




# Queue length asymptotics for the multiple-server queue with heavy-tailed Weibull service times

Mihail Bazhba<sup>1</sup> · Jose Blanchet<sup>2</sup> · Chang-Han Rhee<sup>3</sup> · Bert Zwart<sup>1,4</sup> 

Received: 27 September 2018 / Revised: 18 October 2019 / Published online: 10 November 2019  
© The Author(s) 2019

## Abstract

We study the occurrence of large queue lengths in the  $GI/GI/d$  queue with heavy-tailed Weibull-type service times. Our analysis hinges on a recently developed sample path large-deviations principle for Lévy processes and random walks, following a continuous mapping approach. Also, we identify and solve a key variational problem which provides physical insight into the way a large queue length occurs. In contrast to the regularly varying case, we observe several subtle features such as a non-trivial trade-off between the number of big jobs and their sizes and a surprising asymmetric structure in asymptotic job sizes leading to congestion.

**Keywords** Multiple-server queue · Queue length asymptotics · Heavy tails · Weibull service times

**Mathematics Subject Classification** 60K25 · 68M20

---

✉ Bert Zwart  
bert.zwart@cwil.nl

Mihail Bazhba  
bazhba@cwil.nl

Jose Blanchet  
jose.blanchet@stanford.edu

Chang-Han Rhee  
chang-han.rhee@northwestern.edu

- <sup>1</sup> Centrum Wiskunde & Informatica, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
- <sup>2</sup> Management Science and Engineering, Stanford University 475 Via Ortega, Suite 310, Stanford, CA 94305, USA
- <sup>3</sup> Industrial Engineering and Management Sciences, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA
- <sup>4</sup> Eindhoven University of Technology, Eindhoven, The Netherlands

## 1 Introduction

The queue with multiple servers, known as the  $GI/GI/d$  queue, is a fundamental model in queueing theory. Its use in everyday applications such as call centers and supermarkets is well documented, and, despite being significantly studied over decades, it continues to pose interesting research challenges. Early work [1,2] focused on exact analysis of the invariant waiting-time distribution, but finding tractable solutions has turned out to be challenging. This has led to lines of research that focus on approximations, either considering heavily loaded systems [3,4] or investigating the frequency of rare events, for example the probability of a long waiting time or large queue length. For light-tailed service times, such problems have been considered in [5,6].

The focus on this paper is on rare event analysis of the queue length in the case of heavy-tailed service times, a topic that is more recent. For a single server, the literature on this topic is extensive, as there is an explicit connection between waiting times and first passage times of random walks; a textbook treatment can be found in [7]. Tail asymptotics for the steady-state queue length have been treated in [8].

The earliest paper on heavy tails in the setting of a queue with multiple servers that we are aware of is [9], which stated a conjecture regarding the form of the tail of the waiting time distribution in steady state, assuming that the service-time distribution is sub-exponential. This has led to follow-up work on necessary and sufficient conditions for finite moments of the waiting-time distribution in steady state [10], and on tail asymptotics [11,12]. Most of the results in the latter two papers focus on the case of regularly varying service times. An insight is that if the system load  $\rho$  is not an integer, a large waiting time occurs due to the arrival of  $\lceil d - \rho \rceil$  big jobs. The case of other heavy-tailed service times is poorly understood.

In the present paper, we assume that the service-time distribution has a tail of the form  $e^{-L(x)x^\alpha}$ , where  $\alpha \in (0, 1)$ , and  $L$  is a slowly varying function (a more comprehensive definition is given later on). Tail distributions of this form are also known as semi-exponential. Their analysis poses challenges, as this category of tails falls in between the Pareto (very heavy tailed) case and the classical light-tailed case. In particular, in the case of  $d = 2$  and  $\rho < 1$ , the results in [11] imply that two big jobs are necessary to cause a large waiting time when service times have a Weibull distribution. The arguments in [11] cannot be extended to the case  $\rho > 1$ . In the 2009 Erlang centennial conference, Sergey Foss posed the question “how many big service times are needed to cause a large waiting time to occur, if the system is in steady state?”. He noted that even a physical or heuristic treatment has been absent.

This has motivated us to investigate a strongly related question; namely, we analyze the event that the queue length  $Q(\gamma n)$  at a large time  $\gamma n$  exceeds a value  $n$ . A key result that we utilize in our analysis is a powerful upper bound of Gamarnik and Goldberg, see [13], for  $\mathbf{P}(Q(t) > x)$ . This upper bound can be combined with a recently developed large-deviations principle for random walks with heavy-tailed Weibull-type increments, which is another key result that we use. Consequently, we can estimate the probability of a large queue length of the  $GI/GI/d$  queue with heavy-tailed Weibull-type service times and obtain physical insights into “the most likely way” in which a large queue length builds up.

The main result of this paper, given in Theorem 3.1, states the following: If  $Q(t)$  is the queue length at time  $t$  (assuming an empty system at time zero) and  $\gamma \in (0, \infty)$ , then

$$\lim_{n \rightarrow \infty} \frac{\log \mathbf{P}(Q(\gamma n) > n)}{L(n)n^\alpha} = -c^*, \tag{1.1}$$

with  $c^*$  the value of the optimization problem

$$\begin{aligned} & \min_{x_1, \dots, x_d} \sum_{i=1}^d x_i^\alpha \text{ subject to} \\ & \sup_{s \in [0, \gamma]} \left\{ \lambda s - \sum_{i=1}^d (s - x_i)^+ \right\} \geq 1, \\ & x_1, \dots, x_d \geq 0, \end{aligned} \tag{1.2}$$

where  $\lambda$  is the arrival rate and service times are normalized to have unit mean. Note that this problem is equivalent to an  $L^\alpha$ -norm minimization problem with  $\alpha \in (0, 1)$ . Such problems also appear in applications such as compressed sensing and are strongly NP-hard in general; see [14] and references therein. In our particular case, we can analyze this problem exactly, and if  $\gamma \geq 1/(\lambda - \lfloor \lambda \rfloor)$ , the solution takes the simple form

$$c^* = \min_{l \in \{0, \dots, \lfloor \lambda \rfloor\}} (d - l) \left( \frac{1}{\lambda - l} \right)^\alpha. \tag{1.3}$$

This simple minimization problem has at most two optimal solutions, which represent the most likely number of big jumps that are responsible for a large queue length to occur, and the most likely buildup of the queue length is through a linear path. For smaller values of  $\gamma$ , asymmetric solutions can occur, leading to a piecewise linear buildup of the queue length; we refer to Sect. 3 for more details.

Note that the intuition that the solution to (1.2) yields is qualitatively different from the case in which service times have a power law. In the latter case, the optimal number of big jobs equals the minimum number of servers that need to be removed to make the system unstable. In the Weibull-type case, there is a non-trivial trade-off between the *number* of big jobs and their *size*, and this trade-off is captured by (1.2) and (1.3).

Although we do not make these claims rigorous for  $\gamma = \infty$  (which requires an interchange of limits argument beyond the scope of the paper), it makes a clear suggestion of what the tail behavior of the steady-state queue length should be. This can then be related to the steady-state waiting-time distribution, and the original question posed by Foss, using distributional Little’s law.

As mentioned before, we obtain (1.1) by utilizing a tail bound for  $Q(t)$ , which is derived in [13]. This tail bound is given in terms of functionals of superpositions of renewal processes. We show that these functionals are (almost) continuous in the  $M'_1$  topology (in the sense of being amenable to the use of the extended contraction

principle). The  $M'_1$  introduced in [15] is precisely the topology used in the development of a recently produced large-deviations principle for random walks with Weibull-type increments; see [16]. So, our approach here makes the new large-deviations principle directly applicable.

The paper is organized as follows: Section 2 provides a model description and some useful tools used in our proofs. Section 3 provides our main result and some mathematical insights associated with it. Lastly, Sect. 4 contains the lemmas and proofs needed to construct the main result of this paper, Theorem 3.1.

## 2 Model description and preliminary results

We consider the FCFS  $GI/GI/d$  queueing model with  $d$  servers in which inter-arrival times are independent and identically distributed (i.i.d.) random variables (r.v.'s) and service times are i.i.d. r.v.'s independent of the arrival process. Let  $A \geq 0$  and  $S \geq 0$  be a pair of generic inter-arrival time and service time, respectively. We introduce the following assumptions:

**Assumption 2.1** There exists  $\theta_+ > 0$  such that  $\mathbf{E}(e^{\theta A}) < \infty$  for every  $\theta \leq \theta_+$ .

**Assumption 2.2**  $\mathbf{P}(S \geq x) = e^{-L(x)x^\alpha}$ ,  $\alpha \in (0, 1)$ , where  $L(\cdot)$  is a slowly varying function at infinity and  $L(x)x^{\alpha-1}$  is eventually non-increasing.

Let  $Q(t)$  denote the queue length process at time  $t$  in the FCFS  $GI/GI/d$  queueing system with inter-arrival times being i.i.d. copies of  $A$  and service times being i.i.d. copies of  $S$ . We assume that  $Q(0) = 0$ . Our goal is to identify the limiting behavior of  $\mathbf{P}(Q(\gamma n) > n)$  as  $n \rightarrow \infty$  in terms of the distributions of  $A$  and  $S$ . To simplify the notation, let  $\lambda = 1/\mathbf{E}[A]$  and assume without loss of generality that  $\mathbf{E}[S] = 1$ . To ensure stability, let  $\lambda < d$ . Let  $M$  be the renewal process associated with  $A$ . That is,

$$M(t) = \inf\{s : A(s) > t\},$$

and  $A(t) \triangleq A_1 + A_2 + \dots + A_{\lfloor t \rfloor}$ , where  $A_1, A_2, \dots$  are i.i.d. copies of  $A$ , and  $A(0) = 0$ . Similarly, for each  $i = 1, \dots, d$ , let  $S^{(i)}(t) \triangleq S_1^{(i)} + S_2^{(i)} + \dots + S_{\lfloor t \rfloor}^{(i)}$ , where  $S_1^{(i)}, S_2^{(i)}, \dots$  are i.i.d. copies of  $S$ , and  $N^{(i)}$  be the renewal process associated with  $S$ . Let  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$  be scaled processes of  $M$  and  $N^{(i)}$ . More precisely,  $\bar{M}_n(t) = M(nt)/n$  and  $\bar{N}_n^{(i)}(t) = N^{(i)}(nt)/n$  for  $t \geq 0$ . Our analysis hinges on Corollary 1 of [13], which for the  $GI/GI/d$  queue states

**Result 2.1** For all  $x > 0$  and  $t \geq 0$ ,

$$\mathbf{P}(Q(t) > x) \leq \mathbf{P}\left(\sup_{0 \leq s \leq t} \left\{ (M(t) - M(t-s)) - \sum_{i=1}^d (N^{(i)}(t) - N^{(i)}(t-s)) \right\} > x \right). \tag{2.1}$$

Now, from (2.1) we conclude that, for each  $\gamma \in (0, \infty)$ ,

$$\mathbf{P}(Q(\gamma n) > n) \leq \mathbf{P}\left(\sup_{0 \leq s \leq \gamma} \left\{ \bar{M}_n(\gamma) - \bar{M}_n(s) - \sum_{i=1}^d (\bar{N}_n^{(i)}(\gamma) - \bar{N}_n^{(i)}(s)) \right\} \geq 1\right). \tag{2.2}$$

Though this is only an upper bound, our main result implies that (2.2) is an asymptotically tight upper bound as  $n \rightarrow \infty$ . We establish this later on by deriving a lower bound with the same asymptotic behavior.

In view of the above, a natural way to proceed is to establish large-deviations principles for  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$ ,  $i = 1, \dots, d$ . Before we continue, we start with some general background on large-deviations theory, based on [17, 18]. Let  $(\mathbb{X}, d)$  be a metric space and  $\mathcal{T}$  denote the topology induced by the metric  $d$ . Let  $X_n$  be a sequence of  $\mathbb{X}$ -valued random variables. Let  $I$  be a nonnegative lower semi-continuous function on  $\mathbb{X}$  and  $a_n$  be a sequence of positive real numbers that tends to infinity as  $n \rightarrow \infty$ . We say that  $X_n$  satisfies a large-deviations principle (LDP) in  $(\mathbb{X}, \mathcal{T})$  with speed  $a_n$  and rate function  $I$  if

$$-\inf_{x \in A^\circ} I(x) \leq \liminf_{n \rightarrow \infty} \frac{\log \mathbf{P}(X_n \in A)}{a_n} \leq \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}(X_n \in A)}{a_n} \leq -\inf_{x \in A^-} I(x)$$

for any measurable set  $A$ . Here,  $A^\circ$  and  $A^-$  are, respectively, the interior and the closure of the set  $A$ . If the level sets  $\{y : I(y) \leq a\}$  are compact for each  $a \in \mathbb{R}_+$ , we say that  $I$  is a good rate function. By deriving an LDP, one can have an estimation of the magnitude of probabilities of rare events on an exponential scale: if the upper and lower bounds of the LDP match, then  $P(X_n \in G) \approx e^{-a_n \inf_{x \in G} I(x)}$ . The optimizers of the infimum typically provide insight into the most likely way a rare event occurs (i.e., the conditional distribution given the rare event of interest). For more background, we refer to [18] and [19].

An important factor in establishing an LDP on function spaces is the topology of the space under consideration. Let  $\mathbb{D}[0, T]$  denote the Skorokhod space (i.e., the space of càdlàg functions from  $[0, T]$  to  $\mathbb{R}$ ). We shall use  $\mathcal{T}_{M'_1}$  to denote the  $M'_1$  Skorokhod topology on  $\mathbb{D}[0, T]$ , which is generated by a metric  $d_{M'_1}$  defined in terms of the graphs induced by the elements of  $\mathbb{D}[0, T]$ . The precise definitions of the graph and the metric are as follows:

**Definition 2.1** For  $\xi \in \mathbb{D}[0, T]$ , define the extended completed graph  $\Gamma'(\xi)$  of  $\xi$  as

$$\Gamma'(\xi) \triangleq \{(u, t) \in \mathbb{R} \times [0, T] : u \in [\xi(t-) \wedge \xi(t), \xi(t-) \vee \xi(t)]\},$$

where  $\xi(0-) \triangleq 0$ . Define an order on the graph  $\Gamma'(\xi)$  by setting  $(u_1, t_1) < (u_2, t_2)$ , for every  $(u_1, t_1), (u_2, t_2) \in \Gamma'(\xi)$ , if either

- $t_1 < t_2$ ; or
- $t_1 = t_2$  and  $|\xi(t_1-) - u_1| < |\xi(t_2-) - u_2|$ .

We call a continuous non-decreasing function  $(u, t) = ((u(s), t(s)), s \in [0, T])$  from  $[0, T]$  to  $\mathbb{R} \times [0, T]$  an  $M'_1$  parametrization of  $\Gamma'(\xi)$  if  $\Gamma'(\xi) = \{(u(s), t(s)) : s \in [0, T]\}$ . We also just call it a parametrization of  $\xi$ .

