Rare-Event Simulation for Many-Server Queues

Jose Blanchet and Henry Lam

Abstract

We develop rare-event simulation methodology for the analysis of loss events in a many-server loss system. We focus on the scenario when both the number of servers and the number of customers in the system go to infinity in such a way that traffic intensity remains fixed and is strictly less than one. Particular interest is 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 descriptor.

While there is vast literature on rare-event simulation algorithms for queues with fixed number of servers, few algorithms exist for queueing systems with many servers. In systems with single or a fixed number of servers, random walk representations are often used to analyze associated rare events (see for example Siegmund (1976), Asmussen (1985), Anantharam (1988), Sadowsky (1991) and Heidelberger (1995)). 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 (as we shall demonstrate in the basic development of our ideas) but instead the challenge in their rare-event analysis lies on the fact that the system description is typically infinite dimensional (measure-valued). 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 order to illustrate our ideas we focus on the problem of estimating the steady-state loss probability in many-server loss systems (the precise assumptions and formulation are given in Section 1). More precisely, 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 immediately starts service occupying a server. If the customer finds all the servers busy, he leaves the system immediately and the system incurs a "loss". The steady-state loss probability (i.e. the long term proportion of customers that are lost) is rare if the traffic intensity (arrival rate to 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 (1995), who developed large deviations asymptotics for the queue length process of the infinite-server queue with high arrival rates. Based on this result, Szechtman and Glynn (2002) developed a corresponding rare-event algorithm for the number in system of an infinite-server queue. Related results for first passage time probabilities have also been obtained by Ridder (2009) in the setting of Markovian queues. Blanchet, Glynn and Lam (2009) 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, Glynn and Lam (2009) is the closest in spirit to our methodology here, but the slotted-time nature, the Markovian

structure and the fact that the service times were bounded were used in a crucial way to avoid the main technical complications involved in dealing with measure-valued descriptors.

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 exponentials, Weibull and lognormal distributions. A key element of our algorithm, in addition to the use of measure-valued process, is the application of weak convergence limits by Krichagina and Puhalskii (1997) and Pang and Whitt (2009). As we shall see, the weak convergence results are necessary because via a suitable extension of regenerative simulation (see Section 2) the steady-state loss probability of the system can be transformed to a first passage problem of the measure-valued process starting from an appropriate set, suitably chosen by means of such weak convergence analysis. However, unlike infinite-server system, the capacity constraint (s servers) introduces a boundary that forces us to work with the sample path and to tract the whole process description. We will also see that the properties (and especially "decay" behavior) of the steady-state measure plays an important role in controlling the efficiency of the algorithm in the case of unbounded service time. In fact, new logarithmic asymptotic results of steady-state convergence (in the sense described in Section 4) are derived along our way to prove algorithmic efficiency.

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 rare event. In order 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 when the change of measure resembles the conditional distribution given the occurrence of rare event. Construction of good changes of measure often draws on associated large deviations theory (see Asmussen and Glynn (2007), Chapter 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 (2004), Asmussen and Glynn (2007) and Juneja and Shahabuddin (2006) 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, intra-company 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 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. While 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 measure-valued descriptor becomes valuable. The full development of this application is presented in Blanchet and Lam (2010).

The organization of the paper is as follows. In Section 1 we will lay out our $GI/G/s$ model assumptions and indicate our main results. In Section 2 we will explain and describe in detail our simulation methodology. Section 3 will focus on the proof of algorithmic efficiency and large deviations asymptotics, while Section 4 will be devoted to the use of weak convergence results mentioned earlier for the design of an appropriate regenerative set. Finally, we will provide numerical results in Section 5.

1 Main Results

1.1 Assumptions on Arrivals and Service Time Distribution

First we state the assumptions of our model, namely 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 its counting process of the 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 U_k has exponential moments in a neighborhood of the origin, and let $\kappa_s(\theta) = \log Ee^{\theta U_k}$ be the logarithmic moment generating function of U_k. It is easy to see that $\kappa_s(\theta) = \kappa^0(\theta/s)$ where $\kappa^0(\theta) = \log Ee^{\theta U_k^0}$ is the logarithmic moment generating function of the interarrival time in the base system.

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

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

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

Now we impose a few assumptions on $\psi_N(\cdot)$. First, we assume Dom $\psi_N \supset \mathbb{R}_+$ (that Dom $\psi_N \supset \mathbb{R}_-$ is obvious from the definition of $\psi_N(\cdot)$), and hence Dom $\psi_N = \mathbb{R}$. We also 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 $\psi_N'(\mathbb{R}_+) = [\lambda, \infty)$. Finally, we insist 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. (Its use is in Lemma 4 as a regularity condition to prevent the blow-up of likelihood ratio due to sample paths that hit overflow very early).

Under these assumptions we have for any $0 = t_0 < t_1 < \cdots < t_m < \infty$ and $\theta_1, \ldots, \theta_m \in$ Dom ψ_N,

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

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 (1991) and Glynn (1995). In the case of Poisson arrival, for example, the interarrival times are exponential and we have $\kappa(\theta) = \log(\lambda/(\lambda - \theta))$. This gives $\psi_N(\theta) = \lambda(e^\theta - 1)$ and Dom $\psi_N = \mathbb{R}$.

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 complementary distribution function $\bar{F}(\cdot)$. We assume that $F(\cdot)$ has a density $f(\cdot)$ that satisfies

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

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,

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

for some $c_1, c_2 > 0$. In other words, $\bar{F}(\cdot)$ decays faster than any power law. Thus assumption (4) covers Weibull and log-normal service times, which have been observed to be important models in call center analysis (see e.g. ??).

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 Section 3.

1.2 Problem Formulation and Main Results

In this subsection we describe our problem formulation, and point out our main results and contributions, which will be elaborated in subsequent sections.

To start with, we are interested in the steady-state loss probability defined as

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

By regenerative theory, this can be expressed by a generalization of Kac's formula (see Breiman (1968)) as:

$$P_\pi(\text{loss}) = \frac{E_A N_A}{\lambda s E_A \tau_A} \tag{7}$$

Here A is a regenerative set that is visited by the chain infinitely often. $E_A[\cdot]$ denotes the expectation with initial state distributed according to the steady-state distribution conditioned on being in A. N_A is the number of loss before returning to set A, and τ_A is the time back to A.

Our first result is on the large deviations behavior of (7):

Theorem 1. *The steady-state loss probability* (7) *is exponentially decaying in s with decay rate* I^* *defined in* (22)*. In other words,*

$$\lim_{s\to\infty} \frac{1}{s} \log P_\pi(loss) = -I^* \tag{8}$$

A key element of this result is the computability of I^*, as shown in (22). In fact, (22) comes from the fact that I^* is the infimum of the rate functions for the probabilities that a coupled $GI/G/\infty$ system (defined in Section 1.4 below) hits level s at different fixed times, assuming the system starts at any point in a suitably chosen regenerative set A. These individual fixed-time rate functions are readily computable with formula given in (20). Moreover, they elicit monotone properties (see Lemma 3), which leads to the convenient formula for I^* in (22).

Next, let us discuss our simulation methodology. Note that (6) cannot be directly simulated, but formula (7) provides a basis for regenerative simulation (see Asmussen and Glynn (2007), Chapter 4). After identifying a regenerative set 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_A and τ_A every time it hits A, and output the sample means of N_A and τ_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 A-cycle).

However, recognizing that loss is a rare event (with exponential decay rate I^*), this method will take an exponential amount of time to get a specified relative error. This is regardless of the choice of A: if A is large, it takes short time to regenerate i.e. τ_A is small, and consequently the number of losses reported as the numerator $E_A N_A$ of (7) is almost always zero; whereas if A is small, it takes a long time to regenerate. In order to circumvent this issue, our strategy is the following. We choose A to be a "central limit" set so that $E_A \tau_A$ is not exponentially large in s (and not exponentially small either by artificially "discretizing" the time scale; see Section 2.1). This isolates the rarity of loss to the numerator $E_A N_A$. In other words, it is very difficult for the process to reach overflow in an 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 A-cycle.

We point out two practical observations using this approach: First, τ_A and N_A can be estimated separately i.e. one can "split" the process every time it hits A: one of which we apply importance sampling to get one sample of N_A and is then discarded, to the other one we apply the original measure to get one sample of τ_A and also set the initial position for the next A-cycle (see Asmussen and Glynn (2007)). Secondly, to get an estimate of standard deviation one has to use batch estimates since the samples obtained this way possess serial correlations (see again Asmussen and Glynn (2007)). In other words, one has to divide the simulated chain into several segments of equal number of time steps. 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 Section 5 when we discuss numerical results.

We summarize our approach as follows:

Algorithm 1

1. Set a regenerative set A. Initialize the $GI/G/s$ queue's status as any point in A.
2. Run the queue. Each time the queue hits a point in A, say x, do the following: Starting from x,
   (a) Use importance sampling to sample one N_A, the number of loss in a cycle.
   (b) Use crude Monte Carlo to sample one τ_A, the regenerative time. The final position of this queue is taken as the new x.

3. Divide the queue into several segments of equal time length. Compute the estimate of steady-state loss probability using the batch samples.

Along the above discussion, the main contribution of this paper is, in addition to the large deviations asymptotic in Theorem 1, the construction of an efficient importance sampling scheme together with the identification of a suitable regenerative set A:

Theorem 2. *The estimator using the regenerative set A in* (11) *and the importance sampler given by Algorithm 2 is asymptotically optimal.*

The novel feature of our algorithm is that A lies in a measure-valued state space. This is because the large deviations behavior of $GI/G/s$ queue (and also its coupled $GI/G/s$ counterpart introduced in Section 1.4) hitting an overflow at a fixed time depends on the initial measure-valued status, despite that the large deviations probability is only on the queue length. This consequently poses a dependence on the measured-valued description of the process for sampling the numerator of (7) i.e. a choice of A using the queue length description alone is not enough to single out the rarity behavior of $E_A N_A$. This issue will be reported rigorously in Section 2 and 3 when we describe our importance sampling algorithm in detail.

We emphasize, methodologically, that our importance sampling algorithm utilizes the representation of a (coupled) $GI/G/s$ as a point process. This is used for proving the "continuity" of likelihood ratio that prevents a blow-up of "overshoot" at the first passage time (see the proof of Theorem 3 in Section 3). This point process representation, we believe, 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. (2011?). Secondly, our algorithm requires essentially the information of the whole sample path of the system due to a randomization of time horizon, in contrast to the algorithm proposed in Szechtman and Glynn (2002) for estimating fixed-time probability. Moreover, one other interesting feature of our approach is that the large deviations asymptotic (8) is concurrently derived, along with algorithmic efficiency, through Jensen's inequality and the upper bound on the variance of our importance sampler (see the proof of Theorem 1 and 2 in Section 3).

