J. Appl. Prob. **54,** 943–962 (2017)
*doi:*10.1017/jpr.2017.43

RARE EVENTS OF TRANSITORY QUEUES

HARSHA HONNAPPA,* *Purdue University*

Abstract

We study the rare-event behavior of the workload process in a transitory queue, where the arrival epochs (or 'points') of a finite number of jobs are assumed to be the ordered statistics of independent and identically distributed (i.i.d.) random variables. The service times (or 'marks') of the jobs are assumed to be i.i.d. random variables with a general distribution, that are jointly independent of the arrival epochs. Under the assumption that the service times are strictly positive, we derive the large deviations principle (LDP) satisfied by the workload process. The analysis leverages the connection between ordered statistics and self-normalized sums of exponential random variables to establish the LDP. In this paper we present the first analysis of rare events in transitory queueing models, supplementing prior work that has focused on fluid and diffusion approximations.

Keywords: Workload; empirical process; exchangeable increments; large deviation; population acceleration; transitory queue

2010 Mathematics Subject Classification: Primary 60K25; 60F10
Secondary 90B22; 68M20

1. Introduction

We explicate the rare-event behavior of a 'transitory' queueing model by proving a large deviations principle (LDP) satisfied by the workload process of the queue. A formal definition of a transitory queue follows from [15].

Definition 1.1. (*Transitory queue.*) Let $A(t)$ represent the cumulative traffic entering a queueing system. The queue is transitory if $A(t)$ satisfies

$$\lim_{t\to\infty} A(t) < \infty \quad \text{a.s.},$$

where a.s. is short for almost surely.

We consider a specific transitory queueing model where the arrival epochs (or 'points') of a finite but large number of jobs, say n, are 'randomly scattered' over $[0, \infty)$; that is, the arrival epochs $(T_1, \dots, T_n)$ are independent and identically distributed (i.i.d.), and drawn from some distribution with support in $[0, \infty)$. We assume that the service times (or 'marks') $\{\nu_1, \dots, \nu_n\}$ are i.i.d., jointly independent of the arrival epochs, and with log-moment generating function that satisfies $\varphi(\theta) < \infty$ for $\theta \in \mathbb{R}$. We call this queueing model the RS/GI/1 queue (RS standing for *randomly scattered*; this was previously called the $\Delta_{(i)}/\mathrm{GI}/1$ queueing model in [16]).

While the i.i.d. assumption on the arrival epochs implies that this is a homogeneous model, Honnappa *et al.* [16] showed that the workload process displays time-dependencies in the large population fluid and diffusion scales, mirroring those observed for 'dynamic rate' queueing

Received 14 July 2016; revision received 23 May 2017.
* Postal address: School of Industrial Engineering, Purdue University, 315 N. Grant St., West Lafayette, IN 47906, USA. Email address: honnappa@purdue.edu

models where time-dependent arrival rates are explicitly assumed. This indicates that the rare-event behavior of the workload or queue length process should be atypical compared to that of time-homogeneous queueing models (such as the G/G/1 queue; see [11]). Furthermore, while the standard dynamic rate traffic model is a nonhomogeneous Poisson process that necessarily has independent increments, it is less than obvious that it is a reasonable assumption for many service systems. The authors of [12] and [13], for instance, highlighted data analysis and simulation results in the call center context that indicate that independent increments might not be appropriate. A tractable alternative is to assume that the increments are exchangeable [1]. Lemma 10.9 of [1] implies that *any* traffic process over the horizon [0, 1] with exchangeable increments is necessarily equal in distribution to the empirical sum process

$$\sum_{i=1}^{N} \mathbf{1}_{\{T_i \le t\}} \quad \text{for all } t \in [0, 1], \tag{1.1}$$

where the $\{T_i,\ 1 \le i \le N\}$ are independent and uniformly distributed in [0,1]. In [16], the authors defined (1.1) as the traffic count process for the RS/GI/1 queue. This can be considered the canonical model of a transitory traffic process with exchangeable increments. Thus, the results in this paper can also be viewed as explicating the rare events behavior of queueing models with exchangeable increments. To the best of the author's knowledge this has not been reported in the literature before.

Transitory queueing models, and the RS/GI/1 queue in particular, have received some recent interest in the applied probability literature, besides [16]. In a forthcoming work [14], the authors study large deviations and diffusion approximations to the workload process in a 'near-balanced' condition on the offered load to the system. This paper is complementary by not assuming the near-balanced condition. In a recent work, Bet *et al.* [2] established diffusion approximations to the queue length process of the $\Delta_{(i)}/\mathrm{GI}/1$ queue under a uniform acceleration scaling regime, where it is assumed that the 'initial load' near time 0 satisfies $\rho_n = 1 + \beta n^{-1/3}$. This, of course, contrasts with the population acceleration regime considered in this paper, where the offered load is accelerated by the population size at all time instances in the horizon [0, 1]. The same authors also considered the effect of heavy-tailed service in transitory queues [3], and established weak limits to the scaled workload and queue length processes to reflected alpha-stable processes.

In the ensuing discussion, we will largely focus on the case that the arrival epochs are uniformly distributed with support [0, 1]. In our first result, Theorem 3.1, we establish a large deviations result for the ordered statistics process $(T^n(t) := T_{(\lfloor nt \rfloor)}$, for all t $\in [0, 1])$, $n \ge 1$, where $T_{(j)}$ represents the jth order statistic. This result parallels that of [9], where the authors derived a sample-path large deviations result for the ordered statistics of i.i.d. uniform random variables. Our results deviate from this result in a couple of ways. First, we do not require a full sample-path LDP, since we are interested in understanding the large deviations of the workload at a given point in time. Second, our proof technique is different and explicitly uses the connection between ordered statistics and self-normalized sums of exponential random variables. It is also important to note the result of [4], where the author used Sanov's theorem to prove the large deviation principle for L-statistics, which could be leveraged to establish the LDP for the traffic process in (1.1) and, hence, the number-in-system process. The objective of our study, on the other hand, is the workload process. In Corollary 3.1 we use the contraction principle to extend this large deviations result to arrival epochs that have distribution F with positive support, under the assumption that the distribution is absolutely continuous and strictly increasing. However, much of the 'heavy lifting' for the workload LDP can be demonstrated with uniform order statistics arrival epochs, so in the remainder of the paper we do not emphasize the extension to more generally distributed arrival epochs.

In Proposition 3.2 we make use of the proof of Theorem 3.1 and the well-known Cramer's theorem [7, Theorem 2.2.3] in order to derive the large deviation rate function for the offered load process $X^n(t) := S^n(t) - T^n(t)$ for all $t \in [0, 1]$ and $n \geq 1$, where $S^n(t) := \sum_{i=1}^{\lfloor nt \rfloor} \nu_i$ is the partial sum of the service times. Interestingly enough, the LDP (and the corresponding good rate function) shows that the most likely path to a large deviation event depends crucially on both the sample path of the offered load process up to t as well as the path *after* t. This is a direct reflection of the fact that the traffic process is exchangeable and that there is long-range dependence between the interarrival times (which are 'spacings' between ordered statistics, and thus finitely exchangeable).

We prove the LDP for the workload process

$$W^n(t) := \Gamma(X^n)(t) = \sup_{0 \leq s \leq t} (X^n(t) - X^n(s))$$

for fixed $t \in [0, 1]$, by exploiting the continuity of the reflection regulator map $\Gamma(\cdot)$. However, to do so, we first establish two auxiliary results: in Proposition 4.1 we prove the exponential equivalence of the workload process and a linearly interpolated version $\tilde{X}^n$. Then, in Proposition 4.2, we prove the LDP satisfied by the 'partial' sample paths $(\tilde{X}^n(s),\ 0 \leq s \leq t)$ of the offered load process for fixed $t \in [0, 1]$. Then, in Theorem 4.1, we establish the LDP for the workload process by applying the contraction mapping theorem with the reflection regulator map and exploiting the two propositions mentioned above. We conclude the paper with a summary and comments on future directions for this research.

1.1. Notation

We assume that all random elements are defined with respect to an underlying probability sample space $(\Omega, \mathcal{F}, \mathbb{P})$. We denote convergence in probability by '$\xrightarrow{\mathbb{P}}$'. We denote the space $\mathcal{X}$ and topology of convergence $\mathcal{T}$ by the pair $(\mathcal{X}, \mathcal{T})$, where appropriate. In particular, we note $(\mathcal{C}[0, t], \mathcal{U})$, the space of continuous functions with domain $[0, t]$, equipped with the uniform topology. We also designate $\bar{\mathcal{C}}[0, t]$ as the space of all continuous functions that are nondecreasing on the domain $[0, t]$. Also, $\|\cdot\| = \sup_{0 \leq s \leq 1}(\cdot)$ represents the supremum norm on $\mathcal{C}[0, 1]$. Finally, we will use the following standard definitions in the ensuing results.

Definition 1.2. (*Rate function.*) Let $\mathcal{X}$ be the Hausdorff topological space. Then

- a rate function is a lower semicontinuous mapping $I\colon \mathcal{X} \to [0, \infty]$; i.e. the level set $\{x \in \mathcal{X}\colon I(x) \leq \alpha\}$ for any $\alpha \in [0, \infty)$ is a closed subset of $\mathcal{X}$, and
- a rate function is 'good' if the level sets are also compact.