**Definition 2.2** Define the  $M'_1$  metric on  $\mathbb{D}[0, T]$  as follows:

$$d_{M'_1}(\xi, \zeta) \triangleq \inf_{\substack{(u,t) \in \Pi_{M'_1}(\xi) \\ (v,r) \in \Pi_{M'_1}(\zeta)}} \{\|u - v\|_\infty + \|t - r\|_\infty\},$$

where  $\Pi_{M'_1}(\xi)$  is the set of all  $M'_1$  parametrizations of  $\Gamma'(\xi)$ .

Note that we can alternatively define the  $M'_1$  metric in such a way that the infimum in Definition 2.2 is over the parametrizations that are strictly increasing (rather than merely non-decreasing) without changing the resulting distance. An immediate implication of such an alternative definition is that a sequence of functions  $\{\xi_n\}_{n \geq 1}$  converges to  $\xi$  in  $(\mathbb{D}[0, T], \mathcal{T}_{M'_1})$  if and only if there exist parametrizations  $(u, t) \in \Pi_{M'_1}(\xi)$  and  $(u_n, t_n) \in \Pi_{M'_1}(\xi_n)$  for each  $n \geq 1$  such that

$$\sup_{s \in [0, T]} \{|u_n(s) - u(s)| + |t_n(s) - t(s)|\} \rightarrow 0 \tag{2.3}$$

as  $n \rightarrow \infty$ .

The continuity of certain maps w.r.t. the  $M'_1$  topology is a key component in our whole argument. Therefore, we note some important related properties used in our proofs. We refer to Lemma B.2 in the appendix for proofs of these results.

1. The functional  $S : \mathbb{D}[0, T] \rightarrow \mathbb{R}$ , where  $S(\xi) = \sup_{t \in [0, T]} \xi(t)$ , is continuous w.r.t. the  $M'_1$  topology at  $\xi \in \mathbb{D}[0, T]$  such that  $\xi(0) \geq 0$ ;
2. the functional  $E : \mathbb{D}[0, T] \rightarrow \mathbb{R}$ , where  $E(\xi) = \xi(T)$ , is continuous w.r.t. the  $M'_1$  topology on  $\mathbb{D}[0, T]$ ;
3. the addition map  $(\xi, \zeta) \mapsto \xi + \zeta$  is a continuous map w.r.t. the  $M'_1$  topology if the functions  $\xi$  and  $\zeta$  do not have jumps of the opposite sign at the same jump times.

Now, we describe a recent result derived in [16] on sample path large deviations for random walks with heavy-tailed Weibull increments which constitutes an important cornerstone of our whole argument. We say that  $\xi \in \mathbb{D}[0, T]$  is a pure jump function if  $\xi = \sum_{i=1}^\infty x_i \mathbb{1}_{[u_i, T]}$  for some  $x_i$  and  $u_i$  such that  $x_i \in \mathbb{R}$  and  $u_i \in [0, T]$  for each  $i$ , and  $u_i$  are all distinct. Let  $\mathbb{D}_p^\uparrow[0, T]$  be the subspace of  $\mathbb{D}[0, T]$  consisting of non-decreasing pure jump functions that assume nonnegative values at the origin.

**Result 2.2** Let  $S_n, n \geq 1$ , be a mean-zero random walk such that  $\mathbf{E}(e^{-\epsilon S_1}) < \infty$  for some  $\epsilon > 0$ ,  $\mathbf{P}(S_1 \geq x) = e^{-L(x)x^\alpha}$  for some  $\alpha \in (0, 1)$ , and that  $L(x)x^{\alpha-1}$  is eventually non-increasing. Then,  $\bar{S}_n$  satisfies the LDP in  $(\mathbb{D}[0, T], \mathcal{T}_{M'_1})$  with speed  $L(n)n^\alpha$  and good rate function  $I_{M'_1} : \mathbb{D}[0, T] \rightarrow [0, \infty]$  given by

$$I_{M'_1}(\xi) \triangleq \begin{cases} \sum_{t \in [0, 1]} (\xi(t) - \xi(t-))^\alpha & \text{if } \xi \in \mathbb{D}_p^\uparrow[0, T], \\ \infty & \text{otherwise.} \end{cases} \tag{2.4}$$

Note that  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$ 's depend on a random number of  $A_j$ 's and  $S_j^{(i)}$ 's and hence may depend on an arbitrarily large number of  $A_j$ 's and  $S_j^{(i)}$ 's. This does not exactly correspond to the large-deviations framework presented in Result 2.2. To accommodate such a context, we introduce the following maps: Fix  $\gamma > 0$ . For any path  $\xi$ , let  $\Psi(\xi)(t)$  denote the running supremum of  $\xi$  up to time  $t$ :

$$\Psi(\xi)(t) \triangleq \sup_{s \in [0, t]} \xi(s).$$

For each  $\mu$ , define a map  $\Phi_\mu : \mathbb{D}[0, \gamma/\mu] \rightarrow \mathbb{D}[0, \gamma]$  by

$$\Phi_\mu(\xi)(t) \triangleq \varphi_\mu(\xi)(t) \wedge \psi_\mu(\xi)(t),$$

where

$$\begin{aligned} \varphi_\mu(\xi)(t) &\triangleq \inf\{s \in [0, \gamma/\mu] : \xi(s) > t\} \quad \text{and} \\ \psi_\mu(\xi)(t) &\triangleq \frac{1}{\mu} \left( \gamma + [t - \Psi(\xi)(\gamma/\mu)]_+ \right). \end{aligned} \tag{2.5}$$

Here, we denote  $\max\{x, 0\}$  by  $[x]_+$ . In words, between the origin and the supremum of  $\xi$ ,  $\Phi_\mu(\xi)(s)$  is the first passage time of  $\xi$  crossing the level  $s$ ; from there to the final point  $\gamma$ ,  $\Phi_\mu(\xi)$  increases linearly from  $\gamma/\mu$  at rate  $1/\mu$  (instead of jumping to  $\infty$  and staying there). Define  $\bar{A}_n \in \mathbb{D}[0, \gamma/\mathbf{EA}]$  as  $\bar{A}_n(t) \triangleq A(nt)/n$  for  $t \in [0, \gamma/\mathbf{EA}]$  and  $\bar{S}_n^{(i)} \in \mathbb{D}[0, \gamma]$  as  $\bar{S}_n^{(i)}(t) \triangleq S^{(i)}(nt)/n = \frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} S_j^{(i)}$  for  $t \in [0, \gamma]$ . In deriving LDPs for  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$ , we use the fact that  $\Phi_{\mathbf{EA}}(\bar{A}_n)$  is a function of  $\{\bar{A}_n(t) : t \in [0, \gamma/\mathbf{EA}]\}$  (and, hence, the LDP associated with it can be derived from the LDP we have for  $\bar{A}_n$ ) as well as the fact that  $\Phi_{\mathbf{EA}}(\bar{A}_n)$  is close enough to  $\bar{M}_n$  that they satisfy the same LDP. Similarly, we derive the LDP for  $\bar{N}_n^{(i)}$  from the LDP for  $\bar{S}_n^{(i)}$  using the fact that  $\Phi_1(\bar{S}_n^{(i)})$  is close enough to  $\bar{N}_n^{(i)}$  for our purpose.

We now turn to the main result of this paper and discuss its implications.

### 3 Main result

Recall that  $Q(t)$  denotes the queue length of the  $GI/GI/d$  queue at time  $t$ .

**Theorem 3.1** *For each  $\gamma \in (0, \infty)$ , it holds that*

$$\lim_{n \rightarrow \infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}(Q(\gamma n) > n) = -c^*, \tag{3.1}$$

where  $c^*$  is defined as follows: for  $\gamma \geq 1/\lambda$ ,  $c^*$  is equal to

$$\min \left\{ \inf_{0 < k \leq \lfloor \lambda \rfloor; \gamma < 1/(\lambda - k)} \left\{ (d - k) \gamma^\alpha + (1 - \gamma \lambda + \gamma k)^\alpha (k - \lfloor \lambda - 1/\gamma \rfloor)^{1-\alpha} \right\}, \right. \\ \left. \min_{l=0}^{\lfloor \lambda - 1/\gamma \rfloor} \left\{ (d - l) \left( \frac{1}{\lambda - l} \right)^\alpha \right\} \right\}, \tag{3.2}$$

while for  $\gamma < 1/\lambda$ ,  $c^* = \infty$ .

Theorem 3.1 is stated under the assumption that  $ES = 1$  for the sake of simplicity. Following a completely analogous argument with slightly more involved notation, one can obtain the following expression for  $c^*$  for the general case where  $\sigma = 1/ES \neq 1$ :

$$\min \left\{ \min_{0 < k \leq \lfloor \lambda/\sigma \rfloor; \gamma < 1/(\lambda - k\sigma)} \left\{ (d - k) \gamma^\alpha + (1 - \gamma (\lambda - k\sigma))^\alpha \sigma^{-\alpha} \left( k - \left\lfloor \lambda/\sigma - \frac{1}{\gamma\sigma} \right\rfloor \right)^{1-\alpha} \right\}, \right. \\ \left. \min_{l=0}^{\lfloor \lambda/\sigma - \frac{1}{\gamma\sigma} \rfloor} \left\{ (d - l) \left( \frac{1}{\lambda - l\sigma} \right)^\alpha \right\} \right\}.$$

The proof of Theorem 3.1 is provided in Sect. 4 by implementing the following strategy:

1. We first prove that  $\bar{A}_n$  and  $\bar{S}_n^{(i)}$ ,  $i = 1, \dots, d$ , satisfy certain LDPs in Proposition 4.1. The LDPs for the  $\bar{S}_n^{(i)}$  are a consequence of Result 2.2, while the LDP for  $\bar{A}_n$  is deduced by the sample path LDP in [6].
2. We prove that  $\Phi_{EA}(\cdot)$  and  $\Phi_1(\cdot)$  are essentially continuous maps—see Proposition B.1 for the more precise statement—and, hence,  $\Phi_{EA}(\bar{A}_n)$  and  $\Phi_1(\bar{S}_n^{(i)})$  satisfy the LDPs deduced by the extended contraction principle (cf. Appendix A).
3. We show that  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$  are equivalent to  $\Phi_{EA}(\bar{A}_n)$  and  $\Phi_1(\bar{S}_n^{(i)})$ , respectively, in terms of their large deviations (Proposition 4.2); so  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$  satisfy the same LDPs (Proposition 4.3).
4. By applying the contraction principle to the  $\bar{N}_n^{(i)}$  with the continuous maps in Appendix B, we infer the (logarithmic) asymptotic upper bound of  $\mathbf{P}(Q(\gamma n) > n)$ , which can be characterized by the solution of a (non-standard) variational problem. On the other hand, the lower bound is derived by keeping track of the optimal solution associated with the LDP upper bound. The complete argument is presented in Proposition 4.4.
5. We solve the variational problem in Proposition 4.5 to explicitly compute its optimal solution. The optimal solution of the variational problem provides the limiting exponent and information on the trajectory leading to a large queue length.

In the remainder of this section, we further investigate properties of the solution of the optimization problem that defines  $c^*$ . In large-deviations theory, solutions of such problems are known to provide insights into the most likely way a specific rare event occurs. Such insights are physical, and more technical work is typically needed to make such insights rigorous; we refer to Lemma 4.2 of [18] for more background. The latter lemma can be applied in a relatively straightforward manner to derive a rigorous statement for the most likely way that the functional in the Gamarnik and



Goldberg upper bound (cf. Result 2.1) becomes large. The computations below are mainly intended to provide physical insight and highlight differences from the well-studied case where the job sizes follow a regularly varying distribution.

We consider two different cases based on the value of  $\gamma$ . If  $\gamma < 1/\lambda$ , no finite number of large jobs suffice, and we conjecture that the large-deviations behavior is driven by a combination of light-tailed and heavy-tailed phenomena in which the light-tailed dynamics involve pushing the arrival rate by exponential tilting to the critical value  $1/\gamma$ , followed by the heavy-tailed contribution evaluated as we explain in the following development. If  $\gamma > 1/\lambda$ , we observe the following features that come in contrast to the case of regularly varying service-time tails:

1. The large-deviations behavior may not be driven by the smallest number of jumps which drives the queueing system to instability (i.e.,  $\lceil d - \lambda \rceil$ ). In other words, in the Weibull setting, it might be more efficient to block more servers.
2. It is not necessary that the servers are blocked by the same amount, i.e., asymmetry in job sizes may be the most probable scenario in certain cases.

To illustrate the first point, assume  $\gamma > 1/(\lambda - \lfloor \lambda \rfloor)$ , in which case  $\lfloor \lambda \rfloor \leq \lfloor \lambda - 1/\gamma \rfloor$ . In that particular case, the first infimum in (3.2) is over an empty set and we interpret it as  $\infty$ . So the optimal solution of  $c^*$  reduces to

$$\min_{l=0}^{\lfloor \lambda \rfloor} \left\{ (d - l) \left( \frac{1}{\lambda - l} \right)^\alpha \right\}.$$

Let  $l^*$  denote the index associated with the optimal value of the expression above. Intuitively,  $d - l^*$  represents the optimal number of blocked servers so that the queue gets congested. Observe that  $d - \lfloor \lambda \rfloor = \lceil d - \lambda \rceil$  corresponds to the number of servers blocked in the regularly varying case. Note that if we examine

$$f(t) = (d - t)(\lambda - t)^{-\alpha},$$

for  $t \in [0, \lfloor \lambda \rfloor]$ , then the derivative  $\dot{f}(\cdot)$  is equal to  $\dot{f}(t) = \alpha(d - t)(\lambda - t)^{-\alpha-1} - (\lambda - t)^{-\alpha}$ . Hence,

$$\dot{f}(t) < 0 \iff t < \frac{(\lambda - \alpha d)}{(1 - \alpha)},$$

and

$$\dot{f}(t) > 0 \iff t > \frac{(\lambda - \alpha d)}{(1 - \alpha)},$$

with  $\dot{f}(t) = 0$  if and only if  $t = (\lambda - \alpha d) / (1 - \alpha)$ . This observation allows us to conclude that whenever  $\gamma > 1/(\lambda - \lfloor \lambda \rfloor)$  we can distinguish two cases. The first one occurs if

$$\lfloor \lambda \rfloor \leq \frac{(\lambda - \alpha d)}{(1 - \alpha)},$$

in which case  $l^* = \lfloor \lambda \rfloor$ . This case is qualitatively consistent with the way in which large deviations occur in the regularly varying case. On the other hand, if  $\lfloor \lambda \rfloor > \frac{(\lambda - \alpha d)}{(1 - \alpha)}$ , then we must have that

$$l^* = \left\lfloor \frac{(\lambda - \alpha d)}{(1 - \alpha)} \right\rfloor \text{ or } l^* = \left\lceil \frac{(\lambda - \alpha d)}{(1 - \alpha)} \right\rceil.$$

This case is the one highlighted in Feature 1 in which we may obtain  $d - l^* > \lceil d - \lambda \rceil$  and thus more servers are blocked compared to the large-deviations behavior observed in the regularly varying case. However, the blocked servers are symmetric in the sense that they are treated in exactly the same way.