Finally, we find a regenerative set A, given by (11), that possesses the following properties:

Proposition 1. *In the $GI/G/s$ system,*

$$\lim_{s\to\infty} \frac{1}{s} \log E_A \tau_A^p = 0 \tag{9}$$

and

$$\limsup_{s\to\infty} \frac{1}{s} \log E_A N_A^p \leq 0 \tag{10}$$

for any $p > 0$.

When $p = 1$, Proposition 1 in particular states that the expected regenerative time is subexponential in s, a crucial feature to encapsulate the denominator of (7). As discussed above, this ensures the efficiency of Algorithm 1 and helps identify the rate function I^*. The proposition is stated in general p because such logarithmic estimate is also used to describe the property of N_A in the numerator of (7), since N_A involves the first passage behavior of possessing a loss before τ_A (see the proof of Theorem 3). Interestingly, the proof of Proposition 1 requires an invocation of Borell's inequality for Gaussian random field, which arises as the diffusion limit of the coupled $GI/G/\infty$ queue.

1.3 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 $Q(t, y)$ to 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 $B(t) = \inf\{t - A_k : A_k \leq t\}$ 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. While 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. To insist on a Markov description of the 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. In the case of bounded service time over $[0, M]$ the state-space is further restricted to $\mathcal{D}[0, M] \times \mathbb{R}_+$.

1.4 A Coupling $GI/G/\infty$ System

In 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. The purpose of introducing this system is to remove the nonlinear "boundary" condition on the queue, hence leading to tractable analytical results that we can harness, while the coupling provides a link from this system back to the original $GI/G/s$ system. In this paper we shall use the superscript "∞" to denote quantities in the $GI/G/\infty$ system, so 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. $\bar{Q}^\infty(t, y) = Q^\infty(t, y) - Q^\infty(0, t + y)$.

2 Simulation Methodology

As we have discussed, two key issues in our algorithm are the choice of regenerate set and the importance sampling algorithm. We will present them in detail in Section 2.1 and Section 2.2 respectively.

2.1 Regenerative Set

Pick a fixed small time interval Δ (one choice, for example, is say 1/5 of the mean of service time). We choose A to be

$$A = \{Q(t, y) \in J(y) \text{ for all } y \in [0, \infty),\ t \in \{0, \Delta, 2\Delta, \ldots\}\} \tag{11}$$

Here $J(y)$ is the interval

$$J(y) = \left(\lambda s \int_y^\infty \bar{F}(u)du - \sqrt{s}C^*\xi(y),\ \lambda s \int_y^\infty \bar{F}(u)du + \sqrt{s}C^*\xi(y)\right) \tag{12}$$

for some well chosen constant $C^* > 0$ (discussed below and in Section 4) and

$$\xi(y) = \nu(y) + \gamma \int_y^\infty \nu(u)du \tag{13}$$

where

$$\nu(y) = \left(\lambda \int_y^\infty \bar{F}(u)du\right)^{1/(2+\eta)} \tag{14}$$

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

A few comments are in place. First, we define A in a generalized sense that it depends on both $Q(t, \cdot)$ and t; it is easy to check that Kac's formula (7) still holds. We introduce the lattice $\{\Delta, 2\Delta, \ldots\}$ to avoid exponentially small (and even zero) τ_A. In particular, Δ will set a lower bound for the regenerative time. Next, the form of $J(y)$ comes from the heavy traffic limit of $GI/G/\infty$ queue. Pang and Whitt (2009) 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 $R(t, y)$ on the state space $\mathcal{D}[0, \infty)$ with $\text{var}(R(t, y)) \to \lambda c_a^2 \int_y^\infty \bar{F}(u)^2 du + \lambda \int_y^\infty F(u)\bar{F}(u)du$ as $t \to \infty$, where c_a is the coefficient of variation of the interarrival times. Our regenerative set 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 $Q^\infty(\infty, \cdot)$. It can be proved that this choice of A indeed leads to regenerative time that is subexponential in s, stated as in Proposition 1.

The proof, which harnesses a technical property of the limiting Gaussian process and is exposed in Section 4, shows that the coupled $GI/G/\infty$ system is close enough to $GI/G/s$ to give a useful upper bound for τ_A, to which end we can apply the diffusion limit for $GI/G/\infty$. The slower decay rate of the confidence band width is a technical adjustment to enlarge A so that a subexponential regeneration for the $GI/G/\infty$ system is guaranteed. In fact, for the case of bounded service time, it suffices to set $\eta = 0$, $\xi(y) \equiv$ constant, or a truncated $\xi(y)$ (see numerical example in Section 6).

The first limit of Proposition 1 (putting $p = 1$) in particular concludes that it is sufficient to focus on $E_A N_A$ to construct an asymptotically optimal algorithm.

Remark 1. *One can check easily that the interval $J(\cdot)$ is valid for large enough s i.e. the interval contains at least one non-negative integer for any value of y. In fact, observe that the length of $J(y)$ is continuous and decreasing in y, and let*

$$l(s) = \inf\left\{y > 0 : \sqrt{s}C^*\xi(y) < \frac{1}{2}\right\} \tag{15}$$

When the length of $J(y)$ decreases to 1 i.e. $y = l(s)$, we have $\lambda s \int_y^\infty \bar{F}(u)du \le (\lambda/(C^*)^{2+\eta})(\sqrt{s}C^*\xi(y))^{2+\eta}/s^{\eta/2} \le (\lambda/(2^{2+\eta}(C^*)^{2+\eta})s^{\eta/2})$ which is tiny when s is large. Afterwards the length of $J(y)$ decays slower

than its center. So $J(y)$ includes only the point 0 when $y \geq l(s)$ and s is large enough. Obviously it includes at least one point when $y < l(s)$. Therefore $J(y)$ is never empty for any $y \geq 0$ for large enough s. When s is small, we can modify any empty $J(y)$ to include the two closest integers to the two ends of the interval. This obviously does not affect the asymptotic behavior of $J(y)$ as $s \to \infty$.

Remark 2. *One may ask whether it is possible to define A in a finite-dimensional fashion, instead of introducing the functional "confidence band" in* (11)*. For example, one may divide the the domain of y into segments $[y_i, y_{i+1}), i = 0, 2, \ldots, r(s) - 1$ 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 non-identical. One then define the regenerative set as $\{Q(t, \cdot) : Q(t, y_i) - Q(t, y_{i+1}) \in A_i \text{ for } i = 0, \ldots, r(s) - 1\}$ for some well-defined sets A_i's. As we will see in the arguments in the subsequent sections, the important criteria of a good regenerative set is that 1) it consists of a significantly large region in the central limit theorem, so that it is visited infinitely often. 2) its deviation from the mean of $Q(t, y)$ is small, in the sense that the distance between any $Q(t, y)$ in this regenerative set and the mean of the steady-state of $Q(t, y)$, at every point $y \in [0, \infty)$, has order $o(s)$. Otherwise the large deviations of loss starting from two different $Q(t, y)$ in the regenerative set can be different, which causes a breakdown of algorithmic efficiency.*

In this regard, one has to fine-tune the scale of the segments to preserve algorithmic efficiency. On the other hand, central limit theorem no longer holds if the segments are overly fine-tuned. The functional definition of A in (11) *happens to balance both conditions, with the result that the algorithm using such A is asymptotically optimal. Nevertheless, we believe that similar arguments in this paper can show that for every $\epsilon > 0$, one can choose an order $O(1)$ segmentation on the domain of y i.e. each segment except the last one is small but does not depend on s, so that an ϵ-deficient logarithmically efficient algorithm can be constructed i.e. $\liminf_{s\to\infty}(\log EX^2)/(\log P_\pi(loss)^2) \geq 1 - \epsilon$ for the estimator X.*

Remark 3. *The Δ introduced in* (11) *is a natural definition of A. In fact, from our definition of $P_\pi(loss)$ as the long run probability of loss (see the discussion after* (7)*), we get*

$$P_\pi(loss) = \lim_{k\to\infty} \frac{number\ of\ losses\ before\ k\Delta}{number\ of\ arrivals\ before\ k\Delta}$$

In other words, if we define $\tilde{N}_k(\Delta)$ to be the number of loss from $(k-1)\Delta$ to $k\Delta$, and call the Δ-skeleton of $\{W_t\}_{t\geq 0}$ to be $\{W_{k\Delta}, \tilde{N}_k(\Delta)\}_{k\geq 0}$, then applying Kac's formula to this Δ-skeleton and using A as in (11)*, we get*

$$P_\pi(loss) = \frac{E_A N_A}{\lambda s \Delta E_A \rho_A} = \frac{E_A N_A}{\lambda s E_A \tau_A} \tag{16}$$

where ρ_A is the number of Δ-units before reaching A again. This justifies our choice of A that depends on Δ. Of course, one needs to beware of the conditions that allow the use of Kac's formula for this Markov chain. Under our smoothness assumption on the distribution of service time, we believe that such conditions would be satisfied, but a detailed argument will be a significant distraction from the focus of this paper.

Remark 4. *We argue that the loss probability is invariant in Δ. By applying regenerative theory to the Δ-skeleton of $\{W_t\}_{t\geq 0}$,* (16) *is equal to*

$$\frac{E_\pi \tilde{N}(\Delta)}{\lambda s \Delta} \tag{17}$$

where $\tilde{N}(t)$ denotes the number of loss before t. We claim that this is invariant in Δ. Indeed, for any $d > 0$, we have

$$E_\pi \tilde{N}(nd) = E_\pi \left[\sum_{k=1}^{n} (\tilde{N}(kd) - \tilde{N}((k-1)d)) \right] = nE_\pi[\tilde{N}(d)]$$

by the definition of π, and hence $E_\pi \tilde{N}(nd)/nd = E_\pi \tilde{N}(d)/d$ for any $n \in \mathbb{N}_+$. It is then easy to see that (17) *is invariant for all rational Δ, and by a continuity argument it is true also for all $\Delta \in \mathbb{R}_+$.*

2.2 Simulation Algorithm

First we shall explain some heuristic in constructing the algorithm. As we discussed earlier, the choice of A isolates the rarity of steady-state loss probability to $E_A N_A$, which in turn is small because of the difficulty in approaching overflow from A. So on an exponential scale, $E_A N_A \approx P_A(\tau_s < \tau_A)$, where $P_A(\cdot)$ is the probability measure with initial state distributed as the steady-state distribution conditional on A, and $\tau_s = \inf\{t > 0 : Q(t) > s\}$ is the first passage time to overflow. Observe that the probability $P_A(\tau_s < \tau_A)$ 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 $P_A(\tau_s < \tau_A)$ is the probability $P_A(Q^\infty(t^*) > s)$ 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 $P_A(Q^\infty(t^*) > s)$. 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, Glynn and Lam (2009). The likelihood ratio will then comprise of 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 Section 3).

