Large Deviations and Importance Sampling of the M/M/∞ Queue

This paper describes Monte Carlo simulations for the estimation of the transient probability that the infinite server queue with exponentially distributed servers reaches high levels within a predefined time window. Under the light traffic assumption this is a rare event. The simulations are accelerated by applying importance sampling.

Ad Ridder

studeerde wiskunde aan de Universiteit van Amsterdam, en promoveerde in 1987 aan de Universiteit van Leiden. Sinds 1992 is hij verbonden aan de afdeling Econometrie van de Vrije Universiteit Amsterdam als universitair hoofddocent. In de tussenliggende vijf jaren was hij wetenschappelijk onderzoeker bij een Nederlands softwarebedrijf, gastdocent op de University of California at Berkeley en docent bij de faculteit Bedrijfskunde van de Erasmus Universiteit in Rotterdam.

Introduction

The infinite server queue, denoted by M/G/∞, 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 capacaties. 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 server 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 full statistical information about the chances of a full buffer (in the finite system). In this paper we investigate this problem in case of exponentially distributed servers only, i.e., the M/M/∞ model. We shall point out an exact approach which leads to a numerical algorithm, a large deviations approach which gives rough approximations, and an efficient importance sampling simulation for accurate estimates.

The model

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/M/∞ model with Poisson $(n\gamma)$ arrivals, and with exponentially distributed service demands with rate μ. We assume light traffic, i.e., $\gamma < b\mu$.
The first passage times of the n-th system are

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

where $X_n(0)=0$. The target probabilities are

$$l_n^{(1)} = P(T_n([nb]) \leq \tau), \quad l_n^{(2)} = P(T_n([nb]) \in (\tau_0, \tau]),$$

for some time horizon $0<\tau_0<\tau$, some (scaled) overflow level $b>0$, and for large n. In other words, we consider M/M/∞ queues where the arrival rates $\lambda=n\gamma$ and the hitting levels $B=[nb]$ are growing proportionally to fixed constants γ and b.
The Laplace transform of the first passage time density from state 0 to state $[nb]$ is the product of the Laplace transforms of the densities of going just one level higher:

$$\upsilon_{0,[nb]}(s) = \prod_{i=0}^{[nb]-1} \upsilon_{i,i+1}(s).$$

These transforms satisfy a recurrence relation [6],

$$\upsilon_{i,i+1}(s) = \frac{\lambda}{\lambda + i\mu + s - i\mu\upsilon_{i-1,i(s)}}$$

Thus, the Laplace transform of the ccumulative distribution function of the first passage time is $\upsilon_{0,[nb]}(s)/s$, which we invert using an algorithm developed by Den Iseger [5].

Large deviations

Consider the scaled processes: $z_n=\{z_n(t)=X_n(t)/n:0\leq t\leq\tau\}$, where $(n=1,2,...)$. Convergence of these processes are studied in [12]. Let Φ be the set of absolute continuous functions $\phi:[0,\tau]\rightarrow\mathbb{R}_{\geq 0}$. Then the scaled processes converge in probability to a specific $\phi_m\in\Phi$, called the most likely path:

$$\lim_{n\to\infty} P(\|z_n-\phi_m\|<\varepsilon)=1 \qquad (2)$$

The consequence is that when we simulate the queueing model for large n (in the standard way), almost all realizations stay ‘relatively close’ to the function $n\phi_m$ (see Figure 1). In fact,

$$\phi_m(t)=\frac{\gamma}{\mu}(1-e^{-\mu t}) \quad (0\leq t\leq\tau)$$

However, we are interested in the event that the process reaches high levels. Consider any $\phi\in\Phi$ which starts in 0, and reaches b at or before the horizon τ, called an overflow path. The probabilities that the scaled processes stay close to ϕ satisfy a large deviations convergence [12]:

$$\lim_{n\to\infty}\frac{1}{n}\log P(\|z_n-\phi\|<\varepsilon)=-J(\phi),$$

where $J(\phi)$ is a functional on the set Φ. Among all these overflow paths, there is a unique one that minimizes the functional. It is called the optimal path to overflow, denoted by ϕ^*, and it has the form

$$\phi^*(t)=\frac{c}{\mu}(e^{\mu t}-1)+\frac{\gamma}{\mu}(1-e^{-\mu t}) \quad (0\leq t<\tau) \qquad (3)$$

where c is a constant which takes care of $\phi*(t)=b$. We have shown in [10] that the hitting probabilities converge logarithmically:

$$\lim_{n\to\infty}\frac{1}{n}\log I_n=\lim_{n\to\infty}\frac{1}{n}\log P(\|z_n-\phi^*\|<\varepsilon)=-J(\phi) \quad (4)$$