In contrast, the second feature indicates that the typical trajectory leading to congestion may be obtained by blocking not only a specific amount to drive the system to instability, but also by blocking the corresponding servers by different loads in the large-deviations scaling. To appreciate this, we must assume that

$$1/\lambda < \gamma \leq 1/(\lambda - \lfloor \lambda \rfloor).$$

In this case, the contribution of the infimum in (3.2) becomes relevant. To illustrate that we can obtain solutions with the second feature, consider the case  $d = 2, 1 < \lambda < 2$ , and

$$1/\lambda < \gamma < 1/(\lambda - 1).$$

Choose  $\gamma = 1/(\lambda - 1) - \delta$  and  $\lambda = 2 - \delta^3$  for  $\delta > 0$  sufficiently small. We derive

$$\gamma^\alpha + (1 - \gamma(\lambda - 1))^\alpha = 1 - \delta\alpha + \delta^\alpha + o(\delta^2) \leq 2^{1-\alpha},$$

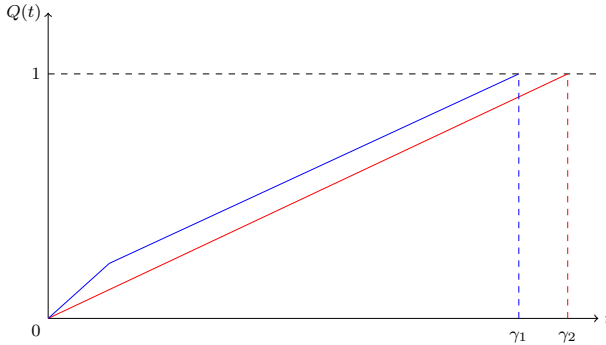
concluding that

$$\gamma^\alpha + (1 - \gamma(\lambda - 1))^\alpha < 2 \left( \frac{1}{\lambda} \right)^\alpha.$$

More explicitly, consider the case  $d = 2, \lambda = 1.49, \alpha = 0.1$  and  $\gamma = \frac{1}{\lambda - 1} - 0.1$ . For these values,  $\gamma_1^\alpha + (1 - \gamma_1(\lambda - 1))^\alpha < 2 \left( \frac{1}{\lambda} \right)^\alpha$ , and the most likely scenario leading to a large queue length is two big jobs arriving at the beginning and blocking both servers with different loads. On the other hand, if  $\gamma = \frac{1}{\lambda - 1}$ , the most likely scenario is a single big job blocking one server. These two scenarios are illustrated in Fig. 1.

We conclude this section by presenting a future research direction. We provide asymptotics only for the transient model of the queue length process  $Q$ . For a queue in steady state, more work is needed to overcome the technicalities arising with the large-deviations framework. Specifically, one has to prove that the interchange of limits as  $\gamma$  and  $n$  tend to infinity,

$$\lim_{\gamma \rightarrow \infty} \lim_{n \rightarrow \infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}(Q(\gamma n) > n) = \lim_{n \rightarrow \infty} \frac{1}{L(n)n^\alpha} \lim_{\gamma \rightarrow \infty} \log \mathbf{P}(Q(\gamma n) > n),$$



**Fig. 1** Most likely path for the queue buildup up to times  $\gamma_1 = \frac{1}{\lambda-1} - 0.1$  and  $\gamma_2 = \frac{1}{\lambda-1}$  when the number of servers is  $d = 2$ , the arrival rate is  $\lambda = 1.49$ , and the Weibull shape parameter of the service time is  $\alpha = 0.1$

is valid. We conjecture that the optimal value, similar to (3.2), of the variational problem associated with the steady-state model will consist solely of the term  $\min_{l=0}^{\lfloor \lambda \rfloor} \left\{ (d-l) \left( \frac{1}{\lambda-l} \right)^\alpha \right\}$ , obtained by taking  $\gamma = \infty$  in (1.2).

### 4 Proof of Theorem 3.1

We follow the general strategy outlined in the previous section. The first step consists of deriving the LDPs for  $\bar{A}_n, \bar{S}_n^{(i)}$  which subsequently provide us with the LDPs for  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$ . Let  $\mathbb{D}_p^\uparrow[0, \gamma/\mu]$  be the subspace of  $\mathbb{D}[0, \gamma/\mu]$  consisting of non-decreasing pure jump functions that assume nonnegative values at the origin, and define  $\zeta_\mu \in \mathbb{D}[0, \gamma/\mu]$  by  $\zeta_\mu(t) \triangleq \mu t$ . Let  $\mathbb{D}^\mu[0, \gamma/\mu] \triangleq \zeta_\mu + \mathbb{D}_p^\uparrow[0, \gamma/\mu]$  be the subspace of non-decreasing piecewise linear functions that have slope  $\mu$  and assume nonnegative values at the origin.

#### 4.1 Sample path large deviations for the components of the queue length upper bound

Recall that  $\bar{A}_n(t) = \frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} A_j$  and  $\bar{S}_n^{(i)}(t) = \frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} S_j^{(i)}$ .

**Proposition 4.1**  $\bar{A}_n$  satisfies the LDP on  $(\mathbb{D}[0, \gamma/\mathbf{E}A], d_{M_1'})$  with speed  $L(n)n^\alpha$  and good rate function

$$I_0(\xi) = \begin{cases} 0 & \text{if } \xi = \zeta_{\mathbf{E}A}, \\ \infty & \text{otherwise,} \end{cases} \tag{4.1}$$

and  $\bar{S}_n^{(i)}$  satisfies the LDP on  $(\mathbb{D}[0, \gamma], d_{M_1'})$  with speed  $L(n)n^\alpha$  and good rate function

$$I_i(\xi) = \begin{cases} \sum_{t \in [0, \gamma]} (\xi(t) - \xi(t-))^\alpha & \text{if } \xi \in \mathbb{D}^1[0, \gamma], \\ \infty & \text{otherwise.} \end{cases}$$

**Proof** In view of Lemma 3.2 of [6], it is easy to deduce that  $\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} (A_j - \mathbf{EA})$  satisfies the LDP on  $(\mathbb{D}[0, \gamma/\mathbf{EA}], d_{M_1'})$  with speed  $L(n)n^\alpha$  and with good rate function

$$I_A(\xi) = \begin{cases} 0 & \text{if } \xi = 0, \\ \infty & \text{otherwise.} \end{cases}$$

On the other hand, due to Result 2.2,  $\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} (S_j^{(i)} - 1)$  satisfies the LDP on  $(\mathbb{D}[0, \gamma], d_{M_1'})$  with good rate function

$$I_{S^{(i)}}(\xi) = \begin{cases} \sum_{t \in [0, \gamma]} (\xi(t) - \xi(t-))^\alpha & \text{if } \xi \in \mathbb{D}_p^\uparrow[0, \gamma], \\ \infty & \text{otherwise.} \end{cases} \tag{4.2}$$

Clearly,  $\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} A_j - t \cdot \mathbf{EA}$  and  $\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} S_j^{(i)} - t$  are exponentially equivalent to  $\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} (A_j - \mathbf{EA})$  and  $\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} (S_j^{(i)} - 1)$ , respectively. (For the definition of exponential equivalence, we refer to Definition A.1 in Appendix A.) Therefore,  $\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} A_j - t \cdot \mathbf{EA}$  and  $\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} S_j^{(i)} - t$  satisfy the LDPs with the good rate functions  $I_A$  and  $I_{S^{(i)}}$ , respectively.

Now, consider the map  $\Upsilon_\mu : (\mathbb{D}[0, \gamma/\mu], \mathcal{T}_{M_1'}) \rightarrow (\mathbb{D}[0, \gamma/\mu], \mathcal{T}_{M_1'})$ , where  $\Upsilon_\mu(\xi) \triangleq \xi + \zeta_\mu$ . Let  $I_0(\zeta) \triangleq \inf\{I_A(\xi) : \xi \in \mathbb{D}[0, \gamma/\mathbf{EA}], \zeta = \Upsilon_{\mathbf{EA}}(\xi)\}$ . From the form of  $I_A$ , it is easy to see that  $I_0$  coincides with the right-hand-side of (4.1). Since this map is continuous (Lemma B.1), the contraction principle (Result A.1) applies showing that  $\bar{A}_n = \Upsilon_{\mathbf{EA}}(\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} A_j - t \cdot \mathbf{EA})$  satisfies the desired LDP with the good rate function  $I_0$ . We next consider  $\bar{S}_n^{(i)}$ . Let  $I_i(\zeta) \triangleq \inf\{I_{S^{(i)}}(\xi) : \xi \in \mathbb{D}[0, \gamma], \zeta = \Upsilon_1(\xi)\}$ . Note that  $I_{S^{(i)}}(\xi) = \infty$  whenever  $\xi \notin \mathbb{D}_p^\uparrow$ , and  $\xi \in \mathbb{D}_p^\uparrow$  if and only if  $\zeta = \Upsilon_1(\xi)$  belongs to  $\mathbb{D}^1[0, \gamma]$ . Again, it is easy to check that  $I_i$  coincides with the right-hand-side of (4.2). We apply the contraction principle once more to conclude that  $\bar{S}_n^{(i)} = \Upsilon_1(\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} S_j^{(i)} - t)$  satisfies the desired LDP with good rate function  $I_i$ . □

To carry out the second step of our approach, we next prove that  $\Phi_{\mathbf{EA}}(\bar{A}_n)$  and  $\Phi_1(\bar{S}_n^{(i)})$  satisfy the same LDPs as  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$  for each  $i = 1, \dots, d$ , respectively. To show this, we next prove that  $\Phi_{\mathbf{EA}}(\bar{A}_n)$  and  $\Phi_1(\bar{S}_n^{(i)})$  are exponentially equivalent to  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$  for each  $i = 1, \dots, d$ , respectively.

**Proposition 4.2**  $\bar{M}_n$  and  $\Phi_{\mathbf{EA}}(\bar{A}_n)$  are exponentially equivalent in  $(\mathbb{D}[0, \gamma], \mathcal{T}_{M_1'})$ .  $\bar{N}_n^{(i)}$  and  $\Phi_1(\bar{S}_n^{(i)})$  are exponentially equivalent in  $(\mathbb{D}[0, \gamma], \mathcal{T}_{M_1'})$  for each  $i = 1, \dots, d$ .

**Proof** We first claim that  $d_{M'_1}(\bar{N}_n^{(i)}, \Phi_1(\bar{S}_n^{(i)})) \geq \epsilon$  implies either

$$\gamma - \Psi(\bar{S}_n^{(i)})(\gamma) \geq \frac{1}{2}\epsilon \quad \text{or} \quad \bar{N}_n^{(i)}(\gamma) - \gamma \geq \epsilon/2.$$

To see this, suppose otherwise. That is,

$$\gamma - \Psi(\bar{S}_n^{(i)})(\gamma) < \frac{1}{2}\epsilon \quad \text{and} \quad \bar{N}_n^{(i)}(\gamma) - \gamma < \epsilon/2. \tag{4.3}$$

By the construction of  $\bar{S}_n^{(i)}$  and  $\bar{N}_n^{(i)}$ , we see that  $\bar{N}_n^{(i)}(\cdot)$  is non-decreasing and  $\bar{N}_n^{(i)}(t) \geq \gamma$  for  $t \geq \Psi(\bar{S}_n^{(i)})(\gamma)$ . Therefore, the second condition of (4.3) implies

$$\sup_{t \in [\Psi(\bar{S}_n^{(i)})(\gamma), \gamma]} |\bar{N}_n^{(i)}(t) - \gamma| < \epsilon/2.$$

On the other hand, since the slope of  $\Phi_1(\bar{S}_n^{(i)})$  is 1 on  $[\Psi(\bar{S}_n^{(i)})(\gamma), \gamma]$ , the first condition of (4.3) implies that

$$\sup_{t \in [\Psi(\bar{S}_n^{(i)})(\gamma), \gamma]} |\Phi_1(\bar{S}_n^{(i)})(t) - \gamma| < \epsilon/2,$$

and hence,

$$\sup_{t \in [\Psi(\bar{S}_n^{(i)})(\gamma), \gamma]} |\Phi_1(\bar{S}_n^{(i)})(t) - \bar{N}_n^{(i)}(t)| < \epsilon. \tag{4.4}$$

Note also that, by the construction of  $\Phi_1$ ,  $\bar{N}_n^{(i)}(\cdot)$  and  $\Phi_1(\bar{S}_n^{(i)})(\cdot)$  coincide on  $[0, \Psi(\bar{S}_n^{(i)})(\gamma))$ . From this, along with (4.4), we see that

$$\sup_{t \in [0, \gamma]} |\Phi_1(\bar{S}_n^{(i)})(t) - \bar{N}_n^{(i)}(t)| < \epsilon,$$

which implies that  $d_{M'_1}(\Phi_1(\bar{S}_n^{(i)}), \bar{N}_n^{(i)}) < \epsilon$ . The claim is proved. Therefore,

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}\left(d_{M'_1}(\bar{N}_n^{(i)}, \Phi_1(\bar{S}_n^{(i)})) \geq \epsilon\right)}{L(n)n^\alpha} \\ & \leq \limsup_{n \rightarrow \infty} \frac{\log \left\{ \mathbf{P}\left(\gamma - \Psi(\bar{S}_n^{(i)})(\gamma) \geq \epsilon/2\right) + \mathbf{P}\left(\bar{N}_n^{(i)}(\gamma) - \gamma \geq \epsilon/2\right) \right\}}{L(n)n^\alpha} \\ & \leq \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}\left(\gamma - \Psi(\bar{S}_n^{(i)})(\gamma) \geq \epsilon/2\right)}{L(n)n^\alpha} \vee \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}\left(\bar{N}_n^{(i)}(\gamma) \geq \gamma + \epsilon/2\right)}{L(n)n^\alpha}, \end{aligned}$$

and we are done for the exponential equivalence between  $\bar{N}_n^{(i)}$  and  $\Phi_1(\bar{S}_n^{(i)})$  if we prove that

$$\limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}\left(\gamma - \Psi(\bar{S}_n^{(i)})(\gamma) \geq \epsilon/2\right)}{L(n)n^\alpha} = -\infty \tag{4.5}$$

and

$$\limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}\left(\bar{N}_n^{(i)}(\gamma) - \gamma \geq \epsilon/2\right)}{L(n)n^\alpha} = -\infty. \tag{4.6}$$

