S.G. Henderson and B.L. Nelson (Eds.), *Handbook in OR & MS, Vol. 13*
Copyright © 2006 Elsevier B.V. All rights reserved
DOI: 10.1016/S0927-0507(06)13011-X

Chapter 11

Rare-Event Simulation Techniques: An Introduction and Recent Advances

S. Juneja

School of Technology and Computer Science, Tata Institute of Fundamental Research, India
E-mail: juneja@tifr.res.in

P. Shahabuddin

Department of Industrial Engineering and Operations Research, Columbia University, USA

Abstract

In this chapter we review some of the recent developments for efficient estimation of rare-events, most of which involve application of importance sampling techniques to achieve variance reduction. The zero-variance importance sampling measure is well known and in many cases has a simple representation. Though not implementable, it proves useful in selecting good and implementable importance sampling changes of measure that are in some sense close to it and thus provides a unifying framework for such selections. Specifically, we consider rare events associated with: (1) multi-dimensional light-tailed random walks, (2) with certain events involving heavy-tailed random variables and (3) queues and queueing networks. In addition, we review the recent literature on development of adaptive importance sampling techniques to quickly estimate common performance measures associated with finite-state Markov chains. We also discuss the application of rare-event simulation techniques to problems in financial engineering. The discussion in this chapter is nonmeasure theoretic and kept sufficiently simple that the key ideas are accessible to beginners. References are provided for more advanced treatments.

1 Introduction

Rare-event simulation involves estimating extremely small but important probabilities. Such probabilities are of importance in various applications: In modern packet-switched telecommunications networks, in order to reduce delay variation in carrying real-time video traffic, the buffers within the switches are of limited size. This creates the possibility of packet loss if the buffers overflow. These switches are modeled as queueing systems and it is important to estimate the extremely small loss probabilities in such queueing systems (see, e.g., Chang et al., 1994; Heidelberger, 1995). Managers of portfolios of loans need to maintain reserves to protect against rare events involving large losses due to multiple loan defaults. Thus, accurate measurement of the probability of large losses is of utmost importance to them (see, e.g., Glasserman and Li, 2005). In insurance settings, the overall wealth of the insurance company is modeled as a stochastic process. This incorporates the incoming wealth due to insurance premiums and outgoing wealth due to claims. Here the performance measures involving rare events include the probability of ruin in a given time frame or the probability of eventual ruin (see, e.g., Asmussen, 1985, 1989, 2000). In physical systems designed for a high degree of reliability, the system failure is a rare event. In such cases the related performance measures of interest include the mean time to failure, and the fraction of time the system is down or the 'system unavailability' (see, e.g., Goyal et al., 1992). In many problems in polymer statistics, population dynamics and percolation, statistical physicists need to estimate probabilities of order 10^{-50} or rarer, often to verify conjectured asymptotics of certain survival probabilities (see, e.g., Grassberger, 2002; Grassberger and Nadler, 2000).

Importance sampling is a Monte Carlo simulation variance reduction technique that has achieved dramatic results in estimating performance measures associated with certain rare events (see, e.g., Glynn and Iglehart, 1989, for an introduction). It involves simulating the system under a change of measure that accentuates paths to the rare-event and then unbiasing the resultant output from the generated path by weighing it with the 'likelihood ratio' (roughly, the ratio of the original measure and the new measure associated with the generated path). In this chapter we primarily highlight the successes achieved by this technique for estimating rare-event probabilities in a variety of stochastic systems.

We refer the reader to Heidelberger (1995) and Asmussen and Rubinstein (1995) for earlier surveys on rare-event simulation. In this chapter we supplement these surveys by focusing on the more recent developments[1]. These include a brief review of the literature on estimating rare events related to multidimensional light-tailed random walks (roughly speaking, light-tailed random variables are those whose tail distribution function decays at an exponential rate or faster, while for heavy-tailed random variables it decays at a slower rate, e.g., polynomially). These are important as many mathematical models of interest involve a complex interplay of constituent random walks, and the way rare events happen in random walks settings provides insights for the same in more complex models.

[1] The authors confess to the lack of comprehensiveness and the unavoidable bias towards their research in this survey. This is due to the usual reasons: Familiarity with this material and the desire to present the authors viewpoint on the subject.

We also briefly review the growing literature on adaptive importance sampling techniques for estimating rare events and other performance measures associated with Markov chains. Traditionally, a large part of rare-event simulation literature has focused on implementing *static* importance sampling techniques. (By static importance sampling we mean that a fixed change of measure is used throughout the simulation, while adaptive importance sampling involves updating and learning an improved change of measure based on the simulated sample paths.) Here, the change of measure is selected that emphasizes the most likely paths to the rare event. In many cases large deviations theory is useful in identifying such paths (for an introduction see, e.g., Dembo and Zeitouni, 1998; Shwartz and Weiss, 1995). Unfortunately, one can prove the effectiveness of such static importance sampling distributions only in special and often simple cases. There also exists a substantial literature highlighting cases where static importance sampling distributions with intuitively desirable properties lead to large, and even infinite, variance. In view of this, adaptive importance sampling techniques are particularly exciting as at least in the finite state Markov chain settings, they appear to be quite effective in solving a large class of problems.

Heidelberger (1995) provides an excellent review of reliability and queueing systems. In this chapter, we restrict our discussion to only a few recent developments in queueing systems.

A significant portion of our discussion focuses on the probability that a Markov process observed at a hitting time to a set lies in a rare subset. Many commonly encountered problems in rare-event simulation literature are captured in this framework. The importance sampling zero-variance estimator of small probabilities is well known, but unimplementable as it involves a priori knowledge of the probability of interest. Importantly, in this framework, the Markov process remains Markov under the zero-variance change of measure (although explicitly determining it remains at least as hard as determining the original probability of interest). This Markov representation is useful as it allows us to view the process of selecting a good importance sampling distribution from a class of easily implementable ones as identifying a distribution that is in some sense closest to the zero-variance measure. In the setting of stochastic processes involving random walks this often amounts to selecting a suitable *exponentially twisted* distribution.

We also review importance sampling techniques for rare events involving heavy-tailed random variables. This has proved to be a challenging problem in rare-event simulation and except for the simplest of cases, the important problems remain unsolved.

In addition, we review a growing literature on application of rare-event simulation techniques in financial engineering settings. These focus on efficiently estimating value-at-risk in a portfolio of investments and the probability of large losses due to credit risk in a portfolio of loans.

The following example[2] is useful in demonstrating the problem of rare-event simulation and the essential idea of importance sampling for beginners.

1.1 An illustrative example

Consider the problem of determining the probability that eighty or more heads are observed in one hundred independent tosses of a fair coin.

Although this is easily determined analytically by noting that the number of heads is binomially distributed (the probability equals 5.58×10^{-10}), this example is useful in demonstrating the problem of rare-event simulation and in giving a flavor of some solution methodologies. Through simulation, this probability may be estimated by conducting repeated experiments or trials of one hundred independent fair coin tosses using a random number generator. An experiment is said to be a success and its output is set to one if eighty or more heads are observed. Otherwise the output is set to zero. Due to the law of large numbers, an average of the outputs over a large number of independent trials gives a consistent estimate of the probability. Note that on average 1.8×10^9 trials are needed to observe one success. It is reasonable to expect that a few orders of magnitude higher number of trials are needed before the simulation estimate becomes somewhat reliable (to get a 95% confidence level of width ±5% of the probability value about 2.75×10^{12} trials are needed). This huge computational effort needed to generate a large number of trials to reliably estimate small probabilities via 'naive' simulation is the basic problem of rare-event simulation.

Importance sampling involves changing the probability dynamics of the system so that each trial gives a success with a high probability. Then, instead of setting the output to one every time a success is observed, the output is unbiased by setting it equal to the likelihood ratio of the trial or the ratio of the original probability of observing this trial with the new probability of observing the trial. The output is again set to zero if the trial does not result in a success. In the coin tossing example, suppose under the new measure the trials remain independent and the probability of heads is set to $p > 1/2$. Suppose that in a trial m heads are observed for $m \geqslant 80$. The output is then set to the likelihood ratio which equals

$$\frac{(1/2)^m (1/2)^{100-m}}{p^m (1-p)^{100-m}}. \tag{1}$$

It can be shown (see Section 2) that the average of many outputs again gives an unbiased estimator of the probability. The key issue in importance sampling is to select the new probability dynamics (e.g., p) so that the resultant output is smooth, i.e., its variance is small so that a small number of trials are needed

[2] This example and some of the discussion appeared in Juneja (2003).

to get a reliable estimate. Finding such a probability can be a difficult task requiring sophisticated analysis. A wrong selection may even lead to increase in variance compared to naive simulation.

In the coin tossing example, this variance reduction may be attained by keeping p large so that success of a trial becomes more frequent. However, if p is very close to one, the likelihood ratio on trials can have a large amount of variability. To see this, consider the extreme case when $p \approx 1$. In this case, in a trial where the number of heads equals 100, the likelihood ratio is $\approx 0.5^{100}$ whereas when the number of heads equals 80, the likelihood ratio is $\approx 0.5^{100}/(1-p)^{20}$, i.e., orders of magnitude higher. Hence, the variance of the resulting estimate is large. An in-depth analysis of this problem in Section 4 (in a general setting) shows that $p = 0.8$ gives an estimator of the probability with an enormous amount of variance reduction compared to the naive simulation estimator. Whereas trials of order 10^{12} are required under naive simulation to reliably estimate this probability, only a few thousand trials under importance sampling with $p = 0.8$ give the same reliability. More precisely, for $p = 0.8$, it can be easily numerically computed that only 7,932 trials are needed to get a 95% confidence level of width ±5% of the probability value, while interestingly, for $p = 0.99$, 3.69×10^{22} trials are needed for this accuracy.

Under the zero-variance probability measure, the output from each experiment is constant and equals the probability of interest (this is discussed further in Sections 2 and 3). Interestingly, in this example, the zero-variance measure has the property that the probability of heads after n tosses is a function of m, the number of heads observed in n tosses. Let $p_{n,m}$ denote this probability. Let $\mathrm{P}(n, m)$ denote the probability of observing at least m heads in n tosses under the original probability measure. Note that $\mathrm{P}(100, 80)$ denotes our original problem. Then, it can be seen that (see Section 3.2)

$$p_{n,m} = \frac{1}{2}\frac{\mathrm{P}(100-n-1, 80-m-1)}{\mathrm{P}(100-n, 80-m)}.$$

Numerically, it can be seen that $p_{50,40} = 0.806$, $p_{50,35} = 0.902$ and $p_{50,45} = 0.712$, suggesting that $p = 0.8$ mentioned earlier is close to the probabilities corresponding to the zero variance measure.

