Fast simulation of the infinite server queue

Ad Ridder
Dept. of Econometrics and Operations Research
Vrije University, 1081 HV Amsterdam

Abstract

Abstract: This paper describes Monte Carlo simulations for the estimation of the transient probability that the infinite server queue reaches high levels within a predefined time window. Under the light traffic assumption this is a rare event. The simulations are speed up by applying importance sampling using a change of measure. In case of Poisson arrivals and exponential servers, the $M/M/\infty$ model, an exponential change of measure with a constant tilting factor gives poor results. The reason being that the optimal path to high levels is not a straight line. However, from the expression of the optimal path we deduce a change of measure under which the arrival and service time distributions are updated after each jump time of the process. We prove that the importance sampling algorithm with this measure is asymptotically optimal. For the general case we adapt this algorithm using a few heuristics that are based on the observation that the importance sampling simulations in the Poisson/exponential model with a few updates of the distributions give good performance. The algorithms are tested by simulation experiments.

1 Introduction

The infinite server queue, denoted by $M/G/\infty$, is a queueing model with infinitely many servers who are accessed by customers arriving according to a Poisson process. The service demands of customers are independent, identically distributed random variables, independent of the Poisson arrival process. Customers leave the system after service. These systems are used, for instance, to model buffer resources in telecommunication systems or in computer networks, where, of course, these resources have finite capacities. However, the capacities are typically large to accomodate a huge number of connections (customers), and as long as the system has not reached its limits, it behaves statistically similar to our infinite server model. Then it may be relevant to know how quickly the full system state will be reached (if at all!) when one observes the system at some arbitrary instant. We will not study this issue in full detail but rather we pick some specific problem related to this matter. We are going to assume that the queue is empty, i.e., all servers are idle, at the time instant at which we observe the queue. Then we like to find the probability distribution function of the first passage time of high levels, because this would give us all the statistical information about the chances of a full buffer (in the finite system). Only in case of exponential servers it is possible to find, numerically, this function for any argument by a numerical inversion [1] of the Laplace Transforms of the first passage time densities which are given by recurrence relations [9, Chapter 5]. However, we run into numerical problems considering the high levels of interest.

To be more specific, we consider a sequence of processes $\{X_n(t) : t \geq 0\}$, $n = 1, 2, \ldots$, where $X_n(t)$ represents the number of busy servers at time t in a $M/G/\infty$ model with Poisson $(n\gamma)$ arrivals, and with general service demands with expectation $1/\mu$. The first passage times of the n-th system are

$$T_n(k) := \inf\{t \geq 0 : X_n(t) = k\} \quad (k = 1, 2, \ldots), \tag{1}$$

where $X_n(0) = 0$. The target probability is

$$\alpha_n := P\Big(T_n([nb]) \in [\tau_1, \tau_2] \Big).$$

In other words, we consider $M/G/\infty$ queues where the arrival rates $\lambda = n\gamma$ and the hitting levels $B = [nb]$ are growing proportionally to fixed constants γ and b. We assume light traffic, i.e., $\gamma < b\mu$. The equilibrium distribution of the process, i.e., the distribution of $\lim_{t\to\infty} X_n(t)$, is insensitive for the second and higher moments of the service demand. In fact, it is Poisson with mean $\lambda/\mu = n\gamma/\mu$ [10, Section 3.3]. Hence, because of the light traffic assumption, the event of reaching level $[nb]$ is unlikely, it is a rare event. In Section 2.1 we shall use large deviations—in case of exponential service demands—to show that $\alpha_n \to 0$ exponentially fast when $n \to \infty$.

Consequently, the execution times of ordinary Monte Carlo simulations will become too long to be practical when we consider the estimation problem for large n, for instance when $\alpha_n \approx 10^{-6}$ or smaller. Various variance reduction techniques exist to overcome this problem. In this paper we shall apply importance sampling. Let Y_n be an unbiased estimator of α_n under the original probability measure P. In importance sampling we simulate under another probability measure, say P^*, such that the original measure P is absolutely continuous relative to this new measure. The new estimator becomes $Y_n^* = LY_n$, where L denotes the likelihood ratio, $L = dP/dP^*$. Clearly, the new estimator is unbiased, $E^*[Y_n^*] = \alpha_n$.

