Importance sampling algorithms for first passage time probabilities in the infinite server queue

Ad Ridder
Department Econometrics and Operations Research
Vrije University
de Boelelaan 1105, 1081 HV Amsterdam, Netherlands
email aridder@feweb.vu.nl

Abstract

This paper applies importance sampling simulation for estimating rare-event probabilities of the first passage time in the infinite server queue with renewal arrivals and general service time distributions. We consider importance sampling algorithms which are based on large deviations results of the infinite server queue, and we consider an algorithm based on the cross-entropy method, where we allow light-tailed and heavy-tailed distributions for the interarrival times and the service times. Efficiency of the algorithms is discussed by simulation experiments

Keywords: Simulation, Queueing, Rare Events, Importance Sampling.

1 Introduction

The infinite server queue, denoted by $G/G/\infty$, is a queueing model with infinitely many servers who are accessed by customers arriving according to a renewal process. The service times of customers are independent, identically distributed random variables, independent of the renewal arrival process. Customers leave the system after service. At time 0 we start with an empty system and then we like to find the probability distribution function of the first passage time of high levels.

Infinite server queues have been studied widely in the queueing literature because of their theoretical importance. However, they have also their practical usefulness, for instance to analyse service systems with a large number of servers such as call centers. The dynamics of the number of busy agents (the name for servers in a call

center) is equivalent to an infinite server queue as long as not all agents are busy. A remarkable application is the modelling of the software failure occurrence process as an infinite server queue (Dohi et al. 2002). When a software product has been developed it is checked for faults by inserting test cases at instants of a renewal process, usually a Poisson process. A test case consists of running a small process of the software product which ends after a random time V (independently of the other processes). Then one considers the number of incompleted processes to be the number of detected faults, this is equivalent to the infinite server queue.

In case of the $M/M/\infty$ model (Poisson arrivals and exponential service times) Keilson (1979, Chapter 5) derives the Laplace transform of the first passage time probability density function, and then it is possible to apply a numerical inversion algorithm. However, when the arrival process and/or the service times have other probability distributions, there are no computable expressions for the first passage time probabilities, and thus we might develop approximation algorithms, or, as we shall do in this paper, efficient simulation algorithms.

The notation in the general model is as follows: the interarrival times are i.i.d. random variables $U_1, U_2, \ldots$ with density function $f(x)$. The j-th renewal (arrival) occurs at time $A(j) = U_1 + \cdots + U_j$, and the number of renewals upto time s is denoted by $N(s)$. The service times are i.i.d. random variables $V_1, V_2, \ldots$ with density function $g(x)$. The cdf of the service time V is denoted by $G(x)$ and its associated complementary cdf by $\overline{G}(x) = 1 - G(x)$. The assumption is that the (generic) interarrival time U and service time V have finite means, and rates λ and μ, respectively. Finally, the cumulant generating function of a random variable X is defined to be $\psi_X(\theta) = \log E[\exp(\theta X)]$ for $\theta \in \mathbb{R}$ (if the expectation exists). We say that X has a heavy-tailed distribution if the cumulant generating function of X does not exist for positive θ. We allow both light-tailed and heavy-tailed interarrival and service times.

We consider a sequence of these infinite server queues, indexed by n: $\{Q_n(s) : s \geq 0\}$, $n = 1, 2, \ldots$, where $Q_n(s)$ represents the number of busy servers at time s in a $G/G/\infty$ model with interarrival times $U_1/n, U_2/n, \ldots$ and service times $V_1, V_2, \ldots$ (in which the (U_j) and (V_j) processes are as before). Notice that in the n-system the renewals occur at times $A_n(j) = A(j)/n$, and that the number of renewals upto time s is $N_n(s) = N(ns)$. Define the first passage times

$$T_n(j) = \inf\{s \geq 0 : Q_n(s) = j\} \quad (j = 1, 2, \ldots),$$

where $Q_n(0) = 0$. The problem of interest in this paper is

$$\ell_n = P\left(T_n(nx) \leq t\right),$$

for some specified time horizon $t > 0$ and large overflow level nx. We shall show in Section 3 that these probabilities decay exponentially fast to 0 as $n \to \infty$, and this says that we deal with rare events. Hence, when we would implement a standard Monte Carlo simulation algorithm for estimating these rare-event probabilities, the execution times will become too long to be practical for large n. Various variance reduction techniques exist to overcome this problem. In this paper we shall apply importance sampling. Let $\overline{Y}_n(k)$ be an unbiased estimator of ℓ_n under the original probability measure P based on k i.i.d. samples. In importance sampling we simulate under another probability measure, say $P^{\rm IS}$, such that the original measure P is absolutely continuous relative to this new measure. The new estimator $\overline{Y_n^*}(k)$ is again unbiased, i.e., $E^{\rm IS}[\overline{Y_n^*}(k)] = \ell_n$, if we incorporate the likelihood ratio $dP/dP^{\rm IS}$. (With the superscript IS we show explicitly that the expectation is taken w.r.t. measure $P^{\rm IS}$.)

Finding a good probability $P^{\rm IS}$ is the main issue in importance sampling. The criterion is to keep the relative error $\sqrt{{\rm Var}^{\rm IS}[\overline{Y_n^*}(k)]}/E^{\rm IS}[\overline{Y_n^*}(k)]$ as small as possible. The best performance is obtained when the relative error remains bounded as $n \to \infty$. Then the number of samples (simulation runs) required to achieve a fixed relative error is constant for all n. However, in practice this is difficult to find and the most frequently used criterion is asymptotical optimality (Bucklew 2004, Heidelberger 1995):

$$\lim_{n\to\infty} \frac{\log E^{\rm IS}[(\overline{Y_n^*}(k))^2]}{\log E^{\rm IS}[\overline{Y_n^*}(k)]} = 2. \tag{1}$$

Basically, it says that the relative error of the estimator $\overline{Y_n^*}(k)$ behaves as $\ell_n^{\epsilon_n}$, where $\epsilon_n = o(1)$ as $n \to \infty$. Consequently, since $\ell_n \to 0$ exponentially fast, the relative error might grow polynomially (or at some other subexponential rate), and thus also the sample sizes grow polynomially in order to obtain a prespecified relative error.

There is ample literature on importance sampling algorithms for efficient simulation in queueing models to estimate rare-event probabilities, see the overviews in Heidelberger (1995), Juneja and Shahabuddin (2006), and Blanchet and Mandjes (2007). The majority of these studies concerns blocking probabilities, buffer overflow (or level crossing) probabilities, tail probabilities, and waiting time tail probabilities. Initially these studies focused on static importance sampling algorithms with fixed tilted probability density functions for the arrival and service time processes to be used throughout the simulation runs (Heidelberger 1995). Since the first counter examples by Glasserman and Wang (1997) it became well known that static algorithms for many rare-event queueing problems cannot be asymptotically optimal, specifically in queueing network models and in queueing models with heavy-tailed

distributions. The focus of research shifted to adaptive algorithms in which during the simulation run the tilted probability density functions are updated based on current state or time. Asymptotically optimal state-dependent importance sampling algorithms have been developed for level-crossing probabilities in Jackson networks by Dupuis et al. (2007), and for waiting time tail probabilities in the single server queue with heavy-tailed distributed service times by Blanchet and Glynn (2008). Szechtman and Glynn (2002) considered a time-dependent importance sampling algorithm for the estimation of the tail probabilities $P(Q_n(t)/n \geq x)$ in the infinite server queue, and they showed asymptotical optimality.

The first passage time of a stochastic process to a barrier is an important issue in insurance and finance. For instance, equity default swaps are financial instruments whose buyers are compensated when some stock process hits a specified low boundary (Asmussen et al. 2008). Usually this targeted boundary is extremely deep which makes hitting the boundary during a certain time horizon a rare event. In the context of hitting large levels the first passage time problem received less attention in the queueing literature. The contribution of this paper has several aspects. In the first part of the paper we consider the $M/M/\infty$ model for which we construct a time-dependent importance sampling algorithm based on sample path large deviations results for the Erlang loss model (Shwartz and Weiss 1995, Chapter 12), and we prove its asymptotic optimality. This importance sampling algorithm simulates interarrival and service times from exponentially tilted distributions with time-dependent tilting parameters given by the optimal path to overflow.

In the second part we consider the general $G/G/\infty$ model and we present three importance sampling algorithms for the first passage time problem. All three are versions or adaptations of existing methods, and they are all time-dependent. These algorithms are investigated empirically by executing simulation experiments.

- An adaptation of the Szechtman and Glynn algorithm. The adaptation was needed to simulate service times whereas in the original algorithm it sufficed to simulate whether an arriving customer would still be present at time t. We allow light- and heavy-tailed distributions for both interarrival and service times.

- An adaptation of the $M/M/\infty$ algorithm. In the adapted version there is no resampling of scheduled event times after a new event as in the $M/M/\infty$ algorithm. The service times must have light-tailed distributions.

- A version of the cross-entropy algorithm introduced in Rubinstein and Kroese

(2004). We allow light- and heavy-tailed distributions for both interarrival and service times.

The paper is organised as follows. First we discuss the $M/M/\infty$ model for which we can implement an asymptotically optimal importance sampling algorithm (Section 2.1). In Section 3 we show the large deviations limit in the general $G/G/\infty$ model, and we sketch the importance sampling algorithm of Szechtman and Glynn (2002) for the tail probabilities. We present the three algorithms for the general model in Section 4, and in Section 5 we consider the cross-entropy based algorithm in case of Pareto distributed service times. Simulation results are given and discussed in Section 6.