The structure of this chapter is as follows: In Section 2 we introduce the rare-event simulation framework and importance sampling in the abstract setting. We also discuss the zero-variance estimator and common measures of effectiveness of more implementable estimators. This discussion is specialized to a Markovian framework in Section 3. In this section we also discuss examples showing how common diverse applications fit this framework. In Section 4 we discuss effective importance sampling techniques for some rare events associated with multidimensional random walks. Adaptive importance sampling methods are discussed in Section 5. In Section 6 we discuss some recent developments in queueing systems. Heavy-tailed simulation is described in Section 7. In Section 8 we give examples of specific rare-event simulation problems in the financial engineering area and discuss the approaches that have been used. Sections 7 and 8 may be read independently of the rest of the paper as long as one has the basic background that is described in Section 2.

2 Rare-event simulation and importance sampling

2.1 Naive simulation

Consider a sample space Ω with a probability measure P. Our interest is in estimating the probability $\mathrm{P}(\mathcal{E})$ of a rare event $\mathcal{E} \subset \Omega$. Let $I(\mathcal{E})$ denote the indicator function of the event $\mathcal{E}$, i.e., it equals 1 along outcomes belonging to $\mathcal{E}$ and equals zero otherwise. Let γ denote the probability $\mathrm{P}(\mathcal{E})$. This may be estimated via naive simulation by generating independent samples $(I_1(\mathcal{E}), I_2(\mathcal{E}), \ldots, I_n(\mathcal{E}))$ of $I(\mathcal{E})$ via simulation and taking the average

$$\frac{1}{n}\sum_{i=1}^{n} I_i(\mathcal{E})$$

as an estimator of γ. Let $\hat{\gamma}_n(\mathrm{P})$ denote this estimator. The law of large numbers ensures that $\hat{\gamma}_n(\mathrm{P}) \to \gamma$ almost surely (a.s.) as $n \to \infty$.

However, as we argued in the introduction, since γ is small, most samples of $I(\mathcal{E})$ would be zero, while rarely a sample equaling one would be observed. Thus, n would have to be quite large to estimate γ reliably. The central limit theorem proves useful in developing a confidence interval (CI) for the estimate and may be used to determine the n necessary for accurate estimation. To this end, let $\sigma_{\mathrm{P}}^2(X)$ denote the variance of any random variable X simulated under the probability P. Then, for large n, an approximate $(1-\alpha)100\%$ CI for γ is given by

$$\hat{\gamma}_n(\mathrm{P}) \pm z_{\alpha/2}\frac{\sigma_{\mathrm{P}}(I(\mathcal{E}))}{\sqrt{n}},$$

where z_x is the number satisfying the relation $\mathrm{P}(\mathrm{N}(0,1) \geqslant z_x) = x$. Here, $\mathrm{N}(0,1)$ denotes a normally distributed random variable with mean zero and variance one (note that $\sigma_{\mathrm{P}}^2(I(\mathcal{E})) = \gamma(1-\gamma)$, and since $\hat{\gamma}_n(\mathrm{P}) \to \gamma$ a.s., $\sigma_{\mathrm{P}}^2(I(\mathcal{E}))$ may be estimated by $\hat{\gamma}_n(\mathrm{P})(1-\hat{\gamma}_n(\mathrm{P}))$ to give an approximate $(1-\alpha)100\%$ CI for γ).

Thus, n may be chosen so that the width of the CI, i.e., $2z_{\alpha/2}\sqrt{\gamma(1-\gamma)/n}$ is sufficiently small. More appropriately, n should be chosen so that the width of the CI relative to the quantity γ being estimated is small. For example, a confidence interval width of order 10^{-6} is not small in terms of giving an accurate estimate of γ if γ is of order 10^{-8} or less. On the other hand, it provides an excellent estimate if γ is of order 10^{-4} or more.

Thus, n is chosen so that $2z_{\alpha/2}\sqrt{(1-\gamma)/(\gamma n)}$ is sufficiently small, say within 5% (again, in practice, γ is replaced by its estimator $\hat{\gamma}_n(\mathrm{P})$, to approximately select the correct n). This implies that as $\gamma \to 0$, $n \to \infty$ to obtain a reasonable level of relative accuracy. In particular, if γ decreases at an exponential rate with respect to some system parameter b (e.g., $\gamma \approx \exp(-\theta b)$, $\theta > 0$; this may be the case for queues with light tailed service distribution where the probability of exceeding a threshold b in a busy cycle decreases at an exponential rate with b) then the computational effort n increases at an exponential rate with b to maintain a fixed level of relative accuracy. Thus, naive simulation becomes an infeasible proposition for sufficiently rare events.

2.2 *Importance sampling*

Now we discuss how importance sampling may be useful in reducing the variance of the simulation estimate and hence reducing the computational effort required to achieve a fixed degree of relative accuracy. Consider another distribution P^* with the property that $\mathrm{P}^*(A) > 0$ whenever $\mathrm{P}(A) > 0$ for $A \subset \mathcal{E}$. Then,

$$\begin{aligned}\mathrm{P}(\mathcal{E}) &= \mathrm{E}_{\mathrm{P}}\big(I(\mathcal{E})\big)\\ &= \int I(\mathcal{E})\,\mathrm{dP} = \int I(\mathcal{E})\frac{\mathrm{dP}}{\mathrm{dP}^*}\,\mathrm{dP}^*\\ &= \int I(\mathcal{E})L\,\mathrm{dP}^* = \mathrm{E}_{\mathrm{P}^*}\big(LI(\mathcal{E})\big),\end{aligned} \tag{2}$$

where the random variable $L = \frac{\mathrm{dP}}{\mathrm{dP}^*}$ denotes the Radon–Nikodym derivative (see, e.g., Royden, 1984) of the probability measure P with respect to P^* and is referred to as the likelihood ratio. When the state space Ω is finite or countable, $L(\omega) = \mathrm{P}(\omega)/\mathrm{P}^*(\omega)$ for each $\omega \in \Omega$ such that $\mathrm{P}^*(\omega) > 0$ and (2) equals $\sum_{\omega\in\mathcal{E}} L(\omega)\mathrm{P}^*(\omega)$ (see Section 3 for examples illustrating the form of the likelihood ratio in simple Markovian settings). This suggests the following alternative importance sampling simulation procedure for estimating γ: Generate n independent samples $(I_1(\mathcal{E}), L_1), (I_2(\mathcal{E}), L_2), \ldots, (I_n(\mathcal{E}), L_n)$ of $(I(\mathcal{E}), L)$ using P^*. Then

$$\hat{\gamma}_n\big(\mathrm{P}^*\big) = \frac{1}{n}\sum_{i=1}^{n} I_i(\mathcal{E})L_i \tag{3}$$

provides an unbiased estimator of γ.

Consider the estimator of γ in (3). Again the central limit theorem may be used to construct confidence intervals for γ. The relative width of the confidence interval is proportional to $\sigma_{\mathrm{P}^*}(LI(\mathcal{E}))/(\gamma\sqrt{n})$. The ratio of the standard deviation of an estimate to its mean is defined as the relative error. Thus, the larger the relative error of $LI(\mathcal{E})$ under P^*, the larger the sample size needed to achieve a fixed relative width of the confidence interval. In particular, the aim of importance sampling is to find a P^* that minimizes this relative error, or equivalently, the variance of the output $LI(\mathcal{E})$.

In practice, the simulation effort required to generate a sample under importance sampling is typically higher compared to naive simulation, thus the ratio of the variances does not tell the complete story. Therefore, the comparison of two estimators should be based not on the variances of each estimator, but on the product of the variance and the expected computational effort required to generate samples to form the estimator (see, e.g., Glynn and Whitt, 1992). Fortunately, in many cases the variance reduction achieved through importance sampling is so high that even if there is some increase in effort to generate a single sample, the total computational effort compared to naive simulation is still orders of magnitude less for achieving the same accuracy (see, e.g., Chang et al., 1994; Heidelberger, 1995).

Also note that in practice, the variance of the estimator is also estimated from the generated output and hence needs to be stable. Thus, the desirable P* also has a well behaved fourth moment of the estimator (see, e.g., Sadowsky, 1996; Juneja and Shahabuddin, 2002, for further discussion on this).

2.3 Zero-variance measure

Note that an estimator has zero variance if every independent sample generated always equals a constant. In such a case in every simulation run we observe $I(\mathcal{E}) = 1$ and $L = \gamma$. Thus, for $A \subset \mathcal{E}$,

$$\mathrm{P}^*(A) = \frac{\mathrm{P}(A)}{\gamma} \tag{4}$$

and $\mathrm{P}^*(A) = 0$ for $A \subset \mathcal{E}^{\mathrm{c}}$ (for any set H, H^{c} denotes its complement). The zero-variance measure is typically unimplementable as it involves knowledge of γ, the quantity that we are hoping to estimate through simulation. Nonetheless, this measure proves a useful guide in selecting a good implementable importance sampling distribution in many cases. In particular, it suggests that under a good change of measure, the most likely paths to the rare set should be given larger probability compared to the less likely ones and that the relative proportions of the probabilities assigned to the paths to the rare set should be similar to the corresponding proportions under the original measure.

Also note that the zero-variance measure is simply the conditional measure under the original probability conditioned on the occurrence of $\mathcal{E}$, i.e., (4) is equivalent to the fact that

$$\mathrm{P}^*(A) = \frac{\mathrm{P}(A \cap \mathcal{E})}{\mathrm{P}(\mathcal{E})} = \mathrm{P}(A|\mathcal{E})$$

for all events $A \in \Omega$.

2.4 Characterizing good importance sampling distributions

Intuitively, one expects that a change of measure that emphasizes the most likely paths to the rare event (assigns high probability to them) is a good one, as then the indicator function $I(\mathcal{E})$ is one with significant probability and the likelihood ratio is small along these paths as its denominator is assigned a large value. However, even a P* that has such intuitively desirable properties may lead to large and even infinite variance in practice, because on a small set in $\mathcal{E}$ the likelihood ratio may take large values, leading to a blow-up in the second moment and the variance of the estimator (see Glasserman and Kou, 1995; Glasserman and Wang, 1997; Andradottir et al., 1995; Juneja and Shahabuddin, 2001; Randhawa and Juneja, 2004). Thus, it is imperative to closely study the characteristics of good importance sampling distributions. We now discuss the different criteria for evaluating good importance sampling distributions and develop some guidelines for such selections. For this purpose we need a more concrete framework to discuss rare-event simulation.

