TAIL ASYMPTOTICS OF A MARKOV-MODULATED INFINITE-SERVER QUEUE

J. BLOM [⋆], K. DE TURCK [†], O. KELLA [+], M. MANDJES [•,⋆]

ABSTRACT. This paper analyzes large deviation probabilities related to the number of customers in a Markov modulated infinite-server queue, with state-dependent arrival and service rates. Two specific scalings are studied: in the first, just the arrival rates are linearly scaled by N (for large N), whereas in the second in addition the Markovian background process is sped up by a factor $N^{1+\epsilon}$, for some $\epsilon > 0$. In both regimes, (transient and stationary) tail probabilities decay essentially exponentially, where the associated decay rate corresponds to that of the probability that the sample mean of i.i.d. Poisson random variables attains an atypical value.

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*. He is also a Postdoctoral Fellow of the same foundation.
The work of O. Kella is partially supported by Israel Science Foundation grant No. 1462/13 and The Vigevani Chair in Statistics.
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.

[⋆] CWI, P.O. Box 94079, 1090 GB Amsterdam, the Netherlands.
[†] TELIN, Ghent University, St.-Pietersnieuwstraat 41, B9000 Gent, Belgium.
[+] Department of Statistics, The Hebrew University of Jerusalem, Jerusalem 91905, Israel.
[•] Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, the Netherlands.

Date: December 6, 2013.

1. Introduction

In [1, 4] we have considered the tail asymptotics of the Markov-modulated infinite-server queue: under two specific scalings large-deviations results were derived for the probability that the number of jobs in the system attains a given (atypical) value. In these queueing systems, both arrival and service processes depend on the state of an external, independently evolving finite-state Markov chain, referred to as the *modulating process;* the feature that there are infinitely many servers entails that customers are served in parallel.
As remarked in [5], however, such Markov-modulated infinite-server queues come in *two* flavors: one in which the service times are sampled upon arrival, and one in which the departure rate at a given point in time depends on the current state of the modulating process. Importantly, the large deviations results in [1, 4] relate to the former model; for the latter model such large deviations asymptotics have not been derived so far, to the best of our knowledge. The primary objective of this paper is to identify these asymptotics.

The model considered in this paper can be specified in greater detail as follows. Let $J(\cdot)$ be an irreducible continuous-time Markov chain, on a finite-state space $\{1,\ldots,d\}$, with transition rate matrix Q and stationary distribution vector $\boldsymbol{\pi}$ (where we follow the convention that vectors are written in bold). When this modulating process, sometimes called the *background process*, is in state i, jobs arrive according to a Poisson process with rate $\lambda_i \geq 0$. In the context of our previous work [1, 4], the service times were sampled upon arrival: if the state of $J(\cdot)$ is i when the job arrives, then the service time is sampled from an exponential distribution with mean $1/\mu_i$ (where it is noticed that the results could be extended to a setting with general state-dependent distributions). In the present paper, however, we consider the model in which the hazard rate of leaving the system at a given point in time, say t, is μ_i if $J(t) = i$. In our model there are infinitely-many servers: there is no waiting. We denote by $M(t)$ the number of customers in the system at time $t \geq 0$; we assume the system being empty at time 0.
The difference between the two models is reflected in a very insightful manner as follows. It was observed in [6] that in the model of [1, 4] the number of customers in the system at time t has a Poisson distribution with *random* parameter

$$\int_0^t \lambda_{J(s)} e^{-\mu_{J(s)}(t-s)} \mathrm{d}s. \tag{1}$$

This property can be intuitively understood by realizing that $e^{-\mu_i(t-s)}$ can be interpreted as the probability that a customer arriving at time $s \in [0, t)$ while the background process is in state i, is still present at time t. For the model to be considered in the present paper, a similar representation is valid: $M(t)$ has again a Poisson distribution, but now with random parameter

$$\int_0^t \lambda_{J(s)} e^{-\int_s^t \mu_{J(r)} \mathrm{d}r} \mathrm{d}s. \tag{2}$$

Observe how the state-dependent departure rate is incorporated in this expression: now $\exp(-\int_s^t \mu_{J(r)}\mathrm{d}r)$ represents the probability that a customer arriving at time $s \in [0,t)$ is still present at time t. As it will turn out, the representation (2) will enable us to derive the large deviations asymptotics that we are aiming for.

The literature on Markov-modulated infinite-server queues is surprisingly small (compared to the literature on Markov-modulated single-server queues); we mention a number of key papers here. O'Cinneide and Purdue [10] provide explicit expressions for the moments of the stationary number of customers, and systems of partial differential equations for the corresponding transient moments, in the context of the model variant studied in the present paper (with state-dependent hazard rate, that is). Related results, for a considerably broader class of models, are given in [9]; we also mention [8] for extensions to a semi-Markovian background process. As mentioned above, [6] presents the useful observation that $M(t)$ has a Poisson law with a random parameter (that depends on the path of the background process in $[0,t]$), as was highlighted in (1) and (2).
In [2, 3] the arrival rates are scaled by N (to become $N\lambda_i$ when $J(\cdot)$ is in state i), while the background process is sped up by a factor N^α. For both model variants introduced above, central-limit type results are derived. A crucial finding is that results in which the background process is faster than the arrival process ($\alpha > 1$, that is) are intrinsically different from those in which the background process is slower ($\alpha < 1$). A similar dichotomy applies in the large-deviations domain for the model in which the service times are sampled upon arrival, corresponding to representation (1); [4] covers the case of a slowly moving background process and [1] the case of a fast background process. The results presented in the present paper show that these qualitative findings carry over to the model variant in which the departure rates depend on the current state of the background process, that is, the variant corresponding to representation (2).

The organization and contributions of this paper are as follows.

- In Section 2 we consider the counterpart of [4]: we study the regime in which only the arrival rates are scaled by a factor N, for N large. It turns out that, with $M^{(N)}(t)$ the number of customers in the system in the N-scaled model, the tail probabilities of $M^{(N)}(t)$ decay exponentially, where the corresponding decay rate is the solution of a specific optimization problem. This optimization problem lends itself to a non-trivial explicit solution, in terms of a closed-form expression for the most likely path followed by the background process in order for the number of customers to reach a high value. Given this explicit result, the large deviations asymptotics follow from a proof that resembles the one in [4].
- Section 3 addresses the counterpart of [1]: we scale the arrival rates by N, but the transition rates of the background process by $N^{1+\epsilon}$ for $\epsilon > 0$. Defining

$$\lambda_\infty := \sum_{i=1}^{d} \pi_i \lambda_i, \quad \mu_\infty := \sum_{i=1}^{d} \pi_i \mu_i, \tag{3}$$

the main intuition is that in this scaling, as N tends to ∞, the arrival process becomes essentially Poisson (with rate $N\lambda_\infty$), while the service times become exponentially distributed with a uniform service rate (with mean μ_∞^{-1}), so that the system behaves as an M/M/∞ queue with these parameters. This explains why the large deviations of the (transient and stationary) number of customers in the system are those of the sample mean of i.i.d. Poisson random variables.

