Proceedings of the 30th Conference
on Decision and Control
Brighton, England • December 1991

Importance Sampling, Jump Distributions and Event-Time Distributions*

Michael R. Frater† Brian D. O. Anderson‡

* Work supported by Australian Telecommunications and Electronics Research Board (ATERB)

† Department of Electrical Engineering, University College, Australian Defence Force Academy, Canberra ACT 2600, Australia.
‡ Department of Systems Engineering, Australian National University, G.P.O. Box 4, Canberra A.C.T., 2601, Australia.

ABSTRACT: Two different methods have been proposed for performing asymptotically optimal simulation to obtain the statistics of buffer overflows in queueing networks, with both using large deviations and importance sampling. In the first, based on heuristic arguments, the distributions of interarrival and virtual service times are analysed to find the simulation system. In the second, it is the distribution of jumps occurring in a Markov chain that is examined. In this paper, we show that the two approaches will produce identical fast simulation systems for an arbitrary GI/GI/1 queue.

1 Introduction

There has been a large amount of interest in obtaining fast simulation methods for obtaining the statistics of rare events in telecommunications systems. In the particular area of queueing systems, two different approaches have been proposed:

1. In [1] a method is proposed, based on a heuristic due to Borovkov, where the distributions of the interarrival and service times are used in conjunction with large deviations theory to obtain a simulation system that is asymptotically optimal, in the sense of minimizing variance.

2. In [2], an alternative approach is used, based on analysing the distribution of jumps on an embedded Markov chain using large deviations.

The first approach has the advantage of using the arrival and service distributions, which are often known, but is based on heuristic arguments. The second approach overcomes this difficulty, but has the disadvantage that it may not be easy to construct an embedded Markov chain for a particular example. In this paper, we show that for an arbitrary GI/GI/1 queue, these two different approaches will always produce the same simulation system.

2 Cramér's Theorem

Definition 1 Let $\xi_1 \ldots \xi_n$ be i.i.d random variables in $\mathbb{R}$. Let $F(\cdot)$ be the distribution function of the $\{\xi_k\}$ and m its mean. The Cramér (or Legendre) transform $h(\cdot)$ is defined as [1, 3]:

$$h(y) = \sup_{s \in \mathbb{R}} [sy - \log M(s)] \tag{1}$$

where $M(s)$ is the moment generating function of $F(\cdot)$, and is assumed to be finite in a neighbourhood of 0.

Theorem 1 Cramér's Theorem Let $\xi_1 \ldots \xi_n$ be i.i.d random variables with $\xi_i \in \mathbb{R}$. Let $F(\cdot)$ be the distribution of the ξ_i, with mean m. Then we have the following:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{P}\left\{\frac{\xi_1 + \ldots + \xi_n}{n} \simeq a\right\} = -h_F(a) \tag{2}$$

where $h_F(\cdot)$ is the Cramér transform of the distribution $F(\cdot)$.

3 Main Result

Let X^A be a random variable whose value denotes the interarrival time at a GI/GI/1 queue, with distribution A, and let $h_A(\cdot)$ be the Cramér transform of the distribution A. Similarly, let X^B be a random variable whose value denotes the virtual service time at a GI/GI/1 queue, with distribution B, and let $h_B(\cdot)$ be the Cramér transform of the distribution B. We will call the number of customers in the queue at any one time the state. Let $h_j(\cdot)$ denote the Cramér transform of the distribution of jumps ξ_i in the state occurring in unit time.

Let

$$P_A = \mathrm{P}\left\{\frac{X_1^A + \ldots + X_{\lambda' T}^A}{\lambda' T} \simeq \frac{1}{\lambda'}\right\} \tag{3}$$

$$P_B = \mathrm{P}\left\{\frac{X_1^B + \ldots + X_{\mu' T}^B}{\mu' T} \simeq \frac{1}{\mu'}\right\} \tag{4}$$

The following are the two results that we will show later to be equivalent.

Theorem 2 Parekh and Walrand [1]. Let λ^* be the average arrival rate and μ^* the average virtual service rate for the optimal simulation system for the GI/GI/1 queue defined above. Then λ^* and μ^* are the unique solutions of

$$h_A\left(\frac{1}{\lambda^*}\right) + h_B\left(\frac{1}{\mu^*}\right) = \left(\frac{1}{\lambda^*} - \frac{1}{\mu^*}\right) h_A'\left(\frac{1}{\lambda^*}\right) \tag{5}$$

$$= \left(\frac{1}{\mu^*} - \frac{1}{\lambda^*}\right) h_B'\left(\frac{1}{\mu^*}\right) \tag{6}$$

with $\lambda^* > \mu^*$.

Theorem 3 Frater, Walrand and Anderson [2]. Let ν^* be the average rate of increase of the state of the optimal simulation system for the GI/GI/1 queue defined above. Then ν^* is the unique positive solution of:

$$h_j(\nu^*) = \nu^* \left.\frac{d}{d\nu} h_j(\nu)\right|_{\nu=\nu^*} \tag{7}$$

with $h_j(\cdot)$ as defined above.

