Fast Simulation of Rare Events in Queueing and Reliability Models

Philip Heidelberger

IBM T.J. Watson Research Center, Hawthorne
P.O. Box 704
Yorktown Heights, New York 10598

Abstract

This paper surveys efficient techniques for estimating, via simulation, the probabilities of certain rare events in queueing and reliability models. The rare events of interest are long waiting times or buffer overflows in queueing systems, and system failure events in reliability models of highly dependable computing systems. The general approach to speeding up such simulations is to accelerate the occurrence of the rare events by using importance sampling. In importance sampling, the system is simulated using a new set of input probability distributions, and unbiased estimates are recovered by multiplying the simulation output by a likelihood ratio. Our focus is on describing asymptotically optimal importance sampling techniques. Using asymptotically optimal importance sampling, only a fixed number of samples are required to get accurate estimates, no matter how rare the event of interest is. In practice, this means that the required run lengths can be reduced by many orders of magnitude, compared to standard simulation. The queueing systems studied include simple queues (e.g., GI/GI/1) and discrete time queues with multiple autocorrelated arrival processes that arise in the analysis of Asynchronous Transfer Mode communications switches. References for results on Jackson networks and and tree structured networks of ATM switches are given. Both Markovian and non-Markovian reliability models are treated.

1 Introduction

This survey paper is concerned with efficient simulation techniques for estimating the probability of certain rare, but important, events that arise in the analysis of queueing and reliability models. For example, consider a switch in a communications system that has a buffer capable of holding at most B packets. A switch designer is interested in the buffer sizing problem; namely determining a value of the buffer size so that the (steady state) packet loss probability is very small,

say less than 10^{-9}. A designer of network session admittance and routing algorithms is faced with a related, but somewhat different problem; namely deciding whether or not to allow a new session to use the switch given the traffic already using the switch and the fixed buffer size B of the switch. For the session admittance problem, one criterion would be to admit the new session provided the packet loss probability remains acceptably low (again, say less than 10^{-9}). Depending on the stochastic characteristics of the sources and the service processes, exact (or numerical) solution of the relevant queueing model may be impossible and (discrete event) simulation may be the only feasible solution approach.

In the reliability context, consider the design of a fault tolerant computer system. The system designer wishes to select components (e.g., processors and disk drives) and configure the system so that it is very reliable, subject to some cost constraints. For example, the designer may wish to find a configuration so that, given a fixed time horizon t (say a month), the probability that the system fails (or operates at an unacceptably low level of performance) within the interval $(0, t)$ is very small. Even under Markovian assumptions on the failure and repair time distribution of the components, such models become difficult to solve using numerical techniques due to the problem of state space size explosion. For generally distributed failure and repair times, effective numerical techniques, for all practical purposes, do not exist. Thus simulation may be the only feasible solution approach.

But how practical is it to use simulation to estimate such small probabilities? For example, to estimate a packet loss probability of less than 10^{-9} would seem to require simulating at least 10^9 packets. As the probability of interest becomes smaller, the required simulation time grows ever larger and becomes excessive, except perhaps on highly parallel computers. *Somewhat remarkably, simple techniques exist that permit very accurate estimation of such small probabilities within a matter of minutes on even modest-sized workstations.* The basic approach is called *importance sampling* [46, 54]; the technique was first applied on computers for performing nuclear physics calculations done during the 1940's in collaborations between von Neumann, Ulam, Fermi, Kahn, Metropolis and their colleagues [53, 54, 63]. In the past several years importance sampling has also been applied to a variety of problems arising in the analysis of computer and communications systems.

The purpose of this paper is to describe how importance sampling can be used to speed up rare event simulations in queueing and reliability models. While analysis of the efficiency of importance sampling techniques can tend to be highly mathematical, this paper will emphasize basic concepts, thereby (hopefully) making the paper accessible to practitioners with a background in performance and/or reliability modeling, but with little or no background in simulation methodology. The rest of the paper is organized as follows. In Section 2, the problem of rare event simulation will be discussed in more detail and the technique of importance sampling introduced. This section will also describe the notion of optimal importance sampling, along with other preliminaries such as a brief description of regenerative simulation. Section 3 will describe the ap-

plication of importance sampling in a variety of queueing systems, including the single server queue with independent interarrival and service time distributions (GI/GI/1), the multiple server queue (GI/GI/m), the single server queue with multiple correlated arrival processes. References for results on and some simple networks will be given. Section 4 will describe results for simulating models of highly dependable systems for the purpose of estimating reliability and availability characteristics. The paper is then summarized in Section 5, highlighting the similarities and differences of simulating queueing and reliability models.

2 Rare Event Simulation and Importance Sampling

2.1 The Problem of Rare Event Simulation

To further demonstrate the difficulty involved in simulating rare events, let us consider a simple example. Let X be a random variable that has a probability density function $p(x)$ and consider estimating the probability, γ, that X is in some set A;

$$\gamma = \int_{-\infty}^{\infty} 1_{\{x \in A\}} p(x) dx = E_p[1_{\{X \in A\}}] \tag{1}$$

where the subscript p denotes sampling from the density p and $1_{\{x \in A\}}$ is the indicator of the set A, i.e., $1_{\{x \in A\}} = 1$ if $x \in A$ and $1_{\{x \in A\}} = 0$ if $x \notin A$. Consider estimating γ by simulation. The standard approach would be to draw N samples, $X_1, \ldots, X_N$ from the density p, set $I_n = 1_{\{X_n \in A\}}$, and form the estimate $\hat{\gamma}_N = (1/N) \sum_{n=1}^{N} I_n$. Note that $E_p[\hat{\gamma}_N] = \gamma$ and that the variance of $\hat{\gamma}_N$ is $\gamma(1-\gamma)/N$. By the central limit theorem, a $100(1-\alpha)\%$ confidence interval for γ is (approximately) $\hat{\gamma}_N \pm z_{\alpha/2}\sqrt{\gamma(1-\gamma)/N}$ where $z_{\alpha/2}$ is defined by the equation $\alpha/2 = P(N(0,1) \geq z_{\alpha/2})$. ($N(0,1)$ denotes a normally distributed random variable with mean zero and variance one.) Suppose we wish to estimate γ to within $\pm 10\%$ (about two significant digits of accuracy), i.e., we want the relative half-width of, say, a 99% confidence interval for γ to be less than 0.1 which implies that $2.576\sqrt{\gamma(1-\gamma)/N}/\hat{\gamma}_N \leq 0.1$. How large must N be in order to achieve this level of accuracy? Since $\gamma_N \to \gamma$ almost surely (a.s.) as $N \to \infty$, we see that $N \sim 100 \times 2.576^2 \times (1-\gamma)/\gamma$, i.e., the required sample size is proportional to $1/\gamma$. The smaller γ is, the larger the sample size must be. For example, if $\gamma = 10^{-6}$, a sample size of 6.64×10^8 is required, while if $\gamma = 10^{-9}$, a sample size of 6.64×10^{11} (664 billion) is required. Note that these sample sizes must be increased by a factor of 100 in order to achieve one additional significant digit of accuracy ($\pm 1\%$ accuracy).

The relative error of the estimate $\hat{\gamma}_N$, $RE(\hat{\gamma}_N)$, is defined to be the standard deviation of the estimate divided by its expected value. Then $RE(\hat{\gamma}_N) \sim 1/\sqrt{\gamma N} \to \infty$ as $\gamma \to 0$. Thus, using standard simulation, the relative error is unbounded as the event becomes rarer.

As we will see, when importance sampling is properly applied, it is possible to construct (unbiased) estimates of γ whose relative error remains bounded even as $\gamma \to 0$. This implies that the required sample size to achieve a given relative error does not blow up as the event becomes rarer.

2.2 Importance Sampling

Importance sampling is based on the following simple observation. (Some of the discussion below basically follows that given in [54].) Consider the integral representation for γ in Equation 1. Multiplying and dividing the integrand by another density function $p'(x)$ we obtain (henceforth, we will suppress the limits of integration unless they are specifically required)

$$\gamma = \int 1_{\{x \in A\}} \frac{p(x)}{p'(x)} p'(x) dx = E_{p'} \left[1_{\{X \in A\}} \frac{p(x)}{p'(x)} \right] = E_{p'}[1_{\{X \in A\}} L(X)] \quad (2)$$

where $L(x) = p(x)/p'(x)$ is called the likelihood ratio and the subscript p' denotes sampling from the density p'. Equation 2 is valid for any density p' provided that $p'(x) > 0$ for all $x \in A$ such that $p(x) > 0$, i.e., a non-zero feasible sample under density p must also be a non-zero feasible sample under density p'. This equation suggests the following estimation scheme, which is called importance sampling. Draw N samples $X_1, \ldots, X_N$ using the density p' and define $Z_n = L(X_n) I_n$ (recall $I_n = 1_{\{X_n \in A\}}$). Then, by Equation 2, $E_{p'}[Z_n] = \gamma$. Thus an unbiased (and strongly consistent as $N \to \infty$) estimate of γ is given by

$$\hat{\gamma}_N(p') = \frac{1}{N} \sum_{n=1}^{N} Z_n = \frac{1}{N} \sum_{n=1}^{N} I_n L(X_n), \quad (3)$$

i.e., γ can be estimated by simulating a random variable with a different density and then unbiasing the output (I_n) by multiplying by the likelihood ratio. Sampling with a different density is sometimes called a change of measure and the density p' is sometimes called the importance sampling density (or if $p'(x) = dP'(x)/dx$, P' is called the importance sampling distribution).

Since essentially any density p' can be used for sampling, what is the optimal density, i.e., what is the density that minimizes the variance of $\hat{\gamma}_N(p')$? Selecting $p'(x) \equiv p^*(x) = p(x)/\gamma$ for $x \in A$ and $p^*(x) = 0$ otherwise has the property of making $Z_n = I_n p(X_n)/p^*(X_n) = \gamma$ with probability one. Since the variance of a constant is zero (and since the variance is always nonnegative), $p^*(x)$ is the optimal change of measure (and one sample from p^* gives you γ exactly). The optimal change of measure thus has the interpretation of being simply the ordinary distribution, conditioned that the rare event has occurred. However, there are several practical problems with trying to sample from this optimal density p^*. First, it explicitly depends upon γ, the unknown quantity that we are trying to estimate. If, in fact, γ were known, there would be no need to run the simulation experiment at all. Second, even if γ were known, it might be impractical to sample efficiently from p^*.

Since the optimal change of measure is not feasible, how should one go about choosing a good importance sampling change of measure? Since $E_{p'}[Z_n] = \gamma$ for any density p', reducing the variance of the estimator corresponds to selecting a density p' that reduces the second moment of Z_n;

$$E_{p'}[Z_n^2] = E_{p'}[(I_n L(X_n))^2] = \int 1_{\{x \in A\}} \left(\frac{p(x)}{p'(x)}\right)^2 p'(x)dx \tag{4}$$
$$= \int 1_{\{x \in A\}} \frac{p(x)}{p'(x)} p(x)dx = E_p[I_n L(X_n)].$$

Thus to reduce the variance, we want to make the likelihood ratio $p(x)/p'(x)$ small on the set A. Since A is a rare event (under density p), roughly speaking, $p(x)$ is small on A. Thus to make the likelihood ratio small on A, we should pick p' so that $p'(x)$ is large on A, i.e., the change of measure should be chosen so as to make the event A likely to occur.

More formally, consider a sequence of rare event problems indexed by a "rarity" parameter ϵ so that $\gamma(\epsilon) \to 0$ as $\epsilon \to 0$. For example, in the buffer sizing problem, we could let $\epsilon = 1/B$ and $\gamma(\epsilon) = P(Q > B)$ where B is the buffer size and Q denotes a random variable having the steady state queue length distribution. In a reliability model, as is done in[57, 72, 88, 89], we could parameterize the failure rates of components by ϵ, e.g., the failure rate of component number i is given by $\lambda_i(\epsilon) = a_i \epsilon^{b_i}$ for some constants a_i and $b_i (b_i \geq 1)$. Then, defining $\gamma(\epsilon)$ to be the probability that the system fails before some fixed time horizon t fits into this framework. Suppose it is known that $\gamma(\epsilon) \sim cf(\epsilon)$ as $\epsilon \to 0$ for some constant c and function $f(\epsilon)$. (Thus $f(\epsilon) \to 0$ as $\epsilon \to 0$.) Now if a p' can be chosen so that $E_{p'}[Z_n^2] \sim df(\epsilon)^2$ for some constant d, then $\sigma_{p'}(Z_n)$, the standard deviation of Z_n, $\sim kf(\epsilon)$ where $k = \sqrt{d - c^2}$. The relative error of the importance sampling estimate thus remains bounded as $\epsilon \to 0$:

$$\lim_{\epsilon \to 0} \frac{\sigma(\hat{\gamma}_N(p'))}{\gamma(\epsilon)} = \lim_{\epsilon \to 0} \frac{\sigma_{p'}(Z_n)}{\gamma(\epsilon)\sqrt{N}} = \frac{k}{\sqrt{N}} < \infty. \tag{5}$$

In such a situation, we say that the importance sampling estimate has "bounded relative error" (see, e.g., [57, 72, 88, 89]). In practice, having bounded relative error is highly desirable, since it implies that only a fixed, bounded, number of samples N are required to estimate $\gamma(\epsilon)$ to within a certain relative precision, *no matter how rare the event of interest is.* For example, suppose $k \leq 10$, then two significant digit accuracy, i.e., making the relative half-width of a 99% confidence for $\gamma(\epsilon)$ less than 0.1, can be achieved in at most $N = 6.64 \times 10^4$ samples, regardless of how small ϵ (and thus $\gamma(\epsilon)$) is. Compared to the sample sizes of 6.64×10^8 and 6.64×10^{11} required by standard simulation to estimate $\gamma(\epsilon) = 10^{-6}$ and 10^{-9}, respectively, we see that orders of magnitude reduction in run lengths, or speedup, could thus be achieved using importance sampling. However, the variance reduction does necessarily not tell the whole story, since the cost of obtaining each sample may be so high as to effectively limit the number of samples that can be collected. It is usually somewhat more expensive to

obtain a sample using importance sampling than standard simulation, however, the massive reduction in sample size generally more than makes up for the increase in the cost per sample. Such sampling costs are taken into account in a number of studies, e.g., [6, 86], and see [47] and the references therein for a theoretical treatment of this issue.