Finding a good new probability P^* is the main issue in importance sampling. The criterion is to keep the relative error $\sqrt{\mathrm{Var}^*[Y_n^*]}/\alpha_n$ 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. Slightly weaker is the concept of logarithmical efficiency [3] or asymptotical optimality [8]:

$$\liminf_{n\to\infty} \frac{\log \mathrm{Var}^*[Y_n^*]}{\log \alpha_n^2} \geq 1. \tag{2}$$

This yields good performance and considerable variance reductions.

A way to find a good new measure P^* is to implement an exponential change of measure [4]. That is, the distribution functions of the random variables are exponentially tilted. The advantage is that the likelihood ratio can be calculated easily. For overflow problems in queueing systems such as $M/G/c$, the optimal tilting factor is found after a large deviations analysis of the model [15]. Optimal in the sense that the IS estimator is asymptotically optimal. The same approach has been successfully applied to other queueing models such as tandem Jackson networks [13], discrete time intree networks [5], and continuous fluid queues with Markov modulated inputs [11]. The optimal tilting factor in all these models is constant. In other words, the new arrival and service time distributions remain the same throughout the simulation. The argument why it works out like that, is the property in these models that the 'optimal path to overflow' is a straight line. The optimal path is a concept in the large deviations analysis of the models where it is the path that has the least 'cost' among all paths that reach overflow [16, Chapter 5] (we come back to this issue in the next section). The interpretation of the optimal path is that when one generates many sample paths of the queueing model in a simulation program, the majority of the sample paths will not show overflow (because overflow is a rare event), but of those samples that do, almost all are 'close' to the optimal path. (There is a scaling involved in the sense that the simulation sample paths have to be normalized.) This interpretation had lead to the heuristic of letting the average behaviour under the importance sampling probability measure to be the optimal path of the rare event. And this is exactly what happens in the studies mentioned above. However, other studies have come up with problems

in which the optimal path approach does not give an asymptotically optimal importance sampling estimator [7]. For instance when the rare event is not a convex set in the sample space the optimal path approach might fail.

In case of exponential servers, the $M/M/\infty$ model, an exponential change of measure with a constant tilting factor gives poor results. The reason being that the optimal path to high levels (explained in Section 2.1) is not a straight line. As we shall show in Sections 2.2 and 2.4, under the optimal change of measure the arrival and service time distributions are updated after each jump time of the process. During the last decade, a growing number of rare event simulation studies contribute to the area of adaptive importance sampling techniques. It appears that adaptive importance sampling algorithms are effective in a larger class of problems and in more complex models than was possible with the static rules. Among these adaptive techniques, we mention the adaptive Monte Carlo method [6], which assumes that the rare event probability is reformulated as an expected average reward in a regenerative Markov chain. Then there is the adaptive stochastic approximation algorithm [2] which assumes also the Markov reward structure of the rare event. In this algorithm, a trajectory of the Markov chain is simulated, where after each new state the transition probabilities are updated based on a stochastic approximation of the reward. And, the cross-entropy method [14] has been developed as a very successful approach of finding a new probability measure. Briefly, the cross-entropy method for rare event simulation considers determining the change of measure that minimises the Kullback-Leibler divergence (or cross-entropy) from the zero-variance probability measure. In order to solve this program the class of probability measures is restricted by parametrisation. The parameterised cross-entropy optimisation is approximately solved by iteration and simulation.

However, our adaptive importance sampling algorithm is based on the more traditional idea of exponential tilting and following the optimal path as for instance in [12] where a fluid queue with a large number of sources is simulated using importance sampling with a time dependent change of measure.

We conclude with the outline of the paper. In Section 2 we treat the exponential case. Applying large deviations results of the Erlang loss model, we find a time dependent change of measure under which the IS estimator is asymptotically optimal. In Section 3 we consider more general service time distributions. From a practical point of view it is intractable to implement an optimal change of measure. However, we give several approximate importance sampling algorithms which yield very good performance. We support our proposals by simulation experiments.

2 Exponentially distributed service demands

In this section we assume that the service demands have an exponential distribution.

2.1 Large deviations

Suppose that arriving customers are lost when they arrive at moments when exactly $[nb]$ customers are busy. This modified queueing system is the classical Erlang loss model, denoted by $M/M/[nb]/[nb]$. Until the first hitting time of level $[nb]$ the infinite server queue and Erlang's model are stochastically equivalent. We shall use this fact to find asymptotics.