2. SLOW TIMESCALE REGIME

In this section, we consider the regime in which the arrival rates λ_i, for $i = 1, \dots, d$ are scaled by N, whereas the generator matrix Q remains unchanged. In Section 2.1 we prove a number of structural properties related to the maximum (and minimum) value that can be attained by the random parameter of the Poisson distribution, cf. representation (2). These results are then used in Section 2.2 when establishing large deviations results.

2.1. **Maximum value attained by Poisson parameter.**

The objective of this section is to find a path $f^+(\cdot)$ for $J(s), 0 \le s \le t$ that maximizes the (random) parameter of the Poisson distribution (2). Let $\mathscr{F}_t$ denote the class of Borel functions $f : [0,t] \mapsto \{1, \dots, d\}$ and, for a given $f \in \mathscr{F}_t$, denote the Poisson parameter by:

$$\kappa_t(f) := \int_0^t \lambda_{f(s)} e^{-\int_s^t \mu_{f(r)}\mathrm{d}r}\mathrm{d}s. \tag{4}$$

We thus want to solve the following *optimization problem*:

$$\sup_{f \in \mathscr{F}_t} \kappa_t(f) \equiv \kappa_t^+ \tag{P}$$

and we seek a maximizing path f^+ satisfying

$$\kappa_t(f^+) = \kappa_t^+.$$

As it turns out in Section 2.2, such a path, which will be shown to exist and is Lebesgue almost-surely unique, plays a crucial role when determining the large deviation asymptotics of the number of customers in the system in our N-scaled model. We will also point out how to identify a path $f^-(\cdot)$ that *minimizes* $\kappa_t(f)$.

In preparation to analyzing the optimization problem (P), denote $\varrho_i := \lambda_i/\mu_i$. An important role is played by an index i^+ (not necessarily unique) that satisfies $\varrho_{i^+} = \varrho^+ := \max_{i \in \{1,\dots,d\}} \varrho_i$.

Lemma 1. *The following claims hold:*

1. *For every $f \in \mathscr{F}_t$,*

$$\kappa_t(f) \le \varrho^+ \left(1 - e^{-\int_0^t \mu_{f(r)}\mathrm{d}r}\right) < \varrho^+.$$

2. *For any $t \ge 0$,*

$$\kappa_t^+ \ge \varrho^+ \left(1 - e^{-\mu_{i^+} t}\right).$$

Proof: Claim 1 is an immediate consequence of

$$\kappa_t(f) = \int_0^t \varrho_{f(s)}\mu_{f(s)} e^{-\int_s^t \mu_{f(r)}\mathrm{d}r}\mathrm{d}s \le \varrho^+ \int_0^t \mu_{f(s)} e^{-\int_s^t \mu_{f(r)}\mathrm{d}r}\mathrm{d}s = \varrho^+ \left(1 - e^{-\int_0^t \mu_{f(r)}\mathrm{d}r}\right)$$

whereas Claim 2 follows from considering the constant function $f(s) = i^+$ for $s \in [0,t]$, so that:

$$\kappa_t^+ \ge \int_0^t \lambda_{i^+} e^{-\mu_{i^+}(t-s)}\mathrm{d}s = \varrho^+ \left(1 - e^{-\mu_{i^+}t}\right) .$$

This proves the claims. □

In the following, it will prove useful to represent the elements of the set of combinations of arrival rates and service rates, i.e., $\{(\mu_i, \lambda_i) \,|\, i \in \{1,\dots,d\}\}$, as points in the (μ,λ)-plane. Also, we define, for any i, j such that $\mu_i \neq \mu_j$,

$$\gamma(i,j) := \frac{\lambda_i - \lambda_j}{\mu_i - \mu_j},$$

denoting the slope between the points (μ_i, λ_i) and (μ_j, λ_j) in the (μ,λ)-plane. Now consider the following algorithm in pseudocode to find ρ^+, i.e., the maximum slope between the origin and any (μ_i, λ_i).

Algorithm 1.
```
Input:  $\lambda_i$, $\mu_i$, $i \in \{1,\dots,d\}$ -- Output:  $I_0,\dots,I_k$ and $i_1,\dots,i_k$.
  0. Set $\ell := 1$, $I_0 := 0$, $\mathscr{A}_0 := \{1,\dots,d\}$, and
         $i_1 := \arg\min\left\{\mu_i : \lambda_i = \max_{j \in \mathscr{A}_0} \lambda_j\right\}$ .
  1. Let
         $\mathscr{A}_\ell := \{i : I_{\ell-1} < \gamma(i,i_\ell) < \varrho_{i_\ell}$ and $\mu_i < \mu_{i_\ell}\}$ .
     If $\mathscr{A}_\ell$ is empty go to step (3), otherwise, go to step (2).
  2. Let
         $I_\ell := \min_{j \in A_\ell} \gamma(j,i_\ell)$,   $i_{\ell+1} := \arg\min_{i \in \mathscr{A}_\ell} \{\mu_i : \gamma(i,i_\ell) = I_\ell\}$ .
     Set $\ell := \ell + 1$ and return to step (1).
  3. Set $I_\ell := \varrho_{i_\ell}$ and STOP.
```

To aid in understanding, see Fig. 1 for an example with $d = 7$ and $k = 2$. Note that i_1 is the index of the node with the largest λ_i and i_2 is that of the node with the largest ϱ_i ($= \varrho^+$). I_0 is the slope zero, I_1 is the slope of the segment connecting $(\mu_{i_1}, \lambda_{i_1})$ and $(\mu_{i_2}, \lambda_{i_2})$, and $I_2 = \varrho^+$. Note that in this case there are two indices j for which $I_1 = \gamma(i_1, j)$, and that i_2 is the one having the smaller value of μ_j. Also note that there is a point (close to (0,0)) that if we connect a segment between it and i_2 the slope is greater than I_1. However, since this slope is also greater than $\varrho_{i_2} = \varrho^+$, the algorithm stops at $\ell = 2$. The bold segments describe a concave function, which is the reason why ϱ_ℓ, the slopes of the segment connecting $(0,0)$ and $(\mu_{i_\ell}, \lambda_{i_\ell})$, increases in ℓ and in particular when the algorithm stops then $\varrho_{i_\ell} = \varrho^+$. Fig. 2 demonstrates that the maximal value of $\lambda_i - c\mu_i$ is given by i_1 for $I_0 < c < I_1$ and $(\mu_i, \lambda_i) \in [\mu_{i_2}, \mu_{i_1}] \times [\lambda_{i_2}, \lambda_{i_1}]$ and by i_2 for $I_1 < c < I_2$ and $(\mu_i, \lambda_i) \in [0, \mu_{i_2}] \times [0, \lambda_{i_2}]$.