How can bounded relative error estimates be obtained? One way is to select p' so as to ensure that the likelihood ratio is always small on A. More specifically, suppose

$$L(X_n) \leq d_1 f(\epsilon) \tag{6}$$

whenever $X_n \in A$, then

$$E_{p'}[Z_n^2] = E_{p'}[I_n L(X_n)^2] \leq d_1^2 f(\epsilon)^2 E_{p'}[I_n] \leq d_1^2 f(\epsilon)^2. \tag{7}$$

Thus, sampling under such a p' produces an estimate with bounded relative error. How likely is the event A to occur under such a p'? The probability of A is given by

$$\begin{aligned} E_{p'}[I_n] &= \int 1_{\{x \in A\}} p'(x) dx = \int 1_{\{x \in A\}} \frac{p'(x)}{p(x)} p(x) dx \\ &\geq \frac{1}{d_1 f(\epsilon)} \int 1_{\{x \in A\}} p(x) dx = \frac{\gamma(\epsilon)}{d_1 f(\epsilon)} \sim \frac{c}{d_1} > 0. \end{aligned} \tag{8}$$

(The inequality in Equation 8 is true by Inequality 6.) Thus, asymptotically, the probability of A under p' does not depend on ϵ, i.e., A is not a rare event under p', and the problem has been transformed from a rare event simulation into a non-rare event simulation.

Note that bounded relative error is obtained by making the second moment, $E_{p'}[Z_n^2]$, approach zero at the same at the same rate, $f(\epsilon)^2$, that $\gamma(\epsilon)^2$ approaches zero. Is it possible to do any better? The answer is essentially no, since suppose $E_{p'}[Z_n^2] \sim d g(\epsilon)^2$ where $g(\epsilon) \to 0$ as $\epsilon \to 0$. Because importance sampling is unbiased and the variance is non-negative, we must have $E_{p'}[Z_n^2] \geq E_{p'}[Z_n]^2 = \gamma(\epsilon)^2$. Thus, $\underline{\lim}_{\epsilon \to 0} g(\epsilon)/f(\epsilon) \geq c/\sqrt{d}$, i.e., $g(\epsilon)$ can approach zero no faster than a constant times $f(\epsilon)$. When in fact, $g(\epsilon) \sim k f(\epsilon)$, some authors (e.g., [16, 17, 21, 26, 36, 85, 87, 95]) term the importance sampling scheme "asymptotically optimal" or "asymptotically efficient." These terms are thus seen to be equivalent to bounded relative error. (As seen from Equation 5, it is possible to obtain a relative error of zero as $\epsilon \to 0$ if one is lucky enough to have $d = c^2$, but, in practice, such situations are the exception rather than the rule.) Note that bounded relative error is also obtained if $\gamma(\epsilon) \geq c f(\epsilon)$ and if $L(X_n) \leq d_1 f(\epsilon)$ for $X_n \in A$. Also, in some applications, it may not be possible to find a constant d_1 such that $L(X_n) \leq d_1 f(\epsilon)$ but it is possible to find a random variable D such that $L(X_n) \leq D f(\epsilon)$ (on A). If $\overline{\lim}_{\epsilon \to 0} E_{p'}[D^2] < \infty$, then bounded relative error is still obtained. Juneja [61] shows that in order to obtain bounded likelihood ratios, one must make the likelihood ratio equal to one on every cycle of states. Although our focus is on asymptotic optimality, it is important to recognize that, for a fixed ϵ, there may be other changes of measure that, because of the (unknown) constants in the variance, result in a lower

variance than that obtained by the asymptotically optimal change of measure; indeed the zero variance estimator is always better, but as mentioned above, it is usually impractical to sample from that distribution. In addition, there may be several different formulations for estimating the same quantity; asymptotic optimality refers only to a particular formulation of a problem. (For example, see Section 3.1 for several formulations to estimate large waiting times in the single server queue.)

As used in this context, the purpose of importance sampling is to reduce the variance of an estimator. Thus importance sampling is called a "variance reduction technique." (See [4, 46, 56, 84] for a different application of importance sampling; so-called "what if" simulations in which importance sampling is used to estimate the performance of a system at many different input parameter settings from a single simulation run.) Does importance sampling always lead to a reduction in variance? The answer is (an emphatic) *NO*, as the following simple example illustrates. Suppose $p(x) = \lambda e^{-\lambda x}$ and $A = \{x > t\}$, i.e., we are interested in estimating $\gamma = P_p(X > t)$ where X has an exponential distribution with rate λ. Suppose we employ importance sampling using an exponential distribution with rate λ', i.e., $p'(x) = \lambda' e^{-\lambda' x}$. From Equation 4, the second moment of the estimator is given by

$$E_{p'}[Z_n^2] = \int_{x=t}^{\infty} \frac{p(x)}{p'(x)} p(x) dx = \frac{\lambda^2}{\lambda'} \int_{x=t}^{\infty} e^{(\lambda' - 2\lambda)x} dx \tag{9}$$

which is infinite if $\lambda' \geq 2\lambda$. Thus, not only can importance sampling result in a variance increase, but it can produce arbitrarily bad results if not applied carefully. (Conditions, and counter-examples, under which importance sampling of discrete time Markov chains over a random time horizon has a finite variance are given in [44, 46, 53].) So, essentially all work on using importance sampling in practical applications deals with choosing an importance sampling distribution (change of measure) that leads to actual variance reduction, with particular emphasis being placed on finding asymptotically optimal changes of measure. As we will see, selecting such asymptotically optimal changes of measure generally involves understanding quite a bit about the structure of the system being simulated. Fortunately, the understanding does not have to so thorough as to imply a complete solution of the problem; there is a rather large class of queueing and reliability models for which exact solutions do not exist (or numerical solutions are impractical) but for which asymptotically optimal importance sampling distributions can be explicitly given.

Note that importance sampling is well suited for parallelization by running independent replications (or replications with different input parameter settings) in parallel on multiple computers. (See [45, 55] and the references therein for a discussion of the statistical properties and efficiencies of a variety of parallel replications approaches.) If the cost of computing each sample is high, and the variance is relatively large, then such parallelization can be very effective; this is especially true if the asymptotically optimal importance sampling distribution is unknown. Several such examples, including the analysis of fault-tolerant routing

algorithms on a hypercube, where parallelization is effective are considered in [77].

2.3 Likelihood Ratios

The discussion of the previous section dealt with a very special situation: using importance sampling for a single random variable that has a probability density function. However, importance sampling is true much more generally as we will describe briefly (without getting too technical). In particular suppose $\gamma = \int H(\omega)dP(\omega)$ for some arbitrary random variable H and arbitrary probability measure P. Then, as in Equation 2, $\gamma = \int H(\omega)L(\omega)dP'(\omega) = E_{P'}[HL]$ for another probability measure P' where $L(\omega) = dP/dP'(\omega)$ is again the likelihood ratio or, in measure-theoretic terms, the Radon-Nikodym derivative. (This requires that the measure P be "absolutely continuous" with respect to the measure P'.) The reason for this brief excursion into measure theory is to make the point that the probabilistic setting in which importance sampling applies is extremely general. For example, it can involve stochastic processes such as discrete or continuous time Markov chains, semi-Markov process, or more general stochastic processes such as generalized semi-Markov processes (see [43, 99] and the references therein). The (rare) event A can also be defined quite generally; typically it is defined in terms of "stopping times" [24]. For example, in a reliability model defined by a continuous time Markov chain (CTMC), suppose τ_F is the first time that the process enters a "bad" set of states F in which the system is considered unavailable. Then an event of interest (which is hopefully rare) is $A = \{\tau_F \leq t\}$ for some given value of t.

In order to apply importance sampling, it must be possible to compute the relevant likelihood ratio. We briefly give several examples; explicit formulas for a variety of stochastic processes are given in [46]. In the case of sampling from a single probability density function, as described above, the likelihood ratio $L(X) = p(X)/p'(X)$. This equation is also valid if X is drawn from a discrete distribution, i.e., if $P(X = a_i) = p(a_i), i = 1, \ldots, m$ and $P'(X = a_i) = p'(a_i), i = 1, \ldots, m$. We require that $p'(a_i) > 0$ if $p(a_i) > 0$, but note that we can have $p'(a_i) > 0$ even if $p(a_i) = 0$ since the likelihood ratio is zero in this case, i.e., no weight is given to an impossible (under p) sample path. Suppose $X = (X_1, \ldots, X_m)$ is a random vector, where X_i is drawn from density $p_i(x)$ and X_i is independent of $X_j (j \neq i)$. If, under importance sampling, X_i is drawn from density $p_i'(x)$, and again X_i is independent of $X_j (j \neq i)$, then

$$L(X) = \prod_{i=1}^{m} \frac{p(X_i)}{p'(X_i)}. \tag{10}$$

Suppose $\{X_i, i \geq 0\}$ is a discrete time Markov chain (DTMC), on the state space of nonnegative integers where X_0 has (the initial) distribution $p_0(i)$ and the one-step transition probabilities are given by $P(i,j) = P(X_n = j | X_{n-1} = i)$. Let $\mathbf{X}_m = (X_0, X_1, \ldots, X_m)$. If, under importance sampling, X_0 is drawn from $p_0'(i)$, and the process is generated with the one-step transition probabilities

$P'(i,j)$, then

$$L(\mathbf{X}_m) = \frac{p_0(X_0)}{p'_0(X_0)} \prod_{i=1}^{m} \frac{P(X_{i-1}, X_i)}{P'(X_{i-1}, X_i)}. \tag{11}$$

We require that $p'_0(i) > 0$ if $p_0(i) > 0$ and $P'(i,j) > 0$ if $P(i,j) > 0$. In fact there is no requirement that the importance sampling distribution correspond to a time homogeneous DTMC; see [30, 50, 51] for examples in which it is advantageous to use a process other than a time homogeneous DTMC for importance sampling to estimate quantities associated with a DTMC. Identities for the variance of $L(\mathbf{X}_m)$ and $L(\mathbf{X}_m)Y_m$ where Y_m is a function of $\mathbf{X}_m$ may be found in [44, 46]; the relationship of such identities to the problems of rare event simulation covered here is beyond the scope of this paper.

For a CTMC, let $p_0(i)$ denote the initial distribution, $\lambda(i)$ denote the total rate out of state i and let $P(i,j) = Q(i,j)/\lambda(i)$ denote the transition probabilities of the embedded DTMC, where $Q(i,j)$ is the rate from state i to state j. If importance sampling is done using a CTMC with initial distribution $p'(i)$, holding rates $\lambda'(i)$, and embedded DTMC transition probabilities $P'(i,j)$, then the likelihood ratio after m transitions is

$$L(X) = \frac{p_0(X_0)}{p'_0(X_0)} \prod_{i=1}^{m} \frac{\lambda(X_{i-1})e^{-\lambda(X_{i-1})t_{i-1}}}{\lambda'(X_{i-1})e^{-\lambda'(X_{i-1})t_{i-1}}} \frac{P(X_{i-1}, X_i)}{P'(X_{i-1}, X_i)} \tag{12}$$

where the sequence of states of the embedded DTMC is $X_0, \ldots, X_m$ and t_i is the holding time in state X_i. (Note that this expression simplifies somewhat since $\lambda(X_{i-1})P(X_{i-1}, X_i) = Q(X_{i-1}, X_i)$.)

Another example, that will be of use in reliability models, concerns using uniformization, or thinning, [60, 71, 93] to sample points from a nonhomogeneous Poisson process (NHPP). Let $\{N(s), s \geq 0\}$ denote a NHPP with intensity rate $\lambda(s)$. Suppose there exists a finite constant β such that $\lambda(s) \leq \beta$ for all $s \geq 0$. Then points in the NHPP can be simulated (without importance sampling) as follows. Let $\{S_n, n \geq 1\}$ denote the points in an ordinary Poisson process $\{N_\beta(s), s \geq 0\}$ with (constant) rate β. Then, S_n is accepted as a point in the NHPP with probability $\lambda(S_n)/\beta$, otherwise it is rejected as a "pseudo event" (with probability $1 - \lambda(S_n)/\beta$). To implement importance sampling, we could simply thin the Poisson process $\{N_\beta(s), s \geq 0\}$ with different probabilities, in effect generating a NHPP with a different intensity rate, say $\lambda'(s)(\lambda'(s) \leq \beta)$. Let $N(t)$ denote the number of accepted (real) events and $P(t)$ denote the number of rejected (pseudo) events; then $N_\beta(t) = N(t) + P(t)$. Let T_n denote the time of the n-th real event ($n \leq N(t)$) and let P_n denote the time of the n-th pseudo event ($n \leq P(t)$). Then, the likelihood ratio (at time t) has a simple form:

$$L(t) = \prod_{n=1}^{N(t)} \frac{\lambda(T_n)}{\lambda'(T_n)} \times \prod_{n=1}^{P(t)} \frac{1 - \lambda(P_n)/\beta}{1 - \lambda'(P_n)/\beta}. \tag{13}$$

This requires that $\lambda'(s) > 0$ whenever $\lambda(s) > 0$ and $1 - \lambda'(s)/\beta > 0$ whenever $1 - \lambda(s)/\beta > 0$. In the reliability context, $\lambda(s)$ might represent the total component

failure rate at time s which, because components are reliable, is very small. Therefore, to see system failure events, we need to accelerate the rate at which components fail, which can be accomplished simply by increasing the component failure rate, i.e., by sampling using failure rate $\lambda'(s)$ where $\lambda'(s) \gg \lambda(s)$.

