Mathematics of Operations Research

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

Rare-Event Simulation for Many-Server Queues

Jose Blanchet, Henry Lam

To cite this article:
Jose Blanchet, Henry Lam (2014) Rare-Event Simulation for Many-Server Queues. Mathematics of Operations Research 39(4):1142-1178. http://dx.doi.org/10.1287/moor.2014.0654

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 39, No. 4, November 2014, pp. 1142–1178
ISSN 0364-765X (print) | ISSN 1526-5471 (online)

http://dx.doi.org/10.1287/moor.2014.0654

Rare-Event Simulation for Many-Server Queues

Jose Blanchet
Department of Industrial Engineering and Operations Research, and Department of Statistics, Columbia University, New York, New York 10027, jose.blanchet@columbia.edu

Henry Lam
Department of Mathematics and Statistics, Boston University, Boston, Massachusetts 02215, khlam@bu.edu

We develop rare-event simulation methodology for the analysis of loss events in a many-server loss system under the quality-driven regime, focusing on the steady-state loss probability (i.e., fraction of lost customers over arrivals) and the behavior of the whole system leading to loss events. The analysis of these events requires working with the full measure-valued process describing the system. This is the first algorithm that is shown to be asymptotically optimal, in the rare-event simulation context, under the setting of many-server queues involving a full measure-valued representation.

Keywords: many-server queues; rare-event simulation; large deviations
MSC2000 subject classification: Primary: 60K25, 65C05; secondary: 60F10, 60F05
OR/MS subject classification: Primary: queues-simulation; secondary: queues-approximations
History: Received January 11, 2012; revised April 8, 2013. Published online in *Articles in Advance* July 2, 2014.

1. Introduction. Although there is vast literature on provably efficient rare-event simulation algorithms for queues with fixed number of servers, few such algorithms exist for queueing systems with the number of servers scaled asymptotically with the incoming traffic, frequently known as many server systems. In models with a single or a fixed number of servers, random walk representations are often used to analyze associated rare events (see for example Siegmund [27], Asmussen [3], Anantharam [2], Sadowsky [26] and Heidelberger [18]). The difficulty in these types of systems arises from the boundary behavior induced by the positivity constraints inherent to queueing systems. Many-server systems are, in some sense, less sensitive to boundary behavior; instead, the challenge in their rare-event analysis lies on the fact that the system description is typically asymptotically infinite dimensional. One of the goals of this paper, broadly speaking, is to propose methodology and techniques that we believe are applicable to a wide range of rare-event problems involving many-server systems. In particular, we will demonstrate how a full Markovian representation, or customarily known as measure-valued representation in the literature, is both necessary and useful for efficient rare-event simulation of steady-state loss probabilities. As far as we know, the algorithm proposed in this paper is the first provably asymptotically optimal algorithm (in a sense that we will explain shortly) that involves such full measure-valued representation in the rare-event simulation literature.

In this paper we focus on the problem of estimating the steady-state loss probability in many-server loss systems. We consider a system with general i.i.d. interarrival times and service times (both under suitable tail conditions). The system has s servers and no waiting room. If a customer arrives and finds a server empty, he/she immediately starts service occupying a server. If the customer finds all the servers busy, he/she leaves the system immediately and the system incurs a "loss" (see Figure 1 for a pictorial description). The steady-state loss probability (i.e., the long term proportion of customers that are lost) is rare if the traffic intensity (arrival rate into the system/total service rate) is less than one and the number of servers is large. This is precisely the asymptotic environment that we consider.

Related large deviations and simulation results include the work of Glynn [16], who developed large deviations asymptotics for the number-in-system of an infinite-server queue with high arrival rates. Based on this result, Szechtman and Glynn [28] developed a corresponding rare-event algorithm for the same quantity of an infinite-server queue, using a sequential tilting scheme that mimics the optimal exponential change of measure. Related results for first passage time probabilities have also been obtained by Ridder [24] in the setting of Markovian queues. Blanchet et al. [9] constructed an algorithm for the steady-state loss probability of a slotted-time $M/G/s$ system with bounded service time. The algorithm in Blanchet et al. is the closest in spirit to our methodology here, but the slotted-time nature, the Markovian structure, and the service times being bounded were used in a crucial way to avoid the main technical complications involved in dealing with measure-valued representations.

In this paper we focus on the steady-state loss estimation of a fully continuous $GI/G/s$ system with service times that accommodate most distributions used in practice, including mixtures of exponential, Weibull, and lognormal distributions. A key element of our algorithm, in addition to the use of a full Markovian representation, is the application of weak convergence limits by Krichagina and Puhalskii [20] and Pang and Whitt [21]. As we

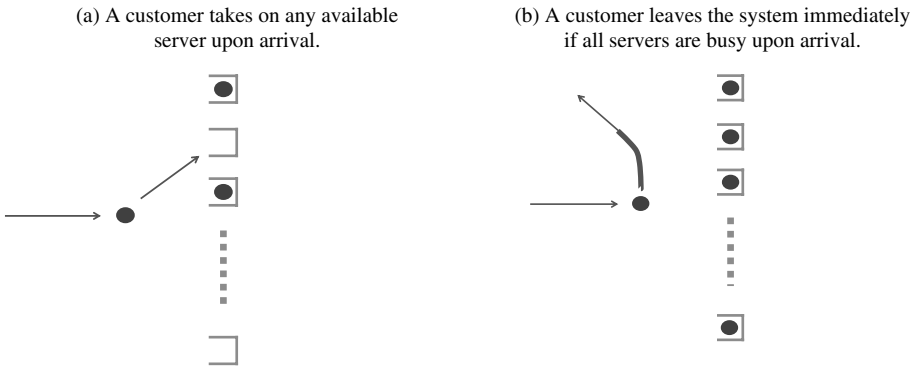

FIGURE 1. Dynamics of many-server loss system.

shall see, the weak convergence results are necessary because via a suitable extension of regenerative-type simulation (see §2.4), the steady-state loss probability of the system can be transformed to a first passage problem of the Markov process starting from an appropriate set, suitably chosen by means of such weak convergence analysis. However, unlike the infinite-server system, the capacity constraint (s servers) introduces a boundary that forces us to work with the sample path and to track the whole process history.

Our main methodology to construct an efficient algorithm is based on importance sampling, which is a variance reduction technique that biases the probability measure of the system (via a so-called change of measure) to enhance the occurrence of the rare event of interest. To correct for the bias, a likelihood ratio is multiplied to the sample output to maintain unbiasedness. The key to efficiency is then to control the likelihood ratio, which is typically small, and hence favorable, when the change of measure resembles the conditional distribution given the occurrence of the rare event. Construction of good changes of measure often draws on associated large deviations theory (see Asmussen and Glynn [5], Chap. 6). We will carry out this scheme of ideas in subsequent sections.

The criterion of efficiency that we will be using is the so-called asymptotic optimality (or logarithmic efficiency). More concretely, suppose we want to estimate some probability $\alpha := \alpha(s)$ that goes to 0 as $s \nearrow \infty$. For any unbiased estimator X of α (i.e., $\alpha = EX$) one must have $EX^2 \geq (EX)^2 = \alpha^2$ by Jensen's inequality. Asymptotic optimality requires that α^2 is also an upper bound of the estimator's variance in terms of exponential decay rate. In other words,

$$\liminf_{s\to\infty} \frac{\log EX^2}{\log \alpha^2} = 1.$$

This implies that the estimator X possesses the optimal exponential decay rate any unbiased estimator can possibly achieve. See, for example, Bucklew [12], Asmussen and Glynn [5], and Juneja and Shahabuddin [19] for further details on asymptotic optimality.

Finally, we emphasize the potential applications of loss estimation in many-server systems. One prominent example is call center analysis. Customer support centers, intracompany phone systems, and emergency rooms, among others, typically have fixed system capacity above which calls would be lost. In many situations losses are rare, yet their implications can be significant. The most extreme example is perhaps a 911 center in which any call loss can be life threatening. In view of this, an accurate estimate (at least to the order of magnitude) of loss probability is often an indispensable indicator of system performance. Although in this paper we focus on i.i.d. interarrival and service times, under mild modifications, our methodology can be adapted to different model assumptions such as Markov-modulation and time inhomogeneity that arise naturally in certain application environments. As a side tale, a rather surprising and novel application of the present methodology is in the context of actuarial loss in insurance and pension funds. In such systems the policyholders (insurance contract or pension scheme buyers) are the "customers," and "loss" is triggered not by an exceedence of the number of customers but rather by a cash overflow of the insurer. Under suitable model assumptions, the latter can be expressed as a functional of the past system history whereby the full Markovian representation becomes valuable. The full development of this application is presented in Blanchet and Lam [7].

The organization of the paper is as follows. In §2 we will indicate our main results and lay out our $GI/G/s$ model assumptions. In §3 we will explain and describe in detail our simulation methodology. Section 4 will focus on the proof of algorithmic efficiency and large deviations asymptotics, whereas §5 will be devoted to the use of weak convergence results mentioned earlier for the design of an appropriate recurrent set. Finally, we will provide numerical results in §6, and technical details are left to the appendix.

2. Main results and contributions. In this section we describe our assumptions and introduce the objects we shall use in this paper. Then we shall discuss our main results. At a general level, our main contribution in this paper is the development of methodology for efficient *rare-event analysis of the steady-state behavior* of many-server systems in a *quality driven regime* (quality driven regime refers to the scenario when the traffic intensity is bounded away from 1 as the number of servers and the arrival rate both grow to infinity). Our methodology, however, is also suitable for transient rare-event analysis assuming the initial condition of the system is within the diffusion scale from the fluid limit of the system.

The main idea of our methodology has four parts, which can be informally summarized as follows. First introduce a coupling with the infinite-server queue (related construction appeared in, e.g., Reed [22]). Second, take advantage of a suitable ratio representation for the associated probability of interest for the system in consideration (in our case a loss system). Third, identify a suitable regenerative-like set based on available results in the literature on diffusion approximations for the system in consideration. A cycle is defined as the period between return times to the regenerative-like set. Finally, the fourth step is to identify a rare-event of interest inside a cycle that is common to both the system in consideration and the infinite-server system. Such rare event of interest must have the same large deviations asymptotics as the probability of interest. It is crucial for the last step to select the regenerative-like set carefully. We concentrate on loss probabilities in this paper, but an almost identical (asymptotically optimal) algorithm can be obtained for the steady-state probability of delay in a many-server queue under the quality driven regime.

Let us now introduce our assumptions on the loss system and develop the four elements outlined in the previous paragraph for the evaluation of steady-state loss probabilities.

2.1. Assumptions on arrivals and service time distribution. Our model of interest is a $GI/G/s$ loss system. There are $s \geq 1$ servers in the system. We assume arrivals follow a renewal process with rate λs; i.e., the interarrival times are i.i.d. with mean $1/(\lambda s)$. More precisely, we introduce a "base" arrival system, with $N^0(t)$, $t \geq 0$ as the counting process of its arrivals from time 0 to t and U_k^0, $k = 0, 1, 2, \ldots$, as the i.i.d. interarrival times with $EU_k^0 = 1/\lambda$ (except the first arrival U_0^0, which can be delayed). We then scale the system so that $N_s(t) := N^0(st)$ is the counting process of the s-th order system, and $U_k := U_k^0/s, k = 0, 1, 2, \ldots$ are the interarrival times. Moreover, we let $A_k, k = 1, 2, \ldots$, be the arrival times; i.e., $A_k := \sum_{i=0}^{k-1} U_i$ (note the convention $U_k = A_{k+1} - A_k$ and $A_0 = 0$). Note that for convenience we have suppressed the dependence on s in U_k and A_k.

We assume that $\kappa_s(\theta) := \log E e^{\theta U_k}$, the logarithmic moment generating function of U_k, is finite for θ in a neighborhood of the origin. It is easy to see that $\kappa_s(\theta) = \kappa^0(\theta/s)$ where $\kappa^0(\theta) := \log E e^{\theta U_k^0}$ is the logarithmic moment generating function of the interarrival time in the base system.

Since $\kappa^0(\cdot)$ is increasing, we can let

$$\psi_N(\theta) := -(\kappa^0)^{-1}(-\theta) \tag{1}$$

where $(\kappa^0)^{-1}(\cdot)$ is the inverse of $\kappa^0(\cdot)$. Note that $\kappa_s^{-1}(\theta) = s(\kappa^0)^{-1}(\theta)$. Also, $\psi_N(\cdot)$ is increasing and convex; this is inherited from $\kappa^0(\cdot)$.

Now we impose a few assumptions on $\psi_N(\cdot)$. We assume that $\psi_N(\cdot)$ is twice continuously differentiable on $\mathbb{R}$, strictly convex, and steep on the positive side; i.e., $\psi_N'(\theta) \nearrow \infty$ as $\theta \nearrow \infty$. Thus $\psi_N'(0) = \lambda$ and $\psi_N'(\mathbb{R}_+) = [\lambda, \infty)$. We also impose the technical condition

$$\theta \frac{d}{d\theta} \log \psi_N(\theta) \to \infty \tag{2}$$

as $\theta \nearrow \infty$. This condition is satisfied by many common interarrival distributions, such as exponential, Gamma, Erlang, etc.

Under these assumptions we have for any $0 = t_0 < t_1 < \cdots < t_m < \infty$ and $\theta_1, \ldots, \theta_m \in \mathbb{R}$,

$$\frac{1}{s} \log E \exp\left\{ \sum_{i=1}^{m} \theta_i (N_s(t_i) - N_s(t_{i-1})) \right\} \to \sum_{i=1}^{m} \psi_N(\theta_i)(t_i - t_{i-1}) \tag{3}$$

as $s \nearrow \infty$. In particular, $\psi_N(\cdot)t$ is the so-called *Gartner-Ellis limit* of $N_s(t)$ for any $t > 0$ as $s \nearrow \infty$ (see Glynn and Whitt [17] and Glynn [16]). In the case of Poisson arrival, for example, the interarrival times are exponential and we have $\kappa(\theta) = \log(\lambda/(\lambda - \theta))$. This gives $\psi_N(\theta) = \lambda(e^\theta - 1)$.

We now state our assumptions on the service times. Denote V_k as the service time of the k-th arriving customer, and let V_k, $k = 1, 2, \ldots$ be i.i.d. with distribution function $F(\cdot)$ and tail distribution function $\bar{F}(\cdot)$. We assume that $F(\cdot)$ has a density $f(\cdot)$ that satisfies

$$\lim_{y \to \infty} y h(y) = \infty \tag{4}$$

where $h(y) := f(y)/\bar{F}(y)$ is the hazard rate function (with the convention that $h(y) = \infty$ whenever $\bar{F}(y) = 0$). In particular, (4) implies that for any $p > 0$ we can find $a > 0$ such that $yh(y) > p$ as long as $y > a$. Hence,

$$\bar{F}(y) = e^{-\int_0^y h(u)\,du} \le c_1 e^{-\int_a^y p/u\,du} = \frac{c_2}{y^p} \tag{5}$$

for some c_1, $c_2 > 0$. In other words, $\bar{F}(\cdot)$ decays faster than any power law. It is worth pointing out that assumption (4) covers Weibull and lognormal service times, which have been observed to be important models in call center analysis (see, e.g., Brown et al. [11]).

Note that service time distribution does not scale with s. Hence the traffic intensity, defined by the ratio of arrival rate to service rate, is λEV (we sometimes drop the subscript k of V_k for convenience). We assume that $\lambda EV < 1$. This corresponds to a *quality-driven regime* and implies that loss is rare. We will see the importance of this assumption in our derivation of efficiency and large deviations results in §4.

2.2. Representation of system status. Let $Q(t)$ be the number of customers in the $GI/G/s$ system at time t. More generally, we let $Q(t, y)$ be the number of customers at time t who have residual service time larger than y, where residual service time at time t for the k-th customer is given by $(V_k + A_k - t)^+$ (defined for customers that are not lost). We also keep track of the age process $B(t) = \inf\{t - A_k\colon A_k \le t\}$ i.e., the time elapsed since the last arrival. We assume right-continuous sample path (i.e., customers who arrive at time t and start service are considered to be in the system at time t, while those who finish their service at time t are outside the system at time t). We also make the assumption that service time is assigned and known upon arrival of each served customer. Although not necessarily true in practice, this assumption does not alter any output from a simulation point of view as far as estimation of loss probabilities is concerned. Figure 2 illustrates a typical realization of $Q(t, \cdot)$ at a time t. To have a Markov process, we let $W_t := (Q(t, \cdot), B(t)) \in \mathcal{D}[0, \infty) \times \mathbb{R}_+$ as the state of the process at time t, where $\mathcal{D}[0, \infty)$ denotes the set of right-continuous-with-left-limit (RCLL) functions defined on $[0, \infty)$. In the case of bounded service time over $[0, M]$ the state-space is further restricted to $\mathcal{D}[0, M] \times \mathbb{R}_+$, where $\mathcal{D}[0, M]$ is the set of RCLL functions defined on $[0, M]$.

2.3. A coupling $GI/G/\infty$ system. As indicated briefly before, multiple times in this paper we shall use a $GI/G/\infty$ system that is naturally coupled with the $GI/G/s$ system under the above assumptions. This $GI/G/\infty$ system has the same arrival process and service time distribution as the $GI/G/s$ system but has infinite number of servers and thus no loss can occur. Furthermore, it labels s of its servers from the beginning. When customer arrives, he would choose one of the idle labeled servers in preference to the rest and only choose unlabeled server if all the s labeled servers are busy. It is then easy to see that the evolution of the $GI/G/\infty$ system restricted to the s labeled servers follows exactly the same dynamic of the $GI/G/s$ system that we are considering. In this paper we shall use the superscript "∞" to denote quantities in the $GI/G/\infty$ system; for example, $Q^\infty(t)$ denotes the number of customers at time t for the $GI/G/\infty$ system, and so on.

Throughout the paper we also use overline to denote quantities that exclude the initial customers. So for example $\bar{Q}^\infty(t, y)$ denotes the number of customers who arrive after time 0 in the $GI/G/\infty$ system and are present at time t having residual service time larger than y; i.e., $\bar{Q}^\infty(t, y) = Q^\infty(t, y) - Q^\infty(0, t + y)$.

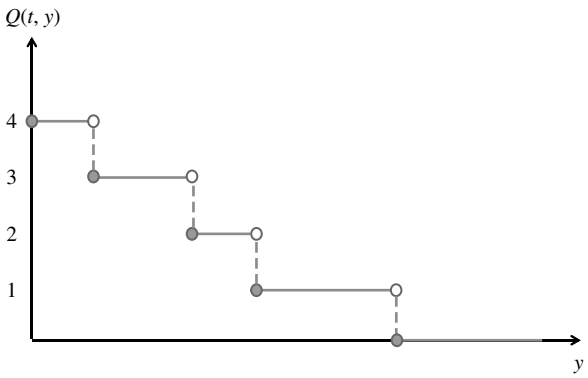

FIGURE 2. A typical realization of $Q(t, \cdot)$; note that $Q(t) = 4$ in this realization.

2.4. Ratio representation for steady-state loss probabilities. Our quantity of interest is the steady-state loss probability, defined as

$$P_\pi(\text{loss}) := \lim_{T\to\infty} \frac{\text{number of losses up to } T}{\text{number of arrivals up to } T}, \tag{6}$$

where π denotes the stationary measure (the existence and uniqueness of the steady-state loss probability, as defined in (6), can be seen by regenerative argument; see Foss and Kalashnikov [15], Example 5). Next, Kac's formula (see Breiman [10]) allows to express the loss probability as

$$P_\pi(\text{loss}) = \frac{E_{\mathscr{A}} N_{\mathscr{A}}}{\lambda s E_{\mathscr{A}} \tau_{\mathscr{A}}}, \tag{7}$$

where $\mathscr{A}$ is a set that is visited by the chain infinitely often, which we call a *recurrent set*. The expectation $E_{\mathscr{A}}[\cdot]$ denotes the expectation with initial state distributed according to the steady-state distribution conditioned on being in $\mathscr{A}$. The quantity $N_{\mathscr{A}}$ is the number of loss before returning to set $\mathscr{A}$, $\tau_{\mathscr{A}}$ is the return time to $\mathscr{A}$, and λs is the arrival rate. Depending on the choice of $\mathscr{A}$, the quantities $E_{\mathscr{A}} N_{\mathscr{A}}$ and $E_{\mathscr{A}} \tau_{\mathscr{A}}$ can be dependent on the parameter s. We shall use the term "$\mathscr{A}$-cycle" to refer to the process from one instance of return on $\mathscr{A}$ to another return on $\mathscr{A}$ on suitably defined lattice points.

Our choice of $\mathscr{A}$ will be given shortly. Before so, let us first explain the difficulty of our problem to motivate our subsequent choice, and along the way we also summarize our simulation approach.

Note that formula (7) provides a basis for regenerative-type simulation (see Asmussen and Glynn [5], Chap. 4). Supposing one can identify a recurrent set $\mathscr{A}$, a straightforward crude Monte Carlo strategy would be to run the system for a long time from some initial state, take a record of $N_{\mathscr{A}}$ and $\tau_{\mathscr{A}}$ every time it hits $\mathscr{A}$, and output the sample means of $N_{\mathscr{A}}$ and $\tau_{\mathscr{A}}$. This strategy is valid as long as the running time is long enough to allow for the system to be close to stationarity. Moreover, this strategy is basically the same as merely outputting the number of loss events divided by the run time times λs (excluding the uncompleted last $\mathscr{A}$-cycle).

However, recognizing that loss is a rare event (with exponential decay rate in s as we will show as a by-product of our analysis), this method will take an exponential amount of time in s to get a specified relative error. This is regardless of the choice of $\mathscr{A}$: if $\mathscr{A}$ is large, it takes short time to regenerate (i.e., $\tau_{\mathscr{A}}$ is small, and consequently the number of losses reported as the numerator $E_{\mathscr{A}} N_{\mathscr{A}}$ of (7) is almost always zero), whereas if $\mathscr{A}$ is small, it takes a long time to regenerate. To dramatically speed up the computation time, our strategy is the following. We choose $\mathscr{A}$ to be a "central limit" set so that $E_{\mathscr{A}} \tau_{\mathscr{A}}$ is not exponentially large in s (and not exponentially small either). This isolates the rarity of loss to the numerator $E_{\mathscr{A}} N_{\mathscr{A}}$. In other words, it is very difficult for the process to reach overflow in an $\mathscr{A}$-cycle. *The key, then, is to construct an efficient importance sampling scheme to induce overflow and to estimate the number of losses in each $\mathscr{A}$-cycle.*

We point out two practical observations using this approach: First, $\tau_{\mathscr{A}}$ and $N_{\mathscr{A}}$ can be estimated separately; i.e., one can "split" the process every time it hits $\mathscr{A}$ in two processes: one to which we apply importance sampling to get one sample of $N_{\mathscr{A}}$ and is then discarded; to the other one, we apply the original measure to get one sample of $\tau_{\mathscr{A}}$ and also set the initial position for the next $\mathscr{A}$-cycle (see Asmussen and Glynn [5], Chap. 4). Secondly, to get an estimate of standard deviation, one has to use batch estimates since the samples obtained this way possess serial correlations (Asmussen and Glynn [5], Chap. 4). In other words, one has to divide the simulated chain into several segments of equal number of time units. Then an estimate of the steady-state loss probability is computed from each chain segment. These estimates are regarded as independent samples of loss probability. The details of batch sampling will be provided in §6 when we discuss numerical results.

We summarize our approach as follows:

Algorithm 1

(i) Choose a recurrent set $\mathscr{A}$. Initialize the $GI/G/s$ queue's status as any point in $\mathscr{A}$.

(ii) Run the queue. Each time the queue hits a point in $\mathscr{A}$, say x, do the following: Starting from x,

(a) Use importance sampling to sample one $N_{\mathscr{A}}$, the number of loss in a cycle.

(b) Use crude Monte Carlo to sample one $\tau_{\mathscr{A}}$, the return time. The final position of this queue is taken as the new x.

(iii) After running the $GI/G/s$ system for a sufficiently long time applying (ii), divide the queue into several segments of equal time length. Compute the estimate of steady-state loss probability applying the batch samples using the ratio (7).

2.5. Recurrent set. We now describe our recurrent set $\mathscr{A}$. First of all, note that one can pick $T = n\Delta$ for some $\Delta > 0$ and $n \in \mathbb{N}$ in the definition of loss probability given by Equation (6) and send $n \to \infty$. The introduction of the lattice of size Δ helps to define return times to the set $\mathscr{A}$ only at lattice points. Let us pick a fixed small time interval Δ (one choice, for example, is say 1/5 of the mean of service time) and define

$$\tau_{\mathscr{A}} := \inf\{t | t = \Delta n, n \in \mathbb{N}, Q(t, \cdot) \in \mathscr{A}\}.$$

The set $\mathscr{A}$ is defined to be

$$\mathscr{A} := \{\omega \in \mathscr{D}[0, \infty) \colon \omega(y) \in J(y), y \in \mathbb{R}_+\}. \tag{8}$$

Here $J(y)$ is the interval

$$J(y) := \left(\lambda s \int_y^\infty \bar{F}(u)\, du - \sqrt{s} C^* \xi(y),\ \lambda s \int_y^\infty \bar{F}(u)\, du + \sqrt{s} C^* \xi(y)\right) \tag{9}$$

for some well-chosen constant $C^* > 0$ (discussed in Remark 2.1 below and in §5) and

$$\xi(y) := \nu(y) + \gamma \int_y^\infty \nu(u)\, du \tag{10}$$

where

$$\nu(y) := \left(\lambda \int_y^\infty \bar{F}(u)\, du\right)^{1/(2+\eta)} \tag{11}$$

with any constants $\eta,\ \gamma > 0$.

