as an approximation for models with many servers. In these systems jobs arrive, are served in parallel, to leave when their service is completed. While rooted in communication networks, where the so-called Erlang model describes the dynamics of the number of calls in progress, applications in various other domains have been explored, such as road traffic [19] and biology [16, 17].

In the standard infinite-server model, referred to as M/G/∞, jobs arrive according to a Poisson process with rate λ, where their service times form a sequence of independent and identically distributed (i.i.d.) random variables (distributed as a random variable B with finite first moment), independent of the call arrival process. In such M/G/∞ systems, a key result states that the stationary number of jobs in the system obeys a Poisson distribution with mean $\lambda\,\mathbb{E}B$ (irrespective of the precise distribution of the service times). This basic infinite-server system may be considered somewhat restrictive, though: in many practical situations the assumptions of a constant arrival rate and the jobs stemming from a single distribution are not realistic. A model that allows the input process to exhibit some sort of 'burstiness' is the so-called *Markov-modulated* infinite-server queue. In this model, a finite-state irreducible continuous-time Markov process (often referred to as the *background process*, or *modulating process*) modulates the input process: if the background

Date: April 15, 2013.

Key words and phrases. Markov-modulated Poisson process, queues, general service times, large deviations.

process is in state i, the arrival process is a Poisson process with rate, say, λ_i, while the service times are distributed as a random variable, say, B_i (while the obvious independence conditions are imposed).

The Markov-modulated infinite-server queue has attracted some attention in recent years (but substantially less than the corresponding Markov-modulated single-server queue). The main focus in the literature so far has been on characterizing (through the derivation of moments, or even the full probability generating function) the steady-state number of jobs in the system. The most striking feature is that the number of jobs in the system still has a Poisson distribution, but now with a *random* parameter; a few key references are [5, 8, 11, 15]. Interestingly, under an appropriate time-scaling [2, 9] in which the transitions of the background process occur at a faster rate than the Poisson arrivals, we retrieve the Poisson distribution (with a *deterministic* parameter, that is) for the steady-state number of jobs in the system. Recently, transient results have been obtained as well, under specific scalings of the arrival rates and transition times of the modulating Markov chain [2, 3].

Contribution. In this paper we focus on Markov-modulated infinite-server queues in a large-deviations setting. More precisely, we study the probability that the number of jobs present in the system at some time t attains some unusually high value. In the past in two short papers we have identified the corresponding tail asymptotics in two specific regimes: (i) one in which the transitions of the background process occur at a considerably slower rate than the job arrivals [1], and (ii) one in which the transitions of the background process occur at a considerably faster rate than the job arrivals [4]. In both cases the large deviations are those of a Poisson random variable; in the former case the (non-trivial) parameter value corresponds to the background process' 'worst-case behavior' (constructed so as to build up as many jobs as possible), whereas in the latter case the system essentially behaves as a standard M/G/∞ queue with appropriately chosen arrival rate and service times (e.g., this arrival rate is a weighted sum of the λ_i, where the weights follow from the equilibrium distribution of the background process). These papers, however, do *not* cover the (technically challenging) case in which the timescale of the jumps of the background process and the timescale of the arrival process grow in a proportional manner, and it is a large deviations analysis of this linear regime that we present in this paper.

More formally, in our analysis we replace the arrival rates λ_i by $N\lambda_i$, whereas the transition rates of the background process ν_{ij} are replaced by $N\nu_{ij}$; the service time distributions are left unchanged. With $M^{(N)}(t)$ denoting the number of jobs in the system (starting empty) at time 0, the decay rate of $\mathbb{P}(M^{(N)}(t) \geq Na)$ is identified, for $a > \mathbb{E}M^{(N)}(t)/N$, in the regime that $N \to \infty$. As it turns out, this decay rate can be expressed in terms of the solution to a variational problem. In the paper we specialize to the case that the dimension d of the background process equals 2; it is indicated, though, how the analysis should be adapted for $d \in \{3, 4, \ldots\}$.

Organization. The organization of the rest of this paper is as follows. In Section 2, we provide a detailed model description and introduce some notation. Section 3 states and proves the main result of this paper, viz. an expression for the decay rate under study as the solution to a variational problem. Next, in Section 4, we discuss techniques for numerically solving this variational problem. Next, Section 5, contains some discussion of the results as well as a number of concluding remarks. Finally, numerical results are provided in Section 6.

2. MODEL DESCRIPTION

As mentioned above, this paper studies an infinite-server queue with Markov-modulated Poisson arrivals and general service times. In full detail, the model can be described as follows.
Consider an irreducible continuous-time Markov process $(J(t))_{t\in\mathbb{R}}$ on a finite state space $\{1,\dots,d\}$, with $d \in \mathbb{N}$. Its rate matrix is given by $(\nu_{ij})_{i,j=1}^{d}$. Let π_i the stationary probability that the background process is in state i, for $i = 1,\dots,d$. The time spent in state i (often referred to as the *transition time*) has an exponential distribution with mean $1/\nu_i$, where $\nu_i := -\nu_{ii}$.
While the process $(J(t))_{t\in\mathbb{R}}$, also called the *background process* or *modulating process*, is in state i, jobs arrive according to a Poisson process with rate $\lambda_i \geq 0$. The queueing model is an *infinite-server queue*: jobs are served in parallel – in other words: the sojourn time of a job equals its service time. The service times are assumed to be i.i.d. samples distributed as a random variable B_i if the job was generated when the background process was in state i. The usual independence assumptions apply. It is noted that we exclude the case that all λ_i as well as the distributions of the B_i coincide (as otherwise the queue is just an ordinary M/G/∞).
In the sequel, we specialize to the case of a two-state background process ($d = 2$), and the random variable B_i corresponding to an exponential distribution with mean μ_i^{-1}. In the discussion section, we indicate how these assumptions can be relaxed.

3. MAIN RESULT

In this paper, we consider the scaling $\nu_i \mapsto N\nu_i$, for $i = 1, 2$. We call the resulting background process $(J^{(N)}(s))_{s\in\mathbb{R}}$; in this scaling the background process jumps N times as fast. In addition, the arrival rates are scaled by N as well: $\lambda_i \mapsto N\lambda_i$. The objective of the section is to identify the tail asymptotics of the number of jobs present in our Markov-modulated infinite server at time t under this scaling. We let $M^{(N)}(t)$ denote the number of jobs in the system at time t, in the N-scaled model, where it is assumed that the system starts empty at time 0.
Let

$$\mathscr{F} := \{f : [0,t] \to [0,1]\};$$

the set $\mathscr{F}$ should be interpreted as trajectories of the empirical process associated to the background process, in that $f(t) = x$ informally means that around time t the process spends a fraction of time x in state 1 (and hence a fraction $1 - x$ in state 2).
Let $L^{(N)}(0,s)$ denote the fraction of time the N-scaled background process spends in state 1 during the interval $[0,s]$. It is known that $L^{(N)}(0,1)$ satisfies a large deviations principle [6, 12] with rate function

$$\mathbb{I}(x) = \left(\sqrt{\nu_1 x} - \sqrt{\nu_2(1-x)}\right)^2.$$

We denote by $L(0,s)$ the fraction of time spent in state 1 by the non-timechanged process.
Then define

$$\kappa(f) := \int_0^t \left(f(s)\lambda_1 e^{-\mu_1(t-s)} + (1-f(s))\lambda_2 e^{-\mu_2(t-s)}\right) \mathrm{d}s,$$

as well as

$$I_p(\lambda, x) = \lambda - x + x\log\frac{x}{\lambda},$$

which is the large-deviations rate function of a Poisson random variable with parameter λ.
In the following theorem, we claim that $M^{(N)}(t)$ decays exponentially, with a decay rate that results from a variational problem.