2.4 Regenerative Simulation

In simulations of stochastic systems, one is often interested in steady state performance measures. For example, in queueing models one might be interested in $P(W > x)$ where W is the steady state waiting time distribution. In a reliability model, one might be interested in u the steady state unavailability of the system, i.e., the long run fraction of the time that system is in a state that is considered failed. In both these cases, one is interested in estimating a quantity associated with a rare event. If the system is regenerative [27, 96], then the "regenerative method" can be used to estimate steady state performance measures. Let X_s be the process at time s. We assume there is a particular state, call it 0, such that the process returns to state 0 infinitely often and that, upon hitting state 0, the stochastic evolution of the system is independent of the past and has the same distribution as if the process were started in state in 0. Arrivals to an empty GI/GI/1 queue constitute regeneration points, as do entrances to a fixed state in a CTMC. Let β_i denote the time of the i-th regeneration ($\beta_0 = 0$) and assume that $X_0 = 0$. Let $\alpha_i = \beta_i - \beta_{i-1}$ denote the length of the i-th regenerative cycle. If $E[\alpha_i] < \infty$, then (under certain regularity conditions) $X_s \Rightarrow X$ as $s \to \infty$ where $\Rightarrow$ denotes convergence in distribution and X has the steady state distribution. Let h be a function on the state space and define $Y_i = \int_{\beta_{i-1}}^{\beta_i} h(X_s)ds$. Then $\{(Y_i, \alpha_i), i \geq 1\}$ are i.i.d. and

$$r = E[h(X)] = \frac{E[Y_i]}{E[\alpha_i]} \tag{14}$$

(provided $E[Y_i]$ exists and is finite). Equation 14 and the i.i.d. structure of regenerative processes forms the basis of the regenerative method; simulate N cycles and estimate $E[h(X)]$ by $\hat{r}_N = \bar{Y}_N/\bar{\alpha}_N$ where $\bar{Y}_N$ and $\bar{\alpha}_N$ are the averages of the Y_i-s and α_i-s, respectively. It is possible to form confidence intervals for r by applying the central limit theorem. (See [27] for a discussion of estimating the variance in this central limit theorem.) For example, to estimate $P(W > x)$, the process is in discrete time, $h(w) = 1_{\{w>x\}}$, α_i is the number of customers to arrive in a busy period and $Y_i = \sum_{k=\beta_{i-1}}^{\beta_i} h(W_k)$. To estimate u, the steady state unavailability, $h(x) = 1_{\{x \in F\}}$ where F is the set of failed states. In both of these cases, $Y_i = 0$ with high probability so obtaining a non-zero Y_i is a rare event. Thus importance sampling can be applied to estimate the numerator, $E[Y_i]$, of Equation 14. Typically importance sampling is used until a non-zero value of Y_i occurs, and then it is "turned off" and the process is allowed to return naturally to the regenerative state. Note that the denominator, $E[\alpha_i]$ is simply the expected cycle time, so importance sampling need not be applied to estimating the denominator. Letting L_i denote the likelihood ratio of the

process over cycle i, then $E[Y_i] = E_{P'}[L_i Y_i]$ where importance sampling is done using distribution P'. (See [11, 42, 46] for an alternative expression that uses partial likelihood ratios in regenerative simulation.)

To obtain variance reduction in estimating $E[Y_i]$, roughly speaking, we want to chose P' so as make the rare event likely to occur during a cycle, thus making the likelihood ratio small. But suppose one is interested in estimating the expected time (starting in state 0) until the rare event occurs, e.g., estimating $E_0[\tau_F]$ where τ_F is the first time the reliability model enters a failed state (in F). Using importance sampling to accelerate the occurrence of τ_F will result in unusually small values of τ_F, thereby increasing the variance. However, this problem can be avoided by exploiting a ratio formula for $E_0[\tau_F]$:

$$E_0[\tau_F] = \frac{E_0[\min(\tau_0, \tau_F)]}{P_0(\tau_F < \tau_0)} \tag{15}$$

where τ_0 is the first time to return to state 0 (see [51, 64, 92]). Equation 15 is valid because the number of cycles until hitting τ_F before τ_0 has a geometric distribution with success probability $\gamma = P_0(\tau_F < \tau_0)$. Now the problem becomes one of estimating γ, which is a rare event problem. Thus quantities such as the mean time to failure and the mean time until buffer overflow can be estimated by using importance sampling to estimate γ and standard simulation to estimate $E_0[\min(\tau_0, \tau_F)]$. The geometric distribution and Equation 15 also form the basis for showing that $\tau_F / E_0[\tau_F] \Rightarrow E$ as $\gamma \to 0$ where E is an exponential random variable with mean one [15, 64].

In some applications, the model may not be regenerative (e.g., a reliability model with non-exponential failure time distributions), yet a ratio formula similar to that of Equation 14 exists. For a subset, A, of the state space define A-cycles to begin whenever the process enters A. Then

$$r = E[h(X)] = \frac{E[Y_i(A)]}{E[\alpha_i(A)]} \tag{16}$$

where $Y_i(A)$ is the integral of the process over an A-cycle and $\alpha_i(A)$ is the length of an A-cycle [14, 25]. In Equation 16, the initial distribution is the stationary distribution conditioned on the process just entering A. Importance sampling can be used to estimate the numerator, while standard simulation can be used to estimate the denominator. A "splitting" technique can be used to obtain the proper initial distributions as follows. Simulate the process (without importance sampling) until it is approximately in steady state. Then whenever the process enters A, simulate two A-cycles; one using importance sampling and one without using importance sampling. These provide samples for the ratio estimate. Also, the A-cycle simulated without importance sampling provides a starting point (with approximately the steady state distribution on A) for the next pair of samples. The method of batch means can be used for variance estimation; see [21, 81] for a discussion of this approach, and see [1] for a similar idea applied to estimating bit error rates over certain communications channels.

3 Queueing Models

3.1 The Single Server Queue

We are now ready to apply importance sampling to some specific applications arising in queueing and reliability theory. We start with the waiting time process in the stable single server, GI/GI/1, queue. This queue has quite a lot of random walk related structure that can be exploited. Let W_n denote the waiting time of the n-th customer, $\{A_n\}$ denote the interarrival time sequence and $\{B_n\}$ denote the service time sequence. Then the waiting time sequence follows the well known Lindley's recursion (see, e.g., [7, 33]): $W_0 = 0$ and $W_{n+1} = (W_n + X_{n+1})^+$ for $n \geq 0$ where $X_{n+1} = B_n - A_{n+1}$ and $x^+ = \max(0, x)$. Let $A(x) = P(A_n \leq x)$ and $B(x) = P(B_n \leq x)$. Then, if $E[X_n] < 0$ (or equivalently $\rho = E[B_n]/E[A_n] < 1$), then the queue is stable and $W_n \Rightarrow W$ where W denotes the steady state waiting time. We will be interested in estimating $P(W > x)$ for large values of x. To do so we will exploit the equivalence between the distribution of W and the maximum of the (time-reversed) random walk. Define $\tilde{X}_1 = X_n, \tilde{X}_2 = X_{n-1}, \ldots, \tilde{X}_n = X_1$, $\tilde{S}_0 = 0$, and $\tilde{S}_k = \tilde{X}_1 + \cdots + \tilde{X}_k$ for $k \geq 1$. Then, it is well known that

$$W_n = \tilde{M}_n \equiv \max\{\tilde{S}_0, \tilde{S}_1, \ldots, \tilde{S}_n\} \qquad n \geq 0. \tag{17}$$

Thus, letting $n \to \infty$, W has the same distribution as $\tilde{M}$, the maximum of the time-reversed random walk (assuming the distribution of $\{X_{n-k}, 0 \leq k \leq n\}$ converges as $n \to \infty$). For the GI/GI/1 queue, $\tilde{X}_n$ has the same distribution as X_n, so we can say that W has the same distribution as $M = \max\{S_k, k \geq 0\}$, the maximum of the random walk (which has negative drift).

We note that $P(W > x) = P(M > x) = P(\tau_x < \infty)$ where τ_x is the first time the random walk exceeds x. Using standard simulation, estimation of $P(\tau_x < \infty)$ is not efficient since, with high probability $\tau_x = \infty$. (Not only is the event rare, but it takes a potentially long time to determine that the rare event did not happen.) However, we will use importance sampling to transform the random walk into one with positive drift, so that $\tau_x < \infty$ with probability one. We follow the queueing oriented development in Asmussen [6, 7]. (Siegmund [95] considers this and related problems in the context of sequential analysis and sequential probability ratio tests. Asmussen also discusses the relevance of this problem to "risk theory" in insurance applications.) We will use a specific importance sampling change of measure known as "exponential twisting," "exponential tilting," or embedding within a "conjugate family." Define the moment generating function $M(\theta) = E[e^{\theta X_n}]$ which we will assume is finite for all $0 \leq \theta < \bar{\theta}$ where $\bar{\theta} > 0$. (For technical reasons, we also assume that $M(\theta) \to \infty$ as $\theta \uparrow \bar{\theta}$.) Note that $M(\theta) = E[e^{\theta(B_{n-1} - A_n)}] = M_B(\theta) M_A(-\theta)$ where $M_B(\theta) = E[e^{\theta B_n}]$ and $M_A(\theta) = E[e^{\theta A_n}]$. Let $F(x)$ denote the distribution function of X_n. Now define the exponentially twisted distribution $F_\theta(x)$ by $dF_\theta(x) = [e^{\theta x}/M(\theta)]dF(x)$; if $F(x)$ has a density function $f(x)$, then $dF_\theta(x) = f_\theta(x) = e^{\theta x} f(x)/M(\theta)$. If $X_1, \ldots, X_n$ are independently sampled from $F_\theta(x)$, then the likelihood ratio has

a very simple form:

$$L_n(\theta) = \frac{M(\theta)^n}{e^{\theta(X_1+\ldots+X_n)}} = M(\theta)^n e^{-\theta S_n}. \tag{18}$$

Now if $\tau_x = n$, then by definition, $S_n > x$ and thus

$$L_n(\theta) = M(\theta)^n e^{-\theta x} e^{-\theta(S_n - x)} \leq M(\theta)^n e^{-\theta x}. \tag{19}$$

What value of θ should be chosen? Note that there exists a θ^* such that $M(\theta^*) = 1$. This is true since $M(0) = 1$, the derivative of $M(\theta)$ is negative at $\theta = 0$ (because $E[X_n] < 0$) and $M(\theta)$ is convex, continuous and approaches ∞ as $\theta \uparrow \bar{\theta}$.) By the above argument, at this value of θ^*, the titled random walk has strictly positive drift so $\tau_x < \infty$ with probability one. Thus, using Equations 2 and 19, we obtain

$$P(W > x) = P(\tau_x < \infty) = e^{-\theta^* x} E_{\theta^*}[e^{-\theta^* O_x}] \tag{20}$$

where O_x is the (random) "overshoot" $O_x = S_{\tau_x} - x$. Note that the decay rate, $-\theta^*$, is known; only $E_{\theta^*}[e^{-\theta^* O_x}]$ is unknown and requires estimation. Why is θ^* a good value with which to do importance sampling? It can be shown that, under suitable regularity conditions, $E_{\theta^*}[e^{-\theta^* O_x}]$ converges to a constant, say $a(\theta^*)$, as $x \to \infty$, thus $P(W > x) \sim a(\theta^*)e^{-\theta^* x}$. (This provides a very simple proof that the stationary waiting time distribution in the GI/GI/1 queue has an exponentially decaying tail, provided the moment generating function is finite.) In addition, by Equation 19, the likelihood ratio is also bounded by $e^{-\theta^* x}$. Thus, by the arguments of Section 2.2, importance sampling with this value of θ^* is asymptotically optimal. (Identify $\epsilon = 1/x$ and $f(\epsilon) = e^{-\theta^*/\epsilon}$.) In fact Siegmund [95] showed that θ^* is the unique asymptotically optimal value within the class of exponentially twisted distributions, while Lehtonen and Nyrhinen [68] showed that it is the unique asymptotically optimal change of measure within the class of (essentially) all distributions with i.i.d. increments. Asmussen [9] also suggests using the identity in Equation 20 for $P(W > x)$ to estimate $E[W] = \int P(W > x)dx$ using importance sampling (and quantities such as higher moments), although it needs to be combined with additional techniques in order to be effective.

