A LARGE-DEVIATIONS ANALYSIS OF MARKOV-MODULATED INFINITE-SERVER QUEUES

J. BLOM *, M. MANDJES $^{\bullet,\star}$

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. Scaling the arrival rates λ_i by a factor N, tail probabilities are examined when letting N tend to ∞; non-standard large deviations results are obtained. An importance-sampling based estimation algorithm is proposed, that is proven to be logarithmically efficient.

KEYWORDS. Queues $\star$ infinite-server systems $\star$ Markov modulation $\star$ large deviations

1. INTRODUCTION

Infinite-server queues have found widespread use in various application domains, often as an approximation for many-server models. While their original motivation may stem from communication networks engineering, where the so-called Erlang model records the dynamics of the number of calls in progress, applications in various other domains have been explored, such as road traffic [14] and biology [12].

In the standard model, 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); the main result is that the stationary number of jobs in the system obeys a Poisson distribution with mean $\lambda\,\mathbb{E}B$. In many practical situations, however, 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 *Markov-modulated* infinite server queue. In this model, a finite-state irreducible continuous-time Markov process (often referred to as the *background process*) modulates the input process: if the background 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.

The Markov-modulated infinite-server queue started to attract attention in recent years. The main focus in the literature so far has been on characterizing (through the derivation of moments, or even the full probability generating function) of the steady-state number of jobs in the system [3, 5, 9, 11]. Interestingly, under an appropriate time-scaling [2, 8] in which the transitions of the background process occur at a faster rate than the Poisson arrivals, we retrieve the Poisson distribution for the steady-state number of jobs in the system.

Less attention is paid to characterizing the probability that the number of customers in the Markov-modulated infinite-server queue attains some unusually high (or low) value. For the case that there is no Markov modulation, large-deviations analyses have been performed. In e.g. [6, 7] rare-event probabilities are considered for the stationary number of jobs in the system under renewal arrivals

Date: October 15, 2012.

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

(that is, not necessarily Poisson arrivals). In [10, 13] the arrival rate λ is scaled by N, and sample-path large-deviations results are derived for the transient of the number of jobs in the system.

The main contribution of the present paper is a large-deviations analysis of the number of jobs in a Markov-modulated infinite server queue. We do so by scaling the arrival rates λ_i by N (that is, the arrival rate while in state i becomes $N\lambda_i$), but leaving the time-scale of the background process unchanged.

After introducing our model in Section 2, Section 3 presents our main result: an expression for the exponential decay rate (in N) of the *transient overflow probability*, which is defined as the probability that the transient of the number of jobs in the system, denoted by $M^{(N)}(t)$, exceeds Na, for a given a larger than the limiting value of $\mathbb{E}M^{(N)}(t)/N$ (which we call ϱ_t). Interestingly, under the scaling proposed the decay rate of interest is affected by the arrival rates and service-time characteristics only; the transition rates of the background process do not play a role. In addition, it turns out that there is an a^+ *strictly larger than* ϱ_t such that for $a \in [\varrho_t, a^+]$ there is subexponential decay (i.e., decay rate equalling 0), while for $a > a^+$ the transient overflow probability *does* decay exponentially.

Then, in Section 4 it is pointed out how to evaluate the decay rate under consideration. We further explain the intuitive meaning of all quantities involved, in terms of the most likely way the rare event occurs. As it turns out, the large-deviations based logarithmic asymptotics provide insight into the order of magnitude of the transient overflow probability, but are too crude to provide an accurate approximation. To remedy this, we propose in Section 5 an efficient importance-sampling based simulation algorithm, which we prove to be logarithmically efficient.

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 is 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 $(\nu_{ij})_{i,j=1}^d$, while $\hat\pi_i$ denotes the invariant distribution of the corresponding jump process. The time T_i spent in state i (often referred to as the *transition time*) has an exponential distribution with mean $1/\nu_i$, where $\nu_i := -\nu_{ii}$. Let

$$\pi_i := \frac{\hat\pi_i \mathbb{E}T_i}{\sum_{j=1}^d \hat\pi_j \mathbb{E}T_j} = \frac{\hat\pi_i/\nu_i}{\sum_{j=1}^d \hat\pi_j/\nu_j}$$

denote the stationary probability that the background process is in state i, for $i = 1,\dots,d$. In other words: the π_i s are nonnegative numbers summing up to 1, such that $\sum_{i=1}^d \pi_i\nu_{ij} = 0$.