Hence our algorithm will take the following steps. Suppose we start from some position in 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 $P_A(Q^\infty(t) > s)$ for the realized time horizon t. This involves sequential tilting of both the arrivals and service times (see below). 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 A is reached. At this time one sample of N_A is recorded together with the likelihood ratio.

The key questions now are: 1) the sequential tilting scheme of arrivals and service times given a realized time horizon 2) the distribution of the random time 3) likelihood ratio of 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 Section 3.

2.2.1 Sequential Tilting Scheme

Denote $P_r(\cdot)$ and $E_r[\cdot]$ as the probability measure and expectation with initial system position r. Suppose we want to estimate $P_r(Q^\infty(t) > s)$ efficiently for a $GI/G/\infty$ system as $s \nearrow \infty$, where $r(\cdot) \in J(\cdot) \subset D[0,\infty)$ (so that $r(y)$ is the number of initial customers still in the system at time y). An important clue is an invocation of Gartner-Ellis Theorem (see Dembo and Zeitouni (1998)) to obtain 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 (1995) and Szechtman and Glynn (2002), which we briefly expose here.

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

$$\psi_t(\theta) := \int_0^t \psi_N(\log(e^\theta \bar{F}(t-u) + F(t-u)))du \tag{18}$$

This is the Gartner-Ellis limit (see for example Dembo and Zeitouni (1998)) 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 (1995) for a proof. It uses (3) and the definition of Riemann sum; alternatively, see Lemma 6 in Section 3 as a generalization of this result). Let us state the following properties of $\psi_t(\cdot)$ for later convenience:

Lemma 1. *$\psi_t(\cdot)$ is defined on $\mathbb{R}$, twice continuously differentiable, strictly convex and steep.*

Next let $a_t = 1 - \lambda \int_t^\infty \bar{F}(u)du$. Note 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,

$$P_r(Q^\infty(t) > s) = P(\bar{Q}^\infty(t) > a_t s + o(s)) \tag{19}$$

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 that $a_t = 1 - \lambda \int_t^\infty \bar{F}(u)du > \lambda \int_0^t \bar{F}(u)du = \psi_t'(0)$. Then under our current assumptions Gartner-Ellis Theorem concludes that $(1/s)\log P_r(Q^\infty(t) > s) \to -I_t$ where

$$I_t = \sup_{\theta \in \mathbb{R}}\{\theta a_t - \lambda_t(\theta)\} = \theta_t a_t - \psi_t(\theta_t) \tag{20}$$

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 2. *θ_t satisfies the following:*

1. *$\theta_t > 0$ is non-increasing in t for all $t > 0$*
2. *$\lim_{t\to 0} \theta_t = \infty$*
3. *$\lim_{t\to\infty} \theta_t = \theta_\infty$ where θ_∞ is the unique positive root of the equation $\psi_\infty'(\theta) = 1$, and*

$$\psi_\infty(\theta) = \int_0^\infty \psi_N(\log(e^\theta \bar{F}(u) + F(u)))du \tag{21}$$

Lemma 3. *I_t satisfies the following:*

1. *I_t is non-increasing in t for $t > 0$.*
2. *$\lim_{t\to\infty} I_t = \inf_{t>0} I_t = I^*$ where*

$$I^* = \theta_\infty - \psi_\infty(\theta_\infty) \tag{22}$$

3. *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 mean of $\bar{Q}^\infty(t)$ 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

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

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

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

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)

$$\frac{P(U_i)}{\tilde{P}^t(U_i)} = \frac{e^{s\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i}}{e^{\theta_t}\bar{F}(t-A_i) + F(t-A_i)} \tag{23}$$

and

$$\frac{P(V_i)}{\tilde{P}^t(V_i)} = \frac{e^{\theta_t}\bar{F}(t-A_i) + F(t-A_i)}{e^{\theta_t I(V_i > t-A_i)}} \tag{24}$$

We tilt the process using (23) and (24) until the time that we know overflow will happen at time t i.e. $t \wedge \tau_s[t]$ where $\tau_s[t] = \inf\{u > 0 : r(t) + \sum_{i=1}^{N_s(u)} I(V_i > t - A_i) > s\}$. The overall likelihood ratio on the set $Q^\infty(t) > s$ will be

$$
\begin{aligned}
L &= \prod_{i=1}^{N_s(\tau_s[t])-1} \frac{e^{s\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))}}{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)} \prod_{i=1}^{N_s(\tau_s[t])} \frac{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)}{e^{\theta_t I(V_i>t-A_i)}} \\
&= \exp\left\{ s \sum_{i=1}^{N_s(\tau_s[t])-1} \psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i - \theta_t \sum_{i=1}^{N_s(\tau_s[t])} I(V_i > t-A_i) \right\} \\
&\quad (e^{\theta_t}\bar{F}(t-A_{\tau_s[t]})+F(t-A_{\tau_s[t]})) \qquad (25)
\end{aligned}
$$

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

Proposition 2.

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

Proof. The proof follows from Szechtman and Glynn (2002), but for completeness (and also due to 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 $0 = A_0 < A_1 < A_2 < \ldots < A_{N_s(\tau_s[t])}$ 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 non-decreasing in u when $\theta_t > 0$, and $\tau_s[t] \le t$ on $Q^\infty(t) > s$, we have

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

on $Q^\infty(t) > s$. Hence (25) gives

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

which yields the proposition.

□

2.2.2 Distribution of Random Horizon

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

$$P(\tau = T + k\delta) = \frac{1}{(k+1)^2} - \frac{1}{(k+2)^2} \quad \text{for} \quad k = 0, 1, 2 \dots \tag{26}$$

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 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 of the fact that $\lim_{t\to 0} \theta_t = \infty$ in Lemma 3 Part 1; see also discussion point 3) below and Section 3). 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 by the steepness assumption for small enough t). This is the rate function of $N_s(t)$ evaluated at $1 - \lambda EV$.

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

$$\tilde{I}_T > 2I^* \tag{27}$$

which always exists by the following lemma:

Lemma 4. *$\tilde{I}_t$ satisfies the following:*

1. *$\tilde{I}_t$ is non-increasing in t for $t < \eta$ for some small $\eta > 0$.*
2. *$\tilde{I}_t \to \infty$ as $t \searrow 0$.*

Remark 5. *In fact by looking at the arguments in the next section, 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 coefficient of variation, 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 Section 6.*

2.2.3 Likelihood Ratio

After sampling the randomized time horizon, we accelerate the process using the sequential tilting scheme (23) and (24) with a realized $\tau = t$. But since we are now interested in the first passage probability, we tilt the process until $t \wedge \tau_s \wedge \tau_A$ (rather than $\tau_s[t]$ defined above). If $t \wedge \tau_s < \tau_A$, we continue the $GI/G/s$ system under the original measure. Also, to prevent a blow-up of likelihood ratio close to $t = 0$, we use the original measure throughout the whole process whenever $\tau = T$ (see the proof of efficiency next section). Now denote $\tilde{E}[\cdot]$ and $\tilde{P}(\cdot)$ as the importance sampling measure. We have

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

(with $\tilde{P}^T(\cdot) = P(\cdot)$). So the overall likelihood ratio $L = L(W_\cdot)$ on the set $\tau_s < \tau_A$ is given by

$$\begin{aligned} L = \frac{dP}{d\tilde{P}} &= \frac{P(W_u, 0 \le u \le \tau_s)}{\sum_{k=0}^{\infty} P(\tau = T + k\delta)\tilde{P}^{T+k\delta}(W_u, 0 \le u \le \tau_s)} \\ &= \frac{1}{\sum_{k=0}^{\infty} P(\tau = T + k\delta)L_{T+k\delta}^{-1}} \end{aligned} \tag{28}$$

where $L_t = L_t(W_\cdot)$ is the individual likelihood ratio as a sequential product of (23) and (24) up to $t \wedge \tau_s$ i.e.

$$L_t = \begin{cases} \exp\left\{s\sum_{i=1}^{N_s(\tau_s)-1}\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i - \theta_t\sum_{i=1}^{N_s(\tau_s)-1}I(V_i > t-A_i)\right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } t \ge \tau_s \\ \exp\left\{s\sum_{i=1}^{N_s(t)-1}\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))U_i - \theta_t\sum_{i=1}^{N_s(t)-1}I(V_i > t-A_i)\right\} \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } t < \tau_s \end{cases} \tag{29}$$

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

2.2.4 The Algorithm

We now state our algorithm. Assuming we start from $r(\cdot) \in J(\cdot)$ with a given initial age $B(0)$, do the following:

Algorithm 2

1. Set $A_0 = 0$. Also initialize $N_A \leftarrow 0$, $L \leftarrow 0$ and $\tau_s \leftarrow \infty$.
2. Sample τ according to (26). Say we get a realization $\tau = t$.
3. Simulate U_0 according to the initial age $B(0)$. Set $A_1 = U_0$. Check if τ_A is reached, in which case go to Step 7.
4. Starting from $i = 1$, repeat the following (setting θ_t as the one in (20) for $t > T$ and 0 for $t = T$):
   (a) Generate V_i according to
   $$\tilde{P}^t(V_i \in dy) := \begin{cases} \frac{f(y)}{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)} & \text{for } 0 \le y \le t - A_i \\ \frac{e^{\theta_t}f(y)}{e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)} & \text{for } y > t - A_i \end{cases}$$
   (b) Generate U_i according to
   $$\tilde{P}^t(U_i \in dy) := e^{-s\psi_N(\log(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i)))y}(e^{\theta_t}\bar{F}(t-A_i)+F(t-A_i))P(U_i \in dy)$$
   (c) Set $A_{i+1} = U_i + A_i$.
   (d) If τ_A is reached in $[A_i, A_{i+1})$, go to Step 7.
   (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 $N_A \leftarrow N_A + 1$, and go to Step 5.

(f) If $A_{i+1} \geq t$, go to Step 5.

(g) Update $i \leftarrow i+1$.

5. Repeat the following:

(a) Generate V_i and U_i under the original measure. Set $A_{i+1} = U_i + A_i$.

(b) If τ_A is reached in $[A_i, A_{i+1})$, go to Step 6.

(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 $N_A \leftarrow N_A + 1$, and set $\tau_s \leftarrow A_{i+1}$ if in addition that $\tau_s = \infty$.

(d) Update $i \leftarrow i+1$.

6. Compute $LI(\tau_s < \tau_A)$ using (28) and (29).

7. Output $N_A LI(\tau_s < \tau_A)$.

3 Algorithmic Efficiency

In this section we will prove asymptotic optimality of the estimator outputted by Algorithm 2. To be more precise, we will identify I^* defined in (22) as the exponential decay rate of $E_A N_A$. The key result is the following:

Theorem 3. *The second moment of the estimator in Algorithm 2 satisfies*

$$\limsup_{s\to\infty} \frac{1}{s} \log \tilde{E}_r[N_A^2 L^2; \tau_s < \tau_A] \leq -2I^*$$

for any $r(\cdot) \in J(\cdot)$.