In [16, Chapter 12] it is proved that the large deviations principle applies to the Erlang loss model. Briefly, this means the following. Let $Z_n(t)$ be the number of occupied customers in the $M/M/[nb]/[nb]$ queue at time t. Fix a time horizon τ and consider the scaled processes

$z_n := \{z_n(t) := Z_n(t)/n : 0 \le t \le \tau\}$ $(n = 1, 2, \ldots)$. We call an absolute continuous function $\phi : [0, \tau] \to \mathbb{R}_{\ge 0}$ a path. Then,

- The scaled processes converge (with respect to the sup norm) in probability to a path $\phi_{\rm m}$, therefore called the most likely path, which satisfies the differential equation [16, (12.2)]

$$\phi'_{\rm m}(t) = \gamma - \mu\phi_{\rm m}(t) \quad \text{with} \quad \phi_{\rm m}(0) = 0. \tag{3}$$

  The consequence is that when we simulate the Erlang loss model for large n (in the standard way), almost all scaled realizations stay 'relatively close' to this path. Because of the light traffic assumption, the most likely path remains below level b, i.e., there is no loss. Hence, also most scaled realizations of the infinite server model stay around this path.

- Let U_τ be the set of all paths which start in 0, remain below b, and reach b at the horizon τ. There is a functional $J(\phi)$ on the set of all paths, such that the large deviations asymptotic of the hitting probability holds:

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \log \alpha_n &= \lim_{n\to\infty} \frac{1}{n} \log P(z_n \in U_\tau,\ \tau \in [\tau_1, \tau_2]) \\ &= - \inf_{\tau\in[\tau_1,\tau_2]} \inf_{\phi\in U_\tau} J(\phi). \end{aligned} \tag{4}$$

  There is a specific path $\phi^* \in U_{\tau_2}$, called the optimal path to overflow, which solves this problem [16, Lemma 12.6]:

$$\phi^*(t) = \frac{c}{\mu}\left(e^{\mu t} - 1\right) + \frac{\gamma}{\mu}\left(1 - e^{-\mu t}\right) \quad (0 \le t \le \tau_2), \tag{5}$$

  where c is a constant which takes care of $\phi^*(\tau_2) = b$.

Notice that (4) says that the hitting probability converges to 0 exponentially fast at the rate $J(\phi^*)$. The optimal path has the following property. When n is large, then most scaled realizations hitting b in the time window $[\tau_1, \tau_2]$, stay around the optimal path.

2.2 Change of measure

Let P^* be a probability measure that is implemented for the importance sampling simulations of the queueing model in order to estimate α_n. We will investigate wheter we obtain an efficient (or optimal in the sense of (2)) estimator Y_n^* when most realizations (under P^*) follow the optimal path which we have found in the previous Section.

For each n we define an $M/M/\infty$ queue with arrival rates $\lambda^*(t) = n\gamma^*(t)$ and service rates $\mu^*(t)$ dependent on time, as follows. First, we need the local rate function [16, (7.16)] of the original $M/M/\infty$ process:

$$I(x, y) := \sup_{\theta\in\mathbb{R}} \left(\theta y - \gamma\left(e^{\theta} - 1\right) - x\mu\left(e^{-\theta} - 1\right)\right) \quad (x \ge 0, y \in \mathbb{R}). \tag{6}$$

Denote $\theta(x, y)$ for the optimizer. Then we define

$$\begin{aligned} &\theta^*(t) := \theta(\phi^*(t), \phi^{*'}(t)), \\ &\gamma^*(t) := \gamma \exp(\theta^*(t)), \quad \mu^*(t) := \mu \exp(-\theta^*(t)). \end{aligned} \tag{7}$$

For the new queueing processes we determine the most likely path, similarly as in Section 2.1. This path, $\phi^*_{\rm m}$, satisfies the differential equation (cf. (3))

$$\phi^{*\prime}_{\rm m}(t) = \gamma^*(t) - \mu^*(t)\phi^*_{\rm m}(t) \quad \text{with} \quad \phi^*_{\rm m}(0) = 0. \tag{8}$$

It is an easy exercise to check that the optimal path ϕ^* solves (8). In other words, ϕ^* is the most likely path under the new measure: the scaled processes converge in probability P^* to ϕ^*.