2 The $M/M/\infty$ model

When the arrival process is Poisson and the service times are exponentially distributed, the process of the number of busy servers in the n-system $(Q_n(s))_{s\geq 0}$ is a continuous-time Markov chain (CTMC). Scaling the process by n we get a CTMC with jump rate $n\lambda$ in the jump direction $1/n$, and with jump rate $nq\mu$ in the jump direction $-1/n$ if $Q_n(s)/n = q$. For such processes the chapters 5 and 12 in Shwartz and Weiss (1995) develop sample path large deviations which we shall summarise here.

1. Starting point is the Cramér large deviations for Poisson random variables: let $X_1, X_2, \ldots$ be i.i.d. Poisson-λ random variables with cumulant generating function
$$\psi_X(\theta) = \log E[\exp(\theta X_1)] = \lambda(e^\theta - 1) \quad (\theta \in \mathbb{R}),$$
and associated Legendre-Fenchel transform $I(a) = \sup_\theta(\theta a - \psi_X(\theta))$, and let $Y_n = \sum_{k=1}^n X_k$. Then Y_n is a Poisson-$n\lambda$ random variable, for which $\lim_{n\to\infty} \frac{1}{n} \log P(Y_n/n \geq a) = -I(a)$.

2. Item 1 is generalised to the $(Q_n(s)/n)$ process where jumps are governed by two independent Poisson random variables with rates $n\lambda$ (for the $+1$ jump) and $nq\mu$ (for the -1 jump when the current state is $Q_n(s)/n = q$). Thus, the corresponding cumulant generating function is
$$\psi(\theta, q) = \lambda\left(e^\theta - 1\right) + q\mu\left(e^{-\theta} - 1\right) \quad (\theta \in \mathbb{R}). \tag{2}$$

3. The local rate function is the Legendre-Fenchel transform of $\psi(\theta, q)$ in the direction y, defined formally by

$$I(q, y) = \sup_{\theta \in \mathbb{R}} \left(\theta y - \psi(\theta, q)\right). \tag{3}$$

It is an exercise (Exercise 12.3 in Shwartz and Weiss (1995)) to get that the optimising θ satisfies

$$e^{\theta} = \frac{y + \sqrt{y^2 + 4\lambda q \mu}}{2\lambda}. \tag{4}$$

This θ is called the tilting parameter.

4. There exists a nonnegative absolute continuous function q_∞ on $[0, t]$ such that for any $\epsilon > 0$

$$\lim_{n \to \infty} P\left(\sup_{0 \le s \le t} |Q_n(s)/n - q_\infty(s)| < \epsilon\right) = 1. \tag{5}$$

The function q_∞ is usually refered to as the most likely path of the process because almost all sample paths of the scaled process are close to it (in the sup norm). Identification of this path reveals that it does not reach the target level x when $x > \lambda/\mu$ (page 291 in Shwartz and Weiss (1995)).

5. Sample path large deviations for absolute continuous functions ϕ hold:

$$\lim_{\epsilon \downarrow 0} \lim_{n \to \infty} \frac{1}{n} \log P\left(\sup_{0 \le s \le t} |Q_n(s)/n - \phi(s)| < \epsilon\right) = -J_1(\phi),$$

where the functional J_1 satisfies

$$J_1(\phi) = \int_0^t I\left(\phi(s), \phi'(s)\right)\, ds. \tag{6}$$

Thus, sample path large deviations are limiting logarithmic expressions for probabilities that sample paths stay close to some given function ϕ. The functional J_1 is called the large deviations rate function.

For our purposes we consider the set Φ of all functions ϕ that reach the target level x before (or at) time t starting from $\phi(0) = 0$. We have according to Theorem 12.18 in Shwartz and Weiss (1995)

$$\lim_{n \to \infty} \frac{1}{n} \log P\left((Q_n(s)/n)_{0 \le s \le t} \in \Phi\right) = -\inf\{\, J_1(\phi) : \phi \in \Phi \,\}.$$

Corollary 1. *Let* $\phi^* = \arg\min\{\, J_1(\phi) : \phi \in \Phi \,\}$, *then*

$$\lim_{n\to\infty} \frac{1}{n} \log \ell_n = -J_1(\phi^*). \tag{7}$$

Proof. In Mandjes and Ridder (2001) we have shown that there is a unique ϕ^* that minimises J_1 on Φ. Then (7) follows immediately by the principle of the largest term (Dembo and Zeitouni, 1998, Lemma 1.2.15):

$$\lim_{n\to\infty} \frac{1}{n} \log \ell_n = \lim_{n\to\infty} \frac{1}{n} \log P\left((Q_n(s)/n)_{0\le s\le t} \in \Phi\right) = -J_1(\phi^*).$$

□

The minimiser ϕ^* is called commonly the optimal path to overflow. For ease of notation we drop the $*$ to indicate this optimal path since it will be the only path we will consider in the rest of this section. In Mandjes and Ridder (2001) we found its expression:

$$\phi(s) = \frac{c}{\mu}\left(e^{\mu s} - 1\right) + \frac{\lambda}{\mu}\left(1 - e^{-\mu s}\right), \quad 0 \le s \le t, \tag{8}$$

with the constant c obtained by substituting $\phi(t) = x$. To get the large deviations rate $J_1(\phi)$ we need to determine the tilting parameters (4) along the path by substituting $q = \phi(s)$ and $y = \phi'(s)$. Therefore we deal with a tilting function $\theta(s)$, $0 \le s \le t$.

The optimal path and its associated tilting function are plotted in Figure 1.

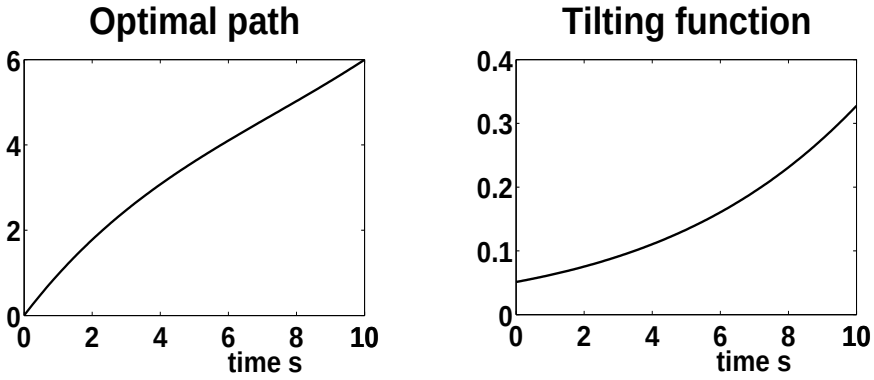

Figure 1. The optimal path ϕ *and the tilting function* θ *for the case* $\lambda = 1, \mu = 0.2, x = 6, t = 10$.

We construct an importance sampling algorithm by exponentially tilting interarrival and service time distributions using the tilting function $\theta(s)$ along the optimal path to overflow.

Algorithm 1.

1. *Compute the tilting function* $\theta(s)$ *in* (4) *where the function* $\phi(s)$ *is given in* (8).

2. *Simulate a sample path of* $Q_n(s), s \geq 0$ *starting at* $Q_n(0) = 0$ *from event to event until either time horizon* t *has reached, or* $Q_n(s) \geq nx$*, whatever comes first.*

3. *Let* s *be an arrival epoch. Then the next interarrival time is set* U/n *with* U *drawn from an exponential distribution with rate* $\lambda e^{\theta(s)}$*. The service times of all customers (present and just arrived) are rescheduled and drawn independently from an exponential distribution with rate* $\mu\, e^{-\theta(s)}$*.*

4. *Let* s *be a departure epoch. Then the ongoing interarrival time is rescheduled to become* U/n *with* U *drawn from an exponential distribution with rate* $\lambda e^{\theta}(s)$*. And the service times of all present customers are rescheduled and drawn independently from an exponential distribution with rate* $\mu\, e^{-\theta(s)}$*.*

2.1 Proof of asymptotic optimality

In this section we shall prove that Algorithm 1 gives an asymptotically optimal estimator. The theorem follows after five lemmas.

Lemma 1. *Recall the tilting function* $\theta(s)$ *in* (4) *and the optimal path* $\phi(s)$ *to overflow in* (8)*. For all* $0 \leq s \leq t$

$$\lambda \left(e^{\theta(s)} - 1 \right) + \phi(s)\mu \left(e^{-\theta(s)} - 1 \right) = \text{constant},$$

with the constant equal to the c *in* (8)*.*

The proof of Lemma 1 is in Appendix A.

Lemma 2. *The large deviations rate function* $J_1(\phi)$ *of* (6) *satisfies*

$$J_1(\phi) = \int_0^t \theta(s)\phi'(s)\, ds - ct.$$

The proof of Lemma 2 is in Appendix A.

Consider a random realisation of the process $(Q_n(s))_{s\geq 0}$: the consecutive jump times are $S_0 = 0 < S_1 < S_2 < \cdots < S_{M_n}$ with associated states $Q_n(S_j)$, $j = 0, 1, \ldots, M_n$, where the last jump brings either the state above nx or the time beyond t (whatever comes first).

Lemma 3. *Define*

$$\begin{aligned} A_n &= \sum_{j=0}^{M_n-1} (S_{j+1} - S_j)\Big(\lambda\left(e^{\theta(S_j)} - 1\right) + \frac{Q_n(S_j)}{n}\mu\left(e^{-\theta(S_j)} - 1\right)\Big) \\ B_n &= \sum_{j=0}^{M_n-1} \theta(S_j)\,\frac{Q_n(S_{j+1}) - Q_n(S_j)}{n}. \end{aligned} \tag{9}$$

Under the importance sampling measure P^{IS} we have for any $\epsilon > 0$:

$$\lim_{n\to\infty} P^{\mathrm{IS}}\left(\,|A_n - ct| < \epsilon\right) = 1,$$

and

$$\lim_{n\to\infty} P^{\mathrm{IS}}\left(\left|B_n - \int_0^t \theta(s)\phi'(s)\,ds\right| < \epsilon\right) = 1.$$

Proof. Recall the most likely path (5) of the scaled process $(Q_n(s)/n)_{s\geq 0}$. The same arguments apply under the importance sampling probability measure P^{IS} with the transition rates given in Algorithm 1:

$$\lim_{n\to\infty} P^{\mathrm{IS}}\left(\sup_{0\leq s\leq t} |Q_n(s)/n - q_\infty^{\mathrm{IS}}(s)| < \epsilon\right) = 1. \tag{10}$$

The most likely path q_∞^{IS} is identified by solving its associated differential equation (see Shwartz and Weiss, 1995, Section 5.1). After doing the calculus we find $q_\infty^{\mathrm{IS}} = \phi$, i.e., the optimal path to overflow under the original probability measure P equals the most likely path under the importance sampling probability measure P^{IS}.