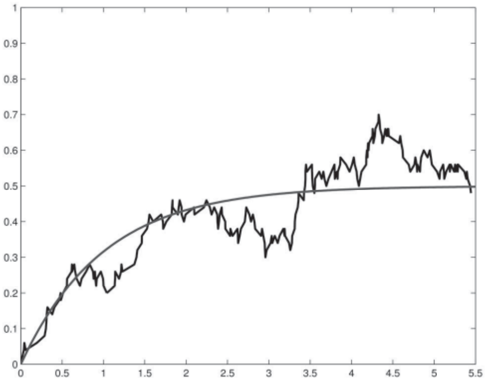

Figure 1. Most likely path and a typical scaled sample path.
Parameters: γ=0.5,μ=1,b=1,τ=5.5,n=50.

(This holds for both target probabilities!) When we simulate the infinite server queueing model for large n, and when we find a realization which hits level nb for the first time at or before horizon τ, then - almost certain - this realization stays ‘close’ to the function $n\phi^*$ all the time (see Figure 2).
The large deviations asymptotic (4) can be used to give a rough approximation of the hitting probability: $I_n^{(2)}=e^{-nJ^*}$, where $J^*=J(\phi^*)$ for which a closed form expression is available [8]. Figure 3 shows the relative errors of this approximation when compared to the exact computations of section 2. That the error increases for larger n might be caused by the numerical instability of the recursion of the Laplace transforms. Thus, for larger n we need other approaches to get more reliable estimates.

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 easy to simulate, since it is a Markov chain with exponential holding times. The unbiased estimator Y_n of $I^{(2),n}$ based on a single realization is defined as

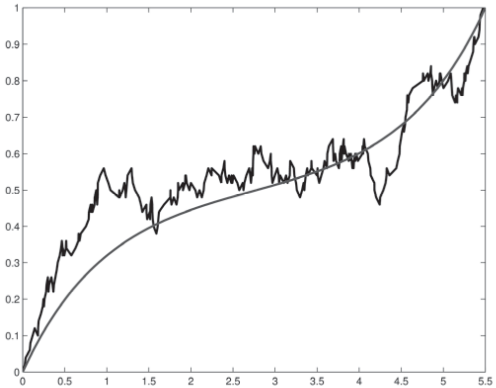

Figure 2. Optimal path and a typical scaled sample path.
Parameters: γ=0.5,μ=1,b=1,τ=5.5,n=50.

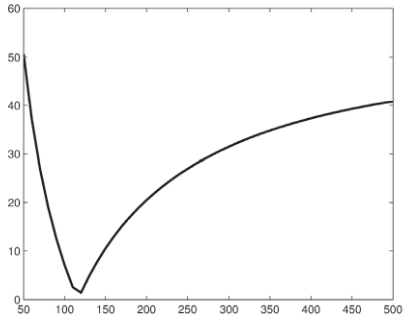

Figure 3. Relative errors (in percentages) made by the large deviations approximations.
Parameters: γ=0.5,μ=1,b=1,
τ_0=5.0,τ=5.5,n=50,...,500

$$Y_n = 1\{T_n(|nb|) \in (\tau_0, \tau])\}.$$

In the crude Monte Carlo (CMC) simulation we draw k i.i.d. copies of Y_n whose average is the CMC estimator. For large n the target probability becomes too small for the CMC to be tractable. For instance, in the scenario of Figure 1, when n=100 the hitting probability is $l^{(2),n}=3.11\times10^{-9}$. Then to obtain an estimate for which the 95% confidence interval has a relative width of less than 20%, we need to simulate about 3×10^{10} samples. It would take on my 2.79 Ghz, 1MB RAM PC about 52 days.