For (4.5), note that  $\Psi(\bar{S}_n^{(i)})(\gamma) \leq \gamma - \epsilon/2$  implies that  $\bar{S}_n^{(i)}(\gamma) \leq \gamma - \epsilon/2$ , and hence,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}\left(\gamma - \Psi(\bar{S}_n^{(i)})(\gamma) \geq \epsilon/2\right)}{L(n)n^\alpha} &\leq \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}\left(\bar{S}_n^{(i)}(\gamma) \leq \gamma - \epsilon/2\right)}{L(n)n^\alpha} \\ &\leq - \inf_{\xi(\gamma) \leq \gamma - \epsilon/2} I_0(\xi) \leq -\infty, \end{aligned}$$

where the second inequality is due to the LDP upper bound for  $\bar{S}_n^{(i)}$  in Proposition 4.1 and the continuity of the map  $\xi \mapsto \xi(\gamma)$  as a functional from  $(\mathbb{D}[0, \gamma], d_{M'})$  to  $\mathbb{R}$ . For (4.6), note that  $\bar{N}_n^{(i)}(\gamma) - \gamma \geq \epsilon/2$  implies  $\bar{S}_n^{(i)}(\gamma + \epsilon/2) \leq \gamma$ . Considering the LDP for  $\bar{S}_n^{(i)}$  on  $\mathbb{D}[0, \gamma + \epsilon/2]$ , we arrive at the same conclusion. This concludes the proof of the exponential equivalence between  $\bar{M}_n$  and  $\Phi_1(\bar{S}_n^{(i)})$ . The exponential equivalence between  $\bar{M}_n$  and  $\Phi_{EA}(\bar{A}_n)$  is essentially identical and, hence omitted.  $\square$

Due to the continuity of  $\Phi_\mu$  over the effective domain of the rate functions  $I_i, i = 1, \dots, d$ —see Proposition B.1—we can appeal to the extended contraction principle—see Remark 1—to establish LDPs for  $\Phi_{EA}(\bar{A}_n)$  and  $\Phi_1(\bar{S}_n^{(i)})$  for each  $i = 1, \dots, d$ . Our next proposition, which constitutes the third step of our strategy, characterizes the LDPs satisfied by  $\Phi_{EA}(\bar{A}_n)$  and  $\Phi_1(\bar{S}_n^{(i)})$ —and, hence, by  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$  as well. Define  $\check{\mathbb{C}}^\mu[0, \gamma] \triangleq \{\zeta \in \mathbb{C}[0, \gamma] : \zeta = \varphi_\mu(\xi) \text{ for some } \xi \in \mathbb{D}^\mu[0, \gamma/\mu]\}$ , where  $\mathbb{C}[0, \gamma]$  is the subspace of  $\mathbb{D}[0, \gamma]$  consisting of continuous paths, and  $\tau_s(\xi) = \max\left\{0, \sup\{t \in [0, \gamma] : \xi(t) = s\} - \inf\{t \in [0, \gamma] : \xi(t) = s\}\right\}$ .

**Proposition 4.3**  $\Phi_{EA}(\bar{A}_n)$  and  $\bar{M}_n$  satisfy the LDP with speed  $L(n)n^\alpha$  and good rate function

$$I'_0(\xi) \triangleq \begin{cases} 0 & \text{if } \xi = \zeta_{1/EA}, \\ \infty & \text{otherwise,} \end{cases}$$

and, for  $i = 1, \dots, d$ ,  $\Phi_1(\bar{S}_n^{(i)})$  and  $\bar{N}_n^{(i)}$  satisfy the LDP with speed  $L(n)n^\alpha$  and good rate function

$$I'_i(\xi) \triangleq \begin{cases} \sum_{s \in [0, \gamma]} \tau_s(\xi)^\alpha & \text{if } \xi \in \check{\mathbb{C}}^1[0, \gamma], \\ \infty & \text{otherwise.} \end{cases}$$

**Proof** Let  $\hat{I}'_0(\zeta) \triangleq \inf\{I_0(\xi) : \xi \in \mathbb{D}[0, \gamma/\mathbf{EA}], \zeta = \Phi_{\mathbf{EA}}(\xi)\}$  and  $\hat{I}'_i(\zeta) \triangleq \inf\{I_i(\xi) : \xi \in \mathbb{D}[0, \gamma], \zeta = \Phi_1(\xi)\}$  for  $i = 1, \dots, d$ . Recall that in Proposition 4.1 we established the LDP for  $\bar{A}_n$  and  $\bar{S}_n^{(i)}$  for each  $i = 1, \dots, d$ . Note that if  $\xi \in \mathcal{D}_{\Phi_{\mathbf{EA}}} \triangleq \{\xi \in \mathbb{D}[0, \gamma/\mathbf{EA}] : \Phi_{\mathbf{EA}}(\xi)(\gamma) - \Phi_{\mathbf{EA}}(\xi)(\gamma-) > 0\}$ , then there has to be  $s, t$  such that  $0 \leq s < t < \gamma/\mathbf{EA}$  and  $\Psi(\xi)(s) = \gamma$ . For such  $\xi$ ,  $I_0(\xi) = \infty$ . From this, along with Proposition B.1, we see that  $\Phi_{\mathbf{EA}}$  is continuous on the effective domain of  $I_0$ . Therefore, the extended contraction principle (see Remark 1 after Result A.1) applies, establishing the LDP for  $\Phi_{\mathbf{EA}}(\bar{A}_n)$  with rate function  $\hat{I}'_0$ . The LDP for  $\Phi_1(\bar{S}_n^{(i)})$  with rate function  $\hat{I}'_i$  follows from the same argument. Due to the exponential equivalence derived in Proposition 4.2,  $\bar{M}_n$  and  $\bar{N}_n^{(i)}$  satisfy the same LDP as  $\Phi_{\mathbf{EA}}(\bar{A}_n)$  and  $\Phi_1(\bar{S}_n^{(i)})$ . Therefore, we are done once we prove that the rate functions  $\hat{I}'_i$  deduced from the extended contraction principle satisfy  $I'_i = \hat{I}'_i$  for  $i = 0, \dots, d$ .

Starting with  $i = 0$ , note that  $I_0(\xi) = \infty$  if  $\xi \neq \zeta_{\mathbf{EA}}$ , and hence,

$$\hat{I}'_0(\zeta) = \inf\{I_0(\xi) : \xi \in \mathbb{D}[0, \gamma/\mathbf{EA}], \zeta = \Phi_{\mathbf{EA}}(\xi)\} = \begin{cases} 0 & \text{if } \zeta = \Phi_{\mathbf{EA}}(\zeta_{\mathbf{EA}}), \\ \infty & \text{otherwise,} \end{cases} \tag{4.7}$$

where it is straightforward to check that  $\Phi_{\mathbf{EA}}(\zeta_{\mathbf{EA}}) = \zeta_{1/\mathbf{EA}}$ . Therefore,  $I'_0 = \hat{I}'_0$ .

Turning to  $i = 1, \dots, d$ , note first that since  $I_i(\xi) = \infty$  for any  $\xi \notin \mathbb{D}^1[0, \gamma]$ ,

$$\hat{I}'_i(\zeta) = \inf\{I_i(\xi) : \xi \in \mathbb{D}^1[0, \gamma], \zeta = \Phi_1(\xi)\}.$$

Note also that  $\Phi_1$  can be simplified on  $\mathbb{D}^1[0, \gamma]$ : it is easy to check that if  $\xi \in \mathbb{D}^1[0, \gamma]$ ,  $\psi_1(\xi)(t) = \gamma$  and  $\varphi_1(\xi)(t) \leq \gamma$  for  $t \in [0, \gamma]$ . Therefore,  $\Phi_1(\xi) = \varphi_1(\xi)$ , and hence,

$$\hat{I}'_i(\zeta) = \inf\{I_i(\xi) : \xi \in \mathbb{D}^1[0, \gamma], \zeta = \varphi_1(\xi)\}.$$

Now, if we define  $\varrho_1 : \mathbb{D}[0, \gamma] \rightarrow \mathbb{D}[0, \gamma]$  by

$$\varrho_1(\xi)(t) \triangleq \begin{cases} \xi(t) & t \in [0, \varphi_1(\xi)(\gamma)) \\ \gamma + (t - \varphi_1(\xi)(\gamma)) & t \in [\varphi_1(\xi)(\gamma), \gamma] \end{cases},$$

then it is straightforward to check that  $I_i(\xi) \geq I_i(\varrho_1(\xi))$  and  $\varphi_1(\xi) = \varphi_1(\varrho_1(\xi))$  whenever  $\xi \in \mathbb{D}^1[0, \gamma]$ . Moreover,  $\varrho_1(\mathbb{D}^1[0, \gamma]) \subseteq \mathbb{D}^1[0, \gamma]$ . From these observations, we see that

$$\hat{I}'_i(\zeta) = \inf\{I_i(\xi) : \xi \in \varrho_1(\mathbb{D}^1[0, \gamma]), \zeta = \varphi_1(\xi)\}. \tag{4.8}$$

Note that  $\xi \in \varrho_1(\mathbb{D}^1[0, \gamma])$  and  $\zeta = \varphi_1(\xi)$  implies that  $\zeta \in \check{\mathbb{C}}^1[0, \gamma]$ . Therefore, in the case  $\zeta \notin \check{\mathbb{C}}^1[0, \gamma]$ , no  $\xi \in \mathbb{D}[0, \gamma]$  satisfies the two conditions simultaneously, and hence,

$$\hat{I}'_i(\zeta) = \inf \emptyset = \infty = I'_i(\zeta). \tag{4.9}$$

Now we prove that  $\hat{I}'_i(\zeta) = I'_i(\zeta)$  for  $\zeta \in \check{\mathbb{C}}^1[0, \gamma]$ . We claim that, if  $\xi \in \varrho_1(\mathbb{D}^1[0, \gamma])$ ,

$$\tau_s(\varphi_1(\xi)) = \xi(s) - \xi(s-)$$

for all  $s \in [0, \gamma]$ . The proof of this claim is provided at the end of the proof of the current proposition. Using this claim,

$$\begin{aligned} \hat{I}'_i(\zeta) &= \inf \left\{ \sum_{s \in [0, \gamma]} (\xi(s) - \xi(s-))^\alpha : \xi \in \varrho_1(\mathbb{D}^1[0, \gamma]), \zeta = \varphi_1(\xi) \right\} \\ &= \inf \left\{ \sum_{s \in [0, \gamma]} \tau_s(\varphi_1(\xi))^\alpha : \xi \in \varrho_1(\mathbb{D}^1[0, \gamma]), \zeta = \varphi_1(\xi) \right\} \\ &= \inf \left\{ \sum_{s \in [0, \gamma]} \tau_s(\zeta)^\alpha : \xi \in \varrho_1(\mathbb{D}^1[0, \gamma]), \zeta = \varphi_1(\xi) \right\}. \end{aligned}$$

Note also that  $\zeta \in \check{\mathbb{C}}^1[0, \gamma]$  implies the existence of  $\xi$  such that  $\zeta = \varphi_1(\xi)$  and  $\xi \in \varrho_1(\mathbb{D}^1[0, \gamma])$ . To see why, note that there exists  $\xi' \in \mathbb{D}^1[0, \gamma]$  such that  $\zeta = \varphi_1(\xi')$  due to the definition of  $\check{\mathbb{C}}^1[0, \gamma]$ . Let  $\xi \triangleq \varrho_1(\xi')$ . Then,  $\zeta = \varphi_1(\xi)$  and  $\xi \in \varrho_1(\mathbb{D}^1[0, \gamma])$ . From this observation, we see that

$$\left\{ \sum_{s \in [0, \gamma]} \tau_s(\zeta)^\alpha : \xi \in \varrho_1(\mathbb{D}^1[0, \gamma]), \zeta = \varphi_1(\xi) \right\} = \left\{ \sum_{s \in [0, \gamma]} \tau_s(\zeta)^\alpha \right\},$$

and hence,

$$\hat{I}'_i(\zeta) = \sum_{s \in [0, \gamma]} \tau_s(\zeta)^\alpha = I'_i(\zeta) \tag{4.10}$$

for  $\zeta \in \check{\mathbb{C}}^1[0, \gamma]$ . From (4.9) and (4.10), we conclude that  $I'_i = \hat{I}'_i$  for  $i = 1, \dots, d$ .

All that remains is to prove that  $\tau_s(\varphi_1(\xi)) = \xi(s) - \xi(s-)$  for all  $s \in [0, \gamma]$ . We consider the cases  $s > \varphi_1(\xi)(\gamma)$  and  $s \leq \varphi_1(\xi)(\gamma)$  separately. First, suppose that  $s > \varphi_1(\xi)(\gamma)$ . Since  $\varphi_1(\xi)$  is non-decreasing, this means that  $\varphi_1(\xi)(t) < s$  for all  $t \in [0, \gamma]$ , and hence  $\{t \in [0, \gamma] : \varphi_1(t) = s\} = \emptyset$ . Therefore,

$$\begin{aligned} \tau_s(\varphi_1(\xi)) &= 0 \vee \left( \sup\{t \in [0, \gamma] : \varphi_1(t) = s\} \right. \\ &\quad \left. - \inf\{t \in [0, \gamma] : \varphi_1(t) = s\} \right) = 0 \vee (-\infty - \infty) = 0. \end{aligned}$$

On the other hand, since  $\xi$  is continuous on  $[\varphi_1(\xi)(\gamma), \gamma]$  by its construction,

$$\xi(s) - \xi(s-) = 0.$$



Therefore,

$$\tau_s(\varphi_1(\xi)) = 0 = \xi(s) - \xi(s-)$$

for  $s > \varphi_1(\xi)(\gamma)$ .

Now we turn to the case  $s \leq \varphi_1(\xi)(\gamma)$ . Since  $\varphi_1(\xi)$  is continuous, this implies that there exists  $u \in [0, \gamma]$  such that  $\varphi_1(\xi)(u) = s$ . From the definition of  $\varphi_1(\xi)(u)$ , it is straightforward to check that

$$u \in [\xi(s-), \xi(s)] \iff s = \varphi_1(\xi)(u). \tag{4.11}$$

Note that  $[\xi(s-), \xi(s)] \subseteq [0, \gamma]$  for  $s \leq \varphi_1(\xi)(\gamma)$  due to the construction of  $\xi$ . Therefore, the above equivalence (4.11) implies that  $[\xi(s-), \xi(s)] = \{u \in [0, \gamma] : \varphi_1(\xi)(u) = s\}$ , which in turn implies that  $\xi(s-) = \inf\{u \in [0, \gamma] : \varphi_1(\xi)(u) = s\}$  and  $\xi(s) = \sup\{u \in [0, \gamma] : \varphi_1(\xi)(u) = s\}$ . We conclude that

$$\tau_s(\varphi_1(\xi)) = \xi(s) - \xi(s-)$$

for  $s \leq \varphi_1(\xi)(\gamma)$ . □

### 4.2 Large deviations for the queue length

Now we are ready to follow step 4) of our outlined strategy and characterize the log asymptotics of  $\mathbf{P}(Q(\gamma n) > n)$ . Recall that  $\tau_s(\xi) \triangleq \max\left\{0, \sup\{t \in [0, \gamma] : \xi(t) = s\} - \inf\{t \in [0, \gamma] : \xi(t) = s\}\right\}$ .