Definition 1.3. (*The LDP.*) The sequence of random elements $\{X_n,\ n \geq 1\}$ taking values in the Hausdorff topological space $\mathcal{X}$ satisfies an LDP with rate function $I\colon \mathcal{X} \to \mathbb{R}$ if,

(i) for each open set $G \subset \mathcal{X}$,

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(X_n \in G) \geq -\inf_{x \in G} I(x);$$

(ii) for each closed set $F \subset \mathcal{X}$,

$$\limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(X_n \in F) \leq -\inf_{x \in F} I(x).$$

Definition 1.4. (*Weak LDP.*) The sequence of random elements $\{X_n,\ n \geq 1\}$ taking values in the Hausdorff topological space $\mathcal{X}$ satisfies a weak large deviation principle with rate function I if,

(i) for each open set $G \subset \mathcal{X}$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X_n \in G) \geq -\inf_{x\in G} I(x);$$

(ii) for each compact set $K \subset \mathbb{R}$,

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X_n \in K) \leq -\inf_{x\in K} I(x).$$

Definition 1.5. (*Large deviation tight.*) A sequence of random elements $\{X_n,\ n \geq 1\}$ taking values in the Hausdorff topological space $\mathcal{X}$ is large deviation tight if, for each $M < \infty$, there exists a compact set K_M such that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(X_n \in K_M^c) \leq -M.$$

2. Model

Let $\{(T_{(i)}, \nu_i),\ i = 1, 2, \dots, n\}$ for $n \in \mathbb{N}$ represent a marked finite point process, where $\{T_{(i)},\ i = 1, 2, \dots, n\}$ are the epochs of the point process and $\{\nu_i,\ i = 1, 2, \dots, n\}$ are the marks. We assume that the two sequences are independent of each other. The epochs $\{T_{(i)},\ i = 1, 2, \dots, n\}$ are the order statistics of n i.i.d. random variables with support $[0, \infty)$ and absolutely continuous distribution F. Also, $\{\nu_i,\ i = 1, 2, \dots, n\}$ are i.i.d. random variables with support $[0, \infty)$, cumulant generating function $\varphi(\theta) < \infty$ for some $\theta \in \mathbb{R}$, and mean $\mathbb{E}[\nu_1] = 1/\mu$. We will also assume that $\mathbb{P}(\nu_1 > 0) = 1$ for technical reasons. Let $\mathbb{D} := \{\theta \in \mathbb{R} \colon \varphi(\theta) < \infty\}$ and we assume that $0 \in \mathbb{D}$. In relation to the queue, $(T_{(j)}, \nu_j)$ represents the arrival epoch and service requirement of job j, and n is the total arrival population. It is useful to think of the n marked points, or $(T_{(i)}, \nu_i)$ pair, being 'scattered' over the horizon following the distribution F.

Let $\{\nu_i^n := \nu_i/n,\ i = 1, 2, \dots, n\}$ be a 'population accelerated' sequence of marks. Assume that $(T_{(0)}, \nu_0^n) = (0, 0)$. The (accelerated) workload ahead of the jth job is $W_j^n = (W_{j-1}^n + \nu_{j-1}^n - (T_{(j)} - T_{(j-1)}))_+$, where $(\cdot)_+ := \max\{0, \cdot\}$. By unraveling the recursion, and under the assumption that the queue starts empty, it can be shown that

$$W_j^n \stackrel{\mathrm{D}}{=} (S_{j-1}^n - T_{(j)}) + \max_{0\leq i\leq j-1} (-(S_i^n - T_{(i+1)})),$$

where $S_{j-1}^n := \sum_{i=0}^{j-1} \nu_i^n$ and '$\stackrel{\mathrm{D}}{=}$' denotes equality in distribution. We define the workload process as $(W^n(t),\ t \in [0, 1]) := (W_{\lfloor nt\rfloor}^n,\ t \in [0, 1])$. Using the unraveled recursion, it can be argued that

$$W^n(t) \stackrel{\mathrm{D}}{=} X^n(t) + \max_{0\leq s\leq t} (-X^n(s)),$$

where $X^n(t) := n^{-1}\sum_{i=0}^{\lfloor nt\rfloor} \nu_i - T_{(\lfloor nt\rfloor)} = S^n(t) - T^n(t)$ (where $T_{(0)} = 0$) for $t \in [0, 1]$ is the offered load process, under the assumption that $S_0^n = 0$ (i.e. the queue starts empty). Thus, it suffices to study $\Gamma(X^n)(t) := X^n(t) + \max_{0\leq s\leq t}(-X^n(s))$, where $\Gamma\colon \mathcal{D}[0, 1] \to \mathcal{D}[0, 1]$ is the so-called Skorokhod regulator map. For future reference, we call $(T^n(t),\ t \in [0, 1]) := (T_{(\lfloor nt\rfloor)},\ t \in [0, 1])$ the ordered statistics process.

We propose to study the workload process in the large population limit and, in particular, understand the rare-event behavior in this limit. As a precursor to this analysis, it is useful to consider what a 'normal deviation' event for this process would be. In particular, in the next proposition we prove a functional strong law of large numbers result for the workload process, that exposes the first-order behavior of the workload sample path, in the large population limit.

Proposition 2.1. *The workload process W^n satisfies*

$$W^n \to \bar{W} = \frac{1}{\mu}\Gamma(F - M) \quad in (\mathcal{C}[0,1], \mathcal{U})\, a.s.$$

as $n \to \infty$, where $M(t) = \mu t$.

Proof. First assume that $\{0 < T_{(1)} \le \cdots \le T_{(n)} < 1\}$ are the ordered statistics of n i.i.d. uniform random variables. Then, by [5, Lemma 5.8], it follows that the ordered statistics process satisfies $(T^n, S^n) \to (e, \mu^{-1}e)in(\mathcal{C}[0,1], \mathcal{U})$ a.s. as $n \to \infty$, where $e\colon \mathbb{R} \to \mathbb{R}$ is the identity map, and the joint convergence follows due to the fact that the arrival epochs and service times are independent sequences. Let $\bar{X} := (\mu^{-1}e - e)$, which is continuous by definition. Since subtraction is continuous under the uniform metric topology, it follows that $X^n \to \bar{X} := \mu^{-1}(e - M)$ in $(\mathcal{C}[0,1], \mathcal{U})$ a.s. as $n \to \infty$. Finally, since $\Gamma(\cdot)$ is continuous under the uniform metric, and the limit function $\Gamma(\bar{X})$ is continuous, it follows that $W^n \to \bar{W}$ in $(\mathcal{C}[0,1], \mathcal{U})$ a.s. as $n \to \infty$. The limit result for generally distributed arrival epochs follows by an application of the quantile transform to the arrival epochs. □

From an operational perspective, it is useful to understand the likelihood that the workload exceeds an abnormally large threshold. More precisely, we are interested in the likelihood that, for a given $t \in [0,1]$, $W^n(t) > w$, where $w \gg \bar{W}(t)$. While this is quite difficult to prove for a fixed n, we prove an LDP for the workload process as the population size n scales to ∞, which will automatically provide an approximation to the likelihood of this event. In the ensuing exposition, we will largely focus on the analysis of a queue where the arrival epochs are modeled as the ordered statistics of i.i.d. uniform random variables on $[0,1]$. However, the results can be straightforwardly extended to a more general case where the arrival epochs have distribution F (with positive support) that is absolutely continuous with respect to the uniform distribution.

Our agenda for proving the workload LDP will proceed in several steps. First, we prove an LDP for the ordered statistics process of i.i.d. uniform random variables. The proof of this result will then be used to establish an LDP for the offered load process $X^n(t)$. Next, we use a projective limit to establish the LDP for the sample path of the offered load process $(X^n(s),\ 0 \le s \le t)$ for each fixed $t \in [0,1]$. Finally, we prove an LDP for the workload process by applying the contraction principle to the LDP for the sample path $(X^n(s),\ 0 \le s \le t)$, transformed through the Skorokhod regulator map $\Gamma(\cdot)$.

3. An LDP for the offered load

3.1. LDP for the ordered statistics process

As a precursor to the LDP for the offered load process, we prove one for the ordered statistics process $(T^n(t),\ t \in [0,1]) := (T_{(\lfloor nt \rfloor)},\ t \in [0,1])$ by leveraging the following well-known relation between the order statistics of uniform random variables and partial sums of unit-mean exponential random variables.

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

Proposition 3.1. *Let $0 < T_{(1)} < T_{(2)} < \cdots < T_{(n)} < 1$ be the ordered statistics of independent and uniformly distributed random variables, and $\{\xi_j,\ 1 \le j \le n+1\}$ independent mean 1 exponential random variables. Then*

$$\{T_{(j)},\ 1 \le j \le n\} \stackrel{\mathrm{D}}{=} \left\{\frac{Z_j}{Z_{n+1}},\ 1 \le j \le n\right\}, \tag{3.1}$$

where $Z_j := \sum_{i=1}^{j} \xi_i$.

Proofs of this result can be found in [10, Lemma 8.9.1]. Now consider the convex, continuous function $I_t \colon [0,1] \to \mathbb{R}$ indexed by $t \in [0,1]$,

$$I_t(x) = t \log\left(\frac{t}{x}\right) + (1-t)\log\left(\frac{1-t}{1-x}\right). \tag{3.2}$$

In Figure 1 we depict (3.2) for different index values $t \in [0,1]$. In Theorem 3.1 below we show that I_t is the good rate function of the LDP satisfied by the ordered statistics process. It is interesting to note that this function is also the rate function satisfied by a sequence of i.i.d. Bernoulli random variables with parameter t; see the references of [8].

Theorem 3.1. (LDP for the ordered statistics process.) *Fix $t \in [0,1]$. The ordered statistics process $T_n(t)$ satisfies the LDP with good rate function (3.2).*

Proof. The proof comprises two parts.

