Fluid Heuristics, Lyapunov Bounds and Efficient Importance Sampling for a Heavy-tailed G/G/1 Queue.

Blanchet, J., Glynn, P. and Liu, J. C.

March, 2007.

Abstract

We develop a strongly efficient rare-event simulation algorithm for computing the tail of the steady-state waiting time in a single server queue with regularly varying service times. Our algorithm is based on a state-dependent importance sampling strategy that is particularly simple to implement. The construction of the algorithm and its asymptotic optimality rely on a Lyapunov-type inequality that is used to bound the second moment of the estimator. The solution to the Lyapunov inequality is constructed using fluid heuristics. Our approach takes advantage of the regenerative ratio formula for the steady-state distribution – and does not use the first passage time representation that is special of the delay in the $G/G/1$ queue. Hence, the strategy has the potential to be applied in more general queueing models.

1 Introduction

Consider a positive recurrent single-server queue under first-in first-out (FIFO) queue discipline. We assume that the service times are heavy-tailed random variables, in particular, regularly varying. Our interest is in the development of an efficient rare-event simulation algorithm, based on the use of importance sampling, for computing tail probabilities associated with steady-state delays. Basically, it is standard in rare-event simulation to consider an algorithm efficient (or, more precisely, strongly efficient) if it produces an estimator that has bounded coefficient of variation – see, Definition 1 in Section 2 for a precise definition; for more on basic notions on rare-event simulation, see Bucklew (2004) and Juneja and Shahabuddin (2006). Efficient rare-event simulation algorithms for heavy-tailed $M/G/1$ queues have been developed during the last years; see, for instance, Asmussen and Binswanger (1997), Asmussen, Binswanger and Hojgaard (2000), Juneja and Shahabuddin (2002), Asmussen and Kroese (2006) and Dupuis, Leder and Wang (2006). All of these algorithms rely

on the Pollaczek-Khintchine representation, which is a special feature of the $M/G/1$ queue and allows one to reduce the problem to that of rare-event simulation for the tail of a finite sum of positive rv's. Consequently, these procedures are not applicable to the $G/G/1$ queue. Recently, Blanchet and Glynn (2007) proposed the first efficient rare-event simulation algorithm for a $G/G/1$ queue with heavy-tailed input (for a large class of sub-exponential distributions). The algorithm proposed by Blanchet and Glynn (2007) takes advantage of the equivalence between the distribution of the steady-state delay and the law of the maximum of a suitably defined random walk. In particular, it explicitly uses the representation of a steady-state tail probability for delay in terms of a level-crossing probability for the associated random walk.

This equivalent representation is unfortunately a feature that does not generalize beyond the $G/G/1$ queue (for instance, to multi-server queues). In addition, the variate generation suggested by the algorithm of Blanchet and Glynn (2007), although statistically efficient, can be challenging to implement. In this paper, we propose a new rare-event simulation algorithm based on state-dependent importance sampling in which the required variate generation is implemented via mixture sampling (a precise description of the mixture distribution is given in Sections 3 and 4, see equation (5)). Secondly, in contrast to Blanchet and Glynn (2007), the algorithm proposed here is based on the regenerative representation of the steady-state waiting time distribution. As a result, we expect the main ideas to be presented here will be applicable to more complex queueing systems. Indeed, the methodology that we propose here generalizes to the $G/G/2$ queue; see Blanchet, Glynn and Liu (2007).

As mentioned above, our algorithm takes advantage of the regenerative ratio formula for steady-state probabilities. The estimator for the numerator corresponds to the number of people who experience long delays within a busy period and the denominator is estimated as the sample mean of the number of customers served in a busy period (note in particular that the overall estimator will typically be biased). Our strategy is to develop a good importance sampling algorithm to estimate the probability of observing, within a busy cycle, at least one customer that experiences a long delay. We then show that this algorithm is actually strongly efficient for the steady-state waiting time itself.

In order to establish the efficiency of our algorithm, we use Lyapunov-type inequalities to upper bound the second moment of the proposed estimator. The use of Lyapunov-type inequalities to prove efficiency was introduced in Blanchet and Glynn (2007). However, here we use the Lyapunov bounds not only as a proof technique but also to construct our algorithm. Indeed, based on asymptotic approximations for the tail of the steady-state delay, Blanchet and Glynn (2007) propose a specific form of the importance sampling distribution that directly approximates the zero-variance change-of-measure. Our approach here is to instead propose a parametric family of importance samplers, based on mixtures. Then, we construct the solution to the Lyapunov inequality using "fluid heuristics" only – in other words, sharp asymptotic approximations are not needed – and derive sufficient conditions on the parameters

of our family in order to satisfy the Lyapunov bound. This Lyapunov function (i.e. the solution to the Lyapunov inequality) provides an upper bound on the second moment of the estimator. If one has access to a lower bound (which typically is easy to obtain by considering simple compound events involving the heavy-tailed rvs) for the probability of interest, a good choice of Lyapunov function permits one to establish bounded relative variance and therefore asymptotic optimality – in the sense of achieving the fastest possible rate of decay for the second moment of the proposed estimator.

The mixture sampling idea we use here is due to Dupuis, Leder and Wang (2006). They use this idea as a means of efficiently computing tail probabilities for heavy-tailed sums. To prove efficiency, they propose a verification procedure that is based on a weak convergence analysis (rather than the approach followed here).

The rest of this paper is organized as follows. In Section 2 we collect basic definitions related to computational efficiency in the context of rare-event simulation. Since, as indicated previously, our proposed estimator will typically be biased, we provide a brief discussion about efficiency of biased rare-event simulation estimators. Sections 3 further discusses the use of state-dependent importance sampling and describes an approach (based on Lyapunov functions) for verifying algorithmic efficiency. Sections 4 and 5 discuss in detail the design of our algorithm and the verification of its efficiency. In particular, Section 4 constructs the algorithm that is designed to be efficient for computing the probability of observing a long delay in a busy period. Then, in Section 5, we prove that such algorithm is also efficient for the steady-state waiting time. An implementation of our algorithm and its empirical performance is given in our last section, namely Section 6.

2 Computing Steady-State Rare Event Probabilities

Let $W = (W_n : n \geq 0)$ be an S-valued Harris recurrent Markov chain with stationary distribution π. For $w \in S$, let $P_w(\cdot)$ and $E_w(\cdot)$ be the probability distribution and expectation operator corresponding to W, conditional on $W_0 = w$. We are interested in computing $\pi(B_b)$, for a decreasing sequence of sets $\{B_b : b > 0\}$ such that $\pi(B_b) \longrightarrow 0$ as $b \nearrow \infty$.

Suppose that there exists a singleton $w_0 \in S$ with the property that W returns to w_0 infinitely often. Then, we have the following ratio formula for the steady-state distribution $\pi(\cdot)$:

$$\pi(B_b) = \frac{E_{w_0}\left(\sum_{j=0}^{T_{w_0}-1} I(W_j \in B_b)\right)}{E_{w_0}(T_{w_0})}, \tag{1}$$

where $T_{w_0} = \inf\{n \geq 1 : W_n = w_0\}$ (see, for example, Asmussen (2003)). Since the denominator in (1) does not depend on b, the rare-event type computation is needed

only for the numerator of the regenerative ratio formula. If we define $T_b = \inf\{n \geq 1 : W_n \in B_b\}$, the event $\{T_b < T_{w_0}\}$ is a rare event for large values of b. This observation suggests that we estimate $\pi(B_b)$ by developing a good importance sampling algorithm for the event $\{T_b < T_{w_0}\}$. Such an algorithm should then efficiently estimate the numerator in (1). One should keep in mind, however, that given $\{T_b < T_{w_0}\}$ the number of visits to B_b prior to returning to w_0 could be large and difficult to control. We shall come back to this issue in Section 4, when we discuss the assumptions imposed for the $G/G/1$ model that we consider here. The denominator, on the other hand, can be estimated using crude Monte Carlo. Since the complexity to evaluate $E_{w_0}(T_{w_0})$ to a given relative precision is insensitive to b, we expect (and verify in a moment) that the overall efficiency of an algorithm for estimating $\pi(B_b)$ for b large using (1) should depend mainly on the quality of the numerator's estimator.

Let us recall the definition of efficiency in the context of rare-event simulation.

Definition 1 An estimator Z_b is said to be *efficient* (or *strongly efficient*) for estimating $z_b \in (0, \infty)$ if

$$\sup_{b>0} \frac{E(Z_b - z_b)^2}{z_b^2} < \infty.$$

In most situations in rare-event simulation, one focuses on estimators that are unbiased (i.e. $EZ_b = z_b$, see, for instance, Juneja and Shahabuddin (2006)). Note that the previous definition does not require Z_b to be unbiased. Efficiency means that the number of replications required to estimate z_b within a prescribed relative accuracy is insensitive to b.

The strategy that we pursue is the following. We will find a good importance sampler, say $\widetilde{P}$, for the event $\{T_b < T_{w_0}\}$ to compute the numerator of (1), and crude Monte Carlo to calculate its denominator.