While the process $(J(t))_{t\in\mathbb{R}}$, often referred to as the *background process* or *modulating process*, is in state i, jobs arrive according to a Poisson process with rate $\lambda_i \geq 0$. 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. We exclude the case that all λ_i s as well as the distributions of the B_i s coincide (as otherwise the queue is just an ordinary M/G/∞).

3. Large deviations

The main goal of the paper is to study the tail asymptotics of the number of jobs in the system at time t, given the system starts off empty at time 0. We do this by scaling the arrival rates by N, that is, we scale $\lambda_i \mapsto N\lambda_i$; at the same time we leave the timescale of the background process unchanged.

3.1. **Transient overflow probability.** The first objective is to study the logarithmic asymptotics of 'transient overflow probabilities' of the type

$$\mathbb{P}\left(M^{(N)}(t) > Na\right),$$

for a larger than the limiting value of $\mathbb{E}M^{(N)}(t)/N$, as well the corresponding 'transient underflow probabilities'. For the moment we assume that the background process is in equilibrium at time 0, that is, the probability that the background process' initial state is i equals π_i; the proof, however, will reveal that the initial distribution does not have any impact on the logarithmic asymptotics under consideration. The first observation [2] is that $M^{(N)}(t)$ has a Poisson distribution with a random parameter that depends on the evolution of the background process. This random parameter equals

$$N\int_0^t \lambda_{J(s)}\mathbb{P}\left(B_{J(s)} > t-s\right)\mathrm{d}s.$$

We found earlier [2] that $\mathbb{E}M^{(N)}(t) = N\varrho_t$, with

$$\varrho_t := \sum_{i=1}^{d}\pi_i\lambda_i\int_0^t \mathbb{P}(B_i > s)\mathrm{d}s.$$

We identify an a^+ (a^-, respectively) such that for all $a < a^+$ ($a > a^-$) the exponential decay rate of the above transient overflow (underflow) probability equals 0; the striking feature, however, is that a^+ *is strictly larger than* a^- (unless all λ_i as well as all the distributions of the B_i match).
In the following we omit, for notational convenience, the dependency on t of functions and variables. Let $P^{(N)}(f)$ denote a Poisson random variable with mean $N\kappa(f)$, where

$$\kappa(f) := \int_0^t g_s(f(s))\mathrm{d}s, \qquad g_s(i) := \lambda_i\mathbb{P}(B_i > t-s),$$

and $\mathscr{F} := \{f : [0,t] \mapsto \{1,\ldots,d\}\}$. Combining the above, we can write, in self-evident notation,

$$\mathbb{P}\left(M^{(N)}(t) > Na\right) = \int_{f\in\mathscr{F}}\mathbb{P}\left(P^{(N)}(f) > Na\right)\mathbb{P}(J(\cdot)\in\mathrm{d}f(\cdot)).$$

Define

$$f^+(s) := \arg\max_{i\in\{1,\ldots,d\}} g_s(i), \quad f^-(s) := \arg\min_{i\in\{1,\ldots,d\}} g_s(i),$$

and

$$d(f) := a - \kappa(f) - a\log\frac{a}{\kappa(f)}.$$

Let for the moment the service times B_i be exponentially distributed with parameter $\mu_i \in (0,\infty)$; below we comment on the case of more general service times.

Theorem 1. *For* $a \geq a^+ := \kappa(f^+)$,

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(M^{(N)}(t) > Na\right) = d(f^+).$$

For $a \leq a^- := \kappa(f^-)$,

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(M^{(N)}(t) < Na\right) = d(f^-).$$

We focus in the proof on the case that $a \geq a^+$; the case $a \leq a^-$ works analogously.