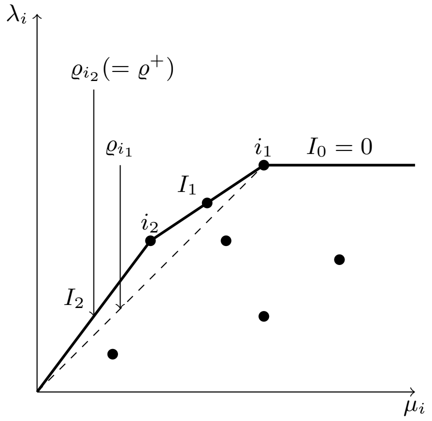

Fig. 1: Example with $d = 7$ and $k = 2$.

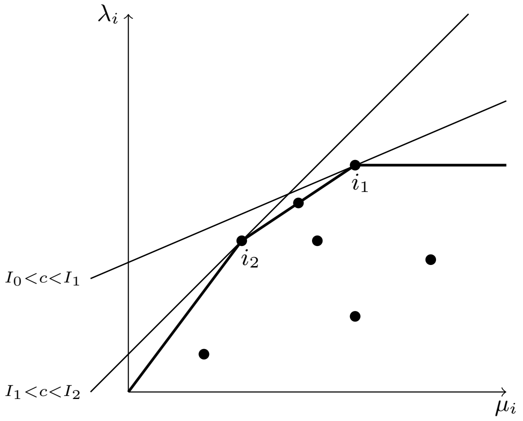

Fig. 2: Maximal value of $\lambda_i - c\mu_i$.

Algorithm 1 is one possible organized method of finding a minimal nonnegative, nondecreasing, concave function g such that that $g(\mu_i) \geq \lambda_i$ for all $i \in \{1, \ldots, d\}$. This function is unique (the infimum of concave functions is concave); $i_1, \ldots, i_k$ are the indices for which $g(\mu_i) = \lambda_i$, that is, the indices of the extreme points of the hypograph.
For any $0 < v < \infty$ there is some supergradient $c(v) \in [0, \rho^+]$, such that for each $0 \leq u < \infty$, $g(u) - g(v) \leq c(v)(u - v)$, that is, $g(u) - c(v)u \leq g(v) - c(v)v$. In particular, if we take $v = \mu_{i_\ell}$, then $g(v) = \lambda_{i_\ell}$ and a supergradient is any $c \in [I_{\ell-1}, I_\ell]$ which gives for each i,

$$\lambda_i - c\mu_i \leq g(\mu_i) - c\mu_i \leq g(\mu_{i_\ell}) - c\mu_{i_\ell} = \lambda_{i_\ell} - c\mu_{i_\ell}. \tag{5}$$

We thus have the following lemma.

Lemma 2. *The output of Algorithm 1 is a sequence of different states $i_1, \ldots, i_k$ and values $0 = I_0 < \ldots < I_k = \varrho^+$ such that for every $0 \leq c < \varrho^+$ we have that*

$$\max_{i \in \{1,\ldots,d\}} (\lambda_i - c\mu_i) = \sum_{\ell=1}^{k} (\lambda_{i_\ell} - c\mu_{i_\ell})\, 1_{[I_{\ell-1}, I_\ell)}(c)\ . \tag{6}$$

In an identical manner, for the purpose of minimization (rather than maximization), we need to find the maximal nondecreasing convex function h such that $h(\mu_i) \leq \lambda_i$ for each i with maximal subgradient $\varrho^- = \min_{i \in \{1,\ldots,d\}} \varrho_i$. In the maximizing case, the restriction on the maximal subgradient was automatically satisfied by the assumption that the function must be nonnegative and thus $g(0) = 0$, which implies that the maximal supergradient is ϱ^+ (the slope of the segment that connects $(0, 0)$ with $(\mu_{i_k}, \lambda_{i_k})$).

Algorithm 2. `Input:` λ_i, μ_i, $i \in \{1, \ldots, d\}$ `-- Output:` $I_0, \ldots, I_k$ `and` $i_1, \ldots, i_k$.

```
0. Set $\ell := 1$, $I_0 := 0$, $\mathscr{A}_0 := \{1, \ldots, d\}$, and
```

$$i_1 := \arg\max \left\{ \mu_i : \lambda_i = \min_{j \in \mathscr{A}_0} \lambda_j \right\}.$$

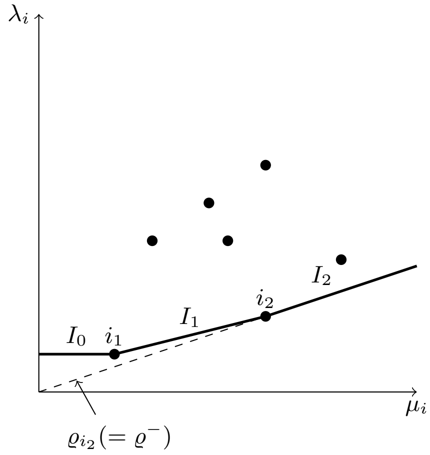

Fig. 3: Example with $d = 7$ and $k = 2$.

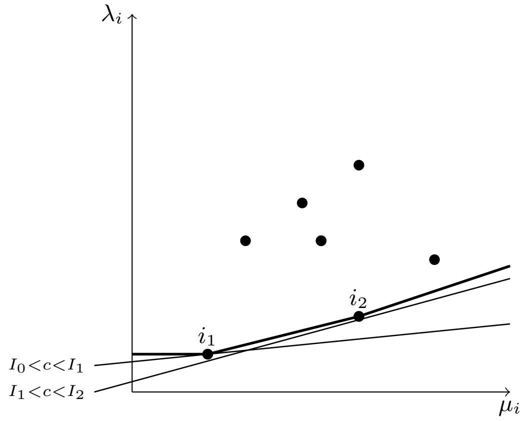

Fig. 4: Minimal value of $\lambda_i - c\mu_i$.

```
1. Let
       $\mathscr{A}_\ell := \{i : I_{\ell-1} < \gamma(i, i_\ell) < \varrho_{i_\ell} \text{ and } \mu_i > \mu_{i_\ell}\}$.
   If $\mathscr{A}_\ell$ is empty go to step (3), otherwise, go to step (2).
2. Let
       $I_\ell := \min_{j\in\mathscr{A}_\ell} \gamma(j, i_\ell), \quad i_{\ell+1} := \arg\max_{i\in\mathscr{A}_\ell} \{\mu_i : \gamma(i, i_\ell) = I_\ell\}$.
   Set $\ell := \ell + 1$ and return to step (1).
3. Set $I_\ell := \varrho_{i_\ell}$ and STOP.
```

The corresponding figures are Figs. 3 and 4. Therefore, we also have the following lemma.