We denote the likelihood ratio associated with simulating W up to $T_b \wedge T_{w_0}$ by the rv L_b. We then generate n iid copies of

$$\sum_{j=0}^{T_{w_0}-1} I(W_j \in B_b) L_b$$

starting from w_0 under the importance sampling distribution $\widetilde{P}$ with modified dynamics up to $T \wedge T_{w_0}$, followed by use of the nominal (original) dynamics (from $T_b \wedge T_{w_0} + 1$ to T_{w_0}) to estimate the numerator of (1). We then run another (independent) set of n independent iid simulations of the rv T_{w_0}, starting from w_0, under W's nominal dynamics to estimate the denominator. Let $\overline{C}_n^{(b)}$ and $\overline{D}_n$ be the sample means for the numerator and denominator respectively (note that $\overline{D}_n$ is independent of b) and

observe that since $\overline{D}_n \geq 1$

$$\begin{aligned} E\left(\frac{\overline{C}_n^{(b)}}{\overline{D}_n} - \pi\left(B_b\right)\right)^2 &\leq E\left(\overline{C}_n^{(b)} - \pi\left(B_b\right)\overline{D}_n\right)^2 \\ &= Var\left(\overline{C}_n^{(b)} - \pi\left(B_b\right)\overline{D}_n\right) \\ &= Var\left(\overline{C}_n^{(b)}\right) + \pi\left(B_b\right)^2 Var\left(\overline{D}_n\right). \end{aligned} \tag{2}$$

The previous equation clearly indicates that if we are able to construct an efficient estimator for the numerator in (1) (in the traditional sense of unbiased estimators), then the estimator is automatically efficient in the sense indicated in Definition 1 and the relative (mean squared) accuracy within which the ratio estimator evaluates $\pi\left(B_b\right)$ is insensitive to b. As a consequence, we conclude that developing efficient rare-event simulation for $\pi\left(B_b\right)$ using the ratio formula (1) boils down to developing an efficient rare-event simulation algorithm for the numerator in the regenerative ratio formula.

3 State-dependent Importance Sampling

To compute the rare-event probability $u_b^*\left(w\right) = P_w\left(T_b < T_{w_0}\right)$, we note that $\left(u_b^*\left(w\right) : w \in S\right)$ solves the equation

$$u_b^*\left(w\right) = E_w u_b^*\left(W_1\right) \triangleq \int_S P\left(w, dy\right) u_b^*\left(y\right) \tag{3}$$

subject to the boundary condition $u_b^*\left(w\right) = 1$ for $w \in B_b$ and $u_b^*\left(w\right) = 0$ for $w \in w_0$. The conditional distribution of W given the event $\{T_b < T_{w_0}\}$, is that W's conditional dynamics form a Markov chain with modified transition kernel

$$R_b\left(w, dy\right) = P\left(w, dy\right) u_b^*\left(y\right) / u_b^*\left(w\right)$$

for $w, y \notin w_0$. Given that u_b^* is unknown, one possible means to developing a rare-event simulation algorithm is to substitute an approximation v_b for u_b^*. Since v_b is not an exact solution to (3), the normalization constant

$$\varpi_b\left(w\right) = \int_S P\left(w, dy\right) v_b\left(y\right)$$

does not equal $v_b\left(w\right)$, and the approximating transition kernel $\widetilde{R}_b$ takes the form

$$\widetilde{R}_b\left(w, dy\right) = P\left(w, dy\right) v_b\left(y\right) / \varpi_b\left(w\right).$$

This approach to developing an importance sampler for computing $P_{w_0}(T_b < T_{w_0})$ was recently implemented in the $G/G/1$ case by Blanchet and Glynn (2007) (using for $v_b(y)$ a classical heavy-tailed approximation sometimes cited in the literature as the Pakes-Veraverbeke theorem; see, for instance, Foss et al (2007)). However, one difficulty with this idea (that could be troublesome specially in higher dimensional problems) is the need to develop an efficient algorithm for simulating transitions from the kernel $\widetilde{R}_b$. The difficulty is that the kernel $\widetilde{R}_b$ is not a priori constructed in such a way that the path generation problem has an immediate solution.

An alternative is to take advantage of known problem structure relating, in particular, to the probabilistic mechanism that generates a visit to B_b prior to returning to the regeneration state w_0. Let us try to explain this idea by drawing parallels between heavy-tailed situations, which are the focus of our development, and light-tailed environments, for which such probabilistic mechanisms are better understood. For example, in light-tailed queues, one knows that such paths occur when the associated random walk is exponentially twisted according to a twisting parameter θ (that, in principle, can be chosen in a state-dependent way, see Dupuis, Sezer and Wang (2005)). On the other hand, for heavy-tailed queues, one expects that the associated random walk proceeds according to increments that are chosen as a mixture of a big jump, which occurs say with probability p, and a regular size jump, which happens with probability $(1-p)$. This intuition has become standard in the analysis of heavy-tailed systems, see for instance, the paper of Anantharam (1989), which discusses conditional limit theorems for the workload process of a single-server queue given the occurrence of a large delay. The parameter p may be state-dependent (just as we indicated for θ in the light-tailed case). In both the light or heavy tailed cases (or other environments that could involve a mixture of these two cases) a parametric family of state-dependent changes-of-measure induces a one-step transition kernel of the form $Q_\beta = (Q_\beta(w, dz) : w, z \in S)$, where $Q_\beta(w, dz)$ can be represented, given a parameter vector β, as $Q_\beta(w, dz) = q_b^{-1}(\beta, w, y) P(w, dz)$. The function $q_\beta(\cdot)$ is the corresponding (local) likelihood ratio, which is normalized so that $Q_\beta(\cdot)$ is a well defined Markov kernel. The parameter β might include θ in the light-tailed case or p in the heavy-tailed case. Transitions under Q_β could be simulated by exponential twisting in the light-tailed setting and via mixture sampling in the heavy-tailed context. As a consequence, constraining Q_β to be of this form forces the Markov chain to be easily simulatable under the importance distribution.

This key variate generation insight is due to Dupuis and Wang (2004). They further recognized that the state-dependent choice of $(\beta(w) : w \in S)$ that minimizes the second moment of $I(T_b < T_{w_0}) L_b$ under the importance sampler is the optimal control associated with the Hamilton-Jacobi-Bellman equation

$$V_b(w) = \min_\beta E_w[q_b(\beta, w, W_1) V_b(W_1)] \tag{4}$$

subject to $V_b(w) = 1$ on B_b. (They specifically point out this connection in minimizing the variance for light-tailed uniformly geometrically ergodic Markov chains). The

value function $V_b(\cdot)$ represents the lowest second moment that we can achieve for an importance sampling scheme based on the family (indexed by β) of Q_β's. Since the estimators considered are unbiased, (4) equivalently provides a minimum variance estimator among the class of importance sampling estimators indexed by β.

Because (4) is typically difficult to solve, an alternative is to seek a "control" $\beta = (\beta(w) : w \in S)$ that is efficient but not necessarily optimal (in the sense of (4)). To verify efficiency, one must bound $E_{w_0}^{Q_\beta} I(T_b < T_{w_0}) L_b$ (over b). Given a state-dependent selection $(\beta(w) : w \in S)$ of β, this requires bounding the second moment quantity.

$$\begin{aligned} s_b(w_0) &\triangleq E_{w_0}^{Q_\beta}\left(I(T_b < T_{w_0}) \prod_{j=1}^{T_b} q_b(\beta(W_{j-1}), W_{j-1}, W_j)^2\right) \\ &= E_{w_0}\left(I(T_b < T_{w_0}) \prod_{j=1}^{T_b} q_b(\beta(W_{j-1}), W_{j-1}, W_j)\right). \end{aligned}$$

More generally, given that finally we are interested in the efficiency of the waiting time sequence, we are also concerned with bounding

$$s_{b,\chi}(w_0) \triangleq E_{w_0}\left(I(T_b < T_{w_0}) \prod_{j=1}^{T_b} q_b(\beta(W_{j-1}), W_{j-1}, W_j)\chi(W_{T_b})\right),$$

for a $\chi : S \longrightarrow [0, \infty)$. The following proposition allows to obtain the desired bound on $s_{b,\chi}(\cdot)$ (and, consequently, on $s_b(\cdot)$).

Proposition 1 *Suppose that there exists a function $h_b : S \longrightarrow [0, \infty)$ satisfying*

i.)

$$E_w q_b(\beta(w), w, W_1) h_b(W_1) I(W_1 \in w_0^c) \leq h_b(w)$$

for $w \in B_b^c$;

ii) $h_b(w) \geq \varepsilon\chi(w)$ for $w \in B_b$.

Then, $s_{b,\chi}(w) \leq \varepsilon^{-1} h_b(w)$ for $w \in S$.