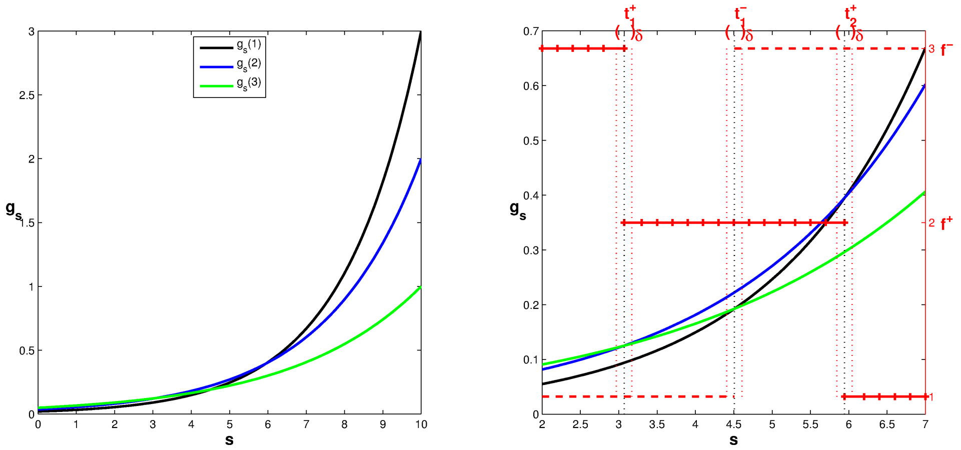

FIGURE 1. Graphical explanation of the nomenclature used in this section. (*Left panel*) The functions $g_s(i)$ for 3 different values of i for exponentially distributed B_i. (*Right panel*) Blow-up confirms that each pair of functions $g_s(i)$ has at most 1 intersection point and shows the g_s maximizing function f^+ (+++) with jump times t^+ and minimizing function f^- (---) with jump time t^-.

Lemma 1. *Both $f^+(\cdot)$ and $f^-(\cdot)$ are piecewise constant functions, taking values in $\{1,\dots,d\}$, that jump at most $d-1$ times in $[0,t]$.*

Proof. This is an immediate consequence of the fact that there is just at most a single s for which the functions

$$\lambda_i e^{-\mu_i(t-s)} \quad \text{and} \quad \lambda_j e^{-\mu_j(t-s)}$$

match (unless all λ_i and all μ_i coincide, but this we ruled out). See Figure 1 for an illustration in the case B_i is exponentially distributed. □

We are now in the position to prove Thm. 1.

Proof of Thm. 1. As indicated above, we concentrate on the transient overflow probability; the transient underflow probability can be dealt with analogously. We start by proving the lower bound. Define the jump epochs in $[0,t]$ corresponding to $f^+(\cdot)$, resulting from Lemma 1, by $t_1^+,\dots,t_k^+$, with $k<d$. Introduce the following set of functions that are 'close to' f^+:

$$\mathscr{F}_\delta := \left\{ f\in\mathscr{F} : f(s)=f^+(s) \text{ for all } s\in[0,t]\setminus\bigcup_{j=1}^{k}(t_j^+-\delta,t_j^++\delta) \right\};$$

choose $\delta>0$ sufficiently small that the intervals $(t_j^+-\delta,t_j^++\delta)$ do not overlap nor cover times 0 and t. Obviously,

$$\mathbb{P}\left(M^{(N)}(t)>Na\right) \geq \left(\min_{f\in\mathscr{F}_\delta}\mathbb{P}\left(P^{(N)}(f)>Na\right)\right)\mathbb{P}(J(\cdot)\in\mathscr{F}_\delta).$$

Notice that $\mathbb{P}(J(\cdot) \in \mathscr{F}_\delta)$ is strictly positive (use that the background process is irreducible), and in addition independent of N. As a consequence,

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)}(t) > Na\right) \geq \liminf_{N\to\infty} \frac{1}{N} \log\left(\min_{f\in\mathscr{F}_\delta} \mathbb{P}\left(P^{(N)}(f) > Na\right)\right).$$

Then observe that , due to Stirling's factorial approximation, if $a \geq \kappa(f)$, for any $\varepsilon > 0$ and N large enough,

$$\begin{aligned} \mathbb{P}\left(P^{(N)}(f) \geq Na\right) &= \sum_{k\geq Na} e^{-N\kappa(f)} \frac{(N\kappa(f))^k}{k!} \\ &\geq e^{-N\kappa(f)} \frac{(N\kappa(f))^{\lceil Na\rceil}}{\lceil Na\rceil!} \\ &\geq e^{Nd(f)} \frac{1-\varepsilon}{\sqrt{2\pi Na}}. \end{aligned}$$