The form of $J(y)$ comes from the heavy traffic limit of $GI/G/\infty$ queue. Pang and Whitt [21] proved the fluid limit $Q^\infty(t, y)/s \to \lambda \int_y^{t+y} \bar{F}(u)\, du$ a.s. and the diffusion limit $(Q^\infty(t, y) - \lambda s \int_y^{t+y} \bar{F}(u)\, du)/\sqrt{s} \Rightarrow R(t, y)$ for some Gaussian process $R(t, y)$ with $\mathrm{var}(R(t, y)) \to \lambda c_a^2 \int_y^\infty \bar{F}(u)^2\, du + \lambda \int_y^\infty F(u)\bar{F}(u)\, du$ as $t \to \infty$, where c_a is the coefficient of variation of the interarrival times in the base system, namely U_k^0. Our recurrent set $\mathscr{A}$ is thus a "confidence band" of the steady state of $Q^\infty(t, y)$, with the width of the confidence band decaying slower than the standard deviation of $Q^\infty(\infty, y)$ as $y \to \infty$. Via a coupling argument, it can be proved (see Proposition 2.1) that this choice of $\mathscr{A}$ indeed leads to a return time for the $GI/G/s$ system that is subexponential in s. The slower decay rate of the confidence band width is a technical adjustment to enlarge $\mathscr{A}$ so that such a subexponential (in s) return time for the $GI/G/s$ system is guaranteed. In fact, for the case of bounded service time, it suffices to set $\eta = 0$.

2.6. Main results. The main result of this paper is the construction and the asymptotic optimality proof of an efficient importance sampling scheme to simulate $N_{\mathscr{A}}$. In order to show the optimality of the algorithm, on our way, we obtain large deviations asymptotics for loss probabilities that might be of independent interest.

THEOREM 2.1. *Under the assumptions in §2.1, the estimator using the recurrent set $\mathscr{A}$ in (8) and the importance sampler given by Algorithm 2 is asymptotically optimal. Moreover, the steady-state loss probability (7) can be seen to be exponentially decaying in s with decay rate I^* defined in (19).*

An important novel feature of the problem we consider (and our solution) is that it requires a construction based on a full Markovian representation of the process. Intuitively, the steady-state loss probability of the $GI/G/s$ system depends on its loss behavior starting from a "normal" or "typical" state under stationarity (which can be identified via a diffusion limit). It turns out that the loss behavior can vary substantially if one defines this initial "normal" state only through the system's queue length (even though a loss event is defined only through the queue length). However, by defining the "normal" state through the full Markovian representation of the system (which includes tracking the residual service time for each of the current customer), the loss behavior starting from this state is characterized by a natural optimal path in the large deviations sense, and as a result we can identify the efficient importance sampling scheme to induce such losses. These observations ultimately translate to the need of a recurrent set $\mathscr{A}$ that is also defined via the full Markovian representation of the system in the simulation of $E_{\mathscr{A}} N_{\mathscr{A}}$ in (7).

We next point out two further methodological observations. First, our importance sampling algorithm utilizes the representation of the coupled $GI/G/\infty$ as a point process. This point process representation can also be used to prove results on sample path large deviations for many-server systems; such development will be reported in Blanchet et al. [8]. Secondly, our algorithm requires essentially the information of the whole sample path of the system because of the introduction of an auxiliary random time that is independent of the system in the algorithm. This random time, as will be discussed in detail in §3.3, is important in establishing the efficiency of our algorithm and will render a likelihood ratio that is measurable with respect to the space of sample paths. This is in sharp contrast to the algorithm proposed in Szechtman and Glynn [28] for estimating fixed-time probability.

Finally, the recurrent set $\mathscr{A}$, given by (8), can be seen to possess the following properties:

PROPOSITION 2.1. *In the $GI/G/s$ system,*

$$\lim_{s\to\infty} \frac{1}{s} \log E_{\mathscr{A}} \tau_{\mathscr{A}}^p = 0 \tag{12}$$

and

$$\limsup_{s\to\infty} \frac{1}{s} \log E_{\mathscr{A}} N_{\mathscr{A}}^p \le 0 \tag{13}$$

for any $p > 0$.

Briefly stated, Proposition 2.1 stipulates that any moments of the time length and number of losses of an $\mathscr{A}$-cycle are subexponential in s. When $p = 1$, it in particular states that the expected time length of a cycle is subexponential in s. As discussed above, this isolates the rarity of loss to the numerator in (7) and ensures the validity of Algorithm 1. The result on general p in Proposition 2.1 is also used in the optimality proof of the importance sampling (as will be seen in §4). Interestingly, the proof of Proposition 2.1 requires the use of the Borell-TIS inequality for Gaussian random fields (Adler [1]). The connection to Gaussian random fields arises in the diffusion limit of the coupled $GI/G/\infty$ queue.

We close this section with two remarks on $\mathscr{A}$.

REMARK 2.1. The interval $J(y)$ in the definition of $\mathscr{A}$ in (9) contains a nonnegative integer for any value of y if C^* is chosen large enough. In fact, observe that the length of $J(y)$ is continuous and decreasing in y, and let

$$l(s) := \sup\{y > 0\colon \sqrt{s} C^* \xi(y) \ge 1/2\}. \tag{14}$$

If y is such that the width of $J(y)$ is equal to 1 (equivalently $y = l(s)$) we have that the center of $J(y)$, namely, $\lambda s \int_y^\infty \bar{F}(u)\,du$, satisfies

$$0 \le \lambda s \int_y^\infty \bar{F}(u)\,du \le (\lambda/(C^*)^{2+\eta})(\sqrt{s} C^* \xi(y))^{2+\eta}/s^{\eta/2} = (\lambda/(C^*)^{2+\eta})(1/2)^{2+\eta}/s^{\eta/2}.$$

The right-hand side is less than 1/2 for $(C^*)^{2+\eta} \ge \lambda$ and this implies that $\{0\} \subset J(y)$ for $y = l(s)$. Now, if $y > l(s)$, we can ensure that the half-width of $J(y)$, namely, $\sqrt{s} C^* \xi(y)$, is larger than the center, if C^* is chosen sufficiently large. To see this, note that a sufficient condition is that

$$\lambda s \int_y^\infty \bar{F}(u)\,du \le \sqrt{s} C^* \left(\lambda \int_y^\infty \bar{F}(u)\,du\right)^{1/(2+\eta)}$$

which is equivalent to

$$s^{1/2} \left(\int_y^\infty \bar{F}(u)\,du\right)^{(1+\eta)/(2+\eta)} \le C^* \lambda^{-(1+\eta)/(2+\eta)}$$

or

$$s^{(1+\eta/2)/(1+\eta)} \int_y^\infty \bar{F}(u)\,du \le (C^*)^{(2+\eta)/(1+\eta)} \lambda^{-1}.$$

Now, choosing $C^* \ge \max(\lambda, 1)$, we have, for $y > l(s)$,

$$s^{(1+\eta/2)/(1+\eta)} \int_y^\infty \bar{F}(u)\,du \le s^{1+\eta/2} \int_y^\infty \bar{F}(u)\,du \le 1/(C^*)^{2+\eta}(1/2)^{2+\eta} \le (C^*)^{(2+\eta)/(1+\eta)} \lambda^{-1}$$

which gives the required implication. So $\{0\} \subset J(y)$ for $y > l(s)$. Obviously it includes at least one point when $y < l(s)$ (because the width of $J(y)$ is larger than 1). Therefore $J(y)$ always contains a nonnegative integer for any $y \ge 0$, and the recurrent set $\mathscr{A}$ is hence well defined.

REMARK 2.2. One may ask whether it is possible to define $\mathscr{A}$ in a finite-dimensional fashion, instead of introducing the functional "confidence band" in (8). For example, one may divide the domain of y into segments $[y_i, y_{i+1})$, $i = 0, 1, 2, \ldots, r(s) - 1$ for some integer $r(s)$ with $y_0 = 0$ and $y_{r(s)} = \infty$, where the length of each segment can be dependent on s and nonidentical. One then defines the recurrent set as $\{Q(t,\cdot)\colon Q(t, y_i) - Q(t, y_{i+1}) \in \mathscr{A}_i \text{ for } i = 0, \ldots, r(s) - 1\}$ for some well-defined sets $\mathscr{A}_i$'s. As we will see in the arguments in the subsequent sections, the important criteria of a good recurrent set are (1) it consists of a significantly large region in the central limit theorem, so that it is visited often enough, and (2) its deviation from the mean of $Q(t, y)$ is small, in

the sense that the distance between any element in this recurrent set and the mean of the steady state of $Q(t, y)$, at every $y \in [0, \infty)$, has order $o(s)$. Criterion (2) is important; otherwise, the large deviations of loss starting from two different elements in the recurrent set can be substantially different. We want to avoid having to consider several substantially different paths that can contribute to the loss event in a significant way as having such variability would complicate the design of the importance sampling estimator.

Keeping criterion (2) in mind, we conclude that it is important to fine-tune the scale of the segments $[y_i, y_{i+1})$ to preserve the efficiency of the algorithm. This suggests that a reasonable description of the recurrent set would involve a dimension that grows at a suitable rate as $s \to \infty$, thereby effectively obtaining a set of the form that we propose. The functional definition of $\mathcal{A}$ in (8) happens to balance both criteria (1) and (2).

3. Simulation methodology.

As discussed, the key idea in our simulation procedure consists of an importance sampling algorithm. We will now present this in detail.

3.1. Overview of the algorithm.

First we shall explain some heuristic in constructing the algorithm. As we discussed earlier, the choice of $\mathcal{A}$ isolates the rarity of steady-state loss probability to $E_{\mathcal{A}}N_{\mathcal{A}}$, which in turn is small because of the difficulty in approaching overflow from $\mathcal{A}$. So on an exponential scale, $E_{\mathcal{A}}N_{\mathcal{A}} \approx P_{\mathcal{A}}(\tau_s < \tau_{\mathcal{A}})$, where $P_{\mathcal{A}}(\cdot)$ is the probability measure with initial state distributed as the steady-state distribution conditional on $\mathcal{A}$, and $\tau_s = \inf\{t > 0\colon Q(t) > s\}$ is the first passage time to overflow. Observe that the probability $P_{\mathcal{A}}(\tau_s < \tau_{\mathcal{A}})$ is identical for $GI/G/s$ and the coupled $GI/G/\infty$ system since the systems are identical before τ_s. The key idea is to leverage our knowledge of the structurally simpler $GI/G/\infty$ system. In fact, one can show that the greatest contribution to $P_{\mathcal{A}}(\tau_s < \tau_{\mathcal{A}})$ is the probability $P_{\mathcal{A}}(Q^{\infty}(t^*) > s)$ for some optimal time t^*, whereas the contribution by other times is exponentially smaller.

In view of this heuristic, one may think that the most efficient importance sampling scheme is to exponentially tilt the process as if we are interested in estimating the probability $P_{\mathcal{A}}(Q^{\infty}(t^*) > s)$. However, doing so does not guarantee a small "overshoot" of the process at τ_s. Instead, we introduce a randomized time horizon following the idea of Blanchet et al. [9]. The likelihood ratio will then comprise a mixture of individual likelihood ratios under different time horizons and a bound on the overshoot is attained by looking at the right horizon (namely, $\lceil \tau_s \rceil$ as explained in §4).

Hence our algorithm will take the following steps. Suppose we start from some position in $\mathcal{A}$. First we sample a randomized time horizon with some well-chosen distribution. Then we tilt the coupled $GI/G/\infty$ process to target overflow over this realized time horizon, i.e., as if we are estimating $P_{\mathcal{A}}(Q^{\infty}(K) > s)$ for the realized time horizon K. This involves sequential tilting of both the arrivals and service times. Once overflow is hit, we switch back to the $GI/G/s$ system, drop the lost customers, and change back to the arrival rate and service times under the original measure to run the $GI/G/s$ system until $\mathcal{A}$ is reached. At this time one sample of $N_{\mathcal{A}}$ is recorded together with the likelihood ratio.

The key questions now are how to determine (1) the sequential tilting scheme of arrivals and service times given a realized time horizon, (2) the distribution of the random time, and (3) the likelihood ratio associated with this mixture scheme. In the following we will explain these ingredients in detail and then lay out our algorithm. The proof of efficiency will be deferred to §4.

3.2. Sequential tilting scheme.

Denote $P_r(\cdot)$ and $E_r[\cdot]$ as the probability measure and expectation with initial system status $r \in \mathcal{D}[0, \infty)$ (so that $r(y)$ is the number of initial customers still in the system at time y). Suppose we want to estimate $P_r(Q^{\infty}(t) > s)$ efficiently for a $GI/G/\infty$ system as $s \nearrow \infty$, where $r \in \mathcal{A}$. An important clue is the use of the Gartner-Ellis Theorem (see Dembo and Zeitouni [13], p. 44, Theorem 2.3.6) to obtain a large deviations result. Although this may not give an immediate importance sampling scheme, it can suggest the type of exponential tilting needed that can be verified to be efficient. This is proposed by Glynn [16] and Szechtman and Glynn [28], which we briefly recall here.

To be more specific, let us introduce more notation. Let, for any $t > 0$,

$$\psi_t(\theta) := \int_0^t \psi_N(\log(e^{\theta}\bar{F}(t-u) + F(t-u)))\,du. \tag{15}$$

This is the Gartner-Ellis limit (see, for example, Dembo and Zeitouni [13]) of $\bar{Q}^{\infty}(t)$ since

$$\frac{1}{s}\log E e^{\theta \bar{Q}^{\infty}(t)} = \frac{1}{s}\log E \exp\left\{\theta \sum_{i=1}^{N_s(t)} I(V_i > t - A_i)\right\} \to \int_0^t \psi_N(\log(e^{\theta}\bar{F}(t-u) + F(t-u)))\,du,$$

where $I(\cdot)$ is the indicator function (see Glynn [16] for a proof). It uses (3) and the definition of Riemann sum; alternatively, see Lemma 4.2 in §4 as a generalization of this result. Let us state the following properties of $\psi_t(\cdot)$ for later convenience:

LEMMA 3.1. *$\psi_t(\cdot)$ is defined on $\mathbb{R}$, twice continuously differentiable, strictly convex, and steep.*

Next let $a_t := 1 - \lambda \int_t^\infty \bar{F}(u)\,du$. $r \in \mathcal{A}$ implies that $a_t s + o(s)$ is the number of customers needed excluding the initial ones to reach overflow at time t. In other words,

$$P_r(Q^\infty(t) > s) = P(\bar{Q}^\infty(t) > a_t s + o(s)). \tag{16}$$

Now denote θ_t as the unique positive solution of the equation $\psi_t'(\theta) = a_t$. Such solution exists because $\psi_t(\cdot)$ is steep and $a_t = 1 - \lambda \int_t^\infty \bar{F}(u)\,du > \lambda \int_0^t \bar{F}(u)\,du = \psi_t'(0)$. Then under our current assumptions Gartner-Ellis Theorem implies that $(1/s)\log P_r(Q^\infty(t) > s) \to -I_t$ where

$$I_t := \sup_{\theta\in\mathbb{R}}\{\theta a_t - \psi_t(\theta)\} = \theta_t a_t - \psi_t(\theta_t). \tag{17}$$

The quantity I_t is the so-called rate function of $\bar{Q}^\infty(t)$ evaluated at a_t.

At this point let us note the following properties of θ_t and I_t when regarded as functions of t:

LEMMA 3.2. *θ_t satisfies the following*:
(i) *$\theta_t > 0$ is nonincreasing in t for all $t > 0$.*
(ii) *$\lim_{t\to 0}\theta_t = \infty$.*
(iii) *$\lim_{t\to\infty}\theta_t = \theta_\infty$ where θ_∞ is the unique positive root of the equation $\psi_\infty'(\theta) = 1$, and*

$$\psi_\infty(\theta) := \int_0^\infty \psi_N(\log(e^\theta \bar{F}(u) + F(u)))\,du. \tag{18}$$

LEMMA 3.3. *I_t satisfies the following*:
(i) *I_t is nonincreasing in t for $t > 0$.*
(ii) *$\lim_{t\to\infty} I_t = \inf_{t>0} I_t = I^*$ where*

$$I^* := \theta_\infty - \psi_\infty(\theta_\infty). \tag{19}$$

(iii) *If V has bounded support over $[0, M]$, then $I^* = I_t$ for any $t \geq M$.*

To construct an implementable efficient importance sampling scheme, one can look at the derivative of $\psi_t(\theta)$,

$$\psi_t'(\theta) = \int_0^t \psi_N'(\log(e^\theta \bar{F}(t-u) + F(t-u)))\frac{e^\theta \bar{F}(t-u)}{e^\theta \bar{F}(t-u) + F(t-u)}\,du,$$

which is the asymptotic mean of $\bar{Q}^\infty(t)/s$ as $s \to \infty$ under the exponential change of measure with parameter θ. When $\theta = 0$, $\psi_t'(0) = \int_0^t \psi_N'(0)\bar{F}(t-u)\,du = \lambda \int_0^t \bar{F}(t-u)\,du$. Comparing with $\psi_t'(\theta_t)$ suggests a build-up of the system by accelerating the arrival rate from λ to $\psi_N'(\log(e^{\theta_t}\bar{F}(t-u) + F(t-u)))$ at time u and changing the service time distributions such that the probability for an arrival at time u to stay in the system at time t is given by $e^{\theta_t}\bar{F}(t-u)/(e^{\theta_t}\bar{F}(t-u) + F(t-u))$. Denote $\tilde{P}^t(\cdot)$ and $\tilde{E}^t[\cdot]$ as the probability measure and expectation under importance sampling. The above changes can be achieved by setting an exponential tilting of the i-th interarrival time U_i by

$$\begin{aligned}
&\tilde{P}^t(U_i \in dy)\\
&\quad := \exp\{\kappa_s^{-1}(-\log(e^{\theta_t}\bar{F}(t-A_i) + F(t-A_i)))y - \kappa_s(\kappa_s^{-1}(-\log(e^{\theta_t}\bar{F}(t-A_i) + F(t-A_i))))\}P(U_i \in dy)\\
&\quad = e^{-s\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))y}(e^{\theta_t}\bar{F}(t-A_i) + F(t-A_i))P(U_i \in dy)
\end{aligned}$$

given the i-th arrival time A_i (recall the convention $U_i = A_{i+1} - A_i$), and for an arrival at A_i its tilted service time distribution follows

$$\tilde{P}^t(V_i \in dy) := \begin{cases} \dfrac{f(y)}{e^{\theta_t}\bar{F}(t-A_i) + F(t-A_i)} & \text{for } 0 \leq y \leq t - A_i \\ \dfrac{e^{\theta_t} f(y)}{e^{\theta_t}\bar{F}(t-A_i) + F(t-A_i)} & \text{for } y > t - A_i \end{cases}.$$

The contribution to likelihood ratio $P(\cdot)/\tilde{P}^t(\cdot)$ by each arrival and service time assignment is accordingly (using slight abuse of notation)

$$\frac{P(U_i)}{\tilde{P}^t(U_i)} = \frac{e^{s\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i}}{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)} \tag{20}$$

and

$$\frac{P(V_i)}{\tilde{P}^t(V_i)} = \frac{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)}{e^{\theta_t I(V_i>t-A_i)}}. \tag{21}$$

We tilt the process using (20) and (21) until the time that we know overflow will happen at time; t i.e., $t \wedge \tau_s[t]$ where $\tau_s[t] := \inf\{u>0\colon r(t)+\sum_{i=1}^{N_s(u)} I(V_i > t-A_i) > s\}$. The overall likelihood ratio on the set $Q^\infty(t) > s$ will be

$$\begin{aligned} L &= \prod_{i=1}^{N_s(\tau_s[t])-1} \frac{e^{s\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))}}{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)} \prod_{i=1}^{N_s(\tau_s[t])} \frac{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)}{e^{\theta_t I(V_i>t-A_i)}} \\ &= \exp\left\{ s \sum_{i=1}^{N_s(\tau_s[t])-1} \psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i - \theta_t \sum_{i=1}^{N_s(\tau_s[t])} I(V_i > t-A_i) \right\} \\ &\quad \cdot (e^{\theta_t}\bar{F}(t-A_{\tau_s[t]})+F(t-A_{\tau_s[t]})). \end{aligned} \tag{22}$$

This estimator $LI(Q^\infty(t) > s)$ can be shown to be asymptotically optimal in estimating $P_r(Q^\infty(t) > s)$:

Proposition 3.1.

$$\limsup_{s\to\infty} \frac{1}{s} \log \tilde{E}_r^t[L^2; Q^\infty(t) > s] \le -2I_t.$$

Proof. The proof follows from Szechtman and Glynn [28], but for completeness (and also because of our introduction of $\tau_s[t]$ that simplifies the argument in their paper slightly), we shall present it here.

Note that $\sum_{i=1}^{N_s(\tau_s[t])} I(V_i > t-A_i) = s+1-r(t) = a_t s + o(s)$ by the definition of $\tau_s[t]$ and $r(t)$. Also, $e^{\theta_t}\bar{F}(t-A_{\tau_s[t]})+F(t-A_{\tau_s[t]}) \le e^{\theta_t}$ since $\theta_t > 0$.

Since ψ_N is continuous, $\sum_{i=1}^{N_s(\tau_s[t])-1} \psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i$ is an approximation to the Riemann integral $\int_0^{\tau_s[t]} \psi_N(\log(e^{\theta_t}\bar{F}(t-u)+F(t-u)))\,du$, with intervals defined by $0 = A_0 < A_1 < A_2 < \cdots < A_{N_s(\tau_s[t])}$ and within each interval the leftmost function value is used as approximation (with the last interval truncated). Since $\psi_N(\log(e^{\theta_t}\bar{F}(t-u)+F(t-u)))$ is nondecreasing in u when $\theta_t > 0$, and $\tau_s[t] \le t$ on $Q^\infty(t) > s$, we have

$$\begin{aligned} &\sum_{i=1}^{N_s(\tau_s[t])-1} \psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i \\ &\quad \le \int_0^{\tau_s[t]} \psi_N(\log(e^{\theta_t}\bar{F}(t-u)+F(t-u)))\,du \\ &\quad \le \int_0^{t} \psi_N(\log(e^{\theta_t}\bar{F}(t-u)+F(t-u)))\,du \\ &\quad = \psi_t(\theta_t) \end{aligned}$$

on $Q^\infty(t) > s$. Hence (22) gives

$$L^2 \le e^{2s\psi_t(\theta_t)-2\theta_t(a_t s+o(s))}$$

which yields the proposition. □

3.3. Distribution of random horizon. Denote τ as our randomized time horizon. We propose a discrete power-law distribution for τ independent of the process:

$$P(\tau = T + k\delta) = \frac{1}{(k+1)^2} - \frac{1}{(k+2)^2} \quad \text{for } k = 0, 1, 2, \ldots \tag{23}$$

where $\delta = \delta(s) = c/s$ for some constant $c > 0$. The power-law distribution of τ is to avoid exponential contribution from the mixture probability to the likelihood ratio that may disturb algorithmic efficiency. Notice that we use a power law of order 2, and in fact we can choose any power law distribution (with finite mean so that it does not take a long time to generate the process up to τ).

T is a constant to avoid tilting the process on a time horizon too close to 0; otherwise, likelihood ratio would blow up for paths that hit overflow very early (because $\lim_{t\to 0}\theta_t = \infty$ in Lemma 3.2 Part ii; see also §4). A good

choice of T is the following. Let $\tilde{I}_t := \sup_{\theta\in\mathbb{R}}\{\theta(1-\lambda EV)-\psi_N(\theta)t\} = \tilde{\theta}_t(1-\lambda EV)-\psi_N(\tilde{\theta}_t)t$ where $\tilde{\theta}_t$ is the solution to the equation $\psi_N'(\theta)t = 1-\lambda EV$ (which exists for small enough t by the steepness assumption). $\tilde{I}_t$ is the rate function of $N_s(t)$ evaluated at $1-\lambda EV$.

We choose $0 < T < \infty$ that satisfies

$$\tilde{I}_T > 2I^* \tag{24}$$

which always exists by the following lemma:

Lemma 3.4. *$\tilde{I}_t$ satisfies the following*:
(i) *$\tilde{I}_t$ is nonincreasing in t for $t < \eta$ for some small $\eta > 0$.*
(ii) *$\tilde{I}_t \to \infty$ as $t \searrow 0$.*

Remark 3.1. In fact by looking at the arguments in §4, one can see that δ being merely $o(1)$ leads to asymptotic optimality. However, the coarser the δ, the larger is the subexponential factor beside the exponential decay component in the variance, with the extreme that when δ is order 1, asymptotic optimality no longer holds. The choice of $\delta = c/s$ is found to perform well empirically, as illustrated in §6.