Consider a sequence of rare events $(\mathcal{E}_b\colon b \geqslant 1)$ and associated probabilities $\gamma_b = \mathrm{P}(\mathcal{E}_b)$ indexed by a rarity parameter b such that $\gamma_b \to 0$ as $b \to \infty$. For example, in a stable single server queue setting, if $\mathcal{E}_b$ denotes the event that the queue length hits level b in a busy cycle, then we may consider the sequence $\gamma_b = \mathrm{P}(\mathcal{E}_b)$ as $b \to \infty$ (in the reliability set-up this discussion may be modified by replacing b with ε, the maximum of failure rates, and considering the sequence of probabilities γ_ε as $\varepsilon \to 0$).

Now consider a sequence of random variables $(Z_b\colon b \geqslant 1)$ such that each Z_b is an unbiased estimator of γ_b under the probability P* (this probability measure may depend upon b). The sequence of estimators $(Z_b\colon b \geqslant 1)$ is said to possess the *bounded relative error property* if

$$\limsup_{b\to\infty} \frac{\sigma_{\mathrm{P}^*}(Z_b)}{\gamma_b} < \infty.$$

It is easy to see that if the sequence of estimators possesses the bounded relative error property, then the number of samples, n, needed to guarantee a fixed relative accuracy remains bounded no matter how small the probability is, i.e., the computational effort is bounded in n for all b.

Example 1. Suppose we need to find $\gamma_b = \mathrm{P}(\mathcal{E}_b)$ for large b through importance sampling as discussed earlier. Let $Z_b = L(b)I(\mathcal{E}_b)$ denote the importance sampling estimator of γ_b under P*, where $L(b)$ denotes the associated likelihood ratio (see (2)). Further suppose that under P*:

(1) $\mathrm{P}^*(\mathcal{E}_b) \geqslant \beta > 0$ for all b.
(2) For each b, the likelihood ratio is constant over sample paths belonging to $\mathcal{E}_b$. Let k_b denote its constant value.

Then, it is easy to see that the estimators $(Z_b\colon b \geqslant 1)$ have bounded relative error. To see this, note that $\gamma_b = \mathrm{E}_{\mathrm{P}^*}(L(b)I(\mathcal{E}_b)) = k_b\mathrm{P}^*(\mathcal{E}_b)$ and $\mathrm{E}_{\mathrm{P}^*}(L(b)^2 I(\mathcal{E}_b)) = k_b^2\mathrm{P}^*(\mathcal{E}_b)$. Recall that

$$\sigma^2_{\mathrm{P}^*}(Z_b) = \mathrm{E}_{\mathrm{P}^*}\bigl(L(b)^2 I(\mathcal{E}_b)\bigr) - \mathrm{E}_{\mathrm{P}^*}\bigl(L(b)I(\mathcal{E}_b)\bigr)^2.$$

Then

$$\frac{\sigma_{\mathrm{P}^*}(Z_b)}{\gamma_b} \leqslant \frac{\sqrt{\mathrm{E}_{\mathrm{P}^*}(L(b)^2 I(\mathcal{E}_b))}}{\gamma_b} \leqslant \frac{1}{\sqrt{\beta}}.$$

The two conditions in Example 1 provide useful insights in finding a good importance sampling distribution, although typically it is difficult to find an implementable P* that has constant likelihood ratios along sample paths to the rare set (Example 8 discusses one such case). Often one finds a distribution such that the likelihood ratios are *almost constant* (see, e.g., Siegmund, 1976; Sadowsky, 1991; Sadowsky and Szpankowski, 1995; Juneja, 2001, and the discussion in Section 4). In such and more general cases, it may be difficult to find a P* that has bounded relative error and we often settle for estimators that are efficient on a 'logarithmic scale'. These are referred to in the literature as *asymptotically optimal* or *asymptotically efficient*. Notable exceptions where P* with bounded relative error are known include rare-event probabilities associated with certain reliability systems (see, e.g., Shahabuddin, 1994) and level crossing probabilities (see, e.g., Asmussen and Rubinstein, 1995). To understand the notion of asymptotic optimality, note that since $\sigma^2_{\mathrm{P}^*}(Z_b) \geqslant 0$ and $\gamma_b = \mathrm{E}_{\mathrm{P}^*}(Z_b)$, it follows that

$$\mathrm{E}_{\mathrm{P}^*}\left(Z_b^2\right) \geqslant \gamma_b^2,$$

and hence $\log(\mathrm{E}_{\mathrm{P}^*}(Z_b^2)) \geqslant 2\log(\gamma_b)$. Since $\log(\gamma_b) < 0$, it follows that

$$\frac{\log(\mathrm{E}_{\mathrm{P}^*}(Z_b^2))}{\log(\gamma_b)} \leqslant 2$$

for all b and for all P*. The sequence of estimators are said to be asymptotically optimal if the above relation holds as an equality in the limit as $b \to \infty$. For example, suppose that $\gamma_b = \mathrm{P}_1(b)\exp(-cb)$ and $\mathrm{E}_{\mathrm{P}^*}(Z_b^2) = \mathrm{P}_2(b)\exp(-2cb)$ where $c > 0$, and $\mathrm{P}_1(\cdot)$ and $\mathrm{P}_2(\cdot)$ are any two polynomial functions of b (of course, $\mathrm{P}_2(b) \geqslant \mathrm{P}_1(b)^2$). The measure P* may be asymptotically optimal, although we may not have bounded relative error.

2.4.1 Uniformly bounded likelihood ratios

In many settings, one can identify a change of measure where the associated likelihood ratio is uniformly bounded along paths to the rare set $\mathcal{E}$ (the subscript b is dropped as we again focus on a single set) by a small constant $k < 1$, i.e.,

$$LI(\mathcal{E}) \leqslant kI(\mathcal{E}).$$

This turns out to be a desirable trait. Note that $\mathrm{E}_{\mathrm{P}^*}(L^2 I(\mathcal{E})) = \mathrm{E}_{\mathrm{P}}(LI(\mathcal{E}))$. Thus,

$$\frac{\sigma^2_{\mathrm{P}^*}(L(I(\mathcal{E}))}{\sigma^2_{\mathrm{P}}(I(\mathcal{E}))} = \frac{\mathrm{E}_{\mathrm{P}}(L(I(\mathcal{E})) - \gamma^2}{\gamma - \gamma^2} \leqslant \frac{k\gamma - \gamma^2}{\gamma - \gamma^2} \leqslant k. \tag{5}$$

Thus, guaranteed variance reduction by at least a factor of k is achieved. Often, a parameterized family of importance sampling distributions can be identified so that the likelihood ratio associated with each distribution in this family is uniformly bounded along paths to the rare set by a constant that may depend on the distribution. Then, a good importance sampling distribution from this family may be selected as the one with the minimum uniform bound. For instance, in the example considered in Section 1.1, it can be seen that the likelihood ratio in (1) is upper bounded by

$$\frac{(1/2)^{100}}{p^{80}(1-p)^{20}}$$

for each $p \geqslant 1/2$ when the experiment is a success, i.e., the number of heads n is greater than or equal to 80 (also see Section 4). Note that this bound is minimized for $p = 0.8$.

In some cases, we may be able to partition the rare event of interest $\mathcal{E}$ into disjoint sets $\mathcal{E}_1, \ldots, \mathcal{E}_J$ such that there exist probability measures $(\mathrm{P}_j^*\colon\ j \leqslant J)$ such that the likelihood ratio $L^{(j)}$ corresponding to each probability measure P_j^* satisfies the relation

$$L^{(j)} \leqslant k_j$$

for a constant $k_j \ll 1$ on the set $\mathcal{E}_j$ (although, the likelihood ratio may be unbounded on other sets). One option then may be to estimate each $\mathrm{P}(\mathcal{E}_j)$ separately using the appropriate change of measure. Sadowsky and Bucklew (1990) propose that a convex combination of these measures may work in estimating $\mathrm{P}(\mathcal{E})$. To see this, let $(p_j\colon\ j \leqslant J)$ denote positive numbers that sum to one, and consider the measure

$$\mathrm{P}^*(\cdot) = \sum_{j \leqslant J} p_j \mathrm{P}_j^*(\cdot).$$

It is easy to see that the likelihood ratio of P w.r.t. P^*, then equals

$$\frac{1}{\sum_{j \leqslant J} p_j / L^{(j)}} \leqslant \max_{j \leqslant J} \frac{k_j}{p_j},$$

so that if the right-hand side is smaller than 1 (which is the case, e.g., if p_j is proportional to k_j and $\sum_{j \leqslant J} k_j < 1$) guaranteed variance reduction may be achieved.

In some cases, under the proposed change of measure, the uniform upper bound on the likelihood ratio is achieved on a substantial part of the rare set and through analysis it is shown that the remaining set has very small probability, so that even large likelihood ratios on this set contribute little to the variance of the estimator (see, e.g., Juneja and Shahabuddin, 2002). This remaining set may be asymptotically negligible so that outputs from it may be ignored (see, e.g., Boots and Shahabuddin, 2001) introducing an asymptotically negligible bias.

3 Rare-event simulation in a Markovian framework

We now specialize our discussion to certain rare events associated with discrete time Markov processes. This framework captures many commonly studied rare events in the literature including those discussed in Sections 4–7.

Consider a Markov process $(S_i\colon\ i \geqslant 0)$ where each S_i takes values in space $\mathcal{S}$ (e.g., $\mathcal{S} = \Re^d$). Often, in rare-event simulation we want to determine the small probability of an event $\mathcal{E}$ determined by the Markov process observed up to a stopping time T, i.e., $(S_0, S_1, \ldots, S_T)$. A random variable (r.v.) T is a stopping time w.r.t. the stochastic process $(S_i\colon\ i \geqslant 0)$ if for any nonnegative integer n, whether $\{T = n\}$ occurs or not can be completely determined by observing $(S_0, S_1, S_2, \ldots, S_n)$. In many cases we may be interested in the probability of a more specialized event $\mathcal{E} = \{S_T \in \mathcal{R}\}$, where $\mathcal{R} \subset \mathcal{S}$ and T denotes the hitting time to a 'terminal' set $\mathcal{T}$, $\mathcal{R} \subset \mathcal{T}$, i.e., $T = \inf\{n\colon\ S_n \in \mathcal{T}\}$. In many cases, the rare-event probability of interest may be reduced to $\mathrm{P}(S_T \in \mathcal{R})$ through state-space augmentation; the latter representation has the advantage that the zero-variance estimator is Markov for this probability. Also, as we discuss in Examples 5 and 6, in a common application, the stopping time under consideration is infinite with large probability and our interest is in estimating $\mathrm{P}(T < \infty)$.