By Lemma 1 we have for any jump time S_j

$$\lambda\left(e^{\theta(S_j)} - 1\right) = c - \phi(S_j)\mu\left(e^{-\theta(S_j)} - 1\right).$$

When we substitute this in A_n we get

$$A_n = \sum_{j=0}^{M_n-1} (S_{j+1} - S_j)c + \sum_{j=0}^{M_n-1} (S_{j+1} - S_j)\left(\frac{Q_n(S_j)}{n} - \phi(S_j)\right)\mu\left(e^{-\theta(S_j)} - 1\right). \tag{11}$$

We analyse the two summations seperately. Clearly, when the rare event does not occur, the last jump is just after t, arbitrary small as $n \to \infty$:

$$\lim_{n\to\infty} P^{\mathrm{IS}}(|S_{M_n} - t| < \epsilon \,|\, T_n(nx) > t) = 1.$$

However, the same holds given that the rare event occurs ($T_n(nx) \leq t$), because the scaled process $\{Q_n(s)/n\}$ is close to the optimal path ϕ which is increasing to x,

i.e., $\phi(s) \uparrow x$ as $s \uparrow t$ which can be seen easily from the explicit formula for ϕ in (8). Thus, $S_{M_n} \overset{P^{\mathrm{IS}}}{\to} t$, and hence

$$\sum_{j=0}^{M_n-1} (S_{j+1} - S_j)c = S_{M_n} c \overset{P^{\mathrm{IS}}}{\to} ct \quad (\text{as } n \to \infty). \tag{12}$$

The intervals between jumps $[0, S_1), [S_1, S_2), \ldots$ form a partitition of $[0, t]$, with interval lengths $S_{j+1} - S_j \to 0$ (in probability). Thus the second summation of (11) can be considered to be a Riemann sum approximation of the integral

$$\int_0^t \left(\frac{Q_n(s)}{n} - \phi(s) \right) \mu \left(e^{-\theta(s)} - 1 \right) ds.$$

This integral is in absolute value less than

$$\int_0^t \left| \frac{Q_n(s)}{n} - \phi(s) \right| \mu \, ds \le \mu t \sup_{0 \le s \le t} \left| \frac{Q_n(s)}{n} - \phi(s) \right|.$$

Hence, with (10) we get that the second part in (11) convergences (in P^{IS}) to 0, and with (12) we have shown the first statement of the lemma.

For the second statement we shall argue that $\{(Q_n(s+h) - Q_n(s))/nh : 0 \le s \le t\}$ converges in probability to ϕ' as $n \to \infty$ and $h \to 0$:

$$\begin{aligned}
&\sup_{0 \le s \le t} \left| \frac{Q_n(s+h)/n - Q_n(s)/n}{h} - \phi'(s) \right| \\
&\qquad \le \sup_{0 \le s \le t} \left| \frac{Q_n(s+h)/n - Q_n(s)/n}{h} - \frac{\phi(s+h) - \phi(s)}{h} \right| + \sup_{0 \le s \le t} \left| \frac{\phi(s+h) - \phi(s)}{h} - \phi'(s) \right|.
\end{aligned}$$

The first sup-term goes to 0 (in P^{IS}) as $n \to \infty$ for any $h \neq 0$. The second sup-term goes to 0 as $h \to 0$.

Next we reason that

$$\begin{aligned}
B_n &= \sum_{j=0}^{M_n-1} \theta(S_j) \, \frac{Q_n(S_{j+1}) - Q_n(S_j)}{n} \\
&= \sum_{j=0}^{M_n-1} (S_{j+1} - S_j)\theta(S_j) \, \frac{Q_n(S_{j+1})/n - Q_n(S_j)/n}{S_{j+1} - S_j} \\
&= \sum_{j=0}^{M_n-1} (S_{j+1} - S_j)\theta(S_j) \, \left(\phi'(S_j) + \Delta_j \right),
\end{aligned} \tag{13}$$

where $\sup_j |\Delta_j| \overset{P^{\mathrm{IS}}}{\to} 0$ as $n \to \infty$. The tilting function $\theta(s)$ is positive and increasing in s (calculus with its expression (4)), thus we can bound $\theta(s)$ for all s by $\theta(t)$ to obtain

$$\left| \sum_{j=0}^{M_n-1} (S_{j+1} - S_j)\theta(S_j)\Delta_j \right| \leq \theta(t) \left(\sup_j |\Delta_j| \right) \sum_{j=0}^{M_n-1} (S_{j+1} - S_j)$$
$$= \theta(t) \sup_j |\Delta_j| S_{M_n} \overset{P^{\mathrm{IS}}}{\to} 0.$$

The remaining term in (13) can be considered to be a Riemann sum approximation of the integral

$$\int_0^t \theta(s)\phi'(s)\, ds.$$

This completes the proof. □

Suppose that an importance sampling simulation of the process $(Q_n(s))_{s\geq 0}$ is executed. Let $L_n = dP/dP^{\mathrm{IS}}$ be the likelihood ratio of a random realisation.

Lemma 4. *The likelihood ratio can be expressed as* $L_n = e^{n(A_n - B_n)}$ *where* A_n *and* B_n *are given in* (9).

Proof. For ease of notation we set $Q_j = Q_n(S_j)$ for the state of the process at the j-th jump time, $\sigma_j = \exp(\theta(S_j))$ for the exponent of the tilting function at the j-th jump time, and $X_j = S_{j+1} - S_j$ for the time between two consecutive jump times.

These interjump times X_j, $j = 0, 1, \ldots$ are independent exponentially distributed random variables (minimum of exponentials) with state dependent rates $n\lambda\sigma_j + Q_j\mu/\sigma_j$. The states $(Q_j)_{j=0,1,\ldots}$ form a Markov chain with ± 1 jumps and transition probability $n\lambda\sigma_j/(n\lambda\sigma_j + Q_j\mu/\sigma_j)$ for the $+1$ jump at time S_{j+1}, and its complement for the -1 jump. Let $\mathcal{A}$ be the set of all $+1$ jumps, and $\mathcal{D}$ be the set of all -1 jumps.

The likelihood ratio $L_n = dP/dP^{\mathrm{IS}}$ of a random realisation is:

$$\begin{aligned}
L_n &= \prod_{j=0}^{M_n-1} \frac{(n\lambda + Q_j\mu)\,\exp\Big(-(n\lambda+Q_j\mu)X_j\Big)}{(n\lambda\sigma_j + Q_j\mu/\sigma_j)\,\exp\Big(-(n\lambda\sigma_j+Q_j\mu/\sigma_j)X_j\Big)} \\
&\quad \times \prod_{j+1\in\mathcal{A}} \frac{n\lambda}{n\lambda+Q_j\mu}\,\frac{n\lambda\sigma_j+Q_j\mu/\sigma_j}{n\lambda\sigma_j} \times \prod_{j+1\in\mathcal{D}} \frac{Q_j\mu}{n\lambda+Q_j\mu}\,\frac{n\lambda\sigma_j+Q_j\mu/\sigma_j}{Q_j\mu/\sigma_j} \\
&= \prod_{j=0}^{M_n-1} \exp\Big(-X_j\Big(n\lambda(1-\sigma_j) + Q_j\mu(1-\sigma_j^{-1})\Big)\Big) \quad \times \prod_{j+1\in\mathcal{A}} \frac{1}{\sigma_j} \times \prod_{j+1\in\mathcal{D}} \sigma_j \\
&= \prod_{j=0}^{M_n-1} \exp\Big(-X_j\Big(n\lambda(1-\sigma_j) + Q_j\mu(1-\sigma_j^{-1})\Big)\Big) \quad \times \prod_{j=0}^{M_n-1} \exp\Big(-(Q_{j+1}-Q_j)\theta(S_j)\Big) \\
&= \exp\left(-\sum_{j=0}^{M_n-1} X_j\Big(n\lambda(1-\sigma_j) + Q_j\mu(1-\sigma_j^{-1})\Big)\right) \times \exp\Big(-\sum_{j=0}^{M_n-1}\theta(S_j)(Q_{j+1}-Q_j)\Big) \\
&= \exp\left(n\sum_{j=0}^{M_n-1}\left(X_j\Big(\lambda(\sigma_j-1) + \frac{Q_j}{n}\mu(\sigma_j^{-1}-1)\Big) - \theta(S_j)\,\frac{Q_{j+1}-Q_j}{n}\right)\right) \\
&= e^{n(A_n-B_n)}.
\end{aligned}$$

□

Let $Y_n = 1\{T_n(nx) \le t\}$ indicate the occurrence of the rare event.