#### Proposition 4.4

$$\lim_{n \rightarrow \infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}(Q(\gamma n) > n) = -c^*,$$

where  $c^*$  is the solution of the following variational problem:

$$\begin{aligned} & \inf_{\xi_1, \dots, \xi_d} \sum_{i=1}^d \sum_{s \in [0, \gamma]} \tau_s(\xi_i)^\alpha \\ & \text{subject to } \sup_{0 \leq s \leq \gamma} \left( \frac{s}{\mathbf{EA}} - \sum_{i=1}^d \xi_i(s) \right) \geq 1; \\ & \xi_i \in \check{\mathcal{C}}^1[0, \gamma] \text{ for } i = 1, \dots, d. \end{aligned} \tag{4.12}$$

**Proof** From Corollary 1 of [13], for any  $\epsilon > 0$ ,

$$\begin{aligned}
 & \mathbf{P}(Q(\gamma n) > n) \\
 & \leq \mathbf{P}\left(\sup_{0 \leq s \leq \gamma} \left\{ \bar{M}_n(\gamma) - \bar{M}_n(s) - \sum_{i=1}^d (\bar{N}_n^{(i)}(\gamma) - \bar{N}_n^{(i)}(s)) \right\} \geq 1\right) \\
 & \leq \mathbf{P}\left(\sup_{0 \leq s \leq \gamma} \left\{ \bar{M}_n(\gamma) - \bar{M}_n(s) - \frac{\gamma - s}{\mathbf{EA}} \right\} \right. \\
 & \quad \left. + \sup_{0 \leq s \leq \gamma} \left\{ \frac{\gamma - s}{\mathbf{EA}} - \sum_{i=1}^d (\bar{N}_n^{(i)}(\gamma) - \bar{N}_n^{(i)}(s)) \right\} \geq 1\right) \\
 & \leq \underbrace{\mathbf{P}\left(\bar{M}_n(\gamma) - \frac{\gamma}{\mathbf{EA}} \geq \epsilon\right)}_{\text{(I)}} \\
 & \quad + \underbrace{\mathbf{P}\left(-\inf_{0 \leq s \leq \gamma} \left(\bar{M}_n(s) - \frac{s}{\mathbf{EA}}\right) \geq \epsilon\right)}_{\text{(II)}} \\
 & \quad + \underbrace{\mathbf{P}\left(\sup_{0 \leq s \leq \gamma} \left\{ \frac{\gamma - s}{\mathbf{EA}} - \sum_{i=1}^d (\bar{N}_n^{(i)}(\gamma) - \bar{N}_n^{(i)}(s)) \right\} \geq 1 - 2\epsilon\right)}_{\text{(III)}}.
 \end{aligned}$$

By the LDP for  $\bar{M}_n$  (Proposition 4.3), it is straightforward to deduce that

$$\limsup_{n \rightarrow \infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}\left(\bar{M}_n(\gamma) - \frac{\gamma}{\mathbf{EA}} \geq \epsilon\right) = -\infty$$

and

$$\limsup_{n \rightarrow \infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}\left(-\inf_{0 \leq s \leq \gamma} \left(\bar{M}_n(s) - \frac{s}{\mathbf{EA}}\right) \geq \epsilon\right) = -\infty.$$

Therefore, by the principle of the maximum term,

$$\begin{aligned}
 & \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}(Q(\gamma n) > n)}{L(n)n^\alpha} \\
 & \leq \max \left\{ \limsup_{n \rightarrow \infty} \frac{\log \text{(I)}}{L(n)n^\alpha}, \limsup_{n \rightarrow \infty} \frac{\log \text{(II)}}{L(n)n^\alpha}, \limsup_{n \rightarrow \infty} \frac{\log \text{(III)}}{L(n)n^\alpha} \right\} \\
 & = \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}\left(\sup_{0 \leq s \leq \gamma} \left\{ \frac{\gamma - s}{\mathbf{EA}} - \sum_{i=1}^d (\bar{N}_n^{(i)}(\gamma) - \bar{N}_n^{(i)}(s)) \right\} \geq 1 - 2\epsilon\right)}{L(n)n^\alpha}.
 \end{aligned}$$

To bound the limit supremum in the equality above, we derive an LDP for

$$\frac{\gamma}{\mathbf{EA}} - \sum_{i=1}^d \bar{N}_n^{(i)}(\gamma) + \sup_{0 \leq s \leq \gamma} \left( \sum_{i=1}^d \bar{N}_n^{(i)}(s) - \frac{s}{\mathbf{EA}} \right).$$

Due to Proposition 4.3 and Theorem 4.14 of [18],  $(\bar{N}_n^{(1)}, \dots, \bar{N}_n^{(d)})$  satisfy the LDP in  $\prod_{i=1}^d \mathbb{D}[0, \gamma]$  (w.r.t. the  $d$ -fold product topology of  $\mathcal{T}_{M'_1}$ ) with speed  $L(n)n^\alpha$  and rate function

$$I'(\xi_1, \dots, \xi_d) \triangleq \sum_{i=1}^d I'_i(\xi_i).$$

Let  $\mathbb{D}^\uparrow[0, \gamma]$  denote the subspace of  $\mathbb{D}[0, \gamma]$  consisting of non-decreasing functions. Since  $\bar{N}_n^{(i)} \in \mathbb{D}^\uparrow[0, \gamma]$  with probability 1 for each  $i = 1, \dots, d$ , we can apply Lemma 4.1.5 (b) of [17] to deduce the same LDP for  $(\bar{N}_n^{(1)}, \dots, \bar{N}_n^{(d)})$  in  $\prod_{i=1}^d \mathbb{D}^\uparrow[0, \gamma]$ . We define  $f_1 : \prod_{i=1}^d \mathbb{D}^\uparrow[0, \gamma] \rightarrow \mathbb{D}[0, \gamma]$  as

$$f_1(\xi_1, \dots, \xi_d) \triangleq \sum_{i=1}^d \xi_i - \zeta_{1/\mathbf{EA}}.$$

Note that  $f_1$  is continuous since all the jumps are in one direction in its domain. Since the supremum functional  $f_2 : \xi \mapsto \sup_{0 \leq s \leq \gamma} \xi(s)$  is continuous in the range of  $f_1$ —see Lemma (B.2)— $f_2 \circ f_1$  is a continuous map as well. The functional  $f_3 : \xi \mapsto \xi(\gamma)$  is also continuous w.r.t. the  $M'_1$  topology on  $\mathbb{D}[0, \gamma]$  due to Lemma B.2. Therefore, the continuous map  $f : \prod_{i=1}^d \mathbb{D}^\uparrow[0, \gamma] \rightarrow \mathbb{R}$ , where

$$f(\xi_1, \dots, \xi_d) \triangleq \frac{\gamma}{\mathbf{EA}} - \sum_{i=1}^d f_3(\xi_i) + f_2 \circ f_1(\xi_1, \dots, \xi_d),$$

is continuous and, hence, we can apply the contraction principle with  $f$  to establish the LDP for  $f(\bar{N}_n^{(1)}, \dots, \bar{N}_n^{(d)}) = \frac{\gamma}{\mathbf{EA}} - \sum_{i=1}^d \bar{N}_n^{(i)}(\gamma) + \sup_{0 \leq s \leq \gamma} \left( \sum_{i=1}^d \bar{N}_n^{(i)}(s) - \frac{s}{\mathbf{EA}} \right)$ . The LDP is controlled by the good rate function

$$I''(x) \triangleq \inf \left\{ I'(\xi_1, \dots, \xi_d) : \frac{\gamma}{\mathbf{EA}} - \sum_{i=1}^d \xi_i(\gamma) + \sup_{0 \leq s \leq \gamma} \left( \sum_{i=1}^d \xi_i(s) - \frac{s}{\mathbf{EA}} \right) = x \right\}.$$

Note that since  $I'(\xi) = \infty$  for  $\xi \notin \check{\mathbb{C}}^1[0, \gamma]$ , and  $\xi(\cdot) \in \check{\mathbb{C}}^1[0, \gamma]$  if and only if  $\xi(\gamma) - \xi(\gamma - \cdot) \in \check{\mathbb{C}}^1[0, \gamma]$ ,

$$\begin{aligned}
 I''(x) &= \inf \left\{ I'(\xi_1, \dots, \xi_d) : \frac{\gamma}{\mathbf{EA}} - \sum_{i=1}^d \xi_i(\gamma) \right. \\
 &\quad \left. + \sup_{0 \leq s \leq \gamma} \left( \sum_{i=1}^d \xi_i(s) - \frac{s}{\mathbf{EA}} \right) = x, \xi_i \in \check{\mathcal{C}}^1[0, \gamma] \right\} \\
 &= \inf \left\{ I'(\xi_1, \dots, \xi_d) : \sup_{0 \leq s \leq \gamma} \left\{ \frac{\gamma - s}{\mathbf{EA}} \right. \right. \\
 &\quad \left. \left. - \sum_{i=1}^d (\xi_i(\gamma) - \xi_i(s)) \right\} = x, \xi_i \in \check{\mathcal{C}}^1[0, \gamma] \right\} \\
 &= \inf \left\{ I'(\xi_1, \dots, \xi_d) : \sup_{0 \leq s \leq \gamma} \left\{ \frac{s}{\mathbf{EA}} \right. \right. \\
 &\quad \left. \left. - \sum_{i=1}^d (\xi_i(\gamma) - \xi_i(\gamma - s)) \right\} = x, \xi_i \in \check{\mathcal{C}}^1[0, \gamma] \right\} \\
 &= \inf \left\{ I'(\xi_1, \dots, \xi_d) : \sup_{0 \leq s \leq \gamma} \left\{ \frac{s}{\mathbf{EA}} \right. \right. \\
 &\quad \left. \left. - \sum_{i=1}^d (\xi_i(s)) \right\} = x, \xi_i \in \check{\mathcal{C}}^1[0, \gamma] \right\}.
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}(Q(\gamma n) > n)}{L(n)n^\alpha} &\leq \limsup_{n \rightarrow \infty} \frac{\log \mathbf{P}(f(\bar{N}_n^{(1)}, \dots, \bar{N}_n^{(d)}) \geq 1 - 2\epsilon)}{L(n)n^\alpha} \\
 &\leq - \inf_{x \in [1 - 2\epsilon, \infty)} I''(x) \\
 &= - \inf \left\{ \sum_{i=1}^d \sum_{s \in [0, \gamma]} \tau_s(\xi_i)^\alpha : \sup_{0 \leq s \leq \gamma} \left( \frac{s}{\mathbf{EA}} - \sum_{i=1}^d \xi_i(s) \right) \geq 1 - 2\epsilon, \xi_i \in \check{\mathcal{C}}^1[0, \gamma] \right\}.
 \end{aligned}$$

Taking  $\epsilon \rightarrow 0$ , we see that  $-c^*$  is the upper bound for the left-hand side.

We move on to the matching lower bound in the case  $\gamma > 1/\lambda$ . Considering the obvious coupling between  $Q$  and  $(M, N^{(1)}, \dots, N^{(d)})$ , one can see that  $M(s) - \sum_{i=1}^d N^{(i)}(s)$  can be interpreted as (a lower bound of) the length of an imaginary queue at time  $s$  where the servers can start working on the jobs that have not arrived yet. Therefore,  $\mathbf{P}(Q((a + s)n) > n) \geq \mathbf{P}(Q((a + s)n) > n | Q(a) = 0) \geq \mathbf{P}(\bar{M}_n(s) - \sum_{i=1}^d \bar{N}_n^{(i)}(s) > 1)$  for any  $a \geq 0$ . Let  $s^*$  be the level crossing time of the optimal solution of (4.12). Then, for any  $\epsilon > 0$ ,

$$\begin{aligned}
 \mathbf{P}(Q(\gamma n) > n) &\geq \mathbf{P}\left(\bar{M}_n(s^*) - \sum_{i=1}^d \bar{N}_n^{(i)}(s^*) > 1\right) \\
 &\geq \mathbf{P}\left(\bar{M}_n(s^*) - s^*/\mathbf{EA} > -\epsilon \text{ and } s^*/\mathbf{EA} - \sum_{i=1}^d \bar{N}_n^{(i)}(s^*) > 1 + \epsilon\right) \\
 &\geq \mathbf{P}\left(s^*/\mathbf{EA} - \sum_{i=1}^d \bar{N}_n^{(i)}(s^*) > 1 + \epsilon\right) - \mathbf{P}\left(\bar{M}_n(s^*) - s^*/\mathbf{EA} \leq -\epsilon\right).
 \end{aligned}
 \tag{4.13}$$

Due to Proposition 4.3,

$$\limsup_{n \rightarrow \infty} \frac{1}{L(n)n^\alpha} \log \mathbf{P}(\bar{M}_n(s^*) - s^*/\mathbf{EA} \leq -\epsilon) = -\infty,$$

and hence, due to (4.13), it is straightforward to deduce that

$$\begin{aligned}
 \liminf_{n \rightarrow \infty} \frac{\log \mathbf{P}(Q(\gamma n) > n)}{L(n)n^\alpha} &\geq \liminf_{n \rightarrow \infty} \frac{\log \mathbf{P}(s^*/\mathbf{EA} - \sum_{i=1}^d \bar{N}_n^{(i)}(s^*) > 1 + \epsilon)}{L(n)n^\alpha} \\
 &\geq - \inf_{(\xi_1, \dots, \xi_d) \in A^\circ} I'(\xi_1, \dots, \xi_d),
 \end{aligned}$$

where  $A = \{(\xi_1, \dots, \xi_d) : s^*/\mathbf{EA} - \sum_{i=1}^d \xi_i(s^*) > 1 + \epsilon\}$ . Note that the optimizer  $(\xi_1^*, \dots, \xi_d^*)$  of (4.12) satisfies  $s^*/\mathbf{EA} - \sum_{i=1}^d \xi_i^*(s^*) \geq 1$ . Consider  $(\xi'_1, \dots, \xi'_d)$  obtained by increasing one of the job sizes of  $(\xi_1^*, \dots, \xi_d^*)$  by  $\delta > 0$ . One can always find a small enough such  $\delta$  since  $\gamma > 1/\lambda$ . Note that there exists  $\epsilon > 0$  such that  $s'/\mathbf{EA} - \sum_{i=1}^d \xi'_i(s') > 1 + \epsilon$ . Therefore,

$$\liminf_{n \rightarrow \infty} \frac{\log \mathbf{P}(Q(\gamma n) > n)}{L(n)n^\alpha} \geq -I'(\xi'_1, \dots, \xi'_d) \geq -c^* - \delta^\alpha,$$

where the second inequality is from the subadditivity of  $x \mapsto x^\alpha$ . Since  $\delta$  can be chosen arbitrarily small, letting  $\delta \rightarrow 0$ , we arrive at the matching lower bound.  $\square$

### 4.3 Explicit solution of the variational problem associated with the queue length

We now simplify the expression of  $c^*$  given in Proposition 4.4.