Theorem 1. *Let a be larger than*

$$\varrho_t := \frac{\mathbb{E}M^{(N)}(t)}{N} = \frac{1}{\nu_1+\nu_2}\left(\frac{\nu_2\lambda_1}{\mu_1}(1-e^{-\mu_1 t})+\frac{\nu_1\lambda_2}{\mu_2}(1-e^{-\mu_2 t})\right).$$

Then

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(M^{(N)}(t)\geq Na\right) = -\inf_{f\in\mathscr{F}}\left(I_p(\kappa(f),a)+\int_0^t \mathbb{I}(f(s))\mathrm{d}s\right). \tag{1}$$

In the sequel, we denote by $f^\star(\cdot)$ the optimizing path in the right-hand side of (1). Later we establish the claim that it solves the equation

$$A(s)+\frac{aA(s)}{B} = \nu_1-\nu_2+\frac{1-2f(s)}{\sqrt{f(s)(1-f(s))}}\sqrt{\nu_1\nu_2},$$

where

$$A(s) := \lambda_1 e^{-\mu_1(t-s)}-\lambda_2 e^{-\mu_2(t-s)}, \quad B := \int_0^t\left(f(r)\lambda_1 e^{-\mu_1(t-r)}+(1-f(r))\lambda_2 e^{-\mu_2(t-r)}\right)\mathrm{d}r.$$

As $f(\cdot)$ also appears in the expression for B, this is an implicit equation for $f^\star(\cdot)$, which is not directly solvable. We mention that in 4 we discuss a numerical method to evaluate $f^\star(\cdot)$, as well as a perturbation approach which yields closed-form expressions for the first few expansion terms. Owing to its intuitive interpretation, the path $f^\star(\cdot)$ is often referred to as the *optimal path* or *most likely path*; this in means that, conditional on the rare event under consideration, the state frequencies follow, with overwhelming probability, a path close to $f^\star(\cdot)$. This intuition is made more explicit in the proof below.

Proof. of Thm. 1. The starting point of the proof is the insight that $M^{(N)}(t)$ has a Poisson distribution with random mean, cf. [2, 5]. More specifically, it holds that

$$\mathbb{P}\left(M^{(N)}(t)\geq Na\right) = \mathbb{P}\left(P^{(N)}\left(\varphi\left(J^{(N)}\right)\right)\geq Na\right),$$

with

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

here $P^{(N)}(\lambda)$ is a Poisson random variable with mean $N\lambda$.

Lower bound. Let $\boldsymbol{x}$ be an arbitrary vector in $[0,1]^{\sqrt{N}}$. Pick an arbitrary $\delta>0$. Define the event, with $\Delta(x_i) := (x_i-\delta, x_i+\delta)$,

$$\mathscr{E}(\boldsymbol{x},N) := \left\{L^{(N)}\left(0,\frac{t}{\sqrt{N}}\right)\in\Delta(x_1),\ldots,L^{(N)}\left(t-\frac{t}{\sqrt{N}},t\right)\in\Delta(x_{\sqrt{N}})\right\}.$$

(For notational convenience we here assume that t is a multiple of $1/\sqrt{N}$ and that $\sqrt{N}$ is an integer; it is readily checked, though, that these assumptions have no impact on the validity of the claims.) We obtain the obvious lower bound

$$\mathbb{P}\left(P^{(N)}\left(\varphi\left(J^{(N)}\right)\right)\geq Na\,\middle|\,\mathscr{E}(\boldsymbol{x},N)\right)\mathbb{P}\left(\mathscr{E}(\boldsymbol{x},N)\right),$$

for any $\boldsymbol{x} \in [0,1]^{\sqrt{N}}$. Let us first consider the decay rate of $\mathbb{P}\left(\mathscr{E}(\boldsymbol{x},N)\right)$. It is clear that

$$\begin{aligned}\mathbb{P}\left(\mathscr{E}(\boldsymbol{x},N)\right) &\geq \prod_{i=1}^{\sqrt{N}} \max_{j_i\in\{1,2\}} \mathbb{P}\left(L^{(N)}\left(0,\frac{t}{\sqrt{N}}\right) \in \Delta(x_i) \,\middle|\, J^{(N)}(0)=j_i\right)\\ &= \prod_{i=1}^{\sqrt{N}} \max_{j_i\in\{1,2\}} \mathbb{P}\left(L\left(0,t\sqrt{N}\right) \in \Delta(x_i) \,\middle|\, J(0)=j_i\right).\end{aligned}$$

Because of Prop. 1 (use the uniformity!), we have that N large enough, there is a sequence $(C_i)_i$, bounded away from 0, such that for all $i=1,\ldots,\sqrt{N}$,

$$\max_{j_i\in\{1,2\}} \mathbb{P}\left(L\left(0,t\sqrt{N}\right) \in \Delta(x_i) \,\middle|\, J(0)=j_i\right) \geq \frac{C_i}{t\sqrt{N}} e^{-t\sqrt{N}\max\{\mathbb{I}(x_i-\delta),\mathbb{I}(x_i+\delta)\}}.$$

It follows that

$$\begin{aligned}\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(\mathscr{E}(\boldsymbol{x},N)\right) &\geq \liminf_{N\to\infty}\sum_{i=1}^{\sqrt{N}}\left(\frac{1}{N}\log\left(\frac{C_i}{t\sqrt{N}}\right)-\frac{t}{\sqrt{N}}\max\{\mathbb{I}(x_i-\delta),\mathbb{I}(x_i+\delta)\}\right)\\ &= \liminf_{N\to\infty}\sum_{i=1}^{\sqrt{N}}\left(-\frac{t}{\sqrt{N}}\max\{\mathbb{I}(x_i-\delta),\mathbb{I}(x_i+\delta)\}\right).\end{aligned}$$

Realize that this lower bound holds for any x_i; letting $x_i := f^\star(it/\sqrt{N})$, we obtain the lower bound, after letting $\delta\downarrow 0$,

$$\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(\mathscr{E}(\boldsymbol{x},N)\right) \geq -\int_0^t \mathbb{I}(f^\star(s))\mathrm{d}s.$$

Let us now determine a lower bound on the decay rate of

$$\mathbb{P}\left(P^{(N)}\left(\varphi\left(J^{(N)}\right)\right) \geq Na \,\middle|\, \mathscr{E}(\boldsymbol{x},N)\right). \tag{2}$$

Recall the stochastic monotonicity of the Poisson random variable with respect to its mean value:

$$P^{(N)}(\eta_1) \geq_{\mathrm{st}} P^{(N)}(\eta_2) \;\text{ if }\; \eta_1\geq\eta_2.$$

We therefore need to find a lower bound on $\varphi(J^{(N)})$ conditional on $\mathscr{E}(\boldsymbol{x},N)$; it is immediate that the following lower bound applies:

$$\sum_{i=1}^{\sqrt{N}}\int_{(i-1)t/\sqrt{N}}^{it/\sqrt{N}} N\lambda_{J^{(N)}(s)} e^{-\mu_{J^{(N)}(s)}\cdot(t-s)}\mathrm{d}s \;\geq\; t\sqrt{N}\,\psi_\delta(\boldsymbol{x},N),$$

where $\psi_\delta(\boldsymbol{x},N)$ is defined by

$$\sum_{i=1}^{\sqrt{N}}\left((x_i-\delta)\lambda_1\exp\left(-\mu_1 t\left(1-\frac{i-1}{\sqrt{N}}\right)\right)+(1-x_i-\delta)\lambda_2\exp\left(-\mu_2 t\left(1-\frac{i-1}{\sqrt{N}}\right)\right)\right).$$

It follows that the decay rate of (2) is bounded from below by

$$-\frac{t}{\sqrt{N}}\psi_\delta(\boldsymbol{x},N)+a\liminf_{N\to\infty}\log\left(t\sqrt{N}\psi_\delta(\boldsymbol{x},N)\right)-\limsup_{N\to\infty}\frac{1}{N}\log\lceil Na\rceil!.$$

By a straightforward application of 'Stirling', plugging in $x_i := f^\star(it/\sqrt{N})$, and letting $\delta\downarrow 0$, we have that this converges to

$$a-\kappa(f^\star)-a\log\frac{a}{\kappa(f^\star)},$$

as desired. This completes the lower bound.