This result, together with Theorem 4 in the sequel, will expose a loop of inequality 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 $\lceil x \rceil = \min\{T + k\delta,\ k = 0, 1, \ldots : x \leq T + k\delta\}$. Also recall the definition $a_t = 1 - \lambda \int_t^\infty \bar{F}(u)du$.

Consider the likelihood ratio in (28):

$$
\begin{aligned}
& LI(\tau_s < \tau_A) = \frac{1}{\sum_{k=0}^{\infty} P(\tau = T + k\delta) L_{T+k\delta}^{-1}} I(\tau_s < \tau_A) \le \frac{L_{\lceil \tau_s \rceil}}{P(\tau = \lceil \tau_s \rceil)} I(\tau_s < \tau_A) \\
= \;& P(\tau = T)^{-1} I(\tau_s \le T; \tau_s < \tau_A) + P(\tau = \lceil \tau_s \rceil)^{-1} \exp \Bigg\{ s \sum_{i=1}^{N_s(\tau_s)-1} \psi_N(\log(e^{\theta_{\lceil \tau_s \rceil}} \bar{F}(\lceil \tau_s \rceil - A_i) \\
& + F(\lceil \tau_s \rceil - A_i))) U_i - \theta_{\lceil \tau_s \rceil} \sum_{i=1}^{N_s(\tau_s)-1} I(V_i > \lceil \tau_s \rceil - A_i) \Bigg\} I(\tau_s > T; \tau_s < \tau_A) \\
\le \;& C_1 I(\tau_s \le T; \tau_s < \tau_A) + \frac{C_2 \tau_s^3}{\delta^3} \exp\left\{ s\psi_{\lceil \tau_s \rceil}(\theta_{\lceil \tau_s \rceil}) - \theta_{\lceil \tau_s \rceil}(\bar{Q}^{\infty}(\tau_s, \lceil \tau_s \rceil - \tau_s) - 1) \right\} \\
& I(\tau_s > T; \tau_s < \tau_A) \\
\le \;& C_1 I(\tau_s \le T; \tau_s < \tau_A) + \frac{C_2 \tau_s^3}{\delta^3} \exp\left\{ -sI^* + \theta_{\lceil \tau_s \rceil} \left(s a_{\lceil \tau_s \rceil} + 1 - \bar{Q}^{\infty}(\tau_s, \lceil \tau_s \rceil - \tau_s) \right) \right\} \\
& I(\tau_s > T; \tau_s < \tau_A)
\end{aligned}
$$

where C_1 and C_2 are positive constants. Note that the second inequality comes from the fact that $\sum_{i=1}^{N_s(\tau_s)-1} \psi_N(\log(e^{\theta_{\lceil \tau_s \rceil}} \bar{F}(\lceil \tau_s \rceil - A_i) + F(\lceil \tau_s \rceil - A_i)))U_i$ 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 non-decreasing 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 Part 2. Now we have

$$
\begin{aligned}
& \tilde{E}_r[N_A^2 L^2; \tau_s < \tau_A] = E_r[N_A^2 L; \tau_s < \tau_A] \\
\le \;& C_1 E_r[N_A^2; \tau_s \le T; \tau_s < \tau_A] + \frac{C_2}{\delta^3} e^{-sI^*} E_r \Bigg[N_A^2 \tau_s^3 \exp\left\{ \theta_{\lceil \tau_s \rceil} \left(s a_{\lceil \tau_s \rceil} + 1 - \bar{Q}^{\infty}(\tau_s, \lceil \tau_s \rceil - \tau_s) \right) \right\}; \\
& \tau_s > T; \tau_s < \tau_A \Bigg] \qquad (30)
\end{aligned}
$$

Consider the first summand. By Holder's inequality $E_r[N_A^2; \tau_s \le T; \tau_s < \tau_A] \le (E_r[N_A^{2p}])^{1/p} (P_r(\tau_s \le T))^{1/q}$ 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 a straightforward invocation of Gartner-Ellis Theorem yields $\lim_{s\to\infty} \frac{1}{s} \log P(N_s(T) > s(1 - \lambda EV) + o(s)) = -\tilde{I}_T < -2I^*$ by our choice of T in (27). Combining these observations, and using Lemma 1, we get

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

for q close enough to 1.

In view of (30) and Dembo and Zeitouni (1998) Lemma 1.2.15, the proof will be complete once we can prove that

$$
\limsup_{s\to\infty} \frac{1}{s} \log E_r \left[N_A^2 {\tau_s}^3 \exp\left\{ \theta_{\lceil \tau_s \rceil} \left(s a_{\lceil \tau_s \rceil} + 1 - \bar{Q}^{\infty}(\tau_s, \lceil \tau_s \rceil - \tau_s) \right) \right\}; \tau_s > T; \tau_s < \tau_A \right] \le -I^* \qquad (31)
$$

To this end, we write

$$
\begin{aligned}
& E_r\left[N_A^2{\tau_s}^3\exp\left\{\theta_{\lceil\tau_s\rceil}\left(sa_{\lceil\tau_s\rceil}+1-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s)\right)\right\};\tau_s>T;\tau_s<\tau_A\right] \\
= \;& E_r\left[N_A^2{\tau_s}^3\exp\left\{\theta_{\lceil\tau_s\rceil}\left(s+1-\lambda s\int_{\lceil\tau_s\rceil}^{\infty}\bar{F}(u)du-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s)\right)\right\};\tau_s>T;\tau_s<\tau_A\right] \\
\leq \;& e^{C\theta_T\sqrt{s}}E_r\left[N_A^2\tau_s^3\exp\left\{\theta_{\lceil\tau_s\rceil}\left(s+1-r(\lceil\tau_s\rceil)-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s)\right)\right\};\tau_s>T;\tau_s<\tau_A\right] \\
= \;& e^{C\theta_T\sqrt{s}}\sum_{k=1}^{\infty}E_r\Bigg[N_A^2\tau_s^3\exp\left\{\theta_{\lceil\tau_s\rceil}\left(s+1-r(\lceil\tau_s\rceil)-\bar{Q}^\infty(\tau_s,\lceil\tau_s\rceil-\tau_s)\right)\right\};\lceil\tau_s\rceil=T+k\delta; \\
& \tau_A>T+(k-1)\delta\Bigg] \\
\leq \;& e^{C\theta_T\sqrt{s}}\sum_{k=1}^{\infty}(E_rN_A^{2p})^{1/p}(E_r{\tau_A}^{3q})^{1/q}(P_r(\tau_A>T+(k-1)\delta))^{1/h} \\
& \left(E_r\left[\exp\left\{l\theta_{T+k\delta}\left(s+1-r(T+k\delta)-\bar{Q}^\infty(\tau_s,T+k\delta-\tau_s)\right)\right\};T+(k-1)\delta<\tau_s\leq T+k\delta\right]\right)^{1/l} \\
= \;& e^{O(\sqrt{s})}\sum_{k=1}^{\infty}(E_rN_A^{2p})^{1/p}(E_r{\tau_A}^{3q})^{1/q}(P_r(\tau_A>T+(k-1)\delta))^{1/h} \\
& \left(E_r\left[\exp\left\{l\theta_{T+k\delta}\left(s+1-r(\tau_s)-\bar{Q}^\infty(\tau_s,T+k\delta-\tau_s)\right)\right\};T+(k-1)\delta<\tau_s\leq T+k\delta\right]\right)^{1/l} \quad (32)
\end{aligned}
$$

where C is a positive constant and $1/p+1/q+1/h+1/l=1$. The first inequality follows from the fact that $r(\cdot)\in J(\cdot)$ and Lemma 3 Part 1 while the second inequality follows from generalized Holder's inequality. The last equality holds because $r(\tau_s)-r(T+k\delta)=o(s)$, again since $r(\cdot)\in J(\cdot)$, for $T+(k-1)\delta<\tau_s\leq T+k\delta$.

We now analyze

$$
E_r\left[\exp\left\{l\theta_{T+k\delta}\left(s+1-r(\tau_s)-\bar{Q}^\infty(\tau_s,T+k\delta-\tau_s)\right)\right\};T+(k-1)\delta<\tau_s\leq T+k\delta\right] \quad (33)
$$

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. Such plot has been used in the study of $M/G/\infty$ system (see for example Foley (1982)). In this representation it is easy to see that the departure time of an arriving customer is the 45° projection of the point onto the x-axis. 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 and a downward 45° line joining at the point $(t,0)$. See Figure 1.

For notational convenience we denote $\bar{Q}^\infty_{t_1,t_2}[t_3,t_4]:=\sum_{i=N_s(t_1)+1}^{N_s(t_2)}I(t_3-A_i<V_i\leq t_4-A_i)$ 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 $T+k\delta\geq\tau_s$.

Figure 2 shows the region filled in 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.

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

$$\begin{aligned} & E_r[\exp\{l\theta_{T+k\delta}(s+1-r(\tau_s) - \bar{Q}^\infty_{0,\tau_s}[T+k\delta,\infty])\}; T+(k-1)\delta < \tau_s \le T+k\delta] \\ \le\ & E_r[e^{l\theta_{T+k\delta}H_k}; G_k > s - r(T+(k-1)\delta)] \end{aligned} \tag{34}$$

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 4:

$$\begin{array}{lcl} H^1_k(z) := \bar{Q}^\infty_{0,z}[T+(k-1)\delta, T+k\delta] & \subset & G^1_k(z) := \bar{Q}^\infty_{0,z}[T+(k-1)\delta,\infty] \\ H^2_k(z) := \bar{Q}^\infty_{z,T+k\delta}[T+(k-1)\delta, T+k\delta] & \subset & G^2_k(z) := \bar{Q}^\infty_{z,T+k\delta}[T+(k-1)\delta,\infty] \end{array}$$

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

Moreover, define $A^k_i, i = 1,\ldots,G_k$ to be the arrival times of all the customers that G_k is counting. Note that given the arrival times $A^k_i, i = 1,\ldots,G_k$, the events whether each of these customers falls into H_k are independent Bernoulli random variables with probability

$$p^k_i := \frac{\bar{F}(T+(k-1)\delta - A^k_i) - \bar{F}(T+k\delta - A^k_i)}{\bar{F}(T+(k-1)\delta - A^k_i)} \tag{35}$$