Choose an arbitrary $f \in \mathscr{F}_\delta$. Then, with $\lambda^+ := \max_i \lambda_i$, using the definition of the set $\mathscr{F}_\delta$,

$$|\,\kappa(f) - \kappa(f^+)\,| \leq \left| \sum_{j=1}^{k} \int_{t_j^+-\delta}^{t_j^++\delta} \left(g_s(f(s)) - g_s(f^+(s))\right) \mathrm{d}s \right| \leq k \cdot 2\delta \cdot \lambda^+,$$

which goes to 0 as $\delta \downarrow 0$. As a consequence, also $|\,d(f) - d(f^+)\,|$ vanishes as $\delta \downarrow 0$. From this, we conclude that for $a \geq \kappa(f^+)$,

$$\liminf_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)}(t) > Na\right) \geq d(f^+).$$

The corresponding upper bound is less involved. Note that if $a > a^+$, then for all $f \in \mathscr{F}$ we have that $\mathbb{E}P^{(N)}(f) \leq Na$. Then observe the trivial inequality

$$\mathbb{P}\left(M^{(N)}(t) > Na\right) \leq \max_{f\in\mathscr{F}} \mathbb{P}\left(P^{(N)}(f) > Na\right).$$

Based on the Chernoff bound [4], we have

$$\mathbb{P}\left(P^{(N)}(f) > Na\right) \leq e^{Nd(f)}.$$

Combining the above, we obtain

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)}(t) > Na\right) \leq \max_{f\in\mathscr{F}} d(f).$$

As $\kappa(f^+)$ maximizes $\kappa(f)$, and $d(f)$ is increasing in $\kappa(f)$ (for $\kappa(f) \leq a$), we observe that

$$\limsup_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)}(t) > Na\right) \leq d(f^+).$$

This proves the upper bound. □

Corollary 1. *For all* $a \in [a^-, a^+]$,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)}(t) > Na\right) = \lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)}(t) < Na\right) = 0.$$

Remark 1. From the proof of Thm. 1 it becomes immediately clear that for $a < a^+$ the transient overflow probability $\mathbb{P}\left(M^{(N)}(t) > Na\right)$ can be bounded from below, uniformly in N, by a constant. Likewise, for $a > a^-$, the transient underflow probability $\mathbb{P}\left(M^{(N)}(t) < Na\right)$ can be bounded from below, uniformly in N by a constant.

Remark 2. We now comment on extending the result for exponential service times to more general distributions. It is readily verified that the claim remains true as long as the distributions of the B_is (for $i = 1, \dots, d$) are such that Lemma 1 applies, that is, that $f^+(\cdot)$ and $f^-(\cdot)$ jump only finitely often. For example, in the case of the B_i having a 'shifted Pareto' distribution, that is

$$\mathbb{P}(B_i > s) = (s+1)^{-\alpha_i},$$

for $\alpha_i > 0$ (in addition assuming that not all of them identical), the claim of Thm. 1 applies as well, as it does for all standard distributions.

Remark 3. Observe that the exponential decay rate of the transient overflow and underflow probabilities is fully determined by the arrival rates λ_i and the distributions of the service times B_i. In other words, the Markovian background process (with rates ν_{ij}) does *not* have any impact, which may look counterintuitive at first sight. The reason for this effect is reflected in the above proof, and can be further understood as follows. The most likely scenario is such that (i) the Markov chain $J(\cdot)$ essentially mimics the function $f^+(\cdot)$ (or $f^-(\cdot)$), while (ii) being in any of the states an unusually high number of jobs is generated, and an unusually low number of jobs leaves. Then observe that aspect (i) is not affected by the value N, while aspect (ii) is. As a result, the transition rates ν_{ij} do not appear in the asymptotics, while the arrival and departure rates do affect the decay rate.

Evidently, this 'insensitivity property' does not apply if the background process were scaled by N as well (that is, if the transition rates ν_{ij} were replaced by $N\nu_{ij}$). This scaling is substantially more complicated, and beyond the scope of the current work.

3.2. **Stationary overflow probability.** The above arguments can be used to analyze the stationary overflow probability as well. In this case, it is used that the stationary number of users in the system has a Poisson distribution with random mean [3]. This random parameter equals

$$N \int_{-\infty}^{0} \lambda_{J(s)} \mathbb{P}\left(B_{J(s)} > -s\right) \mathrm{d}s.$$

We focus on the case of exponentially distributed service times, but the result can be extended to more general random variables, analogously to the transient overflow probability.

