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

August 26, 2008

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.

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^{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., $E^{\mathrm{IS}}[\overline{Y_n^*}(k)] = \ell_n$, if we incorporate the likelihood ratio dP/dP^{IS}. (With the superscript IS we show explicitly that the expectation is taken w.r.t. measure P^{IS}.)

Finding a good probability P^{IS} is the main issue in importance sampling. The criterion is to keep the relative error $\sqrt{\mathrm{Var}^{\mathrm{IS}}[\overline{Y_n^*}(k)]}/E^{\mathrm{IS}}[\overline{Y_n^*}(k)]$ 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):

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

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 grows polynomially (or at some other subexponential rate), and thus also the sample sizes grow polynomially in order to obtain a prespecified relative error.

Related to our study is the work of Szechtman and Glynn (2002) who 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 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 asymptotically optimal importance sampling algorithm based on sample path large deviations results for the Erlang loss model (Shwartz and Weiss 1995, Chapter 12). 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.

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 have applied here, resulting in

$$\lim_{\epsilon \downarrow 0} \lim_{n\to\infty} \frac{1}{n} \log P\left(\sup_{0\le s\le t} |Q_n(s)/n - \phi(s)| < \epsilon \right) = -J_1(\phi), \tag{2}$$

for absolute continuous functions ϕ. The functional J_1 satisfies

$$J_1(\phi) = \int_0^t I\left(\phi(s), \phi'(s)\right)\, ds, \tag{3}$$

with $I(\cdot,\cdot)$ the local rate function defined by

$$I(q,y) = \sup_{\theta\in\mathbb{R}} \left(\theta y - \psi(\theta, q)\right), \tag{4}$$

where $\psi(\cdot,\cdot)$ is the cumulant generating function of the jumps,

$$\psi(\theta, q) = \lambda\left(e^{\theta} - 1\right) + q\mu\left(e^{-\theta} - 1\right) \quad (\theta \in \mathbb{R}). \tag{5}$$

It is an exercise (Exercise 12.6 in Shwartz and Weiss (1995)) to get that the optimising θ (called tilting parameter) satisfies

$$e^{\theta} = \frac{y + \sqrt{y^2 + 4\lambda q\mu)}}{2\lambda}. \tag{6}$$

The main point is that sample path large deviations (2) 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 (6) along the path by substituting $q = \phi(s)$ and $y = \phi'(s)$. Therefore we deal with a tilting function $\theta(s)$, $0 \le s \le t$.

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* (6) *where the function $\phi(s)$ is given in* (8).
2. *Simulate a sample path of $Q_n(s), s \ge 0$ starting at $Q_n(0) = 0$ from event to event until either time horizon t has reached, or $Q_n(s) \ge nx$, 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 $\lambda e^{\theta(s)}$. And 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 $\mu\, e^{-\theta(s)}$.*

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

The proof is given in Ridder (2008).

3 The general model

The general model comprises a renewal process for arrivals and i.i.d. service times. We apply the large deviations results for the sequence of scaled variables $(Q_n(t)/n)_{n=1}^{\infty}$ at the horizon t (taken fixed), developed by Glynn (1995):

$$\lim_{n\to\infty} \frac{1}{n} \log P\left(\, Q_n(t)/n \ge x \,\right) = -J_2(t,x) \quad (x > m(t)). \tag{9}$$

The large deviations rate function $J_2(\cdot,\cdot)$ is the Legendre-Fenchel transform

$$J_2(t,x) = \sup_{\theta\in\mathbb{R}} \left(\theta x - \psi_Q(\theta,t)\right), \tag{10}$$

where $\psi_Q(\cdot,\cdot)$ is the limiting cumulant generating function

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

The optimising θ^* in (10) is the positive root of

$$\frac{\partial}{\partial\theta}\,\psi_Q(\theta,t) = x, \tag{11}$$

and is called the optimal tilting parameter. We need the following lemma before proving the large deviations for the first passage time probabilities $\ell_n = P(T_n(nx) \le t)$.

Lemma 1. *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 (11), 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.

The algorithm of Szechtman and Glynn (2002) for estimating the tail probability $P\left(\, Q_n(t) \ge nx \,\right)$ in the general model is based on the fact that it suffices to generate arrivals, and along with each arrival to generate a Bernoulli random variable that indicates whether the arriving customer would still be present at time t. Suppose that the previous arrival was at time s, then the interarrival time to the next arrival has exponentially tilted density

$$f^{\alpha}(u) = \exp\left(\alpha u - \psi_U(\alpha)\right)\, f(u), \tag{12}$$

where the tilting parameter $\alpha = \alpha(s)$ solves

$$\psi_U(\alpha) = -\log\Big(e^{\theta^* \overline{G}}(t-s) + G(t-s)\Big). \tag{13}$$

And when s is an arrival epoch, the arriving customer has a service time that takes longer than $t-s$ with probability

$$\frac{e^{\theta^* \overline{G}}(t-s)}{e^{\theta^* \overline{G}}(t-s) + G(t-s)}. \tag{14}$$

4 Importance sampling algorithms

In this section we present the three importance algorithms for the first-passage time problem in the general $G/G/\infty$ queue.

4.1 Adapted Szechtman-Glynn algorithm

The first algorithm for the general model is an adaptation of the algorithm of Szechtman and Glynn (2002), 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 service time V for the customer arriving at time s has a cdf $G^*(v)$ that we define by its complementary

$$\overline{G^*}(v) = \frac{e^{\theta^*}\overline{G}(v)}{e^{\theta^*}\overline{G}(v) + G(v)}. \tag{15}$$

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 2. [SG]

1. *Compute the optimal tilting parameter θ^* in* (11).
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* (12).
4. *Let s be an arrival epoch. Then the arriving customer receives service time V drawn from cdf G^* given above in* (15).
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 the Bernoulli probability (14). Also notice that because $\theta^* > 0$, the righthand side in (13) is negative, and thus this equation is also solvable for heavy-tailed interarrival times. In that case the solution to (13) is a negative α for which $\psi_U(\alpha) < \infty$. Hence, Algorithm 2 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 3. [MM]

1. *Compute the tilting function $\theta(s)$ in* (6).

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*
$$f^{\alpha}(u) = \exp\left(\alpha u - \psi_U(\alpha)\right) f(u), \tag{16}$$
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*
$$g^{\beta}(v) = \exp\left(\beta v - \psi_V(\beta)\right) g(v), \tag{17}$$
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 3 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 4. [MMint]

1. *Compute the tilting function $\theta(s)$ in* (6).

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*
$$\psi'_U(\alpha_m) = \frac{e^{-\theta(t_m)}}{\lambda}, \quad \psi'_V(\beta_m) = \frac{e^{\theta(t_m)}}{\mu}. \tag{18}$$

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. 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 4 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):

$$\max_{\alpha,\beta}\ E\left[Y_n \log\, dP^{\alpha,\beta}(Q_n)\right], \tag{19}$$

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 (16) and (17), respectively. Then

$$\begin{aligned}\log dP^{\alpha,\beta}(Q_n) &= \sum_{m=1}^{M}\sum_{j=1}^{N_m}\Big(\log n f^{\alpha_m}(U_j) + \log g^{\beta_m}(V_j)\Big) \\ &= \sum_{m=1}^{M}\sum_{j=1}^{N_m}\Big(\log n + \alpha_m U_j - \psi_U(\alpha_m) + \log f(U_j) + \beta_m V_j - \psi_V(\beta_m) + \log g(V_j)\Big).\end{aligned} \tag{20}$$

The maximum likelihood program (19) is solved by considering its first order condition (FOC). After interchanging expectation and differentation in the FOC (allowed!), we get for $m = 1, \ldots, M$:

$$\begin{aligned}\frac{\partial}{\partial \alpha_m} E\left[Y_n \log dP^{\alpha,\beta}(Q_n)\right] = 0 \;\Leftrightarrow\; \psi_U'(\alpha_m) = \frac{E\left[Y_n \sum_{j=1}^{N_m} U_j\right]}{E\left[Y_n N_m\right]} \\ \frac{\partial}{\partial \beta_m} E\left[Y_n \log dP^{\alpha,\beta}(Q_n)\right] = 0 \;\Leftrightarrow\; \psi_V'(\beta_m) = \frac{E\left[Y_n \sum_{j=1}^{N_m} V_j\right]}{E\left[Y_n N_m\right]}.\end{aligned} \tag{21}$$

This solution to the FOC is estimated by simulation. Notice that the expectations in (21) 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 (21) 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 5. [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* (21).

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 (18) in Algorithm 4. 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 (19) 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. (20),

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

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. For the Pareto distribution we fit just the first moment. It is not possible to implement Algorithms 3 and 4 for the Pareto case as explained in Section 5.

1. 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 4 and 5, 5000 samples per CE iteration, and CE was iterated until two consecutive solution vectors differ less that 0.1 (in 2-norm).

2. 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 4 and 5, 5000 samples per CE iteration, and CE was iterated until two consecutive solution vectors differ less that 0.1 (in 2-norm).

Extensive simulation results can be found in Ridder (2008). From these experiments we see that Algorithm 2 (adapted Szechtman-Glynn) gives good performance for low-variable service times, but that the relative error degrades when the variability grows. Algorithms 3 and 4 (using the $M/M/\infty$ parameters) are applicable for low-variable service times but perform in most cases worse than Algorithm 2. Algorithm 5 (cross-entropy) gives in all cases excellent results and outperforms (in most cases) the other algorithms.

References

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.

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.

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.

Ridder A., 2008. Importance sampling algorithms for first passage time probabilities in the infinite server queue. Submitted. Available at

`http://staff.feweb.vu.nl/aridder/papers/gginf.pdf`

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.