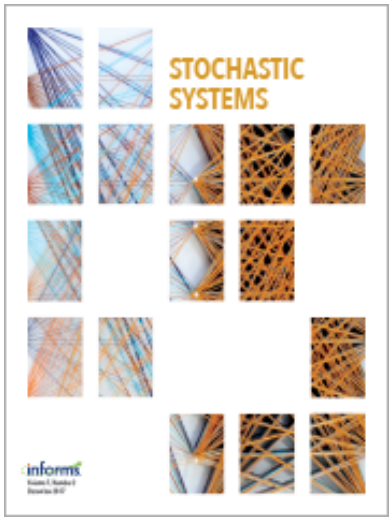


Stochastic Systems

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Two-parameter Sample Path Large Deviations for Infinite-Server Queues

Jose Blanchet, Xinyun Chen, Henry Lam

To cite this article:
Jose Blanchet, Xinyun Chen, Henry Lam (2014) Two-parameter Sample Path Large Deviations for Infinite-Server Queues. Stochastic Systems 4(1):206-249. https://doi.org/10.1287/12-SSY080

Stochastic Systems
2014, Vol. 4, No. 1, 206–249
DOI: 10.1214/12-SSY080

TWO-PARAMETER SAMPLE PATH LARGE DEVIATIONS FOR INFINITE-SERVER QUEUES

By Jose Blanchet*, Xinyun Chen† and Henry Lam‡

Boston University, Stony Brook University*† and Columbia University*‡

Let $Q_\lambda(t, y)$ be the number of people present at time t with at least y units of remaining service time in an infinite server system with arrival rate equal to $\lambda > 0$. In the presence of a non-lattice renewal arrival process and assuming that the service times have a continuous distribution, we obtain a large deviations principle for $Q_\lambda(\cdot)/\lambda$ under the topology of uniform convergence on $[0, T] \times [0, \infty)$. We illustrate our results by obtaining the most likely paths, represented as surfaces, to overflow in the setting of loss queues, and also to ruin in life insurance portfolios.

1. Introduction. The asymptotic analysis of queueing systems with many servers in heavy-traffic has received substantial attention, especially in recent years. Among the earliest references that come to mind in connection to this topic is the work of Iglehart (1965) on heavy-traffic limits for the infinite-server queue. Another highly influential paper in the area is Halfin and Whitt (1981) in the context of many server Markovian queues, which introduced a scaling that is now known as the "Quality and Efficiency Driven" regime. The ideas in these papers have fueled more recent results in the asymptotic analysis of many server systems such as: Puhalskii and Reiman (2000), Jelenkovic, Mandelbaum and Momcilovic (2004), Puhalskii and Reiman (2009), Kaspi and Ramanan (2010), Kaspi and Ramanan (2011), in the setting of many server queues, and Glynn and Whitt (1991), Decreusefond and Moyal (2008), Pang and Whitt (2010), Reed and Talreja (2012), in the setting of the infinite server queue. The asymptotic analysis of queueing systems with many servers has been motivated by applications in service engineering, in particular in the context of call centers and healthcare operations. Another set of application areas that is also very relevant, but that is infrequently mentioned in the analysis of many server systems is that of insurance mathematics. It is clear, for instance, that a portfolio of insurance policies can be directly modeled as an infinite server system;

Received July 2012.
AMS 2000 subject classifications: 60F10, 60K25.
Keywords and phrases: Large deviations, infinite-server queues, two-parameter processes, rare-event tail estimation, life insurance portfolio management.

casting insurance portfolios in this framework is particularly appealing in the setting of life insurance as we shall illustrate in Section 5.

So far most of the asymptotic analysis of many server systems has concentrated mainly on fluid and diffusion approximations on central limit scaling. Meanwhile, the literature on large deviations analysis for many server queues is not as extensive relatively; despite the fact that it is clearly of interest to understand the large deviations behavior of these types of systems. For instance, consider the consequences of dropping calls in an emergency call center or being unable to satisfy the demand for critically ill patients in the context of health-care applications. As another application, in the insurance setting, it is of interest to estimate ruin probabilities and, perhaps even more importantly, understanding the most likely path (or set of paths) to ruin. Risk theory typically concentrates on ruin probabilities for aggregated models, such as the classical ruin model (see Asmussen and Albrecher (2010)); the results in this paper, as we shall illustrate, provide a systematic way for assessing ruin probabilities for a natural class of bottom-up models.

Our main contribution in this paper is to provide the first sample-path large deviations analysis of the state descriptor of the infinite server queueing model in heavy-traffic (i.e. as the arrival rate increases to infinity without introducing any scaling on the service times). The statement of our main result, which is given in Theorem 1, features a convenient representation of a good large deviations rate function, under a strong topology. To illustrate the strength of our result, we apply it to compute the most likely path to overflow in a loss system, and also the most likely path to ruin for a life insurance portfolio that embeds an infinite server queue with a particular service cost structure. It is important to emphasize that our result takes advantage of a convenient representation of the system's description that facilitates the representation of the rate function; detailed discussion on this system's representation is given in Section 2.1. Previous large deviations analysis of the infinite server queue has concentrated on queue-length characteristics only; see, for instance, Glynn (1995) who develops large deviations for marginal quantities in the case of renewal arrivals, and Zajic (1998) who develops sample path large deviations for the queue length process of infinite server queues in tandem in the case of Poisson arrivals.

Our large deviations analysis complements results on fluid analysis and diffusion approximations recently obtained for infinite server systems. For example, Pang and Whitt (2010) have shown that the state descriptor of the infinite server queue, suitably parameterized in terms of a two-parameter stochastic process, converges after centering and re-scaling to a Gaussian and Markov process; see also Reed and Talreja (2012) who interpret the

state descriptor of the infinite server queue as a measure valued process acting on the space of tempered distributions. These recent results, in turn, extend prior work by Glynn and Whitt (1991) in the context of discrete and bounded service time distributions, and Decreusefond and Moyal (2008) for the case of Poisson arrivals. We also mention the growing literature on large deviations of measure valued processes, see for example, Léonard (2000), and Feng and Kurtz (2006) for general theory. This literature is relevant as the state of the infinite server queue at time t can be represented as a measure with point masses representing the remaining service times of the customers currently in the system. This approach requires to define the right topology on the space of measures, just as in the weak convergence analysis in Decreusefond and Moyal (2008) and Reed and Talreja (2012). It appears that the resulting topologies, however, would not be as strong as the ones that we consider here (see, for instance, the discussion on the resulting topologies in p. 3 of Léonard (2000)). Our topology is basically the same as that in Pang and Whitt (2010), which in turn is stronger than that in Decreusefond and Moyal (2008) and Reed and Talreja (2012) (which do not include the queue length process as a continuous function, for example). In addition, the rate function would involve a different representation than the one we obtain here. We believe that our representation is more convenient for applications in queueing, as we illustrate in our examples in Section 5.

The analysis of the infinite server queue is important as it serves as a building block for other models of interest. For instance, in the setting of loss models one can clearly couple the loss systems with associated infinite server systems, and in the setting of many server queues Puhalskii and Reiman (2009) shows how one can precisely understand queues with multiple servers as a perturbation of infinite server queues. Furthermore, the infinite server model is a classical model in queueing theory that serves as a direct model in important applications. Of particular interest to us, as mentioned earlier, are the applications to insurance mathematics.

The rest of the paper is organized as follows. In Section 2 we introduce our problem setting and provide a statement of our main result. This is a fundamental section and it is divided into three parts. We first shall introduce our assumptions and define our notation. Then we provide the precise mathematical statement of our result, and, finally, we will provide a heuristic argument that allows us to gain some intuition behind it. The next two sections then provide proofs. We first show our result for bounded service times in Section 3. Then, in Section 4, we apply a truncation argument to extend our result to unbounded service times. In Section 5 we apply our result to computing the most likely paths to rare events in the setting of

loss queueing systems and also in the setting of ruin probabilities for large life insurance portfolios. Finally, we include an appendix that contains the construction of a continuous function which approximates the underlying two parameter queueing process in uniform norm.

2. Assumptions, notation, and main result. The purpose of this section is threefold. First we shall clearly state our assumptions and introduce necessary notation for our development. Second, we shall explain the main large deviations result and provide a heuristic derivation of the rate function that we obtain. Finally, we shall provide a road map for the strategy behind the proof which will be presented in subsequent sections.

2.1. *Assumptions and notation.* We shall describe an underlying system corresponding to an arrival rate λ. We call the system with $\lambda = 1$, i.e. one customer per unit time, our "base system"; eventually we shall send λ to infinity in our asymptotic analysis. We collect our assumptions as follows.

Assumptions and notation concerning the arrival process. For the base system, we assume the interarrival times are i.i.d. positive random variables $(U_n : n \geq 1)$ with $E[U_n] = 1$ and finite exponential moments in a neighborhood of the origin; in precise words, $\kappa(\theta) := \log Ee^{\theta U_n} < \infty$ for some $\theta > 0$. Besides, we also assume that $(U_n : n \geq 1)$ are non-lattice in the sense that there does not exist any constant $\alpha > 0$ such that the value of U_n lies in $\{\alpha k : k = 0, 1, 2, \ldots\}$. In our λ-scaled system, the arrivals come λ times faster (i.e. the n-th interarrival times becomes U_n/λ). The associated logarithmic moment generating function of the λ-scaled service times is then $\kappa_\lambda(\theta) := \log Ee^{\theta U_n/\lambda} = \kappa(\theta/\lambda)$. Hence, following the assumptions on the base system, $\kappa_\lambda(\theta) < \infty$ for some $\theta > 0$.

The time at which the n-th arrival occurs in the base system is $A_n = U_1 + \cdots + U_n$ for $n \geq 1$. We simply define $A_0 := 0$ and then let $N(t) := \max\{n \geq 0 : A_n \leq t\}$ be the number of arrivals that have occurred up to time t in the base system. It is important to keep in mind that $N(\cdot)$ increases by one unit at discontinuity points since we are assuming that the U_n's are positive.

Eventually, we shall increase the arrival rate, so it is sensible to define $N_\lambda(t) := N(\lambda t)$.

Define the so-called infinitesimal logarithmic moment generating function for the arrival process via $\psi_N(\theta) = -\kappa^{-1}(-\theta)$ (see Glynn and Whitt (1994)). This definition is motivated by the fact that

$$\lim_{t\to\infty} \lambda^{-1} \log E \exp\left(\theta[N_\lambda(t+\delta) - N_\lambda(t)]\right) = \psi_N(\theta)\,\delta \tag{1}$$

for any $\delta > 0$. Since the U_n's are positive with probability one we have that $\psi_N(\cdot)$ is continuous and strictly convex on the positive line. We also assume that $\psi_N(\cdot)$ is continuously differentiable throughout $\mathbb{R}$. This assumption is satisfied for most arrival processes, certainly for interarrival times that are strictly positive and such that $\sup\{\kappa(\theta) : \kappa(\theta) < \infty\} = \infty$.

Assumptions and notation concerning the service times. We assume that the n-th customer that arrives to the base system (i.e. at time A_n) brings up a service requirement of size V_n that is independent of the arrival process. The sequence $(V_n : n \geq 1)$ is assumed to be i.i.d. and is independent of the arrivals $(U_n : n \geq 1)$. We write $F(x) = P(V_n \leq x)$ to denote the associated distribution function evaluated at x, and set $\bar{F}(x) := 1 - F(x)$ to be the tail distribution. Moreover, we assume that $F(\cdot)$ is continuous.

Two-parameter representation of system status. For any fixed $0 < T < \infty$, let $\bar{Q}_\lambda(t, y)$ denote the number of customers who arrived before or at time t and leave after time y in the λ-scaled system for all $(t, y) \in [0, T] \times [0, \infty)$. In detail,

$$\bar{Q}_\lambda(t, y) = \begin{cases} \bar{Q}_\lambda(0, y - t) + \sum_{n=1}^{N_\lambda(t)} I(V_n + A_n/\lambda > y) & t \leq y, \\ \bar{Q}_\lambda(y, y) + N_\lambda(t) - N_\lambda(y) & t > y. \end{cases}$$

We shall assume that the system is initially empty at the beginning. This is done for simplicity. Since we have infinitely many servers, we can incorporate the initial configuration by keeping track of its evolution independently of what occurs subsequently. Given our assumption of an initial empty system we then have that $\bar{Q}_\lambda(0, u) = 0$ for all $u \geq 0$. Note that for all $(t, y) \in [0, T] \times [0, \infty)$,

$$\bar{Q}_\lambda(t, y) = \bar{Q}_\lambda(t \wedge y, y) + N_\lambda(t) - N_\lambda(t \wedge y). \tag{2}$$

It is worth comparing the current system representation with the more common one involving the quantity $Q_\lambda(t, u)$ defined as the number of customers in the system currently at time t who have residual service time larger than $u \geq 0$; more precisely,

$$Q_\lambda(t, u) = \bar{Q}_\lambda(t, u + t). \tag{3}$$

These two system representations are equivalent in the sense that $(Q_\lambda(t, u) : t \in [0, T], u \geq 0)$ encodes the evolution of the infinite server systems and thus, such evolution can be used in principle to retrieve $(\bar{Q}_\lambda(t, u) : t \in [0, T], u \geq 0)$. We have chosen the representation based on $\bar{Q}_\lambda$ to facilitate the representation of the rate function; a more detailed discussion is given towards the end of Section 2.2.1. In addition, the representation based on $\bar{Q}_\lambda$

allows to obtain a rich large deviations principle to which one can apply the contraction principle directly to several continuous functions of interest. For instance, it follows immediately that the arrival process $N_\lambda(t) = \bar{Q}_\lambda(t,0)$, and the departure process, $D_\lambda(t) := N_\lambda(t) - \bar{Q}_\lambda(t,t)$ are continuous functions under the topology that we consider (and that we shall discuss in the next paragraphs). More applications of the contraction principle will be discussed in Section 5.

Discussion about the topological space. Let $\mathcal{D} = \{(t,y) : 0 \leq t \leq T, y \geq 0\}$ and let us write $||\cdot||_{\mathcal{C}}$ to denote the supremum norm over any set $\mathcal{C}$. The space of functions that we consider for our large deviations principle shall be denoted by $L_{+,\infty}(\mathcal{D})$ and it corresponds to bounded functions with domain in $\mathcal{D}$, such that $x(0,u) = 0$ for $u \geq 0$, $x(t,\cdot)$ is non increasing, and $x(t,\cdot)$ vanishes at infinity. We will develop the large deviations principle for the family of stochastic processes $(\bar{Q}_\lambda/\lambda : \lambda > 0)$ on the space $L_{+,\infty}(\mathcal{D})$ endowed with the topology generated by the supremum norm. Following Dembo and Zeitouni (1998) p. 4, the probability measures in path space in our development are assumed to have been completed.

Our large deviations principle for $\bar{Q}_\lambda/\lambda$ immediately implies in particular a large deviations principle in the Skorokhod topology in the space $D_{D_{\mathbb{R}}[0,\infty)}[0,T]$ which is the space of right-continuous-with-left-limits (RCLL) functions x, with domain on $[0,T]$, that take values on the space of RCLL functions taking values on $\mathbb{R}$. That is, on each time point t in $x = (x(t) : t \in [0,T]) \in D_{D_{\mathbb{R}}[0,\infty)}[0,T]$ is a function $x(t) \in D_{\mathbb{R}}[0,\infty)$. This is precisely the topology considered in Pang and Whitt (2010), who also provide a discussion on the benefits of using this topology relative to other natural (but weaker) alternative options (see Section 2.3 in Pang and Whitt (2010)).

An alternative approach that one might consider given the available results on functional weak convergence analysis of the infinite server queue, such as Reed and Talreja (2012), is to interpret the space descriptor of the infinite server queue as acting on the space of tempered distributions. We believe, however, that this approach, although elegant, has important limitations in terms of assumptions and the class of functions to which the contraction principle can be directly applied to obtain other large deviations principles of interest.