(i) Let $F \subset [0,1]$ be closed. There are two cases to consider. First, if $t \in F$ then $I_t(F) := \inf_{x\in F} I_t(x) = 0$, by definition. Thus, we assume that $t \notin F$. Let $x_+ := \inf\{x \in F : x > t\}$ and $x_- := \sup\{x \in F : x < t\}$. If $\sup F < t$ then we define $x_+ = 1$, and if $\inf F > t$ we set $x_- = 0$. Since $t \notin F$, there exists a connected open set $F^c \supseteq (x_-, x_+) \ni t$.

Now let $0 \le a < t$. Proposition 3.1 implies that

$$\begin{aligned}\mathbb{P}(T_{(\lfloor nt \rfloor)} < a) &= \mathbb{P}\left(\frac{Z_{\lfloor nt \rfloor}}{Z_{n+1}} < a\right)\\ &= \mathbb{P}\left(Z_{n+1-\lfloor nt \rfloor} > \frac{1-a}{a} Z_{\lfloor nt \rfloor}\right)\\ &= \int_0^\infty \mathbb{P}\left(Z_{n+1-\lfloor nt \rfloor} > \frac{(1-a)x}{a}\right)\mathbb{P}\left(Z_{\lfloor nt \rfloor} \in \mathrm{d}x\right).\end{aligned} \tag{3.3}$$

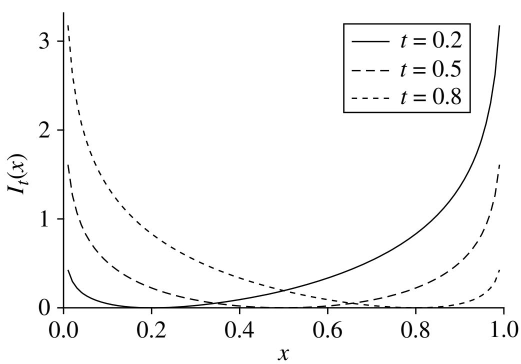

FIGURE 1: Rate function for the ordered statistics process.

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

Now, Chernoff's inequality implies that

$$\mathbb{P}\bigg(Z_{n+1-\lfloor nt\rfloor} > \frac{(1-a)x}{a}\bigg) \le \exp\bigg(-\theta_1 \frac{(1-a)x}{a}\bigg)\mathbb{E}[\exp(\theta_1 Z_{n+1-\lfloor nt\rfloor})].$$

Since $Z_{n+1-\lfloor nt\rfloor} = \sum_{i=1}^{n+1-\lfloor nt\rfloor} \xi_i$, it follows that $\mathbb{E}[\exp(\theta_1 Z_{n+1-\lfloor nt\rfloor})] = (1-\theta_1)^{-(n+1-\lfloor nt\rfloor)}$ for $\theta_1 < 1$. Substituting this into (3.3), we obtain

$$\mathbb{P}(T_{(\lfloor nt\rfloor)} < a) \le (1-\theta_1)^{-(n+1-\lfloor nt\rfloor)} \int_0^\infty \exp\bigg(-\theta_1 \frac{1-a}{a} x\bigg)\mathbb{P}(Z_{\lfloor nt\rfloor} \in \mathrm{d}x).$$

Recognize that the integral above represents the moment generating function of $Z_{\lfloor nt\rfloor} = \sum_{i=1}^{\lfloor nt\rfloor} \xi_i$. Since $(1-a)/a > 0$, if $1 > \theta_1 > a/(a-1)$, it follows that

$$\int_0^\infty \exp\bigg(-\theta_1 \frac{1-a}{a} x\bigg)\mathbb{P}(Z_{\lfloor nt\rfloor} \in \mathrm{d}x) = \bigg(1 + \theta_1 \frac{1-a}{a}\bigg)^{-\lfloor nt\rfloor}.$$

Combining these results, it follows that

$$\mathbb{P}(T_{(\lfloor nt\rfloor)} < a) \le (1-\theta_1)^{-(n+1-\lfloor nt\rfloor)}\bigg(1 + \theta_1 \frac{1-a}{a}\bigg)^{-\lfloor nt\rfloor}.$$

Similarly, it can be shown that, for any $1 \ge b > t$,

$$\mathbb{P}(T^n(t) > b) \le (1-\theta_1)^{-(n+1-\lfloor nt\rfloor)}\bigg(1 + \theta_1 \frac{1-b}{b}\bigg)^{-\lfloor nt\rfloor}$$

if $1 > \theta_1 > b/(b-1)$.

Thus, it follows that

$$\begin{aligned}
\mathbb{P}(T^n(t) \in F) &\le \mathbb{P}(T^n(t) \in (x_-, x_+)^c) \\
&\le \mathbb{P}(T^n(t) \le x_-) + \mathbb{P}(T^n(t) \ge x_+) \\
&\le (1-\theta_1)^{-(n+1-\lfloor nt\rfloor)}\bigg[\bigg(1 + \theta_1 \frac{1-x_-}{x_-}\bigg)^{-\lfloor nt\rfloor} + \bigg(1 + \theta_1 \frac{1-x_+}{x_+}\bigg)^{-\lfloor nt\rfloor}\bigg] \\
&\le 2 \max_{x \in F}\bigg\{(1-\theta_1)^{-(n+1-\lfloor nt\rfloor)}\bigg(1 + \theta_1 \frac{1-x}{x}\bigg)^{-\lfloor nt\rfloor}\bigg\}. \qquad (3.4)
\end{aligned}$$

Now, for any $x \in [0, 1]$, it can be seen that $(1-\theta_1)^{-(n+1-\lfloor nt\rfloor)}(1 + \theta_1((1-x)/x))^{-\lfloor nt\rfloor}$ has a unique maximizer at $\theta_1^* = (t-x)(1-x)^{-1}$. Substituting this into (3.4), it follows that

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(T_n(t) \in F) \le \max_{x \in F}\bigg\{-(1-t)\log\bigg(\frac{1-t}{1-x}\bigg) - t\log\bigg(\frac{t}{x}\bigg)\bigg\} = -\inf_{x \in F} I_t(x).$$

(ii) Next, let $G \subset [0, 1]$ be an open set, such that $t \notin G$ and $t < \inf\{G\}$. For each point $x \in G$, there exists a $\delta > 0$ (small) such that $(x-\delta, x+\delta) \subset G$. Once again, appealing to Proposition 3.1, we have

$$\begin{aligned}
&\mathbb{P}(T^n(t) \in (x-\delta, x+\delta)) \\
&\quad = \mathbb{P}\bigg(\frac{\bar{Z}_{\lfloor nt\rfloor}}{\bar{Z}_{\lfloor nt\rfloor} + \bar{Z}_{n+1-\lfloor nt\rfloor}} \in (x-\delta, x+\delta)\bigg) \\
&\quad = \int_{z_1=0}^\infty \mathbb{P}(\bar{Z}_{\lfloor nt\rfloor} \in \mathrm{d}z_1)\mathbb{P}\bigg(\bar{Z}_{n+1-\lfloor nt\rfloor} \in z_1\bigg(-1 + \frac{1}{x+\delta}, -1 + \frac{1}{x-\delta}\bigg)\bigg),
\end{aligned}$$

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

where $\bar{Z}_{m(n)} := n^{-1} Z_{m(n)}$ for $m(n) \in \{\lfloor nt \rfloor, n+1-\lfloor nt \rfloor\}$. Let $v > t > 0$, implying that the right-hand side of the equation above is greater than or equal to

$$\mathbb{P}(\bar{Z}_{\lfloor nt \rfloor} \geq v)\mathbb{P}\bigg(\bar{Z}_{n+1-\lfloor nt \rfloor} \in v\bigg(-1+\frac{1}{x+\delta}, -1+\frac{1}{x-\delta}\bigg)\bigg). \tag{3.5}$$

Now let $\theta_1 > 0$ and consider the partial sum of 'twisted' random variables $\{\xi_1^{\theta_1}, \ldots, \xi_{\lfloor nt \rfloor}^{\theta_1}\}$, $Z_{\lfloor nt \rfloor}^{\theta_1} = \sum_{i=1}^{\lfloor nt \rfloor} \xi_i^{\theta_1}$, where the distribution of $\xi_1^{\theta_1}$ is (by an exponential change of measure)

$$\frac{\mathbb{P}(\xi_1^{\theta_1} \in \mathrm{d}x)}{\mathbb{P}(\xi_1 \in \mathrm{d}x)} = \frac{\mathrm{e}^{\theta_1 x}}{\mathbb{E}[\mathrm{e}^{\theta_1 \xi_1}]},$$

and, by induction,

$$\frac{\mathbb{P}(Z_{\lfloor nt \rfloor}^{\theta_1} \in \mathrm{d}x)}{\mathbb{P}(Z_{\lfloor nt \rfloor} \in \mathrm{d}x)} = \frac{\mathrm{e}^{\theta_1 x}}{(\mathbb{E}[\mathrm{e}^{\theta_1 \xi_1}])^{\lfloor nt \rfloor}}.$$

Define $\Lambda_n(\theta_1) := \log(\mathbb{E}[\mathrm{e}^{\theta_1 \xi_1}])^{\lfloor nt \rfloor}$, and consider $\mathbb{P}(\bar{Z}_{\lfloor nt \rfloor} > v)$. From the proof of Cramér's theorem (see [7, Chapter 2]), we have, for $\theta_1 > 0$,

$$\frac{1}{n}\log \mathbb{P}(Z_{\lfloor nt \rfloor} > nv) \geq -\theta_1 v - \frac{\lfloor nt \rfloor}{n}\log(1-\theta_1) + \frac{1}{n}\log \mathbb{P}(Z_{\lfloor nt \rfloor}^{\theta_1} > nv).$$