To this end, we now use $\mathscr{F} := \{f : (-\infty, 0] \mapsto \{1, \dots, d\}\}$. Define $g_s(i) := \lambda_i e^{\mu_i s}$, for $s \leq 0$, and again

$$f^+(s) := \arg \max_{i \in \{1,\dots,d\}} g_s(i).$$

It is easily checked that the counterpart of Lemma 1 applies: $f^+(\cdot)$ jumps at most $d-1$ times in $(-\infty, 0]$; define the jump epochs $t_1^+, \dots, t_k^+$, with $k < d$. Let $P^{(N)}(f)$ be a Poisson random variable with mean $N\kappa(f)$, where

$$\kappa(f) := \int_{-\infty}^{0} g_s(i) \mathrm{d}s.$$

We define

$$\mathscr{F}_{\delta,T} := \left\{ f \in \mathscr{F} : f(s) = f^+(s) \text{ for all } s \in [-T, 0] \setminus \bigcup_{j=1}^{k} (t_j^+ - \delta, t_j^+ + \delta) \right\}.$$

Let $M^{(N)}$ be the stationary number of jobs in the system. Based on the above, we evidently have

$$\mathbb{P}\left(M^{(N)} > Na\right) = \int_{f \in \mathscr{F}} \mathbb{P}\left(P^{(N)}(f) > Na\right) \mathbb{P}(J(\cdot) \in \mathrm{d}f(\cdot)).$$

This identity can then be used to establish an upper and lower bound. The upper bound is as in the transient case. Regarding the lower bound, realize that, for any $f \in \mathscr{F}_{\delta,T}$,

$$|\,\kappa(f) - \kappa(f^+)\,| \leq \left| \sum_{j=1}^{k} \int_{t_j-\delta}^{t_j+\delta} \big(g_s(f(s)) - g_s(f^+(s))\big)\,\mathrm{d}s \right| + 2\int_{-\infty}^{T} \max_{i\in\{1,\dots,d\}} g_s(i)\mathrm{d}s,$$

which vanishes as $\delta \downarrow 0$ and $T \to -\infty$; the rest of the lower bound works as in the transient case. We have thus proved the following result; $f^-(\cdot)$ is defined in the obvious way.

Proposition 1. *For $a \geq a^+ := \kappa(f^+)$,*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)} > Na\right) = d(f^+).$$

For $a \leq a^- := \kappa(f^-)$,

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)} < Na\right) = d(f^-).$$

Example 1. We now consider the case of $d = 2$. Without loss of generality, we can restrict ourselves to two situations: (i) $\lambda_1 \geq \lambda_2$ and $\mu_1 \leq \mu_2$, and (ii) $\lambda_1 \geq \lambda_2$ and $\mu_1 > \mu_2$. In case (i) overflow is essentially caused by staying relatively long in state 1; it follows that

$$\kappa(f^+) = \frac{\lambda_1}{\mu_1}.$$

In case (ii) the background process jumps from state 2 to state 1 at time

$$t_1^+ = \frac{\log(\lambda_1/\lambda_2)}{\mu_2 - \mu_1} < 0.$$

It requires a bit of standard calculus to verify that

$$\kappa(f^+) = \frac{\lambda_1}{\mu_1} + \lambda_1^{\frac{\mu_2}{\mu_2-\mu_1}} \lambda_2^{-\frac{\mu_1}{\mu_2-\mu_1}} \left(\frac{1}{\mu_2} - \frac{1}{\mu_1}\right).$$

4. Computational issues and numerical examples

In this section we focus on the computational issues related to the results of the previous section; for ease we just consider just the transient overflow probability, for a larger than $\kappa(f^+)$.

Finding $f^+(\cdot)$. It is straightforward to compute t_1^+ up to t_k^+, as follows. Focusing on the (exponential) case that $\mathbb{P}(B_i > t - s) = e^{-\mu_i(t-s)}$, it is clear that

$$i_0 := f^+(0) = \arg \max_{i\in\{1,\dots,d\}} \lambda_i e^{-\mu_i t}.$$

Then compute the intersections s_j of $g_s(i_0)$ and $g_s(j)$ (for $j \neq i_0$):