2.3 Monte Carlo Simulation

The process $\{X_n(t) : t \geq 0\}$ of the number of busy servers in the infinite server queue with Poisson arrivals and exponential servers is a Birth & Death process on the nonnegative integers. A realization of $\{X_n(t) : 0 \leq t \leq \tau_2\}$ is generated by drawing the consecutive holding times and new states at the jump times. The unbiased estimator Y_n of α_n based on a single realization is defined as

$$Y_n := 1\{T_n([nb]) \in [\tau_1, \tau_2]\}, \tag{9}$$

where $T_n([nb])$ is the first passage time (1). In the crude Monte Carlo (CMC) experiments we draw N i.i.d. copies of Y_n in the way we sketched above. The CMC estimator is the sample average.

2.4 Importance Sampling

After the change measure of measure of Section 2.2 the queueing process becomes a nonhomogeneous Birth & Death process with transition rates dependent on time:

$$q^*(i, i+1|t) = n\gamma^*(t), \quad q^*(i, i-1|t) = i\mu^*(t).$$

We cannot implement a simulation procedure which generates realizations with this continuous change of rate. As an approximation we update the rates only after the jumps of the process. Specifically, suppose that t is a jump time and that the number of busy servers is i. Then we implement the holding time until the next jump to be exponentially distributed with rate $q^*(i|t) := q^*(i, i+1|t) + q^*(i, i-1|t)$. The next state (at the new jump time) is $i+1$ with probability $q^*(i, i+1|t)/q^*(i|t)$ or $i-1$ with probability $q^*(i, i-1|t)/q^*(i|t)$. In this way we have implemented a new measure P^*.

We denote the new queueing process as $\{X^*_n(t) : t \geq 0\}$. The unbiased estimator Y^*_n of α_n based on a single realization is

$$Y^*_n := L1\{T^*_n(nb]) \in [\tau_1, \tau_2]\}, \tag{10}$$

where $T^*_n(nb])$ is the first passage time of level $[nb]$ of the new queueing process, and $L = dP/dP^*$ the likelihood ratio. The importance sampling Monte Carlo (IS) estimator is the sample average of i.i.d. copies of Y^*_n.

Theorem 2.1. *The estimator Y^*_n is asymptotically optimal.*

Proof. To show (2) it suffices

$$\liminf_{n\to\infty} \frac{\log E^*[(Y^*_n)^2]}{2\log E^*[Y^*_n]} \geq 1.$$

Let $\tau := \tau_2$. Consider a realization of the B & D process under P^*, starting at time $t_0 := 0$ and state $x_0 := 0$. Say that there are m jumps, at times $0 < t_1 < t_2 < \cdots < t_m \approx \tau$. Let $x_1, x_2, \ldots, x_m$

be the consecutive states immediately after these epochs. The likelihood ratio of the realization is

$$\ell = \left(\prod_{k:x_{k+1}-x_k=1} \ell_k[\text{arrival}] \right) \times \left(\prod_{k:x_{k+1}-x_k=-1} \ell_k[\text{departure}] \right). \tag{11}$$

The likelihood ratio factors for arrivals and departures are calculated using the transition rates of the B & D processes:

$$\ell_k[\text{arrival}] = \frac{n\gamma \, \exp\Big(-(n\gamma + x_k\mu)(t_{k+1}-t_k)\Big)}{n\gamma^*(t_k) \, \exp\Big(-(n\gamma^*(t_k) + x_k\mu^*(t_k))(t_{k+1}-t_k)\Big)},$$

and

$$\ell_k[\text{departure}] = \frac{x_k\mu \, \exp\Big(-(n\gamma + x_k\mu)(t_{k+1}-t_k)\Big)}{x_k\mu^*(t_k) \, \exp\Big(-(n\gamma^*(t_k) + x_k\mu^*(t_k))(t_{k+1}-t_k)\Big)}.$$

Substituting these in (11), applying the new rates (7) and using [16, Appendix C]

$$n\gamma + x_k\mu - n\gamma^*(t_k) - x_k\mu^*(t_k) = nc$$

(where c is the constant in (5)), we get

$$\ell = \exp\left[nc\tau - \sum_{k=1}^{m} \theta^*(t_k) \Big(1\{\text{arrival at } t_k\} - 1\{\text{departure at } t_k\} \Big) \right].$$

