Queueing Syst (2007) 57: 99–113
DOI 10.1007/s11134-007-9047-4

Fluid heuristics, Lyapunov bounds and efficient importance sampling for a heavy-tailed G/G/1 queue

J. Blanchet · P. Glynn · J.C. Liu

Published online: 14 November 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 constructed so as to be straightforward 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 particular to the delay in the $G/G/1$ queue. Hence, the strategy has the potential to be applied in more general queueing models.

Keywords State-dependent importance sampling · Rare-event simulation · Heavy-tails · Fluid heuristics · Lyapunov bounds · Single-server queue · Change-of-measure

Mathematics Subject Classification (2000) Primary 60G50 · 60J05 · 68W40 · Secondary 60G70 · 60J20

J. Blanchet (✉) · J.C. Liu
Science Center, Harvard University, 7th Floor, 1 Oxford St., Cambridge, MA 02138, USA
e-mail: blanchet@fas.harvard.edu

P. Glynn
Terman Engineering Center, Stanford University, 3rd Floor, 380 Panama Way, Stanford, CA 94305-4026, USA

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; an algorithm is said to be (strongly) efficient if it produces an estimator that has a bounded coefficient of variation uniformly in how far out into the tail the computation is done; the concept of strong efficiency is made precise in Definition 1. Efficient rare-event simulation algorithms for heavy-tailed $M/G/1$ queues have been developed by, for example [3–5, 12, 17]. All of these algorithms rely upon 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 [6] 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 [6] 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, variate generation from the Markov transition kernel associated with the algorithm of [6] 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". Mixture sampling means that the random variates required by the algorithm can be generated as a mixture (i.e. a convex combination) of distributions, each of which permits easy variate generation. A precise description of the mixture distribution is given in Sects. 3 and 4; see (5). Secondly, in contrast to [6], 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 [8].

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. This implies, in particular, that the overall estimator will typically be biased for any finite sample size. 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 will later 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 [6]. 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 [6] 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 similar to that proposed by [12]. 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 Sect. 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 of efficiency for biased rare-event simulation estimators. Section 3 further discusses the use of state-dependent importance sampling and describes an approach (based on Lyapunov functions) for verifying strong efficiency. Sections 4 and 5 discuss in detail the design of our algorithm and the verification of its efficiency. In particular, Sect. 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 Sect. 5, we prove that this 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 Sect. 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 $w_0 \in B_b^c$ with the property that W returns to w_0 infinitely often. Then, we have the following ratio formula for the steady-state distribution $\pi(\cdot)$:

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

where $T_{w_0} = \inf\{n \geq 1 : W_n = w_0\}$ (see, for example [2]). For the $G/G/1$ model, a particularly convenient choice for $\{w_0\}$ is $w_0 = 0$, so that regenerative cycles correspond to busy cycles.

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 Sect. 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 independent of b, we expect (and verify shortly) the overall efficiency of an algorithm for estimating $\pi(B_b)$ for b large using (1) to 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 previously studied rare-event simulation settings, the focus has been on unbiased estimators (i.e. $E\,Z_b = z_b$); see, for instance [9] or [18] for more on standard notions of efficiency in rare-event simulation. Note that Definition 1 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 roughly insensitive to b.

The importance sampling 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

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

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 $\overline{D}_n \geq 1$

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

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 to which the ratio estimator computes $\pi(B_b)$ is insensitive to b. As a consequence, we conclude that developing efficient rare-event simulation for $\pi(B_b)$ 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 $u_b^*(w) = P_w(T_b < T_{w_0})$, we note that $(u_b^*(w) : w \in S)$ solves the equation

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

subject to the boundary condition $u_b^*(w) = 1$ for $w \in B_b$ and $u_b^*(w) = 0$ 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

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

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

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

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

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

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 [6] (using for $v_b\,(\cdot)$ a classical heavy-tailed approximation that becomes asymptotically exact as $b \nearrow \infty$ and that is sometimes cited in the literature as the Pakes-Veraverbeke theorem; see, for instance [13]). However, one difficulty with this idea (that could be troublesome 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 under the new measure 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 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 [11]). 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 is standard in the rare-event analysis of heavy-tailed systems; see for instance, the paper of [1], 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-measures 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 $q_b(\cdot)$ 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 can then be simulated by exponential twisting in the light-tailed setting or 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 sampling distribution.