Proof. Let $M = (M_n : n \geq 1)$ be defined via

$$M_n = \prod_{j=1}^{T_b \wedge n} q_b(\beta(W_{j-1}), W_{j-1}, W_j) h_b(W_{T_b \wedge n}) I(T_{w_0} > T_b \wedge n).$$

Note that, because of condition i), we have that M is a non-negative supermartingale adapted to the filtration generated by the chain W. Since $P_w(T_{w_0} < \infty) = 1$ for $w \in S$ we have that

$$M_n \longrightarrow \prod_{j=1}^{T_b} q_b(\beta(W_{j-1}), W_{j-1}, W_j) h_b(W_{T_b}) I(T_{w_0} > T_b)$$

as $n \nearrow \infty$. Fatou's lemma and the supermartingale property imply that

$$\begin{aligned} & E\left(\prod_{j=1}^{T_b} q_b\left(\beta\left(W_{j-1}\right), W_{j-1}, W_j\right) h_b\left(W_{T_b}\right) I\left(T_{w_0} > T_b\right)\right) \\ \leq \; & EM_0 = h_b(w). \end{aligned}$$

The previous inequality, combined with condition ii), yield the statement of the result. ■

The function h_b is called a Lyapunov function.

The next result shows how the previous Lyapunov bounds immediately yield upper bounds on rare-event probabilities for heavy-tailed models. Such upper bounds are often the most challenging part of those types of asymptotic calculations.

Corollary 1 *The Lyapunov function satisfying Proposition 1 yields an upper bound on* $P_w\left(T_b < T_{w_0}\right)$*, namely* $P_w\left(T_b < T_{w_0}\right) \leq \left(h_b(w)/\varepsilon\right)^{1/2}$.

Proof.

$$\begin{aligned} P_w\left(T_b < T_{w_0}\right)^2 &= \left(E_w^{Q_\beta} I\left(T_b < T_{w_0}\right) L_b\right)^2 \\ &\leq E_w^{Q_\beta} I\left(T_b < T_{w_0}\right) L_b^2 \\ &= s_b(w) \leq h_b(w)/\varepsilon. \end{aligned}$$

■

Note that the zero-variance change-of-measure for $\{T_b < T_{w_0}\}$ is Markovian and (obviously) efficient, so that $s_b(w)$ is then given by $P_w\left(T_b < T_{w_0}\right)$. Since we are developing our (hopefully) efficient change-of-measure so as to mimic the zero-variance Markovian conditional distribution, this suggests that $E_w^{Q_\beta} I\left(T_b < T_{w_0}\right) L_b^2$ should behave (roughly) like $P_w\left(T_b < T_{w_0}\right)^2$. In the presence of good intuition (or known asymptotics) for the model, this recommends the choice of Lyapunov function $h_b(w) = v_b(w)^2$, where $v_b(w)$ is our approximation to $P_w\left(T_b < T_{w_0}\right)$. Note that our chosen approximation will often be poor when w is close to B_b. Because Proposition 1 demands that the appropriate inequality be satisfied everywhere on B_b^c, it will often be useful to introduce some additional parameters into $v_b(w)^2$ so as to provide more flexibility in satisfying the Lyapunov inequality. The development of a practically implementable and theoretically efficient importance sampler then comes down to choosing β and the parameters of the Lyapunov function in such a way that the Lyapunov inequality is satisfied (and so that $v_b(w)$ is of the order of magnitude of $P_w\left(T_b < T_{w_0}\right)$). This suggests the following general approach to building efficient importance samplers:

Step 1: Guess an appropriate parametric functional form for h_b, typically based on intuition or asymptotics available for v_b.

Step 2: Find a feasible (possibly state-dependent) choice for β and for the parameters present in h_b that jointly satisfy the Lyapunov inequality of Proposition 1.

In the next section, we illustrate the use of the above ideas by showing how Steps 1 and 2 lead to an efficient mixture-based importance sampling algorithm for computing steady-state tail probabilities for the single-server queue.

4 Mixture-based Importance Sampling for the $G/G/1$ Queue

Let $W = (W_n : n \geq 0)$ be the waiting time sequence (exclusive of service) for a single-server queue having an infinite capacity waiting room under a first-in first-out (FIFO) queue discipline. We assume that the interarrival times between successive customers form an iid sequence that is independent of the service requirements that themselves are assumed to form an iid sequence. Accordingly, it is well known that W satisfies the recursion

$$W_{n+1} = (W_n + X_{n+1})^+,$$

where $(X_n : n \geq 1)$ is iid (see, for example, Asmussen (2003) p. 267). The sequence W forms a Markov chain on the state space $S = [0, \infty)$. We require that $EX_1 < 0$, so that the queue is stable, and the Markov chain W is a positive recurrent Harris chain. Let W_∞ be a rv having the stationary distribution of W. Our goal is to efficiently compute the steady-state tail probability $P(W_\infty > b)$ when b is large in the presence of heavy-tailed service time requirements. In particular, we shall require X_1 to have a continuous regularly varying density f with index $(\alpha + 1) > 0$, so that

$$f(t) = L(t)\, t^{-(\alpha+1)}$$

for $t > 0$. (The function $L(\cdot)$ is slowly varying, i.e. $L(tm)/L(t) \longrightarrow 1$ as $t \nearrow \infty$ for each $m > 0$). In addition, we shall assume that $Var(X) < \infty$. We shall write $\overline{F}(t) = P(X_1 > t)$ for all $t \in \mathbb{R}$, set $G(x) = \int_x^\infty \overline{F}(t)\, dt$, and let X be a generic rv having the same distribution as X_1. (It is worth noting that by Karamata's theorem (see Embrechts et al (1997), p. 567) $\overline{F}(\cdot)$ is regularly varying with index α and $G(\cdot)$ is regularly varying with index $\alpha - 1$.)

Let us provide some discussion regarding the previous assumptions. First, assuming the existence of a density is a technical condition imposed to facilitate the handling of a remainder term in a Taylor expansion (which will be shown in equation (7) few paragraphs below). Such condition may be removed by introducing a smoothing technique but we shall not pursue this idea here. The most important assumption

concerns regular variation. Turns out that a simple mixture of two components may not provide a large enough family to produce a strongly efficient estimator for other types of heavy-tailed random variables (such as Weibull or lognormal). However, the basic techniques explained in this paper, based on parameter tuning guided by a Lyapunov inequality, still apply provided an appropriate (parametric) family of importance sampling distributions is selected. Finally, given that the stability condition of the system can be guaranteed even when the variance is infinite, it is also important to explain what is the reason for requiring $Var\,(X) < \infty$. First, let us frame our current $G/G/1$ setting in the context of Section 2. In view of the steady-state probability of interest here and our discussion of Section 2, we put $B_b = [b, \infty)$. Since W regenerates whenever W visits state 0, a convenient choice for w_0 is $w_0 = \{0\}$. As pointed out in Section 2, the key to a good importance sampler for the tail probability $P(W_\infty > b)$ is to construct an efficient sampler for computing $P\,(T_b < T_{w_0})$. Certainly, this event (which becomes rarer and rarer as $b \nearrow \infty$ if $EX < 0$), drives the occurrence of delays larger than b within a busy period. However, the overshoot over the boundary b is asymptotically, as $b \nearrow \infty$, Pareto with index α if the X_k's are regularly varying (see Embrechts et al (1997), Appendix A). Therefore, if $Var\,(X) = \infty$, even the use of the zero-variance change-of-measure for the event $\{T_b < T_{w_0}\}$ could produce an infinite variance estimator for the numerator of the regenerative ratio – this is the type of situation whose importance was stressed in Section 2.

Given the heavy tails that are present here, our prior discussion in Section 3 suggests that, in order to design a good importance sampler for $P\,(T_b < T_{w_0})$, we should consider using mixture distributions that will induce the large jumps associated with the zero-variance conditional distribution of W given $\{T_b < T_{w_0}\}$. More precisely, we consider a change-of-measure, for the transition kernel of W, taking the form

$$\begin{aligned} & Q_{a,p}\,(w, dy) \\ = \; & p\frac{P\,(w + X \in y + dy)}{\overline{F}\,(a\,(b-w))} I\,(y - w > a\,(b-w)) \\ & + (1-p)\,\frac{P\,(w + X \in y + dy)}{P\,(-w < X \leq a\,(b-w))} I\,(y > 0; y - w \leq a\,(b-w))\,, \end{aligned} \tag{5}$$

for $p, a \in (0,1)$. This particular mixture form was introduced in Dupuis, Leder and Wang (2006)) in the setting of tail probability computation for sums of heavy-tailed rvs. We shall permit the mixture probability $p = p(w)$ to be state-dependent, but shall let a be state-independent. We require that $a \in (0,1)$ to reflect the fact that there are paths of significant probability leading to $\{T_b < T_{w_0}\}$ that involve large jumps but take W to a position below b.