Example 2. The coin tossing example discussed in the introduction fits this framework by setting $T = 100$ and letting $(X_i\colon\ i \geqslant 1)$ be a sequence of i.i.d. random variables where each X_i equals one with probability 0.5 and zero with probability 0.5. Here, $\mathcal{E} = \{\sum_{i=1}^{100} X_i \geqslant 80\}$. Alternatively, let S_n denote the vector $(\sum_{i=1}^{n} X_i, n)$. Let $\mathcal{T}$ denote the event $\{(x, 100)\colon\ x \geqslant 0\}$, $T = \inf\{n\colon\ S_n \in \mathcal{T}\}$ and let $\mathcal{R} = \{(x, 100)\colon\ x \geqslant 80\}$. Then the probability of interest equals $\mathrm{P}(S_T \in \mathcal{R})$.

Note that a similar representation may be obtained more generally for the case where $(X_i\colon\ i \geqslant 1)$ is a sequence of generally distributed i.i.d. random variables, and our interest is in estimating the probability $\mathrm{P}(S_n/n \in \mathcal{R})$ for $\mathcal{R}$ that does not include $\mathrm{E}X_i$ in its closure.

Example 3. The problem of estimating the small probability that the queue length in a stable $M/M/1$ queue hits a large threshold b in a busy cycle (a busy cycle is the stochastic process between the two consecutive times that an arrival to the system finds it empty), fits this framework as follows: Let λ denote the arrival rate to the queue and let μ denote the service rate. Let $p = \lambda/(\lambda + \mu)$. Let S_i denote the queue length after the ith state change (due to an arrival or a departure). Clearly $(S_n\colon\ n \geqslant 0)$ is a Markov process. To denote that the busy cycle starts with one customer we set $S_0 = 1$. If $S_i > 0$, then $S_{i+1} = S_i + 1$ with probability p and $S_{i+1} = S_i - 1$ with probability $1 - p$. Let $T = \inf\{n\colon\ S_n = b \text{ or } S_n = 0\}$. Then $\mathcal{R} = \{b\}$ and the probability of interest equals $\mathrm{P}(S_T \in \mathcal{R})$.

Example 4. The problem of estimating the small probability that the queue length in a stable $GI/GI/1$ queue hits a large threshold b in a busy cycle is important from an applications viewpoint. For instance, Chang et al. (1994) and Heidelberger (1995) discuss how techniques for efficient estimation of this probability may be used to efficiently estimate the steady state probability of buffer overflow in finite-buffer single queues. This probability also fits in our framework, although we need to keep in mind that the queue length process observed at state change instants is no longer Markov and additional variables are needed to ensure the Markov property. Here, we assume that the arrivals and the departures do not occur in batches of two or more. Let $(Q_i\colon\ i \geqslant 0)$ denote the queue-length process observed *just before* the time of state change (due to arrivals or departures). Let J_i equal 1 if the ith state change is due to an arrival. Let it equal 0, if it is due to a departure. Let R_i denote the remaining service time of the customer in service if $J_i = 1$ and $Q_i > 0$. Let it denote the remaining interarrival time if $J_i = 0$. Let it equal zero if $J_i = 1$ and $Q_i = 0$. Then, setting $S_i = (Q_i, J_i, R_i)$, it is easy to see that $(S_i\colon\ i \geqslant 0)$ is a Markov process. Let $T = \inf\{n\colon\ (Q_i, J_i) = (b, 1) \text{ or } (Q_i, J_i) = (1, 0)\}$. Then $\mathcal{R} = \{(b, 1, x)\colon\ x \geqslant 0\}$ and the probability of interest equals $\mathrm{P}(S_T \in \mathcal{R})$.

Example 5. Another problem of importance concerning small probabilities in a $GI/GI/1$ queue setting with first-come-first-serve scheduling rule involves estimation of the probability of large delays in the queue in steady state. Suppose that the zeroth customer arrives to an empty queue and that $(A_0, A_1, A_2, \ldots)$ denotes a sequence of i.i.d. nonnegative r.v.'s where A_n denotes the interarrival time between customer n and $n + 1$. Similarly, let $(B_0, B_1, \ldots)$ denote the i.i.d. sequence of service times in the queue so that the service of customer n is denoted by B_n. Let W_n denote the waiting time of customer n in the queue. Then $W_0 = 0$. The well-known Lindley recursion follows:

$$W_{n+1} = \max(W_n + B_n - A_n, 0)$$

for $n \geqslant 0$ (see, e.g., Asmussen, 2003). We assume that $\mathrm{E}(B_n) < \mathrm{E}(A_n)$, so that the queue is stable and the steady state waiting time distribution exists. Let $Y_n = B_n - A_n$. Then, since $W_0 = 0$, it follows that

$$W_{n+1} = \max(0, Y_n, Y_n + Y_{n-1}, \ldots, Y_n + Y_{n-1} + \cdots + Y_0).$$

Since the sequence $(Y_i\colon\ i \geqslant 0)$ is i.i.d., the right-hand side has the same distribution as

$$\max(0, Y_0, Y_0 + Y_1, \ldots, Y_0 + Y_1 + \cdots + Y_n).$$

In particular, the steady-state delay probability $\mathrm{P}(W_\infty > u)$ equals $\mathrm{P}(\exists n\colon \sum_{i=0}^{n} Y_i > u)$. Let $S_n = \sum_{i=0}^{n} Y_i$ denote the associated random walk with a negative drift. Let $T = \inf\{n\colon\ S_n > u\}$ so that T is a stopping time w.r.t.

$(S_i\colon i \geqslant 0)$. Then $\mathrm{P}(W_\infty > u)$ equals $\mathrm{P}(T < \infty)$. The latter probability is referred to as the level-crossing probability of a random walk. Again, we need to generate $(S_0, S_1, \ldots, S_T)$ to determine whether the event $\{T < \infty\}$ occurs or not. However, we now have an additional complexity that $\mathrm{P}(T = \infty) > 0$ and hence generating $(S_0, S_1, \ldots, S_T)$ may no longer be feasible. Importance sampling resolves this by simulating under a suitable change of measure P^* under which the random walk has a positive drift so that $\mathrm{P}^*(T = \infty) = 0$ (see Siegmund, 1976). This is also discussed in Section 4 in a multidimensional setting when the X_i's have a light-tailed distribution.

Example 6. The problem of estimating ruin probabilities in the insurance sector also fits this framework as follows: Suppose that an insurance company accumulates premiums at a deterministic rate p. Further suppose that the claim interarrival times are an i.i.d. sequence of r.v.'s $(A_1, A_2, \ldots)$. Let $N(t) = \sup\{n\colon \sum_{i=1}^{n} A_i \leqslant t\}$ denote the number of claims that have arrived by time t. Also, assume that the claim sizes are again another i.i.d. sequence of r.v.'s $(B_1, B_2, \ldots)$ independent of the interarrival times (these may be modeled using light or heavy-tailed distributions). Let the initial reserves of the company be denoted by u. In such a model, the wealth of the company at time t is denoted by

$$W(t) = u + pt - \sum_{i=1}^{N(t)} B_i.$$

The probability of eventual ruin therefore equals $\mathrm{P}(\inf_t W(t) \leqslant 0)$. Note that a ruin can occur only at the times of claim arrivals. The wealth at the time of arrival of claim n equals

$$W\left(\sum_{i=1}^{n} A_i\right) = u + p\sum_{i=1}^{n} A_i - \sum_{i=1}^{n} B_i.$$

Let $Y_i = B_i - pA_i$ and $S_n = \sum_{i=1}^{n} Y_i$. The probability of eventual ruin then equals $\mathrm{P}(\max_n S_n > u)$ or equivalently $\mathrm{P}(T < \infty)$, where $T = \inf\{n\colon S_n > u\}$. Hence, the discussion at the end of Example 5 applies here as well.

Example 7 (Highly reliable Markovian systems). These reliability systems have components that fail and repair in a Markovian manner, i.e., they have exponentially distributed failure and repair times. High reliability is achieved due to the highly reliable nature of the individual components comprising the system. Complex system interdependencies may be easily modeled in the Markov framework. These interdependencies may include failure propagation, i.e., failure of one component with certain probability leads to failure of other components. They may also include other features such as different modes of component failure, repair and operational dependencies, component switch-over times, etc. See, e.g., Goyal and Lavenberg (1987) and Goyal et al. (1992) for further discussion on such modeling complexities.

A mathematical model for such a system may be built as follows: Suppose that the system has d distinct component-types. Each component type i has m_i identical components for functional and spare requirements. Let λ_i and μ_i denote the failure and repair rate, respectively, for each of these components. The fact that each component is highly reliable is modeled by letting $\lambda_i = \Theta(\varepsilon^{r_i})$[3] for $r_i \geqslant 1$, and letting $\mu_i = \Theta(1)$. The system is then analyzed as $\varepsilon \to 0$.

Let $(Y(t)\colon\ t \geqslant 0)$ be a continuous time Markov chain (CTMC) of this system, where $Y(t) = (Y_1(t), Y_2(t), \ldots, Y_d(t), R(t))$. Here, each $Y_i(t)$ denotes the number of failed components of type i at time t. The vector $R(t)$ contains all configurational information required to make $(Y(t)\colon\ t \geqslant 0)$ a Markov process. For example, it may contain information regarding the order in which the repairs occur, the failure mode of each component, etc. Let $\mathcal{A}$ denote the state when all components are 'up' (let it also denote the set containing this state). Let $\mathcal{R}$ denote the set of states deemed as failed states. This may be a rare set for small values of ε. The probability that the system starting from state $\mathcal{A}$, hits the set $\mathcal{R}$ before returning to $\mathcal{A}$ is important for these highly reliable systems as this is critical to efficient estimation of performance measures such as system unavailability and mean time to failure. Let $(S_i\colon\ i \geqslant 0)$ denote the discrete time Markov chain (DTMC) embedded in $(Y(t)\colon\ t \geqslant 0)$. For estimating this probability, the DTMC may be simulated instead of the CTMC as both give identical results. Set $S_0 = \mathcal{A}$. Then, the process $(S_1, \ldots, S_T)$ may be observed where $T = \inf\{n \geqslant 1\colon\ S_n \in \mathcal{T}\}$, where $\mathcal{T} = \mathcal{A} \cup \mathcal{R}$. The set $\mathcal{E}$ equals $\{S_T \in \mathcal{R}\}$.