Lemma 3. *The output of Algorithm 2 is a sequence of different states $i_1, \dots, i_k$ and values $0 = I_0 < \dots < I_k = \varrho^-$ such that for every $0 \le c < \varrho^-$ we have that*

$$\min_{i\in\{1,\dots,d\}} (\lambda_i - c\mu_i) = \sum_{\ell=1}^{k} (\lambda_{i_\ell} - c\mu_{i_\ell})\, 1_{[I_{\ell-1},I_\ell)}(c)\,. \tag{7}$$

We are now ready to state the main result of this subsection, presenting the maximizing path explicitly; later we also point out how the corresponding minimizing path can be constructed. We first introduce some notation. Let the sequence $i_1, \dots, i_k$ be as in Lemma 2. Let the time epochs $t_0^+, \dots, t_{k-1}^+$ be defined recursively through $t_0^+ := 0$, $t_k^+ := \infty$ and, for $\ell \in \{1, \dots, k-1\}$,

$$t_\ell^+ := t_{\ell-1}^+ + \frac{1}{\mu_{i_\ell}} \log \frac{\varrho_{i_\ell} - I_{\ell-1}}{\varrho_{i_\ell} - I_\ell}; \tag{8}$$

Also, set $f^+(s) = i_\ell$ for $s \in [t_{\ell-1}^+, t_\ell^+)$ and $\ell \in \{1, \dots, k\}$.

Theorem 1. *f^+ is optimal for* (P) *for each $t \ge 0$. If $t_{\ell-1}^+ \le t < t_\ell^+$ then the optimal value is given by*

$$\kappa_t^+ = \kappa_t(f^+) = I_{\ell-1} e^{-\mu_{i_\ell}(t-t_{\ell-1}^+)} + \varrho_{i_\ell} \left(1 - e^{-\mu_{i_\ell}(t-t_{\ell-1}^+)}\right)\,. \tag{9}$$

Proof: We observe that

$$\kappa_t(f) = \int_0^t (\lambda_{f(s)} - \mu_{f(s)}\kappa_s(f))\mathrm{d}s, \tag{10}$$

since the derivatives of the integral above and (4) are equal and $\kappa_t(0) = 0$. Thus (P) is equivalent to an optimal control problem (P′) of the form

$$\begin{cases} \text{maximize} & x(t) \\ \text{subject to:} & x'(s) = \lambda_{f(s)} - \mu_{f(s)}x(s),\ s \in [0,t], \\ & x(0) = 0, \\ & f \in \mathscr{F}_t. \end{cases} \tag{P$'$}$$

The Hamiltonian — as used in optimal control theory — for (P′) is $H(x,p,f) = p(\lambda_f - \mu_f x)$ where $p(s) = e^{-\int_s^t \mu_{f(r)}\,\mathrm{d}r}$ is the unique solution of

$$p'(s) = -\partial_x H(x,p,f) = \mu_{f(s)}p(s)$$

and $p(t) = 1$. As p is positive, f^+ maximizes $H(x,p,f)$ if and only if it maximizes $\lambda_f - \mu_f x$. As a consequence, the Pontryagin maximum principle hints at the guess that an optimal solution should satisfy for each $s \in [0,t]$:

$$\lambda_{f^+(s)} - \mu_{f^+(s)}\kappa_s^+ = \max_{i\in\{1,\ldots,d\}}(\lambda_i - \mu_i\kappa_s^+).$$

Assuming that this is the correct guess, and recalling that $i^+ = \arg\max \varrho_i$, we have that

$$\max_{i\in\{1,\ldots,d\}}(\lambda_i - \mu_i\kappa_s^+) \geq \lambda_{i^+} - \mu_{i^+}\kappa_s^+ = \mu_{i^+}(\varrho^+ - \kappa_s^+) > 0$$

and it follows by combining part 1 of Lemma 1 with Eqn. (10) that κ_s^+ is strictly increasing, continuous, with $\kappa_0^+ = 0$ and $\kappa_s^+ \to \varrho^+$ as $s \to \infty$.
By Lemma 2, for every $\ell \in \{1,\ldots,k\}$ and s such that $I_{\ell-1} \leq \kappa_s^+ < I_\ell$, we have that

$$\arg\max_{i\in\{1,\ldots,d\}}(\lambda_i - \kappa_s^+\mu_i) = i_\ell,$$

and in order to describe the optimal control and the value of κ_t^+ (the optimal value of $\kappa_t(f)$), for each $t \geq 0$, it remains to find $t_0^+ < \ldots < t_k^+$ such that $I_{\ell-1} \leq \kappa_s^+ < I_\ell$ if and only if $t_{\ell-1}^+ \leq s < t_\ell^+$. For this purpose it is straightforward to show that

$$I_\ell = I_{\ell-1}e^{-\mu_{i_\ell}(t_\ell^+ - t_{\ell-1}^+)} + \varrho_{i_\ell}\left(1 - e^{-\mu_{i_\ell}(t_\ell^+ - t_{\ell-1}^+)}\right) \tag{11}$$

and some simple manipulations result in (8). In particular, for $\ell = k$ the equality in (11) can be achieved only with $t_k^+ = \infty$. For $I_{\ell-1} \leq t < I_\ell$, replacing I_ℓ on the left by κ_t^+ and t_ℓ^+ on the right by t, results in (9).
To complete the proof, we need to show that our guess is indeed the correct one. If not, then there would be some choice of $f \in \mathscr{F}_t$ such that $\kappa_t(f) > \kappa_t^+$. Note that both κ_t^+ and $\kappa_t(f)$ are (absolutely) continuous functions of t, and satisfy $\kappa_0^+ = \kappa_0(f) = 0$. Now introduce $\tau_t(f) := \sup\{s : s \leq t,\ \kappa_s^+ = \kappa_s(f)\}$. By continuity it follows that $\kappa_{\tau_t(f)}^+ =$