For GI/GI/1, the equation $M(\theta^*) = 1$ is equivalent to $M_A(-\theta^*)M_B(\theta^*) = 1$. For the M/M/1 queue with arrival rate λ and service rate μ, $M_A(-\theta) = \lambda/(\lambda+\theta)$ and $M_B(\theta) = \mu/(\mu-\theta)$ for $\theta < \mu$. Solving for θ^* yields $\theta^* = \mu - \lambda$. At this value of θ^*, the interarrival time density $a_{\theta^*}(x) = \mu e^{-\mu x}$ and the service time density $b_{\theta^*}(x) = \lambda e^{-\lambda x}$, i.e., the process is simulated with arrival rate μ and service rate λ. Thus, with the asymptotically optimal importance sampling distribution, the simulated queue is indeed unstable. This flipping of arrival and service rates will also occur in other rare event problems as will be discussed later.

Further interpretation as to why exponential twisting with parameter θ^* is a good idea can be found in [2, 5]. Suppose one is told that $W_n > na$ for some constant a and large value of n. Then how did the queue come to be in

such a (bad) state? Let $\mu^* > 0$ denote the mean of the tilted random walk, i.e., the random $S_k(\theta^*)$ with increment distribution $F_{\theta^*}(x)$. The titled random walk builds up linearly at rate μ^*, i.e., by the strong law of large numbers, $S_{[nt]}(\theta^*)/n \to t\mu^*$ a.s. as $n \to \infty$ for $0 \leq t \leq 1$. Now suppose $a \leq \mu^*$. Then, the titled random walk is capable of reaching the level na by time n. In fact, the tilted random walk reaches level na at approximately time na/μ^*. Define $t(a) = 1 - a/\mu^*$. Anantharam [2] shows that in this case as $n \to \infty$, given $W_n > na$, with probability one, $W_{[nt]}/n \to 0$ for all $t \leq (1 - t(a))$, and then $W_{[nt]}/n \to \mu^*(t - t(a))$ for all t such that $t(a) \leq t \leq 1$. In other words, W_k remains stable until a certain time, $t(a)n$, and then starts building up linearly at the same rate that the titled random walk builds up until reaching level na at time n. (If $a > \mu^*$, then the tilted random walk with parameter θ^* cannot reach level na by time n. In this case, the waiting times build up linearly at rate a, corresponding to exponential twisting with a parameter θ_a such that the drift at θ_a is a.) In fact, related conditional limit theorems for the GI/GI/1 queue (see [5] and the references therein) show that during such a buildup, the waiting time behavior is (asymptotically) identical to that of the tilted random walk. Recall that in Section 2.2 we stated that the zero variance estimator was obtained by sampling from the ordinary distribution given that the rare event has occurred. The above results show, roughly speaking, that given a large waiting time has occurred, the ordinary distribution *is* the twisted distribution. Thus it is very natural that exponential twisting with parameter θ^* should be very effective. Indeed, experimental results in [68], show that for the M/M/1 queue (a problem not requiring simulation but convenient as a simulation benchmark) with $\rho = 0.5$, $P(W > 20) \approx 10^{-9}$ can be estimated to within $\pm 5\%$ in only 129 replications. In fact, if a non-optimal value of $\theta \neq \theta^*$ is used, then good estimates are obtained within reasonable time even if θ is as much as 30% away from θ^*.

(Interestingly, if $M_B(\theta) = \infty$ for all $\theta > 0$, it is shown in [2] that large waiting times are essentially the result of a single customer experiencing a very large service time. The assumption that $M_B(\theta) < \infty$ for some $\theta > 0$ implies that the tail of the distribution $B()$ goes to zero exponentially fast. Thus if $M_B(\theta) = \infty$ for all $\theta > 0$, the tail of the distribution is "fat" and very large values are not that unlikely.)

The above analysis makes use of the fact that, in the GI/GI/1 queue, the stationary waiting time has the same distribution as the maximum of a random walk with negative drift. This relationship does not occur more generally, e.g., in multiserver queues, queues with finite buffers, or networks of queues. Therefore it is desirable to consider techniques that work directly with the queueing processes. The first such problem we consider is estimating the mean time until a GI/GI/1 queue reaches a queue length of n (not including the customer in service). Recalling the ratio formula of Equation 14, we see that we need to estimate $\gamma_n = P(\tau_n < \tau_0)$ where τ_n is the first time the queue length reaches n and τ_0 is the first time the queue empties. (This problem is related to determining the distribution of the maximum waiting time during a busy period

in the GI/G/1 queue, see [59].) We assume that the initial conditions are that a customer (customer number 0) has just arrived to an empty queue. We basically follow the analysis in Sadowsky [85]. This analysis provides a rigorous justification for a heuristic developed by Parekh and Walrand [83], which will be described in Section 3.2. Let $S(k) = \{\tau_n = k, \tau_0 > k\}$. On $S(k)$, the arrival time of customer number k customer is less than the departure time of the $k-n$-th customer to depart the queue (customer number $k-n-1$), i.e.,

$$\sum_{j=1}^{k} A_j < \sum_{j=1}^{k-n} B_{j-1}. \tag{21}$$

For $k \geq n$ define $Z_k(n) = \sum_{j=1}^{k} A_j - \sum_{j=1}^{k-n} B_{j-1}$. Consider using exponential twisting with parameter θ^*. Then, on $S(k)$, $Z_k(n) \leq 0$ by Equation 21 and the likelihood ratio $L(\theta^*)$ is bounded as follows:

$$L(\theta^*) = M_B(\theta^*)^{k-n} M_A(-\theta^*)^k e^{\theta^* Z_k(n)} \leq M_B(\theta^*)^{-n} = M_A(-\theta^*)^n \tag{22}$$

since $Z_k(n) \leq 0$ and the fact that θ^* satisfies $M_A(-\theta^*)M_B(\theta^*) = 1$. Since the likelihood ratio is bounded by $M_A(-\theta^*)^n$ whenever $\tau_n < \tau_0$, taking expectations we have $\gamma_n \leq M_A(-\theta^*)^n$. Asymptotic optimality will be established if we can show that $\gamma_n \geq cM_A(-\theta^*)^n$ for some constant c. To do so, we will make the simplifying assumption that the service times are bounded, i.e., $B_j \leq b$ for all j and some constant b. On $S(k)$,

$$Z_k(n) = Z_{k-1}(n) + A_k - B_{k-n-1} \geq Z_{k-1}(n) + A_k - b \geq -b \tag{23}$$

where the second inequality is true because, on $S_k(n)$, $Z_{k-1}(n) \geq 0$. Combining the equality part of Equation 22 with Inequality 23 we obtain

$$L(\theta^*) \geq M_A(-\theta^*)^n e^{-\theta^* b} \tag{24}$$

whenever $\tau_n > \tau_0$. Taking expectations yields

$$\gamma_n = E_{\theta^*}[L(\theta^*) 1_{\{\tau_n < \tau_0\}}] \geq M_A(-\theta^*)^n e^{-\theta^* b} P_{\theta^*}(\tau_n < \tau_0). \tag{25}$$

But since, at θ^*, the queueing process is unstable, $P_{\theta^*}(\tau_n < \tau_0)$ is bounded away from 0 as $n \to \infty$, thereby establishing the result.

Note that $\lim_{n\to\infty}(1/n)\log(\gamma_n) = \log(M_A(-\theta^*))$. In addition, it is true that $\lim_{n\to\infty}(1/n)\log V_n(\theta^*) = 2\log(M_A(-\theta^*))$ where $V_n(\theta^*)$ is the variance of a single observation using importance sampling with parameter θ^*. For queueing models, asymptotic optimality is often stated in these terms.

Sadowsky shows that the bounded service time assumption can be removed. He furthermore shows that exponential twisting with parameter θ^* is the unique asymptotically optimal change of measure within the class of all simulations having i.i.d. interarrival and service time distributions. Further optimality results, concerning higher moments of the estimate (not just the variance), are established in [86].

3.2 A Large Deviations Approach to the Single Server Queue

The above results are in fact closely related to the theory of large deviations [16], which basically deals with estimating the probability that S_n/n is far from its mean. We will give an overly simplified, and somewhat heuristic, description of certain basic results and show how they can be applied to fast simulation of certain queues; see [16] for a rigorous (and readable) treatment of this topic in much greater detail. (There are a number of technical issues and difficulties which tend to obscure what's going on, so we will not deal with them here.) Consider a random walk, S_n, whose increments, X_k, have a distribution function F, finite moment generating function $M(\theta)$, and twisted distribution function $dF_\theta(x) = e^{\theta x}dF(x)/M(\theta)$ as described earlier. Cramér's theorem, which dates to 1937, roughly states that

$$P(S_n/n \approx y) \approx e^{-nI(y)} \tag{26}$$

where $I(y)$ is the Cramér transform (or large deviation rate function):

$$I(y) = \sup_\theta[\theta y - \log(M(\theta))]. \tag{27}$$

(More precisely, by Equation 26 we mean $\lim_{\delta\to 0}\lim_{n\to\infty}(1/n)\log[P(|S_n/n - y| < \delta)] = -I(y)$ provided $I(y)$ is finite and continuous in a neighborhood about y.) At first glance this result looks somewhat mysterious, but the intuition behind it is actually quite simple. Pick a small positive δ and consider estimating $P(\mathcal{S}(n,y)) = P(y-\delta \le S_n/n \le y+\delta)$ by simulating the random walk with exponential twisting. If $y \neq E[X_k]$, then $\mathcal{S}(n,y)$ is a rare event. It is natural to chose the twisting parameter $\theta = \theta_y$ so that $P_\theta(\mathcal{S}(n,y)) \to 1$, i.e., select θ_y so that $y = E_{\theta_y}[X_k] = \int x dF_{\theta_y}(x)$. Then

$$\begin{aligned} P(\mathcal{S}(n,y)) &= E_{\theta_y}[L(\theta_y)1_{\{\mathcal{S}(n,y)\}}] = M(\theta_y)^n E_{\theta_y}[e^{-\theta_y S_n}1_{\{\mathcal{S}(n,y)\}}] \qquad (28)\\ &\approx M(\theta_y)^n e^{-\theta_y n y} E_{\theta_y}[1_{\{\mathcal{S}(n,y)\}}] = e^{-n[\theta_y y - \log(M(\theta_y))]} \end{aligned}$$

where the $\approx$ is (approximately) true because $S_n \approx ny$ on $\mathcal{S}(n,y)$. It thus remains to be shown that $I(y) = \theta_y y - \log(M(\theta_y)$ which is established by setting the derivative of the function $h(\theta) = \theta y - \log(M(\theta))$ equal to zero. (It can be shown that $I(y)$ is convex and nonnegative on its domain of finiteness.) Note that if $y = E[X_k](= E_0[X_k])$, then, by definition $\theta_y = 0$, $I(y) = 0$, and thus Equation 26 is equivalent to $P(S_n/n \approx E[X_k]) \approx 1$ as it should be.

Let us see how this result can be used (heuristically) in the GI/GI/1 queue. Following Parekh and Walrand [83], again consider estimating $\gamma_n = P(\tau_n < \tau_0)$ where τ_n is the first time the queue length reaches n and τ_0 is the first time the queue empties. Suppose this event happens at some time T and that during the interval $(0,T)$, the arrival rate is some constant λ' and the service rate is some constant μ'. Thus the number of arrivals is approximately $\lambda' T$, the number of

departures is approximately $\mu' T$ and the queue builds up at rate $\lambda' - \mu'$. Now, by Equation 26, the probability of observing such rates λ' and μ' is approximately

$$p(\lambda', \mu') = \exp\left[-\lambda' T I_A(1/\lambda') - \mu' T I_B(1/\mu')\right] \tag{29}$$

where $I_A()$ and $I_B()$ are the Cramér transforms of the interarrival and service time distributions respectively. The values of λ' and μ' should be chosen so as to maximize the probability $p(\lambda', \mu')$ of this event. Since $\tau_n = T$, we have the (approximate) constraint

$$(\lambda' - \mu')T = n. \tag{30}$$

Substituting this into Equation 29, differentiating the exponent with respect to λ' and μ', setting the derivatives equal to 0 and solving yields the result that, as before, the optimal λ^* and μ^* correspond to twisting with rates $-\theta^*$ and θ^* such that $M_A(-\theta^*)M_B(\theta^*) = 1$. An extension of this heuristic approach to certain networks of queues was also described in [83] and will be discussed in Section 3.5.

Parekh and Walrand also relate this approach to the asymptotically optimal simulation of slow Markov walks as developed by Cottrell, Fort and Malgouyres [26]. These results are based on the Wentzell-Freidlin theory of large deviations for appropriately scaled diffusion processes (see [16] for a discussion), the details of which are beyond the scope of this paper. The basic model is

$$X^\epsilon_{n+1} = X^\epsilon_n + \epsilon V_n(X^\epsilon_n) \tag{31}$$

for some small ϵ (and $X^\epsilon_0 = 0$). Thus $\{X^\epsilon_n, n \geq 0\}$ is a Markov chain (with a general state space) in discrete time with small increments. (The distribution of the increments can depend on the current value of chain.) Let $F^x(y) = P(V_n \leq y | X^\epsilon_n = x)$ and assume that $M^x(\theta) = \int e^{\theta y} dF^x(y) < \infty$. Under appropriate technical conditions, there is a large deviations result, similar in spirit to Equation 26, but in which the decay rate $I(y)$ is replaced by integral of an "action functional." Define the twisted increment distribution $dF^x_\theta(y) = e^{\theta y} dF^x(y)/M^x(\theta)$. Assume the process is one-dimensional and let $\gamma(\epsilon) = P(X^\epsilon_n$ hits 1 before 0). In [26], it is shown that the asymptotically optimal method for estimating $\gamma(\epsilon)$ is obtained by using exponential twisting with parameter θ_x whenever $X^\epsilon_n = x$ where $M^x(\theta_x) = 1$. Cottrell, Fort and Malgouyres apply their method to certain Aloha systems. Parekh and Walrand discuss how it can be applied to the M/M/1 queue, but point out difficulties in trying to use this approach for networks of queues. The application of this approach to certain other simple simple single server queues (e.g., M/D/1) is considered in [37].