A straightforward calculation yields

$$\frac{1}{n}\mathbb{E}\bigg[\sum_{i=1}^{\lfloor nt \rfloor} \xi_i^{\theta_1}\bigg] = \frac{\lfloor nt \rfloor}{n}\frac{1}{1-\theta_1}.$$

Thus, we want to twist the random variables such that $t/(1-\theta_1) > v$, in which case $\mathbb{P}(Z_{\lfloor nt \rfloor}^{\theta_1} > nv) \to 1$ as $n \to \infty$ by the weak law of large numbers. It follows that

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(\bar{Z}_{\lfloor nt \rfloor} > v) \geq -\theta_1 v - t\log(1-\theta_1). \tag{3.6}$$

On the other hand, consider the second probabilistic statement in (3.5),

$$\mathbb{P}\bigg(\bar{Z}_{n+1-\lfloor nt \rfloor} \in v\bigg(-1+\frac{1}{x+\delta}, -1+\frac{1}{x-\delta}\bigg)\bigg).$$

Following a similar argument as above, we consider $\{\xi_1^{\theta_2}, \ldots, \xi_{n+1-\lfloor nt \rfloor}^{\theta_2}\}$, the twisted random variables , and define $\tilde{\Lambda}_n(\theta_2) := \log(\mathbb{E}[\mathrm{e}^{\theta_2 \xi_1}])^{n+1-\lfloor nt \rfloor} = -(n+1-\lfloor nt \rfloor)\log(1-\theta_2)$ so that

$$\begin{aligned}
&\mathbb{P}\bigg(\bar{Z}_{n+1-\lfloor nt \rfloor} \in v\bigg(-1+\frac{1}{x+\delta}, -1+\frac{1}{x-\delta}\bigg)\bigg) \\
&\quad = \int_{v(-1+1/(x+\delta))}^{v(-1+1/(x-\delta))} \mathbb{P}(\bar{Z}_{n+1-\lfloor nt \rfloor} \in \mathrm{d}y) \\
&\quad = \int_{v(-1+1/(x+\delta))}^{v(-1+1/(x-\delta))} \exp(-n\theta_2 y)\exp(\tilde{\Lambda}_n(\theta_2))\mathbb{P}(\bar{Z}_{n+1-\lfloor nt \rfloor}^{\theta_2} \in \mathrm{d}y)
\end{aligned}$$

$$\geq \exp\biggl(-n\theta_2 v\biggl(-1+\frac{1}{(x-\delta)}\biggr)\biggr)\exp(\tilde{\Lambda}_n(\theta_2))$$
$$\times\,\mathbb{P}\biggl(\bar{Z}^{\theta_2}_{n+1-\lfloor nt\rfloor}\in v\biggl(-1+\frac{1}{(x+\delta)},-1+\frac{1}{(x-\delta)}\biggr)\biggr). \tag{3.7}$$

Observe that

$$\frac{1}{n}\mathbb{E}[Z^{\theta_2}_{n+1-\lfloor nt\rfloor}]=\frac{n+1-\lfloor nt\rfloor}{n}\frac{1}{1-\theta_2}.$$

Thus, we should twist the random variables such that

$$\frac{1-t}{1-\theta_2}\in v\biggl(-1+\frac{1}{(x+\delta)},-1+\frac{1}{(x-\delta)}\biggr),$$

implying that

$$\mathbb{P}\biggl(\bar{Z}^{\theta_2}_{n+1-\lfloor nt\rfloor}\in v\biggl(-1+\frac{1}{(x+\delta)},-1+\frac{1}{(x-\delta)}\biggr)\biggr)\to 1\quad\text{as } n\to\infty$$

as a consequence of the weak law of large numbers. From (3.7), it follows that

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}\biggl(\bar{Z}_{n+1-\lfloor nt\rfloor}\in v\biggl(-1+\frac{1}{(x+\delta)},-1+\frac{1}{(x-\delta)}\biggr)\biggr)$$
$$\geq -\theta_2 v\biggl(-1+\frac{1}{x+\delta}\biggr)-(1-t)\log(1-\theta_2). \tag{3.8}$$

Using the limits in (3.6) and (3.8), it follows that, for any $0<\varepsilon<\delta$,

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}(T^n(t)\in(x-\varepsilon,x+\varepsilon))$$
$$\geq -\theta_1 v - t\log(1-\theta_1)-\theta_2 v\biggl(-1+\frac{1}{x+\varepsilon}\biggr)-(1-t)\log(1-\theta_2).$$

This is valid for any $v>t$. In particular, setting $v=x-\varepsilon$, we obtain

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbb{P}(T^n(t)\in(x-\varepsilon,x+\varepsilon))$$
$$\geq -\theta_1(x-\varepsilon)-t\log(1-\theta_1)-\theta_2\biggl(\frac{x-\epsilon}{x+\varepsilon}\biggr)(1-(x+\varepsilon))-(1-t)\log(1-\theta_2).$$

Now consider the function

$$I(\theta_1,\theta_2):=\theta_1(x-\varepsilon)+t\log(1-\theta_1)+\theta_2\biggl(\frac{x-\varepsilon}{x+\varepsilon}\biggr)(1-(x+\varepsilon))+(1-t)\log(1-\theta_2).$$

For $\theta_2,\theta_1<1$, it is straightforward to see that the Hessian is positive semidefinite, implying it is convex. The unique minimizer of $I(\theta_1,\theta_2)$ is

$$(\theta_1^*,\theta_2^*)=\biggl(\frac{1-t}{x-\varepsilon},1-\biggl(\frac{x+\varepsilon}{x-\varepsilon}\biggr)\biggl(\frac{1-t}{1-(x+\varepsilon)}\biggr)\biggr).$$

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

Letting $\epsilon \to 0$, it follows that

$$I(\theta_1^*, \theta_2^*) = t \log\left(\frac{t}{x}\right) + (1-t)\log\left(\frac{1-t}{1-x}\right) = I_t(x).$$

Thus, we have

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(T^n(t) \in (x-\delta, x+\delta)) \geq -I_t(x).$$

Next, by definition, it follows that, for small enough $\delta > 0$,

$$\frac{1}{n}\log \mathbb{P}(T^n(t) \in G) \geq \sup_{x\in G} \frac{1}{n}\log \mathbb{P}(T^n(t) \in (x-\delta, x+\delta)),$$

implying that

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(T^n(t) \in G) \geq -\inf_{x\in G} I_t(x).$$

On the other hand, if $t > \sup\{G\}$, we will now consider a $v > 1-t > 0$ in the lower bounding argument used in (3.5). Since the remaining arguments are identical to the previous steps, we will not repeat them. This proves the large deviation lower bound.

Finally, observe that the rate function I_t is continuous and convex. Consider the level set $L(c) = \{x \in [0, 1]\colon I_t(x) \leq c\}$ for $c > 0$. Let $\{x_n,\ n \geq 1\}$ be a sequence points in the set $L(c)$ such that $x_n \to x^* \in (0, 1)$ as $n \to \infty$. Since I_t is continuous, it follows that $I_t(x_n) \to I_t(x^*)$ as $n \to \infty$. Suppose that $I_t(x^*) > c$, then the only way this can happen is if there is a singularity at x^*. However, this contradicts the fact that I_t is continuous on the domain (0, 1), implying that $x^* \in L(c)$. Therefore, it is the case that $L(c)$ is closed. Furthermore, this level set is bounded (by definition), implying that it is compact. Thus, I_t is a good rate function as well. □

Now suppose that $\{\tilde{T}_{(i)},\ i \leq n\}$ are the ordered statistics of random variables with distribution F (assumed to have positive support) that is absolutely continuous with respect to the Lebesgue measure, and strictly increasing. Define

$$\tilde{I}_t(y) := \inf_{x\in[0,1]\colon F^{-1}(x)=y} I_t(x).$$

In the following corollary, we establish an LDP for the corresponding order statistics process.

Corollary 3.1. *Fix $t \in [0, 1]$. Then the ordered statistics process corresponding to $\{\tilde{T}_{(i)},\ i \leq n\}$ satisfies the LDP with good rate function $\tilde{I}_t$.*

Since F^{-1} maps [0, 1] to $[0, \infty)$, which are Hausdorff spaces, the proof is a simple application of the contraction principle [18, Equation (2.12)]. For the remainder of the paper, however, we will operate under the assumption that the arrival epochs are i.i.d. uniform random variables. The analysis below can be straightforwardly extended to the more general case where the distribution is absolutely continuous with respect to the Lebesgue measure.

3.2. LDP for the offered load

Next, recall that $\{\nu_i,\ i \geq 1\}$ is a sequence of i.i.d. random variables with cumulant generating function $\varphi(\theta) = \log \mathbb{E}[\mathrm{e}^{\theta\nu_1}] < \infty$ for some $\theta \in \mathbb{R}$. In the next theorem we show that the service process $(S^n(t),\ t \in [0, 1])$ satisfies the LDP.

Lemma 3.1. (Cramér's theorem [7].) *Fix* $t \in [0, 1]$. *Then the sequence of random variables* $\{S^n(t),\ n \geq 1\}$ *satisfies the LDP with good rate function* $\Lambda_t^*(x) := \sup_{\theta \in \mathbb{R}}\{\lambda x - t\varphi(\theta)\}$.