3.4. Likelihood ratio. After sampling the randomized time horizon, we accelerate the process using the sequential tilting scheme (20) and (21) with a realized $\tau = T + k\delta$, under a modification: As discussed in §3.1, we are interested in approximating the first-passage-type probability $P_{\mathcal{A}}(\tau_s < \tau_{\mathcal{A}})$; consequently, we tilt the process until $(T+k\delta)\wedge\tau_s\wedge\tau_{\mathcal{A}}$ (rather than $\tau_s[T+k\delta]$ defined in §3.2). If $(T+k\delta)\wedge\tau_s < \tau_{\mathcal{A}}$, we continue the $GI/G/s$ system under the original measure until $\tau_{\mathcal{A}}$. Also, to prevent a blow-up of likelihood ratio close to time 0, we use the original measure throughout the whole process whenever the realization of τ is T (the proof of efficiency in §4 will illustrate this in detail). We refer $\tilde{E}[\,\cdot\,]$ and $\tilde{P}(\,\cdot\,)$ to the overall importance sampling measure under this scheme, which is depicted rigorously as follows. Recall from §2.2 that $W_u = (Q(u,\cdot), B(u))$ represents the state of the process at time u. We have

$$\tilde{P}(\{W_u, 0\le u\le\tau_s\wedge\tau_{\mathcal{A}}\}\in S) = \sum_{k=0}^{\infty} P(\tau = T+k\delta)\tilde{P}^{T+k\delta}(\{W_u, 0\le u\le\tau_s\wedge\tau_{\mathcal{A}}\}\in S),$$

where $\tilde{P}^T(\,\cdot\,)$ is set to equate $P(\,\cdot\,)$, and for $k\ge 1$, $\tilde{P}^{T+k\delta}$ is the probability measure under the sequential tilting scheme introduced in §3.2, using time horizon $T+k\delta$, with tilting stopped at time $(T+k\delta)\wedge\tau_s\wedge\tau_{\mathcal{A}}$. So the overall likelihood ratio $L := L(W_u, 0\le u\le\tau_s)$ on the set $\tau_s < \tau_{\mathcal{A}}$ is given by (with slight abuse of notation)

$$\begin{aligned} L = \frac{dP}{d\tilde{P}} &= \frac{P(W_u, 0\le u\le\tau_s)}{\sum_{k=0}^{\infty} P(\tau = T+k\delta)\tilde{P}^{T+k\delta}(W_u, 0\le u\le\tau_s)} \\ &= \frac{1}{\sum_{k=0}^{\infty} P(\tau = T+k\delta)L_{T+k\delta}^{-1}}, \end{aligned} \tag{25}$$

where $L_t := L_t(W_u, 0\le u\le\tau_s)$ is the individual likelihood ratio as a sequential product of (20) and (21) up to $t\wedge\tau_s$; i.e.,

$$L_t = \begin{cases} \exp\left\{ s\sum_{i=1}^{N_s(\tau_s)-1}\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i - \theta_t\sum_{i=1}^{N_s(\tau_s)-1} I(V_i > t-A_i)\right\} & \text{for } t\ge\tau_s \\ \exp\left\{ s\sum_{i=1}^{N_s(t)-1}\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i - \theta_t\sum_{i=1}^{N_s(t)-1} I(V_i > t-A_i)\right\} & \text{for } t < \tau_s \end{cases} \tag{26}$$

for $t > T$ and is 1 for $t = T$.

3.5. The algorithm. We now state our algorithm. Assuming we start from $r\in\mathcal{A}$ with a given initial age $B(0)$, do the following:

Algorithm 2

(i) Set $A_0 := 0$. Also initialize $N_{\mathcal{A}}\leftarrow 0$, $L\leftarrow 0$, and $\tau_s\leftarrow\infty$.
(ii) Sample τ according to (23). Say we get a realization $\tau = T+k\delta$.
(iii) Simulate U_0 according to the initial age $B(0)$. Set $A_1 := U_0$. Check if $\tau_{\mathcal{A}}$ is reached, in which case go to Step vii.

(iv) Starting from $i=1$, repeat the following:
(a) Generate V_i according to $\tilde{P}^{T+k\delta}(\cdot)$, where

$$\tilde{P}^t(V_i \in dy) := \begin{cases} \dfrac{f(y)}{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)} & \text{for } 0 \le y \le t-A_i \\ \dfrac{e^{\theta_t} f(y)}{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)} & \text{for } y > t-A_i \end{cases}$$

with θ_t defined as in (17) for $t > T$ and 0 for $t = T$.
(b) Generate U_i according to $\tilde{P}^{T+k\delta}(\cdot)$, where

$$\tilde{P}^t(U_i \in dy) := e^{-s\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))y}(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i))P(U_i \in dy)$$

with θ_t defined as in (17) for $t > T$ and 0 for $t = T$.
(c) Set $A_{i+1} := U_i + A_i$.
(d) If $\tau_{\mathcal{A}}$ is reached in $[A_i, A_{i+1})$, go to Step vi.
(e) Compute $Q^{\infty}(A_{i+1})$. If $Q^{\infty}(A_{i+1}) > s$, then set $\tau_s \leftarrow A_{i+1}$, remove the new arrival at A_{i+1}, update $N_{\mathcal{A}} \leftarrow N_{\mathcal{A}} + 1$, and go to Step v.
(f) If $A_{i+1} \ge t$, go to Step v.
(g) Update $i \leftarrow i+1$.
(v) Repeat the following:
(a) Generate V_i and U_i under the original measure. Set $A_{i+1} := U_i + A_i$.
(b) If $\tau_{\mathcal{A}}$ is reached in $[A_i, A_{i+1})$, go to Step vi.
(c) Compute $Q(A_{i+1})$. This includes the removal of new arrival A_{i+1} from the system in case it is a loss; in such case update $N_{\mathcal{A}} \leftarrow N_{\mathcal{A}} + 1$, and set $\tau_s \leftarrow A_{i+1}$ if $\tau_s = \infty$.
(d) Update $i \leftarrow i+1$.
(vi) Compute $LI(\tau_s < \tau_{\mathcal{A}})$ using (25) and (26).
(vii) Output $N_{\mathcal{A}} LI(\tau_s < \tau_{\mathcal{A}})$.

4. Algorithmic efficiency. In this section we will prove asymptotic optimality of the estimator outputted by Algorithm 2. We will also identify I^* defined in (19) as the exponential decay rate of $E_{\mathcal{A}} N_{\mathcal{A}}$. The key result is the following:

THEOREM 4.1. *The second moment of the estimator in Algorithm 2 satisfies*

$$\limsup_{s\to\infty} \frac{1}{s} \log \tilde{E}_r[N_{\mathcal{A}}^2 L^2; \tau_s < \tau_{\mathcal{A}}] \le -2I^*$$

for any $r \in \mathcal{A}$.

This result, together with Theorem 4.2 in the sequel, will expose a loop of inequalities that leads to asymptotic optimality and large deviations asymptotic simultaneously. The main technicality of this result is an estimate of the continuity of the likelihood ratio, or intuitively the "overshoot" at the time of loss. It draws upon a two-dimensional point process description of the system, in which the geometry of the process plays an important role in estimating this "overshoot."

PROOF. Denote $\lceil x \rceil := \min\{T + k\delta\colon k \in \mathbb{N}, x \le T + k\delta\}$. Also recall the definition $a_t := 1 - \lambda \int_t^{\infty} \bar{F}(u)\,du$.
Consider the likelihood ratio in (25). We provide an upper bound by isolating the term $\tau = \lceil \tau_s \rceil$ in the involved summation. This technique has been used in analyzing rare-event estimators that involve hitting sets coverable by several half-spaces (see Bucklew [12], Chap. 5, p. 112); a similar idea has also been used in Blanchet et al. [9]. We have

$$LI(\tau_s < \tau_{\mathcal{A}}) = \frac{1}{\sum_{k=0}^{\infty} P(\tau = T + k\delta) L_{T+k\delta}^{-1}} I(\tau_s < \tau_{\mathcal{A}}) \le \frac{L_{\lceil \tau_s \rceil}}{P(\tau = \lceil \tau_s \rceil)} I(\tau_s < \tau_{\mathcal{A}}). \tag{27}$$

We denote $g(\cdot) := P(\tau = \cdot)$, a deterministic function defined on $\mathbb{N}$, and hence $g(\lceil \tau_s \rceil) = P(\tau = \lceil \tau_s \rceil)$ is a random variable generated by τ_s. Then (27) is a.s. bounded from above by

$$g(T)^{-1} I(\tau_s \le T; \tau_s < \tau_{\mathcal{A}}) + g(\lceil \tau_s \rceil)^{-1} \exp\left\{ s \sum_{i=1}^{N_s(\tau_s)-1} \psi_N(\log(e^{\theta_{\lceil \tau_s \rceil}} \bar{F}(\lceil \tau_s \rceil - A_i)\right.$$

$$+F(\lceil\tau_s\rceil-A_i)))U_i-\theta_{\lceil\tau_s\rceil}\sum_{i=1}^{N_s(\tau_s)-1}I(V_i>\lceil\tau_s\rceil-A_i)\Big\}I(\tau_s>T;\tau_s<\tau_{\mathcal{M}})$$

$$\le C_1I(\tau_s\le T;\tau_s<\tau_{\mathcal{M}})+\frac{C_2\tau_s^3}{\delta^3}\exp\{s\psi_{\lceil\tau_s\rceil}(\theta_{\lceil\tau_s\rceil})-\theta_{\lceil\tau_s\rceil}(\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s)-1)\}I(\tau_s>T;\tau_s<\tau_{\mathcal{M}})$$

$$\le C_1I(\tau_s\le T;\tau_s<\tau_{\mathcal{M}})+\frac{C_2\tau_s^3}{\delta^3}\exp\{-sI^*+\theta_{\lceil\tau_s\rceil}(sa_{\lceil\tau_s\rceil}+1-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s))\}I(\tau_s>T;\tau_s<\tau_{\mathcal{M}}),$$

where C_1 and C_2 are positive constants. Note that the second inequality is because $\sum_{i=1}^{N_s(\tau_s)-1}\psi_N(\log(e^{\theta_{\lceil\tau_s\rceil}}\cdot\bar{F}(\lceil\tau_s\rceil-A_i)+F(\lceil\tau_s\rceil-A_i)))U_i$ is a Riemann sum of the integral $\psi_{\lceil\tau_s\rceil}(\theta_{\lceil\tau_s\rceil})=\int_0^{\lceil\tau_s\rceil}\psi_N(\log(e^{\theta_{\lceil\tau_s\rceil}}\bar{F}(\lceil\tau_s\rceil-u)+F(\lceil\tau_s\rceil-u)))\,du$ (excluding the intervals at the two ends) and that $\psi_N(\log(e^\theta_{\lceil\tau_s\rceil}\bar{F}(\lceil\tau_s\rceil-u)+F(\lceil\tau_s\rceil-u)))$ is a nondecreasing function in u. Also note that $\sum_{i=1}^{N_s(\tau_s)}I(V_i>\lceil\tau_s\rceil-A_i)=\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s)$ is the number of customers who arrive before τ_s and leave after $\lceil\tau_s\rceil$. The last inequality follows from the definition of $I_{\lceil\tau_s\rceil}$ and Lemma 3.3 Part ii. Now we have

$$\begin{aligned}\tilde{E}_r[N_{\mathcal{M}}^2L^2;\tau_s<\tau_{\mathcal{M}}]&=E_r[N_{\mathcal{M}}^2L;\tau_s<\tau_{\mathcal{M}}]\\&\le C_1E_r[N_{\mathcal{M}}^2;\tau_s\le T;\tau_s<\tau_{\mathcal{M}}]\\&\quad+\frac{C_2}{\delta^3}e^{-sI^*}E_r[N_{\mathcal{M}}^2\tau_s^3\exp\{\theta_{\lceil\tau_s\rceil}(sa_{\lceil\tau_s\rceil}+1-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s))\};\ \tau_s>T;\tau_s<\tau_{\mathcal{M}}].\end{aligned}\tag{28}$$

Consider the first summand. By Holder's inequality $E_r[N_{\mathcal{M}}^2;\tau_s\le T;\tau_s<\tau_{\mathcal{M}}]\le(E_r[N_{\mathcal{M}}^{2p}])^{1/p}(P_r(\tau_s\le T))^{1/q}$ for $1/p+1/q=1$. Also, $P_r(\tau_s\le T)\le P(N_s(T)>s-r(T))\le P(N_s(T)>s(1-\lambda EV)+o(s))$ and Gartner-Ellis Theorem yield $\lim_{s\to\infty}(1/s)\log P(N_s(T)>s(1-\lambda EV)+o(s))=-\tilde{I}_T<-2I^*$ by our choice of T in (24). Combining these observations, and using Proposition 2.1, we get

$$\limsup_{s\to\infty}\frac{1}{s}\log E_r[N_{\mathcal{M}}^2;\tau_s\le T;\tau_s<\tau_{\mathcal{M}}]\le\limsup_{s\to\infty}\frac{1}{sp}\log E_r[N_{\mathcal{M}}^{2p}]+\limsup_{s\to\infty}\frac{1}{sq}\log P_r(\tau_s\le T)\le-2I^*$$

for q close enough to 1.

In view of (28) and Dembo and Zeitouni [13], Lemma 1.2.15, the proof will be complete once we can prove that

$$\limsup_{s\to\infty}\frac{1}{s}\log E_r[N_{\mathcal{M}}^2\tau_s^3\exp\{\theta_{\lceil\tau_s\rceil}(sa_{\lceil\tau_s\rceil}+1-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s))\};\tau_s>T;\tau_s<\tau_{\mathcal{M}}]\le-I^*.\tag{29}$$

The derivation of (29) requires the analysis of the "overshoot" at the time of loss, briefly discussed after the statement of Theorem 4.1. To this end, we write

$$\begin{aligned}&E_r[N_{\mathcal{M}}^2\tau_s^3\exp\{\theta_{\lceil\tau_s\rceil}(sa_{\lceil\tau_s\rceil}+1-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s))\};\tau_s>T;\tau_s<\tau_{\mathcal{M}}]\\&=E_r\left[N_{\mathcal{M}}^2\tau_s^3\exp\left\{\theta_{\lceil\tau_s\rceil}\left(s+1-\lambda s\int_{\lceil\tau_s\rceil}^\infty\bar{F}(u)\,du-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s)\right)\right\};\tau_s>T;\tau_s<\tau_{\mathcal{M}}\right]\\&\le e^{C\theta_T\sqrt{s}}E_r[N_{\mathcal{M}}^2\tau_s^3\exp\{\theta_{\lceil\tau_s\rceil}(s+1-r(\lceil\tau_s\rceil)-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s))\};\tau_s>T;\tau_s<\tau_{\mathcal{M}}]\\&=e^{C\theta_T\sqrt{s}}\sum_{k=1}^\infty E_r[N_{\mathcal{M}}^2\tau_s^3\exp\{\theta_{\lceil\tau_s\rceil}(s+1-r(\lceil\tau_s\rceil)-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s))\};\lceil\tau_s\rceil=T+k\delta;\tau_A>T+(k-1)\delta]\\&\le e^{C\theta_T\sqrt{s}}\sum_{k=1}^\infty(E_rN_{\mathcal{M}}^{2p})^{1/p}(E_r\tau_{\mathcal{M}}^{3q})^{1/q}(P_r(\tau_A>T+(k-1)\delta))^{1/h}\\&\quad\cdot(E_r[\exp\{l\theta_{T+k\delta}(s+1-r(T+k\delta)-\bar{Q}^\infty(\tau_s,T+k\delta-\tau_s))\};T+(k-1)\delta<\tau_s\le T+k\delta])^{1/l}\\&=e^{O(\sqrt{s})}\sum_{k=1}^\infty(E_rN_{\mathcal{M}}^{2p})^{1/p}(E_r\tau_A^{3q})^{1/q}(P_r(\tau_{\mathcal{M}}>T+(k-1)\delta))^{1/h}\\&\quad\cdot(E_r[\exp\{l\theta_{T+k\delta}(s+1-r(\tau_s)-\bar{Q}^\infty(\tau_s,T+k\delta-\tau_s))\};T+(k-1)\delta<\tau_s\le T+k\delta])^{1/l},\end{aligned}\tag{30}$$

where C is a positive constant and $1/p+1/q+1/h+1/l=1$. The first inequality follows because $r(\cdot)\in J(\cdot)$ and Lemma 3.3 Part I, whereas the second inequality follows from generalized Holder's inequality (e.g., Wheeden and Zygmund [29]). The last equality holds because $r(\tau_s)-r(T+k\delta)=o(s)$, again since $r\in\mathcal{M}$, for $T+(k-1)\delta<\tau_s\le T+k\delta$.

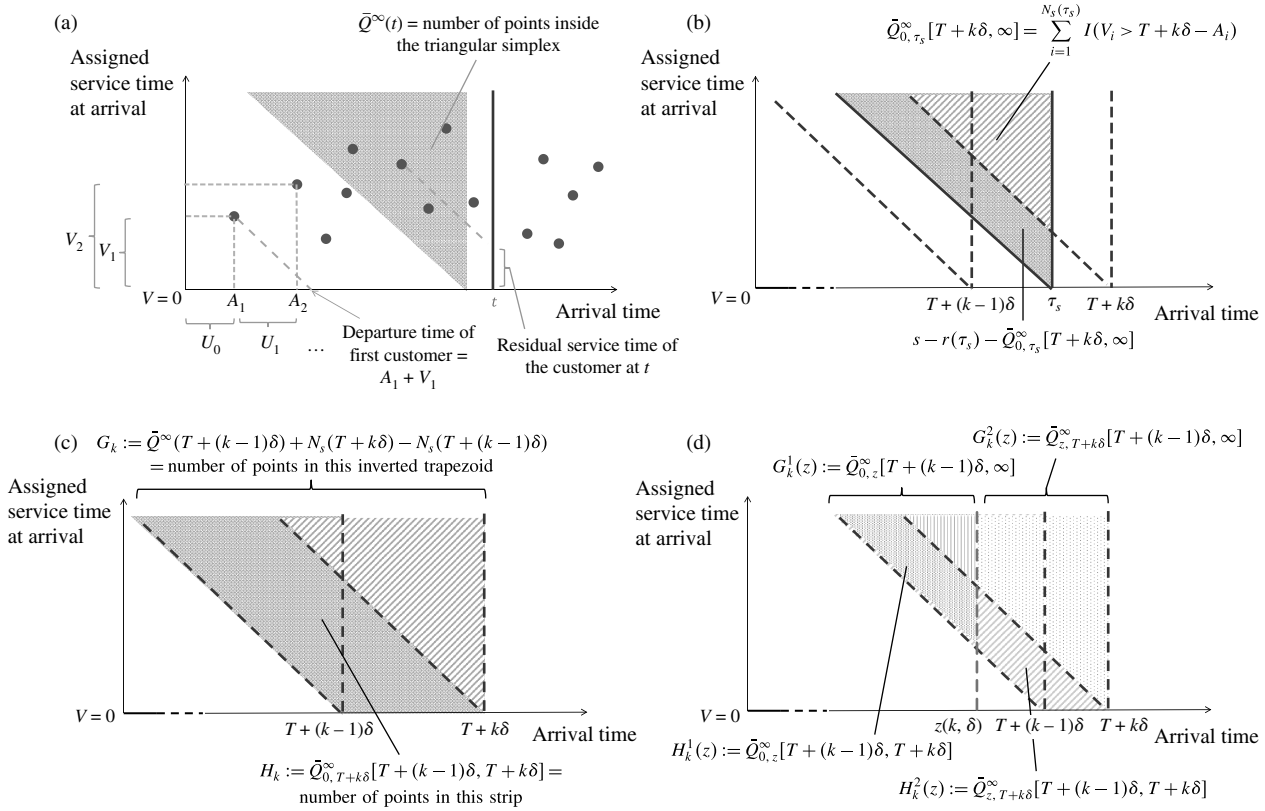

FIGURE 3. Two-dimensional plots of the point process representing customer statuses.

We now analyze

$$E_r[\exp\{l\theta_{T+k\delta}(s+1-r(\tau_s)-\bar{Q}^{\infty}(\tau_s, T+k\delta-\tau_s))\}; T+(k-1)\delta < \tau_s \le T+k\delta]. \tag{31}$$

We plot the arrivals on a two-dimensional plane, with x-axis indicating the time of arrival and y-axis indicating the assigned service time at the time of arrival. Each arrival is represented by a point with a specified (x, y)-coordinate on this two-dimensional plane. Such plot has been used in the study of $M/G/\infty$ system (see, for example, Foley [14]). In this representation it is easy to see that the departure time of an arriving customer is the x-intercept of a straight line, with slope -1, that passes through the customer's representing point. As a result, $\bar{Q}^{\infty}(t)$, for example, will be the number of all the points inside the triangular simplex created by a vertical line that passes through the point $(t, 0)$ and a straight line with slope -1 that also passes through $(t, 0)$. See Figure 3(a).

For notational convenience we denote $\bar{Q}^{\infty}_{t_1,t_2}[t_3, t_4] := \sum_{i=N_s(t_1)+1}^{N_s(t_2)} I(t_3 - A_i < V_i \le t_4 - A_i)$ as the number of customers in the $GI/G/\infty$ system who arrive sometime in $(t_1, t_2]$ and leave the system sometime in $(t_3, t_4]$. It is easy to see, for example, that $\bar{Q}^{\infty}(\tau_s, T+k\delta-\tau_s) = \bar{Q}^{\infty}_{0,\tau_s}[T+k\delta,\infty]$ for $T+k\delta \ge \tau_s$.

To proceed with our proof, the key idea is to bound the number of customers involved in (31) by identifying convenient geometric objects to cover the involved areas in the two-dimensional plot. Figure 3(b) shows the region filled by $\bar{Q}^{\infty}(\tau_s, T+k\delta-\tau_s) = \bar{Q}^{\infty}_{0,\tau_s}[T+k\delta,\infty]$ as a shifted simplex starting from the point $(\tau_s, T+k\delta-\tau_s)$. Note that by definition $\bar{Q}^{\infty}(\tau_s) = s+1-r(\tau_s)$, and so $s+1-r(\tau_s) - \bar{Q}^{\infty}_{0,\tau_s}[T+k\delta,\infty]$ corresponds to the downward strip ending at $(\tau_s, 0)$ and $(\tau_s, T+k\delta-\tau_s)$, which is obviously smaller than the region represented by $H_k := \bar{Q}^{\infty}_{0,T+k\delta}[T+(k-1)\delta, T+k\delta]$ in Figure 3(c).

Define $G_k := \bar{Q}^{\infty}(T+(k-1)\delta) + N_s(T+k\delta) - N_s(T+(k-1)\delta)$, which is represented by the trapezoidal area depicted in Figure 3. Observe that $T+(k-1)\delta < \tau_s \le T+k\delta$ implies that one of the triangular simplex corresponding to $\bar{Q}^{\infty}(t)$, for $T+(k-1)\delta < t \le T+k\delta$, has number of points larger than $s-r(T+(k-1)\delta)$. This in turn implies that the region represented by G_k has more than $s-r(T+(k-1)\delta)$ number of points.

The above observations lead to

$$\begin{aligned} &E_r[\exp\{l\theta_{T+k\delta}(s+1-r(\tau_s)-\bar{Q}^{\infty}_{0,\tau_s}[T+k\delta,\infty])\}; T+(k-1)\delta < \tau_s \le T+k\delta] \\ &\quad \le E_r[e^{l\theta_{T+k\delta}H_k}; G_k > s-r(T+(k-1)\delta)]. \end{aligned} \tag{32}$$

From now on we focus on the case when service time has unbounded support (the bounded support case is simpler and will be presented later in the proof). We introduce a time point $z = z(k, s)$ and consider the divisions of areas represented by H_k and G_k in Figure 3(d):

$$H_k^1(z) := \bar{Q}^{\infty}_{0,z}[T+(k-1)\delta, T+k\delta] \subset G_k^1(z) := \bar{Q}^{\infty}_{0,z}[T+(k-1)\delta, \infty],$$
$$H_k^2(z) := \bar{Q}^{\infty}_{z,T+k\delta}[T+(k-1)\delta, T+k\delta] \subset G_k^2(z) := \bar{Q}^{\infty}_{z,T+k\delta}[T+(k-1)\delta, \infty].$$

Note that $H_k = H_k^1(z) + H_k^2(z)$ and $G_k = G_k^1(z) + G_k^2(z)$.

Moreover, define A_i^k, $i = 1, \ldots, G_k$ to be the arrival times of all the customers that G_k is counting. Note that given the arrival times A_i^k, $i = 1, \ldots, G_k$, the events whether each of these customers falls into H_k are independent Bernoulli random variables. Indeed, the probability of each of these Bernoulli variables is the conditional probability that the customer, with arrival time A_i^k, falls into the region H_k, given that he/she falls into G_k. Note from Figure 3 that the probability of a customer with arrival time A_i^k falling into G_k is $\bar{F}(T+(k-1)\delta - A_i^k)$ and the probability of falling into H_k is $\bar{F}(T+(k-1)\delta - A_i^k) - \bar{F}(T+k\delta - A_i^k)$. Hence the Bernoulli probability that corresponds to arrival A_i^k is

$$p_i^k := \frac{\bar{F}(T+(k-1)\delta - A_i^k) - \bar{F}(T+k\delta - A_i^k)}{\bar{F}(T+(k-1)\delta - A_i^k)}. \tag{33}$$

As a result, we can write (32) as

$$\begin{aligned}
&E_r[e^{l\theta_{T+k\delta}(H_k^1(z)+H_k^2(z))}; G_k > s - r(T+(k-1)\delta)] \\
&\quad = E_r[E_r[e^{l\theta_{T+k\delta}(H_k^1(z)+H_k^2(z))} \mid A_i^k, i = 1, \ldots, G_k]; G_k > s - r(T+(k-1)\delta)] \\
&\quad = E_r[E_r[e^{l\theta_{T+k\delta}H_k^1(z)} \mid A_i^k, i = 1, \ldots, G_k^1(z)] E_r[e^{l\theta_{T+k\delta}H_k^2(z)} \mid A_i^k, i = G_k^1(z)+1, \ldots, G_k^1(z)+G_k^2(z)]; \\
&\qquad G_k^1(z) + G_k^2(z) > s - r(T+(k-1)\delta)] \\
&\quad \le E_r\left[e^{l\theta_{T+k\delta}G_k^1(z)} \prod_{i=G_k^1(z)+1}^{G_k^1(z)+G_k^2(z)} (1 + (e^{l\theta_{T+k\delta}} - 1)p_i^k); G_k^1(z) + G_k^2(z) > s - r(T+(k-1)\delta)\right].
\end{aligned} \tag{34}$$

Let

$$p_k(z) := \sup_{A_i^k > z} p_i^k \le \frac{C\delta}{\bar{F}(T+k\delta - z)} \tag{35}$$

for some constant $C > 0$, where the inequality follows from (33). Also let

$$\psi^1_{s,z,k}(\theta) := \log E e^{\theta G_k^1(z)} = s\int_0^z \psi_N(\log(e^{\theta}\bar{F}(T+(k-1)\delta - u) + F(T+(k-1)\delta - u)))\, du + o(s)$$
$$\psi^2_{s,z,k}(\theta) := \log E e^{\theta G_k^2(z)} = s\int_z^{T+k\delta} \psi_N(\log(e^{\theta}\bar{F}(T+(k-1)\delta - u) + F(T+(k-1)\delta - u)))\, du + o(s)$$

where $o(s)$ is uniform in θ, k, and z. This is due to the following lemma, whose proof will be deferred to the appendix.