2.2. *Statement of our main result.* We are now ready to state our main result. Let $\bar{q} := (\bar{q}(t,y) : (t,y) \in \mathcal{D}) \in L_{+,\infty}(\mathcal{D})$. We say that $\bar{q} \in AC_+(\mathcal{D})$ if the following conditions hold:

i) $\bar{q}$ is absolutely continuous on $\mathcal{D}$ in the sense that $\forall\ \epsilon > 0$, $\exists\ \gamma > 0$ such that $\forall\ (t,y)$ and $(t',y') \in \mathcal{D}$, $|\bar{q}(t,y) - \bar{q}(t',y')| < \epsilon$ if both $|t - t'|$ and

$|y - y'| < \gamma$. Besides,

$$\int_0^T \int_0^\infty \left| \frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right| dy dt < \infty,$$

ii) $\partial^2 \bar{q}(t,y)/(\partial t \partial y) = 0$ almost everywhere for $(t,y) \in \{(t,y) : 0 \le y \le t \le T\}$.

For $\bar{q} \in AC_+(\mathcal{D})$, we define $I(\bar{q})$ via the following expression

$$\sup_{\theta(\cdot,\cdot) \in C_b(\mathcal{D})} \int_0^T \left[\int_t^\infty \theta(t, y-t) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy - \psi_N \left(\log \left(\int_0^\infty e^{\theta(t,y)} dF(y) \right) \right) \right] dt \tag{4}$$

where $C_b(\mathcal{D})$ is the set of all bounded continuous functions on $\mathcal{D}$. On the other hand, if $\bar{q} \in L_{+,\infty}(\mathcal{D})$ fails to satisfy any of the conditions i) to ii), simply let $I(\bar{q}) = \infty$.

We now can state our main result.

THEOREM 1. *Under the set of assumptions discussed in Section 2.1, $(\bar{Q}_\lambda/\lambda : \lambda > 0)$ satisfies a large deviations principle with good rate function $I(\cdot)$ on the space $(L_{+,\infty}(\mathcal{D}), ||\cdot||_{\mathcal{D}})$. In precise terms, for each open set O we have that*

$$\underline{\lim}_{\lambda \to \infty} \log \frac{1}{\lambda} P(\bar{Q}_\lambda/\lambda \in O) \ge - \inf_{q \in O} I(q),$$

and for each closed set C

$$\overline{\lim}_{\lambda \to \infty} \log \frac{1}{\lambda} P(\bar{Q}_\lambda/\lambda \in C) \le - \inf_{q \in C} I(q).$$

As mentioned earlier, an immediate corollary that we can obtain is a large deviations principle for $(Q_\lambda/\lambda : \lambda > 0)$ under the Skorokhod topology in the space $D_{D_{\mathbb{R}}[0,\infty)}[0,T]$, discussed in the previous section and introduced in Pang and Whitt (2010).

We shall explain the strategy behind the proof of Theorem 1. First, we shall introduce an auxiliary continuous process $\tilde{Q}_\lambda/\lambda$, defined in Section 2.3, that is exponentially equivalent to $\bar{Q}_\lambda/\lambda$. Then, the proof strategy composes of two parts. First, in addition to the assumptions imposed in Section 2.1 we will assume that there exists a deterministic constant $K \in (0,\infty)$ such that $P(V_n \in [0,K]) = 1$. In the second part of the argument we will relax this truncation assumption.

In turn, the first part of the argument (i.e. assuming truncation) is divided into several steps. The first step consists in developing the large deviations

principle for $\tilde{Q}_\lambda/\lambda$ with rate $I(\cdot)$ under the topology of pointwise convergence using the Dawson-Gartner projective limit theorem. The second step involves showing that $\tilde{Q}_\lambda/\lambda$ is exponentially tight as $\lambda \to \infty$ under the uniform topology on the compact set $[0,T] \times [0,K]$. The third and last step involves lifting the large deviations principle to the uniform topology.

In the second part of our argument we introduce an approximation scheme that proceeds by ignoring the customers that arrive to the system with a service time larger than K. Using a coupling argument, the process that is obtained using this scheme is shown to be a good approximation to the original system for the purpose of computing large deviations probabilities.

However, before we do this let us provide a heuristic argument in order to guess the form of the rate function. Later we will explain what are the technical difficulties that need to be addressed.

2.2.1. *Guessing the rate function: A heuristic approach.* One can take advantage of the point process representation of the input process (i.e. the arrivals and the service times represented as a marked point process). Let us start with the case of Poisson arrivals. We shall briefly explain how to adapt the development that follows to the more general case of renewal arrivals.

Consider the scaled system with arrival rate λ and suppose that $F(\cdot)$ has a density $f(\cdot)$. The amount of customers that arrive during the time interval $[t, t+dt]$ and that bring a service requirement of size $[r, r+dr]$ is denoted by the quantity $\mathcal{M}_\lambda(t+dt, r+dr)$, which is governed by a Poisson distribution with rate $\lambda f(r)dtdr$. It follows then by elementary considerations involving the Poisson distribution that for a fixed value of (t,r), $\mathcal{M}_\lambda(t+dt, r+dr)/\lambda$ satisfies a large deviations principle as $\lambda \to \infty$. In particular, we formally obtain that

$$P(\mathcal{M}_\lambda(t+dt, r+dr)/\lambda \approx \mu(t,r)\,dtdr) = \exp(-\lambda J(\mu(t,r))\,dtdr),$$

with

$$J(\mu(t,r)) = \sup_{\eta(t,r)} [\eta(t,r)\,\mu(t,r) - \psi_N(\eta(t,r))\,f(r)],$$

and $\psi_N(\eta) = \exp(\eta) - 1$. The supremum above is obtained formally with $\eta_*(t,r) := \log(\mu(t,r)/f(r))$.

So, by pasting independent regions of the form $[t, t+dt]\times[r, r+dr]$ together one expects that the Poisson random measure $\mathcal{M}_\lambda(\cdot)/\lambda$ would satisfy a large deviations principle under a suitable topology, so that

$$P\Big(\mathcal{M}_\lambda(A\times B)/\lambda \approx \int_{A\times B} \mu(t,r)dtdr, \text{ for a large class } A\times B\Big) \approx \exp(-\lambda \mathbf{J}(\mu))$$

with

$$\begin{aligned}\mathbf{J}(\mu) &= \int_{[0,T]\times[0,\infty)} [\eta_*(t,r)\,\mu(t,r) - \psi_N(\eta_*(t,r))\,f(r)]dtdr \\ &= \sup_{\eta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_{\mathcal{D}} [\eta(t,r)\,\mu(t,r) - \psi_N(\eta(t,r))\,f(r)]dtdr.\end{aligned} \tag{5}$$

Now, observe that for all $y \geq t$

$$\bar{Q}_\lambda(t,y) = \int_0^t \int_{y-s}^\infty \mathcal{M}_\lambda(s+ds, r+dr), \tag{6}$$

and $\bar{Q}_\lambda(0,y) = 0$ for $y \geq 0$. If $\bar{q}(\cdot,\cdot)$ can also be expressed as

$$\bar{q}(t,y) = \int_0^t \int_{y-s}^\infty \mu(s,r)\,drds \tag{7}$$

with

$$\mu(s,r) = -\left.\frac{\partial^2}{\partial y \partial t}\bar{q}(t,y)\right|_{t=s,y=s+r},$$

we can develop the large deviations results for $\bar{q}(\cdot,\cdot)$ based on a similar idea as the contraction principle. In fact, for $\bar{q}(\cdot,\cdot)$ that is absolutely continuous and $\bar{q}(0,y) = 0$ for all $y \geq 0$, the representation (7) is applicable. Therefore, one can formally compute the rate function of $\bar{Q}_\lambda(\cdot,\cdot)/\lambda$ evaluated at $\bar{q}(\cdot,\cdot)$ by evaluating $\mathbf{J}(\mu)$ for $s \in [0,T]$ and $r \in [0,\infty)$. In particular, this analysis yields that $I(\bar{q})$ is equal to

$$\begin{aligned}&\sup_{\eta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_{\mathcal{D}} [\eta(s,r)\,\mu(s,r) - \psi_N(\eta(s,r))\,f(r)]drds \\ &= \sup_{\eta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_{\mathcal{D}} \left[\eta(s,r)\left(-\frac{\partial^2}{\partial y\partial t}\bar{q}(s,s+r)\right) - \psi_N(\eta(s,r))\,f(r)\right] drds \\ &= \sup_{\eta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_{\mathcal{D}} \Bigg[\eta(s,u-s)\left(-\frac{\partial^2}{\partial y\partial t}\bar{q}(s,u)\right) \\ &\qquad\qquad - (\exp(\eta(s,u-s)) - 1) f(u-s)\Bigg] duds,\end{aligned}$$

which is, of course, equivalent to (4) in the Poisson case assuming the existence of a density $f(\cdot)$ for the distribution of the service times. The previous form of the rate function was heuristically obtained assuming that $y \geq t$. However, since all the information of the infinite server queue is contained in the evolution of the process $(Q_\lambda(t,u) : (t,u) \in \mathcal{D})$ defined in Section 2.1 with $Q_\lambda(t,u) = \bar{Q}_\lambda(t,t+u)$, we must have that the rate function should be specified only over $\bar{q}(t,y)$ for $y \geq t$. Indeed, one can check that

$\partial^2\bar{q}(t,y)/(\partial y\partial t) = 0$ for $0 \le y < t \le T$ as $\bar{Q}_\lambda(t+\Delta t, y+\Delta y) - \bar{Q}_\lambda(t+\Delta t, y) - \bar{Q}(t, y+\Delta y) + \bar{Q}(t,y) = 0$ for all $y < t$.

For the non-Poisson case one can argue using renewal arguments. We need to compute the log-moment generating function of the vertical strip $(\mathcal{M}_\lambda(t+dt, r_i + dr) : 1 \le i \le n)$, where $r_1 < r_2 < \cdots < r_n$ for an arbitrary partition $(r_i : 1 \le i \le n)$. We obtain, using elementary properties of the multinomial distribution together with an application of the key renewal theorem as in Glynn (1995),

$$
\begin{aligned}
&E\left[\exp\left(\sum_{i=1}^{n}\theta\left(t,r_i\right)\mathcal{M}_\lambda\left(t+dt,r_i+dr\right)\right)\right]\\
&\quad= E\left[\left(\sum_{i=1}^{n}\exp\left(\theta\left(t,r_i\right)\right)P\left(V_1\in[r_i,r_i+dr]\right)\right)^{N(\lambda(t+dt))-N(\lambda t)}\right]\\
&\quad= E\left[\exp\left(\left[N\left(\lambda(t+dt)\right)-N\left(\lambda t\right)\right]\right.\right.\\
&\qquad\left.\left.\cdot\log\left(\sum_{i=1}^{n}\exp\left(\theta\left(t,r_i\right)\right)P\left(V_1\in[r_i,r_i+dr]\right)\right)\right)\right]\\
&\quad= \exp\left(\lambda\psi_N\left(\log\left(\sum_{i=1}^{n}\exp\left(\theta\left(t,r_i\right)\right)P\left(V_1\in[r_i,r_i+dr]\right)\right)\right)+o\left(\lambda\right)\right)
\end{aligned}
$$

as $\lambda\to\infty$.

So, by pasting together vertical strips (i.e. ranging the parameter t) we obtain that the family of random measures $\mathcal{M}_\lambda(\cdot)/\lambda$ is expected to satisfy a large deviations principle under a suitable topology with rate function

$$
\begin{aligned}
\mathbf{J}\left(\mu\right) = \sup_{\theta(\cdot,\cdot)\in C_b(\mathcal{D})}\int_{[0,T]}\left[\int_0^\infty\theta\left(t,r\right)\mu\left(t,r\right)dr\right.\\
\left.-\psi_N\left(\log\left(\int_0^\infty\exp\left(\theta\left(t,r\right)\right)dF\left(r\right)\right)\right)\right]dt.
\end{aligned}
$$

The rest of the formal analysis proceeds similarly as in the Poisson case.

The formal argument just outlined, even if heuristic, suggests a potential approach to developing sample path large deviations for $\bar{Q}_\lambda/\lambda$. Namely, first develop a large deviations for the random measures $\mathcal{M}_\lambda(\cdot)/\lambda$, and then apply the contraction principle to obtain the desired large deviations result for $\bar{Q}_\lambda/\lambda$. This approach, although intuitive, will not be followed in our development. We found it easier to directly work with the topology that we wish to impose. Part of the problem involved in making the argument based on random measures rigorous in the setting of the topology that is of interest to

us is that indicator functions are not continuous, so the contraction principle is not directly applicable if one is to endow the space of measures with the weak convergence topology. Of course, one can proceed by trying a different topology (stronger than weak convergence) or by trying to use the extended contraction principle. However, the technical development, we believe, would end up being more involved than the direct approach that we will follow.

An additional concern that might arise at this point is our selection of $\bar{Q}_\lambda/\lambda$ in order to represent the system status; as opposed to Q_λ/λ, which might appear more natural at first sight. Let us explain why $\bar{Q}_\lambda/\lambda$ is a more convenient object to consider. Note that if $q(s,r) = \bar{q}(s, s+r)$, then

$$\begin{aligned}\frac{\partial^2}{\partial s \partial r} q\left(s,r\right) &= \frac{\partial^2}{\partial s \partial r}\bar{q}(s,s+r) + \frac{\partial^2}{\partial r^2}\bar{q}(s,s+r) \\ &= \frac{\partial^2}{\partial s \partial r}\bar{q}(s,s+r) + \frac{\partial^2}{\partial r^2} q\left(s,r\right)\end{aligned}$$

and therefore

$$\frac{\partial^2}{\partial s \partial r}\bar{q}(s,s+r) = \frac{\partial^2}{\partial s \partial r} q\left(s,r\right) - \frac{\partial^2}{\partial r^2} q\left(s,r\right).$$

Since $Q_\lambda(t,u)/\lambda = \bar{Q}_\lambda(t,u+t)/\lambda$ and our heuristic analysis suggests that the candidate rate function of $\bar{Q}_\lambda(t,y)/\lambda$ is given by

$$\sup_{\theta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_0^T \left[\int_t^\infty \theta(t,y-t)\left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right) dy \right. \\ \left. - \psi_N\left(\log\left(\int_0^\infty e^{\theta(t,y)} dF\left(y\right)\right)\right)\right] dt,$$

it is then sensible to conjecture, making $y = u + t$, a representation based on $q(t,u) = \bar{q}(t,t+u)$ via

$$\begin{aligned}&\sup_{\theta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_0^T \left[\int_0^\infty \theta(t,u)\left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,u+t)\right) du \right. \\ &\qquad \left. - \psi_N\left(\log\left(\int_0^\infty e^{\theta(t,u)} dF\left(u\right)\right)\right)\right] dt \\ &= \sup_{\theta(\cdot,\cdot)\in C_b(\mathcal{D})} \int_0^T \left[\int_0^\infty \theta(t,u)\left(-\frac{\partial^2}{\partial t \partial u} q\left(t,u\right) + \frac{\partial^2}{\partial u^2} q\left(t,u\right)\right) du \right. \\ &\qquad \left. - \psi_N\left(\log\left(\int_0^\infty e^{\theta(t,u)} dF\left(u\right)\right)\right)\right] dt.\end{aligned} \tag{8}$$

This representation, in turn, suggests that for the rate function to be finite at $q(\cdot)$, one might need to impose as a necessary condition the existence of $\partial^2 q(t,u)/\partial^2 u$. Nevertheless, as we shall see in our examples, one might have a finite-valued rate function even in cases in which $\partial q(t,\cdot)/\partial u$ is not even continuous for every value of $t \in (0,T)$.

2.3. *An auxiliary continuous process.* In order to prove Theorem 1 we introduce an auxiliary approximating continuous process, $\tilde{Q}_\lambda$, which shall be shown to be exponentially equivalent to the process of interest $\bar{Q}_\lambda$ in the uniform norm. The construction of $\tilde{Q}_\lambda$ which is based on simple polygonal interpolations is explicitly given in the Appendix. First, we show that one can construct a continuous process $(Q^*_\lambda(t,y) : t \in [0,T], y \geq 0)$ such that $Q^*_\lambda(t,\cdot)$ is non-increasing for each $t \in [0,T]$ and satisfying $\|Q^*_\lambda - Q_\lambda\| \leq 2$. Then, we define our auxiliary process $\tilde{Q}_\lambda(t,y)$ for $y \geq t$ via

$$\tilde{Q}_\lambda(t,y) = Q^*_\lambda(t, y-t), \tag{9}$$

which is the analogue of (3). Finally, we define $\tilde{Q}_\lambda(t,y)$ for $0 \leq y \leq t \leq T$ as follows. First let $\tilde{N}_\lambda(\cdot)$ be the continuous process obtained by the polygonal interpolation of $N_\lambda(\cdot)$, so that $\tilde{N}_\lambda(0) = 0$ and $\tilde{N}_\lambda(A_k/\lambda) = N_\lambda(A_k/\lambda)$ for all $k \geq 1$. Then, for $y < t$ define

$$\tilde{Q}_\lambda(t,y) = \tilde{Q}_\lambda(y,y) + \tilde{N}_\lambda(t) - \tilde{N}_\lambda(y), \tag{10}$$

analogous to (2). Observe that $\|\tilde{N}_\lambda - N_\lambda\| \leq 1$. It follows from the triangle inequality and expressions (3), (9) and (10) that

$$||\tilde{Q}_\lambda - \bar{Q}_\lambda||_{\mathcal{D}} \leq 4, \tag{11}$$

where $||\cdot||_{\mathcal{D}}$ represents the uniform norm over the set $\mathcal{D}$.