In this chapter we do not pursue highly reliable systems further. Instead we refer the reader to Heidelberger (1995) and Nakayama et al. (2001) for surveys on this topic.

3.1 Importance sampling in a Markovian framework

Let P_n denote the probability P restricted to the events associated with $(S_0, S_1, \ldots, S_n)$ for $n = 1, 2, \ldots.$ Then

$$\gamma := \mathrm{P}(\mathcal{E}) = \sum_n \mathrm{P}_n(\mathcal{E}_n),$$

where $\mathcal{E}_n = \mathcal{E} \cap \{T = n\}$. Consider another distribution P^* and let P^*_n denote its restriction to the events associated with $(S_0, S_1, \ldots, S_n)$ for $n = 1, 2, \ldots.$ Suppose that for each n, $\mathrm{P}^*_n(A_n) > 0$ whenever $\mathrm{P}_n(A_n) > 0$ for $A_n \subset \mathcal{E}_n$.

[3] A nonnegative function $f(\varepsilon)$ is said to be $\mathrm{O}(\varepsilon^r)$ for $r \geqslant 0$ if there exists a positive constant K such that $f(\varepsilon) \leqslant K\varepsilon^r$ for all ε sufficiently small. It is said to be $\Theta(\varepsilon^r)$ for $r \geqslant 0$ if there exist positive constants K_1 and K_2 ($K_1 < K_2$), such that $K_1\varepsilon^r \leqslant f(\varepsilon) \leqslant K_2\varepsilon^r$ for all ε sufficiently small.

Then, proceeding as in (2),

$$\mathrm{P}(\mathcal{E}) = \sum_n \int_{\mathcal{E}_n} L_n \, \mathrm{d}P_n^*,$$

where $L_n = \frac{\mathrm{dP}_n}{\mathrm{dP}_n^*}$. For example, if the sequence $(S_0, S_1, \ldots, S_n)$ has a density function $f_n(\cdot)$ for each n under P ($f_n^*(\cdot)$ under P*) such that $f_n^*(x_0, x_1, \ldots, x_n) > 0$ whenever $f_n(x_0, x_1, \ldots, x_n) > 0$, then

$$L_n(S_0, S_1, \ldots, S_n) = \frac{f_n(S_0, S_1, \ldots, S_n)}{f_n^*(S_0, S_1, \ldots, S_n)} \tag{6}$$

for each n a.s.

Thus, $\gamma = \mathrm{E}_{\mathrm{P}^*}(L_T I(\mathcal{E}))$ where $\mathrm{E}_{\mathrm{P}^*}$ is an expectation operator under the probability P*. To further clarify the discussion, we illustrate the form of the likelihood ratio for Examples 3 and 4.

Example 8. In Example 3, suppose the queue is simulated under a probability P* under which it is again an $M/M/1$ queue with arrival rate λ^* and service rate μ^*. Let $p^* = \lambda^*/(\lambda^* + \mu^*)$. Consider a sample path $(S_0, S_1, \ldots, S_T)$ that belongs to $\mathcal{E}$, i.e., $\{S_T \in \mathcal{R}\}$. Let N_A denote the number of arrivals and N_S denote the number of service completions up to time T along this sample path. Thus, $N_A = b + N_S - 1$ where b denotes the buffer size. The likelihood ratio L_T along this path therefore equals

$$\left(\frac{p}{p^*}\right)^{N_A} \left(\frac{1-p}{1-p^*}\right)^{N_S}.$$

In the case $\lambda < \mu$, it can be seen that $\lambda^* = \mu$ and $\mu^* = \lambda$ achieves the two conditions discussed in Example 1 (with $k_b = (\lambda/\mu)^{b-1}$) and hence the associated importance sampling distribution has the bounded relative error property.

Example 9. In Example 4, let $f(\cdot)$ and $g(\cdot)$ denote the p.d.f. of the interarrival times and the service times, respectively under the probability P. Let P* be another probability under which the queue remains a $GI/GI/1$ queue with the new p.d.f.'s for interarrival and service times denoted by $f^*(\cdot)$ and $g^*(\cdot)$, respectively. Consider a sample path $(S_0, S_1, \ldots, S_T)$ that belongs to $\mathcal{E}$, i.e., $\{Q_T = b\}$. Let N_A denote the number of arrivals and N_B denote the number of service initiations up to time T along this sample path. Let $(A_1, A_2, \ldots, A_{N_A})$ denote the N_A interarrival times generated and let $(B_1, B_2, \ldots, B_{N_B})$ denote the N_B service times generated along this sample path. The likelihood ratio L_T along this path therefore equals

$$\prod_{i=1}^{N_A} \frac{f(A_i)}{f^*(A_i)} \prod_{i=1}^{N_B} \frac{g(B_i)}{g^*(B_i)}.$$

Thus, from the simulation viewpoint the computation of the likelihood ratio in Markovian settings is straightforward and may be done iteratively as follows: Before generation of a sample path of $(S_0, S_1, \ldots, S_T)$ under the new probability, the likelihood ratio may be initialized to 1. Then, it may be updated at each transition by multiplying it with the ratio of the original probability density function of the newly generated sample(s) at that transition and the new probability density function of this sample(s). The probability density function may be replaced by the probability values when discrete random variables are involved.

3.2 Zero-variance measure in Markovian settings

For probabilities such as $\mathrm{P}(S_T \in \mathcal{R})$, the zero-variance measure has a Markovian representation. For $\mathcal{E} = \{S_T \in \mathcal{R}\}$, let $\mathrm{P}_x(\mathcal{E})$ denote the probability of this event, conditioned on $S_0 = x$. Recall that $T = \inf\{n\colon S_n \in \mathcal{T}\}$. For simplicity suppose that the state space $\mathcal{S}$ of the Markov chain is finite (the following discussion is easily extended to more general state spaces) and let $P = (p_{xy}\colon x, y \in \mathcal{S})$ denote the associated transition matrix. In this setting,

$$\mathrm{P}_x(\mathcal{E}) = \sum_{y\in\mathcal{R}} p_{xy} + \sum_{y\in\mathcal{S}-\mathcal{T}} p_{xy}\mathrm{P}_y(\mathcal{E}).$$

Thus, $p^*_{xy} = p_{xy}/\mathrm{P}_x(\mathcal{E})$ for $y \in \mathcal{R}$ and $p^*_{xy} = p_{xy}\mathrm{P}_y(\mathcal{E})/\mathrm{P}_x(\mathcal{E})$ for $y \in \mathcal{S} - \mathcal{T}$ is a valid transition probability. It is easy to check that in this case

$$L_T = \frac{p_{S_0,S_1} p_{S_1,S_2} \cdots p_{S_{T-1},S_T}}{p^*_{S_0,S_1} p^*_{S_1,S_2} \cdots p^*_{S_{T-1},S_T}}$$

equals $\mathrm{P}_{S_0}(\mathcal{E})$ a.s., i.e., the associated P^* is the zero-variance measure. The problem again is that determining p^*_{xy} requires knowledge of $\mathrm{P}_x(\mathcal{E})$ for all $x \in \mathcal{S}$.

Consider the probability $\mathrm{P}(S_n/n \geqslant a)$, where $S_n = \sum_{i\leqslant n} X_i$, the $(X_i\colon i \geqslant 0)$ are i.i.d. r.v.'s taking values in $\Re$, and $a > \mathrm{E}X_i$. From the above discussion and using the associated augmented Markov chain discussed at the end of Example 2, it can be seen that the zero-variance measure conditioned on the event that $S_m = s_m < na$, $m < n$, has transition probabilities

$$p^*_{m,s_m}(y) = \mathrm{P}(X_{m+1} = y)\frac{\mathrm{P}(S_n \geqslant na | S_{m+1} = s_m + y)}{\mathrm{P}(S_n \geqslant na | S_m = s_m)}.$$

More generally,

$$\begin{aligned}&\mathrm{P}^*(X_{m+1} \in \mathrm{d}y | S_m = s_m)\\ &\quad = \mathrm{P}(X_{m+1} \in \mathrm{d}y)\frac{\mathrm{P}(S_{n-m-1} \geqslant na - s_m - y)}{\mathrm{P}(S_{n-m} \geqslant na - s_m)}.\end{aligned} \qquad (7)$$

Such an explicit representation of the zero-variance measure proves useful in adaptive algorithms where one adaptively learns the zero-variance measure (see Section 5). This representation is also useful in developing simpler implementable importance sampling distributions that are in an asymptotic sense close to this measure (see Section 3.3).

3.3 Exponentially twisted distributions

Again consider the probability $\mathrm{P}(S_n/n \geqslant a)$. Let $\Psi(\cdot)$ denote the log-moment generating function of X_i, i.e., $\Psi(\theta) = \log \mathrm{E}(\exp(\theta X_i))$. Let $\Theta = \{\theta : \Psi(\theta) < \infty\}$. Suppose that Θ^{o} (for any set H, H^{o} denotes its interior) contains the origin, so that X_i has a light-tailed distribution. For $\theta \in \Theta^{\mathrm{o}}$, consider the probability P_θ under which the $(X_i\colon\ i \geqslant 1)$ are i.i.d. and

$$\mathrm{P}_\theta(X_i \in \mathrm{d}y) = \exp\bigl(\theta y - \Psi(\theta)\bigr)\,\mathrm{P}(X_i \in \mathrm{d}y).$$

This is referred to as the probability obtained by exponentially twisting the original probability by θ. We now show that the distribution of X_{m+1} conditioned on $S_m = s_m$ under the zero-variance measure for the probability $\mathrm{P}(S_n/n \geqslant a)$ (shown in (7)) converges asymptotically (as $n \to \infty$) to a suitable exponentially twisted distribution independent of s_m, thus motivating the use of such distributions for importance sampling of constituent r.v.'s in random walks in complex stochastic processes.