Almost all realizations are close to path ϕ^*_τ (under P^*). Consequently:

$$\sum_{k=1}^{m} \theta^*(t_k) \frac{1\{\text{arrival at } t_k\} - 1\{\text{departure at } t_k\}}{n} \to \int_0^\tau \theta^*(t)\phi^{*'}(t)\, dt \tag{12}$$

and $1\{T^*_n(nb]) \in [\tau_1, \tau_2]\} \to 1$ in probability (as $n \to \infty$). When we denote the randomization of (12) by using capital letters, we get

$$\begin{aligned}
\frac{1}{n} \log E^*[Y^*_n] &= \frac{1}{n} \log E^*[L 1\{T^*_n(nb]) \in [\tau_1, \tau_2]\} \\
&= c\tau + \frac{1}{n} \log E^* \exp\left[-n \sum_{k=1}^{M} \theta^*(T_k) \right. \\
&\quad \left. \times \frac{1\{\text{arrival at } T_k\} - 1\{\text{departure at } T_k\}}{n} \right] 1\{T^*_n(nb]) \in [\tau_1, \tau_2]\} \\
&\to c\tau - \int_0^\tau \theta^*(t)\phi^{*'}(t)\, dt.
\end{aligned}$$

The same line of reasoning gives

$$\frac{1}{n} \log E^*[(Y^*_n)^2] \to 2c\tau - 2\int_0^\tau \theta^*(t)\phi^{*'}(t)\, dt.$$

□

n	$\hat{\alpha}_n$	RE(CMC)	RE(IS)	ratio(IS)	gain
10	3.19e-02	0.006	0.029	0.794	5.37e+00
20	7.76e-03	0.011	0.021	0.920	3.89e+01
30	1.50e-03	0.026	0.018	0.964	2.91e+02
40	2.24e-04	0.067	0.017	0.975	1.73e+03
50	3.70e-05	0.164	0.017	0.984	1.14e+04
75	3.53e-07	1.684	0.018	0.984	1.16e+06
100	3.08e-09	18.03	0.020	0.983	1.10e+08
150	2.16e-13	2151.4	0.026	0.979	8.80e+11
200	1.37e-17	270142	0.038	0.974	6.63e+15

Table 1: Estimates and estimator performance for exponential servers.

2.5 Results

The data are

$$\gamma = 0.5,\ \mu = 1.0,\ b = 1.0,\ \tau_1 = 5.0,\ \tau_2 = 5.5. \tag{13}$$

Simulations were executed for increasing values of the scaling factor n. The performance of the CMC and IS estimators (9) and (10) were compared threefold.

- The relative error of the sample average estimator, where the number N of runs is 1,000,000 for the CMC and 5000 for the IS. The relative error of the CMC estimator grows exponentially (in n). This is a consequence of (4). Since the IS estimator is asymptotically optimal we may expect that its relative error grows at most polynomially.
- The ratio in the left handside of (2). This ratio tends to 0.5 for the CMC estimator. The closer to 1, the better the estimator is, i.e., giving more variance reduction.
- The efficiency gain, which is defined by

$$\frac{\mathrm{Var}[Y_n] \times \mathrm{CPU}[Y_n]}{\mathrm{Var}^*[Y_n^*] \times \mathrm{CPU}[Y_n^*]},$$

  where CPU$[Y_n]$ means the computing time of a single realization in the CMC simulations. Clearly, the larger the gain is, the more variance reduction we have.

Table 1 summarizes the results. The given estimates $\hat{\alpha}_n$ of the hitting probability are the CMC estimates until $n = 50$ and the IS estimates for $n \geq 60$. For these larger n the CMC estimates are unreliable because the sample size $N = 1,000,000$ is too small, which gives no or a few observations per run. However, the relative errors RE(CMC) can be estimated using $\mathrm{Var}[Y_n] = \alpha_n(1-\alpha_n)$ and using the IS estimates of α_n. Similarly we estimate the gain.

Table 1 shows that the IS relative errors remain 'almost' bounded. There is a slight increase for very large n. The ratio (2) becomes almost 1, showing asymptotic optimality. The gain is huge for large n.