In order to find the parameters that make the change-of-measure efficient, we need to apply Proposition 1. For this purpose, given a function $h : [0, \infty) \longrightarrow [0, \infty)$ we

define

$$\begin{aligned} J_1(w) &= \frac{E(h(w+X); X > a(b-w))\overline{F}(a(b-w))}{h(w)\,p(w)}, \\ J_2(w) &= \frac{E(h(w+X); -w \le X \le a(b-w))P(X \in (-w, a(b-w)))}{h(w)(1-p(w))}. \end{aligned}$$

Since the indicators that appear in expression (5) it follows easily that verifying the Lyapunov inequality from Proposition 1 is equivalent to showing that

$$J_1(w) + J_2(w) \le 1, \tag{6}$$

for $w \in (0, b]$. In the sequel, in order to simplify the notation we will drop the dependence of w in J_1 and J_2.

To find a good choice of $p(w)$ and the parameter a, we follow the two step procedure suggested in Section 3.

Step 1: Note that W tends to drift down to w_0 linearly. At each such step along the path to w_0 there is an approximate probability $P(X_1 > b - w)$ of entering B_b on that step (given current position w). This suggests the following fluid approximation

$$P_w(T_b < T_0) \approx \int_0^{(-w/EX)} \overline{F}(b - w - sEX)\, ds.$$

Fluid approximations such as the previous one are standard in the heavy-tailed literature, see for instance, Zwart (2001), Chapter 2 and references therein). The previous approximation, in turn, suggests using a Lyapunov bound taking the parametric form $h(w) = h_0(w) \wedge 1$, where $h_0(w) = k \cdot v_b(w)^2$ and

$$\begin{aligned} v_b(w) &= \int_0^{w+d} \overline{F}(b - w + s)\, ds \\ &= G(b-w) - G(b+d), \end{aligned}$$

for some constants d, $k > 0$ to be determined in the execution of Step 2 explained below. Before moving on to Step 2 and in order to enhance the intuition of the role played by d and k, let us mention that d is introduced to deal with boundary effects close to zero and k controls effects close to b.

Step 2: Let us define

$$\Delta \triangleq (w+X)^+ - w = \max(-w, X).$$

It follows (since $h(\cdot)$ is absolutely continuous) that

$$\begin{aligned} & E(h(w+X); -w < X \le a(b-w)) \\ = \; & h(w)\,P(-w < X \le a(b-w)) + E(\partial_w h(w + U\Delta)\,\Delta; -w < X \le a(b-w)), \end{aligned}$$

where U is uniformly distributed on the interval (0,1) and independent of X. In view of the above Taylor-type representation, observe that the required Lyapunov bound can be approximately written as

$$\begin{aligned} 1 &\geq J_1 + J_2 \\ &\approx \frac{\overline{F}\left(a\left(b-w\right)\right)^2}{p\left(w\right)h\left(w\right)} + \frac{P\left(-w<X\right)^2}{1-p\left(w\right)} + \frac{\partial_w h\left(w\right)}{h\left(w\right)}\frac{E\left(\Delta;-w<X\right)}{1-p\left(w\right)} \end{aligned} \tag{7}$$

when $b-w$ is large enough. In addition, if $h\left(w\right)\leq 1$, then

$$\frac{\overline{F}\left(a\left(b-w\right)\right)^2}{p\left(w\right)h} = \frac{\overline{F}\left(a\left(b-w\right)\right)^2}{p\left(w\right)v_b\left(w\right)^2 k}.$$

It seems natural, in order to cancel the squares in the previous expression, to select

$$p\left(w\right) = \theta\frac{\overline{F}\left(a\left(b-w\right)\right)}{v_b\left(w\right)} = \theta\frac{\overline{F}\left(a\left(b-w\right)\right)}{G\left(b-w\right)-G\left(b+d\right)}, \tag{8}$$

for some appropriately chosen value of θ. With this choice of p, (7) can be approximated as

$$\frac{\overline{F}\left(a\left(b-w\right)\right)}{\theta v_b\left(w\right)k} + \frac{P\left(-w<X\right)}{1-p\left(w\right)} + 2\frac{\overline{F}\left(b-w\right)}{v_b\left(w\right)}\frac{E\left(\Delta;-w<X\right)}{1-p\left(w\right)}. \tag{9}$$

Given the need for this expression to satisfy the Lyapunov bound, our objective is to show that it can be made less than one by selecting θ and k appropriately. Of course, since we expect to have $E\left(\Delta;-w<X\right)<0$ when w is bounded away from zero, it is clear that (9) can be upper bounded by one if we select first θ sufficiently small and then k large enough.

In order to provide further intuition into the choice of p given by (8), note that conditional on $\{W_n = w, T_b \wedge T_{w_0} > n\}$, an idealized choice for $p\left(w\right)$ would be to select it according to

$$P_w\left(W_{n+1}>b\middle|\, T_b<T_{w_0}\right) = \frac{P\left(X>b-w\right)}{P_w\left(T_b<T_{w_0}\right)}.$$

Of course, the right-hand side is the hazard rate at which the rare event occurs when the current position is w. Note that if $v_b\left(w\right)$ is a good approximation to $P_w\left(T_b<T_{w_0}\right)$, the right-hand side behaves roughly like

$$\frac{\overline{F}\left(b-w\right)}{v_b\left(w\right)} = \frac{\partial_w v_b\left(w\right)}{v_b\left(w\right)} = \partial_w \log v_b\left(w\right). \tag{10}$$

But (10) is clearly consistent with (8). Hence, the form of $p\left(w\right)$ given by (10) can be interpreted, in the presence of a good approximation v_b, as being proportional to the hazard rate at which the rare event occurs when the current position is w.

The previous paragraph indicates the main ideas underlying the choice of algorithm parameters and Lyapunov function parameters on that part of the state space that is not close to the boundaries at 0 and b (i.e. on a region of the form $w > \kappa_1$, $b - w \geq \kappa$, for some constants κ_1 and κ) and assuming that $h(w) < 1$. To handle the case in which $w \leq \kappa_1$, we again use (9) and note that

$$v_b(w) = \int_0^{w+d} \overline{F}(b - w + s)\, ds \approx \overline{F}(b - w)(w + d).$$

Therefore, (9) is bounded (using $\Delta \leq X^+$) by a quantity that is roughly equal to

$$\frac{m_1}{k\theta(w+d)} + P(-\kappa_1 < X)^2 \left(1 + \frac{m_1}{w+d}\right) + 2\frac{1}{w+d} E\left(X^+\right),$$

for some constant $m_1 > 0$. Suppose now that κ_1 such that $P(-\kappa_1 < X) < 1$. One can then select $d > 0$ large enough in order to make the previous quantity less or equal to 1. This approach is therefore appropriate for dealing with the part of the state space close to the boundary at 0.

To deal with the part of the complement of B_b that is close to the boundary at b, note that the Lyapunov function has been analyzed above on that part of the state space where $h(w) < 1$. However, on the region $\{w : h(w) = 1\}$, a good choice of p is simple. Indeed, when we are close to the boundary at b, we can select $p = 0$ (that is, we do not apply importance sampling at all; alternatively we can also pick $p = P(X > a(b - w))$). Since $h \leq 1$ globally, the Lyapunov bound is automatically satisfied on this region. This completes the second step of our heuristic construction of an efficient importance sampler for this $G/G/1$ problem.

Let us provide a brief summary of the parameters to be selected in order to carry through the previous two step procedure. First, after selecting the form of $p(w)$, we need to choose θ in order to satisfy the Lyapunov bound whenever $h(w) < 1$. The parameter k, which is selected together with θ, introduces a boundary layer close to b (i.e. a region of the form $\{w : 0 \leq b - w \leq \kappa\}$ for $\kappa > 0$), for which $h(w) = 1$ (recall that on the boundary layer we do not apply importance sampling). Finally, the parameter $d > 0$ is chosen in order to satisfy the Lyapunov bound close to zero (i.e. on a region of the form $\{w : 0 \leq w \leq \kappa_1\}$).

The rest of this section provides rigorous support for the above heuristic derivation. The restrictions on various parameters indicated in the previous paragraphs will be collected throughout conditions **C1** to **C4** which are introduced in the next pages. The bound involving J_1, namely

$$J_1 \leq \frac{\overline{F}(a(b-w))^2}{h_0(w)\, p(w)} \tag{11}$$

is automatic in view of the fact that $h(w) \leq 1$. Hence, the technical details of the construction lie in the analysis of J_2.

We first provide a result that summarizes the drift behavior of the waiting time sequence. The proof is straightforward and therefore omitted.

Lemma 1 *There exists $\varepsilon_0 > 0$ and constants κ_0, $\kappa_1 > 0$ such that for all $y_0 \geq \kappa_0$ and $y_1 \geq \kappa_1$ we have*

$$E(X; -y_1 < X \leq y_0) \leq -\varepsilon_0 < 0. \tag{12}$$

In addition, $\kappa_1 > 0$ can be chosen so that $P(X > -\kappa_1) < 1$.