3.3 The Multiple Server Queue

Sadowsky [85] also obtains asymptotic optimality results for the multiple server GI/GI/m queue ($m > 1$). In this case, the probability of interest is $\gamma_n = P$(queue length exceeds n during a "cycle") where a cycle is defined to start whenever a customer arrives to find all but one of the servers busy. (The initial

distribution is assumed to be the stationary distribution at such instances.) Under similar conditions to those described in Section 3.1, it is shown that the unique asymptotically optimal change of measure is given by

$$dA_{\theta^*}(x) = \frac{e^{-m\theta^* x} dA(x)}{M_A(-m\theta^*)}, \qquad dB_{\theta^*}(x) = \frac{e^{\theta^* x} dB(x)}{M_B(\theta^*)} \tag{32}$$

where θ^* satisfies $M_A(-m\theta^*)M_B(\theta^*) = 1$. Some experimental results are given in [85]; in one example with $m = 4$ servers and $n = 20$, γ_n, which is approximately 2.2×10^{-8}, could be estimated to within $\pm 10\%$ accuracy with only 602 samples using asymptotically optimal importance sampling, whereas over 10^9 samples would be required using standard simulation.

3.4 Discrete Time Queues With Correlated Arrival Processes

We next turn to analysis of queueing systems that arise in ATM (Asynchronous Transfer Mode) communications systems. In such systems, it is important to model bursty and/or correlated arrival processes, such as those that might occur in transmitting video data streams across a network. We begin by considering a particularly simple model of such a queueing system. The model is in discrete time. Let a_t denote the number of packets that arrive during time slot t and let $A_T = a_0 + \cdots + a_t$ denote the total number of arrivals by time T. We assume that the server has the capacity to serve up to c packets/time slot. Letting Q_t denote the queue length at time t, then Q_t obeys Lindley's recursion $Q_{t+1} = (Q_t + a_{t+1} - c)^+$. For simplicity, we assume that the arrival process is Markovian, i.e., $P(a_{t+1} = j | a_t = i) = P(i,j)$, which we assume is aperiodic, irreducible and has a finite state space $(0 \le i, j \le b)$. We again consider $\gamma_n = P(\tau_n < \tau_0)$ given that the queue is empty at time 0. The appropriate change of measure can be inferred from large deviations results for Markov additive processes [16]. Define the matrix $A_\theta(i,j) = e^{\theta j} P(i,j)$. By the Perron-Frobenius theorem, there exists a real valued eigenvalue $\lambda(\theta)$ (the spectral radius) such that if λ is any other eigenvalue, then $|\lambda| < \lambda(\theta)$. Corresponding to $\lambda(\theta)$ is a positive (right) eigenvector $h(i,\theta)$ satisfying

$$\sum_{j=0}^{b} A_\theta(i,j) h(j,\theta) = \lambda(\theta) h(i,\theta). \tag{33}$$

From Equation 33, we see that

$$P_\theta(i,j) = \frac{A_\theta(i,j)h(j,\theta)}{\lambda(\theta)h(i,\theta)} = \frac{e^{\theta j} P(i,j) h(j,\theta)}{\lambda(\theta) h(i,\theta)} \tag{34}$$

defines the transition matrix of a Markov chain, which is called the conjugate process. Now apply importance sampling using the conjugate process. For simplicity assume the initial distribution using importance sampling is the same

as in the original system. Thus by Equation 11, the likelihood ratio after T transitions is given by

$$L(\theta, T) = \prod_{t=1}^{T} \frac{P(a_{t-1}, a_t)}{P_\theta(a_{t-1}, a_t)} = \lambda(\theta)^T e^{-\theta A_T} H(\theta, T) \tag{35}$$

where $H(\theta, T) = h(a_0, \theta)/h(a_T, \theta)$. Let $S(T)$ denote the event that $Q_T \geq n$ and the queue has not emptied before time T. Then $A_T \geq Tc + n$ on $S(T)$, i.e., as in Equations 19 and 21, $A_T = Tc + n + O_T$ where O_T is the (nonnegative) overshoot at time T. Also, on $S(T)$, the likelihood ratio is given by

$$L(\theta, T) = \lambda(\theta)^T e^{-\theta(Tc+n+O_T)} H(\theta, T) = e^{-\theta n - \theta O_T + T[\log(\lambda(\theta)) - \theta c]} H(\theta, T). \tag{36}$$

Which value of θ should be chosen? Similar to the GI/GI/1 case in which we set $M(\theta^*) = 1$, by selecting $\theta = \theta^*$ so that $\log(\lambda(\theta^*)) - \theta^* c = 0$ (assuming it exists), we obtain the simplification that

$$L(\theta^*, T) = e^{-\theta^* n - \theta^* O_T} H(\theta^*, T). \tag{37}$$

Because the state space of the arrival process is finite, $O_T \leq b$ and there exist positive finite constants $\underline{H}$ and $\overline{H}$ such that $\underline{H} \leq H(\theta^*, T) \leq \overline{H}$. Thus, on $S(T)$ we have

$$e^{-\theta^*(n-b)} \underline{H} \leq L(\theta^*, T) \leq e^{-\theta^* n} \overline{H}. \tag{38}$$

Equation 38 implies both that $\lim_{n\to\infty}(1/n)\log(\gamma_n) = -\theta^*$ and that simulating the conjugate process with $\theta = \theta^*$ is asymptotically optimal. Thus finding the asymptotically optimal change of measure involves solving the nonlinear equation

$$\frac{\log(\lambda(\theta^*))}{\theta^*} = c. \tag{39}$$

If the source actually has independent arrivals, i.e., if $P(i, j) = p(j)$ for all i, then the model reduces to the waiting time in the GI/D/1 queue. In this case it can be shown that $\lambda(\theta) = M_A(\theta)$ and that Equation 39 is equivalent to $M(\theta^*) = 1$.

These results extend to the case when there are multiple independent arrival sources feeding the same queue. Suppose now that we have K independent Markovian sources defined by transition matrices $P_k(i, j)$ for $1 \leq k \leq K$. Note that a numerical solution of this model is typically infeasible. As above, let $\lambda_k(\theta)$ denote the spectral radius of the matrix defined by $A_{k,\theta}(i, j) = e^{\theta j} P_k(i, j)$ and let $h_k(i, \theta)$ denote the corresponding eigenvectors. If each source is twisted by the same parameter θ, i.e.,

$$P_{k,\theta}(i, j) = \frac{A_{k,\theta}(i, j) h_k(j, \theta)}{\lambda_k(\theta) h_k(i, \theta)}, \tag{40}$$

then as in Equation 36, on $S(T)$ the likelihood ratio at time T is given by

$$L(\theta, T) = \exp\left[-\theta n - \theta O_T + T\left(\sum_{k=1}^{K} \log(\lambda_k(\theta)) - \theta c\right)\right] H(\theta, T) \tag{41}$$

where now $H(\theta, T)$ is the (bounded) term involving the product of the eigenvector ratios as defined above. Thus, simulating each process with parameter θ^* where

$$\sum_{k=1}^{K} \frac{\log(\lambda_k(\theta^*))}{\theta^*} = c \tag{42}$$

(assuming it exists) results in the likelihood ratio given by Equation 37. This again implies that $\lim_{n\to\infty}(1/n)\log(\gamma_n) = -\theta^*$ and yields an asymptotically optimal simulation scheme. Note that each source is still simulated independently, however the distributions of the sources are related by Equation 42.

More general types of arrival processes are also possible. For example, suppose source k is a Markov arrival process. In this model, there is an environment variable X_t^k and the distribution of arrivals is described by $P_k(a, j|i) = P(a_t^k = a, X_t^k = j|X_{t-1}^k = i)$. Let $\lambda_k(\theta)$ be the spectral radius of the matrix $A_{k,\theta}(i,j) = \sum_{a=0}^{b} e^{\theta a} P_k(a, j|i)$ and let $h_k(i, \theta)$ be the corresponding eigenvector. The twisted distribution is now given by

$$P_{k,\theta}(i,j) = P_\theta(a_t^k = a, X_t^k = j|X_{t-1}^k = i) = \frac{e^{\theta a} P_k(a, j|i) h_k(j, \theta)}{\lambda_k(\theta) h_k(i, \theta)}. \tag{43}$$

Again, simulating each arrival process with twisting parameter θ^* where θ^* satisfies Equation 42 is asymptotically optimal.

The analysis above follows that described Chang, Heidelberger, Juneja and Shahabuddin [21, 22], however there are a number of related results and papers [8, 16, 17, 69, 87] that deal with large deviations of Markov additive processes. For example, Asmussen [8] considers the M/G/1 queue with Markov modulated arrivals. Although not primarily a simulation paper, Asmussen suggests using exponential twisting on the time reversed arrival process along with the maximum representation (in this case for the virtual waiting time) as described in Section 3.1. Descriptions of large deviations and asymptotically optimal simulation results for Markov additive process may be found [16, 17, 87]. Here the problem is estimating $P(S_n/n > a)$ $(a \geq 0)$ where $S_n = f(X_0) + \ldots + f(X_n)$ and $\{X_n\}$ is a Markov chain such that $\lim_{n\to\infty} S_n/n < 0$. These papers show that the twisted distribution is the unique asymptotically optimal change of measure within the class of all time-homogeneous Markov chains. Lehtonen and Nyrhinen [69] show that this is also the unique asymptotically optimal change of measure for estimating the distribution of the maximum of a Markov additive process with negative drift. Conditional limit theorems also exist for Markov additive processes (see [28]) showing that given a large value of S_n/n has occurred, the process got there according to the same distribution as the conjugate process. Similar results for a fluid model of a queue with a Markovian fluid-type arrival process are obtained in [10]. This provides additional interpretation as to why exponential twisting is asymptotically optimal.

These results are closely related to the theory of "effective bandwidths" in telecommunications; see [41, 52, 66] for early papers in this area. A comprehensive treatment of effective bandwidths is given in Chang [20]; see also Whitt

[100] for additional results and references. The effective bandwidth (also called the minimum envelope rate in [20]) of a source is given by

$$a^*(\theta) = \lim_{T\to\infty} \frac{\log\left(E[e^{\theta A_T}]\right)}{\theta T} \tag{44}$$

(assuming the limit exists). The existence of this limit (plus some technical assumptions) is sufficient to show that the stationary queue length distribution has an exponentially decreasing tail with rate θ^* where $a^*(\theta^*) = c$, i.e., $\lim_{n\to\infty}(1/n)\log(P(Q > n)) = -\theta^*$ where Q has the stationary queue length distribution. (If the limit in Equation 44 does not exist, but the lim sup is given by $a^*(\theta)$, then the tail of the stationary queue length distribution decreases at least as fast as an exponential with rate $-\theta^*$.)

The relationship between these effective bandwidth results and fast simulation is explored in [21, 22]. Consider, for example, the Markovian arrival process in which $P(a_t = j|a_{t-1} = i) = P(i,j)$. Then

$$E[e^{\theta A_T}] = E_\theta[L(\theta,T)e^{\theta A_T}] = \lambda(\theta)^T E_\theta[H(\theta,T)] \tag{45}$$

where the second equality follows from Equation 35. Thus,

$$\lim_{T\to\infty} \frac{\log\left(E[e^{\theta A_T}]\right)}{\theta T} = \frac{\log(\lambda(\theta))}{\theta} = a^*(\theta). \tag{46}$$

Thus the effective bandwidth equation $a^*(\theta^*) = c$ is identical to the asymptotically optimal fast simulation Equation 39. The key idea is found in Equation 35 which implies that

$$\underline{H}(\theta) \le \frac{L(\theta,T)e^{\theta A_T}}{e^{\theta a^*(\theta)T}} \le \overline{H}(\theta). \tag{47}$$

More generally, if a family of processes (called envelope processes in [21, 22]) can be found such that the likelihood ratio with respect to the original distribution satisfies the likelihood ratio bounds of Inequalities 47, then Equation 44 holds, simulation with parameter θ^* such that $a^*(\theta^*) = c$ is asymptotically optimal for estimating γ_n, and $\lim_{n\to\infty}(1/n)\log(\gamma_n) = -\theta^*$. For multiple independent arrival processes, the coupling equation is the same as Equation 42, i.e., $a_1^*(\theta^*) + \ldots + a_K^*(\theta^*) = c$ and simulation with this value of θ^* is asymptotically optimal. This is true for essentially *any* type of arrival process (e.g., autoregressive), provided 47 is true, in which case the conditions for the Gärtner-Ellis theorem (see [16, 32, 38]) relating to large deviations results for A_T/T are basically satisfied, where A_T is the sum of random variables that may have a very general dependency structure. Thus, the Gärtner-Ellis theorem, effective bandwidths, and fast simulation are all closely related.

It is observed in [22] that envelope processes can be constructed by minimizing the relative entropy rate, or Kullback-Leibler distance (subject to a drift constraint); see also [16, 26] for applications of the Kullback-Leibler distance in large deviations, and [67] in which a related fast simulation heuristic

for on-off Markovian fluid sources is proposed. (This distance is defined as $K(P', P) = \int \log(dP'/dP)dP'$.) The splitting technique described in Section 2.4 can be used to estimate quantities such as the steady state buffer loss probability. Define an A-cycle to begin whenever the queue is empty. Note that the expected number of packets lost during an A-cycle is zero unless $\tau_n < \tau_0$ where n is the buffer size. Thus, (asymptotically optimal) importance sampling can be used to bring the queue up to level n at which point it can be turned off allowing the queue to empty again. This technique was shown to be very effective in [21, 22].