Lemma 4.1. *We have*

$$\frac{1}{s}\log E e^{\theta \bar{Q}^{\infty}_{w,z}[t,\infty]} \to \int_w^z \psi_N(\log(e^{\theta}\bar{F}(t-u) + F(t-u)))\, du$$

uniformly over $\theta \in [\theta_\infty, \theta_T]$, $t \ge T$ *and* $0 \le w \le z \le t + \eta$ *for any* $\eta > 0$.

When $p_k(z)$ is small enough, (34) is less than or equal to

$$\begin{aligned}
&E_r[e^{l\theta_{T+k\delta}G_k^1(z)}(1 + (e^{l\theta_{T+k\delta}} - 1)p_k(z))^{G_k^2(z)}; G_k^1(z) + G_k^2(z) > s - r(T+(k-1)\delta)] \\
&\quad = E_r[E_r[e^{l\theta_{T+k\delta}G_k^1(z) + \log(1+(e^{l\theta_{T+k\delta}}-1)p_k(z))G_k^2(z)}; G_k^2(z) > s - r(T+(k-1)\delta) - G_k^1(z) \mid G_k^1(z), B(z)]] \\
&\quad \le E_r[\exp\{l\theta_{T+k\delta}G_k^1(z) - \theta_{T+(k-1)\delta}(s - r(T+(k-1)\delta) - G_k^1(z)) \\
&\qquad + \psi^2_{s,z,k}(\log(1 + (e^{l\theta_{T+k\delta}} - 1)p_k(z)) + \theta_{T+(k-1)\delta})\}]
\end{aligned}$$

$$
\begin{aligned}
&= \exp\{\psi^1_{s,z,k}(l\theta_{T+k\delta} + \theta_{T+(k-1)\delta}) - \theta_{T+(k-1)\delta}(s - r(T+(k-1)\delta)) \\
&\qquad + \psi^2_{s,z,k}(\log(1 + (e^{l\theta_{T+k\delta}} - 1)p_k(z)) + \theta_{T+(k-1)\delta})\} \\
&= \exp\Bigg\{ s\int_0^z \psi_N(\log(e^{l\theta_{T+k\delta}+\theta_{T+(k-1)\delta}}\bar{F}(T+(k-1)\delta-u) + F(T+(k-1)\delta-u)))\,du \\
&\quad - s\int_0^z \psi_N(\log(e^{\log(1+(e^{l\theta_{T+k\delta}}-1)p_k(z))+\theta_{T+(k-1)\delta}}\bar{F}(T+(k-1)\delta-u) + F(T+(k-1)\delta-u)))\,du \\
&\quad - \theta_{T+(k-1)\delta}(s - r(T+(k-1)\delta)) + s\psi_{T+(k-1)\delta}(\log(1 + (e^{l\theta_{T+k\delta}} - 1)p_k(z)) + \theta_{T+(k-1)\delta}) + o(s)\Bigg\}, \qquad (36)
\end{aligned}
$$

where the inequality follows by Chernoff's inequality (see for example Bucklew [12], Chap. 8, p. 151), and the last equality follows from

$$
\psi^2_{s,z,k}(\theta) = s\psi_{T+(k-1)\delta}(\theta) - s\int_0^z \psi_N(\log(e^{\theta}\bar{F}(T+(k-1)\delta-u) + F(T+(k-1)\delta-u)))\,du + o(s)
$$

uniformly, by Lemma 4.1.

The expression (36) is important in controlling each term in the summation in (30) over the index $k \in \mathbb{N}$. We shall show that (36) is bounded by the quantity $e^{-sI_{T+(k-1)\delta}+o(s)}$ uniformly over $k \in \mathbb{N}$ by choosing, for each $k \in \mathbb{N}$, a suitable value of $z = z(k,s)$ as $s \to \infty$. First, let $\rho_s \nearrow \infty$ be a sequence satisfying $s\bar{F}(\rho_s) \nearrow \infty$, whose existence is guaranteed by the unbounded support assumption. For a given s, we divide into two cases: For k such that $T + k\delta \le \rho_s$, we put $z = 0$ and consequently (36) becomes

$$
\exp\{-\theta_{T+(k-1)\delta}(s - r(T+(k-1)\delta)) + s\psi_{T+(k-1)\delta}(\log(1 + (e^{l\theta_{T+k\delta}} - 1)p_k(0)) + \theta_{T+(k-1)\delta}) + o(s)\}. \qquad (37)
$$

Since $T + k\delta \le \rho_s$, we have $p_k(0) \le C\delta/\bar{F}(\rho_s)$ from (35), and hence (37) is bounded by

$$
\exp\{-\theta_{T+(k-1)\delta}(s - r(T+(k-1)\delta)) + s\psi_{T+(k-1)\delta}(\log(1 + (e^{l\theta_{T+k\delta}} - 1)C\delta/\bar{F}(\rho_s)) + \theta_{T+(k-1)\delta}) + o(s)\}. \qquad (38)
$$

For k such that $T + k\delta > \rho_s$, we put $z = T + k\delta - \rho_s$ so that $T + k\delta - z = \rho_s$. Hence again $p_k(z) \le C\delta/\bar{F}(\rho_s)$. Also, now we have

$$
\begin{aligned}
&\int_0^z \psi_N(\log(e^{l\theta_{T+k\delta}+\theta_{T+(k-1)\delta}}\bar{F}(T+(k-1)\delta-u) + F(T+(k-1)\delta-u)))\,du \\
&\qquad = \int_{T+(k-1)\delta-z}^{T+(k-1)\delta} \psi_N(\log(e^{l\theta_{T+k\delta}+\theta_{T+(k-1)\delta}}\bar{F}(u) + F(u)))\,du \\
&\qquad \le \int_{T+(k-1)\delta-z}^{\infty} C_1\lambda(e^{l\theta_{T+k\delta}+\theta_{T+(k-1)\delta}} - 1)\bar{F}(u)\,du \\
&\qquad = C_2\lambda\int_{\rho_s-\delta}^{\infty} \bar{F}(u)\,du
\end{aligned}
$$

for some constants C_1, $C_2 > 0$, where the inequality holds for large enough s, since $T + (k-1)\delta - z = \rho_s$ and that $\log(1+x) \le x$ for $x > 0$ and $\psi_N'(0) = \lambda$. Hence (36) is bounded by

$$
\begin{aligned}
&\exp\Bigg\{ C_2\lambda s\int_{\rho_s-\delta}^{\infty} \bar{F}(u)\,du - \theta_{T+(k-1)\delta}(s - r(T+(k-1)\delta)) \\
&\qquad + s\psi_{T+(k-1)\delta}(\log(1 + (e^{l\theta_{T+k\delta}} - 1)C\delta/\bar{F}(\rho_s)) + \theta_{T+(k-1)\delta})\Bigg\} \qquad (39)
\end{aligned}
$$

for large enough s.

Recall that $\delta = O(1/s)$, and so $C\delta/\bar{F}(\rho_s) \searrow 0$. Hence both (38) and (39) become $e^{-sI_{T+(k-1)\delta}+o(s)}$. Consequently, (30) is less than or equal to

$$
\begin{aligned}
&e^{-sI^*/l+o(s)}\sum_{k=1}^{\infty}(E_r N_{\mathcal{A}}^{2p})^{1/p}(E_r\tau_{\mathcal{A}}^{3q})^{1/q}(P_r(\tau_{\mathcal{A}} > T + (k-1)\delta))^{1/h} \\
&\le e^{-sI^*/l+o(s)}(E_r N_{\mathcal{A}}^{2p})^{1/p}(E_r\tau_{\mathcal{A}}^{3q})^{1/q}\left((P_r(\tau_{\mathcal{A}} > T))^{1/h} + \frac{1}{\delta}\int_T^{\infty}(P_r(\tau_{\mathcal{A}} > u))^{1/h}\,du\right).
\end{aligned}
$$

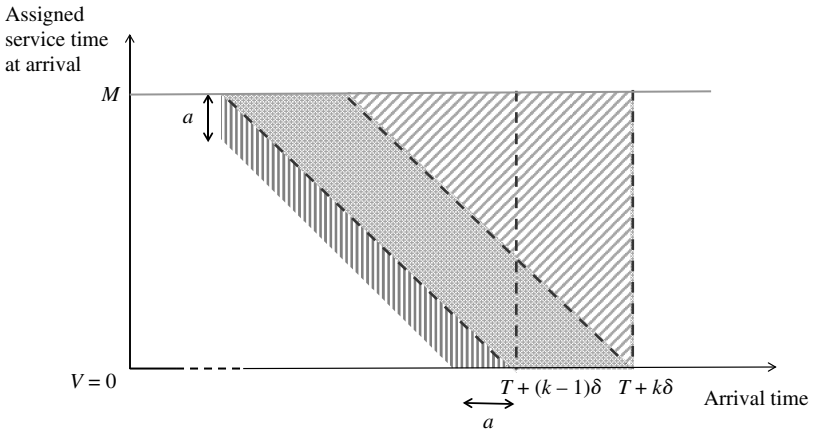

FIGURE 4. Two-dimensional plot for the case of bounded support service time.

From this, and using Proposition 2.1, we get

$$\limsup_{s\to\infty}\frac{1}{s}\log E_r[N_{\mathcal{M}}^2\tau_s^2\exp\{\theta_{\lceil\tau_s\rceil}(sa_{\lceil\tau_s\rceil}+1-\bar{Q}^{\infty}(\tau_s,\lceil\tau_s\rceil-\tau_s))\};\tau_s>T;\tau_s<\tau_{\mathcal{M}}]\le-\frac{I^*}{l}.$$

Since l can be chosen arbitrarily close to 1, we have proved (29).

Finally, we consider the case when V has bounded support over $[0,M]$. Pick a small constant $a>0$, and consider the set of customers $\tilde{G}_k:=\bar{Q}_{(T+(k-1)\delta-M)\vee 0,\,T+k\delta}[T+(k-1)\delta-a,\infty]$ that consists of G_k and a trapezoidal strip of width a running through $(T+(k-1)\delta-a,0)$, $(T+(k-1)\delta,0)$, $((T+(k-1)\delta-M)\vee 0, M\wedge(T+(k-1)\delta))$ and $((T+(k-1)\delta-M)\vee 0, M\wedge(T+(k-1)\delta)-a)$. See Figure 4.

Denote $\tilde{A}_i^k$, $i=1,\ldots,\tilde{G}_k$ as the arrival times of customers falling in $\tilde{G}_k$. Then we have

$$\begin{aligned}
&E_r[e^{l\theta_{T+k\delta}H_k};G_k>s-r(T+(k-1)\delta)]\\
&\quad\le E_r[e^{l\theta_{T+k\delta}H_k};\tilde{G}_k>s-r(T+(k-1)\delta)]\\
&\quad=E_r[E_r[e^{l\theta_{T+k\delta}H_k}\mid\tilde{A}_i^k,i=1,\ldots,\tilde{G}_k];\tilde{G}_k>s-r(T+(k-1)\delta)]\\
&\quad=E_r\Bigg[\prod_{i=1}^{\tilde{G}_k}(1+(e^{l\theta_{T+k\delta}})\tilde{p}_i^k);\tilde{G}_k>s-r(T+(k-1)\delta)\Bigg],
\end{aligned}\tag{40}$$

where

$$\tilde{p}_i^k:=\frac{\bar{F}(T+(k-1)\delta-\tilde{A}_i^k)-\bar{F}(T+k\delta-\tilde{A}_i^k)}{\bar{F}(T+(k-1)\delta-a-\tilde{A}_i^k)}\le\tilde{p}_k:=\sup_{i=1,\ldots,\tilde{G}_k}\tilde{p}_i^k\le\frac{C\delta}{\bar{F}(M-a)}.$$

Hence (40) is less than or equal to

$$\begin{aligned}
&E_r[e^{\log(1+(e^{l\theta_{T+k\delta}})\tilde{p}_k)\tilde{G}_k};\tilde{G}_k>s-r(T+(k-1)\delta)]\\
&\qquad\le e^{-\theta_{T+(k-1)\delta}(s-r(T+(k-1)\delta))+\tilde{\psi}_k(\log(1+(e^{l\theta_{T+k\delta}}-1)\tilde{p})+\theta_{T+(k-1)\delta})},
\end{aligned}\tag{41}$$

where $\tilde{\psi}_k(\theta):=\log Ee^{\theta\tilde{G}_k}$, by Chernoff's inequality. Now note that by Lemma 4.1 we have

$$\begin{aligned}
\tilde{\psi}_k(\theta)&=s\int_{(T+(k-1)\delta-M)\vee 0}^{T+k\delta}\psi_N(\log(e^{\theta}\bar{F}(T+(k-1)\delta-a-u)+F(T+(k-1)\delta-a-u)))\,du+o(s)\\
&=s\int_0^{(M-a)\wedge(T+(k-1)\delta-a)}\psi_N(\log(e^{\theta}\bar{F}(u)+F(u)))\,du+s\psi_N(\theta)(a+\delta)+o(s)\\
&\le s\psi_{T+(k-1)\delta}(\theta)+saC+o(s)
\end{aligned}$$

for some constant $C>0$, uniformly in θ and k. Hence (41) is less than or equal to

$$\begin{aligned}
&e^{-\theta_{T+(k-1)\delta}(s-r(T+(k-1)\delta))+s\psi_{T+(k-1)\delta}(\theta_{T+(k-1)\delta})+saC+o(s)}\\
&\qquad=e^{-sI_{T+(k-1)\delta}+saC+o(s)}.
\end{aligned}$$

Thus (30) is less than or equal to

$$e^{-sI^*/l+saC/l+o(s)}\sum_{k=1}^{\infty}(E_r N_{\mathcal{A}}^{2p})^{1/p}(E_r\tau_{\mathcal{A}}^{3q})^{1/q}(P_r(\tau_{\mathcal{A}}>T+(k-1)\delta))^{1/h}.$$

This gives

$$\limsup_{s\to\infty}\frac{1}{s}\log E_r[N_{\mathcal{A}}^2\tau_s^3\exp\{\theta_{\lceil\tau_s\rceil}(sa_{\lceil\tau_s\rceil}+1-\bar{Q}^{\infty}(\tau_s,\lceil\tau_s\rceil-\tau_s))\};\tau_s>T;\tau_s<\tau_{\mathcal{A}}]\le-\frac{I^*}{l}+\frac{aC}{l}.$$

Since l and a can be chosen arbitrarily close to 1 and 0, respectively, (29) holds and conclusion follows. □

Remark 4.1. The proof can be simplified in the $M/G/s$ system. In particular, there is no need to condition on A_i^k nor introduce the constant a in the case V has bounded support. Since arrival is Poisson, the two-dimensional description of arrivals via the arrival time and the required service time at the time of arrival leads to a Poisson random measure. Hence all the points in G_k are independently sampled, each with probability of falling into H_k being

$$p_k:=\frac{\int_0^{T+k\delta}(\bar{F}(T+(k-1)\delta-u)-\bar{F}(T+k\delta-u))\,du}{\int_0^{T+k\delta}\bar{F}(T+(k-1)\delta-u)\,du}\le\frac{C\delta(M+\delta)}{\int_0^{T+(k-1)\delta}\bar{F}(u)\,du+N_s((k-1)\delta,k\delta)}=O(\delta)$$

for some constant $C>0$; then (31) immediately becomes

$$\begin{aligned}&E_r[(p_ke^{l\theta_{T+k\delta}}+1-p_k)^{G_k};G_k>s-r(T+(k-1)\delta)]\\&\quad=E_r[e^{O(\delta)G_k};G_k>s-r(T+(k-1)\delta)].\end{aligned}$$

The rest follows similarly as in the proof.

Remark 4.2. Note that the result coincides with Erlang's loss formula in the case of $M/G/s$ (see for example Asmussen [4]), which states that the loss probability is exactly given by

$$P_\pi(\text{loss})=\frac{(\lambda sEV)^s/s!}{1+\lambda sEV+\cdots+(\lambda sEV)^s/s!}.$$

A simple calculation reveals that $(1/s)\log P_\pi(\text{loss})\to\log(\lambda EV)+1-\lambda EV=-I^*$.

The next result we will discuss is the lower bound:

Theorem 4.2. *For any $r\in\mathcal{A}$, we have*

$$\liminf_{s\to\infty}\frac{1}{s}\log P_r(\tau_s<\tau_{\mathcal{A}})\ge-I^*.$$

It suffices to prove that $\liminf_{s\to\infty}(1/s)\log P_r(\tau_s<\tau_{\mathcal{A}})\ge-I_{t_n}$ for a sequence $t_n\nearrow\infty$ thanks to Lemma 3.3, Parts i and ii. In fact we will take $t_n=n\Delta$. In the case of bounded support V, it suffices to only consider $n\Delta=\lceil M\rceil$ because of Lemma 3.3 Part iii. For each $n\Delta$, the idea then is to identify a so-called optimal sample path (or more precisely a neighborhood of such a path) that possesses a rate function $I_{n\Delta}$ and has the property $\tau_s<\tau_{\mathcal{A}}$. Note that the probability in consideration is the same for $GI/G/s$ and $GI/G/\infty$ systems. Henceforth we will consider paths in $GI/G/\infty$.

The way we define $\mathcal{A}$ in (8) implies that it suffices to focus on the process on the time-grid $\{0,\Delta,2\Delta,\dots\}$ for checking the condition $\tau_s<\tau_{\mathcal{A}}$. For a path to reach s at time $n\Delta$, the form of $\psi'_{n\Delta}(\theta_{n\Delta})$ hints that $E[\bar{Q}^{\infty}_{(k-1)\Delta,k\Delta}[(j-1)\Delta,j\Delta]\mid Q^{\infty}(n\Delta)>s]=s\alpha_{kj}+o(s)$ and $E[\bar{Q}^{\infty}_{(k-1)\Delta,k\Delta}[n\Delta,\infty]\mid Q^{\infty}(n\Delta)>s]=s\beta_k+o(s)$, where

$$\alpha_{kj}:=\int_{(k-1)\Delta}^{k\Delta}\psi'_N(\log(e^{\theta_{n\Delta}}\bar{F}(n\Delta-u)+F(n\Delta-u)))\frac{F(j\Delta-u)-F((j-1)\Delta-u)}{e^{\theta_{n\Delta}}\bar{F}(n\Delta-u)+F(n\Delta-u)}\,du$$

and

$$\beta_k:=\int_{(k-1)\Delta}^{k\Delta}\psi'_N(\log(e^{\theta_{n\Delta}}\bar{F}(n\Delta-u)+F(n\Delta-u)))\frac{e^{\theta_{n\Delta}}\bar{F}(n\Delta-u)}{e^{\theta_{n\Delta}}\bar{F}(n\Delta-u)+F(n\Delta-u)}\,du$$

for $k=1,\dots,n$, $j=k,\dots,n$. Our goal is to rigorously justify that such a path is the optimal sample path discussed above.

We now state two useful lemmas. The first is a generalization of Glynn [16], whose proof resembles this earlier work and is deferred to the appendix. The second one argues that the path we identified indeed satisfies $\tau_s < \tau_{\mathcal{A}}$:

LEMMA 4.2. *Let* $\boldsymbol{\Theta} := (\theta_{kj}, \theta_{k\cdot})_{k=1,\ldots,n,\, j=k,\ldots,n} \in \mathbb{R}^{n(n+1)/2+n}$, *and define*

$$\bar{\psi}(\boldsymbol{\Theta}) := \sum_{k=1}^{n} \int_{(k-1)\Delta}^{k\Delta} \psi_N\left(\log\left(\sum_{j=k}^{n} e^{\theta_{kj}} P((j-1)\Delta - u < V \le j\Delta - u) + e^{\theta_{k\cdot}} \bar{F}(n\Delta - u)\right)\right) du.$$

We have

$$\frac{1}{s} \log E \exp\left\{\sum_{k=1}^{n}\left(\sum_{j=k}^{n} \theta_{kj} \bar{Q}^{\infty}_{(k-1)\Delta,\, k\Delta}[(j-1)\Delta, j\Delta] + \theta_{k\cdot} \bar{Q}^{\infty}_{(k-1)\Delta,\, k\Delta}[n\Delta, \infty]\right)\right\} \to \bar{\psi}(\boldsymbol{\Theta}).$$

LEMMA 4.3. *Starting with any* $r \in \mathcal{A}$, *the sample path with* $Q^{\infty}_{(k-1)\Delta,\, k\Delta}[(j-1)\Delta, j\Delta] \in ((\alpha_{kj} + \gamma_{kj})s, (\alpha_{kj} + \epsilon)s)$, $Q^{\infty}_{(k-1)\Delta,\, k\Delta}[n\Delta, \infty] \in ((\beta_k + \gamma_k)s, (\beta_k + \epsilon)s)$ *for all* $k = 1, \ldots, n$ *and* $j = k, \ldots, n$ *satisfies* $\tau_s < \tau_{\mathcal{A}}$. *Here* γ_{kj}, $\gamma_k > 0$, $\sum_{\substack{k=1,\ldots,n \\ j=k,\ldots,n}} \gamma_{kj} + \sum_{k=1,\ldots,n} \gamma_k = \gamma < \infty$ *and* $\epsilon > \gamma_{kj}$, $\epsilon > \gamma_k$.

PROOF. For $l = 1, \ldots, n$, consider

$$\begin{aligned}
\bar{Q}^{\infty}(l\Delta) &= \sum_{k=1}^{l} Q^{\infty}_{(k-1)\Delta,\, k\Delta}[l\Delta, \infty] \\
&> \sum_{k=1}^{l}\left(\sum_{j=l+1}^{n} a_{kj} s + b_k s\right) + \sum_{k=1}^{l}\left(\sum_{j=l+1}^{n} \gamma_{kj} s + \gamma_k s\right) \\
&= s \sum_{k=1}^{l}\left(\sum_{j=l+1}^{n} \int_{(k-1)\Delta}^{k\Delta} \psi_N'(\log(e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u))) \frac{F(j\Delta - u) - F((j-1)\Delta - u)}{e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u)}\, du \right. \\
&\quad \left. + \int_{(k-1)\Delta}^{k\Delta} \psi_N'(\log(e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u))) \frac{e^{\theta_{n\Delta}} \bar{F}(n\Delta - u)}{e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u)}\, du\right) \\
&\quad + s \sum_{k=1}^{l}\left(\sum_{j=l+1}^{n} \gamma_{kj} + \gamma_k\right) \\
&= s \int_{0}^{l\Delta} \psi_N'(\log(e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u))) \frac{e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u) - F(l\Delta - u)}{e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u)}\, du \\
&\quad + s \sum_{k=1}^{l}\left(\sum_{j=l+1}^{n} \gamma_{kj} + \gamma_k\right) \\
&> \lambda s \int_{0}^{l\Delta} \bar{F}(l\Delta - u)\, du + C_1 \sqrt{s}
\end{aligned}$$

for any given constant C_1, when s is large enough. The last inequality follows from the monotonicity of ψ_N'. Note that we then have $Q^{\infty}(l\Delta) = \bar{Q}^{\infty}(l\Delta) + r(l\Delta) > \lambda s + C_2 \sqrt{s}$ for any given constant C_2 and large enough s. Hence $\tau_{\mathcal{A}}$ is not reached in time $n\Delta$ when s is large.