Suppose that $\theta_a \in \Theta^{\mathrm{o}}$ solves the equation $\Psi'(\theta) = a$. In that case, when the distribution of X_i is nonlattice, the following exact asymptotic is well known (see Bahadur and Rao, 1960; Dembo and Zeitouni, 1998):

$$\mathrm{P}\left(\frac{S_n}{n} \geqslant a + \frac{k}{n} + \mathrm{o}\left(\frac{1}{n}\right)\right) \sim \frac{c}{\sqrt{n}} \exp\bigl[-n\bigl(\theta_a a - \Psi(\theta_a)\bigr) - \theta_a k\bigr], \quad (8)$$

where $c = 1/(\sqrt{2\pi\Psi''(\theta_a)}\theta_a)$ ($a_n \sim b_n$ means that $a_n/b_n \to 1$ as $n \to \infty$) and k is a constant. Usually, the exact asymptotic is developed for $\mathrm{P}(S_n/n \geqslant a)$. The minor generalization in (8) is discussed, e.g., in Borkar et al. (2004). This exact asymptotic may be inaccurate if n is not large enough especially for certain sufficiently 'nonnormal' distributions of X_i. In such cases, simulation using importance sampling may be a desirable option to get accurate estimates.

Using (8) in (7) as $n \to \infty$, for a fixed m, it can be easily seen that

$$\lim_{n\to\infty} \mathrm{P}^*(X_{m+1} \in \mathrm{d}y | S_m = s_m) = \mathrm{P}(X_{m+1} \in \mathrm{d}y) \exp\bigl(\theta_a y - \Psi(\theta_a)\bigr),$$

i.e., asymptotically the zero-variance measure converges to P_{θ_a}. This suggests that P_{θ_a} may be a good importance sampling distribution to estimate $\mathrm{P}(S_n/n \geqslant a)$ for large n. We discuss this further in Section 4. Also, it is easily seen through differentiation that the mean of X_i under P_θ equals $\Psi'(\theta)$. In particular, under P_{θ_a}, the mean of X_i equals a, so that $\{S_n/n \geqslant a\}$ is no longer a rare event.

4 Large deviations of multidimensional random walks

In this section we focus on efficient estimation techniques for two rare-event probabilities associated with multidimensional random walks, namely: (1) the probability that the random walk observed after a large time period n, lies in a rare set; (2) the probability that the random walk ever hits a rare set. We provide a heuristic justification for the large deviations asymptotic in the two cases and identify the asymptotically optimal changes of measures. We note that the ideas discussed earlier greatly simplify the process of identifying a good change of measure. These include restricting the search for the change of measure to those obtained by exponentially twisting the original measure, selecting those that have constant (or almost constant) likelihood ratios along paths to the rare set, or selecting those whose likelihood ratios along such paths have the smallest uniform bound.

4.1 Random walk in a rare set

Consider the probability $\mathrm{P}(S_n/n \in \mathcal{R})$, where $S_n = \sum_{i=1}^{n} X_i$, the X_i's are i.i.d. and each X_i is a random column vector taking values in $\Re^d$. Thus, $X_i = (X_{i1}, \dots, X_{id})^{\mathrm{T}}$ where the superscript "T" denotes the transpose operation. The set $\mathcal{R} \subset \Re^d$ and its closure does not include $\mathrm{E}X_i$. The essential ideas for this discussion are taken from Sadowsky and Bucklew (1990) (also see Sadowsky, 1996) where this problem is studied in a more general framework. We refer the reader to these references for rigorous analysis, while the discussion here is limited to illustrating the key intuitive ideas in a simple setting.

For simplicity suppose that the log moment generating function,

$$\Psi(\theta) = \log \mathrm{E}\bigl(\exp\bigl(\theta^{\mathrm{T}} X\bigr)\bigr),$$

exists for each column vector $\theta \in \Re^d$. This is true, e.g., when X_i is bounded or has a multivariate Gaussian distribution. Further suppose that X_i is nondegenerate, i.e., it is not a.s. constant in any dimension. Define the associated rate function

$$J(\alpha) = \sup_{\theta}\bigl(\theta^{\mathrm{T}}\alpha - \Psi(\theta)\bigr)$$

for $\alpha \in \Re^d$. Note that for each θ, $\theta^{\mathrm{T}}\alpha - \Psi(\theta)$ is a convex function of α, hence, $J(\cdot)$ being a supremum of convex functions, is again convex. It can be shown that it is strictly convex in the interior $\mathcal{J}^{\mathrm{o}}$, where

$$\mathcal{J} = \bigl\{\alpha\colon J(\alpha) < \infty\bigr\}.$$

From large deviations theory (see, e.g., Dembo and Zeitouni, 1998), we see that

$$\mathrm{P}\left(\frac{S_n}{n} \approx a\right) \approx \exp\bigl(-nJ(a)\bigr). \tag{9}$$

Here, $S_n/n \approx a$ may be taken to be the event that S_n/n lies in a small ball of radius ε centered at a. The relation (9) becomes an equality when an appropriate $\mathrm{O}(\varepsilon)$ term is added to $J(a)$ in the exponent in the right-hand side. It is instructive to heuristically see this result. Note that

$$\mathrm{P}\left(\frac{S_n}{n} \approx a\right) = \int_{x \approx na} \mathrm{d}F_n(x),$$

where $F_n(\cdot)$ denotes the distribution function (d.f.) of S_n (obtained by convolution of the d.f. of X_i n times). Let $F_\theta(\cdot)$ denote the d.f. obtained by exponentially twisting $F(\cdot)$ by θ, i.e.,

$$\mathrm{d}F_\theta(x) = \exp\bigl(\theta^{\mathrm{T}} x - \Psi(\theta)\bigr)\,\mathrm{d}F(x).$$

It follows (heuristically speaking) that

$$\mathrm{P}\left(\frac{S_n}{n} \approx a\right) \approx \exp\bigl[-n\bigl(\theta^{\mathrm{T}} a - \Psi(\theta)\bigr)\bigr]\,\mathrm{P}_\theta\left(\frac{S_n}{n} \approx a\right), \tag{10}$$

where P_θ denotes the probability induced by $F_\theta(\cdot)$. Since the left-hand side is independent of θ, for large n it is plausible that the θ which maximizes $\mathrm{P}_\theta(S_n/n \approx a)$, also maximizes $\theta^{\mathrm{T}} a - \Psi(\theta)$. Clearly, for large n, $\mathrm{P}_\theta(S_n/n \approx a)$ is maximized by $\tilde{\theta}_a$ such that $\mathrm{E}_{\tilde{\theta}_a} X_i = a$, so that by the law of large numbers this probability tends to 1 as $n \to \infty$ (E_θ denotes the expectation under the measure P_θ). Indeed

$$\theta_a = \arg\max_\theta\bigl(\theta^{\mathrm{T}} a - \Psi(\theta)\bigr),$$

uniquely satisfies the relation $\mathrm{E}_{\theta_a} X_i = a$. To see this note that θ_a is the solution to $\nabla\Psi(\theta) = a$ (it can be shown that such a θ_a uniquely exists for each $a \in \mathcal{J}^{\mathrm{o}}$). Also via differentiation, it is easily checked that for each θ,

$$\mathrm{E}_\theta X_i = \nabla\Psi(\theta).$$

In particular, $J(a) = \theta_a^{\mathrm{T}} a - \Psi(\theta_a)$ and (9) follows from (10).

For any set H, let $\overline{H}$ denote its closure. Define the rate function of the set $\mathcal{R}$,

$$J(\mathcal{R}) = \inf_{\alpha \in \mathcal{R}} J(\alpha).$$

For $\mathcal{R}$ that is sufficiently 'nice' so that $\overline{\mathcal{R}} = \overline{\mathcal{R}^{\mathrm{o}}}$ (e.g., in two dimensions $\mathcal{R}$ does not contain any isolated points or lines) and $\mathcal{R} \cap \mathcal{J}^{\mathrm{o}} \neq \emptyset$ so that there exist open intervals in $\mathcal{R}$ that can be reached with positive probability, the following large deviations relation holds

$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{P}\left(\frac{S_n}{n} \in \mathcal{R}\right) = -J(\mathcal{R}). \tag{11}$$

Note that there exists a point a^* on the boundary of $\overline{\mathcal{R}}$ such that $J(a^*) = J(\mathcal{R})$. Such an a^* is referred to as a minimum rate point. Intuitively, (11) may be seen quite easily when $\mathcal{R}$ is compact. Loosely speaking, the lower bound follows since, $\mathrm{P}(S_n/n \in \mathcal{R}) \geqslant \mathrm{P}(S_n/n \approx a^*)$ (where, in this special case, $S_n/n \approx a$ may be interpreted as the event that S_n/n lies in the intersection of $\mathcal{R}$ and a small ball of radius ε centered at a). Now if one thinks of $\mathcal{R}$ as covered by a finite number $m(\varepsilon)$ balls of radius ε centered at $(a^*, a_2, \ldots, a_{m(\varepsilon)})$, then

$$\begin{aligned}\mathrm{P}\left(\frac{S_n}{n} \in \mathcal{R}\right) &\leqslant \mathrm{P}\left(\frac{S_n}{n} \approx a^*\right) + \sum_{i=2}^{m(\varepsilon)} \mathrm{P}\left(\frac{S_n}{n} \approx a_i\right)\\ &\overset{(\approx)}{\leqslant} m(\varepsilon)\exp\bigl(-nJ\bigl(a^*\bigr)\bigr) = m(\varepsilon)\exp\bigl(-nJ(\mathcal{R})\bigr)\end{aligned}$$

and thus (11) may be expected.

Recall that from zero-variance estimation considerations, the new change of measure should assign high probability to the neighborhood of a^*. This is achieved by selecting $F_{\theta_{a^*}}(\cdot)$ as the IS distribution (since $\mathrm{E}_{\theta_{a^*}}(X_i) = a^*$). However, this may cause problems if the corresponding likelihood ratio

$$L_n = \exp\bigl(-n\bigl(\theta_{a^*}^{\mathrm{T}} x - \Psi(\theta_{a^*})\bigr)\bigr)$$

becomes large for some $x \in \mathcal{R}$, i.e., some points are assigned insufficient probability under $F_{\theta_{a^*}}(\cdot)$.

If all $x \in \mathcal{R}$ have the property that

$$\theta_{a^*}^{\mathrm{T}} x \geqslant \theta_{a^*}^{\mathrm{T}} a^*, \tag{12}$$

then the likelihood ratio for all $x \in \mathcal{R}$ is uniformly bounded by

$$\exp\bigl(-n\bigl(\theta_{a^*}^{\mathrm{T}} a^* - \Psi(\theta_{a^*})\bigr)\bigr) = \exp\bigl(-nJ(\mathcal{R})\bigr).$$

Hence $\mathrm{P}(S_n/n \in \mathcal{R}) = \mathrm{E}_{\theta_{a^*}}(L_n I(\mathcal{R})) \leqslant \exp(-nJ(\mathcal{R}))$ and $\mathrm{E}_{\theta_{a^*}}(L_n^2 I(\mathcal{R})) \leqslant \exp(-2nJ(\mathcal{R}))$ so that asymptotic optimality of $F_{\theta_{a^*}}(\cdot)$ follows.