3.5 Networks of Queues

There are fewer papers on fast simulation of rare events in networks of queues, and the situation is currently not as well understood as in the single queue case. Rather than giving a detailed description of results for networks, a set of references will simply be given. Papers dealing with overflows in Jackson networks include [3, 34, 35, 36, 58, 83, 97]. In such networks, roughly speaking, a buildup to an overflow occurs along the same path that the corresponding time-reversed network (see [65]) empties from such an overflow, but in the opposite direction.

Large deviations results for the the way in which rare certain rare events happen in a class of multidimensional Markov jump processes are given in [94, 98]. Tree-structured networks of discrete time queues such as those considered in Section 3.4 have been studied in [21, 22].

Heuristics for simulating certain queueing networks that arise in high speed communications switches, such as a Clos switch, have been developed in Devetsikiotis and Townsend [29, 30, 31]. They consider a family of importance sampling change of measures parameterized by a (perhaps multidimensional) parameter θ; exponential twisting is one such example. Let $\sigma^2(\theta)$ denote the (unknown) variance of the estimator using importance sampling with parameter θ. The basic idea is to run relatively short pilot studies to obtain estimates of $\sigma^2(\theta)$ and then select the value of θ^* that optimizes the estimated variance for use in longer runs. This can be done dynamically within stochastic optimization procedures similar to simulated annealing, although some care has to be taken to discard values of θ with exceptionally small estimates of $\sigma^2(\theta)$, since in finite samples "overbiasing" the simulation can lead to small estimates of $\sigma^2(\theta)$ with high probability even though the actual value of $\sigma^2(\theta)$ is high. Other heuristics for some networks have been proposed in [12].

4 Reliability Models

We now turn to a discussion of fast simulation techniques for models of highly reliable systems. An additional overview of this material may be found in [82]. The general class of models that we will be interested in simulating are basically those that can be described by the SAVE (System AVailability Estimator)

modeling language [49, 48] that has been developed at IBM Research (initially in cooperation with Kishor Trivedi of Duke University for modeling language and numerical algorithm development and subsequently in cooperation with Peter Glynn of Stanford University for development of importance sampling techniques). The models are a class of generalized machine repairmen models. A system consists of components and repairmen; components may fail and repairmen repair failed components. When a component fails, it may "affect" other components (with some probability), thereby causing the other components to fail simultaneously. Components may fail in a variety of failure modes (according to some probability distribution). Associated with each failure mode is a set of components (possibly empty) to be affected, a repairman, and a repair distribution; all of these may be different for different failure modes. The operation of a component may "depend upon" the operation of certain other components. For example, a disk drive may need a control unit in order to access data; if the control unit fails, the disk drive has not failed (thereby requiring repair) but rather is said to be "dormant" to reflect its inoperable state. Similarly, the repair of a component may depend upon other components, e.g., repairing (restarting) an operating system requires an operational processor. The modeler describes Boolean-type conditions involving the states of the components under which the system as a whole is considered to be operational, or operating at a reduced level of performance. Thus the system as a whole can tolerate the failure of certain combinations of component failures. In SAVE, all failure and repair distributions are assumed to be exponentially distributed, although we will consider importance sampling for systems with non-exponential distributions as well. For small models, the generator matrix of the underlying CTMC can be constructed and numerical algorithms can be used to solve for the relevant output measures, e.g., steady state availability, mean time to system failure, system failure time distribution, etc. However, for large models consisting of many components, numerical techniques are not feasible and simulation is used. Since system failure events are (hopefully) rare, importance sampling can be used, and SAVE automatically incorporates a provably good importance sampling heuristic called "balanced failure biasing," which will be described in this section.

While the general approach of importance sampling can be applied, it must be done somewhat differently for reliability models than for queueing models. The main reason is that the rare events of interest happen in very different ways in these two types of systems. In queueing models, events such as buffer overflows happen because many events, none of which are particularly rare (e.g., shorter interarrival times and longer service times), combine to produce an event that is extremely rare. In highly dependable computing systems, there is a limited amount of redundancy, so system failure events happen because a few events, each of which is relatively rare (e.g., component failures), combine to produce an event that is extremely rare. Since importance sampling is effective when the simulation follows typical failure paths, different importance sampling approaches are required for queueing and reliability models.

4.1 Markovian Models

We begin by describing results for Markovian reliability models. We assume that the system can be described by CTMC with a finite space; let i and j denote generic states in the system and let $Q(i,j)$ be the generator matrix. We assume there are two types of transitions possible: failures and repairs. For state i, let $\lambda(i) = -Q(i,i)$ denote the total rate out of state i, let $\lambda_F(i)$ denote the total failure rate out of state i and let $\mu_R(i)$ denote the total repair rate out of state i, i.e., $\lambda_F(i) = \sum_{j \in F(i)} Q(i,j)$ where $F(i)$ denotes the set of failure transitions from state i and $\mu_R(i) = \sum_{j \in R(i)} Q(i,j)$ where $R(i)$ denotes the set of repair transitions from state i. Note that $\lambda(i) = \lambda_F(i) + \mu_R(i)$. We assume that there is a single state, 0, in which all components are operational (and thus there are no repair transitions); all other states are assumed to have at least one repair transition. In order to analyze importance sampling approaches, Shahabuddin [88, 89] parameterizes the generator matrix by a rarity parameter ϵ as follows: $Q(i,j) = q(i,j)\epsilon^{d(i,j)}$ for failure transitions where $d(i,j) \geq 1$ and $q(i,j)$ does not depend on ϵ. Also, $Q(i,j)$ is assumed not to depend upon ϵ for repair transitions. Allowing different exponents $d(i,j)$ in the failure rate transitions permits modeling of systems in which either:

- some components are much more reliable than others, or
- some failure modes are very unlikely, e.g., when a component fails, with some small probability it also causes other components to fail with it.

If $d(i,j)$ is the same for all failure transitions the system is said to be balanced; otherwise it is unbalanced. With this parameterization, for small ϵ, failure transitions are unlikely (except out of state 0 from which all transitions are failure transitions). The transition matrix of the embedded DTMC, $P(i,j) = Q(i,j)/\lambda(i)$ then has a similar form to $Q(i,j)$ (for $i \neq 0$), i.e., $P(i,j) = p(i,j)\epsilon^{d(i,j)} + o(\epsilon^{d(i,j)})$ for failure transitions and $P(i,j) = p(i,j) + o(1)$ for repair transitions. Since all transitions from state 0 are failures, the above representation for $P(0,j)$ does not hold; for state 0, $P(0,j) = p(i,j)\epsilon^{d(j)} + o(\epsilon^{d(j)})$, where $d(j) = d(0,j) - \min_k\{d(0,k)\}$. For state 0, it is assumed that any transitions directly into a system failure state are unlikely, i.e., $d(j) \geq 1$ for $j \in F$ (otherwise system failure events are not rare).

Regenerative simulation, starting in state 0, can be used to estimate steady state measures and the mean time to failure using the ratio formula 15 to estimate $\gamma(\epsilon) = P_0(\tau_F < \tau_0)$, the probability, starting in state 0, of hitting a failure state before returning to 0. Shahabuddin shows that there exists an $r \geq 1$ such that $\gamma(\epsilon) = c\epsilon^r + o(\epsilon^r)$. This generalizes a result in [40], although the value of r is typically unknown. Thus we are faced with a rare event simulation. Note that without importance sampling, a typical cycle consists of a failure transition followed by one or more repair transitions until returning to state 0.

In general, we would like to apply importance sampling so as to sample most often from the most likely paths to failure. However, in complex systems, it may not be easy to identify these most likely paths. Thus, heuristics that are

simple to implement need to be developed that search many different paths to failure, yet sample often enough from the most likely failure paths so as to obtain variance reduction. The basic class of techniques known as failure biasing is designed to do this. The first such technique, known as simple failure biasing, was introduced in [70]. In simple failure biasing, if $i \neq 0$, the probability of failure events is increased by selecting a failure transition with some fixed probability p that does not depend on ϵ. Note that in the original system (without importance sampling) the probability of this event is $\lambda_F(i)/\lambda(i)$ which is very small. If a failure transition is selected, then the failure transition to state j is selected proportionally to the original transition rates, i.e., with probability $Q(i,j)/\lambda_F(i)$. If a repair transition is selected, with probability $(1-p)$, then the repair transition to state j is selected with probability $Q(i,j)/\mu_R(i)$. In the original system, the probability of a repair event is very close to one. Note that to estimate $\gamma(\epsilon)$, only the embedded DTMC need be simulated; the likelihood ratio is obtained from Equation 11. Typically, p is chosen so that $0.25 \leq p \leq 0.9$.

Using simple failure biasing, the system ends up in a system failure state, at least some appreciable fraction of the time. However, it may not follow the most likely path to system failure, as the following simple example of an unbalanced system illustrates. Consider a system with two types of components. There is one component of type 1 that has failure rate ϵ^2 and three components of type 2 that each have failure rate ϵ. The system is considered operational if least one component of each type is operational. Under simple failure biasing, given that a failure event occurs, it is a type 1 failure with probability $\epsilon^2/(N_2\epsilon + \epsilon^2)$ where N_2 is the number of operational type 2 components; this probability is of order ϵ. Similarly, the (conditional) probability of a type 2 failure is $(1 - O(\epsilon))$. Thus, under simple failure biasing, when the system ends up in a failure state, most of the time it gets there by having three type 2 component failures. It only rarely (with probability of order ϵ) ends up in the state in which component 1 is failed. However, the path with a single component 1 failure is the most likely path to system failure; its probability is of order ϵ whereas any other system failure path has a much smaller probability of order ϵ^2. Thus while simple failure biasing takes the system to the set of failure states with reasonable probability, it does not push the system along the right failure path often enough. The result of this is that simple failure biasing applied to unbalanced systems may result in estimates having unbounded relative error. On the other hand, when simple failure biasing is applied to balanced systems, bounded relative error (asymptotically optimal as $\epsilon \to 0$) estimates are obtained [88, 89].

To overcome this problem,"balanced failure biasing" was introduced in [51, 88]; its asymptotic optimality for estimating $\gamma(\epsilon)$ was established in [88, 89] and experimental results are presented in [51]. In balanced failure biasing, a failure transition is again selected with probability p. Now however, given a failure event has occurred, the probabilities of all failure transitions are equalized, i.e., a failure transition from i to j is (conditionally) selected with probability $1/|F(i)|$ where $|F(i)|$ is the number of failure transitions from state i (or more generally selected from a distribution that is independent of ϵ). As in simple failure

biasing, given a repair transition has been selected, the repair transition to state j is selected with probability $Q(i,j)/\mu_R(i)$. The proof technique in [88, 89] is matrix algebraic in nature, however a proof based on bounded likelihood ratios is also possible [61, 72, 74]; if $\tau_F < \tau_0$, then the likelihood ratio $L(X) \leq D[\epsilon^r + o(\epsilon^r)]$ for some random variable D that does not depend on ϵ. For estimating steady state unavailability, balanced failure biasing can be used until the set of system failure states is hit, and then importance sampling can be turned off until the start of the next cycle.

Note that in the example above, balanced failure biasing brings the system to the most likely failure state with probability 0.5 in one transition. Mathematically, balancing prevents the denominator of the likelihood ratio from getting too small, thereby producing stable estimates. Under balanced failure biasing, many unlikely paths to system failure may be generated, but enough of the most likely such paths are generated so as to guarantee good estimates. Nakayama [72, 73] formalizes the notion of most likely failure paths in this setting by showing that, given $\tau_F < \tau_0$, there exists a limiting (conditional) distribution on the set of paths. He also shows that balanced failure biasing results in asymptotically optimal estimates of the derivative of $\gamma(\epsilon)$ with respect to the failure rate of a component, as $\epsilon \to 0$. (See also [76] for empirical results.) Further analysis that characterizes when these and more general failure biasing schemes are efficient is given in Nakayama [74, 75].