**Proposition 4.5** *If  $\gamma < 1/\lambda$ ,  $c^* = \infty$ . If  $\gamma \geq 1/\lambda$ ,  $c^*$  can be computed via*

$$\begin{aligned}
 &\min_{x_1, \dots, x_d} \sum_{i=1}^d x_i^\alpha \\
 &\text{subject to } \sup_{s \in [0, \gamma]} \left\{ \lambda s - \sum_{i=1}^d (s - x_i)^+ \right\} \geq 1,
 \end{aligned}$$

$$x_1, \dots, x_d \geq 0, \tag{4.14}$$

which in turn equals

$$\min \left\{ \inf_{0 < k \leq \lfloor \lambda \rfloor; \gamma < 1/(\lambda - k)} \left\{ (d - k) \gamma^\alpha + (1 - \gamma \lambda + \gamma k)^\alpha (k - \lfloor \lambda - 1/\gamma \rfloor)^{1-\alpha} \right\}, \right. \\ \left. \min_{l=0}^{\lfloor \lambda - 1/\gamma \rfloor} \left\{ (d - l) \left( \frac{1}{\lambda - l} \right)^\alpha \right\} \right\}. \tag{4.15}$$

**Proof** Recall that  $\mathbb{D}^1[0, \gamma]$  is the subspace of the Skorokhod space and consists of non-decreasing piecewise linear functions with slope 1 almost everywhere over the time horizon  $[0, \gamma]$  and nonnegative values at the origin. Recall  $\varphi_1(\cdot)$  defined in (2.5) as well. From these definitions, it is easy to see that Proposition 4.4 implies that the constant  $c^*$  is equal to

$$\inf_{\zeta_1, \dots, \zeta_d} \sum_{i=1}^d \sum_{s \in [0, \gamma]} \tau_s(\zeta_i)^\alpha \\ \text{subject to } \sup_{0 \leq s \leq \gamma} \left( \lambda s - \sum_{i=1}^d \zeta_i(s) \right) \geq 1, \\ \zeta_i = \varphi_1(\xi_i), \quad \xi_i \in \mathbb{D}^1[0, \gamma] \quad \text{for } i = 1, \dots, d. \tag{4.16}$$

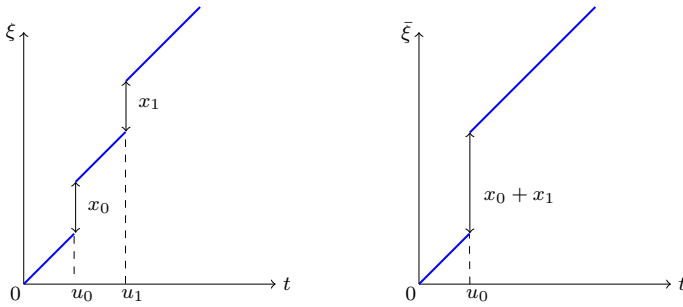
Note that this is an infinite-dimensional (functional) optimization problem. We reduce this optimization problem to a more standard problem in two main steps:

1. We first show that it suffices to optimize over  $\xi_i$  of the form  $\xi_i(t) = t + x_0$  for some  $x_0 \geq 0$ .
2. Next, we reduce the infinite-dimensional problem over the previously mentioned set into a finite-dimensional optimization problem where the aim is to minimize a concave function over a compact polyhedral set. This allows us to invoke Corollary 32.3.1 of [20], which enables us to calculate the optimal solution by finding the extreme points of the feasible region.

*Step 1* Suppose that  $(\zeta_1, \dots, \zeta_d)$  is an optimal solution associated with (4.16) and recall that  $\zeta_i = \varphi_1(\xi_i)$ . We now claim that the corresponding functions  $\xi_1, \dots, \xi_d$  have at most one jump. We prove this by contradiction. Assume that at least one of the  $\xi_i$  exhibits two jumps at times  $u_0$  and  $u_1$  of size  $x_0$  and  $x_1$ , respectively, with  $0 \leq u_0 < u_1 \leq \gamma$ . Let

$$\bar{\xi}_i(\cdot) = \xi_i(\cdot) - x_1 \mathbb{I}_{[u_1, \gamma]}(\cdot) + x_1 \mathbb{I}_{[u_0, \gamma]}(\cdot).$$

Intuitively, we constructed a new path,  $\bar{\xi}_i(\cdot)$  by merging the two jumps into a big jump at time  $u_0$ . Since  $x_0, x_1$  are nonnegative, then we have that



**Fig. 2** The two figures above depict the graphs of two jump functions,  $\xi$  and  $\bar{\xi}$ . By merging the two jumps of  $\xi$  into one big jump, at time  $u_0$ , the resulting step function  $\bar{\xi}$  is bigger than or equal to  $\xi$

$$\bar{\xi}_i(t) \geq \xi_i(t), \quad \forall t > 0.$$

Figure 2 illustrates this.

Now, let  $\bar{\zeta}_i = \varphi_1(\bar{\xi}_i)$ . From the definition of  $\varphi_1$ , we obviously have that

$$\bar{\zeta}_i(s) \leq \zeta_i(s) \quad \text{for } s \in [0, \gamma]. \tag{4.17}$$

Therefore, due to (4.17),  $(\zeta_1, \dots, \zeta_{i-1}, \bar{\zeta}_i, \zeta_{i+1}, \dots, \zeta_d)$  is also a feasible solution for (4.16). Moreover, by the following observation:

$$\sum_{s \in [0, \gamma]} \tau_s (\bar{\zeta}_i)^\alpha = \sum_{s \in [0, \gamma]} \tau_s (\zeta_i)^\alpha + (x_0 + x_1)^\alpha - x_0^\alpha - x_1^\alpha,$$

along with the fact that  $(x_0 + x_1)^\alpha < x_0^\alpha + x_1^\alpha$ , we deduce that  $(\zeta_1, \dots, \zeta_{i-1}, \bar{\zeta}_i, \zeta_{i+1}, \dots, \zeta_d)$  strictly improves the value of the objective function in (4.16). That is,  $(\zeta_1, \dots, \zeta_d)$  cannot be an optimal solution. The argument can be iterated when  $\xi_i$  exhibits more than two jumps.

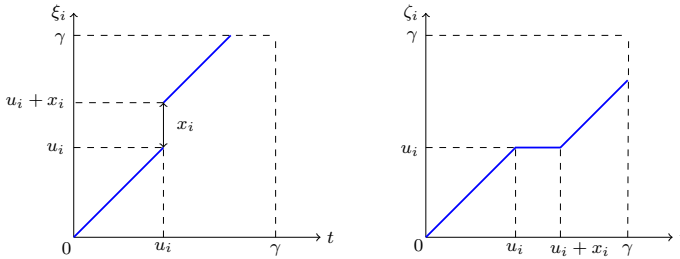
In conclusion, we proceed assuming that every  $\xi_i(\cdot)$  has a single jump of size  $x_i > 0$  at some time  $u_i \in [0, \gamma]$ , and hence we can use the following representation (Fig. 3):

$$\zeta_i(s) = \min(s, u_i) + (s - x_i - u_i)^+, \quad \text{for } i = 1, \dots, d. \tag{4.18}$$

To complete the first step of our construction, we show that, without loss of generality, jumps can be assumed to occur at time 0. Suppose that  $u_i > 0$  for some  $i \in \{1, \dots, d\}$ . Define

$$\xi'_i(s) = \xi_i(s) - x_i \mathbb{I}_{[u_i, \gamma]}(s) + x_i \mathbb{I}_{[0, \gamma]}(s).$$

We constructed a new path  $\xi'$  by moving a jump time to 0. Again, it is easy to verify that  $\xi'(s) \geq \xi(s)$  for all  $s \in [0, \gamma]$ , and if we let  $\zeta'_i = \varphi_1(\xi'_i)$ , then  $\zeta'_i(s) \leq \zeta_i(s)$  for all  $s \in [0, \gamma]$ . Consequently, we preserve feasibility without increasing the value of the objective function in (4.16). Therefore, w.l.o.g. we can assume that  $\xi_i$  that correspond to the optimal solution of (4.16) are those paths that have at most one discontinuity at



**Fig. 3** The pictures above depict the graph of a function  $\xi_i$  in  $\mathbb{D}^1[0, \gamma]$  and the graph of the function  $\zeta_i = \varphi_1(\xi_i)$ . The function  $\xi_i$  has one jump of size  $x_i$  and this translates to a flat line under the transformation  $\varphi_1$ . In conclusion, we infer that  $\zeta_i$  has the representation  $\zeta_i(s) = \min(s, u_i) + (s - x_i - u_i)^+$

time zero and then they linearly increase with slope 1. That is, the solution  $(\zeta_1, \dots, \zeta_d)$  takes the following form: for each  $i = 1, \dots, d$ ,

$$\zeta_i(s) = (s - x_i)^+ \quad \text{for some } x_i \geq 0. \tag{4.19}$$

*Step 2* Thanks to the reduction in (4.19), we see that for each  $i = 1, \dots, d$  we have that  $\tau_0(\zeta_i) = x_i$ , while  $\tau_s(\zeta_i) = 0$  for every  $s > 0$ . Thus, we see that (4.12) takes the form

$$\begin{aligned} & \min_{x_1, \dots, x_d} \sum_{i=1}^d x_i^\alpha \\ & \text{subject to} \quad \sup_{s \in [0, \gamma]} \left\{ \lambda s - \sum_{i=1}^d (s - x_i)^+ \right\} \geq 1, \\ & \quad x_1, \dots, x_d \in [0, \gamma]. \end{aligned} \tag{4.20}$$

We continue simplifying the optimization problem in (4.20), reducing it to a polyhedral optimization problem. Let  $x = (x_1, \dots, x_d)$  be an optimal solution so that its coordinates are sorted in increasing order:  $0 \leq x_1 \leq \dots \leq x_d \leq \gamma$ . Note that the supremum of  $l(s; x) \triangleq \lambda s - \sum_{i=1}^d (s - x_i)^+$  over  $s \in [0, \gamma]$  cannot be obtained strictly before  $x_d$ , since in such a case, a sufficiently small perturbation of  $x_d$  to its left leads to a strictly smaller value of the objective function without changing the supremum of  $l(s; x)$ , which is a contradiction to the assumption that  $x$  is an optimal solution. On the other hand, from the stability assumption  $\lambda < d$ , the slope of  $l(s; x)$  is negative after  $x_d$ , and hence its supremum cannot be obtained strictly after  $x_d$ . Therefore, the supremum of  $l(s; x)$  has to be attained at  $s = x_d$ . Now, set  $a_1 = x_1$  and  $a_i = x_i - x_{i-1}$  for  $i = 2, \dots, d$ . Then,  $x_i = a_1 + \dots + a_i$  for  $i = 1, \dots, d$ , and  $l(x_d; x) = \lambda(a_1 + \dots + a_d) - \sum_{i=1}^d (a_1 + \dots + a_d - \sum_{j=1}^i a_j)$ , and hence (4.20) is equivalent to



$$\begin{aligned} & \min_{a_1, \dots, a_d} \sum_{i=1}^d \left( \sum_{j=1}^i a_j \right)^\alpha \\ & \text{subject to } \lambda (a_1 + \dots + a_d) - \sum_{i=1}^d (a_1 + \dots + a_d - \sum_{j=1}^i a_j) \geq 1, \\ & a_1 + \dots + a_d \leq \gamma, a_1, \dots, a_d \geq 0, \end{aligned}$$

and by simplifying the constraints we arrive at

$$\begin{aligned} & \min_{a_1, \dots, a_d} \sum_{i=1}^d \left( \sum_{j=1}^i a_j \right)^\alpha \\ & \text{subject to } \lambda a_1 + (\lambda - 1) a_2 + \dots + (\lambda - d + 1) a_d \geq 1, \\ & a_1 + \dots + a_d \leq \gamma, a_1, \dots, a_d \geq 0. \end{aligned}$$

Recall  $0 < \lambda < d$ , and let  $m$  be any of the integers in the set  $\{1, \dots, d - 1\}$ . If  $(\lambda - m) < 0$ , we deduce that  $a_{m+1} = 0$ . If this was not the case, we could construct a feasible solution which reduces the value of the objective function and also satisfies the previously mentioned conditions. That is, the variational problem has an even simpler representation than the one above:

$$\min_{a_1, \dots, a_d} \sum_{i=1}^{\lfloor \lambda \rfloor} \left( \sum_{j=1}^i a_j \right)^\alpha + (d - \lfloor \lambda \rfloor) \left( \sum_{j=1}^{\lfloor \lambda \rfloor + 1} a_j \right)^\alpha \tag{4.21}$$

$$\text{subject to } \lambda a_1 + (\lambda - 1) a_2 + \dots + (\lambda - \lfloor \lambda \rfloor) a_{\lfloor \lambda \rfloor + 1} = 1, \tag{4.22}$$

$$a_1 + \dots + a_{\lfloor \lambda \rfloor + 1} \leq \gamma, \tag{4.23}$$

$$a_1, \dots, a_{\lfloor \lambda \rfloor + 1} \geq 0. \tag{4.24}$$

Recall that  $c^* = \infty$  if  $\gamma < 1/\lambda$ . Assuming  $\gamma > 1/\lambda$ , we recover the optimal solution by evaluating the extreme points associated with the polyhedron described by the constraints (4.22), (4.23), and (4.24). The objective function in (4.21) is concave and lower bounded inside the feasible region. In addition, the feasible region is a compact polyhedron. Therefore, the optimizer is achieved at some extreme point in the feasible region (see Corollary 32.3.1 [20]).