Hence we can write (34) as

$$\begin{aligned} & E_r[e^{l\theta_{T+k\delta}(H^1_k(z)+H^2_k(z))}; G_k > s - r(T+(k-1)\delta)] \\ =\ & E_r[E_r[e^{l\theta_{T+k\delta}(H^1_k(z)+H^2_k(z))}|A^k_i, i=1,\ldots,G_k]; G_k > s - r(T+(k-1)\delta)] \\ =\ & E_r[E_r[e^{l\theta_{T+k\delta}H^1_k(z)}|A^k_i, i=1,\ldots,G^1_k(z)]E_r[e^{l\theta_{T+k\delta}H^2_k(z)}|A^k_i, i=G^1_k(z)+1,\ldots,G^1_k(z)+G^2_k(z)]; \\ & G^1_k(z)+G^2_k(z) > s - r(T+(k-1)\delta)] \\ \le\ & E_r\left[e^{l\theta_{T+k\delta}G^1_k(z)} \prod_{i=G^1_k(z)+1}^{G^1_k(z)+G^2_k(z)} (1+(e^{l\theta_{T+k\delta}}-1)p^k_i); G^1_k(z)+G^2_k(z) > s - r(T+(k-1)\delta)\right] \end{aligned} \tag{36}$$

Let

$$p_k(z) := \sup_{A^k_i > z} p^k_i \le \frac{C\delta}{\bar{F}(T+k\delta - z)} \tag{37}$$

for some constant $C > 0$, where the inequality follows from (35). Also let

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

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

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 5. *We have*

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

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

When $p_k(z)$ is small enough, (36) is less than or equal to

$$\begin{aligned}
& E_r[e^{l\theta_{T+k\delta}G^1_k(z)}(1+(e^{l\theta_{T+k\delta}}-1)p_k(z))^{G^2_k(z)}; G^1_k(z)+G^2_k(z) > s-r(T+(k-1)\delta)] \\
= \;& E_r[E_r[e^{l\theta_{T+k\delta}G^1_k(z)+\log(1+(e^{l\theta_{T+k\delta}}-1)p_k(z))G^2_k(z)}; G^2_k(z) > s-r(T+(k-1)\delta)-G^1_k(z)|G^1_k(z), B(z)]] \\
\leq \;& E_r[\exp\{l\theta_{T+k\delta}G^1_k(z)-\theta_{T+(k-1)\delta}(s-r(T+(k-1)\delta)-G^1_k(z)) \\
& +\psi^2_{s,z,k}(\log(1+(e^{l\theta_{T+k\delta}}-1)p_k(z))+\theta_{T+(k-1)\delta})\}] \\
= \;& \exp\left\{\psi^1_{s,z,k}(l\theta_{T+k\delta}+\theta_{T+(k-1)\delta})-\theta_{T+(k-1)\delta}(s-r(T+(k-1)\delta))\right. \\
& \left.+\psi^2_{s,z,k}(\log(1+(e^{l\theta_{T+k\delta}}-1)p_k(z))+\theta_{T+(k-1)\delta})\right\} \\
= \;& \exp\left\{s\int_0^z \psi_N(\log(e^{l\theta_{T+k\delta}+\theta_{T+(k-1)\delta}}\bar{F}(T+(k-1)\delta-u)+F(T+(k-1)\delta-u)))du\right. \\
& -s\int_0^z \psi_N(\log(e^{\log(1+(e^{l\theta_{T+k\delta}}-1)p_k(z))+\theta_{T+(k-1)\delta}}\bar{F}(T+(k-1)\delta-u)+F(T+(k-1)\delta-u)))du \\
& -\theta_{T+(k-1)\delta}(s-r(T+(k-1)\delta))+s\psi_{T+(k-1)\delta}(\log(1+(e^{l\theta_{T+k\delta}}-1)p_k(z))+\theta_{T+(k-1)\delta}) \\
& \left.+o(s)\right\} \qquad (38)
\end{aligned}$$

where the inequality follows by Chernoff's inequality, 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 5.

Now 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. We divide into two cases: For $T+(k-1)\delta \leq \rho_s$, we put $z=0$ and so by (37) and we have $p_k(0) \searrow 0$ as $s \nearrow \infty$ (recall $\delta = O(1/s)$). Consequently (38) becomes

$$\exp\{-\theta_{T+(k-1)\delta}(s-r(T+(k-1)\delta))+s\psi_{T+(k-1)\delta}(\log(1+(e^{l\theta_{T+k\delta}}-1)p_k(z))+\theta_{T+(k-1)\delta})+o(s)\} = e^{-sI_{T+(k-1)\delta}+o(s)}$$

For $T+(k-1)\delta > \rho_s$, we put $z = T+(k-1)\delta-\rho_s$ so that $T+(k-1)\delta-z = \rho_s$. Hence again $p_k(z) \searrow 0$. Also,

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

for large enough $T+(k-1)\delta-z=\rho_s$ and some constants $C_1, C_2>0$, due to the fact that $\log(1+x)\le x$ for $x>0$ and that $\psi_N'(0)=\lambda$. It is now obvious that (38) also becomes $e^{-sI_{T+(k-1)\delta}+o(s)}$ in this case.

Hence (32) is less than or equal to

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

From this, and using Lemma 1, we get

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

Since l is arbitrarily close to 1, we have proved (31).

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 5.

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

$$\begin{aligned} & E_r[e^{l\theta_{T+k\delta}H_k};G_k>s-r(T+(k-1)\delta)] \\ \le\ & E_r[e^{l\theta_{T+k\delta}H_k};\tilde{G}_k>s-r(T+(k-1)\delta)] \\ =\ & E_r[E_r[e^{l\theta_{T+k\delta}H_k}|\tilde{A}_i^k,i=1,\dots,\tilde{G}_k];\tilde{G}_k>s-r(T+(k-1)\delta)] \\ =\ & E_r\left[\prod_{i=1}^{\tilde{G}_k}(1+(e^{l\theta_{T+k\delta}})\tilde{p}_i^k);\tilde{G}_k>s-r(T+(k-1)\delta)\right] \end{aligned} \tag{39}$$

where

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

Hence (39) is less than or equal to

$$\begin{aligned} & E_r[e^{\log(1+(e^{l\theta_{T+k\delta}})\tilde{p}_k)\tilde{G}_k};\tilde{G}_k>s-r(T+(k-1)\delta)] \\ \le\ & e^{-\theta_{T+(k-1)\delta}(s-r(T+(k-1)\delta))+\tilde{\psi}_k(\log(1+(e^{l\theta_{T+k\delta}}-1)\tilde{p})+\theta_{T+(k-1)\delta})} \end{aligned} \tag{40}$$

where $\tilde{\psi}_k(\theta):=\log Ee^{\theta\tilde{G}_k}$, by Chernoff's inequality. Now note that by Lemma 5 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 (40) is less than or equal to

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

Thus (32) is less than or equal to

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

This gives

$$\limsup_{s\to\infty} \frac{1}{s} \log E_r \left[N_A^2 {\tau_s}^3 \exp\left\{\theta_{\lceil \tau_s \rceil} \left(s a_{\lceil \tau_s \rceil} + 1 - \bar{Q}^{\infty}(\tau_s, \lceil \tau_s \rceil - \tau_s)\right)\right\}; \tau_s > T; \tau_s < \tau_A \right] \leq -\frac{I^*}{l} + \frac{aC}{l}$$

Since l and a can be chosen arbitrarily close to 1 and 0 respectively, (31) holds and conclusion follows.

□

Remark 6. *The proof can be simplified in the case of $M/G/s$ system. In particular, there is no need to condition on A_i^k nor introduce the constant a in the case of bounded support V. 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*

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

for some constant $C > 0$. Then (33) immediately becomes

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

The rest follows similarly as in the proof.

Remark 7. *Note that the result coincides with Erlang's loss formula in the case of $M/G/s$ (see for example Asmussen (2003)), which states that the loss probability is exactly given by*

$$P_\pi(loss) = \frac{(\lambda s E V)^s / s!}{1 + \lambda s E V + \cdots + (\lambda s E V)^s / s!}$$

Simple calculation reveals that $(1/s) \log P_\pi(loss) \to \log(\lambda E V) + 1 - \lambda E V = -I^*$.

The next result we will discuss is the lower bound:

Theorem 4. *For any $r(\cdot) \in J(\cdot)$, we have*

$$\liminf_{s\to\infty} \frac{1}{s} \log P_r(\tau_s < \tau_A) \geq -I^*$$

It suffices to prove that $\liminf_{s\to\infty}(1/s)\log P_r(\tau_s < \tau_A) \geq -I_{t_n}$ for a sequence $t_n \nearrow \infty$ thanks to Lemma 3 Part 1 and 2. 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 Part 3. For each $n\Delta$, the idea then is to identify a so-called optimal sample path (or more precisely a neighborhood of such path) that possesses a rate function $I_{n\Delta}$ and has the property $\tau_s < \tau_A$. Note that the probability in consideration is the same for $GI/G/s$ and $GI/G/\infty$ systems. Henceforth we would consider paths in $GI/G/\infty$.

The way we define A in (11) implies that it suffices to focus on the process on the time-grid $\{0, \Delta, 2\Delta, \ldots\}$ for checking the condition $\tau_s < \tau_A$. For a path to reach s at time $n\Delta$, the form of $\psi'_{n\Delta}(\theta_{n\Delta})$ hints that $E[\bar{Q}^{\infty}_{(k-1)\Delta,k\Delta}[(j-1)\Delta, j\Delta]|Q^{\infty}(n\Delta) > s] = s\alpha_{kj} + o(s)$ and $E[\bar{Q}^{\infty}_{(k-1)\Delta,k\Delta}[n\Delta,\infty]|Q^{\infty}(n\Delta) > s] = s\beta_k + o(s)$ 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 $k = 1, \ldots, n$, $j = k, \ldots, n$. 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 (1995), whose proof resembles this earlier work and is deferred to the appendix. The second one argues that the path we identified indeed satisfies $\tau_s < \tau_A$:

Lemma 6. *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*

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

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 7. *Starting with any $r(\cdot) \in J(\cdot)$, 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 $\tau_s < \tau_A$. Here $\gamma_{kj}, \gamma_k > 0$, $\sum_{\substack{k=1,\ldots,n \\ j=k,\ldots,n}} \gamma_{kj} + \sum_{k=1,\ldots,n} \gamma_k = \gamma < \infty$ 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 τ_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:

Proof of Theorem 4. Note that by Lemma 7, for any $r(\cdot) \in J(\cdot)$ and s large enough,

$$\begin{aligned} & P_r(\tau_s < \tau_A) \\ \geq\ & P_r(Q^{\infty}_{(k-1)\Delta,k\Delta}[(j-1)\Delta, j\Delta] \in ((\alpha_{kj}+\gamma_{kj})s, (\alpha_{kj}+\epsilon)s),\ Q^{\infty}_{(k-1)\Delta,k\Delta}[n\Delta,\infty] \in ((\beta_k+\gamma_k)s, (\beta_k+\epsilon)s), \\ & k=1,\ldots,n,\ j=k,\ldots,n) \end{aligned} \tag{41}$$

for large enough s given arbitrary γ_{kj}, γ_k and ϵ satisfying conditions in Lemma 7. Denote $\mathbf{\Gamma} = (\gamma_{kj}, \gamma_k)_{k=1,\ldots,n,\ j=k,\ldots,n}$. Let