For estimating transient quantities, such as the unreliability $U(t) = P(\tau_F \leq t)$, failure biasing needs to be augmented with a technique called "forcing" [70] which basically forces the first transition to occur before time t with high probability (perhaps 1). (For a fixed value of t, the probability that the first transition occurs before time t approaches zero as $\epsilon \to 0$.) If t is "small" (i.e., fixed), then balanced failure biasing and forcing produce bounded relative error estimates of $U(t)$ [91]. For "large" values of t, the empirical effectiveness of this technique decreases [51], and a somewhat different approach must be taken. This problem is analyzed in Shahabuddin and Nakayama [91]. Suppose $\lambda(0) = c\epsilon^{b_0} + o(\epsilon^{b_0})$, i.e., the mean holding time in state 0 is of order $1/\epsilon^{b_0}$. By the ratio formula Equation 15, $E_0[\tau_F]$ is of order $1/\epsilon^{r+b_0}$. Thus if $t_\epsilon = t/\epsilon^b$ where $b < b_0$, then both the first event occurring before time t_ϵ and the event $\tau_F < \tau_0$ are rare and balanced failure biasing with forcing is effective. However, if $t_\epsilon = t/\epsilon^b$ where $b_0 < b < r + b_0$, $\tau_F \leq t_\epsilon$ will still be a rare event, however the previous approach will be inefficient. In this case, importance sampling extends over a long time horizon and it is known that the variance of the likelihood ratio blows up in such cases [44]. In this case, the regenerative structure of the system can be exploited to estimate (tight) upper and lower bounds on $U(t)$, e.g.,

$$U(t) \leq \overline{U}(t) = 1 - e^{-\gamma(\epsilon)\lambda(0)t}. \tag{48}$$

This bound is true because $\tau_F \geq E_1 + \ldots E_N$ where N is geometric with success probability $\gamma(\epsilon)$ and the E_i's are i.i.d. exponentials with rate $\lambda(0)$, independent of N. For $t_\epsilon = t/\epsilon^b$ where $0 < b < r + b_0$, $U(t_\epsilon)/\overline{U}(t_\epsilon) \to 1$ and the upper bound $\overline{U}(t_\epsilon)$ is efficiently estimated by estimating $\gamma(\epsilon)$ using balanced failure biasing. If

$t_\epsilon = t/\epsilon^b$ where $b > r + b_0$, then the problem is no longer a rare event simulation. Conditions under which efficient large t estimates of the derivative of $U(t)$ (with respect to a failure rate) are obtained are also given in [91]. A similar approach to efficient estimation of another transient quantity, the expected fraction of time the system is unavailable during the interval $(0, t)$, is considered in Shahabuddin [90]. A different approach to estimating $U(t)$ is presented in Carrasco [19]. Instead of estimating $U(t)$, its Laplace transform $\hat{U}(s) = E[e^{-sT_F}]$ is estimated at a number of values of s. The regenerative structure is again exploited by deriving a renewal equation for $\hat{U}(s)$ in terms of quantities defined over a single cycle that can easily be estimated. Numerical inversion of the estimated Laplace transform is then used to recover estimates of $U(t)$.

Carrasco [18, 19] considers another failure biasing approach, termed failure distance biasing, that attempts to improve on the efficiency of balanced failure biasing by giving more weight to sample paths that are "closer" to the set of system failure states F. In this approach, failure transitions are grouped into classes based on their estimated distance to F and more weight is given to the classes corresponding to shorter distances. Once a class is chosen, if the probabilities given to individual transitions within the class are chosen proportionally to their original rates (as in simple failure biasing), then unbounded relative error may occur in an unbalanced system (see [75] for an example). However, if the probabilities given to individual transitions within the class are balanced, then a result in [88] implies that the resulting estimate has bounded relative error. In practice, the success of this approach depends on the ability to correctly (and efficiently) assign the distances; the class of systems for which this can be done is unclear. In some cases, significant improvements over balanced failure biasing have been obtained for systems with a large number of component types.

The above papers all assume that the repair rate is nonzero for all states, other than state 0. However, in some models this may not be the case, e.g., if repairs are deferred until a sufficient number of components are failed. Juneja and Shahabuddin [62] show that standard balanced failure biasing need not be asymptotically optimal in such situations. However, a generalization of balanced failure biasing is shown to be asymptotically optimal for balanced systems in [62].

4.2 Non-Markovian Reliability Models

We now turn to a discussion of importance sampling for highly reliable systems in which the failure and repair time distributions are not exponentially distributed. Although other techniques have been proposed (see, e.g., [39, 79, 80]), in this section we will concentrate on showing how balanced failure biasing can be generalized by appropriately applying importance sampling to a uniformization based simulation algorithm. This approach also produces bounded relative error estimates for estimating the unreliability $U(t)$ (under appropriate technical conditions). The discussion here follows that in [56, 57, 78, 81] and is based on the notion of hazard rates [13], e.g., if component i has failure density $f_i(x)$ and distribution function $F_i(x)$, then its failure hazard rate is $h_i(x) = f_i(x)/[1 - F_i(x)]$. For an exponential distribution, the hazard rate is constant. If component i is

new at time 0, then the probability that it fails in the interval $(x, x+dx)$, given that it has not failed before time x, is approximately $h_i(x)dx$. Similarly, let $r_i(x)$ denote the hazard rate of component i's repair distribution. For simplicity, we assume that there are N components that can either be operational or failed (these assumptions can be relaxed). At simulation time s, let $O(s)$ denote the set of operational components and let $F(s)$ denote the set of failed components. Let $\lambda_i(s)$ denote the failure (hazard) rate of component i at time s, $\lambda_i(s) = h_i(A_i(s))$ where $A_i(s)$ is the age of component i at time s if $i \in O(s)$, and $\lambda_i(s) = 0$ if $i \notin O(s)$. Similarly, let $\mu_i(s)$ denote the repair rate of component i at time s, which is zero if i is not undergoing repair at time s. We assume that the component failure hazard rates are small and that repair distribution hazard rates are bounded from above and below, i.e., there exist positive, finite constants $\underline{\lambda}, \overline{\lambda}, \underline{\mu}, \overline{\mu}$ such that $\underline{\lambda}\epsilon^{b_i} \leq \lambda_i(s) \leq \overline{\lambda}\epsilon^{b_i}$ where $b_i \geq 1$ and $\underline{\mu} \leq \mu_i(s) \leq \overline{\mu}$. The total failure and repair rates at time s are given by $\lambda_F(s) \equiv \sum_i^N \lambda_i(s)$ and $\mu_R(s) \equiv \sum_i^N \mu_i(s)$, respectively. The total event rate at time s is $\lambda(s) \equiv \lambda_F(s) + \mu_R(s)$. Note the similarity between this situation and that in the Markovian case.

Consider a uniformization based simulation of this process. Generate a Poisson process $\{N_\beta(s)\}$ with rate β such that $\lambda(s) \leq \beta$ for all $0 \leq s \leq t$. Suppose an event in this process occurs at time S. Then the event is a component i failure with probability $\lambda_i(S)/\beta$, it is a component i repair with probability $\mu_i(S)/\beta$, and it is a pseudo-event (i.e., no event at all) with probability $1 - \lambda(S)/\beta$. Given that an event is real (i.e., failure or repair), it is a failure with probability $\lambda_F(S)/\lambda(S)$, which is small (provided $\mu_R(S) > 0$). The analog of balanced failure biasing is now apparent: given the event is real, increase the probability of a failure to p and decrease the probability of repair to $(1-p)$. Then, if the event is a failure, pick component i to fail with probability $1/|O(S)|$, and if the event is a repair, repair component i with probability $\mu_i(S)/\mu_R(S)$. To be effective, the analog of forcing needs to be done in the transient case when no repairs are ongoing. We have also implicitly assumed that the components affected probabilities are not functions of ϵ and therefore do not require importance sampling, although this assumption can be relaxed. This approach can be described more generally by defining importance sampling failure and repair rates $\lambda_i'(s)$ and $\mu_i'(s)$ with which to do sampling such that $\lambda'(s) = \lambda_F'(s) + \mu_R'(s) \leq \beta$ for $(0 \leq s \leq t)$. Then the likelihood ratio at time t, $L(t) = L_F(t) \times L_R(t) \times L_P(t)$ where $L_F(t)$ is the failure event likelihood ratio, $L_R(t)$ is the repair event likelihood ratio, and $L_P(t)$ is the pseudo event likelihood ratio. For example, if component i fails $N_i(t)$ times in $(0,t)$ and T_{ij} denotes the j'th time that component i fails, then

$$L_F(t) = \prod_{i=1}^{N} \prod_{j=1}^{N_i(t)} \frac{\lambda_i(T_{ij})}{\lambda_i'(T_i j)}. \tag{49}$$

If the importance sampling rates are chosen such that $\underline{\lambda'} \leq \lambda_i'(s) \leq \overline{\lambda'}$ whenever $\lambda_i(s) > 0$, $\underline{\mu'} \leq \mu_i'(s) \leq \overline{\mu'}$ whenever $\mu_i(s) > 0$, and $1 - \lambda'(s) \geq \underline{\beta'} > 0$ whenever $1 - \lambda(s) > 0$, then bounded relative error estimates are obtained for $U(t)$ [57].

The result is obtained by showing that $c(t)\epsilon^r \leq U(t)$ for some function $c(t)$ and $r \geq 1$, and then bounding the likelihood ratio when $\tau_f \leq t$: $L(t) \leq d^{N_B(t)}\epsilon^r$ for some constant $d \geq 1$. This establishes bounded relative error, but indicates that the method will only be effective for "small" values of t. "Large" t results for estimating $U(t)$, similar to the Markovian case, have not been obtained.

Since the hazard rate of some repair distributions may not be bounded (e.g., uniform and discrete distributions), the above importance sampling technique has been generalized so as to sample repairs from their given distributions, but to accelerate failure events with importance sampling. Under appropriate technical conditions, bounded relative error is still obtained when failure events are appropriately accelerated, e.g., either by using uniformization based importance sampling for failure events only as described above, or by sampling the time until the next failure event from an exponential distribution with rate bounded away from zero. These techniques can be adapted to steady state estimation using the ratio formula of Equation 16 and the splitting technique of Section 2.4, although bounded relative error has not been proven in this situation [81].

5 Summary

This paper has surveyed techniques for efficient simulation of rare events in queueing and reliability models. A number of common themes appear in both problem settings:

- Using importance sampling to accelerate the occurrence of the rare events.
- The notion of asymptotically optimal importance sampling.
- Obtaining asymptotically optimal importance sampling by bounding the likelihood ratio on the rare event of interest.

Among the differences between the two application settings are:

- In queueing models, the rare event happens because of a combination of a large number of events, none of which are particularly rare. In reliability models, rare events happen because of the occurrence of only a few events, each of which is itself rare. Thus different importance sampling approaches are required: exponential twisting in queueing models, failure biasing in reliability models. In addition, rare event probabilities in queueing models decrease exponentially at a known rate, whereas they decrease polynomially at an unknown rate in reliability models.
- Performing asymptotically optimal importance sampling of queueing models generally requires understanding quite a bit more about the structure of the problem, as compared to reliability models. Thus, the class of queueing models for which asymptotically optimal importance sampling algorithms are known is more limited.

- There is typically a unique asymptotically optimal change of measure for queueing models: exponential twisting with a certain parameter. Finding the optimal twisting parameter may be fairly complex, e.g., solving a nonlinear equation involving the largest eigenvalues of the sources as in the case of the discrete time queue with Markovian arrival processes. There is much more simplicity and flexibility in simulating reliability models: essentially any generalized form of balanced failure biasing is asymptotically optimal. Furthermore, there is flexibility in selecting the parameters of the failure biasing method, e.g., balanced failure biasing with any fixed p is asymptotically optimal, although the constant term in the variance can certainly be improved by, for example, selecting p appropriately.

- For the queueing models studied, the asymptotically optimal importance sampling distribution is equivalent to the (asymptotic) ordinary distribution, given that the rare event has occurred. For reliability models, it is harder to sample from the (asymptotic) ordinary distribution, given that the rare event has occurred. In practice, failure biasing techniques may devote a large fraction of the samples to highly unlikely sample paths.

- Thus, when the asymptotically optimal change of measure is known for a queueing model, better variance reduction is typically obtained than in reliability models (for estimating a probability of the same order of magnitude).

Acknowledgement

My understanding of importance sampling and rare event simulation has greatly benefited from interactions with my co-authors, whose names are listed in the bibliography. Special appreciation is due to Cheng-Shang Chang and Perwez Shahabuddin. Thanks also to Søren Asmussen for providing helpful comments on the manuscript.

References

[1] Al-Qaq, W.A., M. Devetsikiotis, and K.R. 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.

[2] Anantharam, V. 1988. How large delays build up in a GI/G/1 queue. *Queueing Systems* 5: 345-368.

[3] Anantharam, V., P. Heidelberger, and P. Tsoucas. 1990. Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation. IBM Research Report RC 16280. Yorktown Heights, New York.

[4] Arsham, H., A. Fuerverger, D.L. McLeish, J. Kreimer, and R.Y. Rubinstein. 1989. Sensitivity analysis and the "what if" problem in simulation analysis. *Math. Comput. Modelling* 12: 193-219.

[5] Asmussen, S. 1982. Conditioned limit theorems relating a random walk to its associate. *Advances in Applied Probability* 14: 143-170.

[6] Asmussen, S. 1985. Conjugate processes and the simulation of ruin problems. *Stochastic Processes and their Applications* 20: 213-229.

[7] Asmussen, S. 1987 . *Applied Probability and Queues.* New York, NY: J. Wiley & Sons, Inc.

[8] Asmussen, S. 1989. Risk theory in a Markovian environment. *Scand. Actuarial J.*, 69-100.

[9] Asmussen, S. 1990. Exponential families and regression in the Monte Carlo study of queues and random walks. *The Annals of Statistics* 18: 1851-1867.

[10] Asmussen, S. 1993. Busy period analysis, rare events and transient behaviour in fluid models. To appear in *Journal of Applied Mathematics and Stochastic Analysis.*

[11] Asmussen, S., and R.Y. Rubinstein. 1992. The efficiency and heavy traffic properties of the score function method in sensitivity analysis of queueing models. *Advances in Applied Probability* 24: 172-201.

[12] Asmussen, S., R.Y. Rubinstein, and C.L. Wang. 1992. Efficient regenerative rare events simulation via the likelihood ratio method. Preprint.

[13] Barlow, R.E., and F. Proschan. 1981. *Statistical Theory of Reliability and Life Testing.* New York, NY: Holt, Reinhart and Winston, Inc.

[14] Breiman, L. 1968. *Probability.* Reading, MA: Addison-Wesley.