Depending on the value of  $\gamma$ , we indicate how to compute the basic feasible solutions related to (4.21). Firstly, we treat the case  $\gamma > 1/(\lambda - \lfloor \lambda \rfloor)$ , where  $\lambda$  is not an integer. After that, we treat the general case  $\gamma > 1/\lambda$ . Given that  $\lambda > \lfloor \lambda \rfloor$ , observe that if  $\gamma \geq 1/(\lambda - \lfloor \lambda \rfloor)$ , then any solution satisfying (4.22) and (4.24) automatically satisfies (4.23). That is, we can ignore the constraint (4.23) by assuming that  $\gamma \geq 1/(\lambda - \lfloor \lambda \rfloor)$ . Consequently, we only need to characterize the extreme points of (4.22), (4.24). Let  $\check{a}_i = 1/(\lambda - i + 1)$  for  $i = 1, \dots, \lfloor \lambda \rfloor + 1$ . Let  $\check{x}_i$  denote the vector of the  $i$ th extreme

point. That is,  $\check{x}_i = (0, \dots, \check{a}_i, \dots, 0)$ . Calculating the value of the objective function over all extreme points, assuming that  $\gamma \geq 1/(\lambda - \lfloor \lambda \rfloor)$ , we get

$$\min \left\{ d\check{a}_1^\alpha, (d - 1)\check{a}_2^\alpha, \dots, (d - \lfloor \lambda \rfloor)\check{a}_{\lfloor \lambda \rfloor+1}^\alpha \right\} = \min_{i=1}^{\lfloor \lambda \rfloor+1} \left\{ (d - i + 1) \left( \frac{1}{\lambda - i + 1} \right)^\alpha \right\}. \tag{4.25}$$

Next, we consider the general case  $\gamma > 1/\lambda$ . We show that additional extreme points arise by considering the inclusion of (4.23) and this might potentially give rise to solutions in which large service requirements are not equal across all the servers. Note that if  $\lambda = \lfloor \lambda \rfloor$ , we must have that  $a_{\lfloor \lambda \rfloor+1} = 0$ . To see this, suppose that is not the case. Then, a feasible solution would be of the form  $v = (a_1, \dots, a_i, \dots, a_{\lfloor \lambda \rfloor+1})$ . By setting  $a_{\lfloor \lambda \rfloor+1} = 0$ , we construct another solution,  $v' = (a_1, \dots, a_i, \dots, a_{\lfloor \lambda \rfloor}, 0)$ . Observe that  $v'$  is a feasible solution and it reduces the value of the objective function (4.21) in comparison with  $v$ . Our subsequent analysis also includes the case  $\lambda = \lfloor \lambda \rfloor$ .

We identify the extreme points of (4.22), (4.23), (4.24). For that, we introduce the slack variable  $a_0 \geq 0$ .

$$\lambda a_1 + (\lambda - 1) a_2 + \dots + (\lambda - \lfloor \lambda \rfloor) a_{\lfloor \lambda \rfloor+1} = 1, \tag{4.26}$$

$$a_0 + a_1 + \dots + a_{\lfloor \lambda \rfloor+1} = \gamma, \tag{4.27}$$

$$a_0, a_1, \dots, a_{\lfloor \lambda \rfloor+1} \geq 0. \tag{4.28}$$

From elementary results in polyhedral combinatorics, we know that extreme points correspond to basic feasible solutions. By choosing  $a_{i+1} = 1/(\lambda - i)$  and  $a_0 = \gamma - a_{i+1}$ , we recover basic solutions which correspond to the extreme points identified by the equations above. Recall if  $\lambda = \lfloor \lambda \rfloor$  we must have that  $a_{\lfloor \lambda \rfloor+1} = 0$ . That is, we can safely assume that  $\lambda - i > 0$ . We observe that  $\gamma \geq 1/(\lambda - i)$  implies that  $a_{i+1} = 1/(\lambda - i)$  and  $a_j = 0$  for  $j \neq i + 1$  which is a basic feasible solution for (4.26). Additional basic solutions are obtained by solving

$$1 = (\lambda - k) a_{k+1} + (\lambda - l) a_{l+1},$$

$$\gamma = a_{k+1} + a_{l+1}.$$

Suppose that  $0 \leq l < k < \lambda$ . This system of equations always has a unique solution because the equations are linearly independent, and hence

$$\lambda\gamma - 1 = ka_{k+1} + la_{l+1}.$$

Therefore, the solution  $(\bar{a}_{k+1}, \bar{a}_{l+1})$  is given by

$$(k - l) \bar{a}_{k+1} = (\lambda - l) \gamma - 1,$$

$$(k - l) \bar{a}_{l+1} = 1 - \gamma (\lambda - k).$$

If we want  $(\bar{a}_{k+1}, \bar{a}_{l+1})$  to be both basic and feasible, we must have that  $1/(\lambda - l) \leq \gamma \leq 1/(\lambda - k)$ . Now, we calculate the value of the objective function for  $a_{k+1} = \bar{a}_{k+1}$ ,

$a_{l+1} = \bar{a}_{l+1}$ , and  $a_{i+1} = 0$  for  $i \notin \{k, l\}$ . That is,

$$\begin{aligned} & \sum_{i=1}^{\lfloor \lambda \rfloor} \left( \sum_{j=1}^i a_j \right)^\alpha + (d - \lfloor \lambda \rfloor) \left( \sum_{j=1}^{\lfloor \lambda \rfloor + 1} a_j \right)^\alpha \\ &= \bar{a}_{l+1}^\alpha (k - l) + (\lfloor \lambda \rfloor - k) (\bar{a}_{k+1} + \bar{a}_{l+1})^\alpha + (d - \lfloor \lambda \rfloor) (\bar{a}_{k+1} + \bar{a}_{l+1})^\alpha \\ &= \bar{a}_{l+1}^\alpha (k - l) + (d - k) (\bar{a}_{k+1} + \bar{a}_{l+1})^\alpha. \end{aligned} \tag{4.29}$$

Recall  $1/(\lambda - l) \leq \gamma \leq 1/(\lambda - k)$ . As we mentioned before, if  $\gamma = 1/(\lambda - k)$ , then we have that  $a_{k+1} = 1/(\lambda - k)$  and  $a_i = 0$  for  $i \neq k + 1$  which is a feasible extreme point. Furthermore, we see that under this particular solution the objective function has a smaller value than the solution involving  $\bar{a}_{k+1}$  and  $\bar{a}_{l+1}$ . To illustrate this, observe that

$$\bar{a}_{l+1}^\alpha (k - l) + (d - k) (\bar{a}_{k+1} + \bar{a}_{l+1})^\alpha > (d - k) a_{k+1}^\alpha.$$

Therefore,  $(\bar{a}_{k+1}$  and  $\bar{a}_{l+1})$  would be an optimal solution under the condition  $1/(\lambda - l) \leq \gamma < 1/(\lambda - k)$ . Due to (4.25) and (4.29), we conclude that the optimal value of the variational problem (4.16) is given by

$$\begin{aligned} & \min \left\{ \min_{0 < k \leq \lfloor \lambda \rfloor; \gamma < 1/(\lambda - k)} \left\{ (d - k) \gamma^\alpha + (1 - \gamma (\lambda - k))^\alpha \right\}, \min_{0 \leq l < \lfloor \lambda \rfloor; 1/(\lambda - l) \leq \gamma} \left( \frac{1}{k - l} \right)^\alpha (k - l) \right\}, \\ & \min_{\substack{\lfloor \lambda \rfloor \wedge \lfloor \lambda - 1/\gamma \rfloor \\ l=0}} \left\{ (d - l) \left( \frac{1}{\lambda - l} \right)^\alpha \right\}. \end{aligned}$$

By simplifying the expression above, we arrive at (4.15). □

**List of symbols**

$\mathbb{D}[0, T]$	The Skorokhod space—space of càdlàg functions—over the domain $[0, T]$
$\mathbb{D}^\uparrow[0, T]$	The subspace of $\mathbb{D}[0, T]$ consisting of non-decreasing functions that assume nonnegative values at the origin
$\mathbb{D}_p^\uparrow[0, T]$	The subspace of $\mathbb{D}[0, T]$ consisting of non-decreasing pure jump functions that assume nonnegative values at the origin
$\mathbb{D}^\mu[0, T]$	The subspace of $\mathbb{D}[0, T]$ consisting of non-decreasing piecewise linear functions with slope $\mu$ almost everywhere and nonnegative values at the origin.
$\check{\mathbb{C}}^\mu[0, T]$	The subspace of $\mathbb{D}[0, T]$ consisting of continuous functions which are piecewise linear with slope 0 or $1/\mu$
$\mathcal{T}_{M'_1}$	The $M'_1$ topology
$d_{M'_1}$	The $M'_1$ metric
$Q(t)$	The queue length at time $t$
$d$	The number of servers of the multiple-server queue
$\lambda$	The arrival rate associated with the multiple-server queue

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## A: Some large-deviations theory results

In this appendix, we include some important results and concepts widely used in the field of large deviations as well as in this paper. We have already mentioned the conditions under which a stochastic process  $Y$  satisfies a large-deviations principle. Let  $f$  be a map between two topological spaces. The following result, Theorem 4.2.1 in [17], formulates the conditions so that the transformation  $f(Y)$  satisfies an LDP also.

**Result A.1** (Contraction principle) *Let  $\mathcal{X}$  and  $\mathcal{Y}$  be Hausdorff topological spaces and  $f : \mathcal{X} \rightarrow \mathcal{Y}$  a continuous function. Consider a good rate function  $I : \mathcal{X} \rightarrow [0, \infty]$ .*

(a) *For each  $y \in \mathcal{Y}$ , define*

$$I'(y) \triangleq \inf\{I(x) : x \in \mathcal{X}, y = f(x)\}.$$

*Then,  $I'$  is a good rate function on  $\mathcal{Y}$ , where as usual the infimum over the empty set is taken as  $\infty$ .*

(b) *If  $I$  controls the LDP associated with a family of probability measures  $\mu_\epsilon$  on  $\mathcal{X}$ , then  $I'$  controls the LDP for the family of probability measures  $\{\mu_\epsilon \circ f^{-1}\}$  on  $\mathcal{Y}$ .*

**Remark 1** The theorem above holds under the weaker condition that  $f$  is continuous over the effective domain of the rate function  $I$ —i.e., on  $\{x \in \mathcal{X} : I(x) < \infty\}$ . This particular extension of the contraction principle is called the extended contraction principle (p.367 of [15]; Theorem 2.1 of [6]).

Now, we review the notion of exponential equivalence. We start with the definition.

**Definition A.1** Let  $(\mathcal{Y}, d)$  be a metric space. The probability measures  $\{\mu_\epsilon\}$  and  $\{\tilde{\mu}_\epsilon\}$  are called exponentially equivalent if there exist probability measures  $\{(\Omega, \mathcal{B}_\epsilon, P_\epsilon)\}$  and two families of  $\mathcal{Y}$ -valued random variables  $\{Z_\epsilon\}$  and  $\{\tilde{Z}_\epsilon\}$  with joint laws  $P_\epsilon$  and marginals  $\{\mu_\epsilon\}$  and  $\{\tilde{\mu}_\epsilon\}$ , respectively, such that the following condition is satisfied: For each  $\delta > 0$ , the set  $\{\omega : d(\tilde{Z}_\epsilon, Z_\epsilon) > \delta\}$  is  $\mathcal{B}_\epsilon$  measurable, and

$$\limsup_{\epsilon \rightarrow 0} \epsilon \log \mathbf{P}(d(\tilde{Z}_\epsilon, Z_\epsilon) > \delta) = -\infty.$$

Intuitively, two random variables are exponentially equivalent if their distance is asymptotically negligible.

### B: Continuity of some useful functionals in the $M'_1$ topology

In the next proposition, we prove that the map  $\Phi_\mu$  is sufficiently continuous for the application of extended contraction principle. Define  $\mathcal{D}_{\Phi_\mu} \triangleq \{\xi \in \mathbb{D}[0, \gamma/\mu] : \Phi_\mu(\xi)(\gamma) - \Phi_\mu(\xi)(\gamma-) > 0 \text{ and } \xi(0) \geq 0\}$ .

**Proposition B.1** *For each  $\mu \in \mathbb{R}$ ,  $\Phi_\mu : \mathbb{D}[0, \gamma/\mu] \rightarrow \mathbb{D}[0, \gamma]$  is continuous on  $\mathcal{D}_{\Phi_\mu}^c$  w.r.t. the  $M'_1$  topology.*

**Proof** Note that  $\Phi_\mu = \Phi_\mu \circ \Psi$  and  $\Psi$  is continuous, so we only need to check the continuity of  $\Phi_\mu$  over the range of  $\Psi$ , in particular non-decreasing functions. Let  $\xi$  be a non-decreasing function in  $\mathbb{D}[0, \gamma/\mu]$ . We consider two cases separately:  $\Phi_\mu(\xi)(\gamma) > \gamma/\mu$  and  $\Phi_\mu(\xi)(\gamma) \leq \gamma/\mu$ .

We start with the case  $\Phi_\mu(\xi)(\gamma) > \gamma/\mu$ . Pick  $\epsilon > 0$  such that  $\Phi_\mu(\xi)(\gamma) > \gamma/\mu + 2\epsilon$  and  $\xi(\gamma/\mu) + 2\epsilon < \gamma$ . For such an  $\epsilon$ , it is straightforward to check that  $d_{M'_1}(\zeta, \xi) < \epsilon$  implies  $\Phi_\mu(\zeta)(\gamma) > \gamma/\mu$  and  $\zeta$  never exceeds  $\gamma$  on  $[0, \gamma/\mu]$ . Therefore, the parametrizations of  $\Phi_\mu(\xi)$  and  $\Phi_\mu(\zeta)$  consist of the parametrizations—with the roles of space and time interchanged—of the original  $\xi$  and  $\zeta$  concatenated with the linear part coming from  $\psi_\mu$ . More specifically, suppose that  $(x, t) \in \Gamma(\xi)$  and  $(y, r) \in \Gamma(\zeta)$  are parametrizations of  $\xi$  and  $\zeta$ . Since  $\xi$  is non-decreasing, if we define on  $s \in [0, T]$

$$\begin{aligned} x'(s) &\triangleq \begin{cases} t(2s) & \text{if } s \leq T/2 \\ \frac{1}{\mu}(t'(s) - \Psi(\xi)(\gamma/\mu) + \gamma) & \text{if } s > T/2 \end{cases} \\ t'(s) &\triangleq \begin{cases} x(2s) & \text{if } s \leq T/2 \\ (\gamma - \Psi(\xi)(\gamma/\mu))(2s/T - 1) + \Psi(\xi)(\gamma/\mu) & \text{if } s > T/2 \end{cases} \\ y'(s) &\triangleq \begin{cases} r(2s) & \text{if } s \leq T/2 \\ \frac{1}{\mu}(r'(s) - \Psi(\zeta)(\gamma/\mu) + \gamma) & \text{if } s > T/2 \end{cases} \\ r'(s) &\triangleq \begin{cases} y(2s) & \text{if } s \leq 1/2 \\ (\gamma - \Psi(\zeta)(\gamma/\mu))(2s/T - 1) + \Psi(\zeta)(\gamma/\mu) & \text{if } s > 1/2 \end{cases} \end{aligned}$$

then  $(x', t') \in \Gamma(\Phi_\mu(\xi))$ ,  $(y', r') \in \Gamma(\Phi_\mu(\zeta))$ . Noting that

$$\begin{aligned} &\|x' - y'\|_\infty + \|t' - r'\|_\infty \\ &= \sup_{s \in [0, 1/2]} |t(2s) - r(2s)| \vee \sup_{s \in (1/2, 1]} |x'(s) - y'(s)| \\ &\quad + \sup_{s \in [0, 1/2]} |x(2s) - y(2s)| \vee \sup_{s \in (1/2, 1]} |t'(s) - r'(s)| \\ &= \|t - r\|_\infty \vee \mu^{-1} |\Psi(\zeta)(\gamma) - \Psi(\xi)(\gamma)| \\ &\quad + \|x - y\|_\infty \vee |\Psi(\zeta)(\gamma) - \Psi(\xi)(\gamma)| \\ &\leq \mu^{-1} \|t - r\|_\infty \vee \|x - y\|_\infty + \|x - y\|_\infty \\ &\leq (1 + \mu^{-1})(\|x - y\|_\infty + \|t - r\|_\infty), \end{aligned}$$

and taking the infimum over all possible parametrizations, we conclude that  $d_{M'_1}(\Phi_\mu(\xi), \Phi_\mu(\zeta)) \leq (1 + \mu^{-1})d_{M'_1}(\xi, \zeta) \leq (1 + \mu^{-1})\epsilon$ , and hence  $\Phi_\mu$  is continuous at  $\xi$ .