Lemma 4 The Cramér transform of the distribution of jumps and the Cramér transforms of the distributions of interarrival and virtual service times of the GI/GI/1 queue defined above are related by:

$$h_j(\nu') = \min_{\lambda' - \mu' = \nu'} \left[\lambda' h_A\left(\frac{1}{\lambda'}\right) + \mu' h_B\left(\frac{1}{\mu'}\right)\right] \tag{8}$$

Proof:

Let T be a long time interval. If there are $\lambda' T$ arrivals and $\mu' T$ arrivals in this time, then the number of customers in the queue changes by $\nu' T = \lambda' T - \mu' T$. Hence, it is clear that:

$$\mathrm{P}\left\{\sum_{i=1}^{T} \xi_i \simeq \nu' T\right\} = \sum_{\substack{\lambda' - \mu' \\ = \nu'}} \mathrm{P}\left\{\sum_{i=1}^{\lambda' T} X_i^A \simeq T\right\} \mathrm{P}\left\{\sum_{i=1}^{\mu' T} X_i^B \simeq T\right\} \tag{9}$$

i.e.

$$\frac{1}{T} \log \mathrm{P}\left\{\frac{\xi_1 + \ldots + \xi_T}{T} \simeq \nu'\right\} = \frac{1}{T} \log \sum_{\lambda' - \mu' = \nu'} P_A P_B \tag{10}$$

Now, from Cramér's theorem (Theorem 1), we know that

$$\lim_{T\to\infty} \frac{1}{T} \log P_A P_B = -\lambda' h_A\left(\frac{1}{\lambda'}\right) - \mu' h_B\left(\frac{1}{\mu'}\right) \tag{11}$$

In other words, as the time T becomes large, the asymptotic behaviour of the probability of there being λ' arrivals and μ' services in time T is:

$$P_A P_B \sim \exp\left\{-T\left[\lambda' h_A\left(\frac{1}{\lambda'}\right) + \mu' h_B\left(\frac{1}{\mu'}\right)\right]\right\} \tag{12}$$

Therefore, it is clear that in the limit as $T \to \infty$, the term associated with the smallest value of

$$\lambda' h_A\left(\frac{1}{\lambda'}\right) + \mu' h_B\left(\frac{1}{\mu'}\right) \tag{13}$$

satisfying $\lambda' - \mu' = \nu'$ in the right-hand side of (10) dominates, and hence:

$$\begin{aligned} &\lim_{T\to\infty} \frac{1}{T} \log \mathrm{P}\left\{\frac{\xi_1 + \ldots + \xi_T}{T} \simeq \nu'\right\} \qquad (14)\\ &= \lim_{T\to\infty} \frac{1}{T} \log \sum_{\lambda'-\mu'=\nu'} P_A P_B \\ &= \lim_{T\to\infty} \frac{1}{T} \max_{\lambda'-\mu'=\nu'} \log P_A P_B \end{aligned}$$

Therefore,

$$h_j(\nu') = \min_{\lambda'-\mu'=\nu'} \left[\lambda' h_A\left(\frac{1}{\lambda'}\right) + \mu' h_B\left(\frac{1}{\mu'}\right)\right] \tag{15}$$

as required.

□

We note also that the values of λ', μ' achieving the constrained minimum (with given ν') in (15) are unique. This can be seen from the fact that there is a unique turning point with respect to λ' and μ' in the unconstrained minimization, and that this turning point is in fact a minimum [2], (from the properties of the Cramér transform.) Because the constraint is linear, it is also true that there is a unique solution to the constrained minimization [4].

Theorem 5 The two optimal simulation systems for a GI/GI/1 queue defined by Theorems 2 and 3 are equivalent. Specifically, the average rate of increase of the optimal simulation system $\nu^\star$ satisfies:

$$h_j(\nu^\star) = \nu^\star \frac{\mathrm{d}}{\mathrm{d}\nu'} h_j(\nu')\Big|_{\nu'=\nu^\star} \tag{16}$$

if and only if the arrival rate $\lambda^\star$ and the service rate $\mu^\star$ of the simulation system satisfy

$$\begin{aligned} h_A\left(\frac{1}{\lambda^\star}\right) + h_B\left(\frac{1}{\mu^\star}\right) &= \left(\frac{1}{\lambda^\star} - \frac{1}{\mu^\star}\right) h'_A\left(\frac{1}{\lambda^\star}\right) \qquad (17)\\ &= \left(\frac{1}{\mu^\star} - \frac{1}{\lambda^\star}\right) h'_B\left(\frac{1}{\mu^\star}\right) \end{aligned}$$

with $\nu^\star = \lambda^\star - \mu^\star$.