2.6 Accelerating the importance sampling

The optimal tilting factor $\theta^*(t)$ is the solution of the problem (6):

$$\theta^*(t) = \log \frac{\mu x(t) + y(t) + \gamma + c}{2\gamma}, \tag{14}$$

n	K	gain(i)	gain(ii)	n	K	gain(i)	gain(ii)
50	5	0.48	1.84	150	5	0.21	4.67
	11	1.11	1.26		11	0.84	1.59
	22	1.32	1.21		22	1.15	1.45
	55	1.40	1.21		55	1.32	1.08
100	5	0.31	2.94	200	5	0.38	6.41
	11	1.04	1.42		11	0.65	1.77
	22	1.29	1.25		22	1.55	1.47
	55	1.40	1.27		55	1.44	1.14

Table 2: The gain of the accelarated IS methods with respect to IS.

where $x(t) = \phi^*_{\tau_2}(t)$ the location of the optimal path (5) at time t, $y(t) = x'(t)$ the derivative, and c the constant in (5). Simple calculus shows that $\theta^*(t)$ is an increasing function. We propose two approximation schemes of this function.

(i) Partition $[0, \tau_2]$ in K equal subintervals $[t_k, t_{k+1}]$ ($k = 0, 1, K-1$, $t_k = k\tau_2/K$). Whenever t lies in the k-th subinterval, we apply the tilting factor at the midpoint of the subinterval, i.e., $\theta^*((t_k + t_{k+1})/2)$. In this scheme we calculate 'off-line' the K new arrival and service rates associated with the midpoint tilting factor. At each jump time t in a simulation run we have to detect only the subinterval covering t.

(ii) The function $\theta^*(t)$ is approximated by a piecewise linear interpolation function, where the subintervals are the same as in (i). We calculate 'off-line' the tilting factors $\theta_k := \theta^*(t_k)$ at the knots t_k. The applied tilting factor at time $t \in [t_k, t_{k+1}]$ becomes $\theta_k + (\theta_{k+1} - \theta_k)(t - t_k)/(t_{k+1} - t_k)$.

Both methods give a new probability measure wich approximates the asymptotically optimal P^*, and therefore, they will be suboptimal. They accelerate the computation times of importance sampling realizations against a (slight) variance increase.

Table 2 gives results of experiments with various values of the number K of subintervals. We compare the gains only (the gain with respect to the IS method of Section 2.4). In the experiments we take the same seed of the number generator for each n and modification. From the numbers in Table 2 we conclude that the linear interpolation approximation method with a small number of subintervals gives the best performance. The gain in method (i) becomes higher for larger number of subintervals, but remains below 2.0.

3 General service demands

Now we assume that the service demands are generally distributed. There are no large deviations results known, and thus, the optimal path is not available. We conjecture that an optimal change of measure would require a continuous update of the arrival and the service time distributions of all the customers present in the system, as in the exponential case. However, even when we would know the continuous optimal tilting factors, it is not practical to implement an algorithm because the distributions do not have the memoryless property. It would require a huge amount of computing time to calculate likelihood ratios. Therefore, we have implemented an approximate importance sampling algorithm which is based on the following heuristics.

Heuristic 1
New distributions for the random variables are implemented after an arrival only.

Heuristic 2
A new distribution is implemented only for the next arrival time, and for the service demand of the newly arrived customer. The distributions of all other customers present at the arrival time remain unchanged.

Heuristic 3
The new distributions are calculated by exponentially tilting with a tilting factor which is determined by the tilting factors (14) of the exponential model (see Section 3.1 below).

These heuristics are based on the observation that the IS simulations in the exponential model with a few updates of the distributions give good performance. Furthermore, although the transient probabilities in the $M/G/\infty$ model are sensitive for higher moments of the service demands, we expect this sensitivity to be rather low.

3.1 The new distributions

The service demands have distribution function G and mean $1/\mu$. Consider a typical realization in the IS simulations (with scaling factor n). Suppose that t is an arrival time of a customer, then we draw the next interarrival time similarly as in the exponential model: (i) we calculate the tilting factor $\theta^*(t)$ as in (14), and (ii) we generate a realization of an exponential distribution with rate $n\gamma \exp(\theta^*(t))$. Also, we draw a realization of the service demand of the arrived customer. This new customer has distribution function which is obtained by exponentially tilting in such a way that the service rate becomes $\mu \exp(-\theta^*(t))$. Notice that this distribution depends on the arrival time t. All other ongoing services remain unchanged. Let us work out the new distributions for the following specific distributions.

Deterministic
In this case the service demand is degenerated at $1/\mu$, which does not change in importance sampling simulations.

Gamma
Let G be the Gamma distribution with shape parameter $\nu > 0$ and scale parameter $\alpha > 0$. The density function is

$$g(x) = \frac{1}{\Gamma(\alpha)} \nu(\nu x)^{\alpha-1} \exp(-\nu x) \quad (x \geq 0). \tag{15}$$

Notice that $\nu = \alpha\mu$ in order to have mean $1/\mu$. The exponentially tilted density with tilting factor δ is $g^\delta(x) = \mathit{constant} \times e^{\delta x} g(x)$, where the *constant* makes g^δ a density. A simple calculus exercise shows that we can transform $g^\delta(x)$ in a form like (15), with shape parameter $\nu - \delta$ (and the same scale parameter α). In other words, the service time has mean $\alpha/(\nu - \delta)$ after exponentially δ-tilting. Equating to $1/(\mu \exp(-\theta^*(t)))$ we obtain the tilting factor $\delta = \delta(t) = \alpha\mu(1 - \exp(-\theta^*(t)))$. Hence, the new distribution is Gamma, with shape parameter $\nu \exp(-\theta^*(t))$ and scale parameter α.

Coxian-2
The Coxian-2 distribution with parameters (β, μ_1, μ_2) has density function

$$g(x) = (1-\beta)\mu_1 e^{-\mu_1 x} + \beta\mu_1 e^{-\mu_1 x} * \mu_2 e^{-\mu_2 x} \quad (x \geq 0),$$

where $\beta \in [0,1]$ and $*$ denotes convolution. The mean equals $(1-\beta)/\mu_1+\beta(1/\mu_1+1/\mu_2)$. The exponentially tilted density $g^\delta(x)$ is again Coxian-2 with parameters

$$\beta^\delta = \frac{\beta\mu_2}{b\mu_2+(1-\beta)(\mu_2-\delta)}, \quad \mu_1^\delta = \mu_1-\delta, \quad \mu_2^\delta = \mu_2-\delta.$$

Equating the tilted service rate to $\mu\exp(-\theta^*(t))$ yields a 3-rd degree equation in δ. When we would solve this equation after each jump time of the process, we would get an enormous increase in computing time. Therefore, we have chosen to solve these equations in the knots t_k of a partition $0 = t_0 < t_1 < \cdots < t_K = \tau_2$. The tilting factor δ at an intermediate time t is determined as the linear interpolation of these δ's. This procedure is the same as in Section 2.6

3.2 Results

We assume the same data (13) as in Section 2.5. We distinguish between different service time distributions through their coefficient of variation. Let S be a generic service time, then the squared coefficient of variation of S is denoted by VC, and defined as $\mathrm{VC} := \mathrm{Var}[S]/(E[S])^2$. Given the mean $1/\mu$ and the coefficient of variation VC of the service time S we can easily fit distributions [17, Appendix B]. Specifically: $\mathrm{VC}=0$ gives deterministic, when $0<\mathrm{VC}<1$ we fit Gamma (with $\alpha>1$), $\mathrm{VC}=1$ exponential, and $\mathrm{VC}>1$ Coxian-2. (Other distributions are possible for $\mathrm{VC}>0$.)

We give the results for $\mathrm{VC}=0, 0.25$ and 5. The results of the exponential case $\mathrm{VC}=1$ can be found in Tables 1 and 2. In all three cases we first ran the CMC simulations upto estimates of the order 10^{-4}, and then the IS simulations upto 10^{-15} estimates. The estimates in Table 3 are these IS estimates obtained after N simulation runs, where N differs (10000 deterministic and Gamma, 20000 Coxian). Furthermore, Table 3 reports the relative errors, the ratios (2), and the gains (with respect to the CMC estimates) obtained by the importance sampling algorithms.

Table 3 shows that the IS algorithms give an enormous variance reduction. The relative errors increase, but at a subexponential rate. The ratios are well above 0.5 but remain below 0.9 which says that the variance reduction is suboptimal. As one would expect, the larger the variability of the service time (higher VC), the more the performances of the estimators degrade. Also notice that the relative errors in the Coxian-2 case are getting high. This means that the sample size of 20000 is still too small. Another explanation might be that the optimal (unknown) path in the Coxian case deviates far from the 'exponential' path that we used in the algorithm.

4 Conclusion

We have developed a fast importance sampling algorithm for rare event simulations in the infinite server queue. In case of exponential servers, the algorithm is a linear interpolation approximation of an asymptotically optimal importance sampling algorithm that resulted from the large deviations analysis of the queueing model. In case of general servers, the algorithm exploits the solution to the exponential case and applies three heuristics for simplification.