3. Bounded service times. In addition to the assumptions imposed in Section 2 here we also assume that $P(V_n \in [0,K]) = 1$ for $K \in (0,\infty)$.

We define $\mathcal{D}_K = \{(t,u) : 0 \leq t \leq T, 0 \leq u \leq K+T\}$ and let $C_+(\mathcal{D}_K)$ be the space of functions $(x(t,u) : (t,u) \in \mathcal{D}_K)$ such that $x(\cdot)$ is continuous in both components, $x(t,\cdot)$ is non-increasing on $[0,K+T]$, and $x(0,u) = 0$ for $u \geq 0$. Following the same notation in Section 2.2 we say that $x(\cdot,\cdot) \in AC_+(\mathcal{D}_K)$ if $x(\cdot,\cdot)$ is absolutely continuous, and $\partial^2 x(t,y)/(\partial t \partial y) = 0$ almost everywhere on $0 \leq y \leq t \leq T$.

Our initial goal is to obtain a large deviations principle for $(\tilde{Q}_\lambda/\lambda : \lambda > 0)$ as $\lambda \to \infty$ on the space $(C_+(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K})$; we then will use (11) to obtain the corresponding large deviations principle for $(\bar{Q}_\lambda/\lambda : \lambda > \infty)$.

We start by deriving a large deviations principle in the topology of pointwise convergence. The proof of this result will be given at the end of this section.

LEMMA 1. *Let $\mathcal{X}$ consist of all the maps from $\mathcal{D}_K$ to $\mathbb{R}$, and we equip $\mathcal{X}$ with the topology of pointwise convergence on $\mathcal{D}_K$. Then $\tilde{Q}_\lambda/\lambda$ satisfies a large deviations principle with good rate function $I(\bar{q})$ defined by*

$$\sup_{\theta(\cdot,\cdot)\in C[0,T]\times[0,K]} \int_0^T \left[\int_t^{K+t} \theta(t, y-t) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy - \psi_N \left(\log \left(\int_0^K e^{\theta(t,y)} dF(y) \right) \right) \right] dt \tag{12}$$

if $\bar{q}(\cdot) \in AC_+(\mathcal{D}_K)$, and $I(\bar{q}) = \infty$ otherwise. Here $C[0,T]\times[0,K]$ denotes the set of all continuous functions on $[0,T]\times[0,K]$.

In order to lift the large deviations principle indicated in Lemma 1 to the uniform topology we need the following result on exponential tightness; we shall also give the proof of this result at the end of this section.

LEMMA 2. *$\tilde{Q}_\lambda/\lambda$ is exponentially tight in $C_+(\mathcal{D}_K)$ equipped with the topology of uniform convergence.*

Using the previous two lemmas we are ready to state and prove the main result of this section, which is a version of Theorem 1 for the case of bounded service times.

THEOREM 2. *$\tilde{Q}_\lambda/\lambda$ satisfies a large deviations principle with good rate function defined in (12) under the uniform topology on $\mathcal{D}_K$.*

PROOF. Since the domain of $I(\cdot)$ is a subset of $C_+(\mathcal{D}_K)$, and $\tilde{Q}_\lambda/\lambda \in C_+(\mathcal{D}_K)$ with probability 1, the large deviations principle in Lemma 1 holds in the space $C_+(\mathcal{D}_K)$ with pointwise topology, (Lemma 4.1.5 (b) in Dembo and Zeitouni (1998)). Since by Lemma 2 $\tilde{Q}_\lambda/\lambda$ is exponentially tight in $(C_+(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K})$ the same large deviations principle holds in $(C_+(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K})$ (Corollary 4.2.6 in Dembo and Zeitouni (1998)) and the result follows. □

As a corollary of the previous theorem we obtain that $(\bar{Q}_\lambda/\lambda : \lambda > 0)$ satisfies a large deviations principle on $(L_{+,\infty}(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K})$.

COROLLARY 1. *The process $(\bar{Q}_\lambda/\lambda : \lambda > 0)$ satisfies a large deviations principle on $(L_{+,\infty}(\mathcal{D}_K), ||\cdot||_{\mathcal{D}_K})$ with rate function $I(\cdot)$ defined in (12).*

PROOF. First we verify that $\bar{Q}_\lambda/\lambda$ and $\tilde{Q}_\lambda/\lambda$ are exponentially equivalent according to Definition 4.2.10 in Dembo and Zeitouni (1998)). Since the laws of $(\bar{Q}_\lambda/\lambda, \tilde{Q}_\lambda/\lambda)$ are induced by a separable stochastic process and the underlying topology is induced by the uniform norm, the set

$$\{\omega : ||\bar{Q}_\lambda/\lambda - \tilde{Q}_\lambda/\lambda||_{\mathcal{D}_K} > \eta\}$$

is Borel measurable (see Remark b) following Definition 4.2.10 in Dembo and Zeitouni (1998)). Now recall that by the construction of $\tilde{Q}_\lambda$ that $\|\bar{Q}_\lambda - \tilde{Q}_\lambda\| \leq 4$ a.s. Hence for any $\eta > 0$,

$$P(||\bar{Q}_\lambda/\lambda - \tilde{Q}_\lambda/\lambda||_{\mathcal{D}_K} > \eta) = 0$$

for large enough λ. Hence

$$\limsup_{\lambda\to\infty} \frac{1}{\lambda} \log P(||\bar{Q}_\lambda/\lambda - \tilde{Q}_\lambda/\lambda||_{\mathcal{D}_K} > \eta) = -\infty.$$

The result then follows by applying Theorem 4.2.13 in Dembo and Zeitouni (1998). □

3.1. *Proofs of technical results.* Finally, we provide the proofs of Lemmas 1 and 2.

We start with Lemma 1 which takes advantage of the Dawson-Gartner projective limit theorem and thus requires that we obtain an auxiliary large deviations principle for finite dimensional objects defined via

$$\begin{aligned}\Delta_{ij}(\lambda) &= \sum_{k=N_\lambda(t_{i-1})+1}^{N_\lambda(t_i)} I(y_{j-1} < A_k/\lambda + V_k \leq y_j) \\ &= \bar{Q}_\lambda(t_i, y_{j-1}) - \bar{Q}_\lambda(t_i, y_j) - \bar{Q}_\lambda(t_{i-1}, y_{j-1}) + \bar{Q}_\lambda(t_{i-1}, y_j),\end{aligned} \tag{13}$$

for $t_{i-1} < t_i$, and $y_{j-1} < y_j$.

LEMMA 3. *For $0 = t_0 < t_1 < t_2 < \cdots < t_m \leq T$ and $0 = y_0 < y_1 < \cdots < y_n < y_{n+1} = T + K$, $(\Delta_{ij}(\lambda)/\lambda : 1 \leq i \leq m, 1 \leq j \leq n+1)$ possesses a large deviations principle with a good rate function*

$$\sup_{\theta_{ij}:1\leq i\leq m, 1\leq j\leq n+1} \sum_{i=1}^{m}\sum_{j=1}^{n+1} \theta_{ij}\delta_{ij} - \sum_{i=1}^{m} \int_{t_{i-1}}^{t_i} \psi_N \left(\log \sum_{j=1}^{n+1} e^{\theta_{ij}} P(y_{j-1} - u < V_1 \leq y_j - u) \right) du.$$

PROOF OF LEMMA 3. We use that $\psi_N(\cdot)$ is continuously differentiable over $\mathbb{R}$. Since U_i are non-lattice, the key renewal theorem implies that for any set of $0 \le t_0 < t_1 < t_2 < \cdots < t_m$,

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} \log E \exp\left\{\sum_{i=1}^{m} \theta_i (N_\lambda(t_i) - N_\lambda(t_{i-1}))\right\} = \sum_{i=1}^{m} \psi_N(\theta_i)(t_i - t_{i-1})$$

for any $\theta_i \in \mathbb{R}$; see Glynn and Whitt (1994) p. 115 and Glynn (1995) p. 390, and also (1). Then, from Glynn (1995), the Gartner-Ellis limit $\Lambda(\Theta)$ of $(\Delta_{ij}(\lambda) : 1 \le i \le m, 1 \le j \le n+1)$ equals

$$\begin{aligned}\Lambda(\Theta) &\triangleq \lim_{\lambda\to\infty} \frac{1}{\lambda} \log E \exp\left\{\sum_{i=1}^{m}\sum_{j=1}^{n+1} \theta_{ij}\Delta_{ij}(\lambda)\right\} \\ &= \sum_{i=1}^{m} \int_{t_{i-1}}^{t_i} \psi_N\left(\log \sum_{j=1}^{n+1} e^{\theta_{ij}} P(y_{j-1} - u < V_1 \le y_j - u)\right) du\end{aligned}$$

which is finite for any $\Theta := (\theta_{i,j} : 1 \le i \le m, 1 \le j \le n+1)$. Moreover, for any $t_{i-1} < u \le t_i$,

$$\begin{aligned}&\left|\frac{\partial}{\partial\theta_{ij}} \psi_N\left(\log \sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u)\right)\right| \\ &= \left|\psi_N'\left(\log \sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u)\right)\right| \\ &\quad \cdot \frac{e^{\theta_{ij}} P(y_{j-1} - u < V_1 \le y_j - u)}{\sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u)} \\ &\le \max\{|\psi_N'(\max\{\theta_{ik}, k = 1, \ldots, n+1\})|, |\psi_N'(\min\{\theta_{ik}, k = 1, \ldots, n+1\})|\}\end{aligned}$$

which is uniformly bounded over a neighborhood of θ_{ij} and $t_{i-1} < u \le t_i$, fixing all other θ_{lk}'s. Therefore,

$$\begin{aligned}&\frac{1}{h}\left|\psi_N\left(\log\left(e^{\theta_{ij}+h} P(y_{j-1} - u < V_1 \le y_j - u)\right.\right.\right. \\ &\qquad \left.\left.\left. + \sum_{k\ne j} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u)\right)\right)\right. \\ &\qquad \left. - \psi_N\left(\log \sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u)\right)\right|\end{aligned}$$

is also uniformly bounded on the same region. By dominated convergence theorem, we have

$$\frac{\partial}{\partial\theta_{ij}}\Lambda(\Theta) = \int_{t_{i-1}}^{t_i} \psi_N' \left(\log \sum_{k=1}^{n+1} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u) \right) \cdot \frac{e^{\theta_{ij}} P(y_{j-1} - u < V_1 \le y_j - u)}{\sum_{k=1}^{n_i} e^{\theta_{ik}} P(y_{k-1} - u < V_1 \le y_k - u)} du.$$

Moreover, it is dominated by

$$(t_i - t_{i-1}) \max\{|\psi_N'(\max\{\theta_{ik}, k = 1, \ldots, n+1\})|, |\psi_N'(\min\{\theta_{ik}, k = 1, \ldots, n+1\})|\} < \infty$$

for any given $\Theta \in \mathbb{R}^{m\times(n+1)}$. Since $\Lambda(\cdot)$ is finite and differentiable everywhere on $\mathbb{R}^{m\times(n+1)}$, by the Gartner-Ellis Theorem for the case $\mathcal{D}_\Lambda = \mathbb{R}^{m\times(n+1)}$ (Dembo and Zeitouni (1998), p. 52, Ex 2.3.20 (g)), $\{\Delta_{ij}(\lambda)\}$ possesses a rate function

$$\sup_{\theta_{ij}:1\le i\le m, 1\le j\le n+1} \sum_{i=1}^{m}\sum_{j=1}^{n+1} \theta_{ij}\delta_{ij} - \sum_{i=1}^{m} \int_{t_{i-1}}^{t_i} \psi_N \left(\log \sum_{j=1}^{n+1} e^{\theta_{ij}} P(y_{j-1} - u < V_1 \le y_j - u) \right) du.$$

We argue that the rate function is good. By Dembo and Zeitouni (1998) p. 8, Lemma 1.2.18, it suffices to show that $(\Delta_{ij}(\lambda) : 1 \le i \le m, 1 \le j \le n+1)$ is exponentially tight. Denoting $\|\cdot\|_1$ as the L_1-norm, we have by Chernoff's bound

$$\begin{aligned}\overline{\lim}_{\lambda\to\infty} \frac{1}{\lambda} \log P\left(\|\Delta_{ij}(\lambda)/\lambda\|_1 > \alpha\right) &\le \overline{\lim}_{\lambda\to\infty} \frac{1}{\lambda} \log P(N_\lambda(T) > \alpha\lambda) \\ &\le -\theta\alpha + \psi_N(\theta),\end{aligned}$$

for any $\theta > 0$. Sending $\alpha \to \infty$ we then obtain

$$\overline{\lim}_{\alpha\to\infty} \overline{\lim}_{\lambda\to\infty} \frac{1}{\lambda} \log P\left(\|\Delta_{ij}(\lambda)/\lambda\|_1 > \alpha\right) = -\infty,$$

thereby obtaining exponential tightness and the goodness of the underlying rate function as claimed. □

Proof of Lemma 1. We will use the Dawson-Gartner projective limit theorem. Consider a collection of points in the plane of the form $\kappa = ((t_i, y_j) : 1 \le i \le m, 0 \le j \le n)$, such that $0 := t_0 < t_1 < t_2 < \cdots < t_m \le T$ and $0 :=$

$y_0 < y_1 < \cdots < y_n$. Moreover, we assume that $y_l = t_l$ if $0 \le l \le \min(m,n)$. Let $\mathcal{K}$ be the union of such collection of sets κ. Further, let $\{p_\kappa\}_{\kappa\in\mathcal{K}}$ be the projective system generated by $\mathcal{K}$. We will proceed to obtain a large deviations principle for the projections $(\bar{Q}_\lambda(t,y)/\lambda : (t,y) \in \kappa)$. However, we will do this by first obtaining a large deviations principle for quantities $\Delta_{ij}(\lambda)/\lambda$ and then the large deviations principle for the projections follows using the contraction principle as the $(\bar{Q}_\lambda(t,y)/\lambda : (t,y) \in \kappa)$ will be shown to be continuous functions. Set $y_{n+1} = \infty$, so that $\bar{Q}_\lambda(t, y_{n+1}) = 0$ for every $t \in [0,T]$. It is important to note, given the structure of the partition κ, that if $1 \le i \le m$, $1 \le j \le n$, and $i > j$, then $\Delta_{ij}(\lambda) = 0$. Now, similar to the definition of $\Delta_{ij}(\lambda)$ we define, for $1 \le i \le m$ and $1 \le j \le n+1$,

$$\widetilde{\Delta}_{ij}(\lambda) = \tilde{Q}_\lambda(t_i, y_{j-1}) - \tilde{Q}_\lambda(t_i, y_j) - \tilde{Q}_\lambda(t_{i-1}, y_{j-1}) + \tilde{Q}_\lambda(t_{i-1}, y_j). \tag{14}$$

Once again, observe that $\tilde{Q}_\lambda(t, y_{n+1}) = 0$, and also if $i > j$, for $1 \le i \le m$, $1 \le j \le n$, we have $t_{i-1} \ge y_j$ and therefore

$$\begin{aligned}\widetilde{\Delta}_{ij}(\lambda) &= \tilde{Q}_\lambda(y_{j-1}, y_{j-1}) + \tilde{N}_\lambda(t_i) - \tilde{N}_\lambda(y_{j-1}) \\ &\quad - (\tilde{Q}_\lambda(y_j, y_j) + \tilde{N}_\lambda(t_i) - \tilde{N}_\lambda(y_j)) \\ &\quad - (\tilde{Q}_\lambda(y_{j-1}, y_{j-1}) + \tilde{N}_\lambda(t_{i-1}) - \tilde{N}_\lambda(y_{j-1})) \\ &\quad + (\tilde{Q}_\lambda(y_j, y_j) + \tilde{N}_\lambda(t_{i-1}) - \tilde{N}_\lambda(y_j)) \\ &= 0.\end{aligned}$$