$$s_j = t - \frac{\log \lambda_{i_0} - \log \lambda_j}{\mu_{i_0} - \mu_j}.$$

Pick the smallest positive one among these. If there is no positive intersection, or if the smallest positive intersection is larger than t, we are done (and $f^+(s) = i_0$ for all $s \in [0, t]$); otherwise $f^+(\cdot)$ jumps at $t_1^+ := \min s_j$, from i_0 to $i_1 := \arg\min s_j$ (where the minimum is taken over all $j \neq i_0$). Then we search for the smallest intersection larger than t_1^+ of $g_s(i_1)$ with $g_s(j)$ (with $j \neq i_0, i_1$), etc. This procedure is then repeated until t is reached. The shifted Pareto case mentioned above works completely analogously, with the intersection of the $g_s(i)$ and $g_s(j)$ given by

$$t + 1 - \left(\frac{\lambda_i}{\lambda_j}\right)^{(\alpha_i-\alpha_j)^{-1}}.$$

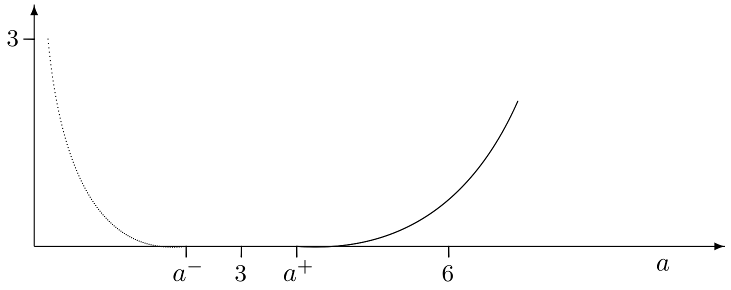

FIGURE 2. Decay rate of the transient overflow and underflow probability for the example in Section 4. Black line corresponds to overflow probability $-\lim_{N\to\infty} N^{-1}\log\mathbb{P}\left(M^{(N)}(t) > Na\right)$, gray line to underflow probability $-\lim_{N\to\infty} N^{-1}\log\mathbb{P}\left(M^{(N)}(t) < Na\right)$.

Computing the decay rates. Again we focus on the decay rate of the transient overflow probability. Put $t_0^+ := 0$ and $t_{k+1}^+ := t$. Then it is evident that

$$\kappa(f^+) = \sum_{\ell=0}^{k} \int_{t_\ell^+}^{t_{\ell+1}^+} \lambda_{i_\ell}\mathbb{P}(B_{i_\ell} > t-s)\mathrm{d}s.$$

In the exponential case, this can be rewritten to

$$\kappa(f^+) = \sum_{\ell=0}^{k} \frac{\lambda_{i_\ell}}{\mu_{i_\ell}} \left(e^{-\mu_{i_\ell}(t-t_{\ell+1}^+)} - e^{-\mu_{i_\ell}(t-t_\ell^+)}\right).$$

In the shifted Pareto case, we have

$$\kappa(f^+) = \sum_{\ell=0}^{k} \frac{\lambda_{i_\ell}}{\alpha_{i_\ell}-1} \left((t-t_{\ell+1}^+ + 1)^{-\alpha_{i_\ell}+1} - (t-t_\ell^+ + 1)^{-\alpha_{i_\ell}+1}\right).$$

Example. Consider the case of $d = 2$ in which the B_is (for $i = 1, 2$) are sampled from exponential distributions (with means $1/\mu_1$ and $1/\mu_2$, respectively). In this example we take $\lambda_1 = e$, $\lambda_2 = e^2$, $\mu_1 = 1$, $\mu_2 = 2$, and $t = 2$. It is readily verified that $k = 1$, $i_0 = 1$, $i_1 = 2$, and $t_1^+ = 1$. As a result

$$a^+ = \kappa(f^+) = e(e^{-1} - e^{-2}) + \frac{e^2}{2}\left(1 - e^{-2}\right) = \frac{e^2}{2} + \frac{1}{2} - \frac{1}{e} \approx 3.8266.$$

It means that the most likely scenario of overflow is that the modulating Markov chain starts in state 1 at time 0, and jumps to state 2 around time 1 (to stay there until $t = 2$).