Throughout the remainder of this paper, we assume that ε_0, κ_0, and κ_1 have been selected as in Lemma 1. We now provide complete details for the Taylor expansion involved in the term J_2 of (7). First, define $m_+, \widetilde{m}_+, \widetilde{m}'_+ \in [1, \infty)$ so that

$$\begin{aligned} \sup_{t \geq 0} \frac{\overline{F}(t(1-a))}{\overline{F}(t)} &\leq m_+, \quad \sup_{t \geq 0} \frac{f(t) G(t)}{\overline{F}(t)^2} \leq \widetilde{m}_+, \\ \sup_{t \geq 0,} \frac{f(t)(t+1)}{\overline{F}(t)} &\leq \widetilde{m}'_+, \end{aligned}$$

and set $m^* = \left(m_+ + m_+ \left(\widetilde{m}_+ \vee \widetilde{m}'_+\right) + \widetilde{m}_+\right) EX^2$. The fact that $\widetilde{m}_+$ and $\widetilde{m}'_+$ are finite follows from Karamata's theorem (see Embrechts et al (1997) p. 567) and the boundedness of $f(\cdot)$. m_+ is finite by definition of regular variation and because $\overline{F}(t) \in (0,1)$ for all $t \geq 0$.

Lemma 2 *For each $\widetilde{\varepsilon} > 0$, there exists $\kappa > 0$ such that if $b - w \geq \kappa$ and $w \geq \kappa_1$ then*

$$\begin{aligned} & E\left(\frac{h_0(w+X)}{h_0(w)}; -w < X \leq a(b-w)\right) \\ \leq\ & P(X \in (-w, a(b-w)]) \\ & + \frac{\partial_w h_0(w)}{h_0(w)} \left(E(\Delta; -\kappa_1 < X \leq a(b-w)) + \widetilde{\varepsilon}\right) + \left(\frac{\partial_w h_0(w)}{h_0(w)}\right)^2 m^*. \end{aligned}$$

Proof. The absolute continuity of h_0 implies that

$$\begin{aligned} & E(h_0(w+X); -w < X \leq a(b-w)) \\ =\ & h_0(w) P(-w < X \leq a(b-w)) \\ & + E(\partial_w h_0(w + U\Delta)\Delta; -w < X \leq a(b-w)). \end{aligned}$$

Now, observe that

$$\begin{aligned}
&\frac{\partial_w h_0(w+U\Delta)}{\partial_w h_0(w)} I(X \in (-w, a(b-w)]) \\
= \;& I(X \in (-w, a(b-w)]) \frac{G(b-w-U\Delta) - G(b+d)}{G(b-w) - G(b+d)} \\
& \cdot \frac{\overline{F}(b-w-U\Delta)}{\overline{F}(b-w)}.
\end{aligned} \tag{13}$$

Because G is also absolutely continuous, we obtain

$$\begin{aligned}
G(b-w-U\Delta) &= G(b-w) \\
&\quad + E\left(\Delta U \overline{F}\left(b-w-U\cdot\widetilde{U}\cdot\Delta\right)\middle| U, \Delta\right),
\end{aligned}$$

where $\widetilde{U}$ is uniformly distributed independent of U and X. Since $\Delta \leq X^+ \leq a(b-w)$, it follows that

$$\begin{aligned}
&\frac{\Delta U E\left(\overline{F}\left(b-w-U\cdot\widetilde{U}\cdot\Delta\right)\middle| U, \Delta\right)}{\overline{F}(b-w)} \\
\leq \;& X^+ \frac{\overline{F}((b-w)(1-a))}{\overline{F}(b-w)} \leq m_+ X^+.
\end{aligned}$$

Thus, we have that

$$\begin{aligned}
& I(X \in (0, a(b-w)]) \frac{G(b-w-U\Delta) - G(b+d)}{G(b-w) - G(b+d)} \\
\leq \;& I(X \in (0, a(b-w)]) \left(1 + m_+ X \frac{\partial_w h_0(w)}{h_0(w)}\right).
\end{aligned}$$

In a similar fashion as in the previous analysis, we obtain that

$$\begin{aligned}
&\frac{I(X \in (0, a(b-w)]) \left(\overline{F}(b-w-U\Delta)\right)}{\overline{F}(b-w)} \\
\leq \;& I(X \in (0, a(b-w)]) \left(1 + \widetilde{m}_+ \frac{\partial_w h_0(w)}{h_0(w)} X\right).
\end{aligned}$$

As a consequence, collecting our previous inequalities and the definition of m^*, we find that, for each $\varepsilon > 0$, it is possible to find $\kappa > 0$ such that if $b - w \geq \kappa$, then

$$\begin{aligned}
& E(\partial_w h_0(w+U\Delta)\Delta; 0 < X \leq a(b-w)) \\
\leq \;& \partial_w h_0(w) \left(E(\Delta; 0 < X \leq a(b-w_1)) + \varepsilon + \frac{\partial_w h_0(w)}{h_0(w)} m^*\right).
\end{aligned} \tag{14}$$

For the case $X \in (-w, 0]$ we argue as follows. First we note (since $\partial_w h_0$ is positive)

$$\begin{aligned} & E\left(\frac{\partial_w h_0(w+U\Delta)}{\partial_w h_0(w)}\Delta; -w < X \leq 0\right) \\ \leq\ & E\left(\frac{\partial_w h_0(w+U\Delta)}{\partial_w h_0(w)}\Delta; -\kappa_1 < X \leq 0\right) \\ \leq\ & \frac{\overline{F}(b-w+\kappa_1)}{\overline{F}(b-w)} E(\Delta; -\kappa_1 < X \leq 0) \\ \leq\ & E(\Delta; -\kappa_1 < X \leq 0)\left(1+\frac{m'}{b-w+1}\right), \end{aligned} \tag{15}$$

where

$$\sup_{t\geq 0}\sup_{0\leq r\leq \kappa_1} \frac{f(t+r)(1+t)}{\overline{F}(t)} \leq m'.$$

The fact that m' is finite is a consequence of Karamata's theorem and the fact that $f(\cdot)$ is bounded. It is clear then that (15) combined with (14) yields the conclusion of the result. ■

Lemmas 1 and 2 give rise to the following constraint on d and κ.

C1 Given $\delta \in (0,1)$, select $d, \kappa > 0$ so that if $b - w \geq \kappa \geq \kappa_0$ and $w > \kappa_1$, then

$$\begin{aligned} & E\left(\frac{h_0(w+X)}{h_0(w)}; -w < X \leq a(b-w)\right) \\ \leq\ & 1 - \frac{\partial_w h_0(w)}{h_0(w)}\varepsilon_0(1-\delta). \end{aligned} \tag{16}$$

Corollary 2 *It is always possible to satisfy* ***C1*** *by appropriately choosing d first and then κ.*

Proof. First, select $\widetilde{\varepsilon} = \varepsilon_0\delta/2$ in Lemma 2. Then, in view of Lemma, 1 it suffices to show that d and κ can be chosen so that

$$\frac{\partial_w h_0(w)}{h_0(w)} m^* \leq \varepsilon_0\delta/2.$$

First, let $\kappa > 0$ such that

$$\sup_{t\geq\kappa} \frac{\overline{F}(t)}{G(t)} \leq \varepsilon_0\delta/4.$$

Then, picking $r_2 < 1/2$, and noting that for each $r_1 \in (0,1)$ and $d > 0$, there exists b_0 such that if $b \geq b_0$ we have

$$r_2 G(b(1-r_1)) \geq G(b) \geq G(b+d).$$

Therefore, if $w \in (br_1, b]$ we have that

$$G(b-w) - G(b+d) \geq (1-r_2)\, G(b-w),$$

which in turn implies that if $b - w \geq \kappa$ then

$$\frac{\overline{F}(b-w)}{G(b-w) - G(b+d)} \leq \frac{\overline{F}(b-w)}{(1-r_2)\, G(b-w)} \leq \frac{\varepsilon_0 \delta}{2}.$$

Now, we consider $w \in [0, br_1)$. Note that

$$G(b-w) - G(b+d) = E\left(\overline{F}(b-w+U(w+d))\right),$$

where U is a uniformly distributed random variable. Therefore,

$$\frac{\overline{F}(b-w)}{G(b-w) - G(b+d)} = \frac{1}{w+d} \frac{\overline{F}(b-w)}{E\left(\overline{F}(b-w+U(w+d))\right)}.$$

Consequently, as long as we have $r_1 b \geq d$, we obtain

$$\frac{\overline{F}(b-w)}{E\left(\overline{F}(b-w+U(w+d))\right)} \leq \frac{\overline{F}(b(1-r_1))}{\overline{F}(b(1+2r_1))} \leq m'_R$$

for some $m'_R > 0$ (by regular variation). It follows that we can pick we can pick d sufficiently large so that for all $b \geq b_0$ and $w \leq br_1$

$$\frac{\overline{F}(b-w)}{G(b-w) - G(b+d)} \leq \frac{m'_R}{d} \leq \varepsilon_0 \delta / 2.$$

■

In order to further characterize $p(\cdot)$ and collect some of our estimates, first let us pick $m_1 \in (0, \infty)$ such that

$$\sup_{t \geq 0} \frac{\overline{F}(at)}{\overline{F}(t)} \leq m_1. \tag{17}$$

(Note that such m_1 exists by virtue of Karamata's theorem and because $\overline{F}(t) \in (0,1)$ for $t \geq 0$). In addition, we introduce the following constrains on θ and k

C2 For $\delta \in (0,1)$ as in **C1**, select θ and k to satisfy

$$\begin{aligned} \theta &\leq \varepsilon_0 (1-\delta) / (4m_1), \\ k &\geq 2m_1 / (\theta \varepsilon_0 (1-\delta)) \end{aligned}$$

respectively.