$\kappa_{\tau_t(f)}(f)$, with $\tau_t(f) < t$, and in addition that $\kappa_s^+ < \kappa_s(f)$ for each $\tau_t(f) < s \leq t$. Hence, for any $\tau_t(f) < s \leq t$,

$$\lambda_{f(s)} - \mu_{f(s)}\kappa_s(f) \leq \max_{i\in\{1,\dots,d\}}(\lambda_i - \mu_i\kappa_s(f)) < \max_{i\in\{1,\dots,d\}}(\lambda_i - \mu_i\kappa_s^+) = \lambda_{f^+(s)} - \mu_{f^+(s)}\kappa_s^+$$

which implies that

$$\begin{aligned}\kappa_t(f) - \kappa_{\tau_t(f)}(f) &= \int_{\tau_t(f)}^t (\lambda_{f(s)} - \mu_{f(s)}\kappa_s(f))\mathrm{d}s \\ &< \int_{\tau_t(f)}^t \left(\lambda_{f^+(s)} - \mu_{f^+(s)}\kappa_s^+\right)\mathrm{d}s = \kappa_t^+ - \kappa_{\tau_t(f)}^+,\end{aligned}$$

and since $\kappa_{\tau_t(f)}(f) = \kappa^+_{\tau_t(f)}$ we have that $\kappa_t(f) < \kappa_t^+$, a contradiction. Thus, $\kappa_t(f) \leq \kappa_t^+ = \kappa_t(f^+)$ for every $f \in \mathscr{F}_t$ (and every $t \geq 0$), and conclude that f^+ is indeed optimal. □

The corresponding minimization (rather than maximization) problem can be dealt with analogously. In Thm. 1 we should now take the sequence $i_1, \dots, i_k$ as in Lemma 3 (that is, the output of Algorithm 2). Let f^- be a minimizing path, which is, like f^+, Lebesgue almost-surely unique.

2.2. **Large deviations results.** We have already noticed that $M^{(N)}(t)$ has a Poisson distribution with (random) parameter $N\kappa(J)$, with $J \equiv (J(s))_{s\in[0,t]}$ the path of the background process. Below we identify two numbers a^+ and a^- 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^-. To keep the notation transparent, we suppress the dependence on t of functions and variables. Let $P^{(N)}(f)$ denote a Poisson random variable with mean $N\kappa(f)$, and let $\mathscr{F}_t$ be as defined before. Combining the above, we can write, in self-evident notation,

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

Define

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

For f^+ and f^-, the following lemma is an immediate consequence of Thm. 1 and its minimization counterpart.

Lemma 4. *Both $f^+(\cdot)$ and $f^-(\cdot)$ are piecewise constant functions, taking values in $\{1,\dots,d\}$, that jump at most $d-1$ times in $[0,t]$.*

We now state and prove the main result of this subsection.

Theorem 2. *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 $a \leq a^- := \kappa(f^-)$,

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

Proof: Although the proof is similar to that of [4, Thm. 1], we include it here for the sake of completeness and readability. We focus in the proof on the case that $a \geq a^+$; the case $a \leq a^-$ works analogously.

▷ We start by proving the lower bound. Recall that the jump epochs in $[0,t]$ corresponding to f^+, resulting from Lemma 4, are denoted by $t_1^+,\dots,t_k^+$, with $k < d$. Introduce the following set of functions that are 'close to' f^+ (i.e., equal to f^+, apart from 'small' intervals around the t_j^+, $j=1,\dots,k$):

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

choose $\delta > 0$ sufficiently small that the intervals $(t_j^+ - \delta, t_j^+ + \delta)$ do not overlap nor cover times 0 and t. Consider the following obvious lower bound:

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

Now it is realized that $\mathbb{P}(J(\cdot) \in \mathscr{F}_{t,\delta})$ is strictly positive (where it is used that $J(\cdot)$ is an irreducible Markov chain on a finite state space), and in addition independent of N. This entails that

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

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 $f \in \mathscr{F}_{t,\delta}$. Then define $\lambda^+ := \max_i \lambda_i$, and $\mu^+ := \max_i \mu_i$. Using the triangle inequality, it is immediate that $|\,\kappa(f) - \kappa(f^+)\,|$ is majorized by

$$\begin{aligned}&\left|\int_0^t \lambda_{f(s)} e^{-\int_s^t \mu_{f(r)}\mathrm{d}r}\mathrm{d}s - \int_0^t \lambda_{f(s)} e^{-\int_s^t \mu_{f^+(r)}\mathrm{d}r}\mathrm{d}s\right| \\ &\quad + \left|\int_0^t \lambda_{f(s)} e^{-\int_s^t \mu_{f^+(r)}\mathrm{d}r}\mathrm{d}s - \int_0^t \lambda_{f^+(s)} e^{-\int_s^t \mu_{f^+(r)}\mathrm{d}r}\mathrm{d}s\right|.\end{aligned}$$

It is readily seen that the latter of these two terms is majorized by, using the definition of the set $\mathscr{F}_{t,\delta}$,

$$\int_0^t |\,\lambda_{f(s)} - \lambda_{f^+(s)}\,|\, e^{-\int_s^t \mu_{f^+(r)}\mathrm{d}r}\mathrm{d}s \leq \int_0^t |\lambda_{f(s)} - \lambda_{f^+(s)}\,|\,\mathrm{d}s \leq 2\lambda^+\delta k.$$

Now focus on the former term. First observe that, for all $s \in [0,t]$,

$$\left|\int_s^t (\mu_{f(r)} - \mu_{f^+(r)})\,\mathrm{d}r\right| \leq 2\mu^+\delta k,$$

so that the term can be bounded by

$$\int_0^t \lambda_{f(s)} e^{-\int_s^t \mu_{f(r)}\mathrm{d}r} \left| 1 - e^{-\int_s^t (\mu_{f^+(r)} - \mu_{f(r)})\mathrm{d}r} \right| \mathrm{d}s \le t\lambda^+ \max\left\{1 - e^{-2\mu^+\delta k}, e^{2\mu^+\delta k} - 1\right\}.$$

We conclude that $|\,\kappa(f) - \kappa(f^+)\,|$ 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 $a \ge \kappa(f^+)$,

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

$\triangleright$ The corresponding upper bound is less involved; the proof is identical to the one of [4, Thm. 1]. Note that if $a > a^+$, then for all $f \in \mathscr{F}_t$ we have that $\mathbb{E}P^{(N)}(f)$ is smaller than or equal to Na. Evidently,

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

Based on the Chernoff bound [7], we have

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

Combining the above inequalities, we obtain

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

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

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

This proves the upper bound. □

Recall that $\varrho^- := \min_{i\in\{1,\dots,d\}} \varrho_i$, where i^- is such that $\varrho_{i^-} = \varrho^-$. Then the following result, featuring the large deviations of the steady-state $M^{(N)}$ of $M^{(N)}(t)$, follows from a small modification in the proof of Thm. 2, fully analogously to [4, Prop. 1].