On the other hand,

$$\begin{aligned}
\bar{Q}^{\infty}(n\Delta) &= \sum_{k=1}^{n} Q^{\infty}_{(k-1)\Delta,\, k\Delta}[n\Delta, \infty] \\
&> \sum_{k=1}^{n} \beta_k s + \sum_{k=1}^{n} \gamma_k s \\
&= s \sum_{k=1}^{m} \int_{(k-1)\Delta}^{k\Delta} \psi_N'(\log(e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u))) \frac{e^{\theta_{n\Delta}} \bar{F}(n\Delta - u)}{e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u)}\, du + s \sum_{k=1}^{n} \gamma_k \\
&= s \int_{0}^{n\Delta} \psi_N'(\log(e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u))) \frac{e^{\theta_{n\Delta}} \bar{F}(n\Delta - u)}{e^{\theta_{n\Delta}} \bar{F}(n\Delta - u) + F(n\Delta - u)}\, du + s \sum_{k=1}^{n} \gamma_k \\
&= s \psi_{n\Delta}'(\theta_{n\Delta}) + s \sum_{k=1}^{n} \gamma_k,
\end{aligned}$$

where the last equality follows from the definition of $\theta_{n\Delta}$. So $Q^\infty(n\Delta) = \bar{Q}^\infty(n\Delta) + r(n\Delta) > s$ when s is large enough. This concludes our proof. □

We now prove Theorem 4.2:

Proof of Theorem 4.2. Note that by Lemma 4.3, for any $r \in \mathcal{A}$ and s large enough,

$$\begin{aligned} &P_r(\tau_s < \tau_{\mathcal{A}}) \\ &\quad \geq P_r(Q^\infty_{(k-1)\Delta,\,k\Delta}[(j-1)\Delta, j\Delta] \in ((\alpha_{kj}+\gamma_{kj})s, (\alpha_{kj}+\epsilon)s),\ Q^\infty_{(k-1)\Delta,\,k\Delta}[n\Delta,\infty] \in ((\beta_k+\gamma_k)s, (\beta_k+\epsilon)s), \\ &\qquad k=1,\ldots,n,\ j=k,\ldots,n) \end{aligned} \tag{42}$$

for large enough s given arbitrary γ_{kj}, γ_k and ϵ satisfying conditions in Lemma 4.3. Denote $\boldsymbol{\Gamma} := (\gamma_{kj}, \gamma_k)_{k=1,\ldots,n,\ j=k,\ldots,n}$. Let

$$S_{\boldsymbol{\Gamma}} := \prod_{k=1}^n \prod_{j=k}^n (\alpha_{kj}+\gamma_{kj}, \alpha_{kj}+\epsilon) \times \prod_{k=1}^n (\beta_k+\gamma_k, \beta_k+\epsilon) \subset \mathbb{R}^{n(n+1)/2+n}.$$

Using Gartner-Ellis Theorem for (42) and Lemma 4.2, we have

$$\begin{aligned} &\frac{1}{s}\log P_r(Q^\infty_{(k-1)\Delta,\,k\Delta}[(j-1)\Delta, j\Delta] \in ((\alpha_{kj}+\gamma_{kj})s, (\alpha_{kj}+\epsilon)s), \\ &\qquad Q^\infty_{(k-1)\Delta,\,k\Delta}[n\Delta,\infty] \in ((\beta_k+\gamma_k)s, (\beta_k+\epsilon)s),\ k=1,\ldots,n,\ j=k,\ldots,n) \\ &\quad \to -I_{\boldsymbol{\Gamma}}, \end{aligned} \tag{43}$$

where $I_{\boldsymbol{\Gamma}} := \inf_{\mathbf{x}\in S_{\boldsymbol{\Gamma}}} I(\mathbf{x})$ and

$$I(\mathbf{x}) := \sup_{\boldsymbol{\Theta}\in\mathbb{R}^{n(n+1)/2+n}} \{\langle \boldsymbol{\Theta}, \mathbf{x}\rangle - \bar{\psi}(\boldsymbol{\Theta})\}$$

with $\bar{\psi}(\boldsymbol{\Theta})$ defined in Lemma 4.2. But note that for $k=1,\ldots,n$, $j=k,\ldots,n$,

$$\begin{aligned} \frac{\partial}{\partial\theta_{kj}}(\langle\boldsymbol{\Theta},\mathbf{x}\rangle - \bar{\psi}(\boldsymbol{\Theta})) &= x_{kj} - \int_{(k-1)\Delta}^{k\Delta} \psi_N'\left(\log\left(\sum_{j=k}^n e^{\theta_{kj}} P((j-1)\Delta - u < V \leq j\Delta - u) + e^{\theta_{k\cdot}}\bar{F}(n\Delta-u)\right)\right) \\ &\quad \cdot \frac{e^{\theta_{kj}} P((j-1)\Delta - u < V \leq j\Delta - u)}{\sum_{j=k}^m e^{\theta_{kj}} P((j-1)\Delta - u < V \leq j\Delta - u) + e^{\theta_{k\cdot}}\bar{F}(n\Delta-u)}\,du \end{aligned} \tag{44}$$

$$\begin{aligned} \frac{\partial}{\partial\theta_{k}}(\langle\boldsymbol{\Theta},\mathbf{x}\rangle - \bar{\psi}(\boldsymbol{\Theta})) &= x_{k} - \int_{(k-1)\Delta}^{k\Delta} \psi_N'\left(\log\left(\sum_{j=k}^n e^{\theta_{kj}} P((j-1)\Delta - u < V \leq j\Delta - u) + e^{\theta_{k\cdot}}\bar{F}(n\Delta-u)\right)\right) \\ &\quad \cdot \frac{e^{\theta_{k}} \bar{F}(n\Delta-u)}{\sum_{j=k}^m e^{\theta_{kj}} P((j-1)\Delta - u < V \leq j\Delta - u) + e^{\theta_{k\cdot}}\bar{F}(n\Delta-u)}\,du. \end{aligned} \tag{45}$$

Define $\mathbf{x}^* := (\alpha_{kj}, \beta_k)_{k=1,\ldots,n,\ j=k,\ldots,n}$. For $\mathbf{x} = \mathbf{x}^*$, it is straightforward to verify that $\Theta^* = (\theta^*_{kj}, \theta^*_{k\cdot})$, where $\theta^*_{kj} = 0$, $\theta^*_{k\cdot} = \theta_{n\Delta}$ for $k=1,\ldots,n$, $j=k,\ldots,n$, satisfies (44) and (45). Since $\langle\Theta,\mathbf{x}\rangle - \bar{\psi}(\Theta)$ is concave in $\boldsymbol{\Theta}$, we have

$$\begin{aligned} I(\mathbf{x}^*) &= \langle\Theta^*,\mathbf{x}^*\rangle - \bar{\psi}(\Theta^*) \\ &= \theta_{n\Delta}\sum_{k=1}^n \beta_k - \sum_{k=1}^n \int_{(k-1)\Delta}^{k\Delta} \psi_N(\log(F(n\Delta-u) - F((k-1)\Delta-u) + e^{\theta_{n\Delta}}\bar{F}(n\Delta-u)))\,du \\ &= \theta_{n\Delta}\psi'_{n\Delta}(\theta_{n\Delta}) - \psi_{n\Delta}(\theta_{n\Delta}) \\ &= I^*. \end{aligned}$$

Now since $\langle\Theta,\mathbf{x}\rangle - \bar{\psi}(\Theta)$ is continuously differentiable in Θ and $\mathbf{x}$, by implicit function theorem, $I(\mathbf{x})$ is continuous in $\mathbf{x}$. This implies that

$$I_{\boldsymbol{\Gamma}} \leq I(\mathbf{x}^* + \boldsymbol{\Gamma}) \to I(\mathbf{x}^*) = I^*$$

as $\boldsymbol{\Gamma} \to 0$. Together with (42) and (43) gives the conclusion. □

Theorems 4.1 and 4.2 together imply both the asymptotic optimality of Algorithm 2 and the large deviations of the loss probability:

Proof of Theorem 2.1. Note that by Jensen's inequality

$$P_r(\tau_s < \tau_{\mathcal{A}})^2 \le (E_r N_{\mathcal{A}})^2 \le \tilde{E}_r[N_{\mathcal{A}}^2 L^2].$$

Hence using Theorems 4.1 and 4.2 yields

$$-2I^* \le \lim_{s\to\infty} \frac{1}{s} \log P_r(\tau_s < \tau_{\mathcal{A}})^2 \le \lim_{s\to\infty} \frac{1}{s} \log (E_r N_{\mathcal{A}})^2 \le \lim_{s\to\infty} \frac{1}{s} \log \tilde{E}_r[N_{\mathcal{A}}^2 L^2] \le -2I^*.$$

Combining Proposition 2.1, we conclude that the steady-state loss probability given by (7) decays exponentially with rate I^* and that Algorithm 2 is asymptotically optimal. □

5. Logarithmic estimate of return time. In this section we will lay out the argument for Proposition 2.1. The first step is to reduce the problem to a $GI/G/\infty$ calculation. Define $x(t) := \sup\{y\colon Q^\infty(t, y) > 0\}$ as the maximum residual service times among all customers present at time t.

Lemma 5.1. *We have* $\tau_{\mathcal{A}} \le \tau'_{\mathcal{A}}$ *where*

$$\tau'_{\mathcal{A}} := \inf\{t \in \{\Delta, 2\Delta, \dots\}\colon x(t-u) \le l,\ Q^\infty(w) < s \text{ for } w \in [t-u, t] \text{ for some } u > l,\ Q^\infty(t, \cdot) \in J(\cdot)\}$$

for any $l > 0$.

Proof. The way we couple the $GI/G/\infty$ system implies that at any point of time the number of customers in the $GI/G/s$ system is at most that of the coupled $GI/G/\infty$ system (in fact, the served customers in the $GI/G/s$ system is a subset of those in $GI/G/\infty$). Suppose at time $t-u$ we have $Q^\infty(t-u) < s$ and $x(t-u) < l$. Then $Q^\infty(w) < s$ for $w \in [t-u, t]$ means that all the arrivals in this interval are not lost; i.e., they all get served in both the $GI/G/\infty$ and the $GI/G/s$ system. Since $x(t-u) \le l$, all the customers present at time t come from arrivals after time $t-u$. This implies that $Q(t, \cdot) \equiv Q^\infty(t, \cdot)$. Hence the result of the lemma. □

The next step is to find a mechanism to identify the instant $t-u$ and set an appropriate value for l so that $\tau'_{\mathcal{A}}$ is small. We use a geometric trial argument. Divide the time frame into blocks separated at $T_0 = 0, T_1, T_2, \dots$ in such a way that (1) a "success" in the block would mean $\tau'_{\mathcal{A}}$ is reached before the end of the block and (2) $\{W_u, T_i < u \le T_{i+1}\}, i = 0, 1, \dots$ are roughly independent, where $W_u = (Q(u, \cdot), B(u))$ is the system status representation defined in §2.2. We then estimate the probability of "success" in a block and also the length of a block to obtain a bound for $\tau'_{\mathcal{A}}$.

At this point let us also introduce a fixed constant t_0 and state the following result:

Lemma 5.2. *For any fixed* $t_0 > 0$,

$$P\left(\bar{Q}^\infty(t, y) \in \left(\lambda s \int_y^{t+y} \bar{F}(u)\, du \pm \sqrt{s} C_1 \nu(y)\right) \text{ for all } t \in [0, t_0],\ y \in [0, \infty) \,\middle|\, B(0)\right) \ge C_2 > 0 \tag{46}$$

and

$$P\left(\bar{Q}^\infty(t, y) \notin \left(\lambda s \int_y^{t+y} \bar{F}(u)\, du \pm \sqrt{s} C_1 \nu(y)\right) \text{ for some } t \in [0, t_0],\ y \in [0, \infty) \,\middle|\, B(0)\right) \ge C_3 > 0 \tag{47}$$

for large enough $C_1 > 0$ *and some constants* C_2 *and* C_3, *all independent of* s, *uniformly for all initial age* $B(0)$. $\nu(y)$ *is defined in* (11).

To prove this lemma, the main idea is to consider the diffusion limit of $Q^\infty(t, y)$ as a two-dimensional Gaussian field and then invoke the Borell-TIS inequality (Adler [1]). By Pang and Whitt [21] we know

$$\frac{Q^\infty(t, y) - \lambda s \int_y^{t+y} \bar{F}(u)\, du}{\sqrt{s}} \Rightarrow R(t, y)$$

in the space $\mathcal{D}_{\mathcal{D}[0,\infty)}[0, \infty)$, where

$$R(t, y) := R_1(t, y) + R_2(t, y) \tag{48}$$

is a two-dimensional Gaussian field given by

$$R_1(t,y) := \int_0^t \int_0^\infty I(u+x > t+y)\, dK(u,x) \tag{49}$$

and

$$R_2(t,y) := \sqrt{\lambda c_a^2} \int_0^t \bar{F}(t+y-u)\, dW(u), \tag{50}$$

where $W(\cdot)$ is a standard Brownian motion, and $K(u,x) := W(\lambda u, F(x)) - F(x)W(\lambda u, 1)$ in which $W(\cdot,\cdot)$ is a standard Brownian sheet on $[0,\infty)\times[0,1]$. $W(\cdot)$ and $K(\cdot,\cdot)$ are independent processes. c_a is the coefficient of variation, i.e., ratio of standard deviation to mean, of the interarrival times in the base system defined in §2.1.

The key step is then to show an estimate of this limiting Gaussian process:

Lemma 5.3. *Fix $t_0 > 0$. For $i = 1, 2$, we have*

$$P(|R(t,y)| \le C_* \nu(y) \text{ for all } t \in [0,t_0],\ y \in [0,\infty)) > 0$$

for a well-chosen constant $C_* > 0$, where $R(\cdot,\cdot)$ and $\nu(\cdot)$ are defined in (48), (49), (50) and (11).

This lemma relies on the Borell-TIS inequality on the Gaussian process $R_i(t,y)$ for $i = 1, 2$. The verification of the conditions for the application of such inequality are tedious but routine and hence will be deferred to the appendix. Here we provide a brief outline of the arguments: For $i = 1, 2$,

Step 1: Define a d_i-metric (in fact a pseudo metric)

$$d_i((t,y),(t',y')) := E(\tilde{R}_i(t,y) - \tilde{R}_i(t,y))^2$$

where $\tilde{R}_i(t,y) := R_i(t,y)/\nu(y)$. Show that the domain $[0,t_0]\times[0,\infty]$ can be compactified under this (pseudo) metric.

Step 2: Use an entropy argument (see for example Adler [1]) to show that $E\sup_S \tilde{R}_i(t,y) < \infty$. In particular, $\tilde{R}_i(t,y)$ is a.s. bounded over S.

Step 3: Invoke the Borell-TIS inequality; i.e., for $x \ge E\sup_S \tilde{R}_i(t,y)$,

$$P\left(\sup_S \tilde{R}_i(t,y) \ge x\right) \le \exp\left\{-\frac{1}{2\sigma_i^2}\left(x - E\sup_S \tilde{R}_i(t,y)\right)^2\right\}$$

where

$$\sigma_i^2 := \sup_S E\tilde{R}_i(t,y)^2.$$

From these steps, it is straightforward to conclude Lemma 5.3. The rest of the proof of Lemma 5.2 is to show the uniformity over U_0 in the weak limit of $\bar{Q}^\infty$ to R. This is done by restricting to the set $U_0 \le x$ for $x = O(1/s)$ and using the light tail property of U_0. The details are provided in the appendix.

We need one more lemma:

Lemma 5.4. *Let V_k be r.v. with distribution function $F(\cdot)$ satisfying the light-tail assumption in (4). For any $p > 0$, we have*

$$E\left(\max_{k=1,\dots,n} V_k\right)^p = O(l_p(n)^p) = o(n^\epsilon),$$

where

$$l_p(n) := \inf\left\{y:\ np\int_y^\infty u^{p-1}\bar{F}(u)\, du < \eta\right\} \tag{51}$$

for a constant $\eta > 0$ and ϵ is any positive number.

Proof. Let $\bar{F}_n(x) = P(\max_{k=1,\dots,n} V_k > x)$. Note that

$$E\left(\max_{k=1,\dots,n} V_k\right)^p = p\int_0^\infty u^{p-1}\bar{F}_n(u)\, du \le y^p + np\int_y^\infty u^{p-1}\bar{F}(u)\, du$$

for any $y \ge 0$. Pick $y = l_p(n)$. Then

$$E\left(\max_{k=1,\dots,n} V_k\right)^p = O(l_p(n)^p).$$

Using (5) we have $O(l_p(n)^p) = O(n^\epsilon)$ for any $\epsilon > 0$. □

We are now ready to prove Proposition 2.1, for which we need the following construction. Pick $\gamma = 1/t_0$ where γ is introduced in (11) and $\xi(y)$ is defined in (10). Recall C_1 as in Lemma 5.2. Define T_i, $i = 0, 1, 2, \ldots$ as follows: Given T_{i-1}, define

$$\begin{aligned}
v(s) &:= \inf\{y\colon \sqrt{s}C_1\xi(y) < 1/2\}\\
z &:= \inf\{kt_0\colon k = 1, 2, \ldots,\ kt_0 \geq v(s) + \Delta\}\\
x_i &:= x(T_{i-1})\\
w_i &:= \inf\{kt_0, k = 1, 2, \ldots : kt_0 \geq x_i\}\\
d_i &:= A_{N_s(T_{i-1}+S_i)+1} - (T_{i-1} + S_i) \quad \text{i.e., } d_i \text{ is the time of the first arrival after } T_{i-1} + S_i\\
T_i &:= T_{i-1} + w_i + d_i + z.
\end{aligned}$$

Note that w_i and z are multiples of t_0. For convenience define, for $u < t$, $\bar{Q}_u^\infty(t, y) := \bar{Q}^\infty(u + t, y) - \bar{Q}^\infty(u, t + y)$ as the number of arrivals after time u that have residual service time larger than y at time $u + t$. We define a "success" in block i to be the event ζ_i that all of the following occur: (1) $\bar{Q}^\infty_{T_{i-1}+(k-1)t_0}(t, y) \in (\lambda s \int_y^{t+y} \bar{F}(u)\,du \pm \sqrt{s}C_1\nu(y))$ for all $t \in [0, t_0]$, for every $k = 1, 2, \ldots, w_i/t_0$. (2) $d_i \leq c/s$ for a small constant $c > 0$. (3) $Q^\infty_{T_{i-1}+w_i+d_i+(k-1)t_0}(t, y) \in (\lambda s \int_y^{t+y} \bar{F}(u)\,du \pm \sqrt{s}C_1\nu(y))$ for all $t \in [0, t_0]$, for every $k = 1, 2, \ldots, z/t_0$.

Roughly speaking, ζ_i occurs when the $GI/G/\infty$ system behaves "normally" for a long enough period so that $Q^\infty(t)$ keeps within capacity for that period and the steady-state confidence band $J(\cdot)$ is reached at the end. More precisely, starting from T_{i-1} and given $x(T_{i-1})$, $T_{i-1} + w_i$ is the time when all customers in the previous block have left. Adjusting for the age at time $T_{i-1} + w_i$, starting from $T_{i-1} + w_i + d_i$, z is a long enough time so that the system would fall into $J(\cdot)$ if it behaves normally in each steps of size t_0 throughout the period. It can be seen by summing up the interval boundaries that the occurrence of ζ_i ensures $\tau'_{\mathcal{A}}$ is reached during the last Δ units of time before T_i.

Proof of Proposition 2.1. We first check that the occurrence of event ζ_i implies that $\tau'_{\mathcal{A}}$ is reached during the last Δ units of time before T_i. As discussed above, since $w_i \geq x_i$, all the customers at time $T_{i-1} + w_i$ will be those arriving after time T_{i-1}. Hence the occurrence of ζ_i implies that

$$\begin{aligned}
&Q^\infty(T_{i-1} + w_i, y)\\
&\quad\in \left(\lambda s \sum_{k=1}^{w_i/t_0} \int_{(k-1)t_0+y}^{kt_0+y} \bar{F}(u)\,du \pm \sqrt{s}C_1 \sum_{k=1}^{w_i/t_0} \nu((k-1)t_0 + y)\right)\\
&\quad\subset \left(\lambda s \int_y^{w_i+y} \bar{F}(u)\,du \pm \sqrt{s}C_1\left[\nu(y) + \frac{1}{t_0}\int_y^\infty \nu(u)\,du\right]\right)\\
&\quad\subset \left(\lambda s \int_y^{w_i+y} \bar{F}(u)\,du \pm \sqrt{s}C_1\xi(y)\right) \qquad (52)
\end{aligned}$$

and

$$Q^\infty(T_{i-1} + w_i + d_i, y) \in \left(\lambda s \int_{d_i+y}^{w_i+d_i+y} \bar{F}(u)\,du \pm \sqrt{s}C_1\xi(d_i + y)\right).$$

For each $t \in ((k-1)t_0, kt_0]$, denote $[t] = t - (k-1)t_0$, for $k = 1, \ldots, z/t_0$. Then

$$\begin{aligned}
&Q^\infty(T_{i-1} + w_i + d_i + t, y)\\
&\quad\in \left(\lambda s \int_y^{t+y} \bar{F}(u)\,du + \lambda s \int_{d_i+y+t}^{w_i+d_i+y+t} \bar{F}(u)\,du\right.\\
&\qquad\left. \pm \sqrt{s}C_1\left[\sum_{j=1}^{w_i/t_0} \nu((j-1)t_0 + d_i + (k-1)t_0 + [t] + y) + \nu(y) + \sum_{j=2}^{k} \nu((j-2)t_0 + [t] + y)I(k > 1)\right]\right)\\
&\quad\subset \left(\lambda s \int_y^{t+y} \bar{F}(u)\,du + \lambda s \int_{d_i+y+t}^{w_i+d_i+y+t} \bar{F}(u)\,du \pm \sqrt{s}C_1\left[\sum_{j=1}^{w_i/t_0+k-1} \nu((j-1)t_0 + [t] + y) + \nu(y)\right]\right)\\
&\quad\subset \left(\lambda s \int_y^{t+y} \bar{F}(u)\,du + \lambda s \int_{d_i+y+t}^{w_i+d_i+y+t} \bar{F}(u)\,du \pm \sqrt{s}C_1\left[2\nu(y) + \frac{1}{t_0}\int_y^\infty \nu(u)\,du\right]\right)\\
&\quad\subset \left(\lambda s \int_y^{t+y} \bar{F}(u)\,du + \lambda s \int_{d_i+y+t}^{w_i+d_i+y+t} \bar{F}(u)\,du \pm \sqrt{s}C'\xi(y)\right), \qquad (53)
\end{aligned}$$

where $C' = 2C_1$ (which depends on γ).

It is now obvious that ζ_i implies $Q^\infty(t) < s$ for $[T_{i-1}+w_i, T_i]$. By the definition of $v(s)$, (52), and that $\lambda s\int_y^\infty \bar F(u)\,du$ is smaller and decays faster than $\sqrt{s}C_1\xi(y)$ for $y \ge v(s)$ when s is large, we get $x(T_{i-1}+w_i) \le v(s) \le z$. Let $\tilde T_i = \sup\{k\Delta\colon k\Delta \le T_i\}$ be the largest time before T_i such that $\mathcal{A}$ can possibly be hit in the Δ-skeleton. It remains to show that $Q^\infty(\tilde T_i, y) \in J(y)$ to conclude that ζ_i implies a hit on $\tau'_{\mathcal{A}}$.

From (53), for $t \in [T_{i-1}+w_i+d_i, T_i]$,

$$Q^\infty(t,y) \in \left(\lambda s\int_y^{t-T_{i-1}+y} \bar F(u)\,du - \lambda s\int_{t-T_{i-1}-w_i-d_i+y}^{t-T_{i-1}-w_i+y} \bar F(u)\,du \pm \sqrt{s}C'\xi(y)\right).$$

In particular,

$$\begin{aligned} Q^\infty(\tilde T_i, y) &\in \left(\lambda s\int_y^{\tilde T_i-T_{i-1}+y} \bar F(u)\,du - \lambda s\int_{\tilde T_i-T_{i-1}-w_i-d_i+y}^{\tilde T_i-T_{i-1}-w_i+y} \bar F(u)\,du \pm \sqrt{s}C'\xi(y)\right)\\ &= \left(\lambda s\int_y^{\infty} \bar F(u)\,du - \lambda s\int_{\tilde T_i-T_{i-1}+y}^{\infty} \bar F(u)\,du - \lambda s\int_{\tilde T_i-T_{i-1}-w_i-d_i+y}^{\tilde T_i-T_{i-1}-w_i+y} \bar F(u)\,du \pm \sqrt{s}C'\xi(y)\right). \end{aligned} \tag{54}$$

Now note that

$$\lambda s\int_{\tilde T_i-T_{i-1}+y}^{\infty} \bar F(u)\,du + \lambda s\int_{\tilde T_i-T_{i-1}-w_i-d_i+y}^{\tilde T_i-T_{i-1}-w_i+y} \bar F(u)\,du \le 2\lambda s\int_{v(s)+y}^{\infty} \bar F(u)\,du$$

and we claim that it is further bounded from above by $\sqrt{s}C\xi(y)$ for arbitrary constant C when s is large enough, uniformly over $y \in [0,\infty)$. In fact, we have $v(s) \ge \inf\{y\colon s\int_y^\infty \bar F(u) \le \alpha\}$ for any $\alpha > 0$ when s is large enough. Now when $\sqrt{s}C\xi(y) < \alpha/(2\lambda)$, $s\int_{v(s)+y}^\infty \bar F(u)\,du \le s\int_y^\infty \bar F(u)\,du$, which is smaller and decays faster than $\sqrt{s}C\xi(y)$ when s is large. When $\sqrt{s}C\xi(y) \ge \alpha/(2\lambda)$, we have $s\int_{v(s)+y}^\infty \bar F(u)\,du \le s\int_{v(s)}^\infty \bar F(u)\,du \le \alpha/(2\lambda)$. Picking $C^* = C' + C$ where C^* is defined in (9), we conclude that ζ_i implies $\tau'_{\mathcal{A}}$ is reached at $\tilde T_i$.