Upper bound. Partition $[0,1]$ into $[0,\delta), [\delta,2\delta), \ldots, [1-\delta,1]$ (assuming 1 is a multiple of δ). In this way we can partition $[0,1]^{\sqrt{N}}$ into $(1/\delta)^{\sqrt{N}}$ 'cubes', which we denote by S_i, with $i = 1, \ldots, (1/\delta)^{\sqrt{N}}$. Define

$$\mathscr{F}(i,N) := \left\{ \left(L^{(N)}\left(0, \frac{t}{\sqrt{N}}\right), \ldots, L^{(N)}\left(t - \frac{t}{\sqrt{N}}, t\right)\right) \in S_i \right\}.$$

We thus obtain the upper bound

$$\sum_{i=1}^{(1/\delta)^{\sqrt{N}}} \mathbb{P}\left(P^{(N)}\left(\varphi\left(J^{(N)}\right)\right) \geq Na \,\middle|\, \mathscr{F}(i,N)\right) \mathbb{P}\left(\mathscr{F}(i,N)\right).$$

The decay rate of this expression is bounded from above by

$$\limsup_{N\to\infty} \frac{1}{N} \log \frac{1}{\delta^{\sqrt{N}}} + \limsup_{N\to\infty} \frac{1}{N} \max_{i=1}^{(1/\delta)^{\sqrt{N}}} \log \mathbb{P}\left(P^{(N)}\left(\varphi\left(J^{(N)}\right)\right) \geq Na \,\middle|\, \mathscr{F}(i,N)\right) \mathbb{P}\left(\mathscr{F}(i,N)\right),$$

where the first $\limsup$ obviously vanishes.
We first find an upper bound on $\varphi(J^{(N)})$ for $\boldsymbol{x} \in S_i$. Similar to what we did above in the lower bound, we obtain the upper bound $t\sqrt{N}\bar{\psi}_\delta(\boldsymbol{x}, \sqrt{N})$, with $\bar{\psi}_\delta(\boldsymbol{x}, N)$ defined as

$$\sum_{i=1}^{N} \left((x_i+\delta)\lambda_1 \exp\left(-\mu_1 t\left(1 - \frac{i+1}{N}\right)\right) + (1-x_i+\delta)\lambda_2 \exp\left(-\mu_2 t\left(1-\frac{i+1}{N}\right)\right)\right).$$

As a result, due to the Chernoff bound,

$$\log \mathbb{P}\left(P^{(N)}\left(\varphi\left(J^{(N)}\right)\right) \geq Na \,\middle|\, \mathscr{F}(i,N)\right) \leq Na - t\sqrt{N}\bar{\psi}_\delta(\boldsymbol{x}, \sqrt{N}) - aN \log \frac{a\sqrt{N}}{t\bar{\psi}_\delta(\boldsymbol{x}, \sqrt{N})}.$$

In addition, define

$$\bar{\mathbb{I}}(\boldsymbol{x}, N) := \sum_{i=1}^{N} \mathbb{I}(x_i).$$

Combining the above, the decay rate of interest is bounded from above by

$$\limsup_{N\to\infty} \max_{i=1}^{(1/\delta)^{\sqrt{N}}} \max_{\boldsymbol{x}\in S_i} \left(a - a\log a - \frac{t\bar{\psi}_\delta(\boldsymbol{x}, \sqrt{N})}{\sqrt{N}} + a \log \frac{t\bar{\psi}_\delta(\boldsymbol{x}, \sqrt{N})}{\sqrt{N}} - \frac{t\bar{\mathbb{I}}(\boldsymbol{x}, \sqrt{N})}{\sqrt{N}}\right).$$

Letting $\delta \downarrow 0$, we obtain the upper bound

$$\limsup_{N\to\infty} \max_{\boldsymbol{x}\in[0,1]^{\sqrt{N}}} \left(a - a\log a - \frac{t\bar{\psi}_0(\boldsymbol{x}, \sqrt{N})}{\sqrt{N}} + a \log \frac{t\bar{\psi}_0(\boldsymbol{x}, \sqrt{N})}{\sqrt{N}} - \frac{t\bar{\mathbb{I}}(\boldsymbol{x}, \sqrt{N})}{\sqrt{N}}\right).$$

It is straightforward to see that this expression coincides with

$$\limsup_{N\to\infty} \max_{\boldsymbol{x}\in[0,1]^N} \xi(\boldsymbol{x}, N), \quad \text{where} \quad \xi(\boldsymbol{x}, N) := a - a\log a - \frac{t\check{\psi}_0(\boldsymbol{x}, N)}{N} + a\log\frac{t\check{\psi}_0(\boldsymbol{x}, N)}{N} - \frac{t\bar{\mathbb{I}}(\boldsymbol{x}, N)}{N},$$

and

$$\check{\psi}_0(\boldsymbol{x}, N) := \sum_{i=1}^{N} \left(x_i\lambda_1 \exp\left(-\mu_1 t\left(1-\frac{i}{N}\right)\right) + (1-x_i)\lambda_2 \exp\left(-\mu_2 t\left(1-\frac{i}{N}\right)\right)\right).$$

Now optimize this expression with respect to x_i, with $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N$ given. To this end, we first define

$$A_i := \lambda_1 \exp\left(-\mu_1 t\left(1-\frac{i}{N}\right)\right) - \lambda_2 \exp\left(-\mu_2 t\left(1-\frac{i}{N}\right)\right),$$

and

$$B_i \equiv B_i(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_N) := \check{\psi}_0(\boldsymbol{x},N) - \sum_{i=1}^N A_i x_i,$$

which does not depend on x_i. It is seen that $\xi(\boldsymbol{x},N)$ can be written in an alternative fashion as

$$\xi(\boldsymbol{x},N) = \sum_{i=1}^N \left(a - a\log a + A_i x_i \cdot \frac{t}{N} + B_i \cdot \frac{t}{N} + a\log\left(A_i x_i \cdot \frac{t}{N} + B_i \cdot \frac{t}{N}\right) - \frac{t\bar{\mathbb{I}}(\boldsymbol{x},N)}{N}\right),$$

such that

$$\frac{\partial \xi}{\partial x_i} = \frac{t}{N}\left(A_i + \frac{aA_i}{A_i x_i + B_i} - \mathbb{I}'(x_i)\right), \quad \frac{\partial^2 \xi}{\partial x_i^2} = \frac{t}{N}\left(\frac{-aA_i^2}{(A_i x_i + B_i)^2} - \mathbb{I}''(x_i)\right).$$