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

Using Gartner-Ellis Theorem for (41) and Lemma 6, we have

$$\begin{aligned} & \frac{1}{s}\log P_r(Q^{\infty}_{(k-1)\Delta,k\Delta}[(j-1)\Delta, j\Delta] \in ((\alpha_{kj}+\gamma_{kj})s, (\alpha_{kj}+\epsilon)s), \\ & Q^{\infty}_{(k-1)\Delta,k\Delta}[n\Delta,\infty] \in ((\beta_k+\gamma_k)s, (\beta_k+\epsilon)s),\ k=1,\ldots,n,\ j=k,\ldots,n) \\ \to\ & -I_{\mathbf{\Gamma}} \end{aligned} \tag{42}$$

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

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

with $\bar{\psi}(\mathbf{\Theta})$ defined in Lemma 6. But note that for $k=1,\ldots,n$, $j=k,\ldots,n$,

$$\begin{aligned} \frac{\partial}{\partial\theta_{kj}}(\langle \mathbf{\Theta}, \mathbf{x}\rangle - \bar{\psi}(\mathbf{\Theta})) &= x_{kj} - \int_{(k-1)\Delta}^{k\Delta} \psi_N'\left(\log\left(\sum_{j=k}^{n} e^{\theta_{kj}} P((j-1)\Delta - u < V \leq j\Delta - u) + e^{\theta_{k\cdot}}\bar{F}(n\Delta-u)\right)\right) \\ & \qquad \frac{e^{\theta_{kj}} P((j-1)\Delta - u < V \leq j\Delta - u)}{\sum_{j=k}^{m} e^{\theta_{kj}} P((j-1)\Delta - u < V \leq j\Delta - u) + e^{\theta_{k\cdot}}\bar{F}(n\Delta-u)} du \end{aligned} \tag{43}$$

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

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 (43) and (44). Since $\langle \Theta, \mathbf{x}\rangle - \bar{\psi}(\Theta)$ is concave in $\mathbf{\Theta}$, 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

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

as $\mathbf{\Gamma} \to 0$. Together with (41) and (42) gives the conclusion.

□

Theorems 3 and 4 together imply both the asymptotic optimality of Algorithm 2 and the large deviations of the loss probability:

Proof of Theorem 1 and 2. Note that by Jensen's inequality

$$P_r(\tau_s < \tau_A)^2 \leq (E_r N_A)^2 \leq \tilde{E}_r[N_A^2 L^2]$$

Hence using Theorems 3 and 4 yields

$$-2I^* \leq \lim_{s\to\infty} \frac{1}{s} \log P_r(\tau_s < \tau_A)^2 \leq \lim_{s\to\infty} \frac{1}{s} \log (E_r N_A)^2 \leq \lim_{s\to\infty} \frac{1}{s} \log \tilde{E}_r[N_A^2 L^2] \leq -2I^*$$

Combining Proposition 1, we conclude that the steady-state loss probability given by (7) decays exponentially with rate I^* and that Algorithm 2 is asymptotically optimal. □

4 Logarithmic Estimate of Regenerative Time

In this section we will lay out the argument for Proposition 1. The first step is to reduce the problem to a $GI/G/\infty$ calculation. Define $x(t) := \sup\{y : Q^\infty(t, y) > 0\}$ as the maximum residual service times among all customers present at time t.

Lemma 8. *We have $\tau_A \leq \tau'_A$ where*

$$\tau'_A = \inf\{t \in \{\Delta, 2\Delta, \ldots\} : x(t-u) \leq l,\ Q^\infty(w) < s \text{ for } w \in [t-u, t] \text{ for some } u > l,\ Q^\infty(t, \cdot) \in J(\cdot)\}$$

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 $x(t-u) \leq l$, 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 τ_A' is small. We use a geometric trial argument. Divide the time frame into blocks separated at $T_0 = 0, T_1, T_2, \ldots$ in such a way that (1) a "success" in the block would mean τ_A' is reached before the end of the block (2) $\{W_u, T_i < u \leq T_{i+1}\}, i = 0, 1, \ldots$ are roughly independent. We then estimate the probability of "success" in a block and also the length of a block to obtain a bound for τ_A'.

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

Lemma 9. *For any fixed $t_0 > 0$.*

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

and

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

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 (14).

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 Borell-TIS inequality (Adler (1990)). By Pang and Whitt (2009) 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 $D_{D[0,\infty)}[0,\infty)$, where

$$R(t,y) = R_1(t,y) + R_2(t,y) \tag{45}$$

is a two-dimensional Gaussian field given by

$$R_1(t,y) = \lambda \int_0^t \int_0^{\infty} I(u+x > t+y) dK(u,x) \tag{46}$$

and

$$R_2(t,y) = \lambda c_a^2 \int_0^t \bar{F}(t+y-u) dW(u) \tag{47}$$

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.

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

Lemma 10. *Fix $t_0 > 0$. For $i = 1, 2$, we have*

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

for well-chosen constant $C_* > 0$, where $R(\cdot,\cdot)$ and $\nu(\cdot)$ are defined in (45), (46), (47) *and* (14).

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

Step 1: Define a d-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 (1990)) 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 Borell-TIS inequality i.e. for $x \geq E\sup_S \tilde{R}_i(t,y)$,

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

where

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

From these steps, it is straightforward to conclude Lemma 10. The rest of the proof of Lemma 9 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 $U_0 \leq x$ for $x = O(1/s)$ and using the light tail property of U_0. Again, the derivation is tedious but straightforward, and hence is skipped.

We need one more lemma:

Lemma 11. *Let V_k be r.v. with distribution function $F(\cdot)$ satisfying the light-tail assumption in (4). For any $p>0$, we have*

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

where

$$l_p(n) = \inf\{y : np\int_y^\infty u^{p-1}\bar{F}(u)du < \eta\} \tag{48}$$

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

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

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

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

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

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

We are now ready to prove Proposition 1, which we need the following construction. Pick $\gamma = 1/t_0$ where γ is introduced in (14) and $\xi(y)$ is defined in (13). Recall C_1 as in Lemma 9. Define $T_i, i = 0, 1, 2, \ldots$ as follows: Given T_{i-1}, define

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

Note that w_i and z are multiples of t_0. For convenience define, for $u < t$, $\bar{Q}^{\infty}_u(t, y) := \bar{Q}^{\infty}(u+t, y) - \bar{Q}^{\infty}(u, t+y)$ 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 occurs: 1) $\bar{Q}^{\infty}_{T_{i-1}+(k-1)t_0}(t, y) \in \left(\lambda s \int_y^{t+y} \bar{F}(u)du \pm \sqrt{s}C_1\nu(y)\right)$ 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) $Q^{\infty}_{T_{i-1}+w_i+d_i+(k-1)t_0}(t, y) \in \left(\lambda s \int_y^{t+y} \bar{F}(u)du \pm \sqrt{s}C_1\nu(y)\right)$ 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 (see the discussion preceding Proposition 1). 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 τ'_A is reached during the last Δ units of time before T_i.

Proof of Proposition 1. We first check that the occurrence of event ζ_i implies that τ'_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 arrive after time T_{i-1}. Hence the occurrence of ζ_i implies that

$$\begin{aligned}
& Q^{\infty}(T_{i-1} + w_i, y) \\
\in & \left(\lambda s \sum_{k=1}^{w_i/t_0} \int_{(k-1)t_0+y}^{kt_0+y} \bar{F}(u)du \pm \sqrt{s}C_1 \sum_{k=1}^{w_i/t_0} \nu((k-1)t_0 + y)\right) \\
\subset & \left(\lambda s \int_y^{w_i+y} \bar{F}(u)du \pm \sqrt{s}C_1\left[\nu(y) + \frac{1}{t_0}\int_y^{\infty} \nu(u)du\right]\right) \\
\subset & \left(\lambda s \int_y^{w_i+y} \bar{F}(u)du \pm \sqrt{s}C_1\xi(y)\right) \qquad (49)
\end{aligned}$$

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

$$\begin{aligned}
& Q^\infty(T_{i-1}+w_i+d_i+t, y) \\
\in\ & \left(\lambda s \int_y^{t+y} \bar{F}(u)du + \lambda s \int_{d_i+y+t}^{w_i+d_i+y+t} \bar{F}(u)du \right. \\
& \left. \pm \sqrt{s} C_1 \left[\sum_{j=1}^{w_i/t_0} \nu((j-1)t_0 + d_i + (k-1)t_0 + [t] + y) + \nu(y) + \sum_{j=2}^{k} \nu((j-2)t_0 + [t] + y) I(k>1) \right] \right) \\
\subset\ & \left(\lambda s \int_y^{t+y} \bar{F}(u)du + \lambda s \int_{d_i+y+t}^{w_i+d_i+y+t} \bar{F}(u)du \pm \sqrt{s} C_1 \left[\sum_{j=1}^{w_i/t_0+k-1} \nu((j-1)t_0 + [t] + y) + \nu(y) \right] \right) \\
\subset\ & \left(\lambda s \int_y^{t+y} \bar{F}(u)du + \lambda s \int_{d_i+y+t}^{w_i+d_i+y+t} \bar{F}(u)du \pm \sqrt{s} C_1 \left[2\nu(y) + \frac{1}{t_0} \int_y^\infty \nu(u)du \right] \right) \\
\subset\ & \left(\lambda s \int_y^{t+y} \bar{F}(u)du + \lambda s \int_{d_i+y+t}^{w_i+d_i+y+t} \bar{F}(u)du \pm \sqrt{s} C' \xi(y) \right) \qquad (50)
\end{aligned}$$

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 $v(s)$, (49) and the fact that $\lambda s \int_y^\infty \bar{F}(u)du$ is smaller and decays faster than $\sqrt{s}C_1\xi(y)$ for $y \geq v(s)$ when s is large, we get $x(T_{i-1}+w_i) \leq v(s) \leq z$. Let $\tilde{T}_i = \sup\{k\Delta : k\Delta \leq T_i\}$ be the largest time before T_i such that A can possibly be hit i.e. in the Δ-skeleton. It remains to show that $Q^\infty(\tilde{T}_i, y) \in J(y)$ in order to conclude that ζ_i implies a hit on τ'_A.