Note that we specifically assume that $0 \in \mathbb{D}$, since, from [7, Lemma 2.2.5], if $\mathbb{D} = \{0\}$ then $\Lambda_t^*(x)$ is equal to 0 for all x. In Lemma 2.2.20 of [7], it was proved that the rate function is good if the interior condition is satisfied. We now establish an LDP for the offered load process $(X^n(t) = S^n(t) - T^n(t),\ t \geq 0)$ by leveraging Theorem 3.1 and Lemma 3.1.

Proposition 3.2. *Fix* $t \in [0, 1]$*, and let* $\mathcal{X} := [0, 1] \times [0, \infty)$. *Then the sequence of random variables* $\{X^n(t),\ n \geq 1\}$ *satisfies the LDP with good rate function*

$$J_t(y) = \inf_{\{x \in \mathcal{X}\colon x_1 = x_2 + y\}} I_t(x_1) + \Lambda_t^*(x_2) \quad \textit{for } y \in \mathbb{R}.$$

Proof. By [18, Lemma 2.6], it is implied that $\{S^n(t),\ n \geq 1\}$ and $\{T^n(t),\ n \geq 1\}$ are large deviation tight (as defined in Definition 1.5). By [18, Corollary 2.9], it follows that $(S^n(t), T^n(t))$, $n \geq 1$, satisfy the LDP with good rate function $\tilde{I}_t(x_1, x_2) = I_t(x_1) + \Lambda_t^*(x_2)$. Now, since subtraction is trivially continuous on the topology of pointwise convergence, it follows that $\{X^n(t) = S^n(t) - T^n(t),\ n \geq 1\}$ satisfies the LDP with rate function J_t as a consequence of the contraction principle; see [18, Equation (2.12)]). □

As an example of the rate function, suppose that the service times are exponentially distributed with mean 1. Then we have

$$J_t(y) = \inf_{x \in [0,1]} \left\{ t \log\left(\frac{t}{x}\right) + (1-t)\log\left(\frac{1-t}{1-x}\right) + t\log\left(\frac{t}{x-y}\right) + (x - y - t) \right\}.$$

Some (tedious) algebra yields that $J_t(y)$ is strictly convex, and, thus, has a unique minimizer, which is the solution to the cubic equation

$$x^3 - yx^2 - 2tx + ty = 0.$$

Unfortunately, the sole real solution to this cubic equation has a complicated form, which we do not present, but can be found by using a symbolic solver.

4. An LDP for the workload

Recall that $W^n(t) = \Gamma(X^n)(t) = \sup_{0 \leq s \leq t}(X^n(t) - X^n(s))$. The key difficulty in establishing the LDP for $W^n(t)$ is the fact that while Γ is continuous on the space $(\mathcal{D}[0, 1], \mathcal{U})$, the latter is not a Polish space. Therefore, it is not possible to directly apply the contraction principle to Γ to establish the LDP. Consider, instead, the continuous process $(\tilde{W}^n(t),\ t \in [0, 1])$, formed by linearly interpolating between the jump levels of W^n; equivalently, $\tilde{W}^n = \Gamma(\tilde{X}^n)$, where $\tilde{X}^n$ is the linearly interpolated version of the offered load. We first show that $(\tilde{W}^n(t),\ t \in [0, 1])$ is asymptotically exponentially equivalent to $(W^n(t),\ t \in [0, 1])$. Next, we prove that, for each fixed $t \in [0, 1]$, $(\tilde{X}^n(s),\ s \in [0, t])$ satisfies the LDP via a projective limit. This enables us to prove that $\tilde{W}^n(t)$ satisfies the LDP by invoking the contraction principle with Γ and using the fact that $(\mathcal{C}[0, t], \mathcal{U})$ is a Polish space. Finally, by [7, Theorem 4.2.13], the exponential equivalence of the processes implies that $W^n(t)$ satisfies the LDP with the same rate function.

4.1. Exponential equivalence

We define the linearly interpolated service time process as

$$\tilde{S}^n(t) := S^n(t) + \left(t - \frac{\lfloor nt \rfloor}{n}\right)\nu_{\lfloor nt \rfloor + 1},$$

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

and the linearly interpolated arrival epoch process as

$$\tilde{T}^n(t) := T^n(t) + \left(t - \frac{\lfloor nt \rfloor}{n}\right)(T_{(\lfloor nt \rfloor+1)} - T_{(\lfloor nt \rfloor)}).$$

Define $\Delta_{n,t} := T_{(\lfloor nt \rfloor+1)} - T_{(\lfloor nt \rfloor)}$ and note that these are spacings of ordered statistics. The process $\tilde{X}^n = \tilde{S}^n - \tilde{T}^n \in \mathcal{C}[0,1]$ can now be used to define the interpolated workload process $\tilde{W}^n = \Gamma(\tilde{X}^n) \in \mathcal{C}[0,1]$. Recall that $\|\cdot\|$ is the supremum norm on $\mathcal{C}[0,1]$.

Proposition 4.1. *The processes $\tilde{W}^n$ and W^n are exponentially equivalent. That is, for any $\delta > 0$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\|\tilde{W}^n - W^n\| > \delta) = -\infty.$$

Proof. First, observe that, for each $t \in [0,1)$,

$$|S^n(t) - \tilde{S}^n(t)| \le \left|\left(t - \frac{\lfloor nt \rfloor}{n}\right)\nu_{\lfloor nt \rfloor+1}\right| \le \frac{\nu_{\lfloor nt \rfloor+1}}{n},$$

and $S^n(1) = \tilde{S}^n(1)$ by definition. Similarly,

$$|T^n(t) - \tilde{T}^n(t)| \le \left(t - \frac{\lfloor nt \rfloor}{n}\right)(T_{(\lfloor nt \rfloor+1)} - T_{(\lfloor nt \rfloor)}) = \frac{nt - \lfloor nt \rfloor}{n}\Delta_{n,t} \quad \text{for all } t \in [0,1),$$

and $T^n(1) = \tilde{T}^n(1)$. Now let $\{E_1, \dots, E_{n+1}\}$ be a collection of independent unit-mean exponential random variables, and define $Z_{n+1} := \sum_{i=1}^{n+1} E_i$. Recall (from [6, pp. 134–136], for instance) that the spacings of the uniform ordered statistics are equal in distribution to the ratio $\Delta_{n,t} \stackrel{\mathrm{D}}{=} E_1/Z_{n+1}$. It follows that

$$\|X^n - \tilde{X}^n\| \le \|S^n - \tilde{S}^n\| + \|T^n - \tilde{T}^n\| \le \left\|\frac{\nu_{\lfloor nt \rfloor+1}}{n}\right\| + \left\|\left(\frac{nt - \lfloor nt \rfloor}{n}\right)\Delta_{n,t}\right\|.$$

Now consider the measure of the event $\{\|X^n - \tilde{X}^n\| > 2\delta\}$, and use the inequality above to obtain

$$\begin{aligned}
\mathbb{P}(\|X^n - \tilde{X}^n\| > 2\delta) &\le \mathbb{P}\left(\left\|\frac{\nu_{\lfloor nt \rfloor+1}}{n}\right\| + \left\|\left(\frac{nt - \lfloor nt \rfloor}{n}\right)\Delta_{n,t}\right\| > 2\delta\right) \\
&\le \mathbb{P}\left(\left\|\frac{\nu_{\lfloor nt \rfloor+1}}{n}\right\| > \delta\right) + \mathbb{P}\left(\left\|\left(\frac{nt - \lfloor nt \rfloor}{n}\right)\Delta_{n,t}\right\| > \delta\right) \\
&\le n\mathbb{P}(\nu_1 > n\delta) + n\mathbb{P}(\Delta_{n,1} > n\delta),
\end{aligned} \tag{4.1}$$

where $\mathbb{P}(\|\nu_{\lfloor nt \rfloor+1}\| > n\delta) = \mathbb{P}((\sup_{0\le t<1} \nu_{\lfloor nt \rfloor+1} > n\delta) = \mathbb{P}(\bigcup_{m=1}^n \{\nu_i > n\delta\}) \le n\mathbb{P}(\nu_1 > n\delta)$ follows from the union bound and the fact that the service times are assumed to be i.i.d. Similarly, since $\Delta_{n,t} \stackrel{\mathrm{D}}{=} E_1/Z_{n+1}$ and $n^{-1} > n^{-1}(nt - \lfloor nt \rfloor)$ for all $t \in [0,1]$ and $n \ge 1$, we obtain the bound on $\Delta_{n,t}$ by similar arguments. Note that we have abused the notation slightly in (4.1) and again used $\Delta_{n,m} = T_{(m)} - T_{(m-1)}$ for $m \in \{1, \dots, n\}$ with the understanding that $T_{(0)} = 0$.

Using Chernoff's inequality, we obtain

$$\mathbb{P}(\nu_1 > n\delta) \le \mathrm{e}^{-n\delta\theta_1}\mathrm{e}^{\varphi(\theta_1)},$$

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

so that

$$\limsup_{n\to\infty} \frac{1}{n} \log(n\mathbb{P}(\nu_1 > n\delta)) \le -\theta_1\delta \quad \text{for all } \theta \in \mathbb{R}. \tag{4.2}$$

On the other hand, we have

$$\mathbb{P}(\Delta_{n,1} > n\delta) = \mathbb{P}(E_1(1-n\delta) > n\delta Z_n) = \int_0^\infty \mathbb{P}\bigg(E_1 > x\frac{n\delta}{1-n\delta}\bigg)\mathbb{P}(Z_n \in \mathrm{d}x),$$

using the fact that $\{E_i,\ 1 \le i \le n\}$ are i.i.d. Again, using Chernoff's inequality, the right-hand side of the equation above is less than or equal to

$$\frac{1}{1-\theta_2}\int_0^\infty \exp\bigg(-\theta_2 x\frac{n\delta}{1-n\delta}\bigg)\mathbb{P}(Z_n \in \mathrm{d}x) \quad \text{for all } \theta_2 < 1$$

or otherwise equal to

$$\frac{1}{1-\theta_2}\bigg(\frac{1-n\delta}{1-n\delta(1-\theta_2)}\bigg)^n.$$

It follows that

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}(\Delta_{n,1} > n\delta) &\le \limsup_{n\to\infty}\frac{\log(1-\theta_2)}{n} + \limsup_{n\to\infty}\log\bigg(\frac{1-n\delta}{1-n\delta(1-\theta_2)}\bigg)\\ &= -\log(1-\theta_2),\end{aligned}$$

and so

$$\limsup_{n\to\infty}\frac{1}{n}\log(n\mathbb{P}(\Delta_{n,1} > n\delta)) \le -\log(1-\theta_2) \quad \text{for all } \theta_2 < 1. \tag{4.3}$$

Now, (4.1)–(4.3), together with the principle of the largest term (see [7, Lemma 1.2.15]), imply that

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}(\|X^n - \tilde{X}^n\|) \le \max\{-\theta_1\delta, -\log(1-\theta_2)\}.$$

Since $\theta_1 \in \mathbb{R}$ and $\theta_2 \in (-\infty, 1)$, by letting $\theta_1 \to \infty$ and $\theta_2 \to -\infty$ simultaneously, it follows that

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}(\|X^n - \tilde{X}^n\| > 2\delta) = -\infty.$$

Finally, using the fact that the map Γ is Lipschitz in $(\mathcal{D}, J_1)$ (see [20, Theorem 13.5.1]), we have

$$\mathbb{P}(\|W^n - \tilde{W}^n\| > 4\delta) \le \mathbb{P}(\|X^n - \tilde{X}^n\| > 2\delta),$$

and, thus,

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbb{P}(\|W^n - \tilde{W}^n\| > 4\delta) = -\infty.$$

Since $\delta > 0$ is arbitrary, the theorem is proved. □

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

4.2. Sample path LDP for the offered load

First, we prove an LDP for the increments of the offered load process. Fix $t \in [0, 1]$, and consider an arbitrary d-partition of $[0, t]$, $j := \{0 \le t_1 < t_2 < \cdots < t_d \le t\}$, so that the increments are $\mathbf{\Delta}_n^X(j) = \mathbf{\Delta}_n^S(j) - \mathbf{\Delta}_n^T(j)$, where

$$\mathbf{\Delta}_n^T(j) := (T^n(t_1), T^n(t_2) - T^n(t_1), \dots, T^n(t) - T^n(t_d)),$$

and

$$\mathbf{\Delta}_n^S(j) = (S^n(t_1), S^n(t_2) - S^n(t_1), \dots, S^n(t) - S^n(t_d)).$$

Now, using (3.1), it follows that

$$\mathbf{\Delta}_n^T(j) \stackrel{\text{D}}{=} \frac{1}{Z_{n+1}}(Z_{\lfloor nt_1 \rfloor}, Z_{\lfloor nt_2 \rfloor} - Z_{\lfloor nt_1 \rfloor}, \dots, Z_{\lfloor nt \rfloor} - Z_{\lfloor nt_d \rfloor}).$$

A straightforward calculation shows that the cumulant generating function of the $(d + 1)$-dimensional random vector $\mathbb{Z}_n := (Z_{\lfloor nt_1 \rfloor}, \dots, Z_{\lfloor nt \rfloor} - Z_{\lfloor nt_d \rfloor}, Z_{n+1})$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}[\exp(\langle \boldsymbol{\lambda}, \mathbb{Z}_n \rangle)] = \Lambda(l) \quad \text{for } \boldsymbol{l} \in \mathbb{R}^{d+1}, \tag{4.4}$$

where

$$\Lambda_j(\boldsymbol{l}) := \begin{cases} -\displaystyle\sum_{i=1}^{d} (t_i - t_{i-1}) \log(1 - l_i - l_{d+1}) - (1 - t) \log(1 - l_{d+1}), & \boldsymbol{l} \in \mathbb{D}_\Lambda, \\ +\infty, & \boldsymbol{l} \notin \mathbb{D}_\Lambda, \end{cases}$$

and $\mathbb{D}_\Lambda := \{\boldsymbol{\lambda} \in \mathbb{R}^{d+1} \colon \max_{1 \le i \le d} \lambda_i + \lambda_{d+1} < 1, \text{ and } \lambda_{d+1} < 1\}$; note, $t_0 := 0$. We also define the function

$$\Lambda_j^*(\boldsymbol{x}) := \sup_{\boldsymbol{l} \in \mathbb{D}_\Lambda} \sum_{i=1}^{d} (l_i + l_{d+1}) x_i + (t_i - t_{i-1}) \log(1 - l_i - l_{d+1})$$
$$+ l_{d+1} x_{d+1} + (1 - t) \log(1 - l_{d+1}).$$

Now define the continuous function $\Phi \colon \mathbb{R}^{d+1} \to \mathbb{R}^d$ as $\Phi(\boldsymbol{x}) = (x_1, \dots, x_d)/\sum_{i=1}^{d+1} x_i$. We can now state the LDP for the increments $\mathbf{\Delta}_n^T(j)$.

Lemma 4.1. *Let* $j := \{0 \le t_1 < t_1 < \cdots < t_d \le t\}$ *be an arbitrary partition of* $[0, t]$*. Then the increments of the ordered statistics process* $\mathbf{\Delta}_n^T(j)$ *satisfy the LDP with good rate function* $\hat{\Lambda}_j(\boldsymbol{y}) = \inf_{\boldsymbol{x} \in \mathbb{R}^{d+1} \colon \Phi(\boldsymbol{x}) = \boldsymbol{y}} \Lambda_j^*(\boldsymbol{x})$ *for all* $\boldsymbol{y} \in (0, 1]^d$*. Furthermore,*

$$\hat{\Lambda}_j(\boldsymbol{y}) = \sum_{i=1}^{d} (t_i - t_{i-1}) \log\left(\frac{t_i - t_{i-1}}{y_i}\right) + (1 - t) \log\left(\frac{1 - t}{1 - \sum_{l=1}^{d} y_l}\right).$$

Proof. Equation (4.4) implies that the sufficient conditions of the Gartner–Ellis theorem [7, Theorem 2.3.6] are satisfied, so that $\boldsymbol{Z}_n$ satisfies the LDP with rate function Λ_j^*. Equivalently, the random vector $(Z_{\lfloor nt_1 \rfloor}, Z_{\lfloor nt_1 \rfloor} - Z_{\lfloor nt_2 \rfloor}, \dots, Z_{\lfloor nt \rfloor} - Z_{\lfloor nt_d \rfloor}, Z_{n+1} - Z_{\lfloor nt \rfloor})$ satisfies the LDP

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

with good rate function Λ_j^*. Now, since $\mathbb{R}^{d+1}$ and $\mathbb{R}^d$ are Polish spaces, the contraction principle applied to the map Φ yields the LDP. Finally, it is straightforward to check that the Hessian of $\Lambda_j^*(\boldsymbol{x})$ is positive semidefinite, implying that the latter is convex. It can now be seen that the minimizer $\boldsymbol{x}^*$ is such that $\sum_{j=1}^{d+1} x_j^* = 1$ and $x_i^* = y_i$, for a given $\boldsymbol{y} \in (0, 1]^d$. The final expression for $\hat{\Lambda}_j(\boldsymbol{y})$ follows. □

As a sanity check, we show that if $d = 1$ the rate function $\hat{\Lambda}_j(\boldsymbol{y})$ is precisely the rate function I_t in (3.2).

Corollary 4.1. *Let $j = \{0 \le t_1 \le t\}$ and $d = 1$, then the rate function is*

$$\hat{\Lambda}_j(y) = t_1 \log\left(\frac{t_1}{y}\right) + (1 - t_1) \log\left(\frac{1 - t_1}{1 - y}\right) \quad \textit{for all } y \in (0, 1).$$

Proof. Since $d = 1$, by definition, we have, for all $\boldsymbol{x} \in \mathbb{R}^2$,

$$\Lambda_j^*(\boldsymbol{x}) = (\lambda_1 + \lambda_2)x_1 + t_1 \log(1 - \lambda_1 - \lambda_2) + \lambda_2 x_2 + (1 - t_1) \log(1 - \lambda_2).$$

Substituting the unique maximizer

$$(\lambda_1^*, \lambda_2^*) = \left(\frac{1 - t_1}{x_2} - \frac{t_1}{x_1}, 1 - \frac{1 - t_1}{x_2}\right),$$

it follows that

$$\Lambda_j^*(\boldsymbol{x}) = (x_1 + x_2 - 1) + t_1 \log\left(\frac{t_1}{x_1}\right) + (1 - t_1) \log\left(\frac{1 - t_1}{x_2}\right).$$

Finally, using the fact that $\boldsymbol{x}^* = \arg\inf\{\Lambda^*(\boldsymbol{x})\}$ satisfies $x_1^* + x_2^* = 1$, the corollary is thus proved. □

As an aside, from this result we see that Theorem 3.1 could also be established as a corollary of Lemma 4.1. However, while the proof is straightforward, it is also somewhat 'opaque': the proof of Theorem 3.1 explicitly demonstrates how the long-range dependence inherent in the order statistics process affects the LDP and is, we believe, more clarifying as a consequence. Next, we use this result to prove a sample-path LDP for the ordered statistics process $(\tilde{T}^n(s),\ s \in [0, t])$ (for each fixed t) in the topology of pointwise convergence on the space $\mathcal{C}[0, t]$. Observe that the exponential equivalence of $\tilde{T}^n$ and T^n implies that the increments of $\tilde{T}^n$ satisfy the LDP in Lemma 4.1.

Let $\mathcal{J}_t$ be the space of all possible finite partitions of $[0, t]$. Note that, for each partition $j = \{0 \le t_1 < t_1 < \cdots < t_d \le t\} \in \mathcal{J}_t$, the increments take values in the space $[0, 1]^d$, which is Hausdorff. Thus, we can appeal to the Dawson–Gartner theorem [7, Theorem 4.6.1] to establish the LDP for the sample path $(\tilde{T}^n(s),\ s \in [0, t])$ via a projective limit. Let $p_j : \mathcal{C}[0, t] \to \mathbb{R}^{|j|}$ be the canonical projections of the coordinate maps, and $\mathcal{X}$ be space of all functions in $\mathcal{C}[0, t]$ equipped with the topology of pointwise convergence. Recall that any nondecreasing continuous function $\phi \in \bar{\mathcal{C}}[0, t]$ is of bounded variation, so that $\phi = \phi^{(a)} + \phi^{(s)}$ by the Lebesgue decomposition theorem; here, $\phi^{(a)}$ is the absolutely continuous component and the $\phi^{(s)}$ is the singular component of ϕ. Recall, too, that a singular component has derivative that satisfies $\dot{\phi}^{(s)}(t) = 0$ for almost every t.

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

Lemma 4.2. *Fix $t \in [0,1]$. Then the sequence of sample paths $\{(\tilde{T}^n(s), s \in [0,t]),\ n \geq 1\}$ satisfies the LDP with good rate function*

$$\hat{\boldsymbol{L}}_t(\phi) = -\int_0^t \log\Big(\dot{\phi}^{(a)}(s)\Big)\,\mathrm{d}s + (1-t)\log\bigg(\frac{1-t}{1-\phi(t)}\bigg) \quad \textit{for all } \phi \in \bar{\mathcal{C}}[0,t].$$

Proof. The proof largely follows that of [7, Lemma 5.1.8]. There are two steps to establishing this result. First, we must show that the space $\mathcal{X}$ coincides with the projective limit $\tilde{\mathcal{X}}$ of $\{\mathcal{Y}_j = \mathbb{R}^{|j|},\ j \in \mathcal{J}_t\}$. This, however, follows immediately from the proof of [7, Lemma 5.1.8]. Second, we must argue that

$$\tilde{\boldsymbol{L}}_t(\phi) := \sup_{0 \leq t_1 < \cdots < t_k \leq t} \sum_{l=1}^k (t_l - t_{l-1}) \log\bigg(\frac{t_l - t_{l-1}}{\phi(t_l) - \phi(t_{l-1})}\bigg) + (1-t)\log\bigg(\frac{1-t}{1-\phi(t)}\bigg)$$

is equal to $\hat{\boldsymbol{L}}_t(\phi)$. Without loss of generality, assume that $t_k = t$. Recall that ϕ has bounded variation, implying that $\phi^{(a)}(t) = \int_0^t \dot{\phi}(s)\,\mathrm{d}s$ or, equivalently, $\dot{\phi}^{(a)}(s) = \dot{\phi}(s)$ almost everywhere (a.e.) $s \in [0,t]$. Since $\log(\cdot)$ is concave, Jensen's inequality implies that

$$\begin{aligned}
\sum_{l=1}^k (t_l - t_{l-1}) \log\bigg(\frac{\phi(t_l) - \phi(t_{l-1})}{t_l - t_{l-1}}\bigg) &= \sum_{l=1}^k (t_l - t_{l-1}) \log\bigg(\frac{\int_{t_l}^{t_{l-1}} \dot{\phi}(r)\,\mathrm{d}r}{t_l - t_{l-1}}\bigg) \\
&\geq \sum_{l=1}^k \int_{t_l}^{t_{l-1}} \log(\dot{\phi}(r))\,\mathrm{d}r \\
&= \int_0^t \log(\dot{\phi}^{(a)}(r))\,\mathrm{d}r,
\end{aligned}$$

so that $\tilde{\boldsymbol{L}}_t(\phi) \leq \hat{\boldsymbol{L}}_t(\phi)$.

Next, define

$$\phi_n(r) = n\bigg(\phi^{(a)}\bigg(\frac{[nr]+1}{n}\bigg) - \phi^{(a)}\bigg(\frac{[nr]}{n}\bigg)\bigg) + n\bigg(\phi^{(s)}\bigg(\frac{[nr]+1}{n}\bigg) - \phi^{(s)}\bigg(\frac{[nr]}{n}\bigg)\bigg),$$

and observe that

$$\lim_{n\to\infty} \phi_n(r) = \dot{\phi}^{(a)}(r) \quad \text{a.e. } r \in [0,t],$$

since

$$n\bigg(\phi^{(a)}\bigg(\frac{[nr]+1}{n}\bigg) - \phi^{(a)}\bigg(\frac{[nr]}{n}\bigg)\bigg) \to \dot{\phi}^{(a)}(r)$$

and

$$n\bigg(\phi^{(s)}\bigg(\frac{[nr]+1}{n}\bigg) - \phi^{(s)}\bigg(\frac{[nr]}{n}\bigg)\bigg) \to \dot{\phi}^{(s)}(r) = 0 \quad \text{a.e. } r \in [0,t] \text{ as } n \to \infty.$$

Now consider the uniform partition $0 = t_0 < t_1 < \cdots < t_n = t$ of $[0,t]$, where $t_l = tl/n$, so that

$$\begin{aligned}
&\liminf_{n\to\infty} \sum_{l=1}^n \frac{1}{n} \log(n(\phi(t_l) - \phi(t_{l-1}))) \\
&\quad = \liminf_{n\to\infty} \sum_{l=1}^n \frac{1}{n} \log(n(\phi^{(a)}(t_l) - \phi^{(a)}(t_{l-1})) + n(\phi^{(s)}(t_l) - \phi^{(s)}(t_{l-1})))
\end{aligned}$$

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

$$
\begin{aligned}
&= \liminf_{n\to\infty} \int_0^t \log(\phi_n(r))\,\mathrm{d}r \\
&\ge \int_0^t \liminf_{n\to\infty} \log(\phi_n(r))\,\mathrm{d}r \\
&= \int_0^t \log(\dot{\phi}^{(a)}(r))\,\mathrm{d}r,
\end{aligned}
$$

where the inequality follows from Fatou's lemma and the last equality is a consequence of the continuity of $\log(\cdot)$. Now, by definition,

$$\tilde{\boldsymbol{L}}_t(\phi) \ge \liminf_{n\to\infty} -\sum_{l=1}^{n} \frac{1}{n} \log(n(\phi(t_l) - \phi(t_{l-1}))) + (1-t)\log\bigg(\frac{1-t}{1-\phi(t)}\bigg),$$

implying that

$$\tilde{\boldsymbol{L}}_t(\phi) \ge -\int_0^t \log(\dot{\phi}^{(a)}(r))\,\mathrm{d}r + (1-t)\log\bigg(\frac{1-t}{1-\phi(t)}\bigg) = \hat{\boldsymbol{L}}_t(\phi). \qquad \square$$

For the service process, we consider the following result implied by [19]. As noted in [9], the form of Mogulskii's theorem presented in [7, Theorem 5.1.2] does not cover the case of exponentially distributed service times, thus we appeal to the generalization proved in [19]. Note that Puhalskii [19] proved the result in the M_1 topology on the space $\mathcal{D}[0, t]$, which implies convergence pointwise as required here.

Lemma 4.3. *Fix $t \in [0, 1]$. Then the sequence of sample paths $\{(\tilde{S}^n(s),\ s \in [0, t])\}$ satisfies the LDP with good rate function, for each $\psi \in \bar{\mathcal{C}}[0, t]$,*

$$\hat{I}_t(\psi) = \int_0^t \boldsymbol{L}^*(\dot{\psi}^{(a)}(s))\,\mathrm{d}s + \psi^{(s)}(t).$$

These two results now imply the LDP for the sequence of sample paths $\{(\tilde{X}^n(s),\ s \in [0, t])\}$.