Lemma 5.

$$\lim_{n\to\infty} \frac{1}{n} \log E^{\mathrm{IS}}[L_n^2 Y_n] = -2J_1(\phi).$$

Proof. Let $Z_n = A_n - B_n$. From Lemmas 2 and 3 we conclude that

$$Z_n \overset{P^{\mathrm{IS}}}{\to} ct - \int_0^t \theta(s)\phi'(s)\,ds = -J_1(\phi).$$

Repeating the details of these lemmas we would obtain that Z_n is uniformly bounded (in n) almost surely. Thus, $Z_n = -J_1(\phi) + \Delta_n$ with

$$|\Delta_n| \le K \quad \text{(a.s.)}, \quad \text{and} \quad \Delta_n \overset{P^{\mathrm{IS}}}{\to} 0.$$

Now,

$$\begin{aligned}
\frac{1}{n}\log E^{\mathrm{IS}}[L_n^2 Y_n] &= \frac{1}{n}\log E^{\mathrm{IS}}[L_n^2|Y_n=1]P^{\mathrm{IS}}(Y_n=1)\\
&= \frac{1}{n}\log E^{\mathrm{IS}}[\exp(2nZ_n)|Y_n=1] + \frac{1}{n}\log P^{\mathrm{IS}}(Y_n=1)\\
&= \frac{1}{n}\log e^{-2nJ(\phi)}E^{\mathrm{IS}}[\exp(2n\Delta_n)|Y_n=1] + \frac{1}{n}\log P^{\mathrm{IS}}(Y_n=1)\\
&= -2J_1(\phi) + \frac{1}{n}\log E^{\mathrm{IS}}[\exp(2n\Delta_n)|Y_n=1] + \frac{1}{n}\log P^{\mathrm{IS}}(Y_n=1)\\
&\to -2J_1(\phi) \quad (n\to\infty),
\end{aligned}$$

because the rare event will occur most likely under the importance sampling measure, i.e., $\lim_{n\to\infty} P^{\mathrm{IS}}(Y_n=1)=1$. □

Theorem 1. *The importance sampling estimator $\overline{Y^*}_n(k)$ obtained by repeating k times Algorithm 1 is asymptotically optimal.*

Proof. It suffices to show asymptotic optimality for the one-run (unbiased) estimator $Y_n^* = L_n Y_n$. And this follows immediately from Lemma 5 and the large deviations result for the rare event probability $\lim_{n\to\infty}\frac{1}{n}\log\ell_n = -J_1(\phi)$:

$$\frac{\log E^{\mathrm{IS}}[L_n^2 Y_n]}{\log E^{\mathrm{IS}}[L_n Y_n]} = \frac{\frac{1}{n}\log E^{\mathrm{IS}}[L_n^2 Y_n]}{\frac{1}{n}\log\ell_n} \to \frac{-2J_1(\phi)}{-J_1(\phi)} = 2.$$

□

3 The general model

The general model comprises a renewal process for arrivals and i.i.d. service times. We refer to Glynn (1995) for the following results concerning the sequence of scaled variables $(Q_n(t)/n)_{n=1}^{\infty}$ at the horizon t which is taken fixed (recall that always $Q_n(0)=0$).

1. The limiting cumulant generating function satisfies

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{n}\log E\left[e^{\theta Q_n(t)}\right] &= \psi_Q(\theta,t)\\
&= -\int_0^t \psi_U^{-1}\left(-\log\left(e^{\theta}\overline{G}(s)+G(s)\right)\right)ds \quad (\theta\in\mathbb{R}).
\end{aligned}$$

   Actually, this limit behaviour is found in two steps: in the first step Glynn (1995) finds the moment generating function as an integral w.r.t. the renewal

process:

$$E\left[e^{\theta Q_n(t)}\right] = E\left[\exp\Big(\int_{[0,t]} \log\left(e^{\theta}\overline{G}(t-s) + G(t-s)\right) N_n(ds)\Big)\right].$$

In the second step Glynn (1995) shows the following limit by bounding the expression of the right hand side

$$\lim_{n\to\infty} \frac{1}{n} \log E\left[e^{\theta Q_n(t)}\right] = \int_0^t \psi_N\left(\log\left(e^{\theta}\overline{G}(s) + G(s)\right)\right) ds, \tag{14}$$

where $\psi_N(\theta) = -\psi_U^{-1}(-\theta)$ in case of a renewal arrival process.

2. The Legendre-Fenchel transform of $\psi_Q(\theta, t)$ is

$$J_2(t,x) = \sup_{\theta\in\mathbb{R}} \left(\theta x - \psi_Q(\theta,t)\right). \tag{15}$$

The optimising θ^* in (15) is the positive root of

$$\frac{\partial}{\partial\theta}\,\psi_Q(\theta,t) = x, \tag{16}$$

and is called the optimal tilting parameter.

3. The sequence of scaled variables $(Q_n(t)/n)$ converges in probability to its 'most likely' mean $m(t) = E[Q_n(t)/n] = \lambda\int_0^t \overline{G}(s)\,ds$:

$$\lim_{n\to\infty} P\left(|Q_n(t)/n - m(t)| < \epsilon\right) = 1. \tag{17}$$

4. A large deviations for the sequence of scaled variables $(Q_n(t)/n)$ holds:

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\, Q_n(t)/n \geq x \,\right) = -J_2(t,x) \quad (x > m(t)). \tag{18}$$

The limits in (14), (17) and (18) exist for the queueing processes we consider in this paper, see Glynn (1995) for the exact formulation of the conditions. We need the following lemma before proving the large deviations for the first passage time probabilities $\ell_n = P(T_n(nx) \leq t)$.

Lemma 6. *The rate function $J_2(t,x)$ is decreasing in t.*

Proof. Denote the optimal tilting parameter θ^* as $\theta(t)$, and recall that x is a constant.

$$\begin{aligned}
\frac{d}{dt}J_2(t,x) &= \theta'(t)x - \frac{d}{dt}\psi_Q(\theta(t),t) \\
&= \theta'(t)x - \frac{\partial}{\partial\theta}\psi_Q(\theta,t)|_{\theta=\theta(t)}\,\theta'(t) - \frac{\partial}{\partial t}\psi_Q(\theta,t)|_{\theta=\theta(t)} \\
&= \theta'(t)\Big(x - \frac{\partial}{\partial\theta}\psi_Q(\theta,t)|_{\theta=\theta(t)}\Big) - \frac{\partial}{\partial t}\psi_Q(\theta,t)|_{\theta=\theta(t)} \\
&\overset{\text{(a)}}{=} -\frac{\partial}{\partial t}\psi_Q(\theta,t)|_{\theta=\theta(t)} \\
&= \frac{d}{dt}\int_0^t \psi_U^{-1}\left(-\log\left(e^{\theta}\overline{G}(s) + G(s)\right)\right)\,ds|_{\theta=\theta(t)} \\
&= \psi_U^{-1}\left(-\log\left(e^{\theta(t)}\overline{G}(t) + G(t)\right)\right) < 0.
\end{aligned}$$

The equality (a) follows from (16), and the final expression is negative because

$$\begin{aligned}
\theta(t) > 0 \;&\Rightarrow\; e^{\theta(t)}\overline{G}(t) + G(t) > 1 \\
&\Rightarrow\; -\log\left(e^{\theta(t)}\overline{G}(t) + G(t)\right) < 0,
\end{aligned}$$

and because interarrival time U is a positive random variable. □

Theorem 2.

$$\lim_{n\to\infty}\frac{1}{n}\log\ell_n = -J_2(t,x).$$

Proof. Apply the principle of the largest term (Dembo and Zeitouni, 1998, Lemma 1.2.15):

$$\lim_{n\to\infty}\frac{1}{n}\log\ell_n = \lim_{n\to\infty}\frac{1}{n}\log P\left(\bigcup_{s\le t}\{Q_n(s)\ge nx\}\right) = -\inf_{s\le t}J_2(s,x) = -J_2(t,x).$$

□

In case of the $M/M/\infty$ model we gave in Corollary 1 the sample path large deviations rate of the ℓ_n as $J_1(\phi)$ with ϕ the optimal path to overflow. Clearly it must hold that $J_2(t,x) = J_1(\phi)$. This equality of the large deviations rate functions follows also by working out their expressions.

We sketch the algorithm of Szechtman and Glynn (2002) for estimating the tail probability $P\,(\,Q_n(t)\ge nx\,)$ in the general model.

Algorithm 2.

1. *Compute the optimal tilting parameter* θ^* *in* (16).

2. *Simulate a sample path of* $Q_n(s), 0 \leq s \leq t$ *starting at* $Q_n(0) = 0$ *from arrival epoch to the next arrival epoch along the following two steps.*

3. *Let* s *be an arrival epoch. Then the next interarrival time is set* U/n *with* U *drawn from an exponentially tilted distribution with probability density function*

$$f^{\alpha}(u) = \exp\left(\alpha u - \psi_U(\alpha)\right) f(u), \tag{19}$$

where the tilting parameter $\alpha = \alpha(s)$ *is time dependent and solves*

$$\psi_U(\alpha) = -\log\Big(e^{\theta^*}\overline{G}(t-s) + G(t-s)\Big). \tag{20}$$

4. *Let* s *be an arrival epoch. Then the arriving customer has a service time that takes longer than* $t - s$ *with probability*

$$\frac{e^{\theta^*}\overline{G}(t-s)}{e^{\theta^*}\overline{G}(t-s) + G(t-s)}.$$

5. *When horizon* t *has been reached, check whether* $Q_n(t) \geq nx$.

Notice that we do not have to simulate the service times but only whether or not an arriving customer is still present at time t. Also notice that because $\theta^* > 0$, the righthand side in (20) is negative, and thus this equation is also solvable for heavy-tailed interarrival times. In that case the solution to (20) is a negative α for which $\psi_U(\alpha) < \infty$. In other words, Algorithm 2 is feasible for light-tailed and heavy-tailed interarrival and service times.

4 Importance sampling algorithms

In this section we present the three importance algorithms for the general $G/G/\infty$ queue. The discussion on their efficiencies is deferred to Section 6.