From (50), 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,

$$\begin{aligned}
Q^\infty(\tilde{T}_i, y) &\in \left(\lambda s \int_y^{\tilde{T}_i-T_{i-1}+y} \bar{F}(u)du - \lambda s \int_{\tilde{T}_i-T_{i-1}-w_i-d_i+y}^{\tilde{T}_i-T_{i-1}-w_i+y} \bar{F}(u)du \pm \sqrt{s} C' \xi(y) \right) \\
&= \left(\lambda s \int_y^\infty \bar{F}(u)du - \lambda s \int_{\tilde{T}_i-T_{i-1}+y}^\infty \bar{F}(u)du - \lambda s \int_{\tilde{T}_i-T_{i-1}-w_i-d_i+y}^{\tilde{T}_i-T_{i-1}-w_i+y} \bar{F}(u)du \pm \sqrt{s} C' \xi(y) \right) \qquad (51)
\end{aligned}$$

Now note that

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

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 $v(s) \geq \inf\{y : s\int_y^\infty \bar{F}(u) \leq \alpha\}$ for any $\alpha > 0$ when s is large enough. Now when $\sqrt{s}C\xi(y) < \alpha/(2\lambda)$, $s\int_{v(s)+y}^\infty \bar{F}(u)du \leq s\int_y^\infty \bar{F}(u)du$ which is smaller and decays faster than $\sqrt{s}C\xi(y)$ when s is large. When $\sqrt{s}C\xi(y) \geq \alpha/(2\lambda)$, 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 (12), we conclude that ζ_i implies τ_A' is reached at $\tilde{T}_i$.

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

$$\begin{aligned} & E(\tau_A')^p \\ = \; & E\left[\sum_{i=1}^{N}(w_i + d_i + z)\right]^p \\ = \; & E\left[\sum_{i=1}^{\infty}(w_i + d_i + z)I(N \ge i)\right]^p \\ \le \; & \left(\sum_{i=1}^{\infty}(E[(w_i + d_i + z)^p; N \ge i])^{1/p}\right)^p \\ \le \; & \left(\sum_{i=1}^{\infty}(E(w_i + d_i + z)^{pq})^{1/(pq)}(P(N \ge i))^{1/(pr)}\right)^p \end{aligned} \tag{52}$$

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

$$E(w_i + d_i + z)^{pq} \le [(Ew_i^{pq})^{1/(pq)} + (Ed_i^{pq})^{1/(pq)} + z]^{pq} \tag{53}$$

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

$$(Ed_i^p)^{1/p} \le d^{(p)} := \sup_{b\ge 0}(E[d_i^p | B(T_{i-1} + w_i) = b])^{1/p} = \frac{1}{s}\sup_{b\ge 0}(E[(U^0 - b)^p | B^0(0) = b])^{1/p} = O\left(\frac{1}{s}\right) \tag{54}$$

and $z \le v(s) + \Delta + t_0 = o(s^\epsilon)$ for any $\epsilon > 0$. The last equality of (54) 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 complementary distribution function of U^0 respectively. This implies that $h(x) \ge c$ for all $x \ge 0$. Then

$$\sup_{b\ge 0} P(U^0 - b > x | U^0 > b) = \sup_{b\ge 0}\exp\left\{-\int_b^{x+b} h(u)du\right\} \le e^{-cx}$$

and so

$$\sup_{b\ge 0} E[(U^0 - b)^p | B^0(0) = b] = \sup_{b\ge 0} p\int_0^\infty x^{p-1}P(U^0 - b > x | U^0 > b)dx \le p\int_0^\infty x^{p-1}e^{-cx}dx < \infty$$

For $i = 1$, $w_1 \le l(s) + t_0 = o(s^\epsilon)$ where $l(s)$ is defined in (15). Hence $E(w_1 + d_1 + z)^p \le [(Ew_1^p)^{1/p} + (Ed_1^p)^{1/p} + z]^p = o(s^\epsilon)$ for any $\epsilon > 0$.

Now

$$\begin{aligned} Ew_i^p &\le E\left[\left(\max_{i=1,\ldots,N_s(T_{i-1})-N_s(T_{i-2})} V_i\right)^p\right] \\ &= E\left[E\left[\left(\max_{i=1,\ldots,N_s(T_{i-1})-N_s(T_{i-2})} V_i\right)^p \middle| N_s(T_{i-1}) - N_s(T_{i-2})\right]\right] \\ &\le CE[l_p(N_s(T_{i-1}) - N_s(T_{i-2}))^p] \text{ for some constant } C = C(p) \text{ and } l_p(\cdot) \text{ defined in (48)} \\ &\le CE[(N_s(T_{i-1}) - N_s(T_{i-2}))^\epsilon] \text{ for constant } C = C(p,\epsilon) \end{aligned} \tag{55}$$

for any $\epsilon > 0$, by Lemma 11. Pick $\epsilon < 1$. By Jensen's inequality and elementary renewal theorem, (55) is less than or equal to

$$\begin{aligned} & C(E[N_s(T_{i-1}) - N_s(T_{i-2})])^\epsilon \\ = \;& C(E[N_s(T_{i-1}) - N_s(T_{i-2})|T_{i-1} - T_{i-2}])^\epsilon \\ \le \;& C(E[\tilde{\lambda} s(T_{i-1} - T_{i-2})])^\epsilon \text{ for some } \tilde{\lambda} > \lambda \\ = \;& C\tilde{\lambda}^\epsilon s^\epsilon (E[T_{i-1} - T_{i-2}])^\epsilon \\ = \;& C\tilde{\lambda}^\epsilon s^\epsilon (E[w_{i-1} + d_{i-1} + z])^\epsilon \end{aligned} \tag{56}$$

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

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

for large enough s, uniformly over i. Hence

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

Now we can write

$$\begin{aligned} y_i &\le Cs^\epsilon y_{i-1}^\epsilon \le Cs^\epsilon (Cs^\epsilon y_{i-2}^\epsilon)^\epsilon = C^{1+\epsilon} s^{\epsilon+\epsilon^2} y_{i-2}^{\epsilon^2} \\ &\cdots \le (C^{1/(1-\epsilon)} \vee 1) s^{\epsilon/(1-\epsilon)} y_1^{\epsilon^{i-1}} = o(s^\rho) \end{aligned} \tag{57}$$

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

Therefore from (53), (56) and (57), we get

$$E(w_i + d_i + z)^{pq} = o(s^\epsilon) \tag{58}$$

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

Now consider

$$\begin{aligned} P(N \ge 1) &= P(\zeta_1^c) = 1 - P(\zeta_1) \\ &\le 1 - P\left(d_1 \le \frac{c}{s}\right) C_2^{(w_1+z)/t_0} \\ &\quad \text{where } C_2 \text{ is defined in Lemma 9 and } c \text{ is defined in the discussion of } \zeta_i \\ &\le 1 - be^{-a(w_1+z)} \\ &= 1 - be^{-o(s^\epsilon)} \end{aligned} \tag{59}$$

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

$$\begin{aligned} P(N \geq i) &= P(N \geq i-1)P(\zeta_{i-1}^c | N \geq i-1) \\ &\leq P(N \geq i-1)E[1 - be^{-a(w_{i-1}+z)} | N \geq i-1] \\ &\leq P(N \geq i-1)(1 - be^{-a(E[w_{i-1}|N \geq i-1]+z)}) \end{aligned} \tag{60}$$

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

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

$$E[w_i | N \geq i] = E[E[w_i | \zeta_{i-1}^c,\ w_{i-1} + d_{i-1} + z] | N \geq i] \tag{61}$$

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

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

where C_3 is defined in Lemma 9, uniformly over $w_{i-1} + d_{i-1} + z$. Hence

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

which gives

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

uniformly over $w_{i-1} + d_{i-1} + z$. Therefore (61) is bounded from above by Ew_i/C_3.

From (56) and (57) we know that $Ew_i = o(s^\epsilon)$ for any $\epsilon > 0$. So (60) is less than or equal to

$$P(N \geq i-1)(1 - be^{-a(Ew_{i-1}/C_3+z)}) = P(N \geq i-1)(1 - be^{-o(s^\epsilon)}) \tag{62}$$

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

By (52), (59), (58) and (62) we get

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

Hence

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

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

$$\frac{1}{s} \log E\tau_A^p \geq \frac{1}{s} \log \Delta^p \to 0$$

Conclusion follows for (9).

For (10), note that $N_A \leq N_s(\tau_A) \leq N_s(\tau'_A)$ and $EN_s(t)^p = O(st)$ since $(1/s)\log Ee^{\theta N_s(t)} \to -\psi_N(\theta)t$. Hence

$$EN_s(\tau'_A)^p \leq O(s^p)E(\tau'_A)^p$$

and the result follows from (9). □

Remark 8. *The proof of Proposition 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.*

5 Numerical Example

We close this paper by 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 $1/2$. Since service time is bounded, we simply set $c^* = 1$ and $C(y) = \mathrm{sd}(R(\infty, y)) \vee C_1 = \sqrt{\lambda \int_y^\infty F(u)\bar{F}(u)du + \lambda c_a^2 \int_y^\infty \bar{F}(u)^2 du} \vee C_1$ with $C_1 = 1.1$ (see the discussion following Lemma 1). Also we choose $\Delta = 1$. 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 Section 2, since we run our importance sampler everytime we hit set 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 (2007)). 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 the tables below, we output the estimates of loss probability, the empirical 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 over 80, importance sampler starts to perform better. When s is above 100, crude Monte Carlo totally breaks down while our importance sampler still gives estimates that have encouragingly small relative error.

	Crude Monte Carlo			Importance Sampler		
s	Estimate	E.R.E.	C.I.	Estimate	E.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	$(-1.4237 \times 10^{-10}, 4.0286 \times 10^{-10})$

We can also analyze the graphical depiction of the sample paths. Figures 6 and 7 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 6 is

a contour plot of $Q(t,y)$, whereas Figure 7 is a three-dimensional plot of another $Q(t,y)$. As we can see, the number of customers (the color at 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.

6 Appendix

6.1 Proof of Lemma 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 (1976), 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 $u \in \text{supp } \bar{F}$ since $\psi_N(\cdot)$ is steep. By monotone convergence theorem we conclude that $\psi_t(\cdot)$ is steep.

6.2 Proof of Lemma 2

1) denote $\theta(t) = \theta_t$ for convenience. Since $\psi_t'(\cdot)$ is continuously differentiable by Lemma 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

$$\begin{aligned}
& \theta'(t) \\
= & \frac{\lambda\bar{F}(t) - \psi'_N(\log(e^{\theta(t)}\bar{F}(t)+F(t)))e^{\theta(t)}\bar{F}(t)/(e^{\theta(t)}\bar{F}(t)+F(t))}{\int_0^t \left[\psi''_N(\log(e^{\theta(t)}\bar{F}(u)+F(u)))\left(\frac{e^{\theta(t)}\bar{F}(u)}{e^{\theta(t)}\bar{F}(u)+F(u)}\right)^2 + \psi'_N(\log(e^{\theta(t)}\bar{F}(u)+F(u)))\frac{F(u)\bar{F}(u)e^{\theta(t)}}{(e^{\theta(t)}\bar{F}(u)+F(u))^2}\right]du} \\
\leq & 0
\end{aligned}$$

The inequality is due to the fact that

$$g_t(\theta) := \psi'_N(\log(e^{\theta}\bar{F}(t)+F(t)))\frac{e^{\theta}\bar{F}(t)}{e^{\theta}\bar{F}(t)+F(t)} \tag{63}$$

is non-decreasing in θ and $g_t(0) = \lambda\bar{F}(t)$, and that $\psi_N(\cdot)$ is non-decreasing and convex. Hence $\theta(t)$ is non-increasing.