Moreover, clearly we have for $1 \le i \le m$, and $1 \le j \le n$

$$\tilde{Q}_\lambda(t_i, y_j) = \sum_{l=1}^{i} \sum_{r=j+1}^{n+1} \widetilde{\Delta}_{lr}(\lambda),$$

so indeed we have that $(\tilde{Q}_\lambda(t_i, y_j) : 1 \le i \le m, 1 \le j \le n+1)$ can be recovered as a continuous function of the $\widetilde{\Delta}_{lr}(\lambda)$'s. Since $\|\tilde{Q} - \bar{Q}\| \le 4$ by (11), it follows by the triangle inequality and from (13) and (14) that $|\widetilde{\Delta}_{ij}(\lambda) - \Delta_{ij}(\lambda)| \le 16$ and hence we have

$$\lim_{\lambda\to\infty} \frac{1}{\lambda} \log E \exp\left\{ \sum_{i=1}^{m} \sum_{j=i}^{n+1} \theta_{ij} \Delta_{ij}(\lambda) \right\} = \lim_{\lambda\to\infty} \frac{1}{\lambda} \log E \exp\left\{ \sum_{i=1}^{m} \sum_{j=i}^{n+1} \theta_{ij} \widetilde{\Delta}_{ij}(\lambda) \right\}.$$

Consequently, from Lemma 3, the rate function for the projections represented by κ (these projections are denoted by $p_\kappa(\bar{q})$) can be written as

$$I(p_\kappa(\bar{q})) = \sup_{\{\theta_{ij} : 1 \le \imath \le m, 1 \le j \le n+1\}} \sum_{i=1}^{m} \sum_{j=1}^{n} \theta_{ij} \delta_{ij}(\kappa)$$

$$-\sum_{i=1}^{m}\int_{t_{i-1}}^{t_i}\psi_N\left(\log\sum_{j=i}^{n+1}e^{\theta_{ij}}P(y_{j-1}-u<V_1\le y_j-u)\right)du.$$

To possess a finite $I(p_\kappa(\bar{q}))$, the quantity $\delta_{ij}(\kappa):=\bar{q}(t_i,y_{j-1})-\bar{q}(t_i,y_j)-\bar{q}(t_{i-1},y_{j-1})+\bar{q}(t_{i-1},y_j)$ must satisfy that

$$\delta_{ij}(\kappa)=0 \tag{15}$$

for $i>j$, and $1\le i\le m$, $1\le j\le n+1$; otherwise, if $\delta_{ij}(\kappa)\neq 0$, the rate function can be made arbitrarily large by picking $\theta_{ij}=c\times sgn(\delta_{ij}(\kappa))$ with arbitrarily large constant $c>0$ for $1\le j<i\le m$, as

$$\int_{t_{i-1}}^{t_i}\psi_N\left(\log\sum_{j=i}^{n+1}e^{\theta_{ij}}P(y_{j-1}-u<V_1\le y_j-u)\right)du$$

is independent of θ_{ij}'s that have $j<i$. In the representation of the rate function $I(p_\kappa(\bar{q}))$ we have also used the fact that $\bar{q}(t_i,y_j)=\sum_{l\le i,r>j}\delta_{lr}(\kappa)$, with $\bar{q}(0,y_j)=0$, so the relation from the $\delta_{ij}(\kappa)$'s to the $\bar{q}(t_i,y_j)$ is a one-to-one, continuous function, so that the contraction principle (Theorem 4.2.1, Dembo and Zeitouni (1998)) is invoked for the above representation for $I(p_\kappa(\bar{q}))$. We want to show that $\sup_{\kappa\in\mathcal{K}}I(p_\kappa(\bar{q}))$ is equal to (12), and hence conclude the proof by Dawson-Gartner Theorem (see Theorem 4.6.1, Dembo and Zeitouni (1998)). Clearly it suffices to concentrate on functions $\bar{q}$ such that $\bar{q}(t,y)=0$ whenever $t>T$ or $y>t+K$ given that we are assuming service times bounded by K. Note that the constraint (15) implies that for any $\bar{q}$, in order that $I(\bar{q})<\infty$, we must have absolute continuity throughout $0\le y\le t\le T$ and, moreover, that

$$\partial^2\bar{q}(t,y)/(\partial y\partial t)=0$$

almost everywhere on $0\le y\le t\le T$ (see Dembo and Zeitouni (1998) p. 189). We now focus on $\bar{q}(t,y)$ that is absolutely continuous on $\mathcal{D}_K$ and has $\partial^2\bar{q}(t,y)/(\partial y\partial t)=0$ almost everywhere on $0\le y\le t\le T$. Observe that

$$\begin{aligned}\delta_{ij}(\kappa)&=\bar{q}(t_i,y_{j-1})-\bar{q}(t_i,y_j)-\bar{q}(t_{i-1},y_{j-1})+\bar{q}(t_{i-1},y_j)\\&=-\int_{t_{i-1}}^{t_i}\int_{y_{j-1}}^{y_j}\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)dydt.\end{aligned}$$

Regarding $\theta(\cdot,\cdot)$ as a step function with jumps at $0=t_1<t_2<\cdots<t_m\le T$ and $0\le y_0<y_1<\cdots<y_n<y_{n+1}=T+K$, and denote $S(\mathcal{C})$ as the set of

all step functions on a given domain $\mathcal{C}$. We can write

$$\sup_{\kappa} I(p_\kappa(\bar{q})) = \sup_{\theta(\cdot,\cdot)\in S(\mathcal{D}_K)} \left\{ \int_0^T \int_t^{t+K} \theta(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) dydt \right. \tag{16}$$
$$\left. - \int_0^T \psi_N \left(\log \int_t^{t+K} e^{\theta(t,y)} dF\,(y-t) \right) dt \right\}$$

To show that $\sup_\kappa I(p_\kappa(\bar{q})) \geq I(\bar{q})$ where $I(\bar{q})$ is as defined in (12), note first that the set of step functions $S(\mathcal{D}_K)$ is dense in $C(\mathcal{D}_K)$, the set of continuous functions equipped with the uniform metric. So for any continuous function $\theta(\cdot,\cdot) \in C(\mathcal{D}_K)$, we can find a sequence $\theta_k(\cdot,\cdot) \in S(\mathcal{D}_K)$ with $\|\theta_k - \theta\|_{\mathcal{D}_K} \to 0$. Note that since θ is continuous, it is bounded and so θ_k is also uniformly bounded i.e. $|\theta_k(t,y)| \leq C$ for all k and some $C > 0$. Consider

$$\int_0^T \int_t^{t+K} \theta(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) dydt - \int_0^T \psi_N \left(\log \int_t^{t+K} e^{\theta(t,y)} dF(y-t) \right) dt$$

with $\theta \in C(\mathcal{D}_K)$. We want to show that this can be approximated by the counterpart in $\theta_k \in S(\mathcal{D}_K)$. Note that

$$\int_0^T \int_t^{t+K} \left| \theta_k(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) \right| dydt \leq C \int_0^T \int_t^{t+K} \left| \frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right| dydt < \infty$$

since $\bar{q}$ is absolutely continuous. By dominated convergence we have

$$\int_0^T \int_t^{t+K} \theta_k(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) dydt \tag{17}$$
$$\to \int_0^T \int_t^{t+K} \theta(t,y) \left(-\frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) \right) dydt.$$

Similarly, since, as mentioned earlier $|\theta_k(t,y)| \leq C$, by the bounded convergence theorem we have

$$\int_t^{t+K} e^{\theta_k(t,y)} dF\,(y-t) \to \int_t^{t+K} e^{\theta_k(t,y)} dF\,(y-t)$$

and so by the continuity of $\psi_N(\log(\cdot))$ we get

$$\psi_N \left(\log \int_t^{t+K} e^{\theta_k(t,y)} dF\,(y-t) \right) \to \psi_N \left(\log \int_t^{t+K} e^{\theta(t,y)} dF\,(y-t) \right)$$

for any t. Furthermore, the obvious inequality

$$e^{-C} = e^{-C}\int_0^K dF(y) \le \int_0^K e^{\theta_k(t,y)}dF(y) \le e^C\int_0^K dF(y) = e^C, \tag{18}$$

yields

$$\left|\psi_N\left(\log\int_t^{t+K} e^{\theta_k(t,y)}f(y-t)dy\right)\right| \le \sup_{\xi\in[-C,C]}|\psi_N(\xi)|.$$

Hence yet another application of dominated convergence gives

$$\begin{aligned}&\int_0^T \psi_N\left(\log\int_t^{t+K} e^{\theta_k(t,y)}dF(y-t)\right)dt\\ &\quad\to\int_0^T \psi_N\left(\log\int_t^{t+K} e^{\theta(t,y)}dF(y-t)\right)dt.\end{aligned} \tag{19}$$

Combining (17) and (19) and using the expression in (16), we conclude that $\sup_\kappa I(p_\kappa(\bar q)) \ge I(\bar q)$ (note a shift of variable y in (12)). For the other direction, consider

$$\int_0^T\int_t^{t+K}\theta(t,y)\left(-\frac{\partial^2}{\partial t\partial y}\bar q(t,y)\right)dydt - \int_0^T\psi_N\left(\log\int_t^{t+K}e^{\theta(t,y)}dF(y-t)\right)dt$$

now with $\theta\in S(\mathcal{D}_K)$. Note that we can find a sequence $\theta_k\in C(\mathcal{D}_K)$ such that $\theta_k\to\theta$ pointwise almost everywhere and that θ_k is uniformly bounded; this sequence can be found, for example, by convolving θ with a sequence of mollifiers (i.e. smooth kernels with bandwidth that tends to zero as $k\to\infty$). Exactly the same argument as above would then yield $\sup_\kappa I(p_\kappa(\bar q)) \le I(\bar q)$. Now, let $\bar q\in C_+(\mathcal{D}_K)$ and suppose that $\bar q$ is not absolutely continuous. That is, it is not of bounded total variation in the sense of Dembo and Zeitouni (1998) p. 189. Then, for every $\gamma>0$ there exists $t_1(\gamma)<\cdots<t_m(\gamma)$ and $y_0(\gamma)<\cdots<y_n(\gamma)$ such that $\sum_{i=1}^m\sum_{j=1}^n|\delta_{ij}^\gamma|\ge\gamma$, where

$$\delta_{ij}^\gamma = \bar q(t_i(\gamma),y_{j-1}(\gamma)) - \bar q(t_i(\gamma),y_j(\gamma)) - \bar q(t_{i-1}(\gamma),y_{j-1}(\gamma)) + \bar q(t_{i-1}(\gamma),y_j(\gamma)).$$

Now observe that

$$\begin{aligned}\sup_{\kappa\in\mathcal{K}} I(p_\kappa(\bar q)) = &\sup_{\substack{\theta_{ij}:1\le i\le m,1\le j\le n\\ \kappa\in\mathcal{K}}}\sum_{i=1}^m\sum_{j=1}^n\theta_{ij}\delta_{ij}(\kappa)\\ &-\sum_{i=1}^m\int_{t_{i-1}}^{t_i}\psi_N\left(\log\sum_{j=1}^{n+1}e^{\theta_{ij}}P(y_{j-1}-u<V_1\le y_j-u)\right)du.\end{aligned}$$

Following Dembo and Zeitouni (1998) p. 192, we can select $\theta_{ij} = sgn(\delta_{ij}^{\gamma})$ for the partition introduced earlier that defines δ_{ij}^{γ}, and obtain

$$\sup_{\kappa\in\mathcal{K}} I(p_\kappa(\bar{q})) \geq \sum_{i=1}^{m}\sum_{j=1}^{n}\left|\delta_{ij}^{\gamma}\right| - T\psi_N(1).$$

Since $\gamma > 0$ is arbitrary we conclude that

$$\sup_{\kappa\in\mathcal{K}} I(p_\kappa(\bar{q})) = \infty$$

as required. □

Proof of Lemma 2. We want to prove that for any η,

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(\left\|\frac{\tilde{Q}_\lambda(0,0)}{\lambda}\right\|_{\mathcal{D}_K} > \eta\right) = -\infty$$

and

$$\lim_{\delta\to 0}\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(w\left(\frac{\tilde{Q}_\lambda}{\lambda},\delta\right) > \eta\right) = -\infty$$

where $w(\tilde{Q}_\lambda/\lambda,\delta)$ is the modulus of continuity of $\tilde{Q}_\lambda/\lambda$ with order δ defined by

$$w\left(\tilde{Q}_\lambda/\lambda,\delta\right) = \sup_{\substack{|t_1-t_2|<\delta\\ |y_1-y_2|<\delta}} |\tilde{Q}_\lambda(t_1,y_1)/\lambda - \tilde{Q}_\lambda(t_2,y_2)/\lambda|.$$

Recall that $\|\tilde{Q}_\lambda - \bar{Q}_\lambda\|_{\mathcal{D}_K} \leq 4$ a.s., and that $\bar{Q}_\lambda(t,y) = Q_\lambda(t,y-t)$ for $y > t$ and $\bar{Q}_\lambda(t,y) = \bar{Q}_\lambda(y,y) + N_\lambda(t) - N_\lambda(y)$ for $0 \leq y \leq t \leq T$. Therefore, it suffices to show that for any $\eta > 0$,

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(\left\|\frac{Q_\lambda(0,0)}{\lambda}\right\|_{\mathcal{D}_K} > \eta\right) = -\infty, \tag{20}$$

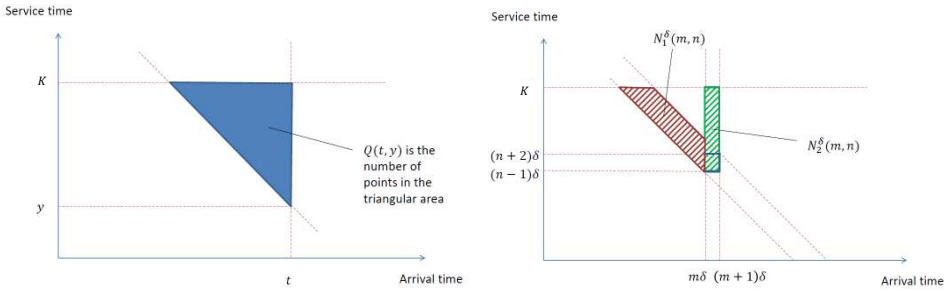
also

$$\lim_{\delta\to 0}\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(w\left(\frac{Q_\lambda}{\lambda},\delta\right) > \eta\right) = -\infty, \tag{21}$$

and finally that

$$\lim_{\delta\to 0}\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(\sup_{0\leq t_2-t_1<\delta}(N_\lambda(t_2)/\lambda - N_\lambda(t_1)/\lambda) > \eta\right) = -\infty. \tag{22}$$

By our assumption that the system is empty, (20) is obvious. Condition (22) will follow as a direct consequence of our analysis of (21). Now, to prove (21)



(a) Representation of $Q(t,y)$ (b) Areas of $\mathcal{N}_1^\delta(m,n)$ and $\mathcal{N}_2^\delta(m,n)$

FIG 1. *Illustrations for* $Q_\lambda(t,y)$.

consider

$$P\left(w\left(\frac{Q_\lambda}{\lambda},\delta\right)>\eta\right)$$
$$\leq \sum_{m=0}^{\lfloor T/\delta\rfloor}\sum_{n=0}^{\lfloor K/\delta\rfloor} P\left(\sup_{\substack{0<t_1-t_2<\delta,\ t_1\in(m\delta,(m+1)\delta]\\ |y_2-y_1|<\delta,\ y_1\in(n\delta,(n+1)\delta]}} |Q_\lambda(t_1,y_1)-Q_\lambda(t_2,y_2)|>\lambda\eta\right)$$