Corollary 1. *For $a \ge a^+ := \varrho^+$,*

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

For $a \le a^- := \varrho^-$,

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

Example 1. Consider, with $d = 2$, the scenario $\lambda_1 = 2$, $\mu_1 = 3$, $\lambda_2 = \mu_2 = 1$; as a consequence $\varrho_1 = \frac{2}{3}$ and $\varrho_2 = 1$, such that $i^+ = 2$. It is readily verified that $k = 2$ (such that $i_1 = 1$ and $i_2 = 2$), whereas $I_1 = \frac{1}{2}$ and $I_2 = 1$.

Using Thm. 1, we find that the 'maximizing path' f^+ is in state 1 until t_1^+, and state 2 thereafter, where t_1^+ is given by

$$t_1^+ = \frac{1}{3} \log\left(\frac{\frac{2}{3} - 0}{\frac{2}{3} - \frac{1}{2}}\right) = \frac{2}{3}\log 2.$$

As a consequence,

$$C^+(t) = \kappa(f^+) = \begin{cases} \dfrac{2}{3}(1-e^{-3t}), & t \in [0, t_1^+), \\ \\ \dfrac{1}{2}e^{-(t-t_1^+)} + (1-e^{-(t-t_1^+)}), & t \in [t_1^+, \infty); \end{cases}$$

the expression for $t \geq t_1^+$ simplifies to $1 - \frac{1}{2}\sqrt[3]{4}\, e^{-t}$. Observe that, in agreement with our results, $C^+(t) \uparrow \varrho_{i^+} = \varrho_2$ as $t \to \infty$. It is easily verified that $C^+(t)$ is continuous in $t = t_1^+$. The corresponding 'minimizing path' can be found analogously; it turns out that f^- is in state 2 until time $\log 2$, and in state 1 thereafter. It requires some elementary algebra to find that

$$C^-(t) := \kappa(f^-) = \begin{cases} 1 - e^{-t}, & t \in [0, \log 2), \\ \\ \dfrac{2}{3} - \dfrac{4}{3}e^{-3t}, & t \in [\log 2, \infty); \end{cases}$$

observe that $C^-(t) \uparrow \varrho_1$ as $t \to \infty$, as expected.
The corresponding large deviations now immediately follow from Thm. 2. ♢

3. Fast timescale regime

In this section, the process $M^{(N)}(t)$ results from scaling $\boldsymbol{\lambda}$, as before, by a factor N, but now also the background process J is sped up. The crucial idea is that J is scaled by a factor $N^{1+\epsilon}$, for some $\epsilon > 0$, and hence jumps at a faster time scale than the arrival process. The key finding of this section is that in this regime, as N tends to ∞, the tail asymptotics of $M^{(N)}(t)$ increasingly behave as those of an M/M/∞ queue with arrival rate $N\lambda_\infty$ and service rate μ_∞, where the definitions on λ_∞ and μ_∞ are given in (3). The results and the proofs in this section are similar as in [1], except of course for the approximation of the Poisson parameter of the scaled background process.
We denote the N-scaled background process by $(J^{(N^{1+\varepsilon})}(t))_{t\in\mathbb{R}}$. Let $\boldsymbol{L}^{(N^{1+\varepsilon})}(t_1,t_2)$ be the empirical distribution of the background process in $[t_1, t_2)$ (with $t_1 < t_2$); its i-th component is the fraction of time spent in state i, for $i = 1, \ldots, d$ (where obviously the d components are non-negative and sum to 1), that is

$$L_i^{(N^{1+\varepsilon})}(t_1,t_2) := \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} 1\{J^{(N^{1+\varepsilon})}(s) = i\}\,\mathrm{d}s.$$

By $\boldsymbol{L}(t_1,t_2)$ we denote the counterpart of $\boldsymbol{L}^{(N^{1+\varepsilon})}(t_1,t_2)$ for the non-scaled background process, where we recall the useful distributional identity:

$$\boldsymbol{L}^{(N^{1+\varepsilon})}(t_1,t_2) \overset{\mathrm{d}}{=} \boldsymbol{L}(N^{1+\varepsilon}t_1, N^{1+\varepsilon}t_2).$$

It is well known that the following law of large numbers applies: for any $\mathscr{S} \subset \mathbb{R}_+^d$ such that $\boldsymbol{\pi}$ is contained in the interior of $\mathscr{S}$, it holds that $\mathbb{P}(\boldsymbol{L}(0,t) \in \mathscr{S}) \to 1$ as $t \to \infty$. It is also a standard result (see e.g. [7, Thm. 3.1.6]) that $\boldsymbol{L}(0,t)$ satisfies a large deviations principle with rate function

$$\mathbb{I}(\boldsymbol{x}) := \sup_{\boldsymbol{u}>\boldsymbol{0}} \left(-\sum_{i=1}^{d} x_i \log \frac{\sum_{j=1}^{d} q_{ij}u_j}{u_i} \right); \tag{12}$$

this function is positive except when $\boldsymbol{x}=\boldsymbol{\pi}$. Under mild regularity conditions imposed on the set $\mathscr{S}$, this large deviations principle means that

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbb{P}(L(0,t) \in \mathscr{S}) = -\inf_{\boldsymbol{x}\in\mathscr{S}} \mathbb{I}(\boldsymbol{x}).$$

Considering the case that $\mathscr{S}$ does not contain $\boldsymbol{\pi}$, then the immediate consequence of this result is that the probability $\mathbb{P}(L(0,t)\in\mathscr{S})$ decays essentially exponentially.
In the sequel, we use the notation

$$\varrho(t) := \frac{\lambda_\infty}{\mu_\infty}(1-e^{-\mu_\infty t}). \tag{13}$$

As in the previous section, we wish to characterize the probability that $M^{(N)}(t)$ exceeds Na, given that the system starts off empty. It is known that $N^{-1}M^{(N)}(t) \to \varrho(t)$, a.s. for $N\to\infty$; see [2, Lemma 3]. In this paper, we are concerned with the rare event that the number of jobs *exceeds* a level Na, with $a \geq \varrho(t)$. The following theorem states that the corresponding large deviations are those of Poisson random variables with parameter $\varrho(t)$.