References

[1] J. Abate and W. Whitt, "Calculating transient characteristics of the Erlang loss model by numerical transform inversion", *Stochastic Models* 14, pp. 663–680, 1998.

	VC = 0 10000 runs	VC = 0.25 10000 runs	VC = 5 20000 runs
n	10	10	10
$\hat{\alpha}_n$	3.13e-02	3.42e-02	1.53e-02
RE	0.029	0.028	0.029
ratio	0.698	0.699	0.660
gain	2.55e+00	3.18e+00	2.37e+00
n	50	50	30
$\hat{\alpha}_n$	4.21e-05	4.65e-05	7.45e-05
RE	0.058	0.072	0.163
ratio	0.826	0.802	0.670
gain	5.84e+02	3.60e+02	1.19e+01
n	100	100	60
$\hat{\alpha}_n$	3.57e-09	3.75e-09	8.89e-09
RE	0.164	0.098	0.903
ratio	0.856	0.882	0.738
gain	7.03e+05	1.86e+06	3.90e+03
n	140	140	80
$\hat{\alpha}_n$	2.42e-12	1.59e-12	2.32e-12
RE	0.288	0.156	0.774
ratio	0.874	0.899	0.825
gain	2.94e+08	1.61e+09	1.93e+07
n	170	170	100
$\hat{\alpha}_n$	1.47e-15	5.41e-15	1.32e-15
RE	0.270	0.189	0.712
ratio	0.904	0.911	0.865
gain	5.03e+11	3.10e+11	3.88e+10

Table 3: IS performances.

[2] T. Ahamed, V. Borkar, and S, Juneja, "Adaptive importance sampling technique for Markov chains using stochastic approximation", *Operations Research* 54, 2006.

[3] S. Asmussen, *Ruin Probabilities*, World Scientific, 2000.

[4] S. Asmussen and R. Rubinstein, "Steady state rare event simulation in queueing models and its complexity properties". In J. Dshalalow (ed.), *Advances in queueing theory, theory, methods and open problems*, pp. 429–461, CRC Press, 1995.

[5] C.-S. Chang, P. Heidelberger, S. Juneja, and P. Shahabuddin, "Effective bandwidth and fast simulation of ATM intree networks", *Performance Evaluation* 20, pp. 45–65, 1994.

[6] P. Desai and P. Glynn, "A Markov chain perspective on adaptive Monte Carlo algorithms". In *Proceedings of 2001 Winter Simulation Conference* (eds. B. Peters, J. Smith, D. Medeiros, and M. Rohrer), IEEE Press 2001.

[7] P. Glasserman and S-G. Kou, "Analysis of an importance sampling estimator for tandem queues", *ACM Transactions on Modeling and Computer Simulation* 5, pp. 22–42, 1995.

[8] P. Heidelberger, "Fast simulation of rare events in queueing and reliability models", *ACM Transactions on Modelling and Computer Simulation* 5, pp. 43–85, 1995.

[9] J. Keilson, *Markov Chain Models — Rarity and Exponentiality*, Springer-Verlag, 1979.

[10] F. Kelly, *Reversibility and Stochastic Networks*, Wiley, 1979.

[11] G. Kesidis and J. Walrand, "Quick simulation of ATM buffers with on-off multiclass Markov fluid sources", *ACM Transactions on Modeling and Computer Simulation* 3, pp. 269–276, 1993.

[12] M. Mandjes and A. Ridder, "A large deviations analysis of the transient of a queue with many Markov fluid inputs: approxiations and fast simulation", *ACM Transactions on Modeling and Computer Simulation* 12, pp. 1–26, 2002.

[13] S. Parekh and J. Walrand, "A quick simulation method for excessive backlogs in networks of queues", *IEEE Transactions of Automatic Control* 34, pp. 54–66, 1989.

[14] R. Rubinstein and D. Kroese, *The cross-entropy method*, Springer, 2004.

[15] J.S. Sadowsky, "Large deviations theory and efficient simulation of excessive backlogs in a $GI/GI/m$ queue", *IEEE Tranactions on Automatic Control* 36, pp. 1383–1394, 1991.

[16] A. Shwartz and A. Weiss, *Large Deviations for Performance Analysis: Queues, Communications and Computing*, Chapman Hall, 1995.

[17] H.C. Tijms, *Stochastic Models: an Algorithmic Approach*, Wiley, 1994.