In order to make sure that $M^{(N)}(t)$ is large, it pays off to spend time in states with (i) a high arrival rate, (ii) a low departure rate. *A priori* it is not clear how this would work out in the above example, as the state with the high (low, respectively) arrival rate is also the state with the high (low) departure rate. Naïvely, one could have thought that the most likely scenario would be to stay in the state i that maximizes the 'load' λ_i/μ_i (this would have been state 2), but apparently it is more likely that state 2 is entered only at time 1. To support this claim, with $f_{\{i\}}(\cdot)$ denoting the path that stays in state i during the entire interval $[0, t]$,

$$\kappa(f_{\{1\}}) = e - \frac{1}{e} = 2.3504, \quad \kappa(f_{\{2\}}) = \frac{1}{2}\left(e^2 - \frac{1}{e^2}\right) = 3.6269,$$

N	$a = 4.0$	$a = 4.2$	$a = 4.4$	$a = 4.6$	$a = 4.8$	$a = 5.0$
30	0.09263	0.11754	0.16085	0.20579	0.25822	0.31742
60	0.05538	0.07807	0.11332	0.15285	0.20024	0.25522
90	0.04130	0.06248	0.09418	0.13184	0.17650	0.23001
120	0.03417	0.05425	0.08374	0.12006	0.16396	0.21623
150	0.02938	0.04918	0.07648	0.11213	0.15590	0.20777
$-d(f^+)$	0.00387	0.01765	0.04096	0.07336	0.11446	0.16390

TABLE 1. $-N^{-1}\log\mathbb{P}\left(M^{(N)}(t) > Na\right)$ for different a and N; $\nu_{12} = \nu_{21} = 1$.

which are both lower than $\kappa(f^+)$ — this means that staying in state 1, or staying in state 2, the entire interval $[0, t]$ is less likely to cause overflow than jumping from 1 to 2 at time 1.
An elementary calculation reveals that $f^-(\cdot)$ jumps from state 2 tot state 1 at time $t_1^- = 1$, leading to

$$a^- = \kappa(f^-) = e - \frac{1}{2e^2} - \frac{1}{2} = 2.1506.$$

Recall that the above calculations are not affected by the values of the transition rates ν_{12} and ν_{21}. In Fig. 2 the resulting decay rates are shown.

5. IMPORTANCE SAMPLING

One could naïvely argue that our theory suggests the use of the approximation, for $a > a^+$,

$$\mathbb{P}\left(M^{(N)}(t) > Na\right) \approx e^{Nd(f^+)}.$$

As we know from the proof of Thm. 1, however, the decay rate $d(f^+)$ essentially follows from the most likely way the background process $J(\cdot)$ behaves to trigger the event that $M^{(N)}(t)$ exceeds Na. As a result, the above approximation $e^{Nd(f^+)}$ is likely to overestimate the transient overflow probability. Put differently, it is expected that $N^{-1}\log\mathbb{P}\left(M^{(N)}(t) > Na\right)$ converges slowly to $d(f^+)$.

To obtain a reliable numerical approximation, we propose to use an efficient simulation-based approach. A direct simulation would be time-consuming due to the rarity of the event under consideration. The idea is therefore to rely on importance sampling — sample under alternative probability measure $\mathbb{Q}$ rather than the actual probability measure $\mathbb{P}$, and then correct the simulation output by weighing it with appropriate likelihood ratios. The choice of our alternative measure $\mathbb{Q}$ guarantees favorable variance properties of the resulting estimator; in particular, we can prove that it is *logarithmically efficient* [1].
The proposed algorithm works as follows. Sample the path $J(\cdot)$ of the background process between 0 and t under the normal measure (which we could denote by $\mathbb{P}$). If it has the realization $f(\cdot)$, we sample the random variable $M^{(N)}(t)$ not from a Poisson distribution with mean $N\kappa(f)$ (which would have been the case under the original measure $\mathbb{P}$), but rather from a Poisson distribution with mean Na. As a consequence, the event under consideration is not rare anymore (it has roughly probability $\frac{1}{2}$ under $\mathbb{Q}$). If the Poisson random variable has the value k, it is trivial to verify that the corresponding likelihood ratio L_N equals

$$L_N = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}} = e^{Na-N\kappa(f)}\left(\frac{\kappa(f)}{a}\right)^k.$$

N	$a = 4.0$	$a = 4.2$	$a = 4.4$	$a = 4.6$	$a = 4.8$	$a = 5.0$
30	0.11809	0.14424	0.18873	0.23462	0.28732	0.34763
60	0.06897	0.09233	0.12774	0.16840	0.21602	0.27130
90	0.05111	0.07260	0.10447	0.14242	0.18855	0.24159
120	0.04145	0.06180	0.09172	0.12847	0.17313	0.22543
150	0.03549	0.05557	0.08340	0.11947	0.16345	0.21518
$-d(f^+)$	0.00387	0.01765	0.04096	0.07336	0.11446	0.16390

TABLE 2. $-N^{-1}\log\mathbb{P}\left(M^{(N)}(t) > Na\right)$ for different a and N; $\nu_{12} = 1$ and $\nu_{21} = 2$.