Importance Sampling

Importance sampling is a variance reduction technique to overcome the problem of many simulation runs without succesful observations. 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.
Finding a good new probability P^* is the main issue in importance sampling. The criterion is to keep the relative error, $\sqrt{Var^*[Y^{*,n}]}\,/\,E^*[Y^{*,n}]$ as small as possible. The best performance is obtained when the relative error remains bounded as $n\to\infty$. Then the number of samples 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 asymptotical optimality [4]:

$$\lim_{n\to\infty}\frac{\log E^*[(Y_n^*)^2]}{\log E^*[Y_n^*]} = 2 \qquad (5)$$

This yields good performance and considerable variance reductions, and typically, the relative error increases polynomially. A way to find a good new measure P^* is to implement an exponential change of measure [1, 4]. That is, the distribution functions of the random variables are exponentially tilted. For overflow problems in queueing systems such as M/G/c it is already a long time well-known how to find an optimal exponential change of measure after a large deviations analysis of the problem, see for instance the early papers of [2, 3, 7, 9, 11]. The resulting importance sampling algorithm is static, in the sense that the change of measure induces fixed new statistical laws to generate the samples throughout the entire simulation.

Dynamic change of measure

We consider a family of twisted probability measures P^θ that are candidates to be implemented for executing importance sampling simulations of the queueing model in order to estimate l_n. The probability measures are induced by constructing the statistical laws of stochastic processes associated with the queueing model. Let $\theta:[0,\tau]\to\mathbb{R}_{>0}$ be a nondecreasing continuous function, called the twisting function. Then we define for $0\le t\le\tau$ the twisted arrival and service rates by

$$\gamma^\theta(t) = \gamma e^{\theta(t)}, \quad \mu^\theta(t) = \mu e^{-\theta(t)}$$

For each n we consider the Markov jump process $\left\{\left(X_n^\theta(t), M_n^\theta(t)\right): t \ge 0\right\}$, where $X_n^\theta(t)$ represents the number of busy servers in the infinite server queue at time t, and $M_n^\theta(t)$ is the time epoch of the last jump of $\left\{X_n^\theta(.)\right\}$ before or at time t. Suppose that $M_n^\theta(t)=s$ was the last jump time of the process before epoch t, bringing the number of busy servers to $X_n^\theta(s)=i$. Then the holding time until the next jump is exponentially distributed with rate $q_n^\theta(i,s) = n\gamma^\theta(s) + i\mu^\theta(s)$. The next state (at the new jump time) is i+1 with probability $n\gamma^\theta(s)/q_n^\theta(i,s)$ or i-1 with probability $i\mu^\theta(s)/q_n^\theta(i,s)$. In this way we have implemented a new measure P^θ which is indirectly determined by these state-time dependent arrival and service rates.
Similar to section 3, the scaled processes $z_n^\theta = \left\{z_n^\theta(t) = X_n^\theta(t)/n : 0 \le t \le \tau\right\}$ converge in probability to a deterministic most likely path