The relation (12) motivates the definition of a *dominating point* (see Ney, 1983; Sadowsky and Bucklew, 1990). A minimum rate point a^* is a dominating point of the set $\mathcal{R}$ if

$$\mathcal{R} \subset \mathcal{H}\bigl(a^*\bigr) = \bigl\{x\colon\ \theta_{a^*}^{\mathrm{T}} x \geqslant \theta_{a^*}^{\mathrm{T}} a^*\bigr\}.$$

Recall that

$$J(a) = \theta_a^{\mathrm{T}} a - \Psi(\theta_a)$$

for $a \in \mathcal{J}^{\mathrm{o}}$. Thus, differentiating with respect to a component-wise and noting that $\theta_{a^*} = \nabla\Psi(\theta_{a^*})$ it follows that $\nabla J(a^*) = \theta_{a^*}$. Hence $\nabla J(a^*)$ is orthogonal to the plane $\theta_{a^*}^{\mathrm{T}} x = \theta_{a^*}^{\mathrm{T}} a^*$. In particular, this plane is tangential to the level set $\{x\colon J(x) = J(a^*)\}$. Clearly, if $\mathcal{R}$ is a convex set, we have $\mathcal{R} \subset \mathcal{H}(a^*)$. Of course, as Figure 1 indicates, this is by no means necessary. Figure 2 illustrates the case where $\mathcal{R}$ is not a subset of $\mathcal{H}(a^*)$. Even, in this case, $F_{\theta_{a^*}}(\cdot)$ may be asymptotically optimal if the region in $\mathcal{R}$ where the likelihood ratio is large has sufficiently small probability. Fortunately, in this more general setting, in

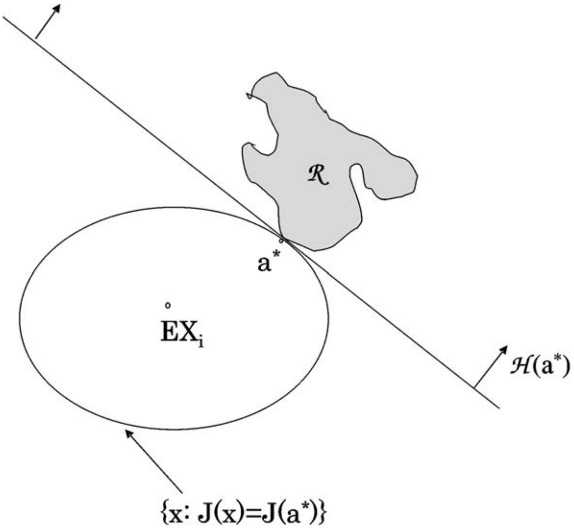

Fig. 1. Set with a dominating point a^*.

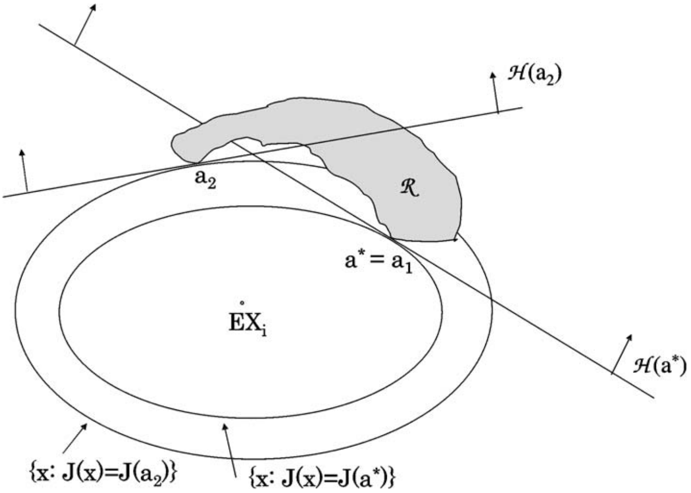

Fig. 2. Set with a minimum rate point a^* which is not a dominating point. Two points (a^*, a_2) are required to cover $\mathcal{R}$ with $\mathcal{H}(a^*)$ and $\mathcal{H}(a_2)$. Note that $J(a_2) > J(a^*)$ so that a_2 is not a minimum rate point.

Sadowsky and Bucklew (1990), sufficient conditions for asymptotic optimality are proposed that cover far more general sets $\mathcal{R}$. These conditions require existence of points $(a_1, \ldots, a_m) \subset \mathcal{J}^o \cap \overline{\mathcal{R}}$ such that $\overline{\mathcal{R}} \subset \bigcup_{i=1}^m \mathcal{H}(a_i)$. Then for any positive numbers $(p_i\colon i \leqslant m)$ such that $\sum_{i \leqslant m} p_i = 1$, the distribution $F^*(\cdot) = \sum_{i \leqslant m} p_i F_{\theta_{a_i}}(\cdot)$ asymptotically optimally estimates $\mathrm{P}(S_n/n \in \mathcal{R})$.

Note that from an implementation viewpoint, generating S_n from the distribution F^* corresponds to generating a r.v. k from the discrete distribution $(p_1, \ldots, p_m)$ and then generating $(X_1, \ldots, X_n)$ using the distribution $F_{\theta_{a_k}}$ to independently generate each of the X_i's.

The fact that F^* is indeed a good importance sampling distribution is easy to see as the corresponding likelihood ratio (F w.r.t. F^*) equals

$$\begin{aligned}\frac{1}{\sum_{i\leqslant m} p_i \exp[n(\theta_{a_i}^{\mathrm{T}} x) + \Psi(\theta_{a_i})]} &\leqslant \frac{\exp[-n(\theta_{a_i}^{\mathrm{T}} x) - \Psi(\theta_{a_i})]}{p_i} \\ &\leqslant \frac{\exp[-nJ(a_i)]}{p_i},\end{aligned}$$

where the upper bound holds for any choice of i. This in turn is upper bounded by

$$\frac{\exp[-nJ(a^*)]}{\min_i p_i}.$$

For large n, this is a uniform upper bound assuring guaranteed variance reduction. It follows that

$$\lim_{n\to\infty} \frac{1}{n} \log \mathrm{E}_{\mathrm{P}^*} L^2 I(\mathcal{R}) \leqslant -2J(\mathcal{R})$$

assuring asymptotic optimality of P^*.

4.2 Probability of hitting a rare set

Let $T_\delta = \inf\{n\colon\ \delta S_n \in \mathcal{R}\}$. We now discuss efficient estimation techniques for the probability $\mathrm{P}(T_\delta < \infty)$ as $\delta \downarrow 0$. This problem generalizes the level crossing probability in the one-dimensional setting discussed by Siegmund (1976) and Asmussen (1989). Lehtonen and Nyrhinen (1992a, 1992b) considered the level crossing problem for Markov-additive processes. (Recall that Examples 5 and 6 also consider this.) Collamore (2002) considered the problem for Markov-additive processes in general state spaces. Again, we illustrate some of the key ideas for using importance sampling for this probability in the simple framework of S_n being a sum of i.i.d. random variables taking values in $\Re^d$, when $\overline{\mathcal{R}^o} = \overline{\mathcal{R}}$.

Note that the central tendency of the random walk S_n is along the ray $\lambda \mathrm{E}X_i$ for $\lambda \geqslant 0$. We further assume that $\mathrm{E}X_i$ does not equal zero and that $\mathcal{R}$ is disjoint with this ray, in the sense that

$$\mathcal{R} \cap \{\lambda x\colon\ \lambda > 0, x \approx \mathrm{E}X_i\} = \emptyset,$$

where $x \approx \mathrm{E}X_i$ means that x lies in a ball of radius $\varepsilon > 0$ centered at $\mathrm{E}X_i$, for some ε. Thus, $\mathrm{P}(T_\delta < \infty)$ is a rare event as $\delta \downarrow 0$. Figure 3 graphically illustrates this problem.

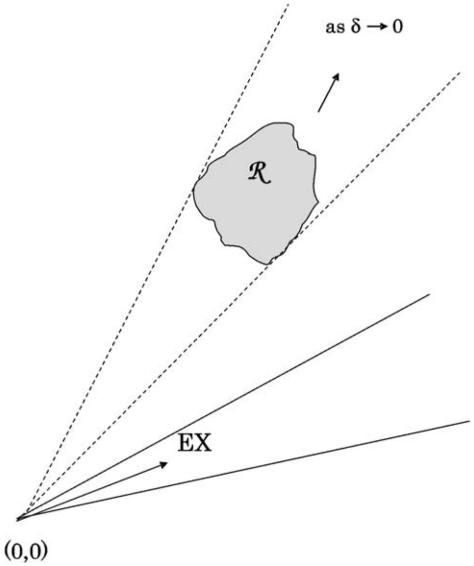

Fig. 3. Estimating the probability that the random walk hits the rare set.

First we heuristically arrive at the large deviations approximation for $P(T_\delta < \infty)$ (see Collamore, 1996, for a rigorous analysis). Let

$$T_\delta(a) = \inf\{n\colon\ \delta S_n \approx a\},$$

where again $\delta S_n \approx a$ may be taken to be the event that δS_n lies in a small ball of radius ε centered at a.

Again, under importance sampling suppose that each X_i is generated using the twisted distribution F_θ. Then the likelihood ratio along $\{T_\delta(a) < \infty\}$ up till time $T_\delta(a)$ equals (approximately)

$$\exp\left[-\theta^{\mathrm{T}}\frac{a}{\delta} + T_\delta(a)\Psi(\theta)\right].$$

Suppose that θ is restricted to the set $\{\theta\colon\ \Psi(\theta) = 0\}$. This ensures that the likelihood ratio is *almost* constant. Thus, for such a θ we may write

$$P\bigl(T_\delta(a) < \infty\bigr) \approx \exp\left[-\theta^{\mathrm{T}}\frac{a}{\delta}\right] P_\theta\bigl(T_\delta(a) < \infty\bigr).$$

Again, the left-hand side is independent of θ so that $\tilde{\theta}$ that maximizes $P_\theta(T_\delta(a) < \infty)$ as $\delta \to 0$ should also maximize $\theta^{\mathrm{T}}a$ subject to $\Psi(\theta) = 0$. Intuitively, one expects such a $\tilde{\theta}$ to have the property that the ray $\lambda E_{\tilde{\theta}}(X_i)$ for $\lambda \geqslant 0$ intersects a, so that the central tendency of the random walk under $F_{\tilde{\theta}}$ is towards a. This may also be seen from the first-order conditions for the relaxed concave programming problem: Maximize $\theta^{\mathrm{T}}a$ subject to $\Psi(\theta) \leqslant 0$.