2) Since $a_t \geq 1 - \lambda EV$, $\theta_t \geq \bar{\theta}_t$ 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 $\psi'_t(\theta) \leq \psi'_N(\theta)t$ for any $\theta, t > 0$ (either by the formula of ψ'_t and ψ'_N or by definition in terms of Gartner-Ellis limit). This implies that $(\psi'_t)^{-1}(y) \geq (\psi'_N)^{-1}(y/t)$ for any y in the domain. Putting $y = 1 - \lambda EV$ gives $\bar{\theta}_t \geq (\psi'_N)^{-1}((1-\lambda EV)/t)$. 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 $t \searrow 0$.

3) Consider $\psi'_t(\theta_t) = a_t$, or $\theta_t = (\psi'_t)^{-1}(a_t)$. Now from (21) 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 (1979), p. 287, or Resnick (2008), 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 (2008), p. 2, this implies continuous convergence, and hence $(\psi'_t)^{-1}(a_t) \to (\psi'_\infty)^{-1}(1)$, or $\theta_t \to \theta_\infty$.

6.3 Proof of Lemma 3

1) As in the proof of Lemma 2 Part 1, 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 non-increasing.

2) Write $I_t = a_t\theta_t - \psi_t(\theta_t)$. By Lemma 2 Part 3, $\theta_t \searrow \theta_\infty$ 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 (2008), p. 2) we have $\psi_t(\theta_t) \to \psi_\infty(\theta_\infty)$. Hence $I_t \to I^*$ defined in (22).

3) 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.

6.4 Proof of Lemma 4

1) Following the spirit of the proof of Lemma 3 Part 1, 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.

2) 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 1).

6.5 Proof of Lemma 5

To prove Lemma 5, we first need the following analytical lemma:

Lemma 12. *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 non-decreasing or non-increasing in y_i fixing $x_1, \ldots, x_{i-1}, x_i, \ldots, x_$ 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 $\{N_\delta(\mathbf{x})\}_{x\in\mathcal{E}}$ be such collection. Note that $h_m \to h$ uniformly over $\mathcal{E}$.

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 $\mathcal{E}$ that satisfies: For $i = 1, \ldots, n$, $\tilde{x}_i \geq x_i$ if h is non-decreasing in the i-th component, and $\tilde{x}_i \leq x_i$ if h is non-increasing 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

$$\begin{aligned}
& |h_m(\mathbf{x}) - h_m(\tilde{\mathbf{x}})| \\
= \; & h_m(\tilde{\mathbf{x}}) - h_m(\mathbf{x}) \quad \text{by our choice of } \tilde{\mathbf{x}} \text{ and monotone property of } h_m \\
\leq \; & h_m(\tilde{\mathbf{x}}) - h_m(\tilde{\tilde{\mathbf{x}}}) \quad \text{where } \tilde{\tilde{\mathbf{x}}} \text{ is chosen to be the closet point to } \mathbf{x} \text{ in } \mathcal{E} \text{ that satisfies:} \\
& \qquad \text{For } i = 1, \ldots, n,\ \tilde{\tilde{x}}_i \leq x_i \text{ if } h \text{ is non-decreasing in the } i\text{-th component, and} \\
& \qquad \tilde{\tilde{x}}_i \geq x_i \text{ if } h \text{ is non-increasing in the } i\text{-th component.} \\
\leq \; & |h_m(\tilde{\mathbf{x}}) - h(\tilde{\mathbf{x}})| + |h(\tilde{\mathbf{x}}) - h(\tilde{\tilde{\mathbf{x}}})| + |h_m(\tilde{\tilde{\mathbf{x}}}) - h(\tilde{\tilde{\mathbf{x}}})| \\
\leq \; & \epsilon + 2\epsilon + \epsilon
\end{aligned}$$

when m is large enough.

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

□

Proof of Lemma 5. For convenience write $\psi_s(\theta; w, z, t) = \log E e^{\bar{Q}^{\infty}_{w,z}[t,\infty]}$ and

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

defined for $\theta \in [\theta_\infty, \theta_T]$, $t \geq T$ and $0 \leq w \leq z \leq t + \eta$ 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 non-decreasing in θ, non-increasing in w, non-decreasing in z and non-increasing 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 12, 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 $t > K$, $z \leq t - K$, we have

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

for some $C_1, C_2 > 0$, uniformly over $\theta \in [\theta_\infty, \theta_T]$. Hence for $z \leq t - K$, $\psi_s(\theta; w, z, t) \leq \psi_s(\theta; 0, t - K, t) \to \psi(\theta; 0, t - K, t) < \epsilon$ 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

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

and bounded from below by

$$\begin{aligned} & \frac{1}{s}\log\left(E e^{\theta \bar{Q}^{\infty}_{0,t-K}[t,\infty] I(w<t-K)} E_{00} e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]}\right) \\ = \;& \psi_s(\theta; w, t-K, t) I(w<t-K) + \frac{1}{s}\log E_{00} e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]} \end{aligned} \tag{64}$$

where $E_0[\cdot]$ denotes the expectation conditioned that a customer arrives at time 0 and is counted in $\bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[t,\infty]$, while $E_{00}[\cdot]$ denotes the expectation conditioned on delayed arrival with complementary distribution (in the basic scale) given by $\sup_b P(U^0 - b > x | U^0 - b)$. Note that $\sup_b P(U^0 - b > x | U^0 > b)$ is a valid complementary distribution because of the light-tail assumption on U^0. Indeed, it is obvious that $\sup_b P(U^0 - b > 0 | U^0 > b) = 1$, and by the same argument following that of (54), we have $\sup_b P(U^0 - b > x | U^0 > b) \le e^{-cx} \to 0$ for some $c > 0$. Moreover, it is obvious that $\sup_b P(U^0 - b > x | U^0 > b)$ is non-increasing. Now by construction this complementary distribution is stochastically at most as large as $P(U^0 - b > x | U^0 > b)$ for any $b \ge 0$, and hence (64). Note that $\frac{1}{s}\log E_0 e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]}$ and $\frac{1}{s}\log E_{00} e^{\theta \bar{Q}^{\infty}_{0,(z-t+K)\wedge(z-w)}[K,\infty]}$ 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.

□

6.6 Proof of Lemma 6

Consider

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

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

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

where $\underline{\zeta}_{kw} = \text{argmin}\{h_k(u) : (k-1)\Delta + (w-1)\Delta/m \le u \le (k-1)\Delta + w\Delta/m\}$ and $\overline{\zeta}_{kw} = \text{argmax}\{h_k(u) : (k-1)\Delta + (w-1)\Delta/m \le u \le (k-1)\Delta + w\Delta/m\}$. The existence of $\underline{\zeta}_{kw}$ and $\overline{\zeta}_{kw}$ 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

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

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.

References

1. Adler, R. (1990), *An Introduction to Continuity, Extrema, and Related Topics for General Gaussian Processes*. IMS Lecture Notes-Monograph Series Vol. 12.

2. Anantharam, V. (1988), How large delays build up in a $GI/GI/1$ queue. *Queueing Systems: Theory and Applications*, **5**, 345368.

3. Asmussen, S. (1985), Conjugate processes and the simulation of ruin problems. *Stochastic Processes and their Applications*, **20**, 213-229.

4. Asmussen, S. (2003), *Applied Probability and Queues*, 2nd Edition. Springer-Verlag, New York.

5. Asmussen, S, and Glynn, P. W. (2007), *Stochastic Simulation: Algorithms and Analysis*. Springer, New York.

6. Blanchet, J., Glynn, P., and Lam, H. (2009), Rare-event simulation for a slotted-time $M/G/s$ model. *Queueing Systems: Theory and Applications*, **63**, 33-57.

7. Blanchet, J. and Lam, H. (2010), Importance sampling for actuarial reserve analysis under a heavy traffic model. *Working paper.*

8. Breiman, L. (1968), *Probability.* Addison-Wesley, Massachusetts.

9. Bucklew, J. (2004), *Introduction to rare-event simulation.* Springer-Verlag, New York.

10. Dembo, A. and Zeitouni, O. (1998), *Large Deviations Techniques and Applications*, 2nd Edition. Springer-Verlag, New York.

11. Foley, R. D. (1982), The non-homogeneous $M/G/\infty$ queue, *OPSEARCH*, **19**(1), 40-48.

12. Glynn, P. W. (1995), Large deviations for the infinite server queue in heavy traffic. *IMA Vol. 71 in Mathematics and its Applications*, Springer-Verlag, 387-395.

13. Glynn, P. W. and Whitt, W. (1991), A new view of the heavy-traffic limit theorem for many-server queues. *Advances in Applied Probability*, **23**, 188-209.

14. Heidelberger, P. (1995), Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, **51**, 43-85.

15. Juneja, S. and Shahabuddin, P. (2006), *Rare Event Simulation Techniques: An Introduction and Recent Advances.* Handbook in Operations Research and Management Sciences, Vol. 13: Simulation. Chapter 11. Elsevier, Henderson, S. and Nelson, B. (Eds.), 291-350.

16. Pang, G. and Whitt, W. (2009), Two-Parameter Heavy-Traffic Limits for Infinite-Server Queues. *Under revision in Queueing Systems: Theory and Applications.*

17. Ridder, A. (2009), Importance sampling algorithms for first passage time probabilities in the infinite server queue. *European Journal of Operational Research*, **199**, 176-186.

18. Resnick, S. (2007), *Extreme Values, Regular Variation, and Point Processes*, Springer-Verlag.

19. Sadowsky, J. (1991), Large deviations and efficient simulation of excessive backlogs in a $GI/G/m$ queue. *IEEE Trans. Autom. Control*, **36**, 579-588.

20. Siegmund, D. (1976), Importance sampling in the Monte Carlo study of sequential tests. *Annals of Statistics*, **4**, 673-684.

21. Szechtman, R., and Glynn., P. W. (2002), Rare event simulation for infinite server queues. *Proceedings of the 2002 Winter Simulation Conference*, Yucesan, E., Chen, C. -H., Snowdon, J. L., and Charnes, J. M. (Eds.), 416-423.