It is best to proceed our analysis by keeping in mind the pictorial representation that we shall describe. One can represent the arrival and status of each customer in a two-dimensional plane, with x-axis representing the arrival time and y-axis the service time at the time of arrival. Under this representation, $Q_\lambda(t,y)$ is the number of points in the triangle formed by a vertical line and a 45° line passing through (t,y) in its northwest direction. Figure 1(a) depicts the shape of this triangle. Consequently, we have

$$P\left(\sup_{\substack{0<t_1-t_2<\delta,\ t_1\in(m\delta,(m+1)\delta],\\ |y_2-y_1|<\delta,\ y_1\in(n\delta,(n+1)\delta]}} |Q_\lambda(t_1,y_1)-Q_\lambda(t_2,y_2)|>\lambda\eta\right)$$
$$\leq P(\mathcal{N}_1^\delta(m,n,\lambda)+\mathcal{N}_2^\delta(m,n,\lambda)>\eta\lambda)$$

where

$$\mathcal{N}_1^\delta(m,n,\lambda)=Q_\lambda(m\delta,(n-1)\delta)-Q_\lambda(m\delta,(n+2)\delta)$$

is the number of customers present at time $m\delta$ who have residual service time between $(n-1)\delta$ and $(n+2)\delta$, and

$$\mathcal{N}_2^\delta(m,n,\lambda)=\sum_{i=N_\lambda(m\delta)+1}^{N_\lambda((m+1)\delta)} I(V_i>(n-1)\delta),$$

is the number of arrivals between $m\delta$ and $(m+1)\delta$ which bring service requirements larger than $(n-1)\delta$. Figure 1(b) depicts the areas under which the points are included in $\mathcal{N}_1^\delta(m,n,\lambda)$ and $\mathcal{N}_2^\delta(m,n,\lambda)$. Fixing $\delta > 0$ and $\theta > 0$ we take the limit as $\lambda \to \infty$ in the following display, obtaining

(23)

$$
\begin{aligned}
&\frac{1}{\lambda}\log Ee^{\theta(\mathcal{N}_1^\delta(m,n,\lambda)+\mathcal{N}_2^\delta(m,n,\lambda))}\\
&\quad=\frac{1}{\lambda}\log E\exp\left\{\theta\left(\sum_{i=1}^{N_\lambda(m\delta)} I(m\delta+(n-1)\delta-A_i/\lambda\right.\right.\\
&\qquad\qquad<V_i\le m\delta+(n+2)\delta-A_i/\lambda)\\
&\qquad\qquad\left.\left.+\sum_{i=N_\lambda(m\delta)+1}^{N_\lambda((m+1)\delta)} I(V_i>(n-1)\delta)\right)\right\}\\
&\quad=\frac{1}{\lambda}\log E\exp\left\{\int_0^{m\delta}\log(e^\theta P(m\delta+(n-1)\delta-u\right.\\
&\qquad\qquad<V_i\le m\delta+(n+2)\delta-u)+1-P(m\delta+(n-1)\delta-u\\
&\qquad\qquad<V_i\le m\delta+(n+2)\delta-u))dN_\lambda(u)\\
&\qquad\qquad+\log(e^\theta\bar F((n-1)\delta)+F((n-1)\delta))\\
&\qquad\qquad\left.\cdot[N_\lambda((m+1)\delta)-N_\lambda(m\delta)]\right\}\\
&\quad\to\int_0^{m\delta}\psi_N(\log(e^\theta P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u)+1\\
&\qquad-P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u)))du\\
&\qquad+\psi_N(\log(e^\theta\bar F((n-1)\delta)+F((n-1)\delta)))\delta.
\end{aligned}
$$

Let us use $\psi_\delta(\theta;m,n)$ to denote the last expression (23). For fixed $\theta \ge 0$, we argue that $\psi_\delta(\theta,m,n) \to 0$ as $\delta \to 0$ uniformly over m,n. Indeed, for any m,n, the first term in (23)

$$
\begin{aligned}
&\int_0^{m\delta}\psi_N(\log(e^\theta P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u)+1\\
&\qquad-P(m\delta+(n-1)\delta-u<V_i\le m\delta+(n+2)\delta-u)))du\\
&\le\int_0^{K}\psi_N(\log(e^\theta P((n-1)\delta+u<V_i\le(n+2)\delta+u)+1\\
&\qquad-P((n-1)\delta+u<V_i\le(n+2)\delta+u)))du\\
&\le K\psi_N(\log(e^\theta\alpha(\delta)+1-\alpha(\delta)))
\end{aligned}
\tag{24}
$$

where $\alpha(\delta) := \sup_{x\in[0,K]} P(x < V_i \le x + 3\delta) = o(1)$ as $\delta \to \infty$ by our assumption that the distribution of V_i is continuous and the fact that a continuous function is uniformly continuous on a compact set. On the other hand, the second term in (23)

$$\psi_N(\log(e^\theta \bar{F}((n-1)\delta) + F((n-1)\delta)))\delta \le \psi_N(\theta)\delta \tag{25}$$

for any m, n. Combining (24) and (25), we get

$$\psi_\delta(\theta, m, n) \le K\psi_N(\log(e^\theta\alpha(\delta) + 1 - \alpha(\delta))) + \psi_N(\theta)\delta. \tag{26}$$

Now fix m and n. By Chernoff's inequality we get

$$P(\mathcal{N}_1^\delta(m,n,\lambda) + \mathcal{N}_2^\delta(m,n,\lambda) > \eta\lambda) \le e^{-\eta\theta\lambda + \psi_\delta(\theta,m,n)\lambda + o(\lambda)}$$

and so

$$\begin{aligned}
&\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(\mathcal{N}_1(m,n,\lambda) + \mathcal{N}_2(m,n,\lambda) > \eta\lambda\right)\\
&\quad\le -\eta\theta + \psi_\delta(\theta,m,n)\\
&\quad\le -\eta\theta + K\psi_N(\log(e^\theta\alpha(\delta) + 1 - \alpha(\delta))) + \psi_N(\theta)\delta
\end{aligned}$$

by (26). Hence

$$\begin{aligned}
&\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(w\left(\frac{Q_\lambda}{\lambda},\delta\right) > \eta\right)\\
&\quad\le \overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log\sum_{m=0}^{\lfloor T/\delta\rfloor}\sum_{n=0}^{\lfloor K/\delta\rfloor} P(\mathcal{N}_1^\delta(m,n,\lambda) + \mathcal{N}_2^\delta(m,n,\lambda) > \eta\lambda)\\
&\quad\le -\eta\theta + K\psi_N(\log(e^\theta\alpha(\delta) + 1 - \alpha(\delta))) + \psi_N(\theta)\delta
\end{aligned}$$

which gives

$$\lim_{\delta\to 0}\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P\left(w\left(\frac{Q_\lambda}{\lambda},\delta\right) > \eta\right) \le -\eta\theta.$$

Since θ can be arbitrarily large, we conclude (21). Finally, condition (22) follows from the analysis of $\mathcal{N}_2^\delta(m,1,\lambda)/\lambda$. □

4. Unbounded service times. In this section, we will extend our result to unbounded service times. The main intuition of the extension beyond the bounded case is to justify that we can ignore in certain sense the customers who arrive with very large service time. Let us first introduce a suitable truncation scheme. For any $K > 0$ and $\bar{q} \in AC_+(\mathcal{D})$ define

$$\phi_K(\bar{q})(t,y) = \int_0^t\int_{y-w}^K -\frac{\partial^2}{\partial s\partial z}\bar{q}(s,z)\Big|_{s=w,z=r+w} dr dw \tag{27}$$

for $t \in [0,T]$ and $y := u + t \ge 0$.

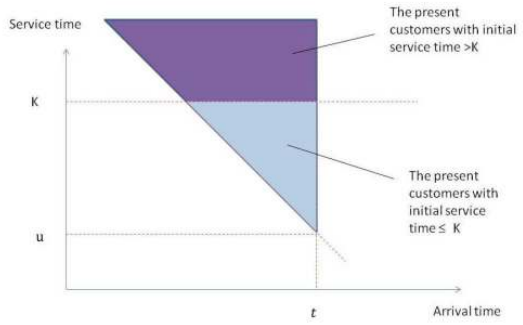


FIG 2. *Illustration for* $\phi_K(\bar{q})(t, t+u)$.

Since $\bar{q}$ is absolutely continuous, $\phi_K(\bar{q})(t, u+t$ is well defined. Moreover, the region over which the integration in (27) is performed corresponds to the triangular area depicted in light color in Figure 2. This region corresponds to the customers that are present at time t, have residual service time greater than y, and whose initial service time is less than K, as illustrated in Figure 2.

Moreover, for a sample path $\bar{Q}_\lambda$, define $\bar{Q}_{\lambda,K}$ as the two-parameter process derived from $\bar{Q}_\lambda$ by ignoring the arrivals with service time greater than K (one way to imagine is that they leave the system immediately upon arrival). Therefore, $\bar{Q}_{\lambda,K}$ is a two-parameter queue length process corresponding to an infinite server system with i.i.d. interarrival times following the law $\bar{U} = \sum_{i=1}^{G} U_i/\lambda$, where G is a geometric r.v. independent of the U_i's such that $P(G = n) = \bar{F}(K)^{n-1}F(K)$, $n \geq 1$. It is easy to check that the arrival process corresponding to $\bar{Q}_{\lambda,K}$, i.e. by ignoring the arrivals with initial service time larger than K, satisfies the conditions in Section 2.1. The service time then has the distribution function $F_K(x) = F(x)/F(K)$ for $x \in [0, K]$. We denote $(V_n^{(K)}, n = 1, \ldots)$ as the sequence of service times in this modified system.

Now recall the continuous version of $\bar{Q}_\lambda$, denoted by $\tilde{Q}_\lambda$ constructed in Section 2.3. Moreover, define $\tilde{Q}_{\lambda,K}$ to be the continuous approximation to $\bar{Q}_{\lambda,K}$ constructed in exactly the same fashion. In addition, for $\bar{q} \in AC_+(\mathcal{D}_K)$ define $I_K(\bar{q})$ as

$$\sup_{\theta(t,\cdot)\in C[0,T]\times[0,T+K]} \int_0^T \left[\int_t^{t+K} \theta(t, y-t) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy \right. \\ \left. - \psi_N^{(K)} \left(\log \left(\frac{1}{F(K)} \int_0^K e^{\theta(t,y)} dF(y) \right) \right) \right] dt, \tag{28}$$

and set $I_K(\bar{q}) = \infty$ otherwise, where $\psi_N^{(K)}$ is the infinitesimal logarithmic moment generating function corresponding to the truncated arrival process.

Theorem 2 yields that $\tilde{Q}_{\lambda,K}/\lambda$ satisfies a full large deviations principle with good rate function $I_K(\cdot)$. For $\bar{q} \in AC_+(\mathcal{D})$ we shall also evaluate $I_K(\bar{q})$ according to the expression (28).

Since the geometric r.v. G is independent of the U_i's, we can compute the associated logarithmic moment generating function of the modified interarrival times

$$\kappa^{(K)}(\theta) := \kappa(\theta) + \log\left(\frac{F(K)}{1 - \bar{F}(K)e^{\kappa(\theta)}}\right),$$

and from which we solve that the associated infinitesimal logarithmic moment generating function of the arrival process is

$$\psi_N^{(K)}(\theta) := \psi_N(\log(F(K)e^{\theta} + \bar{F}(K))).$$

Plugging in the above expressions into (28), we have the following expression of $I_K(\bar{q})$

$$\sup_{\theta(\cdot,\cdot)\in C(\mathcal{D}_K)} \int_0^T \left[\int_t^{T+K} \theta(t, y-t)\left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right) dy \right. \\ \left. - \psi_N\left(\log\left(\bar{F}(K) + \int_0^K e^{\theta(t,y)} dF(y)\right)\right)\right] dt. \tag{29}$$

At this point our strategy involves two steps. First, we want to show that $\bar{Q}_{\lambda,K}/\lambda$ and $\tilde{Q}_{\lambda,K}/\lambda$ are exponentially good approximations as $K \nearrow \infty$ to both $\bar{Q}_\lambda/\lambda$ and $\tilde{Q}_\lambda/\lambda$ respectively. The second step consists in using this fact, together with the properties of $I_K(\bar{q})$ as $K \nearrow \infty$ and also properties of $I(\bar{q})$ to conclude the identification of the rate function of $\tilde{Q}_\lambda/\lambda$.

So, to execute the first step we first define

$$N_\lambda^{(K)}(t) = \sum_{j=1}^{N_\lambda(t)} I(V_j > K),$$

that is, $N_\lambda^{(K)}(t)$ is the number of arrivals with service time larger than K in the λ-scaled system. Then we obtain the following result, which is proved at the end of this section.

LEMMA 4. *For any $\varepsilon > 0$,*

$$\lim_{K\to\infty} \overline{\lim}_{\lambda\to\infty} \frac{1}{\lambda} \log P(N_\lambda^{(K)}(T) > \lambda\varepsilon) = -\infty. \tag{30}$$

Consequently, $\bar{Q}_{\lambda,K}/\lambda$ and $\tilde{Q}_{\lambda,K}/\lambda$ are exponentially good approximations as $K \nearrow \infty$ to both $\bar{Q}_\lambda/\lambda$ and $\tilde{Q}_\lambda/\lambda$ respectively.

Using the previous lemma we obtain the following result. The proof is straightforward, but following our convention we shall give it at the end of the section.

LEMMA 5. *The family* $(\tilde{Q}_\lambda/\lambda : \lambda > 0)$ *satisfies a weak large deviations principle on* $C_+(\mathcal{D})$ *with rate function*

$$I^*(\bar{q}) := \sup_{\delta>0} \underline{\lim}_{K\to\infty} \inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\leq\delta\}} I_K(z).$$

We now extend the weak large deviations principle into a full large deviations principle with a good rate function using exponential tightness.

LEMMA 6. *The family* $(\tilde{Q}_\lambda/\lambda : \lambda > 0)$ *is exponentially tight on* $C_+(\mathcal{D})$ *and therefore it satisfies a full large deviations principle with good rate function* $I^*(\cdot)$.

We proceed to show the identification $I^*(\bar{q}) = I(\bar{q})$. We now collect useful properties that we will need to show this identification.

LEMMA 7.

i) For any $\bar{q}$ *such that* $I(\bar{q}) < \infty$, *we have* $I(\phi_K(\bar{q})) = I_K(\phi_K(\bar{q})) = I_K(\bar{q}) \nearrow I(\bar{q})$ *as* $K \to \infty$; *the notation* $I_K(\bar{q}) \nearrow I(\bar{q})$ *implies that* $(I_K(\bar{q}) : K > 0)$ *is non-decreasing in* K *and convergent to* $I(\bar{q})$.

ii) For any $\bar{q}$ *such that* $I(\bar{q}) = \infty$, *and each* $M > 0$, *there exists a projection* p_κ *(following the notation introduced in the proof of Lemma 1) such that, for large enough* K,

$$I_K(p_\kappa(\bar{q})) > M.$$

iii) Finally, with κ *from ii) there exists* $\varepsilon > 0$ *such that if* $\hat{q} \in C_+(\mathcal{D})$ *and* $||\hat{q} - \bar{q}||_{\mathcal{D}} < \varepsilon$ *then*

$$I_K(p_\kappa(\hat{q})) > M.$$

We now are ready to prove the following important result of this section.

THEOREM 3. $\tilde{Q}_\lambda/\lambda$ *satisfies a large deviations principle with good rate function defined in (4) under the uniform topology on* $[0,T] \times [0,\infty)$.