[15] Brown, M. 1990. Error bounds for exponential approximations of geometric convolutions. *The Annals of Probability* 18: 1388-1402.

[16] Bucklew, J. 1990. *Large Deviation Techniques in Decision, Simulation and Estimation.* New York, NY: J. Wiley & Sons, Inc.

[17] Bucklew, J.A., P. Ney, and J.S. Sadowsky. 1990. Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. *J. Appl. Prob.* 27: 44-99.

[18] Carrasco, J.A. 1991. 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*, 337-351,

[19] 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, Pisa, Italy.

[20] Chang, C.S. 1992. Stability, queue length and delay, Part II: Stochastic queueing networks. IBM Research Report RC 17709, Yorktown Heights, New York. Part of the report is published in *Proceedings of the IEEE CDC'92 Conference*, 1005-1010, IEEE Computer Society Press, Tucson, Arizona, 1992.

[21] Chang, C.S., P. Heidelberger, S. Juneja, and P. Shahabuddin. 1992. Effective bandwidth and fast simulation of ATM intree networks. IBM Research Report RC 18586, Yorktown Heights, New York. To appear in *Proceedings of the Performance '93 Conference.*

[22] Chang, C.S., P. Heidelberger, S. Juneja, and P. Shahabuddin. 1993. The application of effective bandwidth to fast simulation of communication networks. IBM Research Report RC 18877, Yorktown Heights, New York.

[23] Chen, H., and A. Mandelbaum. 1991. Discrete flow networks: bottleneck analysis and fluid approximations. *Mathematics of Operations Research* 16: 408-446.

[24] Çinlar, E. 1975. *Introduction to Stochastic Processes.* Englewood Cliffs, NJ: Prentice Hall, Inc.,

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

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

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

[28] Csiszár, I., T.M. Cover and B.-S. Choi. 1987. Conditional limit theorems under Markov conditioning. *IEEE Transactions on Information Theory* 33: 788-801.

[29] Devetsikiotis, M., and K.R. Townsend. 1992. On the efficient simulation of large communication networks using importance sampling. In *Proceedings of IEEE Globecom '92*, IEEE Computer Society Press.

[30] Devetsikiotis, M., and K.R. Townsend. 1992. A dynamic importance sampling methodology for the efficient estimation of rare event probabilities in regenerative simulations of queueing systems. In *Proceedings of the IEEE ICC '92 Conference*, 1290-1297, IEEE Computer Society Press.

[31] Devetsikiotis, M., and K.R. Townsend. 1993. Statistical optimization of dynamic importance sampling parameters for efficient simulation of communication networks. To appear in *IEEE/ACM Transactions on Networking.*

[32] Ellis, R. 1984. Large deviations for a general class of random vectors. *Annals of Probability* 12: 1-12.

[33] Feller, W. 1971. *An Introduction to Probability Theory and its Applications.* Vol. 2 (Second Edition). New York, NY: J. Wiley & Sons, Inc.

[34] Frater, M.R., and B.D.O. Anderson. 1989. Fast estimation of the statistics of excessive backlogs in tandem networks of queues. *Australian Telecommunications Research* 23: 49-55.

[35] Frater, M.R., R.R. Bitmead, R.A. Kennedy, and B.D.O. Anderson. 1989. Rare events and reverse time models. In *Proceedings of the 28th Conference on Decision and Control*, 1180-1183, IEEE Press.

[36] Frater, M.R., T.M. Lenon, 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.

[37] Frater, M.R., J. Walrand, and B.D.O. Anderson. 1990. Optimally efficient simulation of buffer overflows in queues with deterministic service times via importance sampling. *Australian Telecommunications Research* 24: 1-8.

[38] Gärtner, J. 1977. On large deviations from invariant measure. *Theory Probab. Appl.* 22: 24-39.

[39] Geist, R.M. and M.K. Smotherman. 1989. Ultrahigh reliability estimates through simulation. In *Proceedings of the Annual Reliability and Maintainability Symposium*, 350-355, IEEE Press.

[40] Gertsbakh, I.B. 1984. Asymptotic methods in reliability theory: A review. *Advances in Applied Probability* 16: 147-175.

[41] Gibbens, R.J., and P.J. Hunt. 1991. Effective bandwidths for the multi-type UAS channel. *Queueing Systems* 9: 17-28.

[42] Glasserman, P. 1993. Stochastic monotonicity and conditional Monte Carlo for likelihood ratios. *Advances in Applied Probability* 25: 103-115.

[43] Glynn, P.W. 1989. A GSMP formalism for discrete event systems. *Proceedings of the IEEE* 77: 14-23.

[44] Glynn, P.W. 1992. Importance sampling for Markov chains: asymptotics for the variance. Technical Report, Department of Operations Research, Stanford University, To appear in *Stochastic Models.*

[45] Glynn, P.W., and P. Heidelberger. 1992. Experiments with initial transient deletion for parallel, replicated steady-state simulations. *Management Science* 38: 400 - 418.

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

[47] Glynn, P.W., and W. Whitt. 1992. The asymptotic efficiency of simulation estimators. *Operations Research* 40: 505-520.

[48] Goyal, A., W.C. Carter, E. de Souza e Silva, S.S. Lavenberg, and K.S. Trivedi. 1986. The System Availability Estimator. In *Proceedings of the Sixteenth International Symposium on Fault-Tolerant Computing*, 84-89, IEEE Computer Society Press.

[49] Goyal, A., and S.S. Lavenberg. 1987. Modeling and analysis of computer system availability. *IBM Journal of Research and Development* 31, 6: 651-664.

[50] Goyal, A., P. Heidelberger, and P. Shahabuddin. 1987. Measure specific dynamic importance sampling for availability simulations. In *1987 Winter Simulation Conference Proceedings*, 351-357, IEEE Press.

[51] 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.

[52] Guérin, R., H. Ahmadi and M. Naghshineh. 1991. Equivalent capacity and its application to bandwidth allocation in high-speed networks. *IEEE J. Select. Areas Commun.* 9: 968-981.

[53] Halton, J.H. 1970. A retrospective and prospective survey of the Monte Carlo method. *SIAM Review* 12: 1-60.

[54] Hammersley, J.M., and D.C. Handscomb, 1964. *Monte Carlo Methods.* London: Methuen and Co., Ltd.

[55] Heidelberger, P. 1988. Discrete event simulations and parallel processing: statistical properties. *SIAM Journal on Scientific and Statistical Computing* 9: 1114-1132.

[56] Heidelberger, P., Nicola, V.F., and Shahabuddin, P. 1992. Simultaneous and efficient simulation of highly dependable systems with different underlying distributions. In *Proceedings of the 1992 Winter Simulation Conference*, 458-465, IEEE Press.

[57] Heidelberger, P, P. Shahabuddin, and V.F. Nicola. 1993. Bounded relative error in estimating transient measures of highly dependable non-Markovian systems. IBM Research Report RC 18794, Yorktown Heights, New York.

[58] Heidelberger, P., and P. Tsoucas. 1991. Reverse time simulation of rare events. *IBM Technical Disclosures Bulletin* 34, No. 3: 163-165.

[59] Iglehart, D.L. 1972. Extreme values in the GI/G/1 queue. *Annals of Mathematical Statistics* 43: 627-635.

[60] Jensen, A. 1953. Markov chains as an aid in the study of Markov processes. *Skand. Aktuarietidskr.* 36: 87-91.

[61] Juneja, S. 1993. *Efficient Rare Event Simulation of Stochastic Systems.* Ph.D. Thesis, Department of Operations Research, Stanford University, California.

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

[63] Kalos, M.H., and P.A. Whitlock. 1986. *Monte Carlo Methods, Volume I: Basics.* New York, NY: John Wiley & Sons, Inc.

[64] Keilson, J. 1979. *Markov Chain Models - Rarity and Exponentiality*, New York, NY: Springer Verlag.

[65] Kelly, F.P. 1979. *Reversibility and Stochastic Networks.* New York, NY: John Wiley & Sons, Inc.

[66] Kelly, F.P. 1991. Effective bandwidths at multi-class queues. *Queueing Systems* 9: 5-16.

[67] Kesidis, G., and J. Walrand. 1993. Quick simulation of ATM buffers with on-off multiclass Markov fluid sources. To appear in *ACM Transactions on Modeling and Computer Simulation.*

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

[69] Lehtonen, T., and H. Nyrhinen. 1992. On asymptotically efficient simulation of ruin probabilities in a Markovian environment. *Scand. Actuarial. J.* 60-75.

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

[71] Lewis, P.A.W., and G.S. Shedler. 1979. Simulation of nonhomogeneous Poisson processes by thinning. *Naval Research Logistics Quarterly* 26, 403-413.

[72] Nakayama, M.K. 1991. *Simulation of Highly Reliable Markovian and Non-Markovian Systems.* Ph.D. Thesis, Department of Operations Research, Stanford University, California.

[73] Nakayama, M.K. 1991. Asymptotics for likelihood ratio derivative estimators in simulations of highly reliable Markovian systems. IBM Research Report RC 17357, Yorktown Heights, NY.

[74] Nakayama, M.K. 1993. A characterization of the simple failure biasing method for simulations of highly reliable Markovian systems. IBM Research Report RC 18721, Yorktown Heights, New York.

[75] Nakayama, M.K. 1993. General conditions for bounded relative error in simulations of highly reliable Markovian systems. IBM Research Report RC 18993. Yorktown Heights, New York.

[76] Nakayama, M.K., A. Goyal and P.W. Glynn. 1990. Likelihood ratio sensitivity analysis for Markovian models of highly dependable systems. IBM Research Report RC 15400, Yorktown Heights, New York. To appear in *Operations Research.*

[77] Nicol, D.M., and D.L. Palumbo. 1993. Reliability analysis of complex models using SURE bounds. ICASE NASA Langley Research Center Technical Report 93-14.

[78] Nicola, V.F., P. Heidelberger and P. Shahabuddin. 1992. Uniformization and exponential transformation: techniques for fast simulation of highly dependable non-Markovian systems. In *Proceedings of the Twenty-Second International Symposium on Fault-Tolerant Computing*, 130-139, IEEE Computer Society Press.

[79] Nicola, V.F., M.K. Nakayama, P. Heidelberger and A. Goyal. 1990. Fast simulation of dependability models with general failure, repair and maintenance processes. In *Proceedings of the Twentieth International Symposium on Fault-Tolerant Computing*, 491-498, IEEE Computer Society Press.

[80] Nicola, V.F., M.K. Nakayama, P. Heidelberger, and A. Goyal. 1991. Fast simulation of highly dependable systems with general failure and repair processes. IBM Research Report RC 16993. Yorktown Heights, New York. To appear in *IEEE Transactions on Computers.*

[81] 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.

[82] Nicola, V.F., P. Shahabuddin, and P. Heidelberger. 1993. Techniques for fast simulation of highly dependable systems. IBM Research Report RC 18956, Yorktown Heights, New York.

[83] 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.

[84] Rubinstein, R.Y. 1991. How to optimize discrete-event systems from a single sample path by the score function method. *Annals of Operations Research* 27: 175-212.

[85] 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.

[86] Sadowsky, J.S. 1993. On the optimality and stability of exponential twisting in Monte Carlo estimation. *IEEE Transactions on Information Theory* 39: 119-128.

[87] Sadowsky, J.S., and J.A. Bucklew. 1990. On large deviations theory and asymptotically efficient Monte Carlo estimation. *IEEE Transactions on Information Theory* 36: 579-588.

[88] Shahabuddin, P. 1990. *Simulation and Analysis of Highly Reliable Systems.* Ph.D. Thesis, Department of Operations Research, Stanford University, California.

[89] Shahabuddin, P. 1991. Importance sampling for the simulation of highly reliable Markovian systems. IBM Research Report RC 16729, Yorktown Heights, New York. To appear in *Management Science.*

[90] Shahabuddin, P. 1993. Fast transient simulation of Markovian models of highly dependable systems. IBM Research Report RC 18587, Yorktown Heights, New York. To appear in *Proceedings of the Performance '93 Conference.*

[91] Shahabuddin, P., and M.K. Nakayama. 1993. Estimation of reliability and its derivatives for large time horizons in Markovian systems. IBM Research Report RC 18864, Yorktown Heights, New York. To appear in *Proceedings of 1993 Winter Simulation Conference.*

[92] Shahabuddin, P., V.F. Nicola, P. Heidelberger, A. Goyal, and P.W. Glynn. 1988. Variance Reduction in Mean Time to Failure Simulations. In *1988 Winter Simulation Conference Proceedings*, 491-499, IEEE Press.

[93] Shanthikumar, J.G. 1986. Uniformization and hybrid simulation/analytic models of renewal processes. *Operations Research* 34: 573-580.

[94] Shwartz, A., and A. Weiss. 1993. Induced rare events: analysis via time reversal and large deviations. To appear in *Advances in Applied Probability.*

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

[96] Smith, W.L. 1955. Regenerative stochastic processes. *Proc. Roy. Soc. Ser. A.* 232: 6-31.

[97] Tsoucas, P. 1989. Rare events in series of queues. IBM Research Report RC 15530, Yorktown Heights, New York.

[98] Weiss, A. 1986. A new technique for analyzing large traffic systems. *Advances in Applied Probability* 18: 506-532.

[99] Whitt, W. 1980. Continuity of generalized semi-Markov processes. *Mathematics of Operations Research* 5: 494-501.

[100] Whitt, W. 1993. Tail probability with statistical multiplexing and effective bandwidths in multi-class queues. To appear in *Telecommunication Systems.*