Turning to the case  $\Phi_\mu(\xi)(\gamma) \leq \gamma/\mu$ , let  $\epsilon > 0$  be given. Due to the assumption that  $\Phi_\mu(\xi)$  is continuous at  $\gamma$ , there has to be a  $\delta > 0$  such that  $\varphi_\mu(\xi)(\gamma) + \epsilon < \varphi_\mu(\xi)(\gamma - \delta) \leq \varphi_\mu(\xi)(\gamma + \delta) \leq \varphi_\mu(\xi)(\gamma) + \epsilon$ . We prove that if  $d_{M'_1}(\xi, \zeta) < \delta \wedge \epsilon$ , then  $d_{M'_1}(\Phi_\mu(\xi), \Phi_\mu(\zeta)) \leq 8\epsilon$ . Since the case where  $\Phi_\mu(\zeta)(\gamma) \geq \gamma/\mu$  is similar to the above argument, we focus on the case  $\Phi_\mu(\zeta)(\gamma) < \gamma/\mu$ ; that is,  $\zeta$  also crosses level  $\gamma$  before  $\gamma/\mu$ . Let  $(x, t) \in \Gamma(\xi)$  and  $(y, r) \in \Gamma(\zeta)$  be such that  $\|x - y\|_\infty + \|t - r\|_\infty < \delta$ . Let  $s_x \triangleq \inf\{s \geq 0 : x(s) > \gamma\}$  and  $s_y \triangleq \inf\{s \geq 0 : y(s) > \gamma\}$ . Then, it is straightforward to check  $t(s_x) = \varphi_\mu(\xi)(\gamma)$  and  $r(s_y) = \varphi_\mu(\zeta)(\gamma)$ . Of course,  $x(s_x) = \gamma$  and  $y(s_y) = \gamma$ . If we set  $x'(s) \triangleq t(s \wedge s_x)$ ,  $t'(s) \triangleq x(s \wedge s_x)$ , and  $y'(s) \triangleq r(s \wedge s_y)$ ,  $r'(s) \triangleq y(s \wedge s_y)$ , then

$$\begin{aligned} \|x' - y'\|_\infty &\leq \|t - r\|_\infty + \sup_{s \in [s_x \wedge s_y, s_x \vee s_y]} \{|t(s_x) - r(s)| \vee |t(s) - r(s_y)|\} \\ &\leq \|t - r\|_\infty \\ &\quad + \sup_{s \in [s_x \wedge s_y, s_x \vee s_y]} \{(|t(s_x) - t(s)| + |t(s) - r(s)|) \vee (|t(s) - t(s_y)| \\ &\quad + |t(s_y) - r(s_y)|)\} \\ &\leq \|t - r\|_\infty \\ &\quad + (|t(s_x) - t(s_y)| + \|t - r\|_\infty) \vee (|t(s_y) - t(s_x)| + \|t - r\|_\infty) \\ &\leq 2\|t - r\|_\infty + 2|t(s_x) - t(s_y)|. \end{aligned}$$

Now we argue that  $t(s_x) - \epsilon \leq t(s_y) \leq t(s_x) + \epsilon$ . To see this, note first that  $x(s_y) < x(s_x) + \delta = \gamma + \delta$ , and hence,

$$t(s_y) \leq \varphi_\mu(\xi)(x(s_y)) \leq \varphi_\mu(\xi)(\gamma + \delta) \leq \varphi_\mu(\xi)(\gamma) + \epsilon = t(s_x) + \epsilon.$$

On the other hand,

$$t(s_x) - \epsilon = \varphi_\mu(\xi)(\gamma) - \epsilon \leq \varphi_\mu(\gamma - \delta) \leq t(s_y),$$

where the last inequality is from  $\xi(t(s_y)) \geq x(s_y) > x(s_x) - \delta = \gamma - \delta$  and the definition of  $\varphi_\mu$ . Therefore,  $\|x' - y'\|_\infty \leq 2\delta + 2\epsilon < 4\epsilon$ . Now we are left with showing that  $\|t' - r'\|_\infty$  can be bounded in terms of  $\epsilon$ .

$$\begin{aligned} \|t' - r'\|_\infty &\leq \|x - y\|_\infty + \sup_{s \in [s_x \wedge s_y, s_x \vee s_y]} \{|x(s_x) - y(s)| \vee |x(s) - y(s_y)|\} \\ &\leq \|x - y\|_\infty + \sup_{s \in [s_x \wedge s_y, s_x \vee s_y]} \{(|x(s_x) - x(s)| \\ &\quad + |x(s) - y(s)|) \vee (|x(s) - x(s_y)| + |x(s_y) - y(s_y)|)\} \\ &\leq \|x - y\|_\infty \\ &\quad + (|t(s_x) - t(s_y)| + \|x - y\|_\infty) \vee (|x(s_x) - x(s_y)| + \|x - y\|_\infty) \end{aligned}$$

$$\begin{aligned} &\leq 2\|x - y\|_\infty + 2|x(s_x) - x(s_y)| \\ &= 2\|x - y\|_\infty + 2|y(s_y) - x(s_y)| \leq 4\|x - y\|_\infty < 4\epsilon. \end{aligned}$$

Therefore,  $d_{M'_1}(\Phi_\mu(\xi), \Phi_\mu(\zeta)) \leq \|x' - y'\|_\infty + \|t' - r'\|_\infty < 8\epsilon$ . □

**Lemma B.1** *The map  $\Upsilon_\mu : \mathbb{D}[0, \gamma/\mu] \rightarrow \mathbb{D}[0, \gamma/\mu]$ , where  $\Upsilon_\mu(\xi) \triangleq \xi + \zeta_\mu$ , is continuous w.r.t. the  $M'_1$  topology on  $\mathbb{D}[0, \gamma/\mu]$ .*

**Proof** Suppose that  $\xi_n \rightarrow \xi$  in  $\mathbb{D}[0, \gamma/\mu]$  w.r.t. the  $M'_1$  topology. As a result, there exist parametrizations  $(u_n(s), t_n(s))$  of  $\xi_n$  and  $(u(s), t(s))$  of  $\xi$  such that

$$\sup_{s \leq \gamma/\mu} \{|u_n(s) - u(s)| + |t_n(s) - t(s)|\} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

This implies that  $\max\{\sup_{s \leq \gamma/\mu} |u_n(s) - u(s)|, \sup_{s \leq \gamma/\mu} |t_n(s) - t(s)|\} \rightarrow 0$  as  $n \rightarrow \infty$ . Observe that if  $(u(s), t(s))$  is a parametrization for  $\xi$ , then  $(u(s) + \mu \cdot t(s), t(s))$  is a parametrization for  $\Upsilon_\mu(\xi)$ . Consequently,

$$\begin{aligned} &\sup_{s \leq \gamma/\mu} \{|u_n(s) + \mu \cdot t_n(s) - u(s) - \mu \cdot t(s)| + |t_n(s) - t(s)|\} \\ &\leq \sup_{s \leq \gamma/\mu} \{|u_n(s) - u(s)|\} + \sup_{s \leq \gamma/\mu} \{(\mu + 1)|t_n(s) - t(s)|\} \rightarrow 0. \end{aligned}$$

Thus,  $\Upsilon_\mu(\xi_n) \rightarrow \Upsilon_\mu(\xi)$  in the  $M'_1$  topology, proving that the map is continuous. □

The next lemma provides the continuity of two functionals used in our large-deviation analysis.

**Lemma B.2** *For any  $T > 0$ ,*

- (i) *The functional  $E : \mathbb{D}[0, T] \rightarrow \mathbb{R}$ , where  $E(\xi) = \xi(T)$ , is continuous w.r.t. the  $M'_1$  topology on  $\mathbb{D}[0, T]$ .*
- (ii) *The functional  $S : \mathbb{D}[0, T] \rightarrow \mathbb{R}$ , where  $S(\xi) = \sup_{t \in [0, T]} \xi(t)$ , is continuous w.r.t. the  $M'_1$  topology on  $\xi \in \mathbb{D}[0, T]$  such that  $\xi(0) \geq 0$ .*

**Proof** Consider a sequence  $\xi_n$  such that  $d_{M'_1}(\xi_n, \xi) \rightarrow 0$ . From (2.3), there exists a parametrization  $(u(s), t(s))$  of the completed graph of  $\xi$  and a parametrization  $(u_n(s), t_n(s))$  of the completed graph of  $\xi_n$  such that

$$\sup_{s \in [0, T]} \{|u_n(s) - u(s)| + |t_n(s) - t(s)|\} \rightarrow 0, \quad \text{as } n \rightarrow \infty. \tag{B. 1}$$

For i), note that  $|u_n(T) - u(T)| \leq \sup_{s \in [0, T]} |u_n(s) - u(s)| \rightarrow 0$ , while  $\xi_n(T) = u_n(T)$  and  $\xi(T) = u(T)$ . Therefore,  $|E(\xi) - E(\xi_n)| = |\xi_n(T) - \xi(T)| \rightarrow 0$  as  $n \rightarrow \infty$ . Therefore,  $E$  is a continuous functional. For ii), suppose that  $\xi(0) \geq 0$ . For any  $\epsilon > 0$ , there exists  $N$  such that  $\xi_n(0) \geq -\epsilon$  for  $n > N$ . Now, from the definition of parametrization and the nonnegativity of  $\xi(0)$ , we see that  $\sup_{s \in [0, T]} u(s) =$

$\sup_{s \in [0, T]} \xi(s)$ . Similarly, we can show that  $|\sup_{s \in [0, T]} u_n(s) - \sup_{s \in [0, T]} \xi_n(s)| < \epsilon$ . Therefore,

$$\begin{aligned} & \limsup_{n \rightarrow \infty} \left| \sup_{s \in [0, T]} \xi_n(s) - \sup_{s \in [0, T]} \xi(s) \right| \\ & \leq \limsup_{n \rightarrow \infty} \left| \sup_{s \in [0, T]} u_n(s) - \sup_{s \in [0, T]} u(s) \right| + \epsilon \leq \limsup_{n \rightarrow \infty} \sup_{s \in [0, T]} |u_n(s) - u(s)| + \epsilon = \epsilon. \end{aligned}$$

Since  $\epsilon$  was arbitrary, this proves the continuity of  $S$  at  $\xi$ .  $\square$

## References

- Pollaczek, Félix: Théorie analytique des problèmes stochastiques relatifs à un groupe de lignes téléphoniques avec dispositif d'attente (1961)
- Kiefer, J., Wolfowitz, J.: On the theory of queues with many servers. *Trans. Am. Math. Soc.* **78**, 1–18 (1955)
- Iglehart, D.L., Whitt, W.: Multiple channel queues in heavy traffic. I. *Adv. Appl. Probab.* **2**, 150–177 (1970)
- Pang, G., Talreja, R., Whitt, W.: Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probab. Surv.* **4**, 193–267 (2007)
- Sadowsky, J.S.: The probability of large queue lengths and waiting times in a heterogeneous multiserver queue. II. Positive recurrence and logarithmic limits. *Adv. Appl. Probab.* **27**(2), 567–583 (1995)
- Puhalskii, A.: Large deviation analysis of the single server queue. *Queueing Syst.* **21**(1), 5–66 (1995)
- Foss, S., Korshunov, D., Zachary, S.: *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer Series in Operations Research and Financial Engineering, 2nd edn. Springer, New York (2013)
- Foss, S., Korshunov, D.: Sampling at a random time with a heavy-tailed distribution. *Markov Process. Relat. Fields* **6**(4), 543–568 (2000)
- Whitt, W.: The impact of a heavy-tailed service-time distribution upon the  $M/GI/s$  waiting-time distribution. *Queue. Syst. Theory Appl.* **36**(1–3), 71–87 (2000)
- Scheller-Wolf, A., Vesilo, R.: Sink or swim together: necessary and sufficient conditions for finite moments of workload components in FIFO multiserver queues. *Queue. Syst.* **67**(1), 47–61 (2011)
- Foss, S., Korshunov, D.: Heavy tails in multi-server queue. *Queue. Syst. Theory Appl.* **52**(1), 31–48 (2006)
- Foss, S., Korshunov, D.: On large delays in multi-server queues with heavy tails. *Math. Oper. Res.* **37**(2), 201–218 (2012)
- Gamarnik, D., Goldberg, D.A.: Steady-state  $GI/G/n$  queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* **23**(6), 2382–2419 (2013)
- Ge, D., Jiang, X., Ye, Y.: A note on the complexity of  $L_p$  minimization. *Math. Program.* **129**(2, Ser. B), 285–299 (2011)
- Puhalskii, A.A., Whitt, W.: Functional large deviation principles for first-passage-time processes. *Ann. Appl. Probab.* 362–381 (1997)
- Bazhba, M., Blanchet, J., Rhee, C.-H., Zwart, B.: Sample-path large deviations for Lévy processes and random walks with Weibull increments. [arXiv:1710.04013](https://arxiv.org/abs/1710.04013) (2017)
- Dembo, A., Zeitouni, O.: *Large Deviations Techniques and Applications*. Springer, Berlin (2010)
- Ganesh, A.J., O'Connell, N., Wischik, D.J.: *Big Queues*. Springer, Berlin (2004)
- Feng, J., Kurtz, T.G.: Large deviations for stochastic processes. Number 131. American Mathematical Soc. (2006)
- Rockafellar, T.: *Convex Analysis*. Princeton Mathematical Series. Princeton University Press, Princeton (1970)

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.