PROOF OF THEOREM 3. Given Lemma 6 all we need to show is that $I^*(\bar{q}) = I(\bar{q})$. Suppose that $\bar{q}$ is such that $I(\bar{q}) = \infty$. Then, parts ii) and iii) in Lemma 7 imply in particular that for every M, there exists K, a projection κ, and $\varepsilon > 0$ such that $I_K(p_\kappa(\hat{q})) > M$ for any $||\hat{q} - \bar{q}||_{\mathcal{D}} < \varepsilon$. Consequently,

we conclude, by using the monotonicity of $I_K(\bar{q})$ as a function of K and taking subsequences, that

$$I^*(\bar{q}) = \sup_{\delta>0} \underline{\lim}_{K\to\infty} \inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\le\delta\}} I_K(z) = \sup_{\delta>0}\sup_{K>0} \inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\le\delta\}} I_K(z) = \infty.$$

If $I(\bar{q}) < \infty$, then note that

$$I^*(\bar{q}) = \sup_{\delta>0}\sup_{K>0} \inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\le\delta\}} I_K(z) = \sup_{K>0}\sup_{\delta>0} \inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\le\delta\}} I_K(z).$$

Further, observe that

$$\sup_{\delta>0} \inf_{\{z:||z-\bar{q}||_{\mathcal{D}}\le\delta\}} I_K(z) = \sup_{\delta>0} \inf_{\{\phi_K(z):||\phi_K(z)-\phi_K(\bar{q})||_{\mathcal{D}_K}\le\delta\}} I_K(\phi_K(z)),$$

Since $I_K(\cdot)$ is a rate function (in particular $I_K(\cdot)$ is lower semicontinuous) we have that

$$\sup_{\delta>0} \inf_{\{\phi_K(z):||\phi_K(z)-\phi_K(\bar{q})||_{\mathcal{D}_K}\le\delta\}} I_K(\phi_K(z)) = I_K(\phi_K(\bar{q}))$$

and then by part i) of Lemma 7 we conclude that $\sup_{K>0} I_K(\phi_K(\bar{q})) = I(\bar{q})$, thus concluding that $I^*(\bar{q}) = I(\bar{q})$ as claimed. □

We finish this section with the proof of Theorem 1.

PROOF OF THEOREM 1. All we need to show is that $\bar{Q}_\lambda/\lambda$ and $\tilde{Q}_\lambda/\lambda$ are exponentially equivalent. This follows exactly as in the proof of Corollary 1 since $\|\bar{Q}_\lambda - \tilde{Q}_\lambda\|_{\mathcal{D}} \le 4$ a.s. The measurability issue again is dealt with using separability. The result then follows by applying Theorem 4.2.13 in Dembo and Zeitouni (1998). □

4.1. *Proofs of technical results.* We now provide the proof of the pending technical results.

PROOF OF LEMMA 5. This result is a direct application of part a) in Theorem 4.2.16 in Dembo and Zeitouni (1998). □

PROOF OF LEMMA 6. This is similar to the case with bounded service time, but the conditions for tightness are slightly different given that our domain $\mathcal{D}$ is not compact. We must show that for any $\eta, \gamma > 0$, we can choose small enough $\rho > 0$, such that for $\delta < \rho$,

$$\frac{1}{\lambda}\log P(w(\tilde{Q}_\lambda/\lambda, \delta) > \eta) < -\gamma$$

when λ is large; this part is indeed basically the same as the case $\mathcal{D}_K$. In addition, however, we also must show that for all $\eta > 0$ and every $a > 0$ there exists $K > 0$ such that

$$P\left(\sup_{t\in[0,T]}\sup_{y\geq K}\tilde{Q}_\lambda(t,y)/\lambda > \eta\right) \leq \exp(-\lambda a). \tag{31}$$

Note that

$$\begin{aligned} P(w(\tilde{Q}_\lambda/\lambda,\delta) > \eta) &\leq P\left(w(\tilde{Q}_\lambda/\lambda,\delta) > \eta, \|\tilde{Q}_\lambda/\lambda - \tilde{Q}_{\lambda,K}/\lambda\| \leq \frac{\eta}{2}\right) \\ &\quad + P\left(\|\tilde{Q}_\lambda/\lambda - \tilde{Q}_{\lambda,K}/\lambda\| > \frac{\eta}{2}\lambda\right) \\ &\leq P\left(w(\tilde{Q}_{\lambda,K}/\lambda,\delta) > \frac{\eta}{2}\right) + P\left(N_\lambda^{(K)}(T) > \lambda\eta/2\right), \end{aligned} \tag{32}$$

and

$$P\left(\sup_{t\in[0,T]}\sup_{y\geq K}\tilde{Q}_\lambda(t,y)/\lambda > \eta\right) \leq P\left(N_\lambda^{(K)}(T) > \lambda\eta\right).$$

By Lemma 4, for every $\gamma > 0$ we can choose K large enough such that

$$\frac{1}{\lambda}\log P\left(N_\lambda^{(K)}(T) > \lambda\eta/2\right) < -2\gamma.$$

for all λ large enough. So, condition (31) is enforced and the second term in the sum in (32) is also appropriately controlled. Now, by a similar argument as in Lemma 2 in the previous section, for a chosen K, we have, for all small enough δ,

$$\frac{1}{\lambda}\log P\left(w(\tilde{Q}_{\lambda,K}/\lambda,\delta) > \frac{\eta}{2}\right) < -2\gamma.$$

for large enough λ. In summary, we get

$$\frac{1}{\lambda}\log P(w(\tilde{Q}_\lambda/\lambda,\delta) > \eta) < -\gamma$$

for large enough λ. Therefore, exponential tightness follows. It follows immediately that a weak large deviations principle and exponential tightness imply a full large deviations principle. The goodness of the rate function then is a consequence of exponential tightness together with the weak large deviations principle; see Lemma 1.2.18, p. 8, part b) of Dembo and Zeitouni (1998). □

Proof of Lemma 7. We start with part i), assuming that $I(\bar{q}) < \infty$. Since

$$\frac{\partial^2}{\partial t\partial y}\phi_K(\bar{q})(t,y) = \frac{\partial^2}{\partial y\partial t}\phi_K(\bar{q})(t,y) = \frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)1_{\{y\leq K+t\}},$$

we have immediately that $I(\phi_K(\bar{q})) = I_K(\phi_K(\bar{q})) = I_K(\bar{q})$. It is obvious that $I_K(\bar{q})$ is non-decreasing in K and that $I_K(\bar{q}) \leq I(\bar{q})$. On the other hand, there exists $\theta^n \in C_b(\mathcal{D})$ (the space of bounded and continuous functions on $\mathcal{D}$) such that

$$I^n(\bar{q}) := \int_0^T \left[\int_t^\infty \theta^n(t, y-t) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy \right. \\ \left. - \psi_N \left(\log \left(\int_0^\infty e^{\theta^n(t,y)} dF(y) \right) \right) \right] dt$$

converges to $I(\bar{q})$ as $n \to \infty$. Since $I(\bar{q}) < \infty$, it follows easily that $\partial^2 \bar{q}(\cdot)/(\partial t \partial y)$ is integrable over $\mathcal{D}$. Therefore, given that $\theta^n(\cdot)$ is bounded,

$$\int_0^T \int_K^\infty \theta^n(t, y-t) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy dt \to 0$$

as $K \to \infty$. Besides, for given n, as ψ_N is uniformly continuous on the bounded set $[-M_n, M_n]$ where $M_n = \sup |\theta^n(t,y)| < \infty$, we have that

$$\psi_N \left(\log \left(\bar{F}(K) + \int_0^K e^{\theta^n(t,y)} dF(y) \right) \right)$$

converges to

$$\psi_N \left(\log \left(\int_0^\infty e^{\theta^n(t,y)} dF(y) \right) \right)$$

uniformly on $t \in [0,T]$. In summary, $\underline{\lim}_{K \to \infty} I_K^n(\bar{q}) = I^n(\bar{q})$ as $K \to \infty$. Therefore, there exists K_n such that $I_{K_n}^n(\bar{q}) \geq I^n(\bar{q}) - 1/n$. Recall that $I_{K_n}^n(\bar{q}) \leq I_{K_n}(\bar{q})$ and consequently we obtain

$$I^n(\bar{q}) - \frac{1}{n} \leq I_{K_n}(\bar{q}) \leq I(\bar{q}).$$

Since $I_K(\bar{q})$ increases in K, we have $I_K(q) \nearrow I(q)$ as claimed.

For part ii), when $I(\bar{q}) = \infty$, there are two cases: a) $\bar{q}$ is not absolutely continuous, and b) $\bar{q}$ is absolutely continuous. Case b) in turn is divided into two subcases: b.1) $\partial^2 \bar{q}(t,y)/(\partial t \partial y)$ is not integrable over $\mathcal{D}$, and b.2) $\partial^2 \bar{q}(t,y)/(\partial t \partial y)$ is integrable over $\mathcal{D}$. We shall proceed to analyze all these cases now. For Case a). We can construct a projection p_κ with $I_K(p_\kappa(\bar{q})) > M$ for K large enough, as we did in the proof of Lemma 1. For Case b), we have that $\bar{q}$ is absolutely continuous, but

$$\sup_{\theta(\cdot,\cdot) \in C_b(\mathcal{D})} \int_0^T \left[\int_0^\infty \theta(t,y) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t, y+t) \right) dy \right.$$

$$- \psi_N \left(\log \left(\int_0^\infty e^{\theta(t,y)} dF(y) \right) \right) \Bigg] dt = \infty,$$

so we proceed to study case b.1): Assume that $\partial^2 \bar{q}(t,y)/\partial t \partial y$ is not integrable on $\mathcal{D}$. We shall assume that

$$\int_0^T \int_t^\infty \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right)^+ dy dt = \infty \tag{33}$$

(if this integral is finite, then integral of the negative part must diverge and the analysis that follows next is identical). As in the proof of Lemma 1, given a projection κ induced by $0 \le t_1 < t_2 < \cdots < t_m \le T$ and $0 \le y_0 < y_1 < \cdots < y_{n+1}$, define

$$\begin{aligned} \delta_{ij}(\kappa) &:= \bar{q}(t_i, y_{j-1}) - \bar{q}(t_i, y_j) - \bar{q}(t_{i-1}, y_{j-1}) + \bar{q}(t_{i-1}, y_j) \\ &= - \int_{t_{i-1}}^{t_i} \int_{y_{j-1}}^{y_j} \frac{\partial^2}{\partial y \partial t} \bar{q}(t,y) dy dt, \end{aligned} \tag{34}$$

as long as $y_{j-1} \ge t_{i-1}$ (otherwise, if $y_{j-1} < t_{i-1}$, $\delta_{ij}(\kappa) = 0$). Then, from (33) and (34), it follows easily that for any M, there exists a partition κ such that

$$\sum_{i,j} \delta_{ij}(\kappa) > M + T\psi_N(1).$$

Therefore, for large enough K,

$$I_K(p_\kappa(\bar{q})) \ge \sum_{i,j} \delta_{ij}(\kappa) - T\psi_N(1) > M.$$

Now, for case b.2) suppose that $\partial^2 \bar{q}(t,y)/\partial t \partial y$ is integrable on $\mathcal{D}$. We can find $\theta(t,y)$ such that

$$\begin{aligned} 3M < \int_0^T \int_t^\infty \theta(t, y-t) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y) \right) dy - \psi_N \left(\log \left(\int_0^\infty e^{\theta(t,y)} dF(y) \right) \right) dt \\ < \infty. \end{aligned}$$

Following the same line of reasoning as in the proof of part i) we can conclude that there exists $K > 0$ such that $I_K(\bar{q}) > 2M$. According to Dawson-Gartner Theorem, $I_K(\bar{q}) = \sup I_K(p_\kappa(\bar{q}))$ where the supremum is taken over all projections restricted to $\{t \in [0,T], 0 \le y \le t + K\}$. As a result, there exists some projection p_κ such that $I_K(p_\kappa(\bar{q})) > M$ and hence we are done.

Now we turn to part iii). So far we proved that for any $\bar{q}$ and $M > 0$, we can find a projection p_κ such that $I_K(p_\kappa(\bar{q})) > 2M$. As discussed in the proof of Lemma 1, we have

(35)
$$\sup_{\{\theta_{ij}:1\le i\le m,1\le j\le n+1\}} \sum_{i=1}^{m}\sum_{j=1}^{n+1}\theta_{ij}\delta_{ij}(\kappa) - \sum_{i=1}^{m}\int_{t_{i-1}}^{t_i}\psi_N^{(K)}\left(\log\sum_{j=1}^{n+1}e^{\theta_{ij}}P(y_{j-1}-u<V_1^{(K)}\le y_j-u)\right)du > 2M$$

where $\delta_{ij}(\kappa)$ is induced by the projection p_κ. From (35), and by the definition of supremum, there exists some θ_{ij} such that

$$\sum_{i=1}^{m}\sum_{j=1}^{n+1}\theta_{ij}\delta_{ij}(q) - \sum_{i=1}^{m}\int_{t_{i-1}}^{t_i}\psi_N^{(K)}\left(\log\sum_{j=1}^{n+1}e^{\theta_{ij}}P(y_{j-1}-u<V_1^{(K)}\le y_j-u)\right)du > 3M/2.$$

For all $\varepsilon > 0$ and $\hat{q} \in B_\varepsilon(p)$, we have

$$|\delta_{ij}(q) - \delta_{ij}(\hat{q})| \le 4\varepsilon.$$

Hence for $\varepsilon = M/(8\sum_{i,j}|\theta_{ij}|)$ and all $\hat{q} \in B_\varepsilon(q)$, we have

$$\begin{aligned} I(p_\kappa(\hat{q})) &\ge \sum_{i=1}^{m}\sum_{j=1}^{n+1}\theta_{ij}\delta_{ij}(\hat{q}) \\ &\quad - \sum_{i=1}^{m}\int_{t_{i-1}}^{t_i}\psi_N^{(K)}\left(\log\sum_{j=1}^{n+1}e^{\theta_{ij}}P(y_{j-1}-u<V_1^{(K)}\le y_j-u)\right)du \\ &\ge \sum_{i=1}^{m}\sum_{j=1}^{n+1}\theta_{ij}\delta_{ij}(q) \\ &\quad - \sum_{i=1}^{m}\int_{t_{i-1}}^{t_i}\psi_N^{(K)}\left(\log\sum_{j=1}^{n+1}e^{\theta_{ij}}P(y_{j-1}-u<V_1^{(K)}\le y_j-u)\right)du \\ &\quad - 4\varepsilon\sum_{i=1}^{m}\sum_{j=1}^{n+1}|\theta_{ij}| \\ &> M. \end{aligned}$$

Thus we conclude the result. □

PROOF OF LEMMA 4. Let $N_\lambda^{(K)}(T)$ be the total number of arrivals from time 0 up to T with service time longer than K, under the λ-scaled system. Then following Glynn (1995) (or as in the proof of Lemma 2) we have that

$$\lim_{\lambda\to\infty}\frac{1}{\lambda}\log Ee^{\theta N_\lambda^{(K)}(T)} = T\psi_N(\log(e^\theta\bar{F}(K)+F(K))).$$

Chernoff's bound yields

$$P(N_\lambda^{(K)}(T)>\lambda\varepsilon)\le\exp\{-\theta\lambda\varepsilon+\lambda T\psi_N(\log(e^\theta\bar{F}(K)+F(K)))+o(\lambda)\}.$$

Hence

$$\begin{aligned}&\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P_\lambda(N_\lambda^{(K)}(T)>\lambda\varepsilon)\\&\quad\le-\theta\varepsilon+T\psi_N(\log(e^\theta\bar{F}(K)+F(K)))\end{aligned}$$

Letting $K\to\infty$ gives

$$\lim_{K\to\infty}\overline{\lim}_{\lambda\to\infty}\frac{1}{\lambda}\log P_\lambda(N_\lambda^{(K)}(T)>\lambda\varepsilon)\le-\theta\varepsilon.$$

Since θ can be arbitrarily large, the result follows. □

5. Examples. This section is devoted to two examples that apply the large deviations principle that we have developed in the previous sections. The first example is on the most likely path to overflow in a loss queue, while the second example is on the ruin of a large life insurance portfolio that embeds an infinite server queue with service cost.

EXAMPLE 1 (*Finite-horizon maximum of queue length process for* $M/G/\infty$). Consider an $M/G/\infty$ queue with Poisson arrivals with rate λ. Suppose that the service times have a density $f(\cdot)$ with respect to the Lebesgue measure. The system initially starts empty.

We want to find the optimal large deviations sample path to attain the event $\{\max_{0\le t\le T}\bar{Q}_\lambda(t,t)/\lambda\ge x\}$, for fixed T and x, as $\lambda\to\infty$; this event corresponds precisely to the event of observing a loss in a queue with λx servers, no waiting room, starting empty. Note that $g(\bar{q}):=\max_{0\le t\le T}\bar{q}(t,t)$ is a continuous function under the uniform norm, so the contraction principle is directly applicable.

We impose the condition that $\int_0^T\bar{F}(t)dt<x$. This condition implies that the probability for the queue to reach λx decreases exponentially fast as $\lambda\to\infty$ (Such condition will be clear when we solve the constrained optimization in a moment).