The following result proves that with these choices, the Lyapunov bound holds on the region $w \geq \kappa_1$ and $b - w \leq \kappa$.

Proposition 2 *Suppose that **C1** and **C2** are in force. Then,*

$$J_1 + J_2 \leq 1$$

as long as $h_0(w) \leq 1$.

Proof. Corollary 2 and the fact that $\partial_w h_0(w)/h_0(w) \leq p/\theta$ imply

$$\begin{aligned} J_1 + J_2 &\leq \frac{\partial_w h_0(w)}{h_0(w)} \frac{m_1}{\theta k} + \frac{1}{1-p(w)} - \frac{\partial_w h_0(w)}{h_0(w)} \frac{\varepsilon_0(1-\delta)}{(1-p(w))} \\ &\leq \frac{\partial_w h(w)}{h(w)} \left(\frac{m_1}{\theta k} + 2m_1\theta - \varepsilon_0(1-\delta) \right) + 1 \end{aligned}$$

Since $\partial_w h/h > 0$, the selection of θ and k automatically imply that the previous quantity is guaranteed to be less or equal to 1. ■

We now proceed to the construction of the Lyapunov bound on the set $\{w \leq \kappa_1\}$. Define $\widetilde{\pi} = P(X > -\kappa_1)$. Note that κ_1 has been selected so that $\widetilde{\pi} < 1$ (see Lemma 1). As in the case $w > \kappa_1$, we now provide a set of additional constraints for the parameters d and κ.

C3 Given $\widetilde{\delta} \in (0,1)$, d must satisfy

$$d \geq \left(1 + \widetilde{\delta}\right) \left(\varepsilon_0(1-\delta) + 2EX^+\right) / (1-\widetilde{\pi}), \tag{18}$$

and $\kappa > 0$ must have the property that for all $w \leq \kappa_1$ and $b - w > \kappa$,

$$\frac{\partial_w h_0(w)}{h_0(w)} = \frac{\overline{F}(b-w)}{G(b-w) - G(b+d)} \leq \frac{\left(1+\widetilde{\delta}\right)}{w+d} \tag{19}$$

and also

$$\begin{aligned} & E\left(h_0(w+X); -w < X \leq a(b-w)\right) \\ \leq\ & h_0(w)\,\widetilde{\pi} + \partial_w h_0(w)\, E\left(X^+\right). \end{aligned} \tag{20}$$

The next result shows that the constraint **C3** can always be satisfied (simultaneously with **C1** and **C2**).

Lemma 3 *The constraints imposed by **C1** to **C3** can always be jointly satisfied.*

Proof. First, given the constraints imposed on $d > 0$ we note that constraint (19) is satisfied given a selection of $d > 0$ because

$$\begin{aligned} & G(b-w) - G(b+d) \\ = & \int_0^{w+d} \overline{F}(b-w+s)\, ds \sim \overline{F}(b-w)(w+d) \end{aligned}$$

as $b \nearrow \infty$ uniformly over $0 \leq w \leq \kappa_1$. The fact that (20) is satisfiable follows directly from Lemma 2. ■

We now are ready to provide the result that summarizes the construction of the Lyapunov bound.

Theorem 1 *Under conditions **C1** to **C3**, we have that if $h_0(w) \leq 1$ respectively, then*

$$J_1 + J_2 \leq 1.$$

Proof. Under **C1** to **C3**, we have that

$$\begin{aligned} J_1 + J_2 \quad \leq \quad & \frac{\partial_w h(w)}{h(w)} \frac{m_1}{\theta k} + \frac{\partial_w h(w)}{h(w)} 2m_1 \theta \widetilde{\pi} + \widetilde{\pi} \\ & + 2\frac{\partial_w h(w)}{h(w)} E\left(X^+\right). \end{aligned} \tag{21}$$

In addition, condition **C3** also yields

$$\frac{\partial_w h(w)}{h(w)} \leq \frac{\left(1+\widetilde{\delta}\right)}{w+d}, \tag{22}$$

for $\widetilde{\delta} \in (0,1)$. Then, in this case, we obtain that (21) is bounded by

$$\frac{\left(1+\widetilde{\delta}\right)}{w+d}\left(\frac{m_1}{\theta k} + 2m_1\theta\widetilde{\pi} + 2EX^+\right) + \widetilde{\pi}.$$

In turn, given the selection of θ and k specified in constraint **C2**, we have that the expression in the previous display is bounded by

$$\frac{\left(1+\widetilde{\delta}\right)}{w+d}\left(\varepsilon_0(1-\delta) + 2EX^+\right) + \widetilde{\pi} \leq \frac{\left(1+\widetilde{\delta}\right)}{d}\left(\varepsilon_0(1-\delta) + 2EX^+\right) + \widetilde{\pi}.$$

Since $\widetilde{\pi} < 1$ we conclude that if we choose

$$d \geq \left(1+\widetilde{\delta}\right)\left(\varepsilon_0(1-\delta) + 2EX^+\right)/(1-\widetilde{\pi}),$$

Then, there exists $b_0 > 0$ selected according to (22) for which the conclusion of the result follows. ■

We close this section with the description of the algorithm suggested by the previous theorem. However, before we provide the final description of the algorithm we impose an additional constraint on the parameter $k > 0$.

C4 Select $k > 0$ so that $h_0(w) \geq 1$ if $b - w \geq \kappa$.

Note that constraint **C4** can clearly be satisfied because one can always increase the size of k given the rest of the parameters. The selection of $k > 0$ in order to satisfy **C4** is important because it determines the region within which to apply importance sampling. We are ready to provide the precise description of our algorithm.

Algorithm 1

Set $b > 0$ and fix $a \in (0,1)$. Initialize $w = 0$, $REACH = 0$ and $L = 1$. Suppose that **C1** to **C4** are in force.

STEP 1

While $REACH = 0$

If $h(w) = 1$ then sample X according to the nominal distribution.

Else set

$$p = \theta \frac{\overline{F}(a(b-w))}{G(b-w) - G(b+d)} \wedge 1/2.$$

Sample X as follows. With probability p generate X with law $\mathcal{L}(X | X \geq a(b-w))$, with probability $1-p$ sample X with law $\mathcal{L}(X | -w < X \leq a(b-w))$. Then, update

$$\begin{aligned} L \longleftarrow\ & L \cdot [p^{-1}\overline{F}(a(b-w))\, I(X > a(b-w)) \\ & + (1-p)^{-1} P(-w < X \leq a(b-w))\, I(-w < X \leq a(b-w))]. \end{aligned}$$

Endif

Update

$$w \longleftarrow (w + X)^+,$$

If $w \notin (0, b]$ then $REACH \longleftarrow 1$

Endif

Loop

STEP 2 Set $L \longleftarrow L \cdot I(w_1 > b)$ and RETURN L.

The following theorem summarizes the statistical efficiency properties of the previous estimator. The analysis of the total efficiency for estimating the waiting time sequence is given in the next section.

Theorem 2 *If* $s_b(0) = E_0^{Q_{a,p}}(L^2)$ *(where* $E_0^{Q_{a,p}}(\cdot)$ *is the probability measure induced by the importance sampling scheme indicated in Algorithm 1 and* L *is the final output indicated in STEP 2), then*

$$\sup_{b>0} \frac{s_b(0)}{P_0(T_b < T_{w_0})^2} < \infty.$$

Proof. Our previous analysis combined with Proposition 1 yields

$$s_b(0) \leq h(0)$$

(note that k was selected so that $h(W_{T_b}) = 1$). On the other hand, it follows (by choosing the first service time in the busy cycle larger than b) that

$$\underline{\lim}_{b \longrightarrow \infty} \frac{P_0(T_b < T_{w_0})}{P(X > b)} > 0.$$

Consequently, using Corollary 1, we obtain that there exists $\delta' > 0$ such that

$$\delta' \overline{F}(b) \leq P_0(T_b < T_{w_0}) \leq h(0)^{1/2}.$$

Since the asymptotic relation

$$h(0)^{1/2} \sim k^{1/2}[G(b) - G(b+d)] \sim k^{1/2} \overline{F}(b)\, d$$

holds as $b \nearrow \infty$, the previous observations imply (by virtue of regular variation) the statement of the theorem. ■

5 Efficiency for the Steady-state Delay

We impose the same assumptions indicated in Section 4. Our goal is to show that the algorithms developed in the previous two sections provide efficient estimators for the tail of the steady-state waiting time, namely $P(W_\infty > b)$, when b is large.