Theorem 3. *For $a \geq \varrho(t)$,*

$$\lim_{N\to\infty} \frac{1}{N} \log \mathbb{P}\left(M^{(N)}(t) \geq Na\right) = -\varrho(t) + a + a\log\frac{\varrho(t)}{a}.$$

Proof: Our starting point is again

$$\mathbb{P}\left(M^{(N)}(t) \geq Na\right) = \mathbb{P}\left(P^{(N)}\left(\kappa\left(J^{(N^{1+\varepsilon})}\right)\right) \geq Na\right).$$

For $\delta>0$, we define $\Delta(\boldsymbol{\pi})$ as a hypercube (of 'radius' δ) around $\boldsymbol{\pi}$:

$$\Delta(\boldsymbol{\pi}) := (\pi_1-\delta,\pi_1+\delta)\times\cdots\times(\pi_d-\delta,\pi_d+\delta).$$

Also introduce, for $\zeta>0$, the event

$$\mathscr{E}_\delta(\zeta,N) := \left\{ \boldsymbol{L}^{(N^{1+\varepsilon})}\left(0,\frac{t}{N^\zeta}\right) \in \Delta(\boldsymbol{\pi}),\ldots,\boldsymbol{L}^{(N^{1+\varepsilon})}\left(\frac{\lceil N^\zeta\rceil - 1}{N^\zeta}t, t\right) \in \Delta(\boldsymbol{\pi}) \right\}.$$

Lower bound. Intersecting the event of our interest with a second event evidently leads to a lower bound. Following this idea, we determine the decay rate of the obvious lower bound

$$\mathbb{P}\left(\left\{P^{(N)}\left(\kappa\left(J^{(N^{1+\varepsilon})}\right)\right) \geq Na\right\} \cap \mathscr{E}_\delta\left(\frac{1}{2},N\right)\right).$$

The idea behind considering this intersection, is that we focus on the scenario that the empirical distribution of the background process is during $[0,t]$ in $\Delta(\boldsymbol{\pi})$, and hence systematically close to $\boldsymbol{\pi}$.

To this end, first realize that, for any $\xi \in (0,1)$ and N sufficiently large, by virtue of the law of large numbers for the empirical distribution of the background process, see e.g. [7, Thm. 3.1.6]:

$$\mathbb{P}\left(\mathscr{E}_\delta\left(\frac{1}{2},N\right)\right) \geq \prod_{i=1}^{\lceil\sqrt{N}\rceil} \min_{j_i\in\{1,\dots,d\}} \mathbb{P}\left(\boldsymbol{L}\left(0,tN^{\frac{1}{2}+\varepsilon}\right)\in\Delta(\boldsymbol{\pi})\,\middle|\, J(0)=j_i\right) \geq (1-\xi)^{\lceil\sqrt{N}\rceil}.$$

It is a direct consequence that

$$\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(\mathscr{E}_\delta\left(\frac{1}{2},N\right)\right)=0.$$

We are thus left with determining a lower bound on the decay rate

$$\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(P^{(N)}\left(\kappa\left(J^{(N^{1+\varepsilon})}\right)\right)\geq Na\,\middle|\,\mathscr{E}_\delta\left(\frac{1}{2},N\right)\right).$$

Now the crucial observation is that the Poisson random variable is stochastically increasing in its parameter. As a result, we need to find a lower bound on $N\kappa(J^{(N^{1+\varepsilon})})$, conditional on the event $\mathscr{E}_\delta(\frac{1}{2},N)$. By picking in every segment and for every state (a) a lower bound on the state probability (still in $\Delta(\boldsymbol{\pi})$), as well as (b) the lower bound on the Poisson rate in this segment, it is readily verified that the following (deterministic!) lower bound applies:

$$\varrho_N(t) := t\sqrt{N}\sum_{j=1}^{d}\sum_{i=1}^{\lfloor\sqrt{N}\rfloor}(\pi_j-\delta)\lambda_j\exp\left(-\frac{t}{\sqrt{N}}\sum_{\ell=1}^{d}\sum_{k=\lfloor 1+(s/t)\sqrt{N}\rfloor}^{\lceil\sqrt{N}\rceil}(\pi_\ell+\delta)\mu_\ell\right). \tag{14}$$

We thus obtain the lower bound

$$\mathbb{P}\left(P^{(N)}\left(\kappa\left(J^{(N^{1+\varepsilon})}\right)\right)\geq Na\,\middle|\,\mathscr{E}_\delta\left(\frac{1}{2},N\right)\right)\geq e^{-\varrho_N(t)}\frac{(\varrho_N(t))^{\lceil Na\rceil}}{\lceil Na\rceil!}.$$

Applying Stirling's factorial approximation, this leads to

$$\liminf_{N\to\infty}\frac{1}{N}\log\left(e^{-\varrho_N(t)}\frac{(\varrho_N(t))^{\lceil Na\rceil}}{\lceil Na\rceil!}\right)\geq\liminf_{N\to\infty}\frac{1}{N}\left(-\varrho_N(t)+Na+Na\log\frac{\varrho_N(t)}{Na}\right).$$

Define $\bar\lambda := \sum_{i=1}^{d}\lambda_i$, and $\bar\mu := \sum_{i=1}^{d}\mu_i$. From the expression for $\varrho_N(t)$ in (14), and realizing that (recognize a Riemann integral!)

$$\frac{t}{\sqrt{N}}\sum_{\ell=1}^{d}\sum_{k=\lfloor 1+(s/t)\sqrt{N}\rfloor}^{\lceil\sqrt{N}\rceil}(\pi_\ell+\delta)\mu_\ell\to(t-s)(\mu_\infty+\delta\bar\mu),$$

it is observed that, as $N\to\infty$,

$$\frac{\varrho_N(t)}{N}\to\int_0^t(\lambda_\infty-\delta\bar\lambda)e^{-(t-s)(\mu_\infty+\delta\bar\mu)}\mathrm{d}s=\frac{\lambda_\infty-\delta\bar\lambda}{\mu_\infty+\delta\bar\mu}\left(1-e^{-t(\mu_\infty+\delta\bar\mu)}\right)=:\varrho^{(\delta)}(t).$$

It follows that

$$\liminf_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(P^{(N)}\left(\kappa\left(J^{(N^{1+\varepsilon})}\right)\right)\geq Na\,\middle|\,\mathscr{E}_\delta\left(\frac{1}{2},N\right)\right)\geq-\varrho^{(\delta)}(t)+a+a\log\frac{\varrho^{(\delta)}(t)}{a}.$$

The claimed lower bound follows by letting $\delta\downarrow 0$.