To proceed, let us first observe that $\psi_N(\theta) = e^\theta - 1$. The maximization problem in (4) can be solved and the rate function is immediately recognized as

$$\int_0^T \int_t^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dydt + T,$$

which is easily seen to be a convex function of $\partial^2\bar{q}(t,y)/\partial t\partial y$. To find the optimal sample path amounts to solving the minimization problem

(36)
$$\begin{aligned}\text{min}\quad & \int_0^T \int_t^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dydt + T\\ \text{subject to}\quad & \max_{0\le u\le T}\int_0^u \int_u^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)dydt \ge x,\end{aligned}$$

which is a convex optimization problem. The integral $\int_0^s \int_s^\infty (-\partial^2\bar{q}(t,y)/(\partial t\partial y))dydt$ is equal to $\bar{q}(s,s)$ when $\bar{q}$ is absolutely continuous and the integral is finite, and $\bar{q}(s,s)$ represents the scaled queue length process at time s.

To solve (36), we first consider a fixed u in the constraint and then optimize over u. When considering u fixed we replace the constraint in (36) by $\bar{q}(u,u) \ge x$. Under this new constraint, it suffices to look at the time 0 to s in the objective function, that is, we now solve

(37)
$$\begin{aligned}\text{min}\quad & \int_0^u \int_t^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dydt + u\\ \text{subject to}\quad & \int_0^u \int_u^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)dydt \ge x.\end{aligned}$$

The solution to (36) is then the optimal sample path from (37), among $0 \le u \le T$, that gives the smallest objective.

We now consider (37). Introducing a Lagrange multiplier $\mu \ge 0$, we minimize

$$\int_0^u \left(\int_t^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dy - \mu\int_u^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)dy\right)dt.$$

By a formal application of Euler-Lagrange equations, we differentiate the integrand with respect to $-\partial^2\bar{q}(t,y)/\partial t\partial y$ to get

$$\begin{cases} \log\left(\dfrac{-\partial^2\bar{q}(t,y)/\partial t\partial y}{f(t-y)}\right)=0 & \text{for } t\le y\le u \\ \log\left(\dfrac{-\partial^2\bar{q}(t,y)/\partial t\partial y}{f(t-y)}\right)-\mu=0 & \text{for } y>u \end{cases}$$

which gives

$$-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)=\begin{cases} f(y-t) & \text{for } t\le y\le u \\ \mu f(y-t) & \text{for } y>u \end{cases}$$

for some $\mu\ge 1$ (we replace e^{μ} by another dummy μ for convenience). Complementary slackness then implies

$$\int_0^u\int_u^\infty(-\partial^2\bar{q}(t,y)/\partial t\partial y)dydt=\int_0^u\int_u^\infty \mu f(y-t)dydt=x,$$

which in turn gives

$$\mu=\frac{x}{\int_0^u\bar{F}(t)dt}$$

(note that we have assumed $\int_0^u\bar{F}(t)dt<x$ and so the condition $\mu>1$ is satisfied). As a result

$$-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)=\begin{cases} f(y-t) & \text{for } t\le y\le u \\ \dfrac{xf(y-t)}{\int_0^u\bar{F}(t)dt} & \text{for } y>u. \end{cases}$$

The optimal sample path $\bar{q}(t,y)$ leading to the constraint $\bar{q}(u,u)\ge x$ is given by

$$\begin{aligned}\bar{q}(t,y)&=\int_0^t\int_y^\infty\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(s,w)\right)dwds \\ &=\int_0^t\left(\int_{y\wedge u}^u f(w-s)dw+\int_{u\vee y}^\infty\frac{xf(w-s)}{\int_0^u\bar{F}(r)dr}dw\right)ds\end{aligned}$$

Transforming into $q(t,y)=\bar{q}(t,y+t)$ and some simple calculus gives the optimal sample path

$$q(t,y)=\int_y^{y+t}\bar{F}(s)ds-\int_{u-t}^u\bar{F}(s)ds+\frac{x\int_{u-t}^u\bar{F}(s)ds}{\int_0^u\bar{F}(r)dr},\ \text{for } y+t\le u,$$

and

$$q(t,y) = \frac{x\int_y^{y+t}\bar{F}(s)ds}{\int_0^u \bar{F}(r)dr}, \text{ for } y+t>u.$$

In connection to our discussion about the direct rate function representation in terms of Q_λ/λ (see equation (8), in Section 2.2.1), one can check that $\partial q(t,y)/\partial y$ is not continuous on the line $y=t$ and therefore $\partial^2 q(t,y)/\partial y^2$ does not exists though $I(\bar{q})$ is finite.

Note also that the objective is

(38)
$$\int_0^u \left(-\int_t^u f(y-t)dy + \int_u^\infty \frac{xf(y-t)}{\int_0^u \bar{F}(t)dt}\left(\log\left(\frac{x}{\int_0^u \bar{F}(t)dt}\right)-1\right)dy\right)dt + u$$
$$= \int_0^u \bar{F}(t)dt + \left(\log\left(\frac{x}{\int_0^s \bar{F}(t)dt}\right)-1\right)x.$$

This is the rate function corresponding to the probability $P(\bar{Q}_\lambda(u,u)\geq \lambda x) = P(Q_\lambda(u,0)\geq \lambda x)$, where $Q_\lambda(u,0)$ is the queue length at time u. This rate of decay is consistent with direct calculation using the fact that $Q_\lambda(u,0)$ is a Poisson random variable with rate $\lambda\int_0^u \bar{F}(t)dt$, which gives

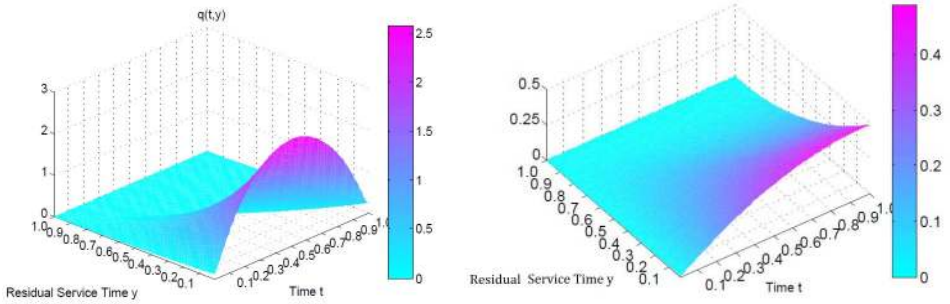
$$P(Q_\lambda(u,0)>\lambda x) = \sum_{n\geq \lambda x} e^{-\lambda\int_0^u \bar{F}(t)dt}\left(\lambda\int_0^u \bar{F}(t)dt\right)^n/n!.$$

For a consistency check, our result here can in fact recover the large deviations for the arrival process itself. If one changes the constraint in (37) to $\bar{q}(u,0) = \int_0^u\int_0^\infty(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y))dydt \geq x$, the optimal value of (37) then becomes $x[\log(x/u)-1]+u$, which coincides with the exponential decay rate of $P(\text{Poisson}(\lambda u)>\lambda x)$ as $\lambda\to\infty$.

Figures 3 and 4 illustrate both the law of large numbers (i.e. the typical path) and the most likely path to the overflow event $\max_{0\leq u\leq T}\bar{Q}_\lambda(u,u)/\lambda \geq x$ for $T=1$, $x=2$. The underlying service time distribution is uniform in the interval $[0,1]$. We can see that the optimal path of $Q(t,y)$ increases gradually over time to overflow at time 1.

It is easy to see that since we assume $\int_0^T \bar{F}(u)du < x$, the rate function (38) is non-decreasing in s, and as a result an optimal time horizon is T. If the service time has bounded support $[0,K]$ with $K<T$, then the selection of any time $s\in[K,T]$ will give an optimal sample path.

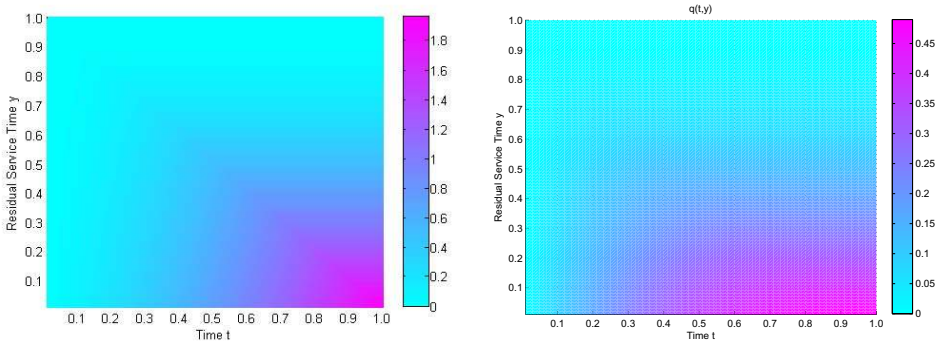
EXAMPLE 2 (*Insurance risk process*). The net reserve of a life insurance company consists of the premium collected from policyholders, deducted by the benefit paid to policyholders in the event of deaths; often all these



(a) $Q(t, y)$ for the most likely path to overflow (surface)

(b) $Q(t, y)$ for the unconditional path (surface)

FIG 3. *Surface plots of the asymptotic surface $Q_\lambda(t, y)/\lambda$, as λ increases, both an optimal (most likely) path leading to overflow, and the unconditional path.*



(a) $Q(t, y)$ for the most likely path to overflow (contour)

(b) $Q(t, y)$ for the unconditional path (contour)

FIG 4. *Contour plots of the asymptotic surface $Q_\lambda(t, y)/\lambda$, as λ increases, both an optimal (most likely) path leading to overflow, and the unconditional path.*

payments are discounted at zero in order to recognize the value of money in time. When policyholders arrive at the insurance company over time (an arrival is interpreted as the moment when a contract is signed), one can model the net assets of the insurer as a function of the underlying arrivals and death events of policyholders. Specifically, we shall assume that policyholders arrive according to a Poisson process with rate λ, and that the time-until-death upon arrival of the policyholders are independent and identically distributed. Moreover, we assume that the time-until-death upon arrival has density $f(\cdot)$, distribution function $F(\cdot)$, and tail distribution $\bar{F}(\cdot)$. The time-until-death in this setting can be thought as the service time in

the queueing context. We shall assume without loss of generality that the initial net reserve of the company is zero.

It is often more convenient to work with the negative net reserve process, also known as the *aggregate loss process*, defined as the total benefit that the insurer has paid up to time t, minus the total premium received up to time t. For a policyholder who arrives at time A_i, and who dies at time $A_i + V_i < t$, the payoff by the insurer, discounted at time zero, is denoted $h_1(A_i, A_i + V_i)$; here A_i and V_i are the arrival time and time-until-death at the time of arrival of the policyholder. This quantity, $h_1(s, y)$, for $y \geq s$, captures the benefit paid at y minus the accumulated premium collected from time s to y. On the other hand, for a policyholder who has arrived prior to t, at time A_i, and who is still alive at time t, the payoff from the insurer to the policyholder is $h_2(A_i, t)$ (typically $h_2(A_i, t)$ will be negative as it represents premium that are paid to the insurer, so the payoff is negative). Here $h_2(s, t)$, for $t \geq s$, captures the premium accumulated from s up to the present time t, discounted to obtain the net present value at time zero.

Consider, for instance, the setting of whole life insurance policies. That is, policies that pay a benefit b to the family of the policyholder, at the time of eventual death, in exchange of a premium which is paid at rate p continuously in time during all the time the policy was held, from arrival, up until the time of death. If the interest rate (or force of interest as it is known in the insurance setting) is constant equal to $\delta > 0$, then

$$h_1(s, y) = be^{-\delta y} - \int_s^y pe^{-\delta r} dr = be^{-\delta y} - p(e^{-\delta s} - e^{-\delta y})/\delta,$$

and

$$h_2(s, t) = -\int_s^t pe^{-\delta r} dr = -p(e^{-\delta s} - e^{-\delta t})/\delta.$$

The aggregate loss process, $S_\lambda(t)$, is represented as the net present value of the sum of the payoffs for all policyholders who arrive before t and it is given by

$$\begin{aligned} S_\lambda(t) &= \sum_{i=1}^{N_\lambda(t)} \left(I\left(A_i + V_i \leq t\right) h_1(A_i, A_i + V_i) + I(A_i + V_i > t) h_2(A_i, t) \right) \\ &= \int_0^t \int_s^t h_1(s, y) d\bar{Q}_\lambda(ds, dy) + \int_0^t \int_t^\infty h_2(s, y) d\bar{Q}_\lambda(ds, dy). \end{aligned}$$

We claim that $S_\lambda(\cdot)$ is a continuous function of $\bar{Q}_\lambda(\cdot)$ under the uniform topology on $\mathcal{D}$. In order to see this, define $D_\lambda(t)$ to be the number of departures by time t, that is,

$$D_\lambda(t) = N_\lambda(t) - \bar{Q}_\lambda(t, t) = \bar{Q}_\lambda(t, 0) - \bar{Q}_\lambda(t, t). \tag{39}$$

Note that $D_\lambda(\cdot)$ and $N_\lambda(\cdot)$ are clearly continuous functions of $\bar{Q}_\lambda(\cdot)$. Moreover, we have that

$$\sum_{i=1}^{N_\lambda(t)} I\left(A_i + V_i \le t\right) h_1(A_i, A_i + V_i) = \int_0^t \int_0^t h_1\left(s,u\right) D_\lambda\left(du\right) N_\lambda\left(ds\right),$$

and therefore

$$S_\lambda\left(t\right) = \int_0^t \int_0^t h_1\left(s,u\right) D_\lambda\left(du\right) N_\lambda\left(ds\right) + \int_0^t \int_t^\infty h_2(s,y)\bar{Q}_\lambda(ds,dy).$$

Now, integration by parts shows that

(40)
$$\begin{aligned}
&\int_0^t \int_0^t h_1\left(s,u\right) D_\lambda\left(du\right) N_\lambda\left(ds\right) \\
&\quad = \int_0^t \int_0^t \frac{\partial^2}{\partial u \partial s} h_1\left(s,u\right) D_\lambda\left(u\right) N_\lambda\left(s\right) dsdu - N_\lambda\left(t\right) \int_0^t D_\lambda\left(u\right) \frac{\partial}{\partial u} h_1\left(t,u\right) du \\
&\qquad - D_\lambda\left(t\right) \int_0^t \frac{\partial}{\partial s} h_1\left(s,t\right) N_\lambda\left(s\right) ds + D_\lambda\left(t\right) N_\lambda\left(t\right) h_1\left(t,t\right).
\end{aligned}$$

A similar development yields

(41)
$$\begin{aligned}
\int_0^t \int_t^\infty h_2(s,y)\bar{Q}_\lambda(ds,dy) = &\int_t^\infty \int_0^t \bar{Q}_\lambda(s,y) \frac{\partial^2 h_2\left(s,y\right)}{\partial s \partial y} dsdy \\
&- \int_t^\infty \bar{Q}_\lambda(t,y) \frac{\partial h_2\left(t,y\right)}{\partial y} dy \\
&+ \int_0^t \frac{\partial h_2(s,t)}{\partial s} \bar{Q}_\lambda(s,t) ds - \bar{Q}_\lambda(t,t) h_2\left(t,t\right).
\end{aligned}$$

It is now not difficult to see from (40) and (41) that indeed $S_\lambda(\cdot)$ is a continuous function of $Q_\lambda(\cdot)$ in the uniform topology on $[0,T] \times [0,\infty)$.