Conclude that, owing to the convexity of $\mathbb{I}(x)$, the function $\xi(\boldsymbol{x},N)$ is concave in $x_i \in [0,1]$. The maximum is attained in the open interval $(0,1)$, as the derivative of ξ is $-\infty$ in 0 and ∞ in 1. The optimizer $\bar{x}_i^{(N)}$ necessarily solves the first-order condition

$$A_i + \frac{aA_i}{t/N(A_i x_i + B_i)} = \nu_1 - \nu_2 - \frac{1-2x_i}{\sqrt{x_i(1-x_i)}}\sqrt{\nu_1\nu_2},$$

and is the unique solution in $(0,1)$. It follows that $\bar{x}^{(N)}_{sN/t}$ converges, as $N\to\infty$ to $f^\star(s)$; realize that $f^\star(s)$ solves

$$A(s)\left(1+\frac{a}{B}\right) = \nu_1 - \nu_2 - \frac{1-2f(s)}{\sqrt{f(s)(1-f(s))}}\sqrt{\nu_1\nu_2},$$

where

$$A(s) := \lambda_1 e^{-\mu_1(t-s)} - \lambda_2 e^{-\mu_2(t-s)}, \quad B := \int_0^t \left(f(r)\lambda_1 e^{-\mu_1(t-r)} + (1-f(r))\lambda_2 e^{-\mu_2(t-r)}\right)\mathrm{d}r.$$

In fact, this convergence is uniform on $s\in[0,t]$, due to the regularity of the functions involved; additional support for this claim can be found in Section 4.1.
We thus get that

$$\limsup_{N\to\infty} \xi(\bar{\boldsymbol{x}}^{(N)},N) = \limsup_{N\to\infty}\left(a - a\log a - \frac{t\check{\psi}_0(\bar{\boldsymbol{x}}^{(N)},N)}{N} + a\log\frac{t\check{\psi}_0(\bar{\boldsymbol{x}}^{(N)},N)}{N} - \frac{t\bar{\mathbb{I}}(\bar{\boldsymbol{x}}^{(N)},N)}{N}\right),$$

which converges to

$$a - \kappa(f^\star) - a\log\frac{a}{\kappa(f^\star)} + \int_0^t \mathbb{I}(f^\star(s))\mathrm{d}s,$$

as desired.

4. Computation of the most likely path

In this section we first present a numerical scheme for identifying the optimal path $f^\star(\cdot)$. Then we consider the situation in which the arrival and service rates are just slightly non-uniform (that is, $\lambda_1 := \lambda + \gamma\varepsilon$, $\mu_1 := \mu + \delta\varepsilon$, $\lambda_2 := \lambda$, and $\mu_2 := \mu$ for ε small); for this situation we expand the decay rate in ε.

4.1. **Numerical procedure.** We first demonstrate how the most likely path $f^\star(\cdot)$ can be evaluated. To this end, write $B = I + g$, with

$$I = \int_0^t f(r)A(r)\mathrm{d}r, \quad g := \frac{\lambda_2}{\mu_2}(1 - e^{-\mu_2 t}).$$

Also, write the first-order condition as $V(s) = h(f^\star(s))$, with

$$V(s) := \frac{1}{\sqrt{\nu_1\nu_2}}\left(\nu_1 - \nu_2 - A(s)\left(1 - \frac{a}{I+g}\right)\right), \quad h(x) = \frac{1-2x}{\sqrt{x(1-x)}}.$$

It is readily verified that

$$h'(x) = -\frac{1}{2(x(1-x))^{3/2}} < 0,$$

and hence $h(\cdot)$ decreases from ∞ to $-\infty$ in the interval $(0,1)$. We conclude that there is a unique solution to the first order equation. It takes some calculus to verify that

$$f^\star(s) = \frac{1}{2}\left(1 - \frac{V(s)}{\sqrt{4+V^2(s)}}\right) = \frac{1}{2} + \frac{1}{2}\frac{A(s)\left(1 + a/(I+g)\right) - \nu_1 + \nu_2}{\sqrt{(A(s)(1+a/(I+g)) - \nu_1 + \nu_2)^2 + 4\nu_1\nu_2}}.$$

It remains to identify the unknown I. This can be done by solving

$$I = \frac{1}{2}\int_0^t A(s)\mathrm{d}s + \frac{1}{2}\int_0^t \frac{A^2(s)\left(1 + a/(I+g)\right) - A(s)(\nu_1 - \nu_2)}{\sqrt{(A(s)(1+a/(I+g)) - \nu_1 + \nu_2)^2 + 4\nu_1\nu_2}}\mathrm{d}s.$$

The integral in the right hand side of the previous display cannot be solved explicitly; we have to resort to a numerical procedure. Bisection can be used; observe that

$$I \in \left[-\frac{\lambda_2}{\mu_2}(1 - e^{-\mu_2 t}), \frac{\lambda_1}{\mu_1}(1 - e^{-\mu_1 t})\right].$$

In the pre-limit case, the optimizer $\bar{x}_i^{(N)}$ can be found similarly. Realize that $-\lambda_2 \le A_i \le \lambda_1$, so that $\bar{x}_i^{(N)}$ solves

$$A_i + \frac{aA_i}{tB_i/N} = \nu_1 - \nu_2 + \frac{1-2x_i}{\sqrt{x_i(1-x_i)}}\sqrt{\nu_1\nu_2} + e(N),$$

where the error term $e(N)$ is (in absolute value) smaller than a constant multiplied by $1/N$, uniformly in i. Interpreting tB_i/N, with $N \to \infty$, as a Riemann integral, and using the fact that the functions involved are well-behaved, it follows that

$$\lim_{N\to\infty} \bar{x}_{sN/t}^{(N)} \to f^\star(s)$$

uniformly in s.

4.2. **Perturbation.** In this subsection we consider the ε-perturbed model:

$$\lambda_1 := \lambda + \gamma\varepsilon, \quad \mu_1 := \mu + \delta\varepsilon, \quad \lambda_2 := \lambda, \quad \mu_2 := \mu.$$

Our objective is to investigate the expansions $f^\star(s) = \sum_{k=0}^\infty \varepsilon^k (f^\star)^{(k)}(s)$, $A(s) = \sum_{k=0}^\infty \varepsilon^k A^{(k)}(s)$, $V(s) = \sum_{k=0}^\infty \varepsilon^k V^{(k)}(s)$, and $I = \sum_{k=0}^\infty \varepsilon^k I^{(k)}$. The ultimate goal is to find explicit expressions for the first terms appearing in the expansion of I.

The zero-th order terms can be found after elementary calculations; as it turns out, $(f^\star)^{(0)}(s) = \nu_2/(\nu_1+\nu_2) = \pi_1$; $A^{(0)}(s) = 0$; $V^{(0)}(s) = (\nu_2 - \nu_1)/\sqrt{\nu_1\nu_2}$ and $I^{(0)} = 0$. We now seek to quantify the effect of deviations from the uniform case (i.e., deviations from the situation that $\varepsilon = 0$). More concretely, we pursue finding the first and second order terms of $f^\star(\cdot)$ and I.