Proof:

Firstly, let $\nu^\star = \lambda^\star - \mu^\star$. We can perform the minimization of Lemma 4 using the method of Lagrange multipliers. We define the lagrangian:

$$\mathcal{L} = \lambda' h_A\left(\frac{1}{\lambda'}\right) + \mu' h_B\left(\frac{1}{\mu'}\right) + k\left(\lambda' - \mu' - \nu^\star\right) \tag{18}$$

where k is the lagrange multiplier. Let $\overline{\lambda}$, $\overline{\mu}$ be the solution. Then we have:

$$\begin{aligned} h_A\left(\frac{1}{\overline{\lambda}}\right) - \frac{1}{\overline{\lambda}} h'_A\left(\frac{1}{\overline{\lambda}}\right) + k &= 0 \qquad (19a)\\ h_B\left(\frac{1}{\overline{\mu}}\right) - \frac{1}{\overline{\mu}} h'_B\left(\frac{1}{\overline{\mu}}\right) - k &= 0 \qquad (19b)\\ \overline{\lambda} - \overline{\mu} &= \nu^\star \qquad (19c) \end{aligned}$$

It is straightforward to check that choosing $\overline{\lambda} = \lambda^\star$ and $\overline{\mu} = \mu^\star$ satisfies the above equations. This solution is unique, since, as we saw above, the solution to the minimization problem of Lemma 4 is unique. Then we have

$$h_j(\nu^\star) = \lambda^\star h_A\left(\frac{1}{\lambda^\star}\right) + \mu^\star h_B\left(\frac{1}{\mu^\star}\right) \tag{20}$$

Now, from the sensitivity theorem ([4], p 230), we have also that

$$\begin{aligned} \frac{\mathrm{d}h_j(\nu^\star)}{\mathrm{d}\nu^\star} &= -k \qquad (21)\\ &= h_A\left(\frac{1}{\lambda^\star}\right) - \frac{1}{\lambda^\star} h'_A\left(\frac{1}{\lambda^\star}\right) \quad \text{by (19a)} \qquad (22)\\ &= \frac{1}{\lambda^\star - \mu^\star}\left[\lambda^\star h_A\left(\frac{1}{\lambda^\star}\right) + \mu^\star h_B\left(\frac{1}{\mu^\star}\right)\right] \qquad (23)\\ &= \frac{1}{\nu^\star} h_j(\nu^\star) \qquad (24) \end{aligned}$$

In order to prove the converse, suppose that (16) holds for some $\nu^\star$. We solve the optimization problem (8) with $\nu' = \nu^\star$ in order to find the associated arrival rate $\overline{\lambda}$ and service rate $\overline{\mu}$. Then (21) holds for some k, and $(\overline{\lambda}, \overline{\mu})$ are the values of (λ', μ')achieving the minimum if and only if $(\overline{\lambda}, \overline{\mu})$ satisfies (19a) through (19c). Thus, using (16), (19a) and (19b), we can write

$$\begin{aligned} &h_A\left(\frac{1}{\overline{\lambda}}\right) - \frac{1}{\overline{\lambda}} h'_A\left(\frac{1}{\overline{\lambda}}\right) - \frac{1}{\overline{\lambda} - \overline{\mu}}\left[\overline{\lambda} h_A\left(\frac{1}{\overline{\lambda}}\right) + \overline{\mu} h_B\left(\frac{1}{\overline{\mu}}\right)\right]\\ &= 0 \qquad (25)\\ &h_B\left(\frac{1}{\overline{\mu}}\right) - \frac{1}{\overline{\mu}} h'_B\left(\frac{1}{\overline{\mu}}\right) + \frac{1}{\overline{\lambda} - \overline{\mu}}\left[\overline{\lambda} h_A\left(\frac{1}{\overline{\lambda}}\right) + \overline{\mu} h_B\left(\frac{1}{\overline{\mu}}\right)\right]\\ &= 0 \qquad (26) \end{aligned}$$

These two equations yield (17), with $\lambda^\star$ replaced by $\overline{\lambda}$ and $\mu^\star$ replaced by $\overline{\mu}$. Now, the values of $(\overline{\lambda}, \overline{\mu})$ satisfying (25) and (26) are turning points (λ', μ') associated with the minimization problem in (8) for the case $\nu' = \nu^\star$. But for given ν' there is a unique extremum, which achieves the minimum. Hence, because $(\overline{\lambda}, \overline{\mu})$ is a solution of both (25) and (26), and of the minimization problem of (8), there is a unique solution $(\overline{\lambda}, \overline{\mu})$ to (25) and (26), with $(\overline{\lambda}, \overline{\mu}) = (\lambda^\star, \mu^\star)$.

References

[1] S. Parekh and J. Walrand, "A quick simulation of excessive backlogs in networks of queues," *IEEE Trans. Automatic Control*, vol. 34, pp. 54 – 66, January 1989.

[2] M. R. Frater, J. Walrand, and B. D. O. Anderson, "Optimally efficient simulation of buffer overflows in queues with deterministic service times," *Australian Telecommunication Research*, vol. 24, no. 1, pp. 1 – 8, 1990.

[3] M. Cottrell, J. C. Fort, and G. Malgouyres, "Large deviations and rare events in the study of stochastic algorithms," *IEEE Trans. Automatic Control*, vol. AC-28, pp. 907 – 918, September 1983.

[4] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*. Reading MA: Addison-Wesley, 1973.

[5] S. R. S. Varadhan, *Large Deviations and Applications*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1984.