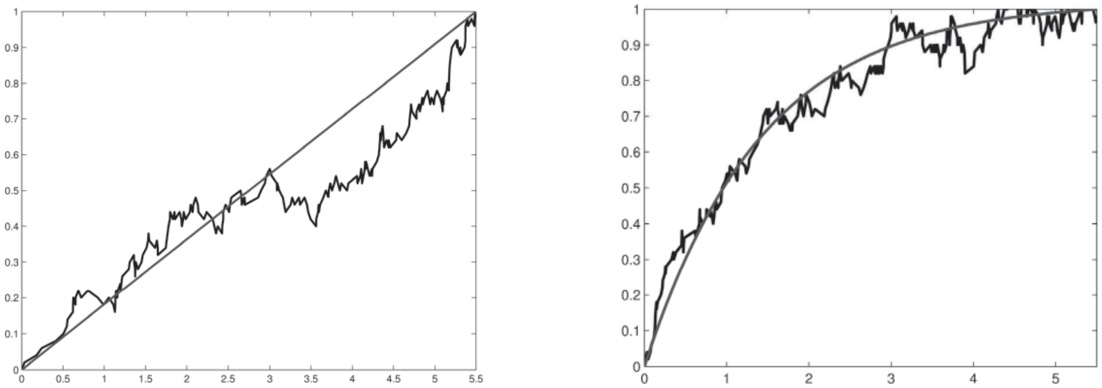

Figure 4. Two implementations of a change of measure, showing the most likely path and a typical scaled sample path. Left: most likely path is a straight line. Right: the twisting function is constant. Parameters: γ =0.5,μ=1,b=1,τ=5.5,n=50.

ϕ_m^θ. The consequence is that when we execute importance sampling simulations of the infinite server queueing model using the new probability measure P^θ, almost all realizations stay 'relatively close' to the function $n\phi_m^\theta$. As an example, we have implemented two different twisting functions $\theta(.)$. In the first implementation the most likely path fq,m is the straight line to overflow at time ϕ_m^θ:

$$\phi_m^\theta(t) = \frac{b}{\tau}t.$$

In the second example we choose the twisting function constant $\theta(.) \equiv \theta$. Both implementations are depicted in Figure 4: the most likely path, and a typical scaled sample path.
Now suppose that we can find a twisting function $\theta^*(.)$, such that for the associated new measure P^* holds that the most likely path is exactly the optimal path ϕ^* to overflow under the original measure P (see (3)):

$$\lim_{n\to\infty} P^*\left(\left\|z_n - \phi^*\right\| < \varepsilon\right) = 1$$

When we do all the calculus, based on the large deviations expressions in [12], we find that

$$\theta^*(t) = \theta\left(\phi_\tau^*(t), \phi_\tau^{*'}(t)\right), \quad 0 \le t \le \tau$$

with $\theta(x,y) = \dfrac{y + \sqrt{y^2 + 4x\gamma\mu}}{2\gamma}$.

In [10] it is proved that the associated importance sampling estimator Y_n^* is asymptotically optimal.

Simulation experiments

We consider here the window problem of estimating $l_n = l_n^{(2)}$. The data are

γ = 0.5, μ = 1.0, b = 1.0, τ_0 = 5.0, τ = 5.5.

Simulations were executed for increasing values of the scaling factor n. The sample sizes were 1,000,000 (CMC) and 5,000 (IS). The performances of the CMC and IS estimators were compared as follows.

- The relative error $\delta(.)$ of the estimate with respect to the exact value, see section 2.
- The relative error $RE(.)$ of the estimator, i.e., the ratio of its standard deviation to its mean of the sample average estimator.
- The ratio in the left handside of (5). This ratio tends to 1 for the CMC estimator. The closer to 2, the better the estimator is, i.e., giving more variance reduction.
- The efficiency gain, which is defined by

  $$\frac{Var[Y_n]xCPU[Y_n]}{Var^*[Y_n^*]xCPU[Y_n^*]}.$$

  Clearly, the larger the gain is, the more variance reduction we have.

Table 1 summarizes the results. The CMC was executed upto n=50, because for larger n there were no overflow observations. However, the relative errors RE(CMC) can be estimated using Var[Y_n]=l_n(1-l_n) and using the exact values of l_n. Similarly we estimated the gain by running small sample sizes for the timing.
Table 1 shows that the IS relative errors increase at a very low pace. The ratio (5) stabilizes around 1.95, close to 2, showing asymptotic optimality. The gain is huge for large n. Also we see that for large n (n ≥ 400) the estimates seem to degrade, but this effect is due to the numerical errors of the recursion and the inversion in the exact method.

n	l_n	δ(CMC)	δ(IS)	RE(CMC)	RE(IS)	Ratio(IS)	gain
10	3.20E^{-002}	0.48	0.12	0.55	2.85	1.53	1.3
20	7.79E^{-003}	0.17	0.19	1.13	2.10	1.76	10
30	1.43E^{-003}	1.77	0.99	2.61	1.82	1.85	69
40	2.39E^{-004}	1.93	2.18	6.40	1.76	1.89	463
50	3.83E^{-005}	4.41	1.90	15.81	1.67	1.91	3069
100	3.11E^{-009}		1.92	1793	2.04	1.94	2.34E^{+007}
150	2.15E^{-013}		0.43	215677	2.75	1.95	1.92E^{+011}
200	1.40E^{-017}		0.29		3.85	1.95	1.53E^{+015}
250	8.77E^{-022}		8.94		4.79	1.95	1.87E^{+019}
300	5.40E^{-026}		7.36		5.89	1.95	1.95E^{+023}
350	3.27E^{-030}		4.47		11.92	1.94	6.20E^{+026}
400	1.97E^{-034}		12.23		6.93	1.96	4.31E^{+031}
450	1.17E^{-038}		19.28		7.20	1.96	7.90E^{+035}
500	6.94E^{-043}		11.68		12.85	1.95	3.51E^{+039}

Table 1: Estimates and estimator performance for exponential servers (relative errors δ and RE are in percentages.

Alternative implementations of importance sampling

In this section we consider the following alternative implementations.

- The twisting function $\theta(.)$ is determined by the the straight line $\phi(t)=bt/\tau$ to overflow, see Figure 4. This gives faster execution time. However, the estimates turn out to be of poor quality and show relative errors δ up to 100%.
- The twisting function $\theta(.)\equiv\theta$ is constant. The estimates are even worse than the previous alternative.
- A fast execution (faster than the optimal of the previous section) with good estimates is obtained by calculating off-line the optimal twisting function $\theta^{*}(.)$ at a finite number of points, and then apply linear interpolation. We have implemented this approach while using 10 upto 50 subintervals of the simulation period [0,τ]. The results are rather "insensitive" to the number of intervals and show the following performance of the estimator. The relative errors and the log ratio of this linear estimater are approximately the same as the corresponding performance of the continuous IS estimator. There is an improvement of the gain, where Table 2 give the gain with respect to the continuous estimator. Notice the outlier which may occur because we ran each simulation experiment just once.

n	δ	*RE*	*ratio*	*gain*
50	1.45	1.68	1.91	3.40
100	2.27	1.98	1.94	4.51
150	8.18	2.75	1.95	4.33
200	0.77	3.74	1.95	4.36
250	4.81	5.53	1.94	2.09
300	7.44	9.30	1.93	9.05
350	7.18	10.66	1.94	26.96
400	5.03	19.17	1.93	1.46
450	5.67	18.76	1.94	0.17
500	10.59	9.17	1.96	5.47

Table 2: Performance of the IS estimator when the optimal twisting function $\theta^{*}(.)$ is approximated lineraly on 30 subintervals. Gain with respect to the original IS estimator

Conclusion

We have developed a fast importance sampling algorithm for rare event simulations in the infinite server queue with 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. The algorithm is generalized to queueing models with generally distributed servers in [10].

References

[1] Asmussen, S. and Rubinstein, R. (1995). *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, CRC-Press, 429–461.

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

[3] Frater, M.R., Lennon, T.M. and Anderson, B.D.O. (1991). Optimally efficient estimation of the statistics of rare events in queueing networks, *IEEE Transactions on Automatic Control* **36**, 1395-1405.

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

[5] den Iseger, P. (2006). Numerical inversion of Laplace transforms using a Gaussian quadrature for the Poisson summation formula, *Probability in the Engineering and Information Sciences*, **20**, 1-44.

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

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

[8] Mandjes, M. and Ridder, A. (2001).A large deviations approach to the transient of the Erlang loss model, *Performance Evaluation* **43**, 181–198.

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

[10] Ridder, A. *Large deviations and importance sampling simulation of first passage times rare-events in the infinite server queue*, submitted, 2007.

[11] Sadowsky, J.S. (1991). Large deviations theory and efficient simulation of excessive backlogs in a GI/GI/m queue, *IEEE Tranctions on Automatic Control* **36**, 1383–1394.

[12] Shwartz, A. and Weiss, A. (1995). L*arge Deviations for Performance Analysis: Queues, Communications and Computing*, Chapman Hall.