Note that

$$A^{(k)}(s) = \frac{(-1)^k}{k!}(\delta(t-s))^{k-1}(\lambda\delta(t-s) - k\gamma)e^{-\mu(t-s)}.$$

In particular, we have that

$$A^{(1)}(s) := (\gamma - \lambda\delta(t-s))e^{-\mu(t-s)}, \quad A^{(2)}(s) := \left(\frac{1}{2}\lambda\delta^2(t-s)^2 - \gamma\delta(t-s)\right)e^{-\mu(t-s)}.$$

In the following, we need the following three integrals that can be computed with routine calculus:

$$\begin{aligned}
\int_0^t A^{(1)}(s)\mathrm{d}s &= \frac{\mu\gamma - \delta\lambda}{\mu^2}\left(1 - e^{-\mu t}\right) - \frac{\delta\lambda t}{\mu}e^{-\mu t}, \\
\int_0^t A^{(1)}(s)^2\mathrm{d}s &= \frac{1}{4\mu^3}\Big(2\mu^2\gamma^2 - 2\delta\lambda\mu\gamma + \delta^2\lambda^2 \\
&\qquad - e^{-2\mu t}\left(2\mu^2\gamma^2 - 4\gamma\delta\lambda\mu^2 t + 2\gamma\delta\lambda\mu + 2\delta^2\lambda^2\mu^2t^2 + 2\delta^2\lambda^2\mu t + \delta^2\lambda^2\right)\Big), \\
\int_0^t A^{(2)}(s)\mathrm{d}s &= -\frac{\delta}{\mu^3}\left(\mu\gamma - \delta\lambda\right)\left(1 - e^{-\mu t}\right) + \frac{\delta}{2\mu^3}\left(2\gamma\mu^2 t - \delta\lambda\mu^2t^2 - 2\delta\lambda\mu t\right)e^{-\mu t};
\end{aligned}$$

in the sequel, we refer to these integrals by A_1, $A_{1,2}$, and A_2, respectively. From the definition of I and the fact that $A(s)$ has no zero-th order term in ε, we immediately find the linear terms corresponding to I:

$$I^{(1)} = \int_0^t A^{(1)}(s)(f^\star)^{(0)}(s)\mathrm{d}s = \frac{\nu_2}{\nu_1+\nu_2}\int_0^t A^{(1)}(s)\mathrm{d}s = \pi_1 A_1.$$

The second term in the expansion of I is harder to find, though. To this end, we first determine the linear term in the expansion of $f^\star(\cdot)$. From the definition of $V(\cdot)$ we directly identify its linear term:

$$V^{(1)}(s) = -\frac{1 + a/g}{\sqrt{\nu_1\nu_2}}A^{(1)}(s).$$

As a consequence,

$$(f^\star)^{(1)}(s) = -\frac{2V^{(1)}(s)}{(4 + V^{(0)}(s)^2)^{\frac{3}{2}}} = 2A^{(1)}(s)\left(1 + \frac{a}{g}\right)\frac{2\nu_1\nu_2}{(\nu_1+\nu_2)^3}.$$

From this, we find that

$$I^{(2)} = \int_0^t \left(A^{(2)}(s)(f^\star)^{(0)}(s) + A^{(1)}(s)(f^\star)^{(1)}(s)\right)\mathrm{d}s = \pi_1 A_2 + 2\left(1 + \frac{a}{g}\right)\frac{2\nu_1\nu_2}{(\nu_1+\nu_2)^3}A_{1,2}.$$

The last two coefficients we determine are $V^{(2)}(s)$ and $(f^\star)^{(2)}(s)$:

$$\begin{aligned}
V^{(2)}(s) &= \frac{1}{\sqrt{\nu_1\nu_2}}\left(-A^{(2)}(s)\left(1 + \frac{a}{g}\right) + \frac{a}{g^2}A^{(1)}(s)I^{(1)}\right), \\
(f^\star)^{(2)}(s) &= \frac{2V^{(0)}(s)^2V^{(2)}(s) + 8V^{(2)}(s) - 3V^{(0)}(s)V^{(1)}(s)^2}{(4 + V^{(0)}(s)^2)^{\frac{5}{2}}}.
\end{aligned}$$

We have now collected all necessary ingredients to find the first terms in the expansion of the decay rate of the probability of our interest. We first focus on the 'Poisson term', and note that $\kappa(f) = I + g$, so that we have

$$\begin{aligned}
I_p(\kappa(f^\star), a) &= I(\varepsilon) + g - a + a\log\frac{a}{g+I} \\
&= g - a + a\log\frac{a}{g} + \left(1 - \frac{a}{g}\right)I^{(1)}\varepsilon + \left(\left(1 - \frac{a}{g}\right)I^{(2)} + \frac{a}{2g^2}(I^{(1)})^2\right)\varepsilon^2 + o(\varepsilon^2).
\end{aligned}$$

As for the 'Markov term', recall that $(f^\star)^{(0)}(s) = \pi_1$. It follows directly that both $\mathbb{I}(\pi_1) = 0$ and $\mathbb{I}'(\pi_i) = 0$, whereas

$$\mathbb{I}''(\pi_1) = \frac{(\nu_1+\nu_2)^3}{2\nu_1\nu_2}.$$

Elementary algebra now yields

$$\begin{aligned}\int_0^t \mathbb{I}(f^\star(s))\mathrm{d}s &= \frac{1}{2}\left(\int_0^t \varepsilon^2\left((f^\star)^{(1)}(s)\right)^2 \mathrm{d}s\right)\mathbb{I}''(\pi_1) + o(\varepsilon^2)\\ &= \varepsilon^2\left(1+\frac{a}{g}\right)^2 \frac{\nu_1\nu_2}{(\nu_1+\nu_2)^3} A_{1,2}.\end{aligned}$$

The expansion can be continued indefinitely. Indeed, from $(f^\star)^{(n)}(s)$, we find $I^{(n+1)}$ via the definition of I. In turn, $(f^\star)^{(n)}(s)$ can be written in terms of $V^{(k)}(s), k = 0, \ldots, n$, which then allows an expansion in terms of $I^{(k)}, k = 1, \ldots, n-1$. Also, it follows by induction that all integrands have the form of a polynomial function multiplied by an exponential, and can thus be evaluated in closed form in terms of incomplete beta functions.

5. Discussion and ramifications

5.1. **Extensions.** In this section, we list a number of extensions to the results presented in the previous sections. Here we do not strive for complete proofs of these statements. Sometimes only a minor variation of the proof in Section 3 is needed, in other instances the proof requires more technical work. In all cases, we provide a heuristic justifications for the presented results.

- ▷ Theorem 1 characterized the decay rate of 'overflow' events, i.e., $a > \rho_t$. We can follow exactly the same reasoning for underflow events, that is

  $$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)}(t) \leq Na\right)$$

  for $a < \rho_t$. In particular, we obtain the same expression for the decay rate.
- ▷ We can extend the results from exponential service times to general service times by substituting th expression for $\kappa(f)$ by

  $$\kappa_B(f) := \int_0^t \left(f(s)\lambda_1(1-B_1(t-s)) + (1-f(s))\lambda_2(1-B_2(t-s))\right)\mathrm{d}s,$$

  where $B_1(\cdot), B_2(\cdot)$ denote the cumulative distribution functions of the service times of jobs corresponding to the first and second background state.
- ▷ In case of more than two background states, the behavior of the background Markov chain is recorded in d functions $f_i : [0,t] \to [0,1], i = 1, \ldots, d$, under the constraint that $\sum_i f_i(s) = 1, \forall s \in [0,t]$. Regarding the 'Poisson part', in this case our 'twodimensional' $\kappa(f)$ needs to be replaced by its 'multidimensional counterpart':

  $$\kappa(f) := \int_0^t \left(\sum_{i=1}^d f_i(s)\lambda_i e^{-\mu_i(t-s)}\right)\mathrm{d}s.$$

  The 'Markov part' changes more pervasively as there is in general no explicit expression like for $\mathbb{I}(\cdot)$ anymore, nor can we hope to find the optimal path $f^\star(\cdot)$ with a relatively simple procedure as the one we presented in Section 4.

Indeed, the equivalent of $\mathbb{I}$ [6, 7] for (irreducible) continuous-time Markov chains with a state space of size d is:

$$\mathbb{I}(x) = \sup_{u>0} \left[-\sum_{i=1}^{d} x_i \frac{(Qu)_i}{u_i} \right],$$

where x, u are the positive d-dimensional vectors, with $\sum_{i=1}^{d} x_i = 1$. In case Q is symmetric, the supremum can be evaluated explicitly:

$$\mathbb{I}(x) = -\sum_{i,j=1}^{d} \sqrt{x_i} \nu_{ij} \sqrt{x_j}.$$

Symmetric Markov chains, which satisfy the detailed balance equation $\pi_i \nu_{ij} = \pi_j \nu_{ji}$, form another class of Markov chains that have an explicit rate function for the occupation measure:

$$\mathbb{I}(x) = -\sum_{i,j=1}^{d} \pi_i \sqrt{\frac{x_i}{\pi_i}} \nu_{ij} \sqrt{\frac{x_j}{\pi_j}},$$

of which, evidently, the expression for the two-dimensional Markov chain is a special case.

▷ We can also find the large deviations of the job populations split based on their 'types': jobs of type i arrive at times when the background chain is in state i and have a corresponding service rate μ_i, with $i = 1, 2$. With a similar proof as in Section 3, we can establish that

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M_i^{(N)}(t) \geq N a_i, i = 1, 2 \right) = -\inf_{f \in \mathscr{F}} \left(\sum_{i=1}^{2} I_p(\kappa_i(f), a_i) + \int_0^t \mathbb{I}(f(s)) \mathrm{d}s \right),$$

where $M_1^{(N)}(t)$ (resp. $M_2^{(N)}(t)$) denotes the user population at time t of type 1 (resp. of type 2), while $\kappa_1(f)$ and $\kappa_2(f)$ are given by

$$\kappa_1(f) = \int_0^t f(s)\lambda_1 \exp(-\mu_1(t-s))\mathrm{d}s; \quad \kappa_2(f) = \int_0^t (1-f(s))\lambda_2 \exp(-\mu_2(t-s))\mathrm{d}s.$$

The corresponding optimum path can be found by a procedure similar to the one presented in Section 4.1.

▷ From the previous result, we can determine the contribution of each population to the overflow event $\{M^{(N)}(t) \geq Na\}$, indeed, by the contraction principle , we have that the decay rate of this overflow event satisfies

$$-\inf_{a_1+a_2=a} \inf_{f\in\mathscr{F}} \left(\sum_{i=1}^{2} I_p(\kappa_i(f), a_i) + \int_0^t \mathbb{I}(f(s)) \mathrm{d}s \right)$$

The optimal function $f^\star(\cdot)$ is the same that optimizes the original variational problem. It takes some calculations involving Lagrange multipliers to see that the optimal a_1 and a_2 equal

$$a_1 = \frac{\kappa_1(f^\star)}{\kappa_1(f^\star) + \kappa_2(f^\star)} a; \qquad a_2 = \frac{\kappa_2(f^\star)}{\kappa_1(f^\star) + \kappa_2(f^\star)} a.$$

Informally, each job type contributes to the overflow event proportionally with the corresponding Poisson rate $\kappa_i(f^\star)$.

▷ Next, we can also consider non-zero starting conditions, that is, at time 0 there is a population of $Na_1(0)$ jobs of type 1 and of $Na_2(t)$ jobs of type 2. The amount of jobs of type i that is still in the system at time t can be seen as a sum of $Na_i(0)$ Bernoulli random variables with success probability $e^{-\mu_i t}$.

Recall that a Bernoulli random variable with success probability p has the following large-deviations rate function:

$$I_b(p,x) = x\log\left(\frac{x}{p}\right) + (1-x)\log\left(\frac{1-x}{1-p}\right).$$

It is now seen that the decay rate of the system with a non-empty starting condition is the solution of the following variational problem:

$$-\inf\left(\sum_{i=1}^{2} I_p(\kappa_i(f), a_i - x_i a_i(0)) + I_b(e^{-\mu_i t}, x_i) + \int_0^t \mathbb{I}(f(s))\mathrm{d}s\right),$$

where the minimization should be performed over $f \in \mathscr{F}$, x_1 and x_2 in $[0,1]$, and $a_i \geq x_i a_i(0)$ (for $i = 1, 2$).

We conclude this subsection by mentioning that the last three extensions can be combined into a procedure for computing the most probable build-up path of each of the job populations; we do not provide the details here.

5.2. **Interpretation.** As observed above, the formula for the decay rate can be decomposed into a 'Poisson contribution' and a 'Markov-chain contribution'. The Poisson contribution is unbounded, while the Markov contribution is bounded by $\max(\nu_1 t, \nu_2 t)$. It can therefore be anticipated that in case of very high value of a, the Poisson contribution will dominate.

The behavior of the system is to a large extent dependent on the precise shapes of the curves (as functions of $s \in [0,t]$)

$$\lambda_i \exp(-\mu_i(t-s)), \quad \text{for } i = 1, 2.$$

These expressions essentially reflect the arrival rate in both states at any time $s \in [0,t]$, but 'thinned' by a fraction $\exp(-\mu_i(t-s))$ (so that we obtain the jobs that are still in the system at time t). When at a specific time s one curve is higher than the other, this means that the corresponding state is more 'favorable' for creating overflows. The intuition is that during build-ups to overflow ($M^{(N)}(t)$ exceeding Na, that is), the background process will reside more frequently in this state than the stationary probability would predict. The extent to which is deviated from the expected behavior depends on both the cost of doing so (as reflected in the function $\mathbb{I}(\cdot)$), as well as the benefit associated with it (quantified by the difference between the two curves). A similar line of reasoning applies in the regime that just the arrival rates λ_i are scaled by N, whereas the background process is not sped up; see [1].

Unless $\mu_1 = \mu_2$, the curves have a unique intersection point $s^\star$, which is equal to

$$s^\star = t - \frac{\log\lambda_1 - \log\lambda_2}{\mu_1 - \mu_2}.$$

The trivial case is that $\lambda_1 > \lambda_2$ and $\mu_1 < \mu_2$ (or $\lambda_2 > \lambda_1$ and $\mu_2 < \mu_1$); then it is always favorable to be in state 1 (2, respectively). Let us therefore assume, without loss of generality, that $\lambda_1 > \lambda_2$ but $\mu_1 > \mu_2$. Hence if $0 < s^\star < t$, then state 2 is favorable in $[0, s^\star]$ and state 1 in $[s^\star, t]$. We observe that in $[0, s^\star]$, $f^\star(\cdot)$ is smaller than π_1, whereas in $[s^\star, t]$, it will be larger.

We can find two easy upper bounds for the decay rate based on two simple paths, as follows.

- ▷ The first path that we can insert is $f(s) = \pi_1$ for all $s \in [0,t]$. This provides us with the following upper bound to the decay rate:

$$\int_0^t \sum_{i=1}^{2} \pi_i \lambda_i e^{-\mu_i(t-s)}\mathrm{d}s = \sum_{i=1}^{2} \pi_i \frac{\lambda_i}{\mu_i} e^{-\mu_i t}.$$

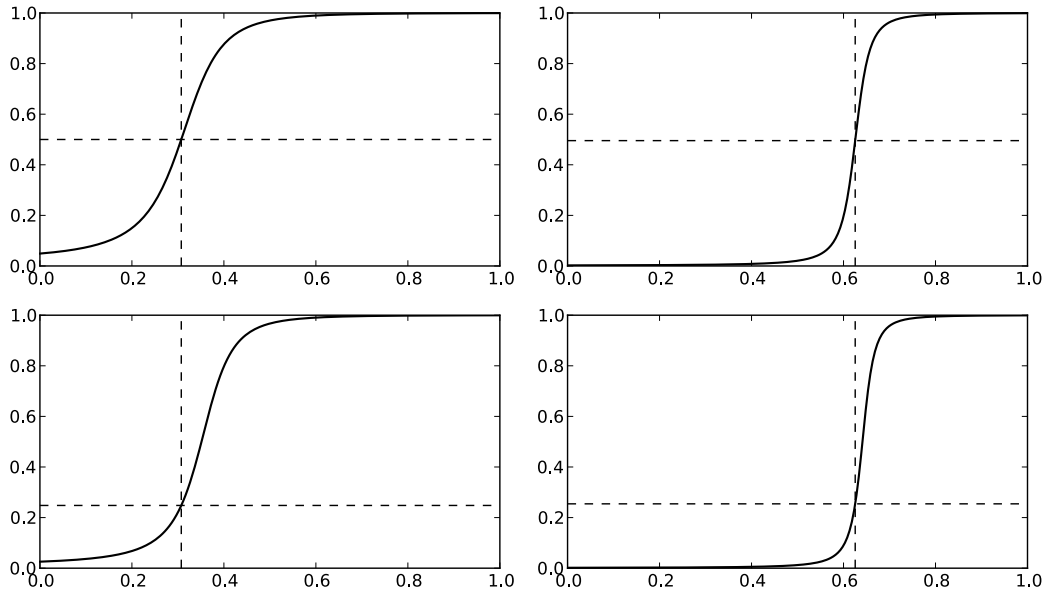

FIGURE 1. Four plots of the optimal path showing the 'switch' around $s = s^\star$.

This path is close to optimal in cases that it is relatively 'expensive' to change the behavior of the Markov chain.

▷ The second path is $f(s) = 1$ for $s \leq s^\star$ and 0 for $s > s^\star$ (or $f(s) = 0$ for $s \leq s^\star$ and 1 for $s > s^\star$; whichever of the two alternatives leads to a lower decay rate). This leads to the decay rate

$$\frac{\lambda_1}{\mu_1}\left(e^{-\mu_1(t-s^\star)} - e^{-\mu_1 t}\right) + \frac{\lambda_2}{\mu_2}\left(1 - e^{-\mu_2(t-s^\star)}\right) + s^\star\nu_1 + (t-s^\star)\nu_2.$$

This path tends to be close to optimal when the Poisson contribution dominates (which is the case for instance for large values of a).

6. NUMERICAL EXAMPLES

In this section, we illustrate the obtained results by means of a series of numerical experiments. We introduce the following alternative characterization of the two-state Markov chain: Let $K = \nu_1 + \nu_2; \pi_1 = \nu_2/K$. As such, π_1 denotes the stationary probability of being in state 1, and K is a measure for the speed at which the continuous-time Markov chain changes states.

In Fig. 1, we show how optimal paths behave around the transition point $s^\star$; we took four typical examples. In the left graphs we took specific values for the arrival and service rates, but changed the background process (thus leaving the position of $s^\star$ unchanged); the same holds for the right graphs. We notice that the optimal path indeed switches behavior at point $s^\star$, and that $f^\star(s^\star) = \pi_1$ as predicted. In the left graphs the value of K is 'moderate', while it is substantially smaller in the right graphs; in the latter case the optimal path tends to look like a step function (cf. the analysis in [1]).

Next, we show in Fig. 2 a plot of the decay rate versus the target level a. It is observed that the decay rate increases rapidly for increasing a. We have included plots of the optimal paths at various places. We see that for smaller a, the optimal paths are fairly flat, whereas curvier graphs are obtained for larger a. This can be intuited from the fact that as a increases, the contribution of

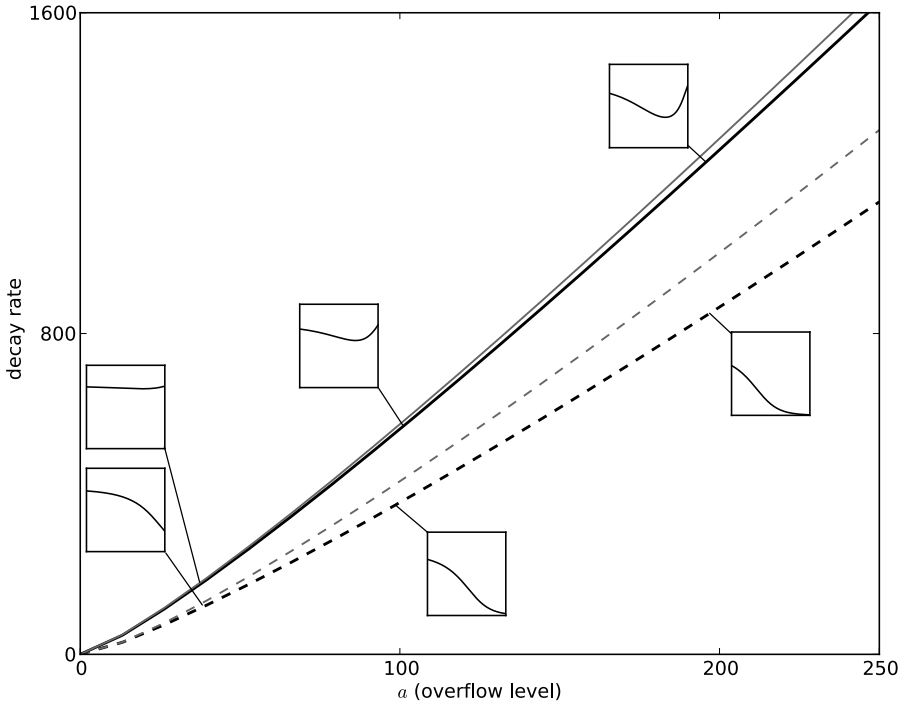

FIGURE 2. Plot of the decay rate versus a for $\lambda_1 = 0.5, \mu_1 = 4, \mu_2 = 3, \pi_1 = 0.75, K = 100, t = 1$; and $\lambda_2 = 4$ for the dashed line; $\lambda_2 = 0.5$ for the full line. The gray lines represent the perturbation-based results.

the Poisson term will get larger, and as such paths that have a more 'expensive' Markov term come into the picture.

Next, we show in Fig. 3 a plot of the decay rate versus K. An increase of K makes the 'Markov contribution' larger, and hence it is in accordance with the intuition that for small K, we see optimal paths that at any time instant almost exclusively reside in the 'most favorable state', whereas for larger K we obtain optimal paths that are close to the stationary probability.

APPENDIX A. ESTIMATES FOR MARKOV FLUID SOURCES

A.1. **Exact asymptotics.** Let $L(0,t)$ be the amount of time the unscaled background process spends in state 1 during the interval $[0,t]$. In this appendix we provide the asymptotics of the probability that $L(0,t)/t$ attains a rare value, that is, values away from the limiting mean $\nu_2/(\nu_1+\nu_2)$.

In the first place, we mention that the usual Chernoff bound estimate gives the upper bound

$$\mathbb{P}\left(\frac{L(0,t)}{t} \geq a\right) \leq \inf_{\vartheta>0} \exp\left(\log \mathbb{E}e^{\vartheta L(0,t)} - \vartheta a t\right).$$

Due to [13, Lemma 2.1], for t large enough, we have that, for a constant K independent of ϑ,

$$\mathbb{E}e^{\vartheta L(0,t)} \leq K e^{\Lambda_+(\vartheta)\, t},$$

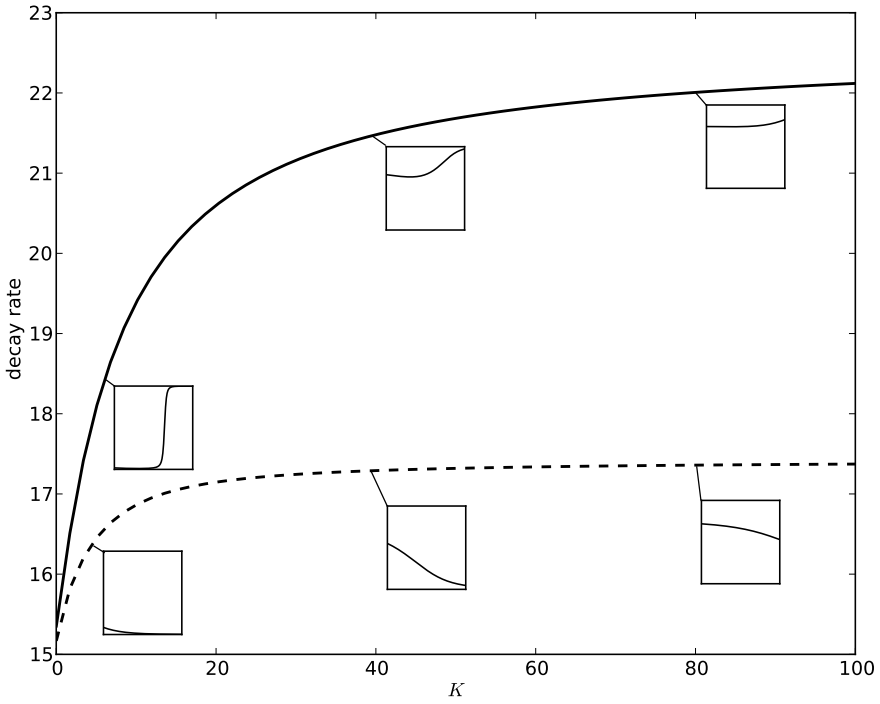

FIGURE 3. Plot of the decay rate versus K for $\lambda_1 = 1, \lambda_2 = 2, \mu_2 = 2, \pi_1 = 0.75, a = 10, t = 1$; and $\mu_1 = 1$ for the dashed line; $\mu_1 = 4$ for the full line.

with

$$\Lambda_{\pm}(\vartheta) := \frac{1}{2}\left(-\nu_1 - \nu_2 + \vartheta \pm \sqrt{(\nu_1 + \nu_2 - \vartheta)^2 + 4\nu_1\vartheta}\right).$$

We thus obtain the uniform upper bound (with K independent of t and a)

$$\mathbb{P}\left(\frac{L(0,t)}{t} \geq a\right) \leq \exp\left(\log K - t \sup_{\vartheta>0}\left(\vartheta a - \Lambda_+(\vartheta)\right)\right) = Ke^{-t\mathbb{I}(a)}.$$

More precise results can be found, though. [13, Lemma 2.1] also provides us with the asymptotics of the moment generating function:

$$\frac{\mathbb{E}e^{\vartheta L(0,t)}}{\kappa(\vartheta)e^{\Lambda_+(\vartheta)t}} \to 1,$$

as $t \to \infty$, with

$$\kappa(\vartheta) := \frac{1}{\nu_1 + \nu_2}\left(\frac{\nu_1\vartheta + (\nu_2 - \nu_1)\Lambda_-(\vartheta)}{\Lambda_+(\vartheta) - \Lambda_-(\vartheta)}\right).$$

Relying on this property, we can identify 'fine asymptotics' of the probability that $L(0,t)/t$ attains a rare value. The idea is that although $L(0,t)$ does not have independent increments, the dependence is sufficiently weak to enable the computation of so-called exact asymptotics, in the spirit of those developed by Bahadur and Rao for sums of independent random variables — this was hinted in Remark c, immediately below [6, Thm. 3.7.4]. The proof is completely analogous to that in [14] for the related model of M/G/∞ input. The uniformity follows as in [10].

Proposition 1. *Let $\eta \equiv \eta(a)$ solve $a = \Lambda'_+(\eta)$. Uniformly in a,*

$$\lim_{t\to\infty} \sqrt{t} e^{t\mathbb{I}(a)} \cdot \mathbb{P}\left(\frac{L(0,t)}{t} \geq a\right) = \frac{1}{\eta\sqrt{2\pi\Lambda''_+(\eta)}} \frac{1}{\kappa(\eta)}.$$

REFERENCES

[1] J. BLOM and M. MANDJES (2013). A large-deviations analysis of Markov-modulated infinite-server queues. Accepted for publication in *OR Letters.*

[2] J. BLOM, O. KELLA, M. MANDJES, and H. THORSDOTTIR (2012). Markov-modulated infinite server queues with general service times. *Submitted.*

[3] J. BLOM, M. MANDJES, and H. THORSDOTTIR (2012). Time-scaling limits for Markov-modulated infinite-server queues. Accepted for publication in *Stochastic Models.*

[4] J. BLOM, K. DE TURCK, and M. MANDJES (2013). Rare event analysis of Markov-modulated infinite-server queues: a Poisson limit. *Submitted.*

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

[6] A. DEMBO and O. ZEITOUNI (1998). *Large Deviations Techniques and Applications*, 2nd edition. Springer, New York.

[7] F. DEN HOLLANDER (2000). *Large Deviations*. Vol. 14 of Fields Institute Monographs, American Mathematical Society, Providence, RI.

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

[9] T. HELLINGS, M. MANDJES, and J. BLOM (2012). Semi-Markov-modulated infinite-server queues: approximations by time-scaling. *Stochastic Models.*

[10] T. HÖGLUND (1979). A unified formulation of the central limit theorem for small and large deviations from the mean. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **49**, 105117.

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

[12] G. KESIDIS, J. WALRAND, and C.-S. CHANG (1993). Effective bandwidths for multiclass Markov fluids and other ATM sources. *IEEE/ACM Transactions on Networking*, **1**, 424–428.

[13] M. MANDJES and P. MANNERSALO (2006). Queueing systems fed by many exponential on-off sources: an infinite-intersection approach. *Queueing Systems*, **54**, 5–20.

[14] M. MANDJES and P. ZURANIEWSKI (2011). M/G/∞ transience, and its applications to overload detection. *Performance Evaluation*, **68**, pp. 507–527.

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

[16] A. SCHWABE, M. DOBRZYŃSKI, K. RYBAKOVA, P. VERSCHURE, and F. BRUGGEMAN (2011). Origins of stochastic intracellular processes and consequences for cell-to-cell variability and cellular survival strategies. *Methods in Enzymology*, **500**, 597-625.

[17] A. SCHWABE, M. DOBRZYŃSKI, and F. BRUGGEMAN (2012). Transcription stochasticity of complex gene regulation models. *Biophysical Journal*, **103**, 1152-1161.

[18] A. SCHWARTZ and A. WEISS (1995). *Large Deviations for Performance Analysis.* Chapman & Hall, London.

[19] T. VAN WOENSEL and N. VANDAELE (2007). Modeling traffic flows with queueing models: an overflow. *Asia-Pacific Journal of Operational Research*, **24**, 235–261.

E-mail address: kdeturck@telin.ugent.be, M.R.H.Mandjes@uva.nl