4.1 Adapted Szechtman-Glynn algorithm

The first algorithm for the general model is an adaptation of Algorithm 2, which cannot be applied directly for estimating the first passage time probabilities because

it gives information on the number of busy servers only at the horizon time t. Thus it cannot decide whether the target level nx might have been hit before t.

The adapted version is a classical discrete-event simulation of the queue by simulating a sample path from event to event, where events are arrivals and departures. The interarrival times are similar as in step 3 of Algorithm 2. The service time V for the customer arriving at time s has a cdf $G^*(v)$ that we define by its complementary

$$\overline{G^*}(v) = \frac{e^{\theta^*}\overline{G}(v)}{e^{\theta^*}\overline{G}(v) + G(v)}. \tag{21}$$

Sampling from G^* is done as in the inverse transform method, by generating y from uniform $U(0,1)$ and solving $\overline{G^*}(v) = y$ (for v). After rewriting we obtain that we have to solve (for v)

$$\overline{G}(v) = \frac{y}{e^{\theta^*}(1-y) + y}.$$

The algorithm follows in detail.

Algorithm 3. [SG]

1. *Compute the optimal tilting parameter θ^* in* (16).
2. *Simulate a sample path of $Q_n(s), s \geq 0$ starting at $Q_n(0) = 0$ from event to event until either time horizon t has reached, or $Q_n(s) \geq nx$, whatever comes first.*
3. *Let s be an arrival epoch. Then the next interarrival time is set U/n with U drawn from the distribution with probability density function given in* (19) *in Algorithm 2.*
4. *Let s be an arrival epoch. Then the arriving customer receives service time V drawn from cdf G^* given above in* (21).
5. *No action, i.e., no rescheduling, is taken after a departure event.*

Notice that a customer arriving at time s is still present at time t with probability

$$P^{\mathrm{IS}}(V > t - s) = \overline{G^*}(t-s) = \frac{e^{\theta^*}\overline{G}(t-s)}{e^{\theta^*}\overline{G}(t-s) + G(t-s)},$$

which coincides with step 4 in Algorithm 2. For the same reasons as we mentioned with respect to Algorithm 2, Algorithm 3 is feasible for light-tailed and heavy-tailed interarrival and service times.

4.2 Adapted $M/M/\infty$ algorithms

The second algorithm in this section uses the optimal tilting parameters of Algorithm 1 for the $M/M/\infty$ model. This is based on the intuition that in the long-run the stationary distributon in the $M/G/\infty$ model is insensitive for the higher moments of the service duration. In the $M/M/\infty$ model, the interarrival time and all service times are rescheduled after each event, however, this is not feasible to execute in the general model, because there is no memoryless property.

Algorithm 4. [MM]

1. *Compute the tilting function $\theta(s)$ in* (4).

2. *Simulate a sample path of $Q_n(s), s \geq 0$ starting at $Q_n(0) = 0$ from event to event until either time horizon t has reached, or $Q_n(s) \geq nx$, whatever comes first.*

3. *Let s be an arrival epoch. Then the next interarrival time is set U/n with U drawn from an exponentially tilted distribution with rate $\lambda e^{\theta(s)}$, i.e., the density function is*
$$f^{\alpha}(u) = \exp\left(\alpha u - \psi_U(\alpha)\right) f(u), \tag{22}$$
and the tilting parameter α solves $\psi'_U(\alpha) = e^{-\theta(s)}/\lambda$.

4. *Let s be an arrival epoch. Then the arriving customer receives a service time V drawn from an exponentially tilted distribution with rate $\mu e^{-\theta(s)}$, i.e., the density function is*
$$g^{\beta}(v) = \exp\left(\beta v - \psi_V(\beta)\right) g(v), \tag{23}$$
and the tilting parameter β solves $\psi'_V(\beta) = e^{\theta(s)}/\mu$.

5. *No action, i.e., no rescheduling, is taken after a departure event.*

Here we notice that the tilting parameters are always $\alpha < 0, \beta > 0$, and thus the interarrival time U may have a heavy-tailed distribution, the service time V must be light tailed.

Also we consider an approximation of Algorithm 4 by partitioning the interval $[0, t]$ into M subintervals and applying the same tilting parameters α_m (for interarrival times) and β_m (for service times) to all arrival instants in the m-th interval.

Algorithm 5. [MMint]

1. *Compute the tilting function $\theta(s)$ in* (4).

2. *Let $I_1, I_2, \ldots, I_M$ be a partition of $[0,t]$ into M subintervals of equal size, and let t_m be the midpoint of the m-th subinterval. For each subinterval I_m determine tilting parameters α_m and β_m by solving*

$$\psi_U'(\alpha_m) = \frac{e^{-\theta(t_m)}}{\lambda}, \quad \psi_V'(\beta_m) = \frac{e^{\theta(t_m)}}{\mu}. \tag{24}$$

3. *Simulate a sample path of $Q_n(s), s \geq 0$ starting at $Q_n(0) = 0$ from event to event until either time horizon t has reached, or $Q_n(s) \geq nx$, whatever comes first.*

4. *Let s be an arrival epoch in the m-th subinterval I_m. Then the next interarrival time is set U/n with U drawn from the exponentially tilted distribution with tilting parameter α_m, and any arriving customer receives service time V drawn from an exponentially tilted distribution with tilting parameter β_m.*

5. *No action is taken after a departure event.*

4.3 A cross-entropy algorithm

Empirically we found that the algorithms of Sections 4.1 and 4.2 perform not so well in case of highly variable interarrival times or service times, see also our Section 6 with the simulation results. In this section we consider the application of the cross-entropy method for improving the tilting vectors $\alpha = (\alpha_m)_{m=1}^M$ and $\beta = (\beta_m)_{m=1}^M$ of Algorithm 5 in such cases. In the next section we consider the heavy-tailed case.

We denote the importance sampling probability measure by $P^{\alpha,\beta}$ when the interarrival times (service times) are exponentially tilted using tilting parameters α_m (β_m). The partitioning of $[0,t]$ into M subintervals of equal size is taken to be fixed throughout this section. The cross-entropy method minimises the Kullback-Leibler divergence of the zero-variance measure P^* from this parameterised family of probability measures $P^{\alpha,\beta}$ (see Rubinstein and Kroese, 2004). This means the following. Recall $Y_n = 1\{T_n(nx) \leq t\}$ the indicator of the rare event. Under the original probability measure P we have $\ell_n = E[Y_n]$. In this way one may view Y_n as an estimator based on a single sample. Consider a random realisation of the process $Q_n(s), 0 \leq s \leq t$, generated under an importance sampling algorithm $P^{\alpha,\beta}$. Its associated likelihood is denoted by $dP^{\alpha,\beta}(Q_n)$. Before we shall calculate the likelihood, we notice that the likelihood ratio is denoted and defined by $L(Q_n;\alpha,\beta) = dP(Q_n)/dP^{\alpha,\beta}(Q_n)$, and thus we have the unbiasedness property

$$\ell_n = E^{\alpha,\beta}\left[L(Q_n;\alpha,\beta)\, Y_n\right].$$

We reason similarly for the zero-variance measure P^* for which

$$\ell_n = E^*\left[L^*(Q_n)\, Y_n\right] \quad \text{and} \quad \mathrm{Var}^*[L^*(Q_n)\, Y_n] = 0.$$

The zero-variance measure P^* is not parameterised, but we might solve

$$\inf_{\alpha,\beta} \int \log \frac{dP^*}{dP^{\alpha,\beta}}\, dP^* = \inf_{\alpha,\beta} \int \frac{dP^*}{dP} \log \frac{dP^*}{dP^{\alpha,\beta}}\, dP.$$

This comes down to solving the following program (see Rubinstein and Kroese, 2004):

$$\max_{\alpha,\beta}\ E\left[Y_n \log\, dP^{\alpha,\beta}(Q_n)\right], \tag{25}$$

where the expectation is taken w.r.t. the original measure P. Because of the independence of the interarrival and service processes we can write down the log likelihood of a sample path. Denote by N_m the number of arrivals during subinterval I_m, with realised interarrival times U_j/n and service times V_j, and their associated densities given in (22) and (23), respectively. Then

$$\begin{aligned}\log dP^{\alpha,\beta}(Q_n) &= \sum_{m=1}^{M} \sum_{j=1}^{N_m} \Big(\log n f^{\alpha_m}(U_j) + \log g^{\beta_m}(V_j) \Big) \\ &= \sum_{m=1}^{M} \sum_{j=1}^{N_m} \Big(\log n + \alpha_m U_j - \psi_U(\alpha_m) + \log f(U_j) + \beta_m V_j - \psi_V(\beta_m) + \log g(V_j) \Big).\end{aligned} \tag{26}$$

The maximum likelihood program (25) is solved by considering its first order condition (FOC). Suppose that the optimising parameter α_m is restricted on $(-\infty, C_\alpha]$ for some positive finite C_α in the domain of ψ_U. Then $0 < \psi_U'(\alpha_m) \le \psi_U'(C_\alpha) < \infty$ because the cumulant generating function is increasing and convex. From (26) we see that

$$\left| \frac{\partial}{\partial \alpha_m} Y_n \log dP^{\alpha,\beta}(Q_n) \right| = \left| Y_n \sum_{j=1}^{N_m} \big(U_j - \psi_U'(\alpha_m) \big) \right| \le \sum_{j=1}^{N_m} U_j + N_m \psi_U'(C_\alpha),$$

with

$$E\left[\sum_{j=1}^{N_m} U_j + N_m \psi_U'(C_\alpha) \right] = (E[U] + \psi_U'(C_\alpha)) E[N_m] < \infty.$$

The same argument holds for the optimising parameter β_m. We assume that the boundary values C_α and C_β are large enough so that the interchange of expectation

and differentation in the FOC is allowed, which yields for $m = 1, \ldots, M$:

$$\begin{aligned} \frac{\partial}{\partial \alpha_m} E\left[Y_n \log dP^{\alpha,\beta}(Q_n)\right] = 0 \;\Leftrightarrow\; \psi_U'(\alpha_m) = \frac{E\left[Y_n \sum_{j=1}^{N_m} U_j\right]}{E\left[Y_n N_m\right]} \\ \frac{\partial}{\partial \beta_m} E\left[Y_n \log dP^{\alpha,\beta}(Q_n)\right] = 0 \;\Leftrightarrow\; \psi_V'(\beta_m) = \frac{E\left[Y_n \sum_{j=1}^{N_m} V_j\right]}{E\left[Y_n N_m\right]}. \end{aligned} \tag{27}$$

This solution to the FOC is estimated by simulation. Notice that the expectations in (27) are with respect to the original probability P and that they involve the rare event (via Y_n), thus we need to simulate with importance sampling densities. The idea is to do this iteratively with changes of measure $P^{\alpha^{(r)},\beta^{(r)}}$ and to use the equations (27) to update the parameters $\alpha_m^{(r)}, \beta_m^{(r)}$. Furthermore, the target level nx is updated adaptively in these iterations by setting it at a level $nx^{(r)}$ where a fraction of at least ρ of all samples gives overflow before (or at) target horizon t. This is the usual implementation of the cross-entropy algorithm as in Rubinstein and Kroese (2004). Hence we obtain the following algorithm.