Now let $N := \inf\{i\colon \zeta_i \text{ occurs}\}$. Consider (suppressing the initial conditions), for any $p > 0$,

$$\begin{aligned} &E(\tau'_{\mathcal{A}})^p\\ &\quad= E\left[\sum_{i=1}^{N}(w_i+d_i+z)\right]^p\\ &\quad= E\left[\sum_{i=1}^{\infty}(w_i+d_i+z)I(N\ge i)\right]^p\\ &\quad\le \left(\sum_{i=1}^{\infty}(E[(w_i+d_i+z)^p; N\ge i])^{1/p}\right)^p\\ &\quad\le \left(\sum_{i=1}^{\infty}(E(w_i+d_i+z)^{pq})^{1/(pq)}(P(N\ge i))^{1/(pr)}\right)^p \end{aligned} \tag{55}$$

where $q, r > 0$ and $1/q + 1/r = 1$, by using Minkowski's inequality and Holder's inequality in the first and second inequality, respectively.

For $i = 2, 3, \ldots,$ we have

$$E(w_i+d_i+z)^{pq} \le [(Ew_i^{pq})^{1/(pq)} + (Ed_i^{pq})^{1/(pq)} + z]^{pq} \tag{56}$$

by Minkowski's inequality again.

We now analyze $E(w_i+d_i+z)^p$ for any $p > 0$. From now on C denotes constant, not necessarily the same every time it appears. First note that

$$(Ed_i^p)^{1/p} \le d^{(p)} := \sup_{b\ge 0}(E[d_i^p \mid B(T_{i-1}+w_i) = b])^{1/p} = \frac{1}{s}\sup_{b\ge 0}(E[(U^0-b)^p \mid B^0(0)=b])^{1/p} = O\left(\frac{1}{s}\right) \tag{57}$$

and $z \le v(s) + \Delta + t_0 = o(s^\epsilon)$ for any $\epsilon > 0$. The last equality of (57) comes from the light-tail assumption on U^0. Indeed, since U^0 is light tailed, we have

$$\exp\left\{-\int_0^x h_U(u)\,du\right\} = \bar F_U(x) \le e^{-cx}$$

for some $c > 0$, where $h_U(\cdot)$ and $\bar{F}(x)$ are the hazard rate function and tail distribution function, respectively, of U^0. This implies that $h(x) \geq c$ for all $x \geq 0$. Then

$$\sup_{b\geq 0} P(U^0 - b > x \mid U^0 > b) = \sup_{b\geq 0} \exp\left\{-\int_b^{x+b} h(u)\,du\right\} \leq e^{-cx}$$

and so

$$\sup_{b\geq 0} E[(U^0-b)^p \mid B^0(0) = b] = \sup_{b\geq 0} p\int_0^\infty x^{p-1} P(U^0 - b > x \mid U^0 > b)\,dx \leq p\int_0^\infty x^{p-1}e^{-cx}\,dx < \infty.$$

For $i = 1$, $w_1 \leq l(s) + t_0 = o(s^\epsilon)$ where $l(s)$ is defined in (14). Hence $E(w_1 + d_1 + z)^p \leq [(Ew_1^p)^{1/p} + (Ed_1^p)^{1/p} + z]^p = o(s^\epsilon)$ for any $\epsilon > 0$.

Now

$$\begin{aligned} Ew_i^p &\leq E\left[\left(\max_{i=1,\ldots,N_s(T_{i-1})-N_s(T_{i-2})} V_i\right)^p\right] \\ &= E\left[E\left[\left(\max_{i=1,\ldots,N_s(T_{i-1})-N_s(T_{i-2})} V_i\right)^p \,\middle|\, N_s(T_{i-1}) - N_s(T_{i-2})\right]\right] \\ &\leq CE[l_p(N_s(T_{i-1}) - N_s(T_{i-2}))^p] \quad \text{for some constant } C = C(p) \text{ and } l_p(\cdot) \text{ defined in (51)} \\ &\leq CE[(N_s(T_{i-1}) - N_s(T_{i-2}))^\epsilon] \quad \text{for constant } C = C(p,\epsilon) \end{aligned} \tag{58}$$

for any $\epsilon > 0$, by Lemma 5.4. Pick $\epsilon < 1$. By Jensen's inequality and the elementary renewal theorem, (58) is less than or equal to

$$\begin{aligned} &C(E[N_s(T_{i-1}) - N_s(T_{i-2})])^\epsilon \\ &\quad = C(E[N_s(T_{i-1}) - N_s(T_{i-2}) \mid T_{i-1} - T_{i-2}])^\epsilon \\ &\quad \leq C(E[\tilde{\lambda}s(T_{i-1} - T_{i-2})])^\epsilon \quad \text{for some } \tilde{\lambda} > \lambda \\ &\quad = C\tilde{\lambda}^\epsilon s^\epsilon (E[T_{i-1} - T_{i-2}])^\epsilon \\ &\quad = C\tilde{\lambda}^\epsilon s^\epsilon (E[w_{i-1} + d_{i-1} + z])^\epsilon. \end{aligned} \tag{59}$$

Let $y_i := E[w_i + d_i + z]$. We then have

$$y_i = Cs^\epsilon y_{i-1}^\epsilon + d^{(1)} + z.$$

By construction $y_i \geq t_0$, and since $v(s) = o(s^\epsilon)$, for any $\epsilon > 0$ we have

$$d^{(1)} + z \leq Cs^\epsilon t_0^\epsilon \leq Cs^\epsilon y_i^\epsilon$$

for large enough s, uniformly over i. Hence

$$y_i \leq Cs^\epsilon y_{i-1}^\epsilon + d^{(1)} + z \leq Cs^\epsilon y_{i-1}^\epsilon.$$

Now we can write

$$\begin{aligned} y_i &\leq Cs^\epsilon y_{i-1}^\epsilon \leq Cs^\epsilon (Cs^\epsilon y_{i-2}^\epsilon)^\epsilon = C^{1+\epsilon} s^{\epsilon+\epsilon^2} y_{i-2}^{\epsilon^2} \\ &\cdots \leq (C^{1/(1-\epsilon)} \vee 1) s^{\epsilon/(1-\epsilon)} y_1^{\epsilon^{i-1}} = o(s^\rho) \end{aligned} \tag{60}$$

for any $\rho > 0$ by choosing ϵ, uniformly over i.

Therefore from (56), (59) and (60), we get

$$E(w_i + d_i + z)^{pq} = o(s^\epsilon) \tag{61}$$

for any $\epsilon > 0$ uniformly over i.

Now consider

$$\begin{aligned} P(N \geq 1) &= P(\zeta_1^c) = 1 - P(\zeta_1) \\ &\leq 1 - P\big(d_1 \leq c/s\big) C_2^{(w_1+z)/t_0} \\ &\qquad \text{where } C_2 \text{ is defined in Lemma 5.2 and } c \text{ is defined in the discussion of } \zeta_i \\ &\leq 1 - be^{-a(w_1+z)} \\ &= 1 - be^{-o(s^\epsilon)} \end{aligned} \tag{62}$$

for some constants $a > 0$ and $0 < b < 1$ and any $\epsilon > 0$. Moreover, for $i = 2, 3, \ldots,$

$$\begin{aligned} P(N \geq i) &= P(N \geq i-1) P(\zeta_{i-1}^c \mid N \geq i-1) \\ &\leq P(N \geq i-1) E[1 - be^{-a(w_{i-1}+z)} \mid N \geq i-1] \\ &\leq P(N \geq i-1)(1 - be^{-a(E[w_{i-1} \mid N \geq i-1]+z)}) \end{aligned} \tag{63}$$

by Jensen's inequality and that the function $1 - be^{-a(\cdot+z)}$ is concave.

Consider $E[w_i \mid N \geq i]$ for any $i = 2, 3, \ldots$. We have

$$E[w_i \mid N \geq i] = E[E[w_i \mid \zeta_{i-1}^c, w_{i-1} + d_{i-1} + z] \mid N \geq i]. \tag{64}$$

Now by singling out failure in the first trial of t_0 (see the discussion on ζ_i), we get

$$P(\zeta_{i-1}^c \mid w_{i-1} + d_{i-1} + z) \geq C_3$$

where C_3 is defined in Lemma 5.2, uniformly over $w_{i-1} + d_{i-1} + z$. Hence

$$\begin{aligned} C_3 E[w_i \mid \zeta_{i-1}^c, w_{i-1} + d_{i-1} + z] &\leq \int P(\zeta_{i-1}^c \mid w_{i-1} + d_{i-1} + z) E[w_i \mid \zeta_{i-1}^c, w_{i-1} + d_{i-1} + z] P(w_{i-1} + d_{i-1} + z \in dx) \\ &\leq E w_i, \end{aligned}$$

which gives

$$E[w_i \mid \zeta_{i-1}^c, w_{i-1} + d_{i-1} + z] \leq \frac{E w_i}{C_3}$$

uniformly over $w_{i-1} + d_{i-1} + z$. Therefore, (64) is bounded from above by $E w_i / C_3$.

From (59) and (60) we know that $E w_i = o(s^\epsilon)$ for any $\epsilon > 0$, so (63) is less than or equal to

$$P(N \geq i-1)(1 - be^{-a(E w_{i-1}/C_3+z)}) = P(N \geq i-1)(1 - be^{-o(s^\epsilon)}) \tag{65}$$

for any $\epsilon > 0$ uniformly over i.

By (55), (62), (61), and (65) we get

$$\begin{aligned} E{\tau'_{\mathcal{A}}}^p &\leq o(s^\epsilon) \left(\sum_{i=1}^{\infty} (P(N \geq i))^{1/(pr)} \right)^p \\ &\leq o(s^\epsilon) \left(\sum_{i=1}^{\infty} (1 - be^{-o(s^\epsilon)})^{i/(pr)} \right)^p \\ &\leq o(s^\epsilon) \frac{1}{[1 - (1 - be^{-o(s^\epsilon)})^{1/(pr)}]^p} \\ &\leq o(s^\epsilon) e^{o(s^\epsilon)}. \end{aligned}$$

Hence

$$\frac{1}{s} \log E{\tau'_{\mathcal{A}}}^p \leq \frac{\epsilon}{s} + \frac{o(s^\epsilon)}{s} \to 0$$

as $s \to \infty$. On the other hand, we pick $\mathcal{A}$ such that $\tau_{\mathcal{A}} \geq \Delta$ and so

$$\frac{1}{s} \log E\tau_{\mathcal{A}}^p \geq \frac{1}{s} \log \Delta^p \to 0.$$

Conclusion follows for (12).

For (13), note that $N_{\mathcal{A}} \leq N_s(\tau_{\mathcal{A}}) \leq N_s(\tau'_{\mathcal{A}})$ and $E N_s(t)^p = O(st)$ since $(1/s) \log E e^{\theta N_s(t)} \to -\psi_N(\theta) t$. Hence

$$E N_s(\tau'_{\mathcal{A}})^p \leq O(s^p) E(\tau'_{\mathcal{A}})^p$$

and the result follows from (12). □

TABLE 1. Simulation results from crude Monte Carlo and importance sampler.

	Crude Monte Carlo			Importance sampler		
s	Estimate	R.E.	C.I.	Estimate	R.E.	C.I.
10	0.05318	0.0265	(0.05252, 0.05384)	0.05412	0.130	(0.05084, 0.05740)
30	0.003174	0.111	(0.003009, 0.003338)	0.003204	0.570	(0.002349, 0.004060)
60	7.0922×10^{-5}	1.388	$(2.4847 \times 10^{-5}, 1.1700 \times 10^{-4})$	6.2585×10^{-5}	2.258	$(-3.5529 \times 10^{-6}, 1.2872 \times 10^{-4})$
80	6.9444×10^{-7}	4.472	$(-7.5904 \times 10^{-7}, 2.1479 \times 10^{-6})$	4.5001×10^{-8}	1.879	$(5.4365 \times 10^{-9}, 8.4565 \times 10^{-8})$
100	0	*N/A*	*N/A*	8.1178×10^{-10}	2.296	$(-6.0511 \times 10^{-11}, 1.6841 \times 10^{-9})$
120	0	*N/A*	*N/A*	1.3025×10^{-10}	4.472	$(-1.4237 \times 10^{-10}, 4.0286 \times 10^{-10})$

REMARK 5.1. The proof of Proposition 2.1 can be simplified when the service time has bounded support, say on $[0, M]$. In this case the $GI/G/\infty$ system is "$M + U_0$-independent"; i.e., W_t^∞, the state of the system at time t, and $W^\infty_{A_{N_s(t)+1}+M}$, the state of the system at M time units after the first arrival since time t, are independent. As a result we can merely set $v(s) := M$ and $x_i := M$ for any i, and the same argument as above will apply.

6. Numerical example. We close this paper with a numerical example for $GI/G/s$. We set the interarrival times in the base system to be Gamma$(1/2, 1/2)$ so $\lambda = 1$. For illustrative convenience we set the service times as Uniform$(0, 1)$. Hence traffic intensity is $1/2$. In this case, we can simply set $C^* = 1$ and $\xi(y) = \mathrm{sd}(R(\infty, y)) \vee C_1 = \sqrt{\lambda \int_y^\infty F(u)\bar{F}(u)\,du + \lambda c_a^2 \int_y^\infty \bar{F}(u)^2\,du} \vee C_1$ with $C_1 = 1.1$ (note that $\eta = 0$, and we use a truncated $\xi(y)$; the validity of this simpler choice than the one displayed in §2.5 can be verified from the arguments in §5 specialized to the case of bounded service time). Also we choose $\Delta = 1$, $\delta = c/s$ with $c = 1$, and $T = 1.54$, which satisfies (24). To test the numerical efficiency of our importance sampling algorithm, we compare it with crude Monte Carlo scheme using increasing values of s; namely, $s = 10$, 30, 60, 80, 100, and 120.

As discussed in §3, since we run our importance sampler every time we hit set $\mathcal{A}$, the initial positions of the importance samplers are dependent. To get an unbiased estimate of standard error, we group the samples into batches and obtain statistics based on these batch samples (see Asmussen and Glynn [5]). To make the estimates and statistics comparable, for each experiment we run the computer for roughly 120 seconds CPU time and always use 20 batches. In Table 1, we output the estimates of loss probability, the relative errors (ratios of sample standard deviation to sample mean), and 95% confidence intervals for both crude Monte Carlo scheme and importance sampler under different values of s.

When s is small we see that crude Monte Carlo performs slightly better than our importance sampler. However, when s is greater than 80, importance sampler starts to perform better. When s is greater than 100, crude Monte Carlo totally breaks down while our importance sampler still gives estimates that have encouragingly small relative error.

We can also analyze the graphical depiction of the sample paths. Figures 5(a) and (b) are two sample paths run by Algorithm 2, initialized at the mean of $Q(t, y)$, i.e., $\lambda s \int_y^\infty \bar{F}(u)\,du$. Figure 5(a) is a contour plot of $Q(t, y)$,

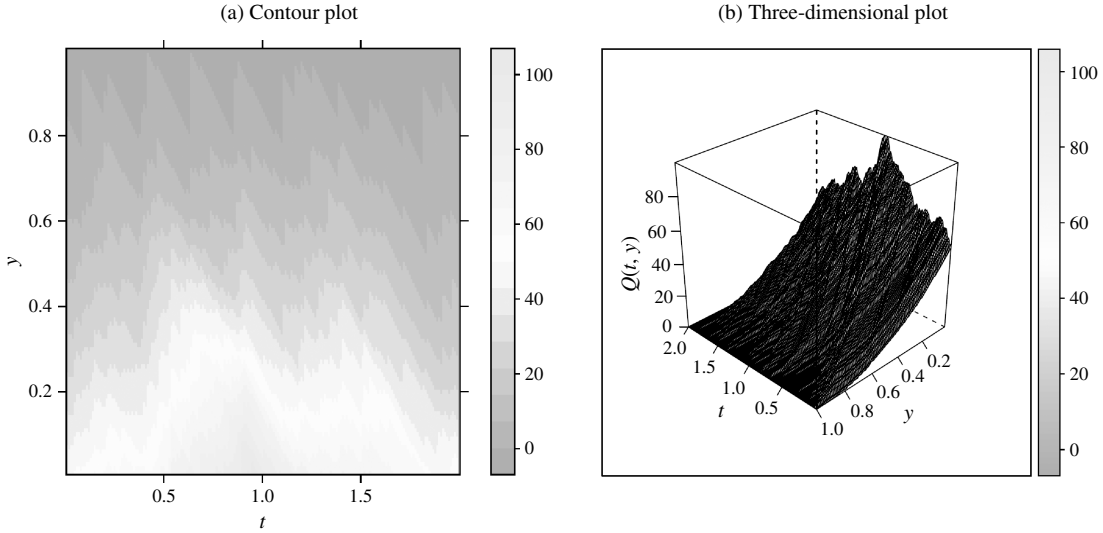

FIGURE 5. Plots of $Q(t, y)$.

whereas Figure 5(b) is a three-dimensional plot of another $Q(t, y)$. As we can see, the number of customers (the shading intensity along the t-axis) increases from time 0 to around 0.95 when it hits overflow in the contour plot. Similar trajectory appears in the three-dimensional plot. These plots are potentially useful for operations manager to judge the possibility of overflow over a finite horizon given the current state.

Acknowledgments. The authors are grateful to the referees for useful comments which improved the quality of the paper. National Science Foundation support [Grant DMS 1320550, CMMI 1069064] for the first author, and National Security Agency [Grant H98230-13-1-0301] for the second author are gratefully acknowledged.

Appendix A. Technical proofs.

A.1. Proof of Lemma 3.1. The domain of $\psi_t(\cdot)$ is easily seen to inherit from $\psi_N(\cdot)$. Write

$$\psi_t(\theta) = \int_0^t \psi_N(\log(e^\theta \bar{F}(u) + F(u)))\, du.$$

Note that

$$\frac{\partial}{\partial\theta}\psi_N(\log(e^\theta \bar{F}(u) + F(u))) = \psi_N'(\log(e^\theta \bar{F}(u) + F(u)))\frac{e^\theta \bar{F}(u)}{e^\theta \bar{F}(u) + F(u)}$$

is continuous in u and θ. Hence

$$\psi_t'(\theta) = \int_0^t \psi_N'(\log(e^\theta \bar{F}(u) + F(u)))\frac{e^\theta \bar{F}(u)}{e^\theta \bar{F}(u) + F(u)}\, du$$

(see Rudin [25], p. 236, Theorem 9.42). Moreover, $\psi_N'(\log(e^\theta \bar{F}(u)+F(u)))e^\theta \bar{F}(u)/(e^\theta \bar{F}(u)+F(u))$ is uniformly continuous in u and a neighborhood of θ, for any $\theta \in \mathbb{R}$. Hence $\psi_t'(\theta)$ is continuous in θ. Also the strict monotonicity of $\psi_N'(\cdot)$ implies that $\psi_t'(\theta)$ too is strictly increasing for any $\theta > 0$.

Following the same argument, we have

$$\psi_t''(\theta) = \int_0^t \left[\psi_N''(\log(e^\theta \bar{F}(u) + F(u)))\left(\frac{e^\theta \bar{F}(u)}{e^\theta \bar{F}(u) + F(u)}\right)^2 + \psi_N'(\log(e^\theta \bar{F}(u) + F(u)))\frac{F(u)\bar{F}(u)e^\theta}{(e^\theta \bar{F}(u) + F(u))^2}\right] du,$$

which is continuous in θ.

Finally, note that as $\theta \nearrow \infty$, $\psi_N'(\log(e^\theta \bar{F}(u)+F(u)))e^\theta \bar{F}(u)/(e^\theta \bar{F}(u)+F(u)) \nearrow \infty$ for any $u \in \operatorname{supp} \bar{F}$ since $\psi_N(\cdot)$ is steep. By monotone convergence theorem we conclude that $\psi_t(\cdot)$ is steep.

A.2. Proof of Lemma 3.2. (i) Denote $\theta(t) = \theta_t$ for convenience. Since $\psi_t'(\cdot)$ is continuously differentiable by Lemma 3.1, by implicit function theorem, we can differentiate $\psi_t'(\theta(t)) = a_t$ with respect to t on both sides to get

$$\psi_N'(\log(e^{\theta(t)} \bar{F}(t) + F(t)))\frac{e^{\theta(t)} \bar{F}(t)}{e^{\theta(t)} \bar{F}(t) + F(t)} + \int_0^t \left[\psi_N''(\log(e^{\theta(t)} \bar{F}(u) + F(u)))\left(\frac{e^{\theta(t)} \bar{F}(u)}{e^{\theta(t)} \bar{F}(u) + F(u)}\right)^2 + \psi_N'(\log(e^{\theta(t)} \bar{F}(u) + F(u)))\frac{F(u)\bar{F}(u)e^{\theta(t)}}{(e^{\theta(t)} \bar{F}(u) + F(u))^2}\right] du\theta'(t) = \lambda \bar{F}(t)$$

which gives

$$\theta'(t) = \frac{\lambda \bar{F}(t) - \psi_N'(\log(e^{\theta(t)} \bar{F}(t)+F(t)))e^{\theta(t)} \bar{F}(t)/(e^{\theta(t)} \bar{F}(t)+F(t))}{\int_0^t [\psi_N''(\log(e^{\theta(t)} \bar{F}(u)+F(u)))((e^{\theta(t)} \bar{F}(u))/(e^{\theta(t)} \bar{F}(u)+F(u)))^2 + \psi_N'(\log(e^{\theta(t)} \bar{F}(u)+F(u)))((F(u)\bar{F}(u)e^{\theta(t)})/(e^{\theta(t)} \bar{F}(u)+F(u))^2)]\, du} \le 0.$$

The inequality holds because

$$g_t(\theta) := \psi_N'(\log(e^\theta \bar{F}(t) + F(t)))\frac{e^\theta \bar{F}(t)}{e^\theta \bar{F}(t) + F(t)} \tag{66}$$

is nondecreasing in θ and $g_t(0) = \lambda \bar{F}(t)$ and $\psi_N(\cdot)$ is nondecreasing and convex. Hence $\theta(t)$ is nonincreasing.

(ii) Since $a_t \ge 1 - \lambda EV$, $\theta_t \ge \bar{\theta}_t$ where $\bar{\theta}_t$ satisfies $\psi_t'(\bar{\theta}_t) = 1 - \lambda EV$, well defined when t is small enough. Moreover, it is easy to check that $\psi_t'(\theta) \le \psi_N'(\theta)t$ for any θ, $t > 0$ (either by the formula of ψ_t' and ψ_N' or by definition in terms of the Gartner-Ellis limit). This implies that $\psi_t'^{-1}(y) \ge (\psi_N'^{-1}(y/t))$ for any y in the domain. Putting $y = 1 - \lambda EV$ gives $\bar{\theta}_t \ge \psi_N'^{-1}((1 - \lambda EV)/t)$. By steepness of ψ_N we have $\psi_N'^{-1}((1 - \lambda EV)/t) \nearrow \infty$ as $t \searrow 0$. So $\theta_t \nearrow \infty$ as $t \searrow 0$.

(iii) Consider $\psi_t'(\theta_t) = a_t$, or $\theta_t = \psi_t'^{-1}(a_t)$. Now from (18) we have

$$\psi_\infty'(\theta) = \int_0^\infty \psi_N'(\log(e^\theta \bar{F}(u) + F(u))) \frac{e^\theta \bar{F}(u)}{e^\theta \bar{F}(u) + F(u)}\, du$$

and that $\psi_\infty'(\theta)$ is increasing in θ, by the same argument as in the proof of (1). Moreover, by monotone convergence we have $\psi_t' \nearrow \psi_\infty'$ as $t \nearrow \infty$.

By Billingsley [6], p. 287, or Resnick [23], p. 5, Proposition 0.1, we have $\psi_t'^{-1} \to \psi_\infty'^{-1}$ as $t \nearrow \infty$. Moreover, since $\psi_t'^{-1}$ is increasing over the compact interval $[\lambda EV, 1]$, the convergence is uniform. By Resnick [23], p. 2, this implies continuous convergence, and hence $\psi_t'^{-1}(a_t) \to \psi_\infty'^{-1}(1)$, or $\theta_t \to \theta_\infty$.

A.3. Proof of Lemma 3.3. (i) As in the proof of Lemma 3.2 Part i, denote $\theta(t) = \theta_t$. Consider

$$\begin{aligned}\frac{d}{dt} I_t &= \theta(t)\lambda\bar{F}(t) + \theta'(t)a_t - \psi_t'(\theta(t))\theta'(t) - \psi_N(\log(e^{\theta(t)}\bar{F}(t) + F(t)))\\ &= \theta(t)\lambda\bar{F}(t) - \psi_N(\log(e^{\theta(t)}\bar{F}(t) + F(t)))\end{aligned}$$

since $\psi_t'(\theta(t)) = a_t$. Note that $h_t(\theta) := \psi_N(\log(e^\theta \bar{F}(t) + F(t)))$ is convex in θ for any $t \geq 0$ and so

$$h_t(\theta(t)) \geq h_t(0) + h_t'(0)\theta(t),$$

which gives

$$\psi_N(\log(e^{\theta(t)}\bar{F}(t) + F(t))) \geq \lambda\bar{F}(t)\theta(t).$$

Hence $(d/dt)I_t \leq 0$ and so I_t is nonincreasing.

(ii) Write $I_t = a_t\theta_t - \psi_t(\theta_t)$. By Lemma 3.2 Part iii, $\theta_t \searrow \theta_\infty$ on $[\theta_\infty, \theta_T]$ for $t \geq T$ for some $T > 0$. Since $\psi_t(\theta)$ is increasing in θ, by continuous convergence (see Resnick [23], p. 2), we have $\psi_t(\theta_t) \to \psi_\infty(\theta_\infty)$. Hence $I_t \to I^*$ defined in (19).