We thus obtain, with I denoting the event that the number of jobs at time t exceeds Na and writing $\bar{L}_N := L_N I$, that

$$\mathbb{E}_{\mathbb{Q}}(\bar{L}_N) = \mathbb{P}\left(M^{(N)}(t) > Na\right).$$

As a consequence, by replicating the above experiment multiple times, we can unbiasedly estimate the probability of our interest by the sample mean of independent realizations of $\bar{L}_N$.

We now study the variance performance of our estimator, showing that it is logarithmically efficient. It is immediate that, as variances are nonnegative, the decay rate of $\mathbb{E}_{\mathbb{Q}}\bar{L}_N^2$ is larger than or equal to twice the decay rate of $\mathbb{E}_{\mathbb{Q}}\bar{L}_N$, which equals $2d(f^+)$. Now we show that this lower bound is actually achieved, and that our estimator therefore has certain optimality properties. To this end, observe that, using that $\kappa(f) \leq a$ for all $f \in \mathscr{F}$, and relying on the definition of $f^+(\cdot)$,

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\bar{L}_N^2 = \mathbb{E}_{\mathbb{Q}}L_N^2 I &\leq \int_{f\in\mathscr{F}} e^{2Na-2N\kappa(f)}\left(\frac{\kappa(f)}{a}\right)^{2Na}\mathbb{P}(J(\cdot)\in \mathrm{d}f(\cdot))\\ &= \int_{f\in\mathscr{F}} e^{2Nd(f)}\mathbb{P}(J(\cdot)\in \mathrm{d}f(\cdot))\\ &\leq e^{2Nd(f^+)}\int_{f\in\mathscr{F}}\mathbb{P}(J(\cdot)\in \mathrm{d}f(\cdot)) = e^{2Nd(f^+)}.\end{aligned}$$

Now standard arguments [1] show that the proposed procedure is logarithmically efficient.

Proposition 2. *The proposed algorithm is logarithmically efficient.*

We end this paper by reporting on a number of simulation experiments. In all these experiments, we generated 'good' estimates (that is, with 95% confidence, and the width of the confidence interval below 10% of the estimate) well within a second on a standard PC. In the first experiment, we took the parameters of the numerical example of Section 4.

Table 1 confirms the slow of convergence of $N^{-1}\log\mathbb{P}\left(M^{(N)}(t) > Na\right)$ to $d(f^+)$, the deviation is large in particular in scenarios in which the event of interest is still relatively likely (a close to a^+, that is). This supports the use of our importance sampling algorithm, rather than relying on the (very crude) approximation $\exp(Nd(f^+))$.

In Table 2 we give results for the case that the value of ν_{21} is changed into 2. Our theory says that this does not affect the decay rate of the transient overflow probability. The simulation output confirms the convergence of $N^{-1}\log\mathbb{P}\left(M^{(N)}(t) > Na\right)$ to its limiting value.

REFERENCES

[1] S. ASMUSSEN and P. GLYNN (2007). *Stochastic Simulation*. Springer, New York.

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

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

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

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

[6] P. GLYNN (1995). Large deviations for the infinite server queue in heavy traffic. *Institute for Mathematics and Its Applications*, **71**, 387–394.

[7] P. GLYNN AND W. WHITT (1991). A new view of the heavy-traffic limit theorem for infinite-server queues. *Advances in Applied Probability*, **23**, 188–209.

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

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

[10] M. MANDJES and A. RIDDER (2001). A large deviations approach to the transient of the Erlang loss model. *Performance Evaluation*, **43**, pp. 181–198.

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

[12] A. SCHWABE, K. RYBAKOVA, and F. BRUGGEMAN (2012). Transcription Stochasticity of Complex Gene Regulation Models. *Biophysical Journal*, **103**, pp. 1152-1161.

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

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

E-mail address: `joke.blom@cwi.nl, M.R.H.Mandjes@uva.nl`