Algorithm 6. [CE]

1. *Choose initial* $\alpha_m^{(0)}$ *and* $\beta_m^{(0)}$, $m = 1, \ldots, M$; $r = 0$.
2. *Simulate* k *sample paths of* $\{Q_n(s) : 0 \le s \le t\}$ *with tilted interarrival and service time distributions with tilting parameters* $\alpha_m^{(r)}$ *and* $\beta_m^{(r)}$, *respectively, and record the maximum attained level* S_i *of each path* $i = 1, \ldots, k$.
3. *Order the attained levels to get* $S_{(1)} \le S_{(2)} \le \cdots \le S_{(k)}$. *Set the target level* $nx^{(r)} = \min\left(nx, S_{([(1-\rho)k])}\right)$, *i.e.,* $Y_n = 1\{T_n(nx^{(r)}) \le t\}$.
4. *Use the same* k *samples to estimate the expections* $E\left[Y_n N_m\right]$, $E\left[Y_n \sum_{j=1}^{N_m} U_j\right]$, *and* $E\left[Y_n \sum_{j=1}^{N_m} V_j\right]$.
5. *Find the updated* $\alpha_m^{(r+1)}$ *and* $\beta_m^{(r+1)}$ *by solving* (27).
6. *Set* $r = r + 1$ *and repeat from 2 until convergence.*

Discussion:

- For the initial parameters $\alpha_m^{(0)}$ and $\beta_m^{(0)}$ we took the parameters given by (24) in Algorithm 5. For the succesfull fraction ρ we took 5%.

- Convergence: in steps 4 and 5 we actually execute a substition rule of the form

$$\alpha_m^{(r+1)} = (\psi_U')^{-1} \left(\frac{E^{(r)} \left[L(Q_n; \alpha^{(r)}, \beta^{(r)}) \, Y_n \sum_{j=1}^{N_m} U_j \right]}{E^{(r)} \left[L(Q_n; \alpha^{(r)}, \beta^{(r)}) \, Y_n N_m \right]} \right),$$

  where $E^{(r)}$ means that the expectation is taken w.r.t. $P^{\alpha^{(r)}, \beta^{(r)}}$. Similarly for the $\beta_m^{(r)}$ parameters. Thus the cross-entropy iteration is a substitution iteration of a fixed point equation. We could not prove analytically the convergence of the substitution rule, but we found empirically that in case of the finite-variance service times a few iterations (upto 10) was sufficient, whereas in case of infinite variability (Pareto distributed service times) the number of iterations could increase up to around 20.

5 Pareto distributed service times

In this section we assume that the service-time distributions are Pareto with form parameter $\kappa > 0$ and scale parameter $\gamma > 0$. The density function is

$$g(v) = \frac{\kappa}{\gamma} \left(1 + \frac{v}{\gamma} \right)^{-\kappa - 1} \qquad (v \geq 0).$$

Specifically we consider the case with $1 < \kappa \leq 2$ for which the mean service $E[V] = \gamma/(\kappa - 1)$ is finite with infinite variance. We apply the cross-entropy method for finding the importance sampling densities on the subintervals I_m. However, exponentially tilted versions of the density with positive tilting parameter β are not defined because the moment generating function does not exist. Instead, we take as importance sampling density on the m-th subinterval a Pareto density with form parameter κ_m and scale parameter γ_m. The interarrival densities during I_m remain as before, i.e., exponentially tilted with tilting parameter α_m. Thus, the maximum likelihood problem (25) becomes

$$\max_{\alpha, \kappa, \gamma} \; E \left[Y_n \log dP^{\alpha, \kappa, \gamma}(Q_n) \right],$$

where the optimisation parameters are the vectors $\alpha = (\alpha_m)_{m=1}^M$, $\kappa = (\kappa_m)_{m=1}^M$ $\gamma = (\gamma_m)_{m=1}^M$. The log likelihood of a sample path is, cf. (26),

$$\log dP^{\alpha, \kappa, \gamma}(Q_n) = \sum_{m=1}^{M} \sum_{j=1}^{N_m} \left(\log n f^{\alpha_m}(U_j) + \log g^{\kappa_m, \gamma_m}(V_j) \right).$$

The first order conditions become

$$\frac{\partial}{\partial \alpha_m} E\left[Y_n \log dP^{\alpha,\kappa,\gamma}(Q_n)\right] = 0 \;\Leftrightarrow\; \psi_U'(\alpha_m) = \frac{E\left[Y_n \sum_{j=1}^{N_m} U_j\right]}{E\left[Y_n N_m\right]} \tag{i}$$

$$\frac{\partial}{\partial \kappa_m} E\left[Y_n \log dP^{\alpha,\kappa,\gamma}(Q_n)\right] = 0 \;\Leftrightarrow\; \kappa_m = \frac{E[Y_n N_m]}{E\left[Y_n \sum_{j=1}^{N_m} \log\left(1 + \frac{V_j}{\gamma_m}\right)\right]} \tag{ii}$$

$$\frac{\partial}{\partial \gamma_m} E\left[Y_n \log dP^{\alpha,\kappa,\gamma}(Q_n)\right] = 0 \;\Leftrightarrow\; \frac{1}{\kappa_m + 1} = \frac{E\left[Y_n \sum_{j=1}^{N_m} \frac{V_j}{\gamma_m + V_j}\right]}{E[Y_n N_m]}. \tag{iii}$$

Equation (i) gives the tilting parameter α_m for the interarrival density. From equation (ii) and (iii) we eliminate κ_m leaving an equation in γ_m which we can solve numerically by bisection. Then any of (ii) and (iii) gives κ_m. The cross-entropy iteration starts with the original parameters.

6 Numerical results

We have executed simulation experiments for various combinations of types of distribution functions F of interarrival time U and of distribution function G of service time V: for arrivals we took Exponential and Hyperexponential distributions; for services we considered Exponential, Deterministic, Gamma, Coxian (two phases), and Pareto (finite mean, infinite variance). Their associated parameters are obtained by fitting the first two moments using mean and squared coefficient of variation (Tijms, 2003, page 448). For the Pareto distribution we fit just the first moment. It is not possible to implement Algorithms 4 and 5 for the Pareto case as explained in Section 5.

The model parameters are

$$E[U] = 1, E[V] = 5, x = 6, t = 10,$$

and the number of simulated sample paths (for the estimation of ℓ_n) is in all experiments $k = 50000$. In the cross-entropy iterations we took 5000 samples. After each simulation experiment we collect three (estimated) performance measures of the importance sampling estimator $\overline{Y^*}_n(k)$ of ℓ_n based on k samples:

- RHW: the relative half width of the 95% confidence interval

$$1.96\sqrt{\mathrm{Var}^{\mathrm{IS}}[\overline{Y_n^*}(k)]}/\overline{Y_n^*}(k).$$

- RAT: the logarithmic efficiency ratio, cf. (1),

$$\log E^{\mathrm{IS}}[\overline{Y_n^*}(k)^2]/\log E^{\mathrm{IS}}[\overline{Y_n^*}(k)].$$

- EFF: the ($-$logarithm of the) effort which takes into account both the variance of the estimator and the total execution time of the simulation (including the iterations in the cross-entropy algorithm):

$$-\log_{10}\left(\mathrm{Var}^{\mathrm{IS}}[\overline{Y_n^*}(k)] \times \mathrm{CPU}[\overline{Y_n^*}(k)]\right).$$

Better performance is obtained by smaller RHW, higher RAT, and larger EFF. In the cross-entropy method we included the time needed to execute the cross-entropy iterations.

1. Poisson arrivals, exponential service times: the $M/M/\infty$ case. Here we compare the importance sampling estimates ℓ_n^{IS} of our algorithm of Section 2 with the numerical values ℓ_n^{NUM} and with the large deviations approximations ℓ_n^{LD}. The numerical values are obtained by numerical inversion of the Laplace transform (Keilson 1979). In Figure 2 we have plotted the logarithm of these values for scalings up to $n = 200$. The constant slope of the line indicates the exponential decay of the probability. The deviation for scalings larger than 150 reflects the appearance of large numerical errors.