We define

$$N_b = \sum_{j=0}^{T_{w_0}-1} I(W_j > b)$$

and let $N_b(w)$ be a rv with the distribution $P_w(N_b \in \cdot)$. Finally, we set $\iota_b(w) = E_w N_b^2$.

We are interested in studying the performance of the estimator

$$Z_b = L_b N_b(W_{T_b})\, I(T_b < T_{w_0}),$$

where L_b is the likelihood ratio obtained by running the importance sampling algorithm described in Section 4, namely **Algorithm 1**. The rv $N_b(W_{T_b})$ is simulated

under the nominal dynamics of the system (i.e. it is not required to apply importance sampling anymore). Note for the generation of $N_b\left(W_{T_b}\right)$ one can apply additional variance reduction techniques, such as control variates (for more information on control variates and other variance reduction techniques see, for instance, Liu (2001)).

We want to establish efficiency, which, as explained in Section 2, involves proving

$$\sup_{b>0} \frac{E_0^{Q_{a,p}}\left(L_b^2 N_b^2\left(W_{T_b}\right) I\left(T_b < T_{w_0}\right)\right)}{E\left(N_b\right)^2} < \infty,$$

where $E_0^{Q_{a,p}}(\cdot)$ denotes the expectation operator induced by the importance sampler selected in Section 4. Now, we have that

$$E_0^{Q_{a,p}}\left(L_b^2 N_b^2\left(W_{T_b}\right) I\left(T_b < T_{w_0}\right)\right) = E_0\left(L_b N_b^2\left(W_{T_b}\right) I\left(T_b < T_{w_0}\right)\right).$$

Our strategy is to study

$$E_w\left(L_b N_b\left(W_{T_b}\right) I\left(T_b < T_{w_0}\right)\right)$$

again using Lyapunov-type arguments.

Note that

$$E_w\left(L_b N_b^2\left(W_{T_b}\right) I\left(T_b < T_{w_0}\right)\right) = E_w\left(L_b \cdot \iota_b\left(W_{T_b}\right) I\left(T_b < T_{w_0}\right)\right).$$

We will complete our program in three steps. First, we will establish the following proposition.

Proposition 3 *There exists a constant $m > 0$ such that*

$$\iota_b\left(W_{T_b}\right) \leq m\left(W_{T_b}\right)^2. \tag{23}$$

Proof. This follows from standard properties for stopped random walks; see Gut (1988) p. 92. ■

This implies that

$$E_w\left(L_b N_b^2\left(W_{T_b}\right) I\left(T_b < T_{w_0}\right)\right) \leq m E_w\left(L_b \cdot \left(W_{T_b}\right)^2 I\left(T_b < T_{w_0}\right)\right), \tag{24}$$

for some $m > 0$. The key issue then becomes finding a convenient bound for

$$e_b(w) = E_w\left(L_b \cdot \left(W_{T_b}\right)^2 I\left(T_b < T_{w_0}\right)\right),$$

which is the content of the following result.

Proposition 4 *Define $\widetilde{e}_b(\cdot)$ via*

$$\widetilde{e}_b(w) = h(w)\left(b^2 I(b > w) + \delta w^2 I(b \leq w)\right).$$

Then, we can find $\delta \in (0,1)$ (independent of b) such that

$$\widetilde{e}_b(w) \geq E_w\left(L_b \cdot \left(W_{T_b}\right)^2 I\left(T_b < T_{w_0}\right)\right)/\delta.$$

Proof. We will apply Proposition 1 using as our Lyapunov function $\widetilde{e}_b(\cdot)$. The strategy proceeds using similar steps as those followed in Section 4. First define

$$\begin{aligned}\widetilde{J}_1 &= \frac{E_w\left(\widetilde{e}_b(w+X); X \geq a(b-w)\right)\overline{F}(a(b-w))}{\widetilde{e}_b(w)\,p(w)},\\ \widetilde{J}_2 &= \frac{E\left(\widetilde{e}_b(w+X); -w < X \leq a(b-w)\right)P(X \in (-w, a(b-w)])}{\widetilde{e}_b(w)\,p(w)}.\end{aligned}$$

Note that

$$\widetilde{J}_2 = \frac{b^2 E\left(h(W); X \leq a(b-w)\right)\overline{F}(a(b-w))}{\widetilde{e}_b(w)\,p(w)}.$$

So, the analysis of $\widetilde{J}_2$ is completely analogous to that of J_2 in Section 4. We just need to analyze $\widetilde{J}_1$ on $w < b$. Note that

$$\begin{aligned}&\widetilde{J}_1\\ \leq\ & E\left(\frac{h(w+X)}{h(w)}; X \geq a(b-w), w+X < b\right)\frac{\overline{F}(a(b-w))}{p(w)}\\ &+\delta E\left(\frac{(w+X)^2}{b^2}\middle| X \geq (b-w)\right)\frac{\overline{F}(a(b-w))^2}{h(w)\,p(w)}.\end{aligned}$$

It follows easily, using the facts that X is regularly varying and that $Var(X) < \infty$, that there exists a constant $m \in (0,\infty)$ such that for all $b \geq 1$

$$E\left(\frac{(w+X)^2}{b^2}\middle| X \geq (b-w)\right) \leq m.$$

Therefore, we obtain that if $w < b$

$$\widetilde{J}_1 + \widetilde{J}_2 \leq J_1 + J_2 + \delta m J_1,$$

Given our analysis of J_1 and J_2 it is clear then that $\delta > 0$ can be chosen so that

$$\widetilde{J}_1 + \widetilde{J}_2 \leq 1$$

on $w \leq b$. The result then follows by applying Proposition 1. ■

Using the previous two propositions we arrive at the last step of our program, which yields the main result of this section, namely

Theorem 3 *Assume that $Var(X) < \infty$. Then,*

$$\sup_{b>0}\frac{E_0^{Q_{a,p}}\left(L_b^2 N_b^2(W_{T_b})\,I(T_b < T_{w_0})\right)}{P(W_\infty > b)^2} < \infty.$$

In addition, let $M(b)$ be the number of variate generations required to produce a single replication of $L_b N_b(W_{T_b})$. Then, $EM(b) \leq \eta(b+1)$ for some $\eta \in (0,\infty)$.

Proof. Proposition 4 together with (24) imply that

$$E_0^{Q_{a,p}}\left(L_b^2 N_b^2\left(W_{T_b}\right) I\left(T_b < T_{w_0}\right)\right) \leq mh\left(0\right) b^2.$$

On the other hand, it is not difficult to develop a lower bound that implies the existence of $\delta > 0$ such that

$$P\left(W_\infty > b\right) \geq \delta b^2 P\left(X > b\right)^2,$$

see, for instance, Zachary (2004), p. 2, where such lower bound is in fact developed in much greater generality, assuming only that X is long-tailed (see, for instance, Embrechts et al (1997) for a detailed discussion on different classes of heavy-tailed distributions). Alternatively, instead of developing a lower bound separately, we can invoke Pakes-Veraverbeke's heavy-tailed exact asymptotic, which applies in the presence of regularly varying tails (see, for instance, Asmussen (2003)). We conclude that

$$\frac{E_0^{Q_{a,p}}\left(L_b^2 N_b^2\left(W_{T_b}\right) I\left(T_b < T_{w_0}\right)\right)}{P\left(W_\infty > b\right)^2} \leq \frac{m\left(G\left(b\right) - G\left(b+d\right)\right)}{\delta P\left(X > b\right)}.$$

The previous quantity is clearly bounded uniformly over $b > 0$, so the conclusion of the first part of the theorem follows. A Lyapunov type argument of the style given in the proof of Proposition 4 shows that $E_0^{Q_{a,p}} T_{w_0} \leq \eta\left(1+b\right)$ for some $\eta > 0$. This, in turn implies that $EM\left(b\right) \leq \eta\left(b+1\right)$. ■

