Fast simulation of retrial queues

Ad Ridder, VU *

Abstract

This paper describes a discrete-time retrial queue and shows how importance sampling simulations can be applied for estimating the probability of large orbit content and the overflow fraction of primary calls.

1 Introduction

A retrial queue is a queueing system consisting of two queueing subsystems or components, a primary and a secondary. Customers arrive from outside the system at the primary component which has finitely many servers and a finite waiting room. When the primary component is fully occupied, arriving customers are directed to the secondary component, also called the orbit. The customers in the orbit retry after a random time to get access to the primary component. Retrialing customers merge with the customers arriving from outside.

This typically describes what is happening in telephone or other communication networks. The customers are calls and when calls cannot be accomodated after set up due to occupied lines in the network, they are usually tried later. They are not lost as is often modeled in so-called loss networks. In these telecommunication networks the primary component consists of a servicing mechanism only, it has no space for buffering or waiting. The orbit is infinitely large, in principle, as every call will retry and not be lost. However, models have been developed to cope with a fraction of calls which do not retry but leave the system. For these and other models, for the analysis of these models, as well as for references to studies of retrial queues, we refer to the excellent book of Falin and Templeton [4]. Another source is the bibliography composed by Artalejo [1].

Our description allows waiting of customers in the primary component. This might be considered to model a service center such as a call center, where calls are put in a buffer (if there is space) when all agents are busy. However, again calls come in an orbit when also the buffer is full, and they try to get access later. Another well known phenomenon in telephony is impatient behavior of callers. Hence, buffered calls leave the primary component when they

*Department of Econometrics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, the Netherlands. Email: aridder@econ.vu.nl

wait too long, enter the orbit and try later. One may model this impatience approximately by setting the buffer sufficiently small.

Traditionally one considers performance measures such as

- the probability of a fully occupied primary component;
- the mean queue length in the orbit;
- the mean waiting time;
- the mean number of busy servers.

For instance, [4] shows how to analyse these performances in models such as $M/M/c/c$ and $M/G/1/1$ (notation is for the primary component). Other studies allow more complicated arrivals (batch, MAP, BMAP) but assume a single server and no buffer for obtaining analytic results.

In our study we focus on other performance characteristics which are analysed hardly in existing literature but which do have practical interest, in our opinion. They are

- the probability of a large orbit content;
- the fraction of external arrivals entering the orbit.

2 The probability of a large orbit content

A large orbit content is similar to a large backlog in traditional queueing systems which takes a long time to clear and causes long waiting times. In the retrial system it means also a large number of retrials with many failing. This frustrates callers. Hence, good management of a retrial system avoids large orbit contents. That is, its probability should be small. Apart of the systems mentioned in the Introduction, the probability is not known and therefore should be calculated by an approximation, or estimated by simulation.

The latter method will be considered here. Since we deal with small probabilities the simulations should be done preferably with a variance reduction technique. We propose to apply importance sampling in order to speed up the simulation runs.

Now let us be more specific about the system. For ease of exposition we assume a time-discrete (or time-slotted) retrial queue. The arrival process is an i.i.d. sequence A_n, $n = 1, 2, \ldots$. A_n is the number of external arrivals during period n, also called primary calls. The primary component consists of a buffer of finite size B. Its content at the beginning of period n is denoted by Q_n. The secondary component is the infinite orbit. Its content is denoted by X_n. During period n a random number R_n retries for getting access at the buffer. This flow contains the so-called secondary calls. R_n depends on the orbit content X_n only. Primary and secondary calls merge into one process of arrivals to the buffer: $I_n = A_n + R_n$. During period n a random number D_n of calls leave the buffer. We allow that calls depart in the same period as they

arrive. However, if there are more arrivals than free spaces after departure, the remaining calls enter the orbit and try later. We call these the overflow calls, denoted by F_n. Notice that part of the overflow stream may be secondary calls, they return to the orbit. Summarizing:

Relations

$$\begin{aligned} I_n &= A_n + R_n, \\ Q^f_{n+1} &= Q_n + I_n - D_n, \\ F_n &= \left(Q^f_{n+1} - B\right)^+, \\ Q_{n+1} &= \left(Q^f_{n+1} - F_n\right)^+, \\ X_{n+1} &= X_n - R_n + F_n. \end{aligned}$$

In these relations Q^f_n stands for the free buffer content. We make the following probabilistic

Assumptions

$$\begin{aligned} A_n &\stackrel{d}{=} \text{Poisson}(\lambda), \\ D_n &\stackrel{d}{=} \text{Constant}(c), \\ R_n &\stackrel{d}{=} \text{Binomial}(x, \alpha) \quad \text{whengiven} \quad X_n = x. \end{aligned}$$

The last assumption says that each call in the orbit retries with probability α. For stability we assume that $\lambda < c$. The point of interest is the steady state probability that the orbit content exceeds K:

$$\gamma(K) := P(X \geq K).$$

We estimate γ by regenerative simulations of the process $\{(Q_n, X_n),\, n = 1, 2, \ldots\}$. The regeneration points are the times of return to state $(0, 0)$. We denote C for the cycle length and T_K for the level-K exceedance time:

$$T_K = \sum_{n=1}^{C} 1\{X_n \geq K\}.$$

Hence,

$$\gamma(K) = \frac{E[T_K]}{E[C]}, \qquad \hat{\gamma}_N(K) = \frac{\sum_{i=1}^N T_K^{(i)}}{\sum_{i=1}^N C^{(i)}},$$

where $T_K^{(i)}$ and $C^{(i)}$ are the i-th independent copies of T_K and C in a simulation experiment of N regeneration cycles.

As mentioned above, we consider applying importance sampling where the arrival distribution is exponentially tilted. This is argued as follows. The only possible way the orbit content

increases, is that the buffer is full ($Q_n = B$) for a period of consecutive periodes. From the relations we see that in such a period the orbit length satisfies the recursion

$$X_{n+1} = X_n + I_n - c - R_n = X_n + A_n - c.$$

In other words, the process behaves as a random walk with jumps of size A_n upwards and jumps of size c downwards. Since $E[A_n] = \lambda < c$, reaching high level K is a rare event for the random walk. It is well known that the random walk satisfies the large deviations principle, e.g. [3]. We use the well known queueing relation

$$P(X \geq K) = P\left(\sup_n \left[\sum_{k=1}^{n} A_k - nc\right] \geq K\right),$$

and apply the relation between the tail of the queue length distribution and the large deviations property of the random walk, e.g. [5]. This gives the large deviations asymptotic

$$\lim_{K\to\infty} \frac{1}{K} \log \gamma(K) = -\theta, \tag{1}$$

where $\theta > 0$ solves

$$\log E[\exp(\theta A)] = \theta c.$$

In fact, we use θ as the tilting factor for an exponential change of measure, e.g. [2, 6]. Under this change of measure the arrivals have a Poisson(λe^{θ}) distributions.

Tables 1 summarizes a simulation experiment. The parameters are

$$\lambda = 10, c = 12, B = 5, \alpha = 0.25, N^{\text{standard}} = 100000, N^{\text{IS}} = 1000.$$

- The first two columns show the relative errors (RE) of the estimators. Notice that the standard RE increases exponentially in K, whereas the importance sampling RE remains bounded. There are no exceedance observations in the standard simulations for large K.

- The third column shows the required amounts of work in the importance sampling simulations. The simulations are executed in MATLAB (on a PC). The work is measured in the number of mega floating-point operations (Mflops) as returned by the MATLAB program. The standard simulations of 100,000 cycles require about 9.5 Mflops (for any K). The work requirement in the importance sampling simulations grows slowly due to longer periods with the tilted distribution. Hence, more calculations are required for the likelihood ratio.

- The fourth column gives the ratios

$$\frac{\log E^{\text{IS}}[\hat{\gamma}^2]}{\log \gamma(K)}$$

which have been realized by the importance sampling simulations until a (95%,15%) efficiency is obtained, i.e., until the relative width of the 95%-confidence interval is within 15% to both sides of the estimate. Variance reduction is obtained when the ratio is larger than 1, and the closer it is to 2, the better the estimator is. Asymptotic optimality would mean that the ratio convergences to 2 for large K. Because of the large deviations asymptotic (1), one can show that the importance sampling estimator is indeed asymptotically optimal [2, 6].

K	RE standard	RE ImpSamp	Mflops ImpSamp	ratio ImpSamp
5	4.03%	8.66%	0.76	1.4451
10	10.23%	11.33%	0.94	1.6033
15	24.38%	8.79%	1.16	1.7182
20	55.30%	16.07%	1.36	1.7471
25		11.90%	1.66	1.8211
30		9.66%	1.74	1.8295
35		8.87%	2.24	1.8555
40		8.01%	2.36	1.8059
45		9.34%	2.69	1.8659
50		14.17%	2.85	1.8937

Table 1. Performance of the importance sampling simulations for estimating the probability of large orbit contents.

3 The overflow fraction

The overflow fraction is the long run fraction of primary calls that is not accepted by the buffer immediately. In most applications the buffer is not present, or has a small size, thus overflow is not a rare event. Then, crude simulations are possible to estimate the fraction. However, let us assume that large buffers are possible.

First we take a closer look at entering the buffer. Above we said that the two arrival streams, primary and secondary, merge. Now we let the primary calls have priority, in other words, the buffer accept as many primary calls as possible before accepting secondary calls (in each period). Then we can decompose the overflow stream, $F_n = F_n^{(1)} + F_n^{(2)}$, where $F_n^{(1)}$ are the primary calls not accepted in period n:

$$F_n^{(1)} = (Q_n + A_n - c - B)^+ . \tag{2}$$

The performance of interest is the long run overflow fraction

$$\gamma(B) = \lim_{n\to\infty} \frac{\sum_{k=1}^n F_k^{(1)}}{\sum_{k=1}^n A_k}.$$

Again we set up regenerative simulations for estimating $\gamma(B)$. Let A_C be the total number of arriving primary calls during a regeneration cycle, and and F_B be the total number of overflowing primary calls:

$$A_C = \sum_{n=1}^{C} A_n, \quad F_B = \sum_{n=1}^{C} F_n^{(1)}.$$

Then, the probability and its estimator are

$$\gamma(B) = \frac{E[F_B]}{E[A_C]}, \qquad \hat{\gamma}_N(B) = \frac{\sum_{i=1}^{N} F_B^{(i)}}{\sum_{i=1}^{N} A_C^{(i)}}.$$

Now we interpret the relation (2) as the evolution of a random walk $Q = \{Q_n,\, n = 1, 2, \ldots\}$ with jumps of size A_n upwards, jumps of size c downwards, and with two boundaries, one at 0 and one at B. For the importance sampling simulations we consider the optimal change of measure in case the event would be to reach high levels B (in stead of the event of counting overflows). This goes similarly as in the previous section. Under the change of measure the arrivals have a Poisson(λe^{θ}) distribution, where the tilting factor θ is determined by solving

$$\log E[\exp(\theta A)] = \theta c.$$

It is not known whether a large deviations asymptotic as (1) holds for $\gamma(B)$. Also, it is not known whether the importance sampling estimator is asymptotically optimal. However, the simulation results (Table 2) strongly suggests asymptotic optimality.

Table 2 shows simulation results similarly to Table 1. The parameters are

$$\lambda = 10, c = 12, \alpha = 0.25, N^{\text{standard}} = 100000, N^{\text{IS}} = 1000.$$

K	RE standard	RE ImpSamp	Mflops ImpSamp	ratio ImpSamp
5	1.82%	7.86%	0.59	1.5502
10	5.38%	7.76%	0.65	1.6890
15	16.00%	7.84%	0.77	1.7739
20	35.08%	7.63%	0.94	1.8051
25	99.98%	7.75%	1.04	1.8348
30		8.10%	1.21	1.8631
35		7.47%	1.39	1.8762
40		7.61%	1.49	1.8895
45		7.65%	1.65	1.9024
50		7.91%	1.88	1.9087

Table 2. Performance of the importance sampling simulations for estimating the overflow fraction.

4 Conclusion

In this paper we have studied rare events in a discrete time retrial queue. The probabilities of the rare events are estimated by simulations. We have approximated the statistical behavior of the retrial queue towards the rare event by a random walk which satisfies the large deviations principle. In this way we have an approximate asymptotic of the probability of the rare event, and we could find the tilting factor for executing importance sampling simulations with the optimal change of measure.

Further investigations are needed to find approximations of the rare event probabilities based on the large deviations asymptotic. Also, continuous-time retrial queues may be studied by the approach exposed in the paper.

References

[1] Artalejo, J.R (1999). A classified bibliography of research on retrial queues: Progress 1990 – 1999. *Top* **7**, 187-211.

[2] Asmussen, S. and R. Rubinstein (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*, 429-461, CRC Press, Boca Raton, USA.

[3] Dembo, A. and O. Zeitouni (1996). *Large deviations techniques and applications*, Second edition, Springer Verlag.

[4] Falin, G.I. and J.G.C. Templeton (1997). *Retrial Queues.* Chapman & Hall, London.

[5] Glynn, P.W. and W. Whitt (1994). Logarithmic asymptotics for steady-state tail prbabilities in a single-server queue. *Journal of Applied Probability* **31**, 131-159.

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