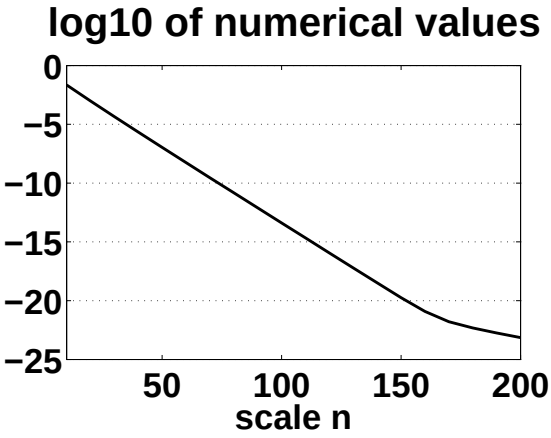

Figure 2. The log10 of the numerical values ℓ_n^{NUM} for scalings up to $n = 200$. For scalings larger than 150 the values become less accurate due to numerical errors.

The large deviations approximations are obtained from Theorem 2 by

$$\ell_n^{\mathrm{LD}} = e^{-nJ_2(t,x)}.$$

The comparisons in Figure 3 are the relative differences

$$\frac{\left|\ell_n^{\mathrm{NUM}} - \ell_n^{\mathrm{IS}}\right|}{\ell_n^{\mathrm{NUM}}} \quad \text{and} \quad \frac{\left|\ell_n^{\mathrm{LD}} - \ell_n^{\mathrm{IS}}\right|}{\ell_n^{\mathrm{LD}}}$$

Furthermore, we have estimated the efficiency ratios RAT which should converge to 2 because the importance sampling algorithm is asymptotic optimal.

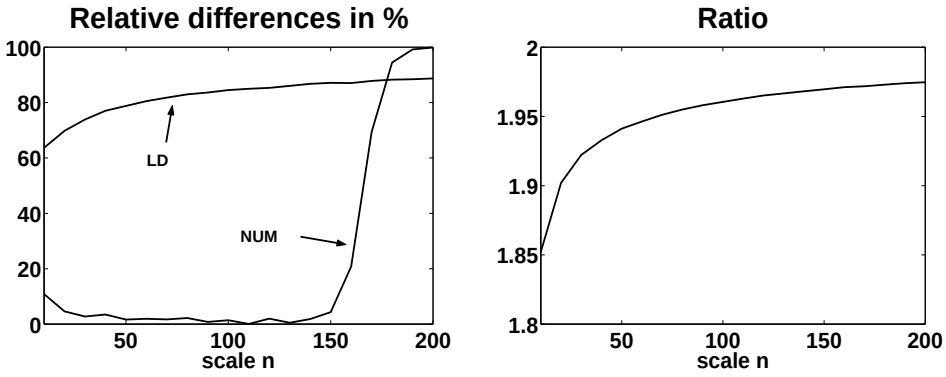

Figure 3. Left: relative differences between importance sampling and numerical inversion resp. large deviations for scalings up to $n = 200$. Right: the estimated logarithmic efficiency ratios.

Notice the large differences between the importance sampling estimates and the numerical values when the scaling factors n are large. A reason might be that numerical errors in the numerical procedures become a crucial issue.

2. Poisson arrivals, Deterministic service times ($c_V^2 = 0$), Gamma service times with $c_V^2 = 0.5$, Coxian service times with $c_V^2 = 4$, and Pareto service times with infinite variance. We chose 20 intervals in Algorithms 5 and 6, 5000 samples per CE iteration, and CE was iterated until two consecutive solution vectors differ less that 0.1 (in 2-norm).

3. Hyperexponential arrivals with $c_U^2 = 5$, Gamma service times with $c_V^2 = 0.5$, Exponential service times, Coxian service times with $c_V^2 = 4$, and Pareto service times with infinite variance. We chose 20 intervals in algorithm 5 and 6, 5000 samples per CE iteration, and CE was iterated until two consecutive solution vectors differ less that 0.1 (in 2-norm).

The simulation results are summarised in the Table 1 (RHW for exponential interarrivals), Table 2 (RHW for hyperexponential interarrivals), Table 3 (RAT for

both cases), and Table 4 (EFF for both cases). A blank in the tables means that there were not enough observations of the rare event for that specific case. From these experiments we see that Algorithm 3 (adapted Szechtman-Glynn) gives good performance for low-variable service times, but that RHW degrades when the variability grows. Algorithms 4 and 5 (using the $M/M/\infty$ parameters) are applicable for low-variable service times but perform in most cases worse than Algorithm 3. Algorithm 6 (cross-entropy) gives in all cases excellent results and outperforms (in most cases) the other algorithms.

The cross-entropy algorithm is based on heuristics, and the importance sampling densities that are finally used, are obtained after simulation experiments. Hence, we have no explicit representation of these denstities, and this makes that we cannot decide upon the asymptotic optimality of its associated estimator by an analytical approach. Instead we assess the optimality empirically by the ratio RAT. Figure 4 plots these ratios in the model with Poisson arrivals and Coxian ditributed service times. We show also these ratios obtained by the (adapted) Szechtman and Glynn algorithm. The scale parameter n has been increased until $n = 200$ in which case $\ell_n \approx 10^{-112}$. The ratios in both algorithms tend to 'creep' to 2, and this indicates asymptotic optimality.

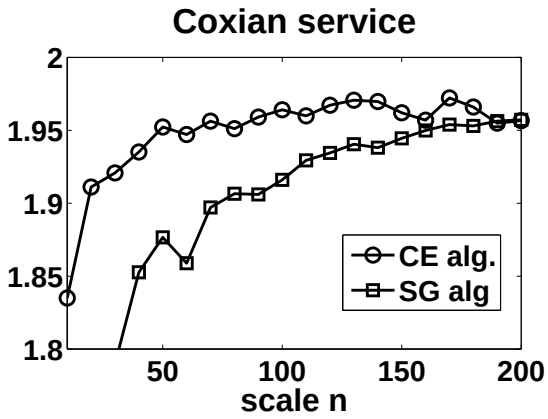

Figure 4. The estimated logarithmic efficiency ratios obtained by Algorithm 3 (SG) and Algorithm 6 (CE) in case of Poisson arrivals and Coxian distributed services.

n	$\hat{\ell}_n$	RHW Alg.3 [SG]	RHW Alg.4 [MM]	RHW Alg.5 [MMint]	RHW Alg.6 [CE]
Deterministic with $c_V^2 = 0$					
100	1.447e-004	0.1840	0.0265	0.0264	0.0265
200	1.595e-008	0.2241	0.2475	0.1859	0.0798
300	1.504e-012	0.5092	0.2968	0.2966	0.1316
400	9.865e-017	0.5937	0.5099	0.5620	0.2355
500	9.175e-021	0.2974	0.2412	0.9042	0.1732
Gamma with $c_V^2 = 0.5$					
50	1.063e-004	0.0496	0.0231	0.0223	0.0232
100	2.992e-008	0.1199	0.0553	0.1026	0.0472
150	8.881e-012	0.2613	0.5384	0.1279	0.0669
200	2.944e-015	0.1778	0.2261	0.2078	0.0869
250	9.145e-019	0.4129	0.4205	0.3093	0.0995
Coxian with $c_V^2 = 4$					
10	1.626e-006	0.0883	0.3500	0.1593	0.0299
20	4.918e-012	0.2684			0.0276
30	1.662e-017	0.5388			0.0538
40	6.152e-023	0.4424			0.0473
50	2.248e-028	0.5535			0.0465
Pareto with $c_V^2 = \infty$					
5	5.941e-004	0.0787			0.0252
10	4.065e-007	0.1368			0.0522
15	2.856e-010	0.4622			0.0598
20	1.996e-013	0.3557			0.1929
25	1.242e-016	0.4233			0.1767

Table 1. The relative halfwidths of the four importance sampling algorithms for Poisson arrivals.

n	$\hat{\ell}_n$	RHW Alg.3 [SG]	RHW Alg.4 [MM]	RHW Alg.5 [MMint]	RHW Alg.6 [CE]
Gamma with $c_V^2 = 0.5$					
100	1.054e-002	0.0103	0.2921	0.3292	0.0163
200	2.538e-004	0.0145	0.4541	0.4200	0.0784
300	6.568e-006	0.0179	0.8641	0.3981	0.0896
400	1.793e-007	0.0223	1.7924	0.3584	0.0435
500	4.938e-009	0.0278	1.6069	1.1525	0.0679
Exponential with $c_V^2 = 1$					
100	1.589e-004	0.0167	0.1955	0.1108	0.0126
200	7.207e-008	0.0299	0.5796	0.2207	0.0170
300	3.689e-011	0.0443	0.3472	0.5233	0.0248
400	1.965e-014	0.0788	0.4530	0.9041	0.0301
500	9.686e-018	0.0927	1.0168	0.8494	0.0806
Coxian with $c_V^2 = 4$					
20	2.976e-005	0.1946	0.2030	0.2008	0.0167
40	1.682e-009	0.4716			0.0223
60	1.076e-013	1.0970			0.0229
80	7.589e-018				0.0539
100	5.180e-022				0.0639
Pareto with $c_V^2 = \infty$					
20	8.131e-006	0.0684			0.0394
30	2.749e-008	0.1474			0.0299
40	9.216e-011	0.2988			0.1157
50	3.136e-013	0.2191			0.1523
60	7.911e-016	0.5032			0.2050

Table 2. The relative halfwidths of the four importance sampling algorithms for hyperexponential arrivals.