This key variate generation insight is due to [10]. 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 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 $(\beta(w) : w \in S)$) of corresponding transition kernels. Since the estimators considered are unbiased, (4) equivalently provides a minimum variance estimator among the class of importance sampling estimators indexed by $(\beta(w) : w \in S)$.

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

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

Given that we are ultimately interested in the efficiency of the waiting time sequence, we must also bound

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

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) $E_w q_b(\beta(w), w, W_1)\, h_b(W_1)\, I(W_1 \in w_0^c) \le h_b(w)$ *for* $w \in B_b^c$;
(ii) $h_b(w) \ge \varepsilon\chi(w)$ *for* $w \in B_b$.

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

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

$$\begin{aligned} M_n = &\prod_{j=1}^{T_b \wedge n} q_b\left(\beta\left(W_{j-1}\right), W_{j-1}, W_j\right) h_b\left(W_{T_b \wedge n}\right) \\ &\times I\left(T_{w_0} > T_b \wedge n\right). \end{aligned}$$

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

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

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

$$E\left(\prod_{j=1}^{T_b} q_b\left(\beta\left(W_{j-1}\right), W_{j-1}, W_j\right) h_b\left(W_{T_b}\right) I\left(T_{w_0} > T_b\right)\right) \leq E M_0 = h_b(w).$$

The previous inequality, combined with condition (ii), yields the statement of the result. □

We call the function h_b 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* $P_w(T_b < T_{w_0})$, *namely* $P_w(T_b < T_{w_0}) \leq (h_b(w)/\varepsilon)^{1/2}$.

Proof By Jensen's inequality,

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

□

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 $P_w(T_b < T_{w_0})^2$. Since we are developing our (hopefully) efficient change-of-measure so as to mimic the zero-variance Markovian conditional distribution, this suggests that $E_w^{Q_\beta} I(T_b < T_{w_0}) L_b^2$ should behave (roughly) like $P_w(T_b < T_{w_0})^2$. 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 $P_w(T_b < T_{w_0})$. 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 $P_w(T_b < T_{w_0})$). 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, [2, 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 and X_1 is assumed to have a continuous regularly varying density f with index $\alpha + 1 > 0$, so that

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

for $t > 0$, where the function $L(\cdot)$ is assumed to be slowly varying, i.e. $L(tm)/L(t) \longrightarrow 1$ as $t \nearrow \infty$ for each $m > 0$. In addition, we shall assume that $\mathrm{Var}(X) < \infty$. 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 [14, p. 567]), $\overline{F}(\cdot)$ is regularly varying with index α and $G(\cdot)$ is regularly varying with index $\alpha - 1$. Another property that we will frequently use is *longtailedness* (i.e. for any fixed $y \in (-\infty, \infty)$, $\overline{F}(t + y) \sim \overline{F}(t)$ as $t \nearrow \infty$. It is a well known fact that regularly varying functions are long tailed; see, for instance [14]).

Regarding the assumptions made above, the most important assumption concerns that of regular variation. Calculations analogous to those used in this paper suggest that a simple mixture of two components (i.e. a regular "jump" component and a "big jump" component) may not provide a rich enough family to contain a strongly efficient estimator for other types of heavy-tailed increment random variables (such as Weibull or lognormal random variables). 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. Turning to the remaining assumptions, it is important to explain the reason for requiring $\mathrm{Var}(X) < \infty$, given that the chain W may be stable even when the variance is infinite. Using the notation of Sect. 2, we put $B_b = [b, \infty)$ and set $w_0 = \{0\}$. As pointed out in Sect. 2, we will build our importance sampler for the tail probability $P(W_\infty > b)$ by constructing an efficient sampler for $\{T_b < T_{w_0}\}$. The difficulty that arises when $\mathrm{Var}(X) = \infty$ is that the conditional overshoot over the boundary b under the zero-variance change of measure for $\{T_b < T_{w_0}\}$ is asymptotically, as $b \nearrow \infty$, Pareto with index α if the X_k's are regularly varying (see [14, Appendix A]). This, in turn, leads to an infinite variance estimator for the numerator expectation of the regenerative ratio (which is clearly undesirable). This is a setting in which a good importance sampler for an expectation can not be reduced to the development of a good importance sampler for an associated rare-event probability (and hence requires a somewhat different theory). Finally, existence of a regularly varying density is a technical condition imposed to facilitate the handling of a Taylor expansion ((14) below) applied to our Lyapunov function. The regular variation on the density is imposed only in order to guarantee that $\overline{F}$ itself is regularly varying while the existence of a density is used to simplify an argument involving the second derivative of the Lyapunov function (which is defined directly in terms of $\overline{F}$ and hence f). The existence of a density could therefore be avoided (at a cost of additional notational and definitional burden) by using a twice continuously differentiable Lyapunov function having the same tail behavior as the current Lyapunov function.

Given the heavy tails that are present here, our discussion of Sect. 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

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

for $p, a \in (0,1)$ and $\kappa > 0$. A similar mixture form was introduced in [12] 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 make a state-independent (i.e., a constant). The parameter $\kappa > 0$ defines a boundary layer of the form $\{w : b - \kappa \le w \le b\}$ at which we "turn off" importance sampling and just make sure that we do not reach w_0 in the next step of the algorithm (i.e., we just keep the process alive). We can think of this boundary layer as a region where the occurrence of the event $\{T_b < T_{w_0}\}$ is no longer rare and therefore it is unnecessary to induce a large jump. 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 remaining parameters that make the change-of-measure efficient, we need to construct a convenient Lyapunov function. The rest of the section is devoted to this construction and is organized as follows. First, we will describe our proposed parametric family of Lyapunov functions as well as our proposed parametric family of changes-of-measure. We will then describe the joint constraints on the two sets of parameters that must be satisfied. These parametric families, as well as the joint constraints, must be obtained from a theoretical analysis of the Lyapunov inequality associated with Proposition 1. So, we conclude this section first with a heuristic analysis (based on fluid ideas) that motivate the parametric forms that we use, as well as the joint constraints that arise in satisfying the Lyapunov inequality, followed by a rigorous analysis that plugs the remaining theoretical holes in our heuristic development.

The proposed family of Lyapunov functions We will take h to be of the form $h(w) = h_0(w) \wedge 1$, where $h_0(w) = k \cdot v_b(w)^2$ and

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

Note that the definition of $h(\cdot)$ involves the parameters k, $d > 0$.

The proposed family of changes-of-measure As noted earlier, the change-of-measure depends on a and the mixture

probability $p(w)$ (which is state dependent). We propose a precise parametric form for $p(w)$, namely

$$p(w) = \theta \frac{\overline{F}(a(b-w))}{v_b(w)} I(b-w \geq \kappa)$$
$$= \theta \frac{\overline{F}(a(b-w))\, I(b-w \geq \kappa)}{G(b-w) - G(b+d)}. \quad (6)$$

Given (5), we therefore have three parameters (a, θ, and κ) that determine the change-of-measure.

Description of the constraints The various constraints on our parameters that arise through the theoretical analysis to be found later in this section are defined in terms of several constants that depend directly on $\overline{F}(\cdot)$. We start by recognizing that because X_1 is continuous with $EX_1 < 0$, there exist positive constants ε_0 and κ_1 such that

$$E(X; -\kappa_1 < X) \leq -\varepsilon_0 < 0 \quad (7)$$

and

$$\widetilde{\pi} \triangleq P(X > -\kappa_1) < 1.$$

In addition, Karamata's theorem implies that there exists $m_1 < \infty$ for which

$$\sup_{t \geq 0} \frac{\overline{F}(at)}{\overline{F}(t)} \leq m_1. \quad (8)$$

The constraints C1 to C4 below involve the constants ε_0, $\widetilde{\pi}$ and m_1 just defined. We can allow the parameter $a \in (0,1)$ to be chosen arbitrarily. Throughout the remainder of this paper, we assume that a has been so chosen. Furthermore, we assume that the two constants δ and $\widetilde{\delta}$ (appearing below) have been chosen from $(0,1)$ by the simulator. (Again, this choice can be arbitrary.)

C1 The parameter $\kappa > 0$ must satisfy

$$E\left(\frac{h_0(w+X)}{h_0(w)}; -w < X \leq a(b-w)\right)$$
$$\leq 1 - \frac{\partial_w h_0(w)}{h_0(w)} \varepsilon_0 (1-\delta). \quad (9)$$

for $\kappa_1 < w \leq b - \kappa$.

C2 The parameters θ and κ must satisfy

$$\theta \leq \varepsilon_0 (1-\delta)/(4m_1),$$
$$k \geq 2m_1/(\theta \varepsilon_0 (1-\delta)).$$

C3 The parameters d and κ must satisfy

$$d \geq (1+\widetilde{\delta})\left(\varepsilon_0(1-\delta) + 2EX^+\right)/(1-\widetilde{\pi}) \quad (10)$$

and

$$\frac{\partial_w h_0(w)}{h_0(w)} = \frac{2\overline{F}(b-w)}{G(b-w) - G(b+d)} \leq \frac{(1+\widetilde{\delta})}{w+d} \quad (11)$$

and also

$$E(h_0(w+X); -w < X \leq a(b-w))$$
$$\leq h_0(w)\widetilde{\pi} + \partial_w h_0(w) E\left(X^+\right), \quad (12)$$

for $w \leq \kappa_1$ and b sufficiently large.

C4 The parameter $\kappa > 0$ must be chosen so that $h_0(w) \geq 1$ when $w \leq b - \kappa$ (for sufficiently large b).

Feasibility of the constraints Note that because h_0 appears linearly in C1 and (12) in C3, the constant κ can be cancelled in both of these constraints. We therefore start by choosing d to satisfy (10), followed by choosing κ so as to satisfy both C1 and the inequalities (11) and (12) in C3 for b sufficiently large; Lemmas 1 and 2 below establish this fact. With κ and d so chosen, we now select θ to satisfy the first inequality in C2, followed by choosing k to satisfy the remaining inequality of C2 and also C4. At the conclusion of this process, we have a feasible set of parameters k, d, θ, and κ that satisfy C1 to C4 for sufficiently large values of b (i.e., for values of $b \geq b_0$ for some $b_0 > 0$). For $b \leq b_0$, we use no importance sampling in our simulation estimators (since the event $\{T_b < T_{w_0}\}$ is not rare in that case). It is only when $b > b_0$ that we use the change-of-measure described above (with parameters θ and κ as determined from C1 to C4).

Heuristic motivation Recall that in Sect. 3, we described our two step procedure for building efficient importance samplers. We follow this procedure to heuristically motivate the form of our candidate Lyapunov function, the parametric selection of $p(w)$ and the nature of the constraints C1 to C4.

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 [20, Chap. 2] and references therein. The previous approximation, in turn, suggests using a Lyapunov function such as $h(\cdot)$. The constraints on the parameters are obtained in the execution of Step 2. Before moving on to Step 2 and in order to enhance the intuition of the roles played by d and k, we point out that d is introduced to deal with boundary effects close to zero and k controls effects close to b.

Step 2 involves testing the Lyapunov bound. For this purpose, we define

$$J_1(w) = \frac{E(h(w+X); X > a(b-w))\overline{F}(a(b-w))}{h(w)\,p(w)},$$

$$J_2(w) = \frac{E(h(w+X); -w \le X \le a(b-w))P(X \in (-w, a(b-w)))}{h(w)(1-p(w))}$$

from which it follows easily that verifying the Lyapunov inequality from Proposition 1 is equivalent to showing that

$$J_1(w) + J_2(w) \le 1, \tag{13}$$

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. Let us define

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

It follows (since $h(\cdot)$ is absolutely continuous) that for all $y, w \in (-\infty, \infty)$ we have

$$h(y) - h(w) = \int_0^1 h'(w + (y-w)u)\,du$$

and therefore we can write

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

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

$$\begin{aligned} 1 \ge J_1 + J_2 &\approx \frac{\overline{F}(a(b-w))^2}{p(w)\,h(w)} + \frac{P(-w < X)^2}{1-p(w)} \\ &\quad + \frac{\partial_w h(w)}{h(w)} \frac{E(\Delta; -w < X)}{1-p(w)} \end{aligned} \tag{14}$$

when $b - w$ is large enough (making rigorous this part involves showing that condition C1 can be satisfied and this is done in Lemma 1 and Corollary 2 below). In addition, if $h(w) \le 1$, then

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

It seems natural, in order to cancel the squares in the previous expression to select $p(w)$ according to (6). (A more compelling way of motivating this selection of $p(w)$ is given in the next paragraph.) With this choice of p, (14) can be approximated as

$$\begin{aligned} &\frac{\overline{F}(a(b-w))}{\theta v_b(w)\,k} + \frac{P(-w < X)}{1-p(w)} \\ &\quad + 2\frac{\overline{F}(b-w)}{v_b(w)} \frac{E(\Delta; -w < X)}{1-p(w)}. \end{aligned} \tag{15}$$

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 have $E(\Delta; -w < X) < 0$ when w is bounded away from zero, it is clear that (15) can be upper bounded by one if we select first θ sufficiently small and then k large enough (or one can even pick $k = 1/\theta^2$ and θ sufficiently small; the details behind the selection of parameters in (15) relates to condition C2).

In order to provide further intuition into the choice of p given by (6), note that conditional on $\{W_n = w, T_b \wedge T_{w_0} > n\}$, the zero-variance choice for the probability $p(w)$ of hitting level b on transition $n+1$ would be to select it according to

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

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 $v_b(w)$ is a good approximation to $P_w(T_b < T_{w_0})$, the right-hand side behaves roughly like

$$\frac{\overline{F}(b-w)}{v_b(w)} = \frac{\partial_w v_b(w)}{v_b(w)} = \partial_w \log v_b(w). \tag{16}$$

But (16) is clearly consistent with (6). Hence, the form of $p(w)$ given by (16) and (15) 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 $w > \kappa_1$, $b - w \ge \kappa$, for some constants κ_1 and κ) and under the assumption that $h(w) < 1$. To handle the case in which $w \le \kappa_1$, we again use (15) and note that

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

Therefore, (15) is bounded (using $\Delta \leq X^+$) by a quantity that is roughly equal to

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

Recall that κ_1 is selected such that $\widetilde{\pi} = P\,(-\kappa_1 < X) < 1$. One can then select $d > 0$ large enough in order to make the previous quantity less or equal to 1. This approach, which is appropriate for dealing with the part of the state space close to the boundary at 0, is related to condition C3.

To deal with the boundary layer (i.e. 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 $\{w : h\,(w) = 1\}$, the analysis is simple. Indeed, 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. Constraint C4 just allows one to translate the condition $h_0\,(w) < 1$ in terms of a clearly defined spatial region where mixture importance sampling is not applied.

The rest of this section provides rigorous support for the above heuristic derivation. The bound involving J_1, namely

$$J_1 \leq \frac{\overline{F}\,(a\,(b-w))^2}{h_0\,(w)\,p\,(w)}, \tag{17}$$

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 now provide complete details for the Taylor expansion involved in the term J_2 of (14). First, define $m_+, \widetilde{m}_+, \widetilde{m}'_+ \in [1, \infty)$ such that

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

and set $m^* = (m_+ + m_+(\widetilde{m}_+ \vee \widetilde{m}'_+) + \widetilde{m}_+)EX^2$. The fact that $\widetilde{m}_+$ and $\widetilde{m}'_+$ are finite follows from Karamata's theorem (see [14, p. 567]) and the boundedness of $f\,(\cdot)$. The quantity m_+ is finite by definition of regular variation and because $\overline{F}\,(t) \in (0, 1)$ for all $t \geq 0$.

Lemma 1 *For each $\widetilde{\varepsilon} > 0$, there exists $\kappa > 0$ such that if $b - w \geq \kappa$ and $w \geq \kappa_1$, then*

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

Proof The absolute continuity of h_0 implies that

$$E\,(h_0\,(w+X)\,; -w < X \leq a\,(b-w)) = h_0\,(w)\,P\,(-w < X \leq a\,(b-w)) + E\,(\partial_w h_0\,(w+U\Delta)\,\Delta; -w < X \leq a\,(b-w))\,.$$

Now, observe that

$$\frac{\partial_w h_0\,(w+U\Delta)}{\partial_w h_0\,(w)} I\,(X \in (-w, a\,(b-w)]) = I\,(X \in (-w, a\,(b-w)]) \times \frac{G\,(b-w-U\Delta) - G\,(b+d)}{G\,(b-w) - G\,(b+d)} \times \frac{\overline{F}\,(b-w-U\Delta)}{\overline{F}\,(b-w)}. \tag{18}$$

Because G is also absolutely continuous, we obtain

$$G\,(b-w-U\Delta) = G\,(b-w) + E\left(\Delta U \overline{F}(b-w-U\cdot\widetilde{U}\cdot\Delta)|U,\Delta\right),$$

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

$$\frac{\Delta U E(\overline{F}(b-w-U\cdot\widetilde{U}\cdot\Delta)|U,\Delta)}{\overline{F}\,(b-w)} \leq X^+ \frac{\overline{F}\,((b-w)\,(1-a))}{\overline{F}\,(b-w)} \leq m_+ X^+.$$

Thus, we have that

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

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

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

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

$$E\left(\partial_w h_0(w + U\Delta)\Delta; 0 < X \leq a(b-w)\right)$$
$$\leq \partial_w h_0(w)\left(E\left(\Delta; 0 < X \leq a(b-w_1)\right)\right.$$
$$\left. + \varepsilon + \frac{\partial_w h_0(w)}{h_0(w)} m^*\right). \tag{19}$$

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

$$E\left(\frac{\partial_w h_0(w+U\Delta)}{\partial_w h_0(w)}\Delta; -w < X \leq 0\right)$$
$$\leq E\left(\frac{\partial_w h_0(w+U\Delta)}{\partial_w h_0(w)}\Delta; -\kappa_1 < X \leq 0\right)$$
$$\leq \frac{\overline{F}(b-w+\kappa_1)}{\overline{F}(b-w)} E(\Delta; -\kappa_1 < X \leq 0)$$
$$\leq E(\Delta; -\kappa_1 < X \leq 0)\left(1 + \frac{m'}{b-w+1}\right), \tag{20}$$

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 (20) combined with (19) yields the conclusion of the result. □

Lemma 1 can now be used to justify constraint C1 imposed on d and κ.

Corollary 2 *It is always possible to satisfy C*1 *by appropriately choosing d first and then* κ.

Proof First, select $\widetilde{\varepsilon} = \varepsilon_0\delta/2$ in Lemma 1. Then, in view of the selection of κ_1 and ε_0, 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,

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

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

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

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

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

The following result verifies the Lyapunov bound on the region $w \geq \kappa_1$ and $b - w \leq \kappa$.

Proposition 2 *Assume that b is large enough so that* C1 *and* C2 *are satisfied by our choice of parameters. 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

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

Since $\partial_w h/h > 0$, the selection of θ and k automatically implies 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 $\{w \le \kappa_1\}$. First, we show that constraint C3 can always be satisfied (simultaneously with C1 and C2).

Lemma 2 *The constraints imposed by* C1 *to* C3 *can always be jointly satisfied for b sufficiently large.*

Proof First, given the constraints imposed on $d > 0$ we note that constraint (11) is satisfied given a selection of $d > 0$ because

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

as $b \nearrow \infty$ uniformly over $0 \le w \le \kappa_1$. The fact that (12) is satisfiable follows directly from Lemma 1. □

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

Theorem 1 *If b is large enough so that constraints* C1 *to* C3 *are satisfied by our choice of parameters, we have that*

$$J_1 + J_2 \le 1$$

holds whenever $h_0(w) \le 1$.

Proof Under C1 to C3, we have that

$$J_1 + J_2 \le \frac{\partial_w h(w)}{h(w)} \frac{m_1}{\theta k} + \frac{\partial_w h(w)}{h(w)} 2m_1\theta\widetilde{\pi} + \widetilde{\pi} + 2\frac{\partial_w h(w)}{h(w)} E\left(X^+\right). \tag{21}$$

In addition, condition C3 also yields

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

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

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

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

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

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

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

then (22) yields the conclusion of the theorem. □

We close this section with the description of the algorithm suggested by the previous theorem for generating a single realization of the random variable L that enters our numerator estimator $\overline{C}_n^{(b)}$ introduced in Sect. 2.

Algorithm 1

Set $b \ge b_0$ 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 $\mathcal{L}(X \mid X \ge a(b-w))$, with probability $1-p$ sample X with law $\mathcal{L}(X \mid -w < X \le a(b-w))$. Then, update

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

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 Sect. 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

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

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

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

where L_b is the likelihood ratio obtained by running the importance sampling algorithm described in Sect. 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); an early reference to this idea, in the dependability modeling setting, is Goyal et al. [15]. Note that for the generation of $N_b(W_{T_b})$, one can apply additional variance reduction techniques, such as control variates.

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

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

where $E_0^{Q_{a,p}}(\cdot)$ denotes the expectation operator induced by the importance sampler selected in Sect. 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 [16, 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 Sect. 4. First define

$$\widetilde{J}_1 = \frac{E_w\left(\widetilde{e}_b(w+X); X \geq a(b-w)\right)\overline{F}(a(b-w))}{\widetilde{e}_b(w)\,p(w)},$$

$$\widetilde{J}_2 = \frac{E(\widetilde{e}_b(w+X); -w < X \leq a(b-w))P(X \in (-w, a(b-w)])}{\widetilde{e}_b(w)\,p(w)}.$$

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 Sect. 4. We just need to analyze $\widetilde{J}_1$ on $w < b$. Note that

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

It follows easily, using the facts that X is regularly varying and that $\mathrm{Var}(X) < \infty$, 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* $\mathrm{Var}(X) < \infty$. *Then,*

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

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

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

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

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

see, for instance [19, p. 2], where such lower bound is in fact developed in much greater generality, assuming only that X is long-tailed (see, for instance [14] 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 [2]). We conclude that

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

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 $E_0^{Q_{a,p}} T_{w_0} \leq \eta(1+b)$ for some $\eta > 0$ (a similar argument is given in Proposition 4 of [6]). This in turn implies that $EM(b) \leq \eta(b+1)$. □

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

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

Moreover, suppose that $\alpha = 5/2$ so that $EV = 1/(\alpha - 1) = 2/3$ 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 $\overline{F}(\cdot) = P(X > \cdot)$, is computed via

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

where

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

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 to high relative accuracy 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 (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

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

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 vari-

Table 1 Simulation result

[Estimation] [Std. Error] [Conf. Interval]	$x = 1000$ PV App.: 3.157e−05	$x = 10^5$ PV App.: 3.162e−08
BGL	3.167e−05 1.130e−06 [2.946e−05, 3.388e−05]	3.065e−08 6.610e−10 [2.935e−08, 3.194e−08]
BL	3.171e−05 9.95e−08 [3.151e−05, 3.190e−05]	3.164e−08 1.01e−10 [3.162e−08, 3.166e−08]
DLW	3.171e−05 9.2e−09 [3.169e−05, 3.173e−05]	3.162e−08 5.65e−12 [3.16e−08, 3.165e−08]
JS	3.143e−05 1.715e−06 2.801e−05, 3.486e−05]	2.932e−08 2.315e−09 [2.469e−08, 3.40e−08]
AK	3.174e−05 1.610e−07 [3.142e−05, 3.206e−05]	3.168e−08 1.599e−10 [3.137e−08, 3.199e−08]

ate 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 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 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. Table 1 summarizes the performance of our algorithm and several other methods. The entries corresponding to PV App. correspond to the Pakes-Veraverbeke heavy-tailed approximation for subexponential increment distribution, which estates that

$$P(W_\infty > b) \sim \frac{1}{-EX}\int_b^\infty P(X_1 > s)\,ds,$$

as $b \nearrow \infty$. BGL corresponds to our algorithm, BL is a variant of the algorithm proposed by [6] and which was adapted to the M/G/1 case in [7]. AK corresponds to the algorithm developed recently by [4] based on conditional Monte Carlo. DLW corresponds to the methods proposed by [12], JS corresponds to the hazard rate twisting procedure of [17]. The output displayed below for all the algorithms except BGL was extracted from [7]. In each of the entries within Table 1, 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 worth discussing 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 algorithm 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 allows one to obtain a precise expression for the distribution of the first-ladder height for the $M/G/1$ queue). These two features, combined with regular variation, enable DLW to solve an optimization problem that allows one to properly select the mixture parameters in order to reduce the coefficient of variation of the estimator. AK's and JS's procedures also use heavily the $M/G/1$ structure in the design of the algorithm. BL's implementation, which also takes advantage of the representation of the tail of the maximum (although it applies to much more general tails than just regularly varying ones), yields a coefficient of variation that is about 10 times lower than ours. BL's performance also relies (in addition to its advantageous use of the maximum representation) on a more direct approximation to the zero-variance change-of-measure (as explained in Sect. 3).

Acknowledgements 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.

References

1. Anantharam, V.: How large delays build up in a GI/G/1 queue. Queueing Syst. Theory Appl. **5**, 345–368 (1989)
2. Asmussen, S.: Applied Probability and Queues. Springer, New York (2003)
3. Asmussen, S., Binswanger, K.: Simulation of ruin probabilities for subexponential claims. Astin Bull. **27**, 297–318 (1997)
4. Asmussen, S., Kroese, D.: Improved algorithms for rare event simulation with heavy tails. Adv. Appl. Probab. **38**, 545–558 (2006)
5. Asmussen, S., Binswanger, K., Hojgaard, B.: Rare event simulation for heavy-tailed distributions. Bernoulli. **2**, 303–322 (2000)
6. Blanchet, J., Glynn, P.: Efficient rare event simulation for the maximum of heavy-tailed random walks. Ann. Appl. Probab. (2007, to appear). See http://www.imstat.org/aap/future_papers.html
7. Blanchet, J., Li, C.: Efficient rare event simulation for geometric sums. In: Proc. of RESIM, Bamberg (2006)
8. Blanchet, J., Glynn, P., Liu, J.C.: Efficient rare event simulation for multiserver queues. Preprint (2007)
9. Bucklew, J.: Introduction to Rare-event Simulation. Springer, New York (2004)
10. Dupuis, P., Wang, H.: Importance sampling, large deviations, and differential games. Stoch. Stoch. Rep. **76**, 481–508 (2004)
11. Dupuis, P., Sezer, A., Wang, H.: Importance sampling for tandem networks. Preprint (2005)
12. Dupuis, P., Leder, K., Wang, H.: Importance sampling for sums of random variables with regularly varying tails. TOMACS **17** (2006)
13. Foss, S., Konstantopoulos, T., Zachary, S.: Discrete and continuous time modulated random walks with heavy-tailed increments. Preprint (2007)
14. Embrechts, P., Klüppelberg, C., Mikosch, T.: Modelling Extremal Events for Insurance and Finance. Springer, New York (1997)
15. Goyal, A., Shahabuddin, P., Heidelberger, P., Nicola, V.F., Glynn, P.W.: A unified framework for simulating Markovian models of highly reliable systems. IEEE Trans. Comput. **41**, 36–51 (1992)
16. Gut, A.: Stopped Random Walks: Limits Theorems and Applications. Springer, New York (1988)
17. Juneja, S., Shahabuddin, P.: Simulating heavy-tailed processes using delayed hazard rate twisting. ACM TOMACS **12**, 94–118 (2002)
18. Juneja, S., Shahabuddin, P.: Rare event simulation techniques: an introduction and recent advances. In: Henderson, S., Nelson, B. (eds.) Handbook on Simulation, pp. 291–350. Elsevier, Amsterdam (2006)
19. Zachary, S.: A note on Veraverbeke's theorem. Queueing Syst. Theory Appl. **46**, 9–14 (2004)
20. Zwart, A.: Queueing Systems with Heavy Tails. Ph.D. Dissertation (2001)