(iii) Note that in case V is supported on $[0, M]$, it is easy to check that $I_t = I_M$ is the same for any $t \geq M$. Hence the conclusion.

A.4. Proof of Lemma 3.4. (i) Following the spirit of the proof of Lemma 3.3 Part i, denote $\tilde{\theta}(t) = \tilde{\theta}_t$ for convenience and consider

$$\frac{d}{dt}\tilde{I}_t = \tilde{\theta}'(t)(1 - \lambda EV) - \psi_N'(\tilde{\theta}(t))t\tilde{\theta}'(t) - \psi_N(\tilde{\theta}(t)) = -\psi_N(\tilde{\theta}(t)) \leq 0$$

for small t, using $\psi_N'(\tilde{\theta}_t)t = 1 - \lambda EV$. Hence the conclusion.

(ii) Consider $\tilde{\theta}_t = \psi_N'^{-1}((1 - \lambda EV)/t)$, well-defined by the strict monotonicity of ψ_N'. By steepness of ψ_N we have $(\psi_N'^{-1}((1 - \lambda EV)/t) \nearrow \infty$ as $t \searrow 0$. So $\tilde{\theta}_t \nearrow \infty$ as $t \searrow 0$.

Now write

$$\tilde{I}_t = \tilde{\theta}_t(1 - \lambda EV) - \psi_N(\tilde{\theta}_t)t = (1 - \lambda EV)\left(\tilde{\theta}_t - \frac{\psi_N(\tilde{\theta}_t)}{\psi_N'(\tilde{\theta}_t)}\right) \to \infty$$

where the convergence follows from (2) and Part i.

A.5. Proof of Lemma 4.1. To prove Lemma 4.1, we first need the following analytical lemma:

LEMMA A.1. *Let h_m: $\mathcal{D} \subset \mathbb{R}^n \to \mathbb{R}$ be a sequence of monotone functions, in the sense that $h_m(x_1, x_2, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n)$ is either nondecreasing or nonincreasing in y_i fixing $x_1, \ldots, x_{i-1}, x_i, \ldots, x_n$, for any $i = 1, \ldots, n$. Moreover, suppose $\mathcal{D}$ is compact. If $h_m \to h$ pointwise, where h is continuous, then the convergence is uniform over $\mathcal{D}$.*

PROOF. Since $\mathcal{D}$ is compact, continuity of h implies uniform continuity. Therefore, given $\epsilon > 0$, there exists $\delta > 0$ such that $\|\mathbf{x}_1 - \mathbf{x}_2\| < \delta$ implies $|h(\mathbf{x}_1) - h(\mathbf{x}_2)| < \epsilon$. Compactness of $\mathcal{D}$ implies that there is a finite collection of these δ-balls to cover $\mathcal{D}$. Let $\{N_\delta(\mathbf{x})\}_{x \in \mathscr{C}}$ be such collection. Note that $h_m \to h$ uniformly over $\mathscr{C}$.

For any $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{D}$, consider

$$|h_m(\mathbf{x} - h(\mathbf{x})| \leq |h_m(\mathbf{x}) - h_m(\tilde{\mathbf{x}})| + |h_m(\tilde{\mathbf{x}}) + h(\tilde{\mathbf{x}})| + |h(\tilde{\mathbf{x}}) - h(\mathbf{x})|$$

where $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ is chosen to be the closet point to $\mathbf{x}$ in $\mathscr{C}$ that satisfies the following: For $i = 1, \ldots, n$, $\tilde{x}_i \geq x_i$ if h is nondecreasing in the i-th component, and $\tilde{x}_i \leq x_i$ if h is nonincreasing in the i-th component.

By construction we have $|h(\tilde{\mathbf{x}}) - h(\mathbf{x})| < 2\epsilon$ and $|h_m(\tilde{\mathbf{x}}) - h(\tilde{\mathbf{x}})| < \epsilon$ when m is large enough.

Now

$$\begin{aligned}
&|h_m(\mathbf{x}) - h_m(\tilde{\mathbf{x}})| \\
&\quad = h_m(\tilde{\mathbf{x}}) - h_m(\mathbf{x}) \quad \text{by our choice of } \tilde{\mathbf{x}} \text{ and monotonicity property of } h_m \\
&\quad \le h_m(\tilde{\mathbf{x}}) - h_m(\tilde{\tilde{\mathbf{x}}}) \quad \text{where } \tilde{\tilde{\mathbf{x}}} \text{ is chosen to be the closet point to } \mathbf{x} \text{ in } \mathscr{E} \text{ that satisfies the following:} \\
&\qquad\qquad \text{For } i = 1, \ldots, n,\ \tilde{\tilde{x}}_i \le x_i \text{ if } h \text{ is nondecreasing in the } i\text{-th component, and} \\
&\qquad\qquad \tilde{\tilde{x}}_i \ge x_i \text{ if } h \text{ is nonincreasing in the } i\text{-th component.} \\
&\quad \le |h_m(\tilde{\mathbf{x}}) - h(\tilde{\mathbf{x}})| + |h(\tilde{\mathbf{x}}) - h(\tilde{\tilde{\mathbf{x}}})| + |h_m(\tilde{\tilde{\mathbf{x}}}) - h(\tilde{\tilde{\mathbf{x}}})| \\
&\quad \le \epsilon + 2\epsilon + \epsilon
\end{aligned}$$

when m is large enough.

Combining the above, we have $|h_m(\mathbf{x}) - h(\mathbf{x})| \le 7\epsilon$ for all $x \in \mathscr{D}$. Hence the conclusion. □

Proof of Lemma 4.1. For convenience write $\psi_s(\theta; w, z, t) := \log E e^{\theta \bar{Q}^{\infty}_{w,z}[t,\infty]}$ and

$$\psi(\theta; w, z, t) := \int_w^z \psi_N(\log(e^{\theta}\bar{F}(t-u) + F(t-u)))\, du$$

defined for $\theta \in [\theta_\infty, \theta_T]$, $t \ge T$ and $0 \le w \le z \le t + \eta$ for some $\eta > 0$. We can extend the domain by putting $\psi_s(\theta; w, z, t) := \psi_s(\theta; w, t+\eta, t)$ and $\psi(\theta; w, z, t) := \psi(\theta; w, t+\eta, t)$ for $z > t + \eta$, and $\psi_s(\theta; w, z, t) = \psi(\theta; w, z, t) := 0$ for $w > z$.

Note that $\psi_s(\theta; w, z, t)$ defined as such is nondecreasing in θ, nonincreasing in w, nondecreasing in z, and nonincreasing in t. Also, $\psi_s(\theta; w, z, t) \to \psi(\theta; w, z, t)$ pointwise with $\psi(\theta; w, z, t)$ continuous. Hence the convergence is uniform over the compact set $\theta \in [\theta_\infty, \theta_T]$ and $(w, z, t) \in [0, K+\eta] \times [0, K+\eta] \times [0, K]$ by Lemma A.1, for any $K > 0$. By our construction we can extend the set of uniform convergence to $(w, z, t) \in [0, \infty)^2 \times [0, K]$.

We now choose K as follows. Given $\epsilon > 0$, there exists $K > 0$ such that for all $t > K$, $z \le t - K$, we have

$$\begin{aligned}
\psi(\theta; w, z, t) &= \int_w^z \psi_N(\log(e^{\theta}\bar{F}(t-u) + F(t-u)))\, du \\
&= \int_{t-z}^{t-w} \psi_N(\log(e^{\theta}\bar{F}(u) + F(u)))\, du \\
&\le \int_K^{\infty} \psi_N(\log(e^{\theta}\bar{F}(u) + F(u)))\, du \\
&\le C_1 \lambda \int_K^{\infty} \log(1 + (e^{\theta} - 1)\bar{F}(u))\, du \\
&\le C_2 \lambda \int_K^{\infty} \bar{F}(u)\, du \\
&< \epsilon
\end{aligned}$$

for some $C_1, C_2 > 0$, uniformly over $\theta \in [\theta_\infty, \theta_T]$. Hence for $z \le t - K$, $\psi_s(\theta; w, z, t) \le \psi_s(\theta; 0, t-K, t) \to \psi(\theta; 0, t-K, t) < \epsilon$ uniformly over $\theta \in [\theta_\infty, \theta_T]$ and so $|\psi_s(\theta; w, z, t) - \psi(\theta; w, z, t)| < 3\epsilon$ for large enough s.

For $z > t - K$, we write

$$\psi_s(\theta; w, z, t) = \frac{1}{s} \log E e^{\theta \bar{Q}^{\infty}_{w,t-K}[t,\infty] I(w < t-K) + \theta \bar{Q}^{\infty}_{(t-K)\vee w, z}[t,\infty]},$$

which is bounded from above by

$$\begin{aligned}
&\frac{1}{s} \log(E e^{\theta \bar{Q}^{\infty}_{w,t-K}[t,\infty] I(w<t-K)} E_0 e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]}) \\
&\qquad = \psi_s(\theta; w, t-K, t) I(w < t-K) + \frac{1}{s} \log E_0 e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]}
\end{aligned}$$

and bounded from below by

$$\begin{aligned}
&\frac{1}{s} \log(E e^{\theta \bar{Q}^{\infty}_{0,t-K}[t,\infty] I(w<t-K)} E_{00} e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]}) \\
&\qquad = \psi_s(\theta; w, t-K, t) I(w < t-K) + \frac{1}{s} \log E_{00} e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]} \qquad (67)
\end{aligned}$$

where $E_0[\,\cdot\,]$ denotes the expectation conditioned that a customer arrives at time 0 and is counted in $\bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[t,\infty]$, whereas $E_{00}[\,\cdot\,]$ denotes the expectation conditioned on delayed arrival with tail distribution (in the base system) given by $\sup_b P(U^0 - b > x \mid U^0 - b)$. Note that $\sup_b P(U^0 - b > x \mid U^0 > b)$ is a valid tail distribution because of the light-tail assumption on U^0. Indeed, it is obvious that $\sup_b P(U^0 - b > 0 \mid U^0 > b) = 1$, and by the same argument following that of (57), we have $\sup_b P(U^0 - b > x \mid U^0 > b) \le e^{-cx} \to 0$ for some $c > 0$. Moreover, it is obvious that $\sup_b P(U^0 - b > x \mid U^0 > b)$ is nonincreasing. Now by construction this tail distribution is stochastically at most as large as $P(U^0 - b > x \mid U^0 > b)$ for any $b \ge 0$, and hence (67). Note that $(1/s) \log E_0 e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]}$ and $(1/s) \log E_{00} e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]}$ both converge to

$\psi(\theta; 0, (z-t+K)\wedge(z-w), K)$ uniformly by the argument earlier (as a special case when $t \le K$). Also we have shown that $\psi_s(\theta; w, t-K, t)$ converges to $\psi_s(\theta; w, t-K, t)$ uniformly for $t > K$ (as a special case when $z \le t-K$ and $t > K$). The sandwich argument concludes the lemma. □

A.6. Proof of Lemma 4.2. Consider

$$\begin{aligned}
&\frac{1}{s}\log E\exp\left\{\sum_{k=1}^{n}\left(\sum_{j=k}^{n}\theta_{kj}Q^{\infty}_{(k-1)\Delta,k\Delta}[(j-1)\Delta, j\Delta] + \theta_{k\cdot}Q^{\infty}_{(k-1)\Delta,k\Delta}[n\Delta,\infty]\right)\right\}\\
&\quad=\frac{1}{s}\log E\exp\left\{\sum_{k=1}^{n}\left(\sum_{j=k}^{n}\theta_{kj}\sum_{i=N_s((k-1)\Delta)+1}^{N_s(k\Delta)} I((j-1)\Delta < V_i + A_i \le j\Delta) + \theta_{k\cdot}\sum_{i=N_s((k-1)\Delta)+1}^{N_s(k\Delta)} I(V_i + A_i > n\Delta)\right)\right\}\\
&\quad=\frac{1}{s}\log E\prod_{k=1}^{n}\prod_{i=N_s((k-1)\Delta)+1}^{N_s(k\Delta)}\left(\sum_{j=k}^{n}e^{\theta_{kj}}P((j-1)\Delta < V_i + A_i \le j\Delta) + e^{\theta_{k\cdot}}\bar{F}(n\Delta - A_i)\right)\\
&\quad=\frac{1}{s}\log E\exp\left\{\sum_{k=1}^{n}\int_{(k-1)\Delta}^{k\Delta} h_k(u)\,dN_s(u)\right\}
\end{aligned}$$

where

$$h_k(u) := \log\left(\sum_{j=k}^{n}e^{\theta_{kj}}P((j-1)\Delta < V_i + u \le j\Delta) + e^{\theta_{k\cdot}}\bar{F}(n\Delta - u)\right).$$

Now

$$\begin{aligned}
&\frac{1}{s}\log E\exp\left\{\sum_{k=1}^{n}\sum_{w=1}^{m} h_k(\underline{\zeta}_{kw})\left[N_s\left((k-1)\Delta + \frac{w\Delta}{m}\right) - N_s\left((k-1)\Delta + \frac{(w-1)\Delta}{m}\right)\right]\right\}\\
&\quad\le\frac{1}{s}\log E\exp\left\{\sum_{k=1}^{n}\int_{(k-1)\Delta}^{k\Delta} h_k(u)\,dN_s(u)\right\}\\
&\quad\le\frac{1}{s}\log E\exp\left\{\sum_{k=1}^{n}\sum_{w=1}^{m} h_k(\bar{\zeta}_{kw})\left[N_s\left((k-1)\Delta + \frac{w\Delta}{m}\right) - N_s\left((k-1)\Delta + \frac{(w-1)\Delta}{m}\right)\right]\right\}
\end{aligned}$$

where $\underline{\zeta}_{kw} := \arg\min\{h_k(u)\colon (k-1)\Delta + (w-1)\Delta/m \le u \le (k-1)\Delta + w\Delta/m\}$ and $\bar{\zeta}_{kw} := \arg\max\{h_k(u)\colon (k-1)\Delta + (w-1)\Delta/m \le u \le (k-1)\Delta + w\Delta/m\}$. The existence of $\underline{\zeta}_{kw}$ and $\bar{\zeta}_{kw}$ is guaranteed by the continuity of $h_k(\cdot)$, which is implied by our assumption that V_i has density.

Letting $s \to \infty$ and by (3) we have

$$\begin{aligned}
\sum_{k=1}^{n}\sum_{w=1}^{m}\psi_N(h_k(\underline{\zeta}_{kw}))\frac{\Delta}{m} &\le \liminf_{s\to\infty}\frac{1}{s}\log E\exp\left\{\sum_{k=1}^{n}\int_{(k-1)\Delta}^{k\Delta} h_k(u)\,dN_s(u)\right\}\\
&\le \limsup_{s\to\infty}\frac{1}{s}\log E\exp\left\{\sum_{k=1}^{n}\int_{(k-1)\Delta}^{k\Delta} h_k(u)\,dN_s(u)\right\}\\
&\le \sum_{k=1}^{n}\sum_{w=1}^{m}\psi_N(h_k(\bar{\zeta}_{kw}))\frac{\Delta}{m}.
\end{aligned}$$

By continuity of $h_k(\cdot)$ and $\psi_N(\cdot)$, $\psi_N(h_k(\cdot))$ is Riemann integrable. Letting $m \to \infty$ yields the conclusion.

A.7. Proof of Lemma 5.2 and 5.3. Our goal here is to prove Lemma 5.2, via Lemma 5.3. For convenience let $G(y) := \int_y^\infty \bar{F}(u)\,du$, so $\nu(y) = G(y)^{1/(2+\eta)}$ where η is defined in (11). Note that by L'Hospital's rule and Assumption (4), we have

$$\lim_{y\to\infty}\frac{y\bar{F}(y)}{G(y)} = \lim_{y\to\infty}\frac{\bar{F}(y) - yf(y)}{-\bar{F}(y)} = \lim_{y\to\infty}(yh(y) - 1) = \infty. \tag{68}$$

As discussed before, the key step to show Lemma 5.2 is an estimate of the limiting Gaussian process given by Lemma 5.3. The proof of this inequality takes three steps. We first consider the case when $i = 1$. The first step is to define a pseudo metric (i.e., a distance function that satisfies all the axioms of a metric except that two distinguishable points can have zero distance)

$$d_1((t,y),(t',y')) := E(\tilde{R}_1(t,y) - \tilde{R}_1(t',y'))^2, \tag{69}$$

where $\tilde{R}_1(t,y) := R_1(t,y)/\nu(y)$, and show that the domain is compact under this (pseudo) metric. For convenience we call (69) the d_1-metric. Then we can prove that the Gaussian process $\tilde{R}_1(t,y)$ is a.s. bounded by an entropy argument. The third step invokes Borell's inequality.

For convenience let $S := [0, t_0] \times [0, \infty)$.

Before these steps, we need an estimate of the d_1-metric:

Lemma A.2. *Let (t, y) and (t', y') be two points on $S = [0, t_0] \times [0, \infty)$. Also, let $t_1 = t \vee t'$ and y_1 be the corresponding y or y', and similarly $t_2 = t \wedge t'$ and y_2 be the corresponding y or y'. Then*

$$\begin{aligned} d_1((t,y),(t',y')) = {} & \frac{\lambda}{\nu(y)^2}\int_0^{t_2} \bar{F}(t+y-u)F(t+y-u)\,du + \frac{\lambda}{\nu(y')^2}\int_0^{t_2} \bar{F}(t'+y'-u)F(t+y-u)\,du \\ & - \frac{2\lambda}{\nu(y)\nu(y')}\int_0^{t_2} (\bar{F}(t+y-u)\vee \bar{F}(t'+y'-u) - \bar{F}(t+y-u)\bar{F}(t'+y'-u)\,du \\ & + \frac{1}{\nu(y_1)^2}\int_{t_2}^{t_1} \bar{F}(t_1+y_1-u)F(t_1+y_1-u)\,du. \end{aligned} \tag{70}$$

The proof of this lemma follows the approach in Lemma 5.1 of Krichagina and Puhalskii [20]. Hence we only sketch the proof here:

Proof (Sketch). Recall that

$$\tilde{R}_1(t,y) := \frac{\int_0^t \int_0^\infty I(u+x>t+y)\,dK(u,x)}{\nu(y)}.$$

For a partition $\{u_0 = 0, u_1, u_2, \ldots, u_k\}$ of $[0, t_0]$, define

$$I^k_{t+y}(u,x) := \sum_{i=1}^k I(u \in (u_{i-1}, u_i])I(x > t+y-u_i).$$

Let

$$\tilde{R}^k_1(t,y) := \frac{\int_0^t \int_0^\infty I^k_{t+y}(u,x)\,dK(u,x)}{\nu(y)}$$

be a discretized version of $\tilde{R}_1(t,y)$. One can check that $\tilde{R}^k_1(t,y)$ converges to $\tilde{R}_1(t,y)$ in mean square as the mesh of the partition goes to 0.

Now take (t,y) and (t',y') in S. Define $t_1 := t \vee t'$ and y_1 be the corresponding y or y', and define $t_2 := t \wedge t'$ and y_2 be the corresponding y or y'. Also define k_2 such that $u_{k_2} \le t_1 < u_{k_2+1}$. Using (5.4) and (5.5) in Krichagina and Puhalskii [20], we have

$$\begin{aligned} & E(\tilde{R}^k_1(t,y) - \tilde{R}^k_1(t',y'))^2 \\ & = E\left[\left(\frac{\int_0^{t_2}\int_0^\infty I^k_{t+y}(u,x)\,dK(u,x)}{\nu(y)} - \frac{\int_0^{t_2}\int_0^\infty I^k_{t'+y'}(u,x)\,dK(u,x)}{\nu(y')}\right)^2\right] + E\left[\left(\frac{\int_{t_2}^{t_1}\int_0^\infty I^k_{t_1+y_1}(u,x)\,dK(u,x)}{\nu(y_1)}\right)^2\right] \\ & = \frac{1}{\nu(y)^2}E\left[\left(\int_0^{t_1}\int_0^\infty I^k_{t+y}(u,x)\,dK(u,x)\right)^2\right] + \frac{1}{\nu(y')^2}E\left[\left(\int_0^{t_1}\int_0^\infty I^k_{t'+y'}(u,x)\,dK(u,x)\right)^2\right] \\ & \quad - \frac{2}{\nu(y)\nu(y')}E\left[\int_0^{t_2}\int_0^\infty I^k_{t+y}(u,x)\,dK(u,x)\int_0^{t_2}\int_0^\infty I^k_{t'+y'}(u,x)\,dK(u,x)\right] \\ & \quad + \frac{1}{\nu(y_1)^2}E\left[\left(\int_{t_2}^{t_1}\int_0^\infty I^k_{t_1+y_1}(u,x)\,dK(u,x)\right)^2\right] \\ & = \frac{1}{\nu(y)^2}\sum_{i=1}^{k_2}\lambda(u_i-u_{i-1})\bar{F}(t+y-u_i)F(t+y-u_i) + \frac{1}{\nu(y')^2}\sum_{i=1}^{k_2}\lambda(u_i-u_{i-1})\bar{F}(t'+y'-u_i)F(t'+y'-u_i) \\ & \quad - \frac{2}{\nu(y)\nu(y')}\sum_{i=1}^{k_2}\lambda(u_i-u_{i-1})[\bar{F}(t+y-u_i)\vee\bar{F}(t'+y'-u_i) - \bar{F}(t+y-u_i)\bar{F}(t'+y'-u_i)] \\ & \quad + \frac{1}{\nu(y_1)^2}\sum_{i=k_2+1}^{k_1}\lambda(u_i-u_{i-1})\bar{F}(t_1+y_1-u_i)F(t_1+y_1-u_i) + o(1) \end{aligned}$$

which converges to (70) as the mesh goes to 0. □

Lemma A.3. *We can compactify the space $[0, t_0] \times [0, \infty]$ with the d_1-metric defined in (69).*

Proof. Consider the mapping $(i, \tan) : [0, t_0] \times [0, \pi/2] \to [0, t_0] \times [0, \infty]$, where i is the identity map. Here the domain is equipped with the Euclidean metric, and the image is equipped with the d_1-metric. We will show that the mapping $(i, \tan)$ is continuous and well-defined over its domain, including the points (t, x) where $x = \pi/2$, and hence its image is compact.

Suppose first that $(t, x) \to (t^*, x^*)$ where $x^* \neq \pi/2$. Since $\tan(\cdot)$ is continuous, and $\int_y^{t+y} \bar{F}(u)\,du$ and $\nu(y)$ are continuous in t and y (under Euclidean metric), it is easy to see that $d_1((t, \tan x), (t^*, \tan x^*)) \to 0$ by using (70).

We now show that $d_1(\cdot,\cdot)$ is still a (pseudo) metric when including the points (t,y) with $y=\infty$. Define, for $y'=\infty$, that

$$d_1((t,y),(t',y')) := \frac{\lambda\int_0^{t_2}\bar F(t+y-u)(1+F(t+y-u))\,du}{\nu(y)^2} + \begin{cases} \dfrac{\lambda\int_{t'}^{t}\bar F(t+y-u)F(t+y-u)\,du}{\nu(y)^2} & \text{if } t>t' \\ 0 & \text{if } t\le t' \end{cases}$$

and $d_1((t,y),(t',y')):=0$ if $y=y'=\infty$. It is straightforward to check that $d_1(\cdot,\cdot)$ is continuous at $y'=\infty$ by using (70) (note that the second term of (70) goes to 0 since for y' large enough it is less than or equal to $\lambda\int_{y'+(t'-t_2)}^{y'+t'}\bar F(du)\,du/\nu(y')^2 \le \lambda G(y)^{1-2/(2+\eta)}\to 0$). Hence both the commutativity and triangle inequality hold also at $y'=\infty$, which implies that $d_1(\cdot,\cdot)$ is a pseudo metric on $[0,t_0]\times[0,\infty]$. Now consider $x^*=\pi/2$. It is now easy to see that $d_1((t,\tan x),(t^*,\infty))\to 0$ as $(t,x)\to(t^*,\pi/2)$. □

Lemma A.4. *$E\sup_S \tilde R_1(t,y)<\infty$. In particular, $\tilde R_1(t,y)$ is a.s. bounded over S.*

Proof. We use C here to denote constants, not necessarily the same every time it appears. We carry out an entropy argument (see, for example, Adler [1])

$$E\sup_S \tilde R_1(t,y) \le K\int_0^\infty H^{1/2}(\epsilon)\,d\epsilon = K\int_0^{\operatorname{diam}(S)/2} H^{1/2}(\epsilon)\,d\epsilon$$

where $K>0$ is a universal constant; $H(\epsilon):=\log N(\epsilon)$, with $N(\epsilon)$ the ϵ-th order entropy of S, i.e., the minimum number of ϵ-balls (under d_1-metric) to cover S; and $\operatorname{diam}(S)$ is the diameter of S given by $\sup_{(t,y),(t',y')\in S} d_1((t,y),(t',y'))$.