6 An $M/G/1$ Example

To illustrate the implementation issues and performance of our algorithm, we consider an $M/G/1$ queue. We do this purely in order to permit comparison of our method with competing algorithms. We recall that our algorithm is more general and does not require Poisson arrivals. We assume that the service times are Pareto distributed with index $\alpha > 0$. In particular, if V denotes a generic service time, then

$$P\left(V > t\right) = \frac{1}{\left(1+t\right)^\alpha}.$$

Moreover, suppose that $\alpha = 5/2$ so that $EV = 1/\left(\alpha - 1\right) = 2/3$ and $EV^2 < \infty$. The inter-arrival times follow an exponential distribution with mean $1/\lambda = 4/3$. Consequently, the traffic intensity $\rho = \lambda EV$ is equal to $1/2$. If τ is a generic inter-arrival time independent of V, then we write $X = V - \tau$. The tail of X, namely $\overline{F}\left(\cdot\right) = P\left(X > \cdot\right)$, is computed via

$$\begin{aligned}
\overline{F}\left(x\right) &= P\left(V > x + \tau\right) = \int_0^\infty \lambda e^{-\lambda s} P\left(V > x + s\right) ds. \\
&= I\left(x < 0\right)\left(1 - e^{\lambda x}\right) + \int_{(-x)\vee 0}^\infty \lambda e^{-\lambda s} \frac{1}{\left(s + x + 1\right)^{5/2}} ds.
\end{aligned}$$

where

$$\begin{aligned}
&\int_y^\infty \lambda e^{-\lambda s} \frac{1}{(s+x+1)^{5/2}} ds \\
= \;& e^{\lambda(x+1)} \lambda^{5/2} \left[\frac{2}{3} \frac{e^{-\lambda(x+y+1)}}{(\lambda(x+y+1))^{3/2}} - \frac{4}{3} \frac{e^{-\lambda(x+y+1)}}{(\lambda(x+y+1))^{1/2}} \right] \\
& + e^{\lambda(x+1)} \lambda^{5/2} \frac{4}{3} \int_{\lambda(x+y+1)}^\infty \frac{e^{-t}}{\sqrt{t}} dt.
\end{aligned}$$

Implementation of the algorithm requires evaluation of the integral

$$\int_{\lambda(x+y+1)}^\infty \frac{e^{-t}}{\sqrt{t}} dt$$

numerically. This integral is an incomplete Gamma function; there are many methods available to evaluate this function efficiently. In general, in the implementation of the proposed algorithms, it will typically be the case that one would need to numerically evaluate one dimensional integrals – which in most cases can be evaluated with high accuracy in relative terms using routine methods.

First, we selected $a = .9$ (recall that $a \in (0,1)$ is the parameter that dictates the fraction of the size of the jump required to reach level b, see equation (5)). In order to avoid the need to numerically evaluate $G(\cdot)$ when implementing the algorithm (which would involve integrating $\overline{F}(\cdot)$), we use here the modified Lyapunov function

$$h(w) = \left(10\left(\left(20(b-w)\right) \wedge (w+5)\right)^2 \overline{F}(b-w)^2\right) \wedge 1.$$

The parameters of the function (together with the selection of p given below) have been selected following the same techniques explained in the previous sections in order to satisfy the Lyapunov bound – note that the function has the same asymptotic behavior as the Lyapunov bound that we studied during our theoretical analysis as $b \nearrow \infty$. The variate generation of each increment is given by (5), as indicated in STEP 2 of Algorithm 1. However, again, the techniques explained in the previous section were adapted in order to obtain a more convenient (in terms of implementation) expressions for the mixture probabilities – basically this involves a style of computation similar to that of Corollary 2). In particular, in our implementation we use the following rule to select the mixture probability p.

- When $h(w) < 1$, then we use

$$p = \max\left(\frac{1}{2(w+5)}, \frac{5}{2(b-w+5)}, \left(1 - \overline{F}(-w)\right)/2 \right)$$

.

- Otherwise, $h(w) = 1$, do not apply importance sampling. Alternatively, in this step, we can also select $p = P(X > a(b - w))$

The likelihood ratio is computed as indicated in Algorithm 1. The following table summarizes the performance of our algorithm and several other methods. BGL corresponds to our algorithm, BL is a variant of the algorithm proposed by Blanchet and Glynn (2007) which is adapted for the M/G/1 case in Blanchet and Li (2006). DLW corresponds to the methods proposed by Dupuis, Leder and Wang (2006), JS corresponds to the hazard rate twisting procedure of Juneja and Shahabuddin (2002) and AB is a conditional Monte Carlo procedure proposed by Asmussen and Binswanger (1997). The output displayed below for all the algorithms except BGL was extracted from Blanchet and Li (2006). In each of the entries inside the table below the first number is the estimate and the second number is the estimated standard deviation using 20,000 samples. An approximate 95% confidence interval is also displayed.

It is interesting to remark some of the differences in performance observed in our experiments. As one can see, DLW's procedure yields a coefficient of variation that is 100 times lower than our proposed procedure. The reason for this performance is that DLW's takes advantage of both, the representation of the tail of the delay as the tail of the maximum of a random walk and the exponential tails (which allow to obtain a precise expression for the distribution of the first-ladder height). These two features, combined with regular variation, enable DLW to solve an optimization problem that allows to select the mixture parameters in order to reduce the coefficient of variation of the estimator. BL's implementation which also takes advantage of the representation of the tail of the maximum (although applies to much more general tails than just regularly varying) yields a coefficient of variation that is about 10 times lower than ours. BL's performance can also be explained (in addition to the advantageous use of the maximum representation) because it uses (as explained in Section 3) a more direct approximation to the zero-variance change-of-measure.

Acknowledgement The authors are grateful to the referees for their careful reading of this paper and their suggested improvements. This research was partially supported by NSF grant DMS 0595595.

[Estimation] [Std. Error] [Conf. Interval]	$x = 1000$	$x = 10^5$
BGL	$6.334e-05$ $2.259e-06$ $[5.891e-05, 6.776e-05]$	$6.129e-08$ $1.322e-09$ $[5.870e-08, 6.388e-08]$
BL	$6.341e-05$ $1.99e-07$ $[6.302e-05, 6.380e-05]$	$6.327e-08$ $2.01e-10$ $[6.323e-08, 6.331e-08]$
DLW	$6.341e-05$ $1.84e-08$ $[6.337e-05, 6.345e-05]$	$6.324e-08$ $2.26e-11$ $[6.32e-08, 6.329e-08]$
JS	$6.286e-05$ $3.429e-06$ $[5.601e-05, 6.972e-05]$	$5.864e-08$ $4.629e-09$ $[4.938e-08, 6.79e-08]$
AB	$6.35e-05$ $3.206e-05$ $[6.285e-05, 6.414e-05]$	$6.352e-08$ $3.199e-10$ $[6.288e-08, 6.416e-08]$

Table 1: Simulation Result

References

[1] Anantharam, V. (1989) How large delays build up in a GI/G/1 queue. *Queueing Systems: Theory and Applications*, Vol. 5, 345 -368.

[2] Asmussen, S. (2003) *Applied Probability and Queues*. Springer-Verlag. New York.

[3] Asmussen, S., and Binswanger, K. (1997) Simulation of ruin probabilities for subexponential claims. *Astin Bulletin* 27, 297-318.

[4] Asmussen, S., Binswanger, K., and Hojgaard, B. (2000) Rare event simulation for heavy-tailed distributions. *Bernoulli* 6, Vol. 2, 303-322.

[5] Asmussen, S. and Kroese, D. (2006) Improved algorithms for rare event simulation with heavy tails. *Advances in Applied Probability* Vol. 38, No. 2, 545-558.

[6] Blanchet, J. and Glynn, P. (2007) Efficient rare event simulation for the maximum of heavy-tailed random walks. Submitted.

[7] Blanchet, J., Glynn, P., and Liu, J. C. (2007) Efficient rare event simulation for multiserver queues. Preprint.

[8] Blanchet, J. and Li, C. (2006) Efficient rare event simulation for geometric sums. *Proc. of RESIM*, Bamberg.

[9] Bucklew, J. (2004) *Introduction to Rare-event Simulation.* Springer, New York.

[10] Dupuis, P. and Wang, H. (2004) Importance sampling, large deviations, and differential games. *Stochastics and Stochastics Reports* 76, 481-508.

[11] Dupuis, P., Leder, K., and Wang, H. (2006) Notes on importance sampling for random variables with regularly varying tails. Preprint.

[12] Dupuis, P., Sezer, A., and Wang, H. (2005) Importance sampling for tandem networks. Preprint.

[13] Foss, S., Konstantopoulos, T. and Zachary, S. (2007) Discrete and continuous time modulated random walks with heavy-tailed increments. Preprint.

[14] Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance.* Springer-Verlag. New York.

[15] Gut, A. (1988) *Stopped Random Walks: Limits Theorems and Applications.* Springer-Verlag, New York.

[16] Juneja, S. and Shahabuddin, P. (2002) Simulating heavy-tailed processes using delayed hazard rate twisting. *ACM TOMACS*, 12 - 2, p. 94 - 118.

[17] Juneja, S. and Shahabuddin, P. (2006) Rare event simulation techniques: An introduction and recent advances. *Handbook on Simulation.* Elsevier. Editors: Shane Henderson and Barry Nelson p. 291-350.

[18] Liu, J. (2001) *Monte Carlo Strategies in Scientific Computing.* Springer-Verlag, New York.

[19] Zachary, S. (2004) A note on Veraverbeke's theorem. *Queueing Systems: Theory and Applications.* Vol. 46, 9-14.

[20] Zwart, A. (2001) *Queueing Systems with Heavy Tails.* Ph.D. Dissertation.