(It can be seen that the solution to the relaxed problem θ_a also satisfies the original constraint $\Psi(\theta) = 0$.) These amount to the existence of a scalar $\lambda > 0$ such that

$$\nabla\Psi(\theta_a) = \lambda a$$

(see, e.g., Luenberger, 1984).

We now heuristically argue that $\mathrm{P}_{\theta_a}(T_\delta(a) < \infty) \to 1$ as $\delta \to 0$. Under P_{θ_a}, from the central limit theorem,

$$S_n \approx n\mathrm{E}_{\theta_a} X_i + \sqrt{n}\,\mathrm{N}(0, C),$$

where $\mathrm{E}_{\theta_a} X_i = \lambda a$ denotes its drift and C denotes the covariance matrix of the components of X_i. In particular,

$$\delta S_{\lfloor \frac{1}{\lambda\delta} \rfloor} \approx a + \sqrt{\frac{\delta}{\lambda}}\,\mathrm{N}(0, C)$$

and this converges to a as $\delta \to 0$ suggesting that $\mathrm{P}_{\theta_a}(T_\delta(a) < \infty) \to 1$.

Thus heuristically,

$$\mathrm{P}\bigl(T_\delta(a) < \infty\bigr) \approx \exp\left[-\theta_a^{\mathrm{T}} \frac{a}{\delta}\right]$$

and

$$\mathrm{P}\bigl(T_\delta(\mathcal{R}) < \infty\bigr) \approx \exp\left[-\frac{H(\mathcal{R})}{\delta}\right],$$

where

$$H(\mathcal{R}) = \inf_{a \in \mathcal{R}} \theta_a^{\mathrm{T}} a = \inf_{a \in \mathcal{R}} \sup_{(\theta: \Psi(\theta)=0)} \theta^{\mathrm{T}} a.$$

Specifically, the following result may be derived

$$\lim_{\delta \to 0} \delta \log \mathrm{P}\bigl(T_\delta(\mathcal{R}) < \infty\bigr) = -H(\mathcal{R}) \tag{13}$$

(see Collamore, 1996). Suppose that there exists an $a^* \in \overline{\mathcal{R}}$ such that $H(\mathcal{R}) = \theta_{a^*}^{\mathrm{T}} a^*$. It is easy to see that such an a^* must be an *exposed point*, i.e., the ray $\{va^*\colon\ 0 \leqslant v < 1\}$ does not touch any point of $\mathcal{R}$. Furthermore, suppose that

$$\mathcal{R} \subset \mathcal{H}\bigl(a^*\bigr) \triangleq \bigl\{x\colon\ \theta_{a^*}^{\mathrm{T}} x \geqslant \theta_{a^*}^{\mathrm{T}} a^*\bigr\}.$$

Then, the likelihood ratio of F w.r.t. $F_{\theta_{a^*}}$ up till time $T_\delta(\mathcal{R})$ equals

$$\exp\bigl(-\theta_{a^*}^{\mathrm{T}} S_{T_\delta(\mathcal{R})}\bigr) \leqslant \exp\left(-\theta_{a^*}^{\mathrm{T}} \frac{a^*}{\delta}\right).$$

Thus, we observe guaranteed variance reduction while simulating under $F_{\theta_{a^*}}$ (note that $\mathrm{P}_{\theta_{a^*}}(T_\delta(\mathcal{R}) < \infty) \to 1$ as $\delta \to 0$). In addition, it follows that

$$\lim_{\delta \to 0} \delta \log \mathrm{E} L^2_{T_\delta(\mathcal{R})} I\bigl(T_\delta(\mathcal{R}) < \infty\bigr) \leqslant -2H(\mathcal{R}).$$

The above holds as an equality in light of (13), proving that $F_{\theta_{a^*}}$ ensures asymptotic optimality.

Again, as in the previous subsection, suppose that $\mathcal{R}$ is not a subset of $\mathcal{H}(a^*)$, and there exist points $(a_1, \ldots, a_m) \subset \overline{\mathcal{R}}$ $(a^* = a_1)$ such that $\overline{\mathcal{R}} \subset \bigcup_{i=1}^{m} \mathcal{H}(a_i)$. Then, for any positive numbers $(p_i\colon i \leqslant m)$ such that $\sum_{i \leqslant m} p_i = 1$, the distribution $F^*(\cdot) = \sum_{i \leqslant m} p_i F_{\theta_{a_i}}(\cdot)$ asymptotically optimally estimates $\mathrm{P}(T_\delta(\mathcal{R}) < \infty)$.

5 Adaptive importance sampling techniques

In this section we restrict our basic Markov process $(S_i\colon i \geqslant 0)$ to a finite state space $\mathcal{S}$. We associate a one-step transition reward $g(x, y) \geqslant 0$ with each transition $(x, y) \in \mathcal{S}^2$ and generalize our performance measure to that of estimating the expected cumulative reward until termination (when a terminal set of states $\mathcal{T}$ is hit) starting from any state $x \in \mathcal{S} - \mathcal{T}$, i.e., estimating

$$J^*(x) = \mathrm{E}_x\left[\sum_{k=0}^{T-1} g(S_k, S_{k+1})\right], \tag{14}$$

where the subscript x denotes that $S_0 = x$, and $T = \inf\{n\colon S_n \in \mathcal{T}\}$. Set $J^*(x) = 0$ for $x \in \mathcal{T}$. Note that if $g(x, y) = I(y \in \mathcal{R})$ with $\mathcal{R} \subseteq \mathcal{T}$, then $J^*(x)$ equals the probability $\mathrm{P}_x(S_T \in \mathcal{R})$.

We refer to the expected cumulative reward from any state as the value function evaluated at that state (this conforms with the terminology used in Markov decision process theory where the framework considered is particularly common; see, e.g., Bertsekas and Tsitsiklis, 1996). Note that by exploiting the regenerative structure of the Markov chain, the problem of estimating steady state measures can also be reduced to that of estimating cumulative reward until regeneration starting from the regenerative state (see, e.g., Fishman, 2001). Similarly, the problem of estimating the expected total discounted reward can be modeled as a cumulative reward until absorption problem after simple modifications (see, e.g., Bertsekas and Tsitsiklis, 1996; Ahamed et al., 2006).

For estimating $(J^*(x)\colon x \in \mathcal{T})$, the expression for the zero-variance change of measure is also well known, but involves knowing a priori these value functions (see Booth, 1985; Kollman et al., 1999; Desai and Glynn, 2001). Three substantially different adaptive importance sampling techniques have been proposed in the literature that iteratively attempt to learn this zero-variance change of measure and the associated value functions. These are: (i) the Adaptive Monte Carlo (AMC) method proposed in Kollman et al. (1999) (our terminology is adapted from Ahamed et al., 2006), (ii) the Cross Entropy (CE) method proposed in De Boer et al. (2000) and De Boer (2001) (also see Rubinstein, 1997, 1999) and (iii) the Adaptive Stochastic Approximation (ASA) based method proposed in Ahamed et al. (2006). We briefly review these methods. We refer the reader to Ahamed et al. (2006) for a comparison of the three methods on a small Jackson network example (this example is known to be difficult to efficiently simulate via static importance sampling).

Borkar et al. (2004) consider the problem of simulation-based estimation of performance measures for a Markov chain conditioned on a rare event. The conditional law depends on the solution of a multiplicative Poisson equation. They propose an adaptive two-time scale stochastic approximation based scheme for learning this solution. This solution is also important in estimating rare-event probabilities associated with queues and random walks involving Markov additive processes as in many such settings the static optimal importance sampling change of measure is known and is determined by solving an appropriate multiplicative Poisson equation (see, e.g., Chang et al., 1994; Beck et al., 1999). We also include a brief review of their scheme in this section.

5.1 The zero-variance measure

Let $P = (p_{xy}\colon\ x, y \in \mathcal{S})$ denote the transition matrix of the Markov chain and let $\mathbf{P}$ denote the probability measure induced by P and an appropriate initial distribution that will be clear from the context. We assume that $\mathcal{T}$ is reachable from all *interior* states $\mathcal{I} \stackrel{\Delta}{=} \mathcal{S} - \mathcal{T}$, i.e., there exists a path of positive probability connecting every state in $\mathcal{I}$ to $\mathcal{T}$. Thus T is an a.s. finite stopping time for all initial values of S_0. Consider another probability measure $\mathbf{P}'$ with a transition matrix $P' = (p'_{xy}\colon\ x, y \in \mathcal{S})$, such that for all $x \in \mathcal{I}$, $y \in \mathcal{S}$, $p'_{xy} = 0$ implies $p_{xy} = 0$. Let E' denote the corresponding expectation operator. Then $J^*(x)$ may be re-expressed as

$$J^*(x) = \mathrm{E}'_x\left[\left(\sum_{n=0}^{T-1} g(S_n, S_{n+1})\right) L(S_0, S_1, \ldots, S_T)\right], \tag{15}$$

where

$$L(S_0, S_1, \ldots, S_T) = \prod_{n=0}^{T-1} \frac{p_{S_n,S_{n+1}}}{p'_{S_n,S_{n+1}}}.$$

Noting that $\mathrm{E}'_x[g(S_n, S_{n+1})L(S_0, S_1, \ldots, S_{n+1})I(T > n)]$ equals

$$\mathrm{E}'_x\bigl[g(S_n, S_{n+1})L(S_0, S_1, \ldots, S_T)I(T > n)\bigr],$$

it may be easily seen that $J^*(x)$ equals

$$\mathrm{E}'_x\left[\left(\sum_{n=0}^{T-1} g(S_n, S_{n+1})L(S_0, S_1, \ldots, S_{n+1})\right)\right].$$

In this framework as well, the static zero-variance change of measure $\mathbf{P}^*$ (with corresponding transition matrix P^*) exists and the process $(S_i\colon\ i \geqslant 0)$ remains a Markov chain under this change of measure. Specifically, consider the transition probabilities

$$p^*_{xy} = \frac{p_{xy}(g(x, y) + J^*(y))}{\sum_{y\in S} p_{xy}(g(x, y) + J^*(y))} = \frac{p_{xy}(g(x, y) + J^*(y))}{J^*(x)}$$

for $x \in \mathcal{I}$ and $y \in \mathcal{S}$.

Then it can be shown that $K = \sum_{n=0}^{T-1} g(S_n, S_{n+1}) L(S_0, S_1, