Proceedings of the 1995 Winter Simulation Conference
ed. C. Alexopoulos, K. Kang, W. R. Lilegdon, and D. Goldsman

RARE EVENT SIMULATION IN STOCHASTIC MODELS

Perwez Shahabuddin

Department of Industrial Engineering & Operations Research
Columbia University
New York, NY 10027, U.S.A.

ABSTRACT

We review fast simulation techniques used for estimating probabilities of rare events and related quantities in different types of stochastic models.

1 INTRODUCTION

In this paper we review some of the fast simulation techniques used for estimating probabilities of rare events and related quantities in different types of stochastic models. This paper is by no means a comprehensive survey of these rare event simulation techniques, nor does it present a complete reference list of all the contributions in this area. However, an attempt has been made to give some of the basic concepts and algorithms used for different types of stochastic models. For those types of models for which this has not been possible due to space constraints, an attempt has been made to point to the latest references, so that an interested reader may follow up. The reader is referred to Heidelberger (1993), Nicola, Shahabuddin and Heidelberger (1993), and Nakayama (1994) for more comprehensive surveys of fast simulation techniques applied to reliability models, and Heidelberger (1993) and Asmussen and Rubinstein (1994) for techniques applied to queueing models.

Estimations of the small probabilities of rare events are required in the design and operation of many engineering systems. Consider the case of a telecommunication network. It is customary to model such systems as network of queues, with each queue having a buffer of finite capacity. Information packets that arrive to a queue when its buffer is full are lost. The rare event of interest may be the event of a packet being lost. Current regulations stipulate that the probability of packet loss should not exceed 10^{-9}. Or in a reliability model of a space craft computer, we may be interested in estimating the probability of the event that the system fails before a mission time. Naturally, one would want this to be extremely low. The main problem with using standard simulation to estimate such small probabilities is that a large number of events have to be simulated in the model before any samples of the rare event may occur. Hence special simulation techniques are needed to make the events of interest occur more frequently.

Importance sampling is a technique that can be used for this purpose. This technique was initially used in the area of Monte Carlo integration (see, e.g., Kahn and Marshall 1953). An extension of the basic concept to stochastic processes may be found in Glynn and Iglehart (1989). In importance sampling, the stochastic model is simulated with a new probability dynamics, that makes the events of interest occur more frequently. The sample value is then adjusted to make the final estimate unbiased. However, choosing any change of measure that makes the event of interest occur frequently is not enough; *how* it is made to happen more frequently is also very important. For example, an arbitrary change of measure that makes the rare event happen more frequently may give an estimator with an infinite variance. Thus the main problem in importance sampling is to come up with an appropriate change of measure for the rare event simulation problem in hand. Different classes of stochastic models may use changes of measure that are totally different in nature.

Another method which makes rare events happen more frequently is a technique introduced in Bayes (1970). In standard simulation, the stochastic process being simulated, wastes a lot of time in a region of the state space which is "far away" from the rare set of interest, i.e, from where the chance of it entering the rare set is extremely low. In Bayes (1970), a region of the state space that is "closer" to the rare set is defined. Each time the process reaches this region, from the "far away" region, many identical copies of this process are generated. In simulation terminology this is called "splitting". Each of the split copies is simulated till it exits back into the "far away" region. From there on, only one of the split copies is continued until another entrance into the "closer" region.

This way we get more instances of the stochastic process spending time in a region where the rare event is more likely to occur.

There are a few software based modeling tools which use these rare event simulation techniques. SAVE (see Blum et al. 1994) incorporates a provably efficient importance sampling heuristic for reliability models called balanced failure biasing. UltraSAN (see Obal and Sanders 1994) gives the user the capability to specify an importance sampling change of measure of his/her choice for certain classes of stochastic models. Importance sampling methods for estimating the normalization constants of multiclass closed queueing networks are incorporated in MonteQueue (see Ross, Tsang and Wang 1994). A version of the splitting method mentioned above has been implemented in ASTRO (see Villen-Altamirano and Villen-Altamirano 1994).

The rest of the paper is organized as follows. The problem statement and the quantities to be estimated are given in Section 2. In Section 3 we illustrate the rare event simulation problem. In Section 4, the general concept of importance sampling is described. Importance sampling for reliability models are presented in Section 5.1, and for queueing models in Section 5.2. Applications of importance sampling for other types of stochastic models have been summarized in Section 5.3. Section 6 describes the splitting technique mentioned above.

2 PROBLEM STATEMENT

Consider a stochastic process $\{X(s) : s \geq 0\}$ on state space $\mathcal{S}$. We partition $\mathcal{S}$ into two subsets: $\mathcal{S} = \mathcal{G} \cup \mathcal{B}$ where $\mathcal{B}$ is the set of system states which are rare and of interest, and $\mathcal{G} = \mathcal{S}/\mathcal{B} \equiv \bar{\mathcal{B}}$. Suppose the process reaches steady state, i.e. $X(s) \Rightarrow X_\infty$ as $s \to \infty$, for some random variable X_∞. One measure of interest is estimating $\alpha = E(f(X_\infty))$ where $f(x) = 1_{\{x \in \mathcal{B}\}}$. From the physical point of view, this is the long run fraction of time the process spends in the rare state $\mathcal{B}$. Sometimes we may also be interested in the mean time between visits to the set $\mathcal{B}$, while the process is in steady state. We denote this by β.

We may also wish to estimate certain transient quantities like the fraction of time during $[0, t]$ the process spends in the set $\mathcal{B}$, i.e., $\alpha(t) = E(\int_{s=0}^{t} f(X(s))ds/t)$. Let $\tau_{\mathcal{B}}$ be the first time the process hits the set $\mathcal{B}$. Then $E(\tau_B)$ and $P(\tau_B < t)$ are also measures of interest in many situations.

3 RARE EVENT SIMULATION

Here we illustrate mathematically the basic problem of rare event simulation. Let Z be a random entity with probability measure $p(\cdot)$ on its sample space Ω and let $\mathcal{R}$ be a rare (under $p(\cdot)$) subset of the sample space. The problem may be to estimate $\gamma = P(\mathcal{R}) \equiv E_p(1_{\{Z \in \mathcal{R}\}})$ where the subscript in the expectation denotes the probability measure assigned to the random variable Z. Systems which have rare events are characterized by a rarity parameter ϵ so that as $\epsilon \to 0$, $\gamma \to 0$. For example, in a reliability system with highly reliable components, ϵ may be the maximum failure rate of components in the system. For a queueing system with buffer size B, we can set $\epsilon = 1/B$ so that as $\epsilon \to 0$, the buffer overflow event becomes rarer. In standard simulation, we generate n samples of the random variable Z, say $Z_1, Z_2, \ldots Z_n$ and estimate γ by using $\hat{\gamma} \equiv \sum_{i=1}^{n} 1_{\{Z_i \in \mathcal{R}\}}/n$. For fixed n, the half width (HW_p) of the confidence interval is (approximately) directly proportional to $\sqrt{Var_p(1_{\{Z \in \mathcal{R}\}})} = \sqrt{\gamma - \gamma^2} \approx \sqrt{\gamma}$ for small γ. Consequently, the relative error $RE_p \equiv HW_p/\gamma$, is directly proportional to $\sqrt{Var_p(1_{\{Z \in \mathcal{R}\}})}/\gamma$. It is then easy to see that $RE_p \to \infty$ as $\epsilon \to 0$. Equivalently, the simulation run length n, required to achieve a fixed relative error, RE_p, goes to ∞ as $\epsilon \to 0$.

A related problem is the estimation of $\psi = E_p(W(Z)1_{\{Z \in R\}})$ where $W(\cdot)$ is some function with domain Ω. Since $\mathcal{R}$ is rare, this expectation tends to be small and difficult to estimate. From the representation of ψ given above, it will not seem surprising that fast simulation techniques that work for the estimation of γ also seem to work for ψ.

4 IMPORTANCE SAMPLING

Let $p'(\cdot)$ be another probability measure on the sample space of Z, so that $p'(z) > 0$ whenever $p(z) > 0$ for all x in $\mathcal{R}$. Then

$$\gamma = \int_{z \in \Omega} 1_{\{z \in \mathcal{R}\}} p(z)dz = E_{p'}(1_{\{Z \in \mathcal{R}\}} L_{p'}(Z)) \quad (1)$$

where the subscript in the expectation denotes the probability measure assigned to Z and $L_{p'}(\cdot)$ is the likelihood ratio, i.e., $L_{p'}(z) \equiv p(z)/p'(z)$ whenever $p'(z) > 0$ and 0 otherwise. Equation (1) suggests that we can use $p'(\cdot)$ instead of $p(\cdot)$ to generate n samples of Z and then use $\hat{\gamma} \equiv \sum_{i=1}^{n} 1_{\{Z_i \in \mathcal{R}\}} L_{p'}(Z_i)/n$ as an unbiased estimator of γ. This is called importance sampling. The problem is to choose a $p'(\cdot)$ so that

$$Var_{p'}(1_{\{Z \in \mathcal{R}\}} L_{p'}(Z)) \ll Var_p(1_{\{Z \in \mathcal{R}\}}).$$

In most rare event simulations with importance sampling, an attempt is made to come up with changes of measure $p'(\cdot)$, so that the relative error, $RE_{p'}$, remains bounded as $\epsilon \to 0$. This is known as the *bounded relative error* (BRE) property.

To apply importance sampling to estimate the measures given in Section 2, one first has to represent

these measure in terms of small probabilities and expectations of the type given in Section 3. First consider the case where $\{X(s) : s \geq 0\}$ is regenerative (see Crane and Iglehart 1975). Pick a regenerative state that is frequently visited and let τ be the corresponding regenerative cycle time. Define W to be the amount of time in a regenerative cycle that the process spends in the rare set $\mathcal{B}$. Then α, may be represented as the ratio $E(W)/E(\tau)$. In most cases $E(\tau)$ is not small and is thus easy to estimate using standard simulation. However most samples of W are zero and thus we need importance sampling to estimate $E(W)$ accurately. Relating to Section 3, the Z corresponds to a random sample path *in a regenerative cycle* of the above stochastic process, the $p(\cdot)$ is the original probability measure on these sample paths, $\mathcal{R}$ is the set of these sample paths that visit $\mathcal{B}$, and $W(Z) \equiv W$ is the amount of time the sample path Z spends in $\mathcal{B}$. Thus the rare event simulation problem becomes estimating $E_p(W(Z)1_{\{Z \in \mathcal{R}\}}) = E_p(W(Z)) \equiv E(W)$. The changes of measure $p'(\cdot)$ used are such that they induce a drift in the stochastic process towards $\mathcal{B}$ so that the chance of the rare event $\mathcal{R}$ happening is increased. However, once $\mathcal{B}$ is visited, the stochastic process is simulated with the usual dynamics so that the regenerative cycle completes soon (Goyal et al. 1992).

Given that at time $t = 0$ the system is in the regenerative state, the $E(\tau_B)$ defined in Section 2 may be represented as the ratio $E(\tau_{min})/P(\mathcal{R})$ (see, e.g., Keilson 1979). Here $P(\mathcal{R})$ is the probability of the rare event $\mathcal{R}$ (i.e., of hitting $\mathcal{B}$ in a regenerative cycle) and τ_{min} is the time to hit either $\mathcal{B}$ or the regenerative state given that the process starts in the regenerative state Again, $E(\tau_{min})$ is easy to estimate using standard simulation. However, we have to use importance sampling to estimate $P(\mathcal{R})$.

The other transient measures are naturally in terms of small probabilities and expectations, if the time horizon is small. However, for large time horizons, the importance sampling variance may be grow exponentially with time (see Glynn 1994). In those cases, some other representations of the transient measures may prove useful (see, e.g., Shahabuddin and Nakayama 1993).

Now consider the case where $\{X(s) : s \geq 0\}$ is non-regenerative (this also applies when the regenerative cycles are very long, so that the regenerative simulation procedure cannot be used effectively). One can still use a ratio representation of the steady state measure. Let $\mathcal{A}$ be a state or a set of states that are visited quite frequently in the simulation. Define an $\mathcal{A}$-cycle to start whenever the process enters the set $\mathcal{A}$. Then the ratio formula $\alpha = E(W)/E(Z)$ still holds where now $W \equiv W(Z)$ is the amount of time the process spents in the set $\mathcal{B}$ in an $\mathcal{A}$-cycle and $\tau = \tau(Z)$ represents the duration of an $\mathcal{A}$-cycle, given that the process is in steady state (e.g., Cogburn 1975). The actual simulation procedure uses a splitting technique combined with batch means (Nicola et al. 1993). We first run a few $\mathcal{A}$-cycles so that the system (approximately) reaches steady state. After that, each time an $\mathcal{A}$-cycle starts, we split a process from the original process. The split process uses the change of measure as prescribed by the importance sampling. We use this split $\mathcal{A}$-cycle to get a sample of $W(Z)$ and $L_{p'}(Z)$ and use the original $\mathcal{A}$-cycle to get a sample of $\tau(Z)$. Since the successive $\mathcal{A}$-cycles are generally not independent, we have to use the procedure of batch means to build confidence intervals. A similar splitting idea was used in Al-Qaq, Devetsikiotis and Townsend (1993) for estimating bit error rates over certain communication channels.

We can also use a ratio-representation to estimate β which was defined in Section 2: $\beta = E(\tau)/E(N)$ where N is the number of visits to $\mathcal{B}$ during an $\mathcal{A}$-cycle (Glynn et al. 1993). Again, $E(N)$ is the small expectation which we have to estimate using importance sampling.

5 APPLICATIONS OF IMPORTANCE SAMPLING

5.1 Reliability Models

Consider the fairly general class of reliability models considered in Blum et al. (1994). These are systems consisting of components that fail and get repaired. Components are not independent in the sense that they share repairmen, they have operational/repair dependencies and there may be failure propagation (i.e., the failure of a component may cause another component to fail instantaneously). If we assume that component failure times and repair times are exponentially distributed, then the system can be modelled as a continuous time Markov chain (CTMC) $\{X(s) : s \geq 0\}$. For example, in the simplest such system, $X(s) = (X_1(s), X_2(s), \ldots, X_N(s))$ where $X_i(s)$ may be considered to be the number of components of type i that are up and N is the total number of component types. A transition of the CTMC corresponds corresponds to either a component failure transition or a component repair transition. In mathematical models of highly reliable systems, the failure rate of a component, say component i, is represented as $\lambda_i \epsilon^{r_i}$ where ϵ is the rarity parameter, and r_i, λ_i are positive constants (i.e, independent of ϵ). The r_i's may be different if the system is "unbalanced", i.e., components have failure rates that are of different orders of magnitude. Since the repair rates are comparatively large, they are represented by a constant.

First consider steady state estimation. For this purpose one can simulate the embedded discrete time Markov chain of the CTMC. Let $\mathbf{P} = \{P_{x,y} : x, y \in$

$\mathcal{S}\}$ denote the transition matrix for this Markov chain. The regenerative state is taken to be the one in which all components are up. Relating to Section 2, the $\mathcal{B}$ corresponds to the set of states of the CTMC in which the system is considered failed. The changes of measure used here are called "failure biasing heuristics" and correspond to simulating the system using a new probability transition matrix $\mathbf{P}' = \{P'_{x,y} : x, y \in \mathcal{S}\}$, with the property that for any states x, y, $P'_{x,y} > 0$ if $P_{x,y} > 0$. If the state $\mathcal{B}$ is visited before the regenerative cycle completes, then the transition matrix $\mathbf{P}$ is used for the remainder of the cycle. The intuitive idea behind these heuristics is to artificially make failure transitions happen much more frequently than in the actual system. The original heuristic was introduced by Lewis and Bohm (1984) and is now called simple failure biasing in the literature. However, this heuristic does not have the BRE property for unbalanced systems (Shahabuddin 1994). A modified technique called balanced failure biasing (Goyal et al. 1992, Shahabuddin 1994) has been proven to have the BRE property in Shahabuddin (1994). The following compact representation has been taken partly from Nakayama (1994).

Algorithm: Balanced Failure Biasing

- From any state x, let $\Lambda_F(x)$ (cf. $\Lambda_R(x)$) be the set of transitions (x, y) that correspond to component failure (cf. repair) transitions. Let $p_F(x) \equiv \sum_{y \in \Lambda_F(x)} P_{x,y}$ and $p_R(x) \equiv \sum_{y \in \Lambda_R(x)} P_{x,y}$. Define $I_{x,y} = 1$ if $P_{x,y} > 0$ and $I_{x,y} = 0$ otherwise. For any state x, define $n_F(x)$ to be the number of failure transitions possible (under $\mathbf{P}$) from x. Let p^*, $0 < p^* < 1$, be a constant. In practice, $0.5 \leq p^* \leq 0.9$.

- If $p_R(x) > 0$ then

$$P'_{xy} = \begin{cases} p_* I_{x,y}/n_F(x) & \text{if } (x,y) \in \Lambda_F(x) \\ (1-p_*)P_{x,y}/p_R(x) & \text{if } (x,y) \in \Lambda_R(x) \\ 0 & \text{otherwise} \end{cases}$$

- If $p_R(x) = 0$, let $P'_{x,y} = I_{x,y}/n_F(x)$ if $(x, y) \in \Lambda_F(x)$ and $P_{x,y} = 0$ otherwise.

A crucial assumption used in Shahabuddin (1994) to prove the BRE property of balanced failure biasing is that all states of the Markov chain, except the state in which all components are up, have at least one component repair transition. In situations where this assumption does not hold (e.g. deferred repair), balanced failure biasing may give infinite variance. This was shown in Shahabuddin and Juneja (1992) who also developed an improved failure biasing scheme for the fast simulation of such systems. Another failure biasing heuristic based on the concept of failure distances may be found in Carrasco (1992). A detailed investigation of the conditions on systems under which failure biasing heuristics give bounded relative error may be found in Nakayama (1993) (and some references therein). However, so far it appears difficult to use these results in practice. Some additional results in this regard may be found in Strickland (1995). For results and references on derivative estimation the reader is referred to Nakayama (1995).

In the case of estimating transient measures in time interval $[0, t]$ like the unreliability and the expected interval unavailability, just using failure biasing is not enough. We also have to use some mechanism to ensure that the first transition happens before time t. This is termed forcing and was introduced in Lewis and Bohm (1984). It is shown in Shahabuddin (1994) and Shahabuddin and Nakayama (1993) that forcing combined with failure biasing produces BRE in the estimation of the reliability and the interval availability for cases where t is small (t is either of the same order as the regenerative cycle time or smaller). However for cases where t is large the relative error tend to infinity. For such cases, a method based on estimating Laplace transform functions of the transient measure is studied in Carrasco (1991) and another approach based on estimating bounds to the transient measure (rather than estimating the actual measure) is studied in Shahabuddin (1994) and Shahabuddin and Nakayama (1993).

For non-Markovian models, an importance sampling approach based on re-scheduling failure events is given in Nicola et al. (1991). Two other approaches, one based on uniformization, and the other which was called exponential transformation, were introduced for estimating the unreliability in Heidelberger, Shahabuddin and Nicola (1994) and shown to have the BRE property under fairly general conditions.

Some work has also been done in the area of estimation of unreliability in a network with independent components. A network is modelled as a graph whose edges represent the components. The network is said to fail if the connectivity between two (disjoint) sets of nodes is lost. Again, if the components are highly reliable, then the chance of a network failure is very small. The reader is referred to Fishman (1986) and Lieber, Elmakis and Rubinstein (1994) for some importance sampling schemes used in this area.

5.2 Queueing Models

5.2.1 Queues with I.I.D. Renewal Input Stream

Changes of measures for queueing models, that have the BRE property, have been proposed and studied in Cottrell, Fort and Malgouvres (1983), Parekh and Walrand (1989), Sadowsky (1991), among many

others. Most of these provably efficient changes of measure in the above papers apply to a single server queueing system where the arrival stream constitutes an i.i.d. renewal process. The measures of interest have been the tail distribution of the steady state waiting time and queue length in systems with infinite buffer; the steady state customer loss probability in systems with finite buffer. The change of measure is based on "exponentially tilting" the arrival and the service distribution. Let $F_A(\cdot)$ (cf. $F_S(\cdot)$) denote the original interarrival time (cf. service time) time distribution and let $M_A(\cdot)$ (cf. $M_S(\cdot)$) be its moment generating function. The new inter-arrival time (cf. service time) distribution $\tilde{F}_A(\cdot)$ (cf. $\tilde{F}_S(\cdot)$) corresponding to the provably efficient change of measure, is determined as follows:

Algorithm: GI/GI/1 Queue

- Let θ^* be the solution of $M_A(-\theta)M_S(\theta) = 1$, $\theta > 0$.
- Then $d\tilde{F}_A(x) = e^{-\theta^* x} dF_A(x)/M_A(\theta^*)$ and $d\tilde{F}_S(x) = e^{\theta^* x} dF_S(x)/M_S(\theta^*)$.

For the M/M/1 queueing system with arrival rate λ and service rate μ (with $\lambda/\mu < 1$) this change of measure corresponds to interchanging the arrival rate and the service rate. Note that this makes the queue unstable so that large queue lengths are reached much faster. Sadowsky (1991) presents a provably efficient change of measure for the GI/GI/m queueing system.

Extensions of these provably efficient changes of measures to networks have been few and apply mainly to Markovian tandem networks (e.g., Glasserman and Kou 1994). Heuristical approaches for fast simulation of more general networks, based on a large deviations approach, may be found in Parekh and Walrand (1989) and Frater, Lennon and Anderson (1991). Ross, Tsang and Wang (1994) have used importance sampling to estimate the normalizing constant which occurs in the solution of multi-class product form closed queueing networks.

5.2.2 Queues with Correlated Arrival Processes

Provably efficient changes of measures for discrete time queues with autocorrelated arrival processes were studied in Chang et al. (1994) and continuous time versions in Juneja (1993) (see, e.g., Lehtonen and Nehriyen 1992 for analogous concepts in the context of risk analysis). Fast simulation of Markov fluid models of such queues have been studied in Kesidis and Walrand (1993) and Mandjes and Ridder (1995). Chang et al. (1994) also linked fast simulation techniques for ATM switches to the concept of effective bandwidth (see, e.g., Chang 1994) of the arrival sources, thus generalizing the class of source models that can be handled and allowing the study to be extended to the class of intree networks. Some critical concepts in Chang et al. (1994), dealing with effective bandwidths in intree networks, were also developed independently in de Veciana et al. (1993). Since the literature on this subject is very vast, we just present the algorithm for a simple discrete time queue that is fed by a Markov modulated arrival processes (MMAP).

Consider a single discrete time queue system that is fed by K external sources, each of which is a MMAP. For simplicity, we consider the simplest form of such an arrival process. Let the kth source be in any of the M_k states $\{0, 1, \ldots M_{k-1}\}$. Let $Y_k(t)$ be the state of the kth source after time t, and let $p_k(i,j) = P(Y_k(t+1) = j | Y_k(t) = i)$. Let the number of packets a source transmits per unit of discrete time, $a_k(t)$, be equal to the current state of the source and let $a(t) = \sum_{k=1}^{K} a_k(t)$ be the total arrival to the queue in that unit of discrete time. We assume that the queue has the capacity to dispatch c packets every unit of discrete time. Let B denote the size of the buffer. Then the number of people in the system at time t is governed by the following Lindley type recursion: $Q(t+1) = (\min(Q(t) + a(t+1), B) - c)^+$. The problem is to estimate the steady state probability of packet loss when $\epsilon \equiv 1/B$ is small. In this case we let the set $\mathcal{A}$ (corresponding to an $\mathcal{A}$-cycle) to be the set of states of the extended Markov chain $(Q(t), Y_1(t) \ldots Y_K(t))$ that have $Q(t) = 0$. Let $\lambda_{k,\theta}$ be the spectral radius of the matrix that has elements $\mathcal{A}_k(i,j) = e^{\theta j} p_k(i,j)$ and let $h_{k,\theta}(j)$ be the corresponding eigenvector. The provably efficient change of measure corresponds to doing a sort of exponential tilting to the MMAPs (the service rate c remains unchanged). The new transition matrix for the kth MMAP, $p'_k(i,j)$, can be determined as follows:

Algorithm: Queue with MMAP arrival

- Let θ^* be the solution of the equation $\sum_{k=1}^{K} log(\lambda_{k,\theta}) = c$ and $\theta > 0$.
- Then $p'_k(i,j) = e^{\theta j} p_k(i,j) h_{k,\theta}(j) / \lambda_{k,\theta} h_{k,\theta}(i)$.

5.3 Other Areas

For references to applications of importance sampling to risk analysis and sequential analysis the reader is referred to Lehtonen and Nyrhinen (1992) and Siegmund (1985) respectively. The concepts used in both the above areas are similar to those used for queueing models. Recently, importance sampling has been used for estimating various measures of service (stockout frequency, fill rate, and average backlogs) in multistage production inventory systems. Standard simulation is no longer effective for estimating these mea-

sures when the inventories become critical, i.e., when either the back order penalty is very large, or the target service level is very high. Glasserman and Liu (1994) developed importance sampling changes of measure for such systems and proved BRE type properties.

Importance sampling has also been used in the estimation of the bit-error rate in digital communication systems. The reader is referred to Al-Qaq, Devetsikiotis and Townsend (1993) for a list of references in this area.

For importance sampling applied to general Markov chains, refer to Andradottir, Heyman and Ott (1995), Glynn (1994) and references therein.

6 THE SPLITTING METHOD

This method was introduced in Bayes (1970). Incidentally, Bayes (1970) referred to this method as "importance sampling", as it does require sampling from a region of importance. However, in the current literature, the definition of importance sampling no longer seems to include this method. Hence we think that a more appropriate term for it may be "importance splitting".

Consider the stochastic process $\{X(s) : s \geq 0\}$ mentioned in Section 2, where the problem is to estimate α. The usual method is to first simulate the process till it (approximately) reaches steady state. After that, we simulate it for a interval of time t. For convenience, assume that at $s = 0$ the process is in steady state. Then $Y = (\int_{s=0}^{t} f(X(s))ds)/t$ gives an unbiased (one sample) estimate of α. One can use either the replication-deletion method or the batch means method to construct confidence intervals.

Let $\mathcal{C} \subset \mathcal{S}$ be such that $\mathcal{B} \subset \mathcal{C}$, and the steady state probability of being in state C is not as small as α. By an "upcrossing" we will mean the stochastic process going from $\bar{\mathcal{C}}(\equiv \mathcal{S}/\mathcal{C})$ to $\mathcal{C}$. A "downcrossing" will mean the opposite. The following is a polished version of the algorithm in Bayes (1970).

Algorithm: Splitting Method

1. Set $j = 0$. Set simulation time $s = 0$, and the cumulator $sum = 0$.

2. Simulate one copy of the process until the next upcrossing. Update $j \leftarrow j + 1$ and set s_j to be the absolute time of this upcrossing. If $s_j < t$, update $s \leftarrow s_j$; otherwise end the simulation and go to Step 5.

3. At s_j generate R split processes, each with the starting state $X(s_j)$ and simulate each split process till a downcrossing. Let Δ_r be the amount of this elapsed time (after s_j) for the rth split process and let Y_r be the amount of time in the interval $[s_j, \min\{s_j + \Delta_r, t\}]$ that the rth split process spends in the set $\mathcal{B}$. Let $\bar{\Delta} = \sum_{r=1}^{R} \Delta_r/R$. Update $sum \leftarrow sum + \sum_{r=1}^{R} Y_r/R$. Advance the simulation time to $s \leftarrow s + \bar{\Delta}$.

4. If $s > t$ then end the simulation and go to Step 5; otherwise set $X(s)$ equal to the state at the downcrossing of the Rth split path. Go to Step 2.

5. Let $\tilde{Y}$ be the amount of time in the interval $[0, \min\{s_1, t\}]$ that the process spent in $\mathcal{B}$. Form the estimator $\hat{\alpha} = (\tilde{Y} + sum)/t$.

Bayes (1970) called the boundary between $\mathcal{C}$ and $\bar{\mathcal{C}}$ the "importance level". A possible generalization of this scheme to the case of multiple importance levels was also mentioned.

Hopmans and Kleijnen (1979) investigated the above algorithm in detail (for the one dimensional, one level case) using a regenerative assumption, i.e., the system regenerates each time we have an upcrossing (or a downcrossing). In that sense the algorithm which they use is slightly different in its execution then what is given above. By conducting a variance analysis they determined the optimum R. They applied it to a complex telecommunication system model but they were not very satisfied with the improvement in efficiency obtained. Villen-Altamirano and Villen-Altamarino (1991) revisited this idea and proposed a slightly modified schemes to the one in Bayes (1970) which they called RESTART. The only difference from the Bayes (1970) algorithm is that in Step 3, instead of s being updated to $s + \bar{\Delta}$, it is updated to $s + \Delta_R$. They also did a variance analysis of their scheme to determine the optimum R and the optimum placement of the level, and then computed the efficiency gain obtained. Experiments using this scheme produced significant variance reduction on particular examples. Generalization of the scheme and the variance analysis to the multi-level case was done in a later paper (see Villen-Altamirano and Villen-Altamirano (1994) for a complete set of references).

7 ACKNOWLEDGEMENT

This work was done while the author was at the IBM T.J. Watson Research Center, Yorktown Heights, New York.

REFERENCES

Al-Qaq, W.A., M. Devetsikiotis, and J.K. Townsend. 1993. Importance sampling methodologies for simulation of communication systems with adaptive

equalizers and time varying channels. *IEEE Journal on Selected Areas in Communications* 11: 317-327.

Andradottir, S., D.P. Heyman and T.J. Ott. 1995. On the choice of alternative measures in importance sampling with Markov chains. *Operations Research* 43(3): 509-519.

Asmussen, S. and R.Y. Rubinstein. 1994. Steady state rare event simulation in queuing models and its complexity properties. Manuscript, Institute of Electronic Systems, Aalborg University, Denmark.

Bayes, A.J. 1970. Statistical techniques for simulation models. *The Australian Computer Journal* 2: 180-184.

Blum, A., A. Goyal, P. Heidelberger, S.S. Lavenberg, M.K. Nakayama, and P. Shahabuddin. 1994. Modeling and analysis of system dependability using the system availability estimator. In *Digest of Papers. The Twenty-Fourth Annual International Symposium of Fault-Tolerant Computing*, 137-141, IEEE Computer Society Press.

Carrasco, J.A. 1991. Efficient transient simulation of failure/repair Markovian models. In *Proceedings of the Tenth Symposium on Reliable and Distributed Computing*, 152-161, IEEE Computer Society Press.

Carrasco, J.A. 1992. Failure distance-based simulation of repairable fault-tolerant systems. In *Proceedings of the Fifth International Conference on Modeling Techniques and Tools for Computer Performance Evaluation*, 351-365.

Chang, C.S. 1994. Stability, queue length and delay of deterministic and stochastic queueing networks. *IEEE Transactions on Automatic Control* 39: 913-931.

Chang, C.S., P. Heidelberger, S. Juneja, and P. Shahabuddin. 1994. Effective bandwidth and fast simulation of ATM intree networks. *Performance Evaluation* 20: 45-65.

Cogburn, R. 1975. A uniform theory for sums of Markov chain transition probabilities. *The Annals of Probability* 3: 191-214.

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

Crane, M.A., and D.L. Iglehart. 1975. Simulating stable stochastic systems III: regenerative processes and discrete event simulation. *Operations Research* 23: 33-45.

de Veciana, G., C. Courcoubetis, and J. Walrand. 1993. Decoupling bandwidth for networks: a decomposition approach for resource management for networks. Manuscript, University of Texas at Austin, TX. Earlier version in *IEEE INFOCOM'94 Proceedings*, 466-473, IEEE Computer Society Press.

Devetsikiotis, M., and J.K. Townsend. 1993. Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks. *IEEE/ACM Transactions on Networking* 1: 293-305.

Fishman, G.S. 1986. A Monte-Carlo sampling plan for evaluating network reliability. *Operations Research* 34(4): 581-594.

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

Glasserman, P., and S.G. Kou. 1994. Analysis of an importance sampling estimator for tandem queues. Manuscript, Columbia University, New York, NY.

Glasserman, P., and T. Liu. 1994. Rare-event simulation for multistage production inventory systems. Manuscript, Columbia University, New York, NY.

Glynn, P.W. 1994. Importance sampling for Markov chains: asymptotics for the variance. *Stochastic Models* 10: 701-717.

Glynn, P.W., and D.L. Iglehart. 1989. Importance sampling for stochastic simulations. *Management Science* 35: 1367-1392.

Glynn, P.W., P. Heidelberger, V.F. Nicola, and P. Shahabuddin. 1993. Efficient estimation of the mean time between failures in non-regenerative dependability models. In *Proceedings of the 1993 Winter Simulation Conference*, 311-316, IEEE Press.

Goyal, A., P. Shahabuddin, P. Heidelberger, V.F. Nicola, and P.W. Glynn. 1992. A unified framework for simulating Markovian models of highly reliable systems. *IEEE Transactions on Computers* C-41: 36-51.

Heidelberger, P. 1993. Fast simulation of rare events in queueing and reliability models. To appear in *ACM Transactions on Modeling and Computer Simulation.*

Heidelberger, P., P. Shahabuddin, and V.F. Nicola. 1994. Bounded relative error in estimating transient measures of highly dependable non-Markovian systems. *ACM Transactions on Modeling and Computer Simulation* 4: 137-164.

Hopmans, A.C.M., and J.P.C. Kleijnen. 1979. Importance sampling in system simulation: a practical failure? *Mathematics and Computing in Simulation XXI* : 209-220.

Juneja, S. 1993. Efficient rare event simulation of stochastic systems. Ph.D. Thesis, Department of Operations Research, Stanford University, California.

Juneja, S., and P. Shahabuddin. 1992. Fast simulation of Markovian reliability/availability models with general repair policies. In *Proceedings of the Twenty-Second Annual International Symposium on Fault Tolerant Computing*, 150-159, IEEE

Computer Society Press.

Kahn, H., and A.W. Marshall. 1953. Methods of reducing sample size in Monte Carlo computations. *Journal of the Operations Research Society 1(5)*: 263-278.

Keilson, J. 1979. *Markov chain models - rarity and exponentiality.* New York, NY: Springer Verlag.

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

Lehtonen, T., and H. Nyrhinen. 1992. Simulating level-crossing probabilities by importance sampling. *Advances in Applied Probability* 24: 858-874.

Lewis, E.E., and F. Bohm. 1984. Monte Carlo simulation of Markov unreliability models. *Nuclear Engineering and Design* 77: 49-62.

Lieber, D., R.Y. Rubinstein, and D. Elmakis. 1994. Quick estimation of rare events in stochastic networks. Manuscript, Technion–Israel Institute of Technology, Haifa.

Mandjes, M., and A. Ridder. 1994. Finding the conjugate of Markov fluid processes. To appear in *Probability in the Engineering and Informational Sciences.*

Nakayama, M.K. 1993. General conditions for bounded relative error in simulation of highly reliable Markovian systems. Manuscript, New Jersey Institute of Technology, Newark, NJ.

Nakayama, M.K. 1994. Fast simulation methods for highly dependable systems. In *1994 Winter Simulation Conference Proceedings*, 221-228, IEEE Press.

Nakayama, M.K. 1995. Likelihood ratio derivative estimators in simulations of reliable Markovian systems. *Management Science* 41: 524-554.

Nicola, V.F., M.K. Nakayama, P. Heidelberger, and A. Goyal. 1991. Fast simulation of highly dependable systems with general failure and repair processes. *IEEE Transactions on Computers* 42(8): 1440-1452.

Nicola, V.F., P. Shahabuddin and P. Heidelberger. 1993. Techniques for fast simulation of highly dependable systems. In *Proceedings of the Second International Workshop on Performability Modelling of Computer and Communication Systems.*

Nicola, V.F., P. Shahabuddin, P. Heidelberger and P.W. Glynn. 1993. Fast simulation of steady-state availability in non-Markovian highly dependable systems. In *Proceedings of the Twenty-Third International Symposium on Fault-Tolerant Computing*, 38-47, IEEE Computer Society Press.

Obal II, W.D., and W. H. Sanders. 1994. Importance sampling simulation in UltraSAN. *Simulation 62*: 98-111.

Parekh, S., and J. Walrand. 1989. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34: 54-56.

Ross, K.W., D.H.K Tsang, and J. Wang. 1994. Monte Carlo summation and integration applied to multiclass queueing networks. *Journal of the ACM* 41(6): 1110-1135.

Sadowsky, J.S. 1991. Large deviations and efficient simulation of excessive backlogs in a GI/G/m queue. *IEEE Transactions on Automatic Control* 36: 1383-1394.

Shahabuddin, P. 1994. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science* 40: 333-352.

Shahabuddin, P. 1994. Fast transient simulation of Markovian models of highly dependable systems. *Performance Evaluation* 20: 267-286.

Shahabuddin, P., and M.K. Nakayama. 1993. Estimation of reliability and its derivatives for large time horizons in Markovian systems. In *Proceedings of the 1993 Winter Simulation Conference*, 422-429, IEEE Press.

Siegmund, D. 1985. *Sequential analysis: tests and confidence intervals.* New York: Springer Verlag.

Strickland, S.G. 1995. Necessary and sufficient conditions for for bounded relative error in importance sampling. Manuscript, University of Virginia, Charlottesville, VA.

Villen-Altamirano, M., and J. Villen-Altamirano. 1991. RESTART: a method for accelerating rare events simulation. In *Proceedings of the 13th International Teletraffic Congress, Queuing performance and control in ATM*, 71-76, North Holland Publishing Company.

Villen-Altamarino, M., and J. Villen-Altamirano. 1994. RESTART: a straightforward method for fast simulation of rare events. In *Proceedings of the 1994 Winter Simulation Conference*, 282- 289, IEEE Press.

AUTHOR BIOGRAPHIES

PERWEZ SHAHABUDDIN has been an Assistant Professor at the Industrial Engineering and Operations Research Department at Columbia University, New York, NY, since Fall 1995. He is currently on a leave of absence from the IBM T.J. Watson Research Center, Yorktown Heights, NY, where he has been a Research Staff Member since 1990. He received his B.Tech in Mechanical Engineering from the Indian Institute of Technology, Delhi, in 1984, followed by a M.S. in Statistics and a Ph.D in Operations Research from Stanford University in 1987 and 1990, respectively. From 1984 to 1985 he worked as a system analyst at Engineers India Limited, India. His research interests include modeling and analysis of stochastic systems, and methods for simulation variance reduction.