Let (t,y) and (t',y') be two points on $[0,t_0]\times[0,\infty]$, and let t_1, t_2, y_1, y_2 be as defined in Lemma A.2. Note that from (70) we have

$$\begin{aligned} d_1((t,y),(t',y')) &\le \frac{\lambda}{\nu(y)^2}\int_0^{t_2}\bar F(t+y-u)\,du + \frac{\lambda}{\nu(y')^2}\int_0^{t_2}\bar F(t'+y'-u)\,du + \frac{\lambda}{\nu(y_1)^2}\int_{t_2}^{t_1}\bar F(t_1+y_1-u)\,du \\ &\le \frac{\lambda\int_y^{t+y}\bar F(u)\,du}{\nu(y)^2} + \frac{\lambda}{\nu(y')^2}\int_y^{t'+y}\bar F(u)\,du + \frac{\lambda}{\nu(y_1)^2}\int_{y_1}^{|t-t'|+y_1}\bar F(u)\,du \\ &\le \lambda G(y)^{\eta/(2+\eta)} + \lambda G(y')^{\eta/(2+\eta)} + \lambda G(y_1)^{\eta/(2+\eta)} \\ &\le 3\lambda G(y\wedge y')^{\eta/(2+\eta)} \end{aligned} \tag{71}$$

which implies that $\operatorname{diam}(S)$ is bounded.

Now pick any $\epsilon>0$. Since $G(\cdot)$ is continuous we can define $G^{-1}(\cdot)$ to be the inverse of $G(\cdot)$. From (71) we have $d_1((t,y),(t',y'))<\epsilon$ for $y,y'>G^{-1}((\epsilon/(3\lambda))^{(2+\eta)/\eta})$.

Now also note that we can write

$$\begin{aligned} d_1((t,y),(t',y')) &= \lambda\left(\frac{1}{\nu(y)^2}-\frac{1}{\nu(y)\nu(y')}\right)\int_0^{t_2}\bar F(t+y-u)F(t+y-u)\,du \\ &\quad+\lambda\left(\frac{1}{\nu(y')^2}-\frac{1}{\nu(y)\nu(y')}\right)\int_0^{t_2}\bar F(t'+y'-u)F(t'+y'-u)\,du \\ &\quad-\frac{\lambda}{\nu(y)\nu(y')}\int_0^{t_2}(2\bar F(t+y-u)\vee\bar F(t'+y'-u)-2\bar F(t+y-u)\bar F(t'+y'-u) \\ &\quad-\bar F(t+y-u)F(t+y-u)-\bar F(t'+y'-u)F(t'+y'-u))\,du \\ &\quad+\frac{\lambda}{\nu(y_1)}\int_{t_2}^{t_1}\bar F(t_1+y_1-u)F(t_1+y_1-u)\,du, \end{aligned}$$

where the integral in the third term can be written as

$$\begin{aligned} &\int_0^{t_2}(2\bar F(t+y-u)\vee\bar F(t'+y'-u)-2\bar F(t+y-u)\bar F(t'+y'-u) \\ &\quad-\bar F(t+y-u)F(t+y-u)-\bar F(t'+y'-u)F(t'+y'-u))\,du \\ &=\int_0^{t_2}(2\bar F(t+y-u)\vee\bar F(t'+y'-u)-2\bar F(t+y-u)\bar F(t'+y'-u) \\ &\quad-\bar F(t+y-u)+\bar F(t+y-u)^2-\bar F(t'+y'-u)+\bar F(t'+y'-u)^2)\,du \\ &=\int_0^{t_2}(2\bar F(t+y-u)\vee\bar F(t'+y'-u)-\bar F(t+y-u)-\bar F(t'+y'-u)+(\bar F(t+y-u)-\bar F(t'+y'-u))^2)\,du \\ &=\int_0^{t_2}(|\bar F(t+y-u)-\bar F(t'+y'-u)|+(\bar F(t+y-u)-\bar F(t'+y'-u))^2)\,du. \end{aligned}$$

Hence

$$\begin{aligned} d_1((t,y),(t',y')) &\le \frac{2\lambda}{\nu(y\wedge y')}\left|\frac{1}{\nu(y)}-\frac{1}{\nu(y')}\right|+\frac{\lambda}{\nu(y_1)^2}|t-t'| \\ &\le \frac{2\lambda}{\nu(y\wedge y')}\frac{\bar F(\bar y)}{G(\bar y)^{1+1/(2+\eta)}}|y-y'|+\frac{\lambda}{\nu(y_1)^2}|t-t'| \\ &\qquad \text{where } \bar y \text{ is between } y \text{ and } y', \text{ by mean value theorem on } 1/\nu(\cdot) \\ &\le \frac{2\lambda}{G(y\wedge y')^{(4+\eta)/(2+\eta)}}|y-y'|+\frac{\lambda}{G(y\wedge y')^{2/(2+\eta)}}|t-t'| \\ &\le \frac{\lambda}{G(y\wedge y')^{(4+\eta)/(2+\eta)}}(2|y-y'|+|t-t'|). \end{aligned}$$

When at least one of y and y' is less than or equal to $G^{-1}((\epsilon/(3\lambda))^{(2+\eta)/\eta})$, we then get

$$d_1((t,y),(t',y')) \le \frac{3^{(4+\eta)/\eta}\lambda^{(4+2\eta)/\eta}}{\epsilon^{(4+\eta)/\eta}}(2|t-t'|+|y-y'|).$$

Hence we can fill up the space S by

$$N(\epsilon)=O\left(\frac{1}{(\epsilon\cdot\epsilon^{(4+\eta)/\eta})^2}\cdot G^{-1}\left(\left(\frac{\epsilon}{3\lambda}\right)^{(2+\eta)/\eta}\right)\right)=O\left(\frac{1}{\epsilon^{4(2+\eta)/\eta}}\cdot G^{-1}\left(\left(\frac{\epsilon}{3\lambda}\right)^{(2+\eta)/\eta}\right)\right)$$

number of ϵ-balls. By (5) we get that $G(y)\le C/y^{1/p}$ for any $p>0$, and so $G^{-1}(\epsilon)\le C/\epsilon^{1/p}$. This gives

$$N(\epsilon)=O\left(\frac{1}{\epsilon^{4(2+\eta)/\eta}}\cdot\frac{1}{\epsilon^p}\right)=O\left(\frac{1}{\epsilon^{4(2+\eta)/\eta+p}}\right)$$

and hence

$$\int_0^{\operatorname{diam}(S)} H^{1/2}(\epsilon)\,d\epsilon = O\left(\int_0^C\sqrt{\log\left(\frac{1}{\epsilon}\right)}\,d\epsilon+C\right)<\infty. \quad \square$$

Lemma A.5. *The Borell-TIS inequality holds; i.e., for* $x\ge E\sup_S\tilde R_1(t,y)$,

$$P\left(\sup_S\tilde R_1(t,y)\ge x\right)\le\exp\left\{-\frac{1}{2\sigma_1^2}\left(x-E\sup_S\tilde R_1(t,y)\right)^2\right\}$$

where

$$\sigma_1^2:=\sup_S E\tilde R_1(t,y)^2.$$

Proof. Note that

$$E\tilde R_1(t,y)^2=\frac{\lambda\int_0^t\bar F(t+y-u)F(t+y-u)\,du}{\nu(y)^2}\le\frac{\lambda\int_y^{t+y}\bar F(u)\,du}{G(y)^{2/(2+\eta)}}\le\lambda G(y)^{\eta/(2+\eta)}$$

and so

$$\sigma_1^2=\sup_S E\tilde R_1(t,y)^2\le C$$

for some constant C. By Lemma A.4 $\tilde R_1(t,y)$ is a.s. bounded and Borell-TIS inequality holds. $\square$

We now carry out the same scheme for $R_2(t,y)$. Let $\tilde R_2(t,y):=R_2(t,y)/\nu(y)$. Indeed it is straightforward to show that the d_2-metric of $\tilde R_2(t,y)$ is given by

$$\begin{aligned} d_2((t,y),(t',y')) &:= E(\tilde R_2(t,y)-\tilde R_2(t',y'))^2 \\ &= \lambda c_a^2\int_0^{t_2}\left(\frac{\bar F(t+y-u)}{\nu(y)}-\frac{\bar F(t'+y'-u)}{\nu(y')}\right)^2du+\lambda c_a^2\int_{t_2}^{t_1}\left(\frac{\bar F(t_1+y_1-u)}{\nu(y_1)}\right)^2du \end{aligned} \tag{72}$$

where again $t_1:=t\vee t'$, $t_2:=t\wedge t'$, and y_1, y_2 are the corresponding y or y'.

Lemma A.6. *We can compactify the space S with the d-metric defined in* (72).

Proof. For $(t,y),(t',y')$ such that $y,y'\ne\infty$, write

$$\begin{aligned} &d_2((t,y),(t',y')) \\ &=\lambda c_a^2\left(\frac{\int_0^{t_2}\bar F(t+y-u)^2\,du}{\nu(y)^2}+\frac{\int_0^{t_2}\bar F(t'+y'-u)^2\,du}{\nu(y')^2}-\frac{2\int_0^{t_2}\bar F(t+y-u)\bar F(t'+y'-u)\,du}{\nu(y)\nu(y')}\right. \\ &\qquad\left.+\frac{\int_{t_2}^{t_1}\bar F(t_1+y_1-u)^2\,du}{\nu(y_1)^2}\right) \end{aligned}$$

and define, for $y'=\infty$, that

$$d_2((t,y),(t',y')):=\lambda c_a^2\int_0^t\frac{\bar F(t+y-u)^2}{\nu(y)^2}\,du$$

and $d_2((t,y),(t',y')):=0$ if both $y,y'=\infty$.

Then $d_2((t,y),(t',y'))$ is continuous at $y'=\infty$ since

$$\frac{\int_0^{t_2}\bar F(t'+y'-u)\,du}{\nu(y')^2}\le\frac{\int_{y'}^{t_0+y'}\bar F(u)\,du}{\nu(y')^2}=G(y')^{\eta/(2+\eta)}\to 0$$

and

$$\begin{aligned}\frac{\int_0^{t_2}\bar F(t+y-u)\bar F(t'+y'-u)\,du}{\nu(y)\nu(y')}&\le\frac{\sqrt{\int_0^{t_2}\bar F(t+y-u)^2\,du\int_0^{t_2}\bar F(t'+y'-u)^2\,du}}{\nu(y)\nu(y')}\\&\le\sqrt{\frac{\int_y^{t_0+y}\bar F(u)\,du}{\nu(y)^2}}\cdot\sqrt{\frac{\int_{y'}^{t_0+y'}\bar F(u)\,du}{\nu(y')^2}}\\&\le G(y)^{\eta/(2(2+\eta))}G(y')^{\eta/(2(2+\eta))}\\&\to 0.\end{aligned}$$

If $t'>t$, then

$$\frac{\int_t^{t'}\bar F(t'+y'-u)^2\,du}{\nu(y')^2}\le\frac{\int_{y'}^{t_0+y'}\bar F(u)\,du}{\nu(y')^2}\le G(y')^{\eta/(2+\eta)}\to 0.$$

Hence $d_2(\cdot,\cdot)$ is continuous at $y'=\infty$. The rest follows as in the proof of Lemma A.3. □

Lemma A.7. *$E\sup_S\tilde R_2(t,y)<\infty$. In particular, $\tilde R_2(t,y)$ is a.s. bounded over S.*

Proof. From (72) we have the estimate

$$\begin{aligned}&d_2((t,y),(t',y'))\\&\quad\le 2\lambda c_a^2\left(\int_0^t\left(\frac{\bar F(t+y-u)}{\nu(y)}\right)^2du\vee\int_0^{t'}\left(\frac{\bar F(t'+y'-u)}{\nu(y')}\right)^2du\right)+\lambda c_a^2\int_{t_1}^{t_2}\left(\frac{\bar F(t_1+y_1-u)}{\nu(y_1)}\right)^2du\\&\quad\le 2\lambda c_a^2(G(y)^{\eta/(2+\eta)}\vee G(y')^{\eta/(2+\eta)})+\lambda c_a^2G(y_1)^{\eta/(2+\eta)}.\end{aligned}\tag{73}$$

On the other hand, using multivariate Taylor series expansion,

$$\begin{aligned}&\frac{\bar F(t+y-u)}{\nu(y)}-\frac{\bar F(t'+y'-u)}{\nu(y')}\\&\quad\le\sup_{t,y}\left|\frac{f(t+y-u)}{\nu(y)}\right||t-t'|+\sup_{t,y}\left|\frac{1}{2+\eta}\frac{\bar F(t+y-u)\bar F(y)}{G(y)^{1+1/(2+\eta)}}-\frac{f(y)}{G(y)^{1/(2+\eta)}}\right||y-y'|\\&\quad\le\frac{C}{G(y)^{(3+\eta)/(2+\eta)}}(|t-t'|+|y-y'|)\end{aligned}$$

and hence

$$d_2((t,y),(t',y'))\le\frac{C}{G(y)^{(3+\eta)/(2+\eta)}}(|t-t'|+|y-y'|)\tag{74}$$

where C are constants not necessarily the same every time they appear. With (73) and (74), the rest follows as in the proof of Lemma A.4. □

Lemma A.8. *The Borell-TIS inequality holds; i.e., for $x\ge E\sup_S\tilde R_2(t,y)$,*

$$P\left(\sup_S\tilde R_2(t,y)\ge x\right)\le\exp\left\{-\frac{1}{2\sigma_2^2}\left(x-E\sup_S\tilde R_2(t,y)\right)^2\right\}$$

where

$$\sigma_2^2:=\sup_S E\tilde R_2(t,y)^2.$$

Proof. Note that

$$E\tilde R_2(t,y)^2=\frac{\lambda c_a^2\int_0^t\bar F(t+y-u)^2du}{\nu(y)^2}\le\frac{\lambda c_a^2\int_y^{t+y}\bar F(u)\,du}{G(y)^{2/(2+\eta)}}\le\lambda c_a^2G(y)^{\eta/(2+\eta)}.$$

The rest follows as in the proof of Lemma A.5. □

Lemma 5.3 is now an immediate corollary of Lemma A.5 and A.8:

Proof of Lemma 5.3.

$$\begin{aligned}
&P(|R(t,y)| \le C_*\nu(y) \text{ for all } t\in[0,t_0],\ y\in[0,\infty))\\
&\quad\ge P\left(\sup_S|\tilde R_1(t,y)|+\sup_S|\tilde R_2(t,y)|\le C_*\right)\\
&\quad\ge P\left(\sup_S|\tilde R_1(t,y)|\le\frac{C_*}{2}\right)P\left(\sup_S|\tilde R_2(t,y)|\le\frac{C_*}{2}\right)\\
&\quad> 0,
\end{aligned}$$

when C_* is large enough, by the independence of $\tilde R_1(\cdot,\cdot)$ and $\tilde R_2(\cdot,\cdot)$ in the second inequality. □

With Lemma 5.3, we now prove Lemma 5.2.

Proof of Lemma 5.2. First consider (46). Take $C_1 = 3C_*$ where C_* is the constant in Lemma 5.3. We have

$$\begin{aligned}
&P\left(\bar Q^\infty(t,y)\in\left(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y)\right)\text{ for all } t\in[0,t_0],\ y\in[0,\infty)\ \middle|\ B(0)\right)\\
&\quad\ge P\left(U_0\le x,\ 0\in\left(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y)\right)\text{ for } t\in[0,U_0],\ y\in[0,\infty),\right.\\
&\qquad\left.\bar Q^\infty(t,y)\in\left(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y)\right)\text{ for all } t\in[U_0,t_0],\ y\in[0,\infty)\ \middle|\ B(0)\right).
\end{aligned}\tag{75}$$

Letting $x = 1/(\lambda s)$, we will show that $0\in(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y))$ for $t\in[0,U_0]$ and $y\in[0,\infty)$ in the expression is redundant. In fact, let $m(s)=\inf\{\sqrt{s}C_*\nu(y)<\frac12\}$. When $y=m(s)$, $\lambda s\int_y^{t+y}\bar F(u)\,du$ is less than 1 for large enough s, and when $y\ge m(s)$, it decays faster than $\sqrt{s}C_1\nu(y)<\frac12$ (see Remark 2.1 for a similar argument). Hence $(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y))$ contains 0 when $y\ge m(s)$. When $y<m(s)$, the choice of x gives

$$\lambda s\int_y^{t+y}\bar F(u)\,du\le\lambda s t\bar F(y)\le\lambda s x=1$$

for $t\in[0,U_0]$ and $U_0\le x$. Hence $(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y))$ also contains 0 when $y<m(s)$.

In fact with the same choice of x, by similar argument we have $(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_*\nu(y))$ contains only 0 for $t\in[0,U_0]$ and $y\ge m(s)$, and that $0\in(\lambda s\int_y^{t+U_0+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y))$ for $t\in[0,U_0]$ and $y\ge m(s)$. This will be useful later on in the proof.

The same choice of x, and because $\bar F(\cdot)$ is decreasing, also guarantees that

$$\lambda s\int_{t+y}^{t+U_0+y}\bar F(u)\,du\le 2C_*\sqrt{s}\nu(y)\tag{76}$$

when s is large enough. In fact, when $y=m(s)$, $\lambda s\int_{t+y}^{t+U_0+y}\bar F(u)\,du$ is less than 1 when s is large enough, and when $y\ge m(s)$ it decays faster than $2C_*\sqrt{s}\nu(y)$. Hence the inequality (76) when $y\ge m(s)$. When $y<m(s)$, $U_0\le x$ leads to $\lambda s\int_{t+y}^{t+U_0+y}\bar F(u)\,du\le 1$, and hence the conclusion. Again this will be useful later on.

Hence (75) is greater than or equal to

$$P(U_0\le x\mid B(0))P\left(\bar Q_0^\infty(t,y)\in\left(\lambda s\int_y^{t+U_0+y}\bar F(u)\,du\pm\sqrt{s}C\nu(y)\right)\text{ for all } t\in[0,t_0]\ \middle|\ U_0\le x, B(0)\right)$$

where $\bar Q_0^\infty(t,y)$ is independent of U_0 and has the same distribution as $\bar Q^\infty(t,y)$ with the first customer arriving at time 0.

For any $U_0\le x$, we have

$$\begin{aligned}
&P\left(\bar Q_0^\infty(t,y)\in\left(\lambda s\int_y^{t+U_0+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y)\right)\text{ for all } t\in[0,t_0],\ y\in[0,\infty)\right)\\
&\quad\ge P\left(\bar Q_0^\infty(t,y)\in\left(\lambda s\int_y^{t+U_0+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y)\right)\text{ for all } t\in[0,t_0],\ y\in[0,m(s)),\right.\\
&\qquad\left.\bar Q_0^\infty(t,y)\in\left(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_*\nu(y)\right)\text{ for all } t\in[0,t_0],\ y\in[m(s),\infty)\right)\\
&\qquad\text{(since the interval }\left(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_*\nu(y)\right)\text{ only contains 0 while}\\
&\qquad 0\in\left(\lambda s\int_y^{t+U_0+y}\bar F(u)\,du\pm\sqrt{s}C_1\nu(y)\right)\text{ when } y>m(s)\text{ as discussed above)}\\
&\quad\ge P\left(\sup_{y\in[0,m(s))}\left|\frac{\bar Q_0^\infty(t,y)-\lambda s\int_y^{t+y}\bar F(u)\,du}{\sqrt{s}}\right|+\sup_{y\in[0,m(s))}\lambda\sqrt{s}\int_{t+y}^{t+U_0+y}\bar F(u)\,du\le C_1\nu(y),\right.\\
&\qquad\left.\bar Q_0^\infty(t,y)\in\left(\lambda s\int_y^{t+y}\bar F(u)\,du\pm\sqrt{s}C_*\nu(y)\right)\text{ for all } t\in[0,t_0],\ y\in[m(s),\infty)\right)
\end{aligned}$$

$$\geq P\left(\left|\frac{\bar{Q}_0^\infty(t,y)-\lambda s\int_y^{t+y}\bar{F}(u)\,du}{\sqrt{s}}\right|\leq C_*\nu(y) \text{ for all } t\in[0,t_0],\ y\in[0,\infty)\right)$$

(by (76))

$$\to P(|R(t,y)|\leq C_*\nu(y) \text{ for all } t\in[0,t_0],\ y\in[0,\infty))>0$$

by Lemma 5.3. The convergence follows from the functional central limit theorem (see Pang and Whitt [21]) and that the set $\{f\colon |f(t,y)|\leq C_*\nu(y) \text{ for all } t\in[0,t_0],\ y\in[0,\infty)\}$ is a continuity set.

Lastly, since U^0 is light tailed, by the argument following (57) in the proof of Proposition 2.1, we have

$$\inf_{b\geq 0}P\left(U_0\leq\frac{1}{\lambda s}\,\middle|\,B(0)=b\right)=\inf_{b\geq 0}P\left(U^0-b\leq\frac{1}{\lambda}\,\middle|\,U^0>b\right)\geq 1-e^{-c/\lambda}>0$$

for some constant $c>0$. Hence (46) holds. Inequality (47) is obvious since one can isolate any point inside S and the projection of the process on the point will possess Gaussian distribution. For example, we can write

$$P\left(\bar{Q}^\infty(t,y)\notin\left(\lambda s\int_y^{t+y}\bar{F}(u)\,du\pm\sqrt{s}C_1\nu(y)\right) \text{ for some } t\in[0,t_0],\ y\in[0,\infty)\,\middle|\,B(0)\right)$$

$$\geq P(U_0\leq x)P\left(\bar{Q}_0^\infty(t^*,y^*)\geq\lambda s\int_{y^*}^{t^*+x+y^*}\bar{F}(u)\,du+\sqrt{s}C_1\nu(y^*)\right)$$

$$>0$$

for any $t^*\in[0,t_0]$ and $y^*\in[0,\infty)$. □

References

[1] Adler RJ (1990) *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes* (Institute of Mathematical Statistics, Hayward, CA).
[2] Anantharam V (1989) How large delays build up in a $GI/G/1$ queue. *Queueing Systems* 5(4):345–367.
[3] Asmussen S (1985) Conjugate processes and the simulation of ruin problems. *Stochastic Processes Their Appl.* 20(2):213–229.
[4] Asmussen S (2003) *Applied Probability and Queues*, 2nd ed. (Springer, New York).
[5] Asmussen S, Glynn PW (2007) *Stochastic Simulation: Algorithms and Analysis*, Stochastic Modelling and Applied Probability, Vol. 57 (Springer, New York).
[6] Billingsley P (2008) *Probability and Measure* (John Wiley & Sons, Hoboken, NJ).
[7] Blanchet J, Lam H (2011) Importance sampling for actuarial cost analysis under a heavy traffic model. Jain S, Creasey RR, Himmelspach J, White KP, Fu MC, eds. *Proc. Winter Simulation Conf.* (IEEE Computer Society, Washington, DC), 3817–3828.
[8] Blanchet J, Chen X, Lam H (2014) Two-parameter sample path large deviations for infinite-server queues. *Stochastic Systems*. Forthcoming.
[9] Blanchet J, Glynn PW, Lam H (2009) Rare event simulation for a slotted time $M/G/s$ model. *Queueing Systems* 63(1–4):33–57.
[10] Breiman L (1992) *Probability*. Classics in Applied Mathematics, Vol. 7 (SIAM, Philadelphia).
[11] Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, Zhao L (2005) Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* 100(469):36–50.
[12] Bucklew JA (2004) *Introduction to Rare Event Simulation*, Springer Series in Statistics (Springer, New York).
[13] Dembo A, Zeitouni O (1998) *Large Deviations Techniques and Applications*, Vol. 2 (Springer, New York).
[14] Foley RD (1982) The non-homogeneous $M/G/\infty$ queue. *Opsearch* 19(1):40–48.
[15] Foss SG, Kalashnikov VV (1991) Regeneration and renovation in queues. *Queueing Systems* 8(1):211–223.
[16] Glynn PW (1995) *Large deviations for the infinite server queue in heavy traffic*, Kelly FP, Williams RJ, eds. *Stochastic Networks*, Mathematics and Its Applications, Vol. 71 (Springer, New York), 387–394.
[17] Glynn PW, Whitt W (1994) Large deviations behavior of counting processes and their inverses. *Queueing Systems* 17(1–2):107–128.
[18] Heidelberger P (1995) Fast simulation of rare events in queueing and reliability models. Nance RE, ed. *ACM Trans. Modeling Comput. Simulation* 5(1):43–85.
[19] Juneja S, Shahabuddin P (2006) Rare-event simulation techniques: An introduction and recent advances. Henderson SG, Nelson BL, eds. *Handbooks in Operations Research and Management Science*, Vol. 13 (North-Holland, Amsterdam), 291–350.
[20] Krichagina EV, Puhalskii AA (1997) A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems* 25(1–4):235–280.
[21] Pang G, Whitt W (2010) Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* 65(4):325–364.
[22] Reed J (2009) The $G/GI/N$ queue in the Halfin-Whitt regime. *Ann. Appl. Probab.* 19(6):2211–2269.
[23] Resnick SI (2007) *Extreme Values, Regular Variation, and Point Processes* (Springer, New York).
[24] Ridder A (2009) Importance sampling algorithms for first passage time probabilities in the infinite server queue. *Eur. J. Oper. Res.* 199(1):176–186.
[25] Rudin W (1976) *Principles of Mathematical Analysis*, Vol. 3 (McGraw-Hill, New York).
[26] Sadowsky JS (1991) Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue. *IEEE Trans. Automatic Control* 36(12):1383–1394.
[27] Siegmund D (1976) Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* 4(4):673–684.
[28] Szechtman R, Glynn PW (2002) Rare event simulation for infinite server queues. Yücesan E, Chen C.-H., Snowdon JL, Charnes JM, eds., *Proc. Winter Simulation Conf.* (IEEE Computer Society, Washington, DC), 416–423.
[29] Wheeden RL, Zygmund A (1977) *Measure and Integral: An Introduction to Real Analysis* (Marcel Dekker, New York).