Upper bound. Again, we focus on scenarios in which the empirical distribution is consistently close to $\boldsymbol{\pi}$. To this end, we consider the obvious upper bound

$$\mathbb{P}\left(\left\{P^{(N)}\left(\kappa\left(J^{(N^{1+\varepsilon})}\right)\right)\geq Na\right\}\cap\mathscr{E}_\delta\left(\frac{\varepsilon}{2},N\right)\right)+\mathbb{P}\left(\mathscr{E}_\delta\left(\frac{\varepsilon}{2},N\right)^c\right).$$

Due to the union bound,

$$\mathbb{P}\left(\mathscr{E}_\delta\left(\frac{\varepsilon}{2},N\right)^c\right)\;\leq\;\lceil N^{\varepsilon/2}\rceil\left(\max_{j\in\{1,\dots,d\}}\mathbb{P}\left(\boldsymbol{L}\left(0,tN^{1+\frac{\varepsilon}{2}}\right)\not\in\Delta(\boldsymbol{\pi})\,\middle|\,J(0)=j\right)\right).$$

Standard large deviations results imply that

$$\lim_{N\to\infty}\frac{1}{N^{1+\frac{\varepsilon}{2}}}\log\mathbb{P}\left(\boldsymbol{L}\left(0,tN^{1+\frac{\varepsilon}{2}}\right)\not\in\Delta(\boldsymbol{\pi})\,\middle|\,J(0)=j\right)=-\inf_{\boldsymbol{x}\not\in\Delta(\boldsymbol{\pi})}\mathbb{I}(\boldsymbol{x})<0,$$

and hence

$$\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(\mathscr{E}_\delta\left(\frac{\varepsilon}{2},N\right)^c\right)=-\infty.$$

Using [7, Lemma 1.2.15], it now suffices to prove that

$$\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(P^{(N)}\left(\kappa\left(J^{(N^{1+\varepsilon})}\right)\right)\geq Na\,\middle|\,\mathscr{E}_\delta\left(\frac{\varepsilon}{2},N\right)\right)\leq-\bar{\varrho}^{(\delta)}(t)+a+a\log\frac{\bar{\varrho}^{(\delta)}(t)}{a},$$

with $\bar{\varrho}^{(\delta)}(t)$ such that $\bar{\varrho}^{(\delta)}(t)\to\varrho(t)$ as $\delta\downarrow 0$. The remainder of the proof settles this issue. To this end, we determine a (deterministic!) upper bound, conditional on $\mathscr{E}_\delta(\frac{\varepsilon}{2},N)$, on the random variable $N\kappa(J^{(N^{1+\varepsilon})})$. Using a similar reasoning as in (14), it is readily verified that the following upper bound applies:

$$\bar{\varrho}_N(t):=tN^{1-\frac{\varepsilon}{2}}\sum_{j=1}^{d}\sum_{i=1}^{\lceil N^{\frac{\varepsilon}{2}}\rceil}(\pi_j+\delta)\lambda_j\exp\left(-\frac{t}{N^{\frac{\varepsilon}{2}}}\sum_{\ell=1}^{d}\sum_{k=\lceil 1+(s/t)N^{\frac{\varepsilon}{2}}\rceil}^{\lfloor N^{\frac{\epsilon}{2}}\rfloor}(\pi_\ell-\delta)\mu_\ell\right).$$

Chebycheff's inequality on the cumulant generating function of Poisson random variables [7, p. 30] gives

$$\begin{aligned}&\limsup_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(P^{(N)}\left(\kappa\left(J^{(N^{1+\varepsilon})}\right)\right)\geq Na\,\middle|\,\mathscr{E}_\delta\left(\frac{\varepsilon}{2},N\right)\right)\\ \leq\;&\limsup_{N\to\infty}\frac{1}{N}\left(-\bar{\varrho}_N(t)+Na+Na\log\frac{\bar{\varrho}_N(t)}{Na}\right).\end{aligned}$$

Pick δ sufficiently small that $\mu_\infty>\delta\bar{\mu}$. Combining the above findings, leads to the desired upper bound, realizing that, using the same reasoning as in the lower bound,

$$\frac{\bar{\varrho}_N(t)}{N}\to\bar{\varrho}^{(\delta)}(t):=\frac{\lambda_\infty+\delta\bar{\lambda}}{\mu_\infty-\delta\bar{\mu}}\left(1-e^{-t(\mu_\infty-\delta\bar{\mu})}\right)$$

as $N\to\infty$; the claim follows immediately from $\bar{\varrho}^{(\delta)}(t)\to\varrho(t)$ as $\delta\downarrow 0$. □

The following corollary follows from the the Gärtner-Ellis theorem, in conjunction with the duality between the cumulant function and the Legendre-Fenchel transform.

Corollary 2. *The limiting cumulant function of $M^{(N)}(t)$ corresponds to that of a Poisson random variable:*

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{E}\exp\left(\vartheta M^{(N)}(t)\right)=\varrho(t)(e^{\vartheta}-1).$$

The above result directly carries over to the steady-state counterpart $M^{(N)}$ of $M^{(N)}(t)$. To this end, we define $\varrho := \lim_{t\to\infty}\varrho(t) = \lambda_\infty/\mu_\infty$, and realize that $M^{(N)}$ has a Poisson distribution with mean

$$N\int_{-\infty}^{0}\lambda_{J(s)}e^{-\int_s^0\mu_{J(r)}\mathrm{d}r}\mathrm{d}s;$$

see e.g. [6]. Then the proof of the corollary below is essentially the same as the one for the transient case.

Corollary 3. *For $a\geq\varrho$,*

$$\lim_{N\to\infty}\frac{1}{N}\log\mathbb{P}\left(M^{(N)}\geq Na\right)=-\varrho+a+a\log\frac{\varrho}{a}.$$

In addition, $N^{-1}\log\mathbb{E}\exp\left(\vartheta M^{(N)}\right)\to\varrho(e^{\vartheta}-1)$ as $N\to\infty$.

Acknowledgment

We thank Rami Atar (Technion) for useful discussions related to the proof of Thm. 1.

References

[1] J. Blom, K. de Turck, and M. Mandjes (2013). Rare-event analysis of Markov-modulated infinite-server queues: a Poisson limit. *Stochastic Models,* **29**, 463–474.

[2] J. Blom, K. de Turck, and M. Mandjes (2013). A central limit theorem for Markov-modulated infinite-server queues. In: *Proceedings ASMTA 2013*, Ghent, Belgium. *Lecture Notes in Computer Science* (LNCS) Series, **7984**, pp. 81-95.

[3] J. Blom, O. Kella, M. Mandjes, and H. Thorsdottir (2013). Markov-modulated infinite server queues with general service times. To appear in *Queueing Systems* (DOI:10.1007/s11134-013-9368-4).

[4] J. Blom and M. Mandjes (2013). A large-deviations analysis of Markov-modulated inifinite-server queues. *Operations Research Letters,* **41**, 220–225.

[5] J. Blom, M. Mandjes, and H. Thorsdottir (2013). Time-scaling limits for Markov-modulated infinite-server queues. *Stochastic Models*, **29**, 112–127.

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

[7] A. Dembo and O. Zeitouni (1998). *Large Deviations Techniques and Applications,* 2nd edition. Springer, New York.

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

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

E-mail address: `joke.blom@cwi.nl, kdeturck@telin.ugent.be, Offer.Kella@huji.ac.il, M.R.H.Mandjes@uva.nl`