Proposition 4.2. *Fix $t \in [0, 1]$. Then the sequence of sample paths $\{(\tilde{X}^n(s),\ s \in [0, t])\}$ satisfies the LDP with good rate function, for $\psi \in \mathcal{C}[0, t]$,*

$$\hat{J}_t(\psi) = \inf_{\phi\in\bar{\mathcal{C}}[0,t],\ \dot{\phi}(s)-\dot{\psi}(s)\ge 0,\ s\in[0,t]} \hat{\boldsymbol{L}}_t(\phi) + \hat{I}_t(\phi - \psi).$$

Proof. The independence of $(\tilde{T}^n(s),\ s \in [0, t])$ and $(\tilde{S}^n(s),\ s \in [0, t])$ for each $n \ge 1$ implies that they jointly satisfy the LDP with good rate function $\hat{\Lambda}_t(f) + \hat{I}_t(g)$ as a consequence of [18, Corollary 2.9], and where $(f, g) \in \bar{\mathcal{C}}[0, t] \times \bar{\mathcal{C}}[0, t]$. Since subtraction is continuous on the Polish space $\mathcal{C}[0, t]$ equipped with the topology of pointwise convergence, applying the contraction principle along with Lemma 4.1 and [7, Lemma 5.1.8] completes the proof. $\square$

As an illustration of the result, suppose that the service times are exponentially distributed with mean 1. Define the function

$$
\begin{aligned}
\check{J}_t(\phi, \psi) := \int_0^t \bigg(\log\Big(\dot{\phi}^{(a)}(s)\Big) + s\log\bigg(\frac{\dot{\phi}^{(a)}(s) - \dot{\psi}^{(a)}(s)}{s}\bigg)\bigg)\,\mathrm{d}s \\
- \bigg(\phi^{(s)}(t) - \psi^{(s)}(t) - \frac{t^2}{2} + (1-t)\log\bigg(\frac{1-t}{1-\phi(t)}\bigg)\bigg).
\end{aligned}
$$

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press

Then the rate function for the offered load sample path is

$$\hat{J}_t(\psi) = \inf_{\phi\in\bar{\mathcal{C}}[0,t],\, \dot{\phi}(t)-\dot{\psi}(s)\geq 0,\, s\in[0,t]} -\check{J}_t(\phi,\psi).$$

We now establish the LDP for the workload process at a fixed $t \in [0, 1]$.

Theorem 4.1. *Fix $t \in [0, 1]$. Then the sequence of random variables $\{W^n(t),\ n \geq 1\}$ satisfy the LDP with good rate function $\tilde{J}_t(y) = \inf_{\{\phi\in\mathcal{X}:\ y=\Gamma(\phi)(t)\}} \hat{J}_t(\phi)$ for all $y \in \mathbb{R}$.*

Proof. Recall that $\Gamma\colon \mathcal{C}[0, t] \to \mathcal{C}[0, t]$ is continuous. Furthermore, $\mathcal{C}[0, t]$ (under the topology of pointwise convergence) and $\mathbb{R}$ are Hausdorff spaces. Therefore, the conditions of the contraction principle [7, Theorem 4.2.1] are satisfied. Thus, it follows that $\{\tilde{W}^n(t),\ n \geq 1\}$ satisfies the LDP with the rate function $\tilde{J}_t$. Finally, the exponential equivalence proved in Proposition 4.1 implies that $\{W^n(t),\ n \geq 1\}$ satisfies the LDP with rate function $\tilde{J}_t$, thus completing the proof. □

4.3. Effective bandwidths

As noted in Section 2, our primary motivation for studying the large deviation principle is to model the likelihood that the workload at any point in time $t \in [0, 1]$ exceeds a large threshold. This is also related to the fact that most queueing models in practice have finite-sized buffers, and so understanding the likelihood that the workload exceeds the buffer capacity is crucial from a system operation perspective. More precisely, if $w \in [0, \infty)$ is the buffer capacity, we are interested in the probability of the event $\{W^n(t) > w\}$. Theorem 4.1 implies that

$$\mathbb{P}(W^n(t) > w) \leq \exp(-n\tilde{J}_t((w,\infty))),$$

where $\tilde{J}_t((w,\infty)) = \inf_{y\in(w,\infty)} \tilde{J}_t(y)$. A reasonable performance measure to consider in this model is to find the 'critical time-scale' at which the large exceedance occurs with probability at most p. That is, we would like to find

$$t^* := \inf\{t > 0 \mid \exp(-n\tilde{J}_t((w,\infty))) \leq p\}.$$

Consider the inequality $\tilde{J}_t((w,\infty)) \geq -(1/n)\log p$. Using the definition of a rate function, we have

$$\begin{aligned}
&\inf_{f\in\mathcal{X}:\ y=\Gamma(f)(t)}\ \inf_{\phi\in\bar{\mathcal{C}}^0_f[0,t]} -\int_0^t (\log(\dot{\phi}^{(a)}(r)) + \boldsymbol{L}^*(\dot{\phi}^{(a)}(r) - \dot{f}^{(a)}(r)))\,\mathrm{d}r \\
&\quad + (1-t)\log\biggl(\frac{1-t}{1-\phi(t)}\biggr) + (\phi^{(s)}(t) - f^{(s)}(t)) \\
&\quad \geq -\frac{\log p}{n},
\end{aligned}$$

where we define $\bar{\mathcal{C}}^0_f[0,t] := \{g \in \bar{\mathcal{C}}[0,t]\colon \dot{g}(s) - \dot{f}(s) \geq 0 \text{ for all } s \in [0,t]\}$ for brevity. The critical time-scale will be the optimizer of this constrained variational problem.

5. Conclusions

The large deviation principle derived for the 'uniform scattering' case in this paper provides the first result on the rare-event behavior of the RS/GI/1 transitory queue, building on the fluid and diffusion approximation results established in [2], [14], [16], and [17]. Our results are an important addition to the body of knowledge dealing with rare-event behavior of queueing models. In particular, a standard assumption is that the traffic model has independent increments, while our model assumed exchangeable increments in the traffic count process by design. We believe that the results in this paper are the first in the large deviations analyses of queueing models under such conditions.

Our next step in this line of research will be to extend the analysis to queues with nonuniform arrival epoch distributions, including distributions that are not absolutely continuous. In this case, the contraction principle cannot be directly applied, complicating the analysis somewhat. In [14] we made initial progress under a 'near-balanced' condition on the offered load process, where the traffic and service effort (on average) are approximately equal. However, it is unclear how to drop the assumption of near-balancedness. In particular, when the distribution is general, it is possible for the queue to enter periods of underload, overload, and critical load in the fluid limit. This must have a significant impact on how the random variables are 'twisted' to rare outcomes. We do not believe it will be possible to exploit (3.1) to establish the LDP. A further problem of interest is to consider a different acceleration regime. In the current setting we assumed that the service times ν_i are scaled by the population size. However, it is possible to entertain alternate scalings, such as $\nu_i^n = n^{-1}\nu_i(1 + \beta n^{1/3})$ as in [2], or scalings that are dependent on the operational time horizon of interest. We leave these problems to future papers.

Acknowledgements

We thank Peter W. Glynn and Vijay Subramanian for many helpful discussions over the course of writing this paper. We also thank the anonymous reviewer for pointing out that it was possible to significantly extend the analysis to prove a general LDP, and providing helpful hints on how to achieve this. This research was supported in part by the National Science Foundation through grant CMMI-1636069.

References

[1] Aldous, D. J. (1985). *Exchangeability and Related Topics*. Springer, Berlin.
[2] Bet, G., van der Hofstad, R. and van Leeuwaarden, J. S. H. (2015). Heavy-traffic analysis through uniform acceleration of queues with diminishing populations. Preprint. Available at https://arxiv.org/abs/1412.5329.
[3] Bet, G., van der Hofstad, R. and van Leeuwaarden, J. S. H. (2016). Finite-pool queues with heavy-tailed services. Preprint. Available at https://arxiv.org/abs/1605.06264.
[4] Boistard, H. (2007). Large deviations for L-statistics. *Statist. Decisions* **25,** 89–125.
[5] Chen, H. and Yao, D. D. (2001). *Fundamentals of Queueing Networks: Performance, Asymptotics, and Optimization*. Springer, New York.
[6] David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*, 3rd edn. John Wiley, Hoboken, NJ.
[7] Dembo, A. and Zeitouni, O. (2010). *Large Deviations Techniques and Applications*. Springer, Berlin.
[8] Dinwoodie, I. H. and Zabell, S. L. (1992). Large deviations for exchangeable random vectors. *Ann. Prob.* **20,** 1147–1166.
[9] Duffy, K. R., Macci, C. and Torrisi, G. L. (2011). Sample path large deviations for order statistics. *J. Appl. Prob.* **48,** 238–257.
[10] Durrett, R. (2010). *Probability: Theory and Examples*, 4th edn. Cambridge University Press.
[11] Ganesh, A., O'Connell, N. and Wischik, D. (2004). *Big Queues*. Springer, Berlin.
[12] Glynn, P. W. (2012). Perspectives on traffic modeling in networks. Stochastic Networks Conference, MIT.
[13] Glynn, P. W. (2014). Perspectives on traffic modeling. Markov Lecture, INFORMS Annual Meeting, San Francisco, CA.
[14] Glynn, P. W. and Honnappa, H. (2016). On Gaussian limits and large deviations for queues fed by high intensity randomly scattered traffic. Working paper.
[15] Honnappa, H., Jain, R. and Ward, A. R. (2016). On transitory queueing. Working paper. Available at https://arxiv.org/abs/1412.2321.
[16] Honnappa, H., Jain, R. and Ward, A. R. (2015). A queueing model with independent arrivals, and its fluid and diffusion limits. *Queueing Systems* **80,** 71–103.

[17] LOUCHARD, G. (1994). Large finite population queueing systems. The single-server model. *Stoch. Process. Appl.* **53,** 117–145.
[18] LYNCH, J. AND SETHURAMAN, J. (1987). Large deviations for processes with independent increments. *Ann. Prob.* **15,** 610–627.
[19] PUHALSKII, A. (1995). Large deviation analysis of the single server queue. *Queueing Systems Theory Appl.* **21,** 5–66. (Erratum: **23** (1996), 337.)
[20] WHITT, W. (2002). *Stochastic-Process Limits*. Springer, New York.

https://doi.org/10.1017/jpr.2017.43 Published online by Cambridge University Press