n	RAT Alg.3 [SG]	Alg.4 [MM]	Alg.5 [MMint]	Alg.6 [CE]	n	RAT Alg.3 [SG]	Alg.4 [MM]	Alg.5 [MMint]	Alg.6 [CE]
Deterministic with $c_V^2 = 0$					Gamma with $c_V^2 = 0.5$				
100	1.31	1.74	1.74	1.73	100	1.81	0.48	0.40	1.67
200	1.65	1.62	1.66	1.75	200	1.84	1.16	1.20	1.47
300	1.70	1.74	1.74	1.80	300	1.86	1.33	1.48	1.61
400	1.77	1.78	1.77	1.85	400	1.87	1.44	1.66	1.79
500	1.85	1.85	1.80	1.87	500	1.87	1.54	1.59	1.78
Gamma with $c_V^2 = 0.5$					Exponential with $c_V^2 = 1$				
50	1.62	1.77	1.78	1.77	100	1.82	1.28	1.42	1.87
100	1.70	1.79	1.71	1.80	200	1.85	1.49	1.62	1.90
150	1.73	1.67	1.79	1.84	300	1.86	1.71	1.68	1.91
200	1.82	1.81	1.81	1.86	400	1.86	1.76	1.72	1.92
250	1.82	1.81	1.83	1.88	500	1.88	1.78	1.79	1.88
Coxian with $c_V^2 = 4$					Coxian with $c_V^2 = 4$				
10	1.66	1.46	1.46	1.81	20	1.41	1.40	1.40	1.85
20	1.74			1.91	40	1.62			1.90
30	1.79			1.91	60	1.70			1.93
40	1.85			1.93	80				1.91
50	1.88			1.95	100				1.92
Pareto with $c_V^2 = \infty$					Pareto with $c_V^2 = \infty$				
5	1.41			1.70	20	1.65			1.74
10	1.63			1.76	30	1.67			1.85
15	1.64			1.82	40	1.69			1.78
20	1.75			1.79	50	1.78			1.80
25	1.79			1.84	60	1.76			1.82

Table 3. The efficiency ratios of the four importance sampling algorithms for Poisson arrivals (left) and hyperexponential arrivals (right).

n	EFF Alg.3 [SG]	Alg.4 [MM]	Alg.5 [MMint]	Alg.6 [CE]	n	EFF Alg.3 [SG]	Alg.4 [MM]	Alg.5 [MMint]	Alg.6 [CE]
Deterministic with $c_V^2 = 0$					Gamma with $c_V^2 = 0.5$				
100	7.56	9.06	9.22	9.16	100	5.79	2.87	3.17	5.61
200	15.18	14.15	14.50	15.37	200	8.31	6.26	6.93	7.06
300	21.61	21.89	21.99	22.56	300	11.02	9.20	11.06	9.69
400	29.90	29.26	29.14	30.61	400	13.72	12.93	16.94	12.99
500	38.55	38.70	36.95	38.18	500	16.46	16.23	17.36	15.49
Gamma with $c_V^2 = 0.5$					Exponential with $c_V^2 = 1$				
50	8.80	9.95	10.09	9.95	100	9.40	6.68	7.83	9.37
100	14.82	15.86	15.29	15.94	200	15.08	12.25	13.96	15.27
150	20.87	20.33	21.92	22.38	300	20.98	20.21	19.87	21.18
200	28.39	28.24	28.37	28.75	400	26.74	26.76	25.90	27.29
250	34.32	34.49	35.02	35.55	500	32.94	33.98	34.20	32.73
Coxian with $c_V^2 = 4$					Coxian with $c_V^2 = 4$				
10	12.62	11.67	8.79	13.49	20	8.94	8.61	9.59	11.38
20	22.50			24.48	40	17.03			19.17
30	32.90			34.39	60	25.71			27.27
40	45.13			45.36	80				34.65
50	56.97			56.02	100				42.60
Pareto with $c_V^2 = \infty$					Pareto with $c_V^2 = \infty$				
5	8.41			8.57	20	11.36			10.83
10	14.09			14.01	30	15.43			15.89
15	18.98			20.20	40	19.41			19.91
20	25.74			25.37	50	24.97			24.43
25	31.78			31.78	60	28.73			29.08

Table 4. The efforts of the four importance sampling algorithms for Poisson arrivals (left) and hyperexponential arrivals (right).

7 Conclusion

In this paper we have developed importance sampling algorithms for the simulation of first-passage time probabilities in the $G/G/\infty$ queueing model. The algorithms that were based on the sample path large deviations in the $M/M/\infty$ model have small applicability and performed poorly in case of highly variable service times.

An adaptation of the algorithm of Szechtman and Glynn (2002) seemed to overcome this difficulty and gave in most cases good performance—even for Pareto servers with infinite variance. We found that the cross-entropy based algorithm gave the best performance although we are not sure whether the associated estimator is asymptotically optimal in all cases. Further investigations on this, as well as on the convergence of the cross-entropy iterations are needed.

Acknowledgements

The author would like to thank two anonymous referees for their numerous comments which helped to improve the presentation of the paper.

Appendix A (Proofs of Lemmas 1 and 2)

Lemma 1. *Recall the tilting function* $\theta(s)$ *in* (4) *and the optimal path* $\phi(s)$ *to overflow in* (8). *For all* $0 \leq s \leq t$

$$\lambda\left(e^{\theta(s)}-1\right) + \phi(s)\mu\left(e^{-\theta(s)}-1\right) = \text{constant},$$

with the constant equal to the c *in* (8).

Proof. Substitute $q = \phi(s)$ in the cumulant generating function (2):

$$\psi(\theta, \phi(s)) = \lambda\left(e^{\theta}-1\right) + \phi(s)\mu\left(e^{-\theta}-1\right).$$

Because ψ is convex in its first argument θ, we can solve the local rate function (3) by its first order condition

$$\frac{\partial}{\partial\theta}\psi(\theta, \phi(s))|_{\theta=\theta(s)} = \phi'(s). \tag{28}$$

The optimal path $\phi(s)$ satisfies (see Shwartz and Weiss, 1995, (C.3)):

$$I(\phi(s), \phi'(s)) - \phi'(s)\frac{\partial}{\partial\phi'(s)}I(\phi(s), \phi' s) = \text{constant}.$$

Working out the differentation and using (28) we get

$$I(\phi(s), \phi'(s)) - \phi'(s)\theta(s) = \text{constant}.$$

And because $I(\phi(s), \phi'(s)) = \phi'(s)\theta(s) - \psi(\theta(s), \phi(s))$, we find that $\psi(\theta(s), \phi(s))$ is a constant with by definition

$$\psi(\theta(s), \phi(s)) = \lambda\left(e^{\theta(s)}-1\right) + \phi(s)\mu\left(e^{-\theta(s)}-1\right).$$

The constant is obtained by substituting $s = 0$ for which we have $\phi(0) = 0$, and $\phi'(0) = c + \lambda$. Thus,

$$\text{constant} = \psi(\theta(0), \phi(0)) = \lambda \left(e^{\theta(0)} - 1 \right) = \lambda \left(\frac{c + \lambda + \sqrt{(c+\lambda)^2}}{2\lambda} - 1 \right) = c.$$

□

Lemma 2. *The large deviations rate function $J_1(\phi)$ of (6) satisfies*

$$J_1(\phi) = \int_0^t \theta(s)\phi'(s)\, ds - ct.$$

Proof. In the proof of Lemma 1 we found that the local rate function satisfies

$$I(\phi(s), \phi'(s)) = \theta(s)\phi'(s) - g(\phi(s), \theta(s)) = \theta(s)\phi'(s) - c.$$

Thus we get the statement immediately by noticing that the large deviations rate function is $J_1(\phi) = \int_0^t I(\phi(s), \phi'(s))\, ds$. □

References

Asmussen S., Madan D., Pistorius M., 2008. Pricing equity default swaps under an approximation to the CGMY Lévy model. Journal of Computational Finance 11, 79-93.

Blanchet J., Glynn P., 2008. Efficient rare-event simulation for the maximum of heavy-tailed random walks. Annals of Applied Probability 18, 1351-1378.

Blanchet J., Mandjes M., 2007. Editorial: rare-event simulation for queues. Queueing Systems 57, 57-59.

Bucklew J.A., 2004. Introduction to Rare Event Simulation. Springer, New York.

Dembo A., Zeitouni O., 1998. Large Deviations Techniques and Applications, second ed. Springer, New York.

Dohi T., Matsuoka T., Osaki S., 2002. An inifinite server queuing model for assessment of the software reliability. Electronics and Communications in Japan 85, 43-51.

Dupuis P., Sezer D., Wang H., 2007. Dynamic importance sampling for queueing network. Annals of Applied Probability 17, 1306-1346.

Glassermann P., Wang Y., 1997. Counter examples in importance sampling for large deviations probabilities, Annals of Applied Probability 7, 731-746.

Glynn P., 1995. Large deviations for the infinite server queue in heavy traffic. In: Kelly F., Williams R. (Eds), Stochastic Networks, IMA Vol. 71. Springer, pp. 387-394.

Heidelberger P., 1995. Fast simulation of rare events in queueing and reliability models. ACM Transactions on Modelling and Computer Simulation 5, 43-85.

Juneja S., Shahabuddin P., 2006. Rare-event simulation techniques: an introduction and recent advances. Chapter 11 in: Henderson S.G., Nelson B.L. (Eds), Handbooks in Operations Research and Management Science Volume 13: Simulation, North-Holland, Amsterdam, pp. 291-350.

Keilson J., 1979. Markov Chain Models — Rarity and Exponentiality. Springer-Verlag, New York.

Mandjes M., Ridder A., 2001. A large deviations approach to the transient of the Erlang loss model. Performance Evaluation 43, 181-198.

Rubinstein R.Y., Kroese D.P., 2004. The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning. Springer, New York.

Shwartz A., Weiss A., 1995. Large Deviations for Performance Analysis: Queues, Communications and Computing. Chapman Hall, London.

Szechtman R., Glynn P., 2002. Rare-event simulation for infinite server queues. Proceedings of the 2002 Winter Simulation Conference, Vol. 1. IEEE Press, pp. 416-423.

Tijms H.C., 2003. A First Course in Stochastic Models. Wiley, Chichester.