Consider the finite-horizon ruin probability that the negative net asset of the insurer rises above the level λx by time T. That is, the event $\{\max_{t\in[0,T]} S_\lambda(t)/\lambda \ge x\}$. We wish to solve for the most likely path that leads to this event and therefore, applying our theory, we must solve the following convex calculus of variations problem.

$$\begin{aligned}
\min \quad & \int_0^T \int_t^\infty \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y)\right) \left(\log\left(\frac{-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)}{f(y-t)}\right) - 1\right) dydt + T \\
\text{subject to} \quad & \max_{0\le u\le T} \int_0^u \left(\int_t^u h_1(t,y) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y)\right) dy \right. \\
& \qquad \left. + \int_u^\infty h_2(t,u) \left(-\frac{\partial^2}{\partial t \partial y} \bar{q}(t,y)\right) dy \right) dt \ge x
\end{aligned}$$

Following the recipe of Example 1, we first consider

$$\begin{aligned}\min \quad & \int_0^u \int_t^\infty \left(-\frac{\partial^2}{\partial t \partial y}\bar{q}(t,y)\right)\left(\log\left(\frac{-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)}{f(y-t)}\right)-1\right)dydt+u \\ \text{subject to} \quad & \int_0^u \left(\int_t^u h_1(t,y)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy \right. \\ & \qquad \left. + \int_u^\infty h_2(t,u)\left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y)\right)dy\right)dt \geq x.\end{aligned}$$

Introducing the Lagrange multiplier $\mu \geq 0$, we get

$$-\frac{\partial^2}{\partial t\partial y}\bar{q}(t,y) = \begin{cases} f(y-t)e^{\mu h_1(t,y)} & \text{for } t \leq y \leq u \\ f(y-t)e^{\mu h_2(t,u)} & \text{for } y > u. \end{cases}$$

When x is large, complementary slackness forces μ to satisfy

$$\int_0^u \left(\int_t^u f(y-t)e^{\mu h_1(t,y)}h_1(t,y)dy + \bar{F}(u-t)e^{\mu h_2(t,u)}h_2(t,u)\right)dt = x \tag{42}$$

for some $\mu > 0$. Denote the integration on the left hand side by $G(\mu)$, then

$$G'(\mu) = \int_0^u \left(\int_t^u f(y-t)e^{\mu h_1(t,y)}h_1^2(t,y)dy + \bar{F}(u-t)e^{\mu h_2(t,u)}h_2^2(t,u)\right)dt > 0.$$

Therefore, for given u, $G(\mu)$ is monotone in μ. Besides, $|G'(\mu)| \to \infty$ as $\mu \to \infty$. As a direct consequence, for any x large enough, equation (42) can be easily fit to many standard numerical solvers, and it admits a unique solution. Given μ, the optimal sample path is given by

$$\begin{aligned}\bar{q}(t,y) &= \int_0^t \int_y^\infty \left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(s,w)\right)dwds \\ &= \int_0^t \left(\int_{y\wedge u}^u f(w-s)e^{\mu(u)h_1(s,w)}dw + \int_{u\vee y}^\infty f(w-s)e^{\mu(u)h_2(s,u)}dw\right)ds\end{aligned}$$

for $y \geq t$, and hence

$$\begin{aligned}q(t,y) &= \int_0^t \int_{y+t}^\infty \left(-\frac{\partial^2}{\partial t\partial y}\bar{q}(s,w)\right)dwds \\ &= \int_0^t \left(\int_{(y+t)\wedge u}^u f(w-s)e^{\mu(u)h_1(s,w)}dw \right. \\ &\qquad \left. + \int_{u\vee(y+t)}^\infty f(w-s)e^{\mu(u)h_2(s,u)}dw\right)ds\end{aligned}$$

for $y \geq 0$. Note that here we highlight the dependence of μ on u. Moreover, the rate function for the fixed-time probability is

$$\int_0^u \left(\int_t^u f(y-t) e^{\mu(u) h_1(t,y)} (\mu(u) h_1(t,y) - 1) dy + \int_u^\infty f(y-t) e^{\mu(u) h_2(t,u)} (\mu(u) h_2(t,u) - 1) dy \right) dt. \tag{43}$$

The optimal time horizon u over $0 \leq u \leq T$ is chosen to minimize (43).

Now we consider a whole life insurance contract with benefit $b = 1.5$, continuous premium $p = 1$, zero interest rate and time-until-death which follows the uniform distribution on $[0, 1]$. Our goal is to compute the optimal sample path for ruin before time $T = 1$, where we set $x = 10$ as the ruin threshold. We solve the constraint equation (42) in Matlab and obtain the optimal $u = 1$ with $\mu = 2.251$.

In this case, we can compute the optimal path

$$q(t,y) = \frac{1}{\mu^2}(e^{\mu b - \mu y} - e^{\mu b - \mu y - \mu t} - e^{\mu b - \mu + \mu t} + e^{\mu b - \mu}) + \frac{t}{\mu} e^{-\mu + \mu t} - \frac{1}{\mu^2}(e^{-\mu + \mu t} - e^{-\mu}), \text{ for } y + t \leq 1,$$

and

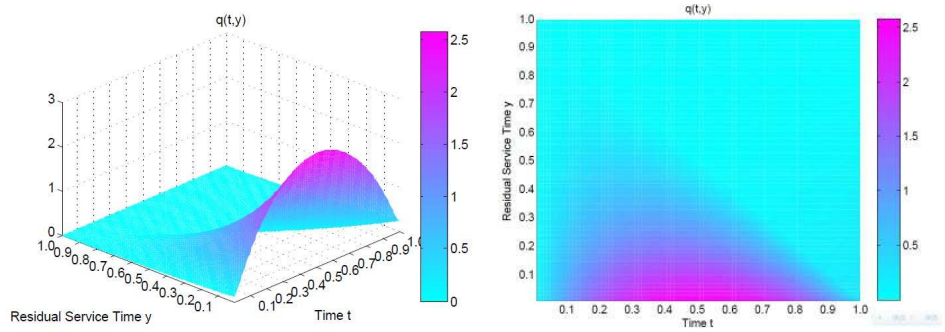
$$q(t,y) = e^{-\mu(2-t-y)} \left(\frac{1-y}{\mu} e^{\mu - \mu y} - \frac{1}{\mu^2}(e^{\mu - \mu y} - 1) \right), \text{ for } y + t > 1.$$

These optimal paths to ruin are shown in Figure 5. We just show the conditional paths, as the unconditional path are identical to the figures illustrated in Example 1. The optimal path here is qualitatively very different from that of Example 1. The value of $Q(t,y)$ is the largest midway between time 0 and 1. Intuitively, it is because it requires the smallest "energy", or distortion from the law of large numbers, at such time point in contributing to a large cash outflow from the insurer.

APPENDIX: CONSTRUCTION OF AN AUXILIARY CONTINUOUS PROCESS

In this section we provide the explicit construction of the process $(Q^*_\lambda(t,y) : 0 \leq t \leq T, y \geq 0)$ introduced in Section 2.3 in order to define our exponentially equivalent continuous process, $\tilde{Q}_\lambda$.

The construction will be based on polygonal interpolations, so it will be convenient to introduce some notation.



(a) $Q(t, y)$ for the most likely path to ruin (surface)

(b) $Q(t, y)$ for the most likely path to ruin (contour)

FIG 5. *The surface and the corresponding contour plot of the asymptotic most likely path to ruin in a portfolio of life insurance policies.*

First, given (t, y) and (t', y') where $t \neq t'$ we write $Q_\lambda(t, y) \leftrightarrow Q_\lambda(t', y')$ to denote the straight line that joins the points $(t, y, Q_\lambda(t, y))$ and $(t', y', Q_\lambda(t', y'))$ in the associated three-dimensional space.

Now, given a sample path of the process $Q_\lambda(\cdot)$, consider the set $\{t_1, \ldots, t_m\}$ of points corresponding to either arrivals or departures in the interval $[0, T]$ (in increasing order); and put $t_0 = 0$ and $t_{m+1} = T$. First let us consider $Q_\lambda(t, \cdot)$ for a fixed time $t \in \{t_0, \ldots, t_{m+1}\}$. Let $\{y_1(t), \ldots, y_{n(t)}(t)\}$ be the set of discontinuities of the function $Q_\lambda(t, \cdot)$ with $n(t)$ equal the number of customers present in the system at time t (recall again that $Q_\lambda(t, \cdot)$ is a right continuous non-increasing step function). Interpolate using straight lines forming the segments $Q_\lambda(t, 0) \leftrightarrow Q_\lambda(t, y_1(t))$, $Q_\lambda(t, y_1(t)) \leftrightarrow Q_\lambda(t, y_2(t)), \ldots, Q_\lambda(t, y_{n(t)-1}(t)) \leftrightarrow Q_\lambda(t, y_{n(t)}(t))$.

The next step is to join the end points of these straight lines to the end points of adjacent (suitably matched in the time axis) end points of straight lines in order to form segments of adjacent planes. In order to do this matching note that for each successive t_i and t_{i+1}, either $Q_\lambda(t_{i+1}, \cdot)$ has one less discontinuous point than $Q_\lambda(t_i, \cdot)$ (i.e. a departure occurs at t_{i+1}) or one more discontinuity point (i.e. an arrival occurs at t_{i+1}); the exception is the last segment from t_m to $t_{m+1} = T$, where there might be no difference between the number of discontinuity points between $Q_\lambda(t_m, \cdot)$ and $Q_\lambda(t_{m+1}, \cdot)$. Note that batch arrivals are not possible since the interarrival times are positive.

According to the notation introduced earlier for discontinuity points, $y_1(t_i), \ldots, y_{n(t_i)}(t_i)$ are the discontinuous points of $Q_\lambda(t_i, \cdot)$ with corresponding values $Q_\lambda(t_i, y_1(t_i))$, $Q_\lambda(t_i, y_2(t_i)), \ldots, Q_\lambda(t_i, y_{n(t_i)}(t_i))$. We will explain how to joint discontinuity points of $Q_\lambda(t_i, \cdot)$ with those from $Q_\lambda(t_{i+1}, \cdot)$.

Suppose a departure occurs at time t_{i+1}. Then we can label the discontinuous points of $Q_\lambda(t_{i+1},\cdot)$ as $y_1(t_{i+1}),\ldots,y_{n(t_{i+1})}(t_{i+1})$, with $n(t_{i+1})=n(t_i)-1$. We form a set of straight lines $Q_\lambda(t_i,0)\leftrightarrow Q_\lambda(t_{i+1},0)\leftrightarrow Q_\lambda(t_i,y_1(t_i))\leftrightarrow Q_\lambda(t_{i+1},y_1(t_{i+1}))\leftrightarrow Q_\lambda(t_i,y_2(t_i))\leftrightarrow Q_\lambda(t_{i+1},y_2(t_{i+1}))\leftrightarrow\cdots\leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1}))\leftrightarrow Q_\lambda(t_i,y_{n(t_i)}(t_i))$ in a zig-zag manner; together with another set of straight lines $Q_\lambda(t_i,0)\leftrightarrow Q_\lambda(t_i,y_1(t_i))\leftrightarrow\cdots\leftrightarrow Q_\lambda(t_i,y_{n(t_i)}(t_i))$, and also the set of straight lines $Q_\lambda(t_{i+1},0)\leftrightarrow Q_\lambda(t_{i+1},y_1(t_{i+1}))\leftrightarrow\cdots\leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1}))$. These three sets describe a series of adjacent triangular planar sections which jointly form a continuous surface.

Similarly, suppose that an arrival occurs at time t_{i+1}. Then we can label the discontinuous points of $Q_\lambda(t_{i+1},\cdot)$ as $Q_\lambda(t_{i+1},y_1(t_{i+1})),\ldots,Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1}))$, with $n(t_{i+1})=n(t_i)+1$. We then form the set of straight lines $Q_\lambda(t_{i+1},0)\leftrightarrow Q_\lambda(t_i,0)\leftrightarrow Q_\lambda(t_{i+1},y_1(t_{i+1}))\leftrightarrow Q_\lambda(t_i,y_1(t_i))\leftrightarrow Q_\lambda(t_{i+1},y_2(t_{i+1}))\leftrightarrow\cdots\leftrightarrow Q_\lambda(t_i,y_{n(t_i)}(t_i))\leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1}))$. Again, together with a second set of straight lines $Q_\lambda(t_i,0)\leftrightarrow Q_\lambda(t_i,y_1(t_i))\leftrightarrow\cdots\leftrightarrow Q_\lambda(t_i,y_{n(t_i)}(t_i))$, and a third set of straight lines, namely $Q_\lambda(t_{i+1},0)\leftrightarrow Q_\lambda(t_{i+1},y_1(t_{i+1}))\leftrightarrow\cdots\leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1}))$. These three sets of straight lines, once again describe a series of adjacent triangular planar sections which jointly form a continuous surface. The last time interval from t_m to T is dealt with similarly, with perhaps one less triangle formed if $n(t_m)=n(T)$.

The continuous function $(Q^*_\lambda(t,y):0\le t\le T, y\ge 0)$ is defined by concatenating all these adjacent triangular planar regions as one varies t_i and t_{i+1} for $i\in\{0,1,\ldots,m\}$, and setting $Q^*_\lambda(t,y)=0$ for the region where y is beyond the boundary of the last triangular plane i.e. beyond the lines $Q_\lambda(t_i,y_{n(t_i)}(t_i))\leftrightarrow Q_\lambda(t_{i+1},y_{n(t_{i+1})}(t_{i+1})), i\in\{0,1,\ldots,m\}$. It is immediate from the previous construction, and the fact that $Q_\lambda(t,\cdot)$ is non-increasing, that $Q^*_\lambda(t,\cdot)$ is also non-increasing for each $t\in[0,T]$.

Acknowledgement. This research was partially supported by the grants DMS-0806145, CMMI-0846816 and CMMI-1069064.

REFERENCES

Asmussen, S. and Albrecher, H. (2010). *Ruin Probabilities*, second ed. World Scientific, New Jersey, US. MR2766220

Decreusefond, L. and Moyal, P. (2008). A functional central limit theorem for the M/GI/∞ queue. *Annals of Applied Probability* **18** 2156–2178. MR2473653

Dembo, A. and Zeitouni, O. (1998). *Large Deviations Techniques and Applications*, Second ed. Springer, New York. MR1619036

Feng, J. and Kurtz, T. (2006). *Large Deviations for Stochastic Processes* **131**. American Mathematical Society. MR2260560

Glynn, P. (1995). Large deviations for the infinite server queue in heavy traffic. In *Stochastic Networks* (F. P. Kelly and R. J. Williams, eds.). *Lecture Notes in Statistics* **71** 387–395. Springer, New York. MR1381021

GLYNN, P. and WHITT, W. (1991). A new view of the heavy-traffic limit theorem for the infinite-server queue. *Annals of Applied Probability* **19** 2211–2269. MR1091098

GLYNN, P. and WHITT, W. (1994). Large deviations behavior of counting processes and their inverses. *Queueing Systems* **17** 107–128.

HALFIN, S. and WHITT, W. (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research* **29** 567–588. MR0629195

IGLEHART, D. (1965). Limit diffusion approximations for the many server queue and the repairman problem. *Journal of Applied Probability* **2** 429–441. MR0184302

JELENKOVIC, P., MANDELBAUM, A. and MOMCILOVIC, P. (2004). Heavy traffic limits for queues with many deterministic servers. *Queueing Systems: Theory and Applications* **47** 53–69. MR2074672

KASPI, H. and RAMANAN, K. (2010). SPDE limits of many server queues. *Preprint arXiv:1010.0330*. MR3059233

KASPI, H. and RAMANAN, K. (2011). Law of large numbers limits for many-server queues. *Annals of Applied Probability* **21** 33–114. MR2759196

LÉONARD, C. (2000). Large deviations for Poisson random measures and processes with independent increments. *Stochastic Processes and Their Applications* **85** 93–121. MR1730616

PANG, G. and WHITT, W. (2010). Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems: Theory and Applications* **65** 325–364. MR2671058

PUHALSKII, A. and REIMAN, M. (2000). The multiclass GI/PH/N queue in the Halfin-Whitt regime. *Advances in Applied Probability* **32** 564–595. MR1778580

PUHALSKII, A. and REIMAN, M. (2009). The G/GI/n queue in the Halfn-Whitt regime. *Annals of Applied Probability* **19** 2211–2269. MR2588244

REED, J. and TALREJA, R. (2012). Distribution-valued heavy-traffic limits for the G/GI/∞ queue. *Preprint* http://people.stern.nyu.edu/jreed/Papers/SubmittedVersionDistribution.pdf.

ZAJIC, T. (1998). Rough asymptotics for tandem non-homogeneous M/G/∞ queues via Poissonized empirical processes. *Queueing Systems: Theory and Applications* **29** 161–174. MR1654488

340 S. W. MUDD BUILDING
500 W. 120 STREET
NEW YORK, NEW YORK 10027
E-MAIL: jose.blanchet@columbia.edu

MATHEMATICS TOWER B-148
STONY BROOK, NEW YORK 11794-3600
E-MAIL: xinyun.chen@stonybrook.edu

MCS 226, 111 CUMMINGTON MALL
BOSTON MA 02215
E-MAIL: khlam@bu.edu