Proceedings of the 1993 Winter Simulation Conference
G.W. Evans, M. Mollaghasemi, E.C. Russell, W.E. Biles (eds.)

SIMULATION OF RARE QUEUEING EVENTS BY SWITCHING ARRIVAL AND SERVICE RATES

Russell C.H. Cheng
Louise Traylor

School of Mathematics
Senghennydd Road
University of Wales College of Cardiff
Cardiff CF2 4YH, U.K.

Janos Sztrik

Institutum Mathematicum
Universitatis Debreceniensis
H-4010 Debrecen Pf. 12
HUNGARY

ABSTRACT

Importance sampling is known to be a powerful method for significantly increasing the efficiency of estimates of the probability of rare events obtained from simulation experiments. However the conditions under which it will be effective require careful checking if the method is to be reliably employed. We show that it is easy to get spurious answers which are apparently accurate when they are in fact quite wrong. For simple queues, there is a simple criterion which guarantees the effective implementation of the method: simply switch the arrival and service rates. This result can be shown using a theorem due to Chernoff, however here we show why the method works by examining sample paths directly, and in particular derive the variance reduction so obtained. In practice the variance reduction can be several orders of magnitude and we give numerical examples demonstrating this.

Keywords: Importance Sampling, Variance Reduction.

1 INTRODUCTION

We consider the estimation of the probability of occurrence of a certain type of rare event. A typical example is the gambler's ruin problem where starting with one unit of money, say, we wish to find the probability that the gambler can build up his/her winnings to a certain level A without going bankrupt beforehand. In the terminology of random walks this is the probability that starting at level 1, the level A will be reached before level 0. A continuous version of essentially the same problem occurs in the single-server queue if we wish to estimate the probability that during a server busy period the level in the queue will reach A, before the busy period ends. Other versions of the problem occur in sequential statistical tests.

Importance sampling is known to be a powerful technique for estimating such probabilities. It was introduced by Siegmund (1976) for handling sequential statistical tests, see also Ripley (1987) and Ross (1990), for some simple examples. An engineering viewpoint for queueing problems is given by Cottrell et al (1983). Fishman (1993) gives full discussion including the above busy period problem. An interesting approach is given by Walrand (1988 a,b) who uses Chernoff's Theorem (Chernoff, 1952) to estimate probabilities of certain sample paths occurring. Below we show how a more direct argument, that does not require Chernoff's Theorem, can give a sharper insight into how importance sampling can be effectively carried out. Our approach enables the variance reduction obtained to be calculated. We also show how to avoid a certain pitfall that can occur when the probability estimate and estimate of its variability are both seriously biased, so that one may be led into thinking that accurate estimation is taking place when exactly the opposite is the case.

In the next section we describe the basic importance sampling technique and give an elementary prototype example. In Sections 3 and 4 we analyse the gambler's ruin problem and the busy period problem, and show that the optimal method is to essentially switch the arrival and service rates, thereby simulating an unstable queue. Our results are illustrated throughout the paper by numerical examples.

2 IMPORTANCE SAMPLING OF RARE EVENTS

2.1 The General Method

The basic method is very simple. We consider terminating processes only. We make a simulation and record if the rare event of interest, call it E say, has occurred or not. Note that the run will follow the process to termination only if E does not occur. The run can stop once E occurs. Simulation runs are thus Bernoulli trials. Let ω denote a typical realization of the process and define

$$\begin{aligned} T(\omega) &= 1 \text{ if E occurs in } \omega \\ &= 0 \text{ otherwise} \end{aligned} \qquad (2.1)$$

Let $dF(\omega)$ denote the probability of occurrence of ω. Clearly the probability of occurrence of E, Pr(E) is given by

$$\Pr(E) = E[T(\omega)] = \int T(\omega)dF(\omega) = \alpha, \text{ say,}$$

This can be estimated by making N runs and taking the sample average, $\overline{T}$, of the observed T's as the estimator.

In importance sampling we simulate the process where the sample space Ω remains the same but where the probability of occurrence of ω is $dF^*(\omega)$, this being different from $dF(\omega)$. We call this the *modified process*. Then, provided $dF^*(\omega) > 0$ whenever $dF(\omega) > 0$, the likelihood ratio

$$L(\omega) = \frac{dF(\omega)}{dF^*(\omega)} \qquad (2.2)$$

remains bounded and

$$E^*(L(\omega)T(\omega)) = \int \frac{dF(\omega)}{dF^*(\omega)} T(\omega) dF^*(\omega) = E(T(\omega)).$$

$$(2.3)$$

We can thus estimate α by simulating the modified process N times and recording:

$$Y_i = L(\omega_i)T(\omega_i), \quad i = 1,\ldots,N. \qquad (2.4)$$

$\overline{Y}$, the sample average of the Y's estimates α. The key trick is to be able to select $dF^*(\omega)$ so that

$$L(\omega) < 1 \text{ whenever } T(\omega) = 1. \qquad (2.5)$$

Then

$$\begin{aligned} E^*[L^2(\omega)T^2(\omega)] &< E^*[L(\omega)T^2(\omega)] \\ &= E[T^2(\omega)], \end{aligned} \qquad (2.6)$$

where E* denotes expectation with respect to dF*; combined with (2.3) this shows that

$$\operatorname{var}(\overline{Y}) < \operatorname{var}(\overline{T}) \qquad (2.7)$$

so that variance reduction is achieved.

2.2 A Prototype Example

We illustrate the method with an elementary example. Let S be a binomial random variable, Bin(n,p), and suppose we wish to find

$$\Pr(S \geq a) = \alpha \qquad (2.8)$$

where a is large. It will be convenient to write $a = \theta n$, with θ close to unity. In this example α can be calculated explicitly of course, but for illustration we consider its estimation by sampling experiment. We can generate S as the sum of n Bernoulli (p) random variables:

$$S = X_1 + X_2 + \ldots + X_n. \qquad (2.9)$$

and from (2.1) define

$$\begin{aligned} T(\omega) &= 1 \text{ if } S \geq a \\ &= 0 \text{ otherwise} \end{aligned}$$

The average $\overline{T}$ of N such values estimates α. We have that

$$\operatorname{var}(\overline{T}) = \alpha(1-\alpha)/N. \qquad (2.10)$$

If importance sampling is used, we generate the X_i's as Bernoulli (p*), rather than Bernoulli (p) variates, and estimate α by Y, with Y as defined in (2.4). Now in this case (2.2) can be written as

$$L(s) = \frac{p^s(1-p)^{n-s}}{p^{*s}(1-p^*)^{n-s}} \qquad (2.11)$$

where s is the observed value of S; and the statistic (2.4) becomes

$$Y = L \text{ if } S \geq a, \quad Y = 0 \text{ otherwise.} \tag{2.12}$$

For variance reduction to occur, (2.5) has to be satisfied. Now, if p* > p , then L(s) as defined in (2.11) decreases as s increases. Thus

$$L(s) \leq \left(\frac{p}{p^*}\right)^a \left(\frac{1-p}{1-p^*}\right)^{n-a} \equiv R(p^*), \text{ for } s \geq a. \tag{2.13}$$

The value of p* which minimises the right hand side is

$$p^* = a / n = \theta. \tag{2.14}$$

Variance reduction is thus guaranteed if this minimized right hand side is less than unity. Under the right circumstances the variance reduction is big. For instance suppose p is small and $\theta = 1 - \delta$ is close to unity so that δ is small. Then $R(\theta) << 1$ and

$$\begin{aligned} E^*\left(Y^2\right) &= \sum_{s \geq a} [L(s)]^2 \binom{n}{s} \theta^s (1-\theta)^{n-s} \\ &\leq R(\theta) \sum_{s \geq a} \binom{n}{s} p^s (1-p)^{n-s} \\ &= R(\theta)\alpha. \end{aligned} \tag{2.15}$$

We can apply Chernoff's theorem to estimate the order of magnitude of α. For our example, the theorem states that

$$\frac{1}{n} \log \Pr(S \geq \theta n) \to -h(\theta)$$

where

$$h(\theta) = \sup_{t \geq 0}[at - \log M(t)];$$

and

$$M(t) = (1-p) + pe^t.$$

A little calculation shows that this is equivalent to

$$\alpha = O[R(\theta)] \quad \text{as } n \to \infty. \tag{2.17}$$

Combined with (2.15) this shows that

$$\text{Var}(Y) = O\left[R^2(\theta)\right]. \tag{2.17}$$

Table 1 summarizes the results of 10,000 experiments in each of which p = 0.7, n = 100. Four values of θ were tried, θ = 0.8, 0.9, .95, and 0.99. It will be seen from the tabulated values of $R(\theta)$ that the variance reduction achieved is much in accordance with (2.17).

3. RANDOM WALK EXAMPLE

Our next example bridges the gap between simple cases where alternative methods, including theoretical analysis, would normally be preferred to importance sampling and genuine cases for which importance sampling is the method of choice but where it is not so obvious how best to implement the method. The random walk example is simple enough for theoretical results to be available for comparison purposes, yet contains the essential difficulties that need to be overcome in more complicated queueing examples if importance sampling is to work properly.

Consider a discrete random walk observed at times t = 0,1,2,... . The only possible positions are X = 0,1,2,3, ... If, at time t, the walk is in position X = j, then

$$p, \ q, \quad r = 1 - p - q \tag{3.1}$$

are respectively the transition probabilities that the walk move to j + 1, j - 1, or remains fixed at j. We shall only consider the case

$$p < q. \tag{3.2}$$

We wish to find the probability

α = pr[Walk reaches level A without first reaching 0, given its position is 1 at time 0].

Thus the straight sampling method comprises generating paths starting at X = 1 at time t = 0 and noting if the level A is reached first or if the level 0 is reached first. We use the notation of Section 2.1 with ω denoting a typical path, and E denoting the event that level A is reached first.

If importance sampling is used we replace (3.1) by the modified transition probabilities

$$p^*, q^*, r^* = 1 - p^* - q^* \tag{3.3}$$

and compensate using (2.4) instead of (2.1). For simplicity we consider only the case where r = r* so that

$$p^* + q^* = p + q. \qquad (3.4)$$

Again the key is to ensure that (2.5) is satisfied.

Consider any path ω starting from 1 which reaches level A before it reaches 0. Its associated probability is

$$dF(\omega) = p^{A-1+b}\, q^b\, r^c \qquad (3.5)$$

for some b, c ≥ 0. Thus (2.2) becomes

$$L(\omega) = \left(\frac{p}{p^*}\right)^{A-1+b} \left(\frac{q}{q^*}\right)^b \qquad (3.6)$$

(as r = r*). If we write $p' = p/(p+q)$, $q' = q/(p+q)$, $p^{*\prime} = p^*/(p^*+q^*)$, $q^{*\prime} = q^*/(p^*+q^*)$ then

$$L(\omega) = \left(\frac{p'}{p^{*\prime}}\right)^{A-1} \left[\frac{p'(1-p')}{p^{*\prime}(1-p^{*\prime})}\right]^b, \text{ some } b \geq 0 \qquad (3.7)$$

To satisfy (2.5) we must have

$$p' \leq p^{*\prime} \leq (1-p'). \qquad (3.8)$$

A possible choice is

$$p^{*\prime} = q^{*\prime} = ½ \text{ (i.e. } p^* = q^* = (p+q)/2), \qquad (3.9)$$

but the most interesting choice is

$$p^{*\prime} = q',\ q^{*\prime} = p \text{ (i.e. } p^* = q,\ q^* = p) \qquad (3.10)$$

when

$$L(\omega) = \left(\frac{p}{q}\right)^{A-1}. \qquad (3.11)$$

Then

$$E^*[L^2(\omega)T^2(\omega)] \leq \left(\frac{p}{q}\right)^{A-1} E\left[T^2(\omega)\right].$$

Now it is known (see Cox and Miller, 1978, for example) that

$$\alpha = E[T(\omega)] = E\left[T^2(\omega)\right] = p^{A-1}(q-p)/(q^A - p^A).$$

Thus for A large we find $\mathrm{Var}(T) \approx (p/q)^A(qp^{-1}-1)$ whilst $\mathrm{Var}(Y) \approx (p/q)^A \mathrm{Var}(T)$.

Table 2 gives the simulation results of 10,000 runs with p = 0.3, q = 0.5, A = 15, which illustrates the variance reduction achieved. Note that the run time for the modified process is longer but this increase is small compared with the variance reduction achieved.

A warning should be noted. If in our example we let p* > (1-p), this will decrease L(ω) for paths ω in which dF(ω), as given by (3.5), has b small, and in consequence reduce the contribution of such paths to Var(Y). This is at the expense of increasing L(ω) for paths where b is large, thus increasing their contribution to Var(Y). The net effect is to reduce Var(Y) if p* is suitably chosen. The effectiveness of a chosen p* will thus require an accurate estimate of Var(Y). An unbiased estimate of Var(Y) will require both types of path to be sampled in proportion to their probability of occurrence. Now increasing p* has the additional effect of reducing the number of paths where b is large, and if p* is chosen inordinately large, then there will be very few such paths. If the sample size is small we may not sample such paths at all and their very large contribution to Var(Y) will not be properly accounted for. This makes Var(Y) negatively biased and will mislead us into thinking that the estimate of E(Y) is much more accurate than it is. This effect is illustrated in Table 2.

4 M/M/1 QUEUE

Our final example is a simple but non-trivial application of the results of Section 3 to the M/M/1 queue, with arrival rate λ and service rate μ. Consider estimation of

$$\alpha = \Pr\{\text{queue level reaches A during a busy period of the server}\}. \qquad (4.1)$$

We use the methodology of Section 2.1 and simulate a modified M/M/1 process with λ* and μ* set differently from λ and μ. We adjust for this by using Y of (2.4) to estimate α. Walrand (1988) has considered this problem using Chernoff's theorem but we use a simpler, more direct approach.

We utilize the analysis given for the random walk. A path ω where the queue level reaches A before it reaches 0 must reach A at some time τ. If there are m departures in this time, then this must be matched by precisely A+m-1 arrivals in the same time period. If t_i are the interarrival times and s_j are the service times, then we have

$$\sum_{i=1}^{A+m-1} t_i = \tau$$

and, because the server is busy over the entire period,

$$\sum_{j=1}^{m} s_j < \tau < \sum_{j=1}^{m+1} s_j .$$

Thus $\sum_1^m s_j$ and τ are nearly equal and we have

$$dF(\omega) = \lambda^{A+m-1} e^{-\lambda\tau} \mu^m e^{-\mu\tau} . \qquad (4.2)$$

It follows that in the case of importance sampling:

$$L(\omega) = \left(\frac{\lambda}{\lambda^*}\right)^{A-1} \left(\frac{\lambda\mu}{\lambda^*\mu^*}\right)^m e^{(\lambda^*-\lambda+\mu^*-\mu)\tau} . \qquad (4.3)$$

If we let $\lambda^*+\mu^* = \lambda+\mu$, the analysis reduces to the random walk case. Variance reduction is obtained if

$$\lambda^*=\mu^* = \tfrac{1}{2}(\lambda+\mu) \qquad (4.4)$$

or more interestingly if

$$\lambda^* = \mu,\ \mu^* = \lambda, \qquad (4.5)$$

when

$$L(\omega) = \left(\frac{\lambda}{\mu}\right)^{A-1} \qquad (4.6)$$

In this latter case

$$\mathrm{Var}(Y) \le \left(\frac{\lambda}{\mu}\right)^{A-1} \alpha - \alpha^2 .$$

Table 3 gives results analogous to those in Table 2 for the random walk case and again they corroborate the analysis given above, showing that the method can be very effective.

5 CONCLUSIONS

Importance sampling allows the probability of rare events to be accurately estimated. In certain queueing situations an attractive method is to simply swap round the arrival and service rates. Variance reduction is not only guaranteed, but the method is robust, and variance reduction of orders of magnitude can be achieved.

The approach is capable of some generalization both in terms of extensions of the methodology itself and to applications involving more complicated queueing situations. A possible generalization is to consider situations where variance reduction is achieved by considering all pairs of states i, j and simply swapping over the transition probability of going from state i to j with that of going from j to i. For example, suppose the states of the system can be ordered and denoted as 0,1,2,. . . and that transitions only occur from i to i+1 with probability λ_i and from i+1 to i with probability μ_i. The M/M/1 queue is the special instance where $\lambda_i = \lambda$, $\mu_i = \mu$, and the well-known repairman problem can be formulated with $\lambda_i = i\lambda$, $\mu_i = \mu$. Then the method in effect makes the swap $\lambda_i = \mu$, $\mu_i = \lambda$ for the M/M/1 model. A similar exchange is possible in the repairman problem by setting $\lambda_i = \mu$, $\mu_i = i\lambda$. We hope to address these extensions elsewhere.

Table 1: Estimation of $\alpha = \Pr(S \ge \theta n)$ from 10,000 Runs, where $S \sim \mathrm{Bin}(n,p)$ with n=100, p=0.7

p	θ	p*	True α	$\hat{\alpha}$	SD($\hat{\alpha}$)	SD ratio	R(θ)
0.7	0.8	0.7	1.646×10^{-2}	2.36×10^{-2}	1.52×10^{-1}	1.00	-
0.7	0.8	0.8	1.646×10^{-2}	1.57×10^{-2}	2.36×10^{-2}	6.44	7.63×10^{-2}
0.7	0.9	0.9	1.556×10^{-6}	1.46×10^{-6}	2.88×10^{-6}	5.28×10^{4}	8.88×10^{-6}
0.7	0.95	0.95	3.993×10^{-10}	3.70×10^{-10}	7.15×10^{-10}	2.13×10^{8}	1.96×10^{-9}
0.7	0.99	0.99	1.419×10^{-14}	1.53×10^{-14}	1.81×10^{-14}	8.40×10^{12}	3.75×10^{-14}

Note: the first row contains results for the standard method, when $\theta = 0.8$. For all other cases of θ, the standard method gives $\hat{\alpha} = 0$ with the S.D. not defined.

Table 2: Estimation of α = Pr (Random Walk reaches A before 0; Starting from 1) based on 10,000 Runs, with p = 0.3, q = 0.5, A = 15. True α = 0.000314.

	p*	q*	$\hat{\alpha}$	SD($\hat{\alpha}$)	SD ratio	Time (in secs)
†	.3	.5	.000500	2.24×10^{-4}	1.00	3215
	.4	.4	.000273	0.131×10^{-4}	17.10	3800
††	.5	.3	.000319	0.039×10^{-4}	57.44	4366
†††	.7	.1	.000086	0.032×10^{-4}	70.00	4450

† standard method
†† recommended method using p-q swap
††† pitfall method, p* too large.

Table 3: Estimation of α = Pr (queue level reaches A in a busy period) from 10,000 Busy Periods, in an M/M/1 Queue with $\lambda = 0.3$, $\mu = 0.5$, A = 10

	λ*	μ*	$\hat{\alpha}$	SD($\hat{\alpha}$)	SD ratio	Time (in secs)
†	0.3	0.5	.00350	5.91×10^{-4}	1.00	2883
	0.4	0.4	.00299	1.73×10^{-4}	3.42	3633
††	0.5	0.3	.00243	0.33×10^{-4}	17.91	4200
†††	0.1	0.1	.00062	0.47×10^{-4}	12.57	4100

† standard method
†† recommended method using λ-μ swap
††† pitfall method, λ* too large.

REFERENCES

Chernoff, H. 1952. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics* 23:493-507.

Cottrell, M., Fort, J., and Malgouyres, G. 1983. Large deviations and rare events in the study of stochastic algorithms. *IEEE Transactions on Automatic Control* AC-28:907-920.

Cox, D.R. and Miller. 1978. *The Theory of Stochastic Processes.* London: Methuen.

Fishman, G.S. 1993. *Principles of the Monte Carlo Method.* To appear.

Ripley, B.D. 1987. *Stochastic Simulation.* New York: Wiley.

Ross, S.M. 1990. *A Course in Simulation.* New York: Macmillan.

Siegmund, D. 1976. Importance sampling in the Monte Carlo study of sequential tests. *Annals of Statistics* 4: 673-684.

Walrand, J. 1988a. Quick simulation of queueing networks: an introduction. *Proceedings of the 2nd International Workshop on Applied Maths, Computer Performance and Reliability*, 275-286.

Walrand, J. 1988b. *An Introduction to Queueing Networks.* Prentice Hall.

AUTHOR BIOGRAPHIES

RUSSELL C.H. CHENG obtained a B.A. from Cambridge University, England, in 1968, and the diploma in Mathematical Statistics in 1969. He obtained his Ph.D. in 1972 from Bath University working on computer simulation models of industrial chemical plants in association with ICI. He joined the Mathematics Department of the University of Wales Institute of Science and Technology in 1972 and was appointed Reader in 1988. His main fields of interest include: computer generation of random variates, variance reduction methods, parametric estimation methods and more recently: ship simulation.

LOUISE TRAYLOR received her B.Sc.(Hons) degree in Computing and Statistics in 1990 from the University of Wales. She is currently enrolled as a doctoral student in the School of Mathematics, University of Wales Cardiff, where she holds a University of Wales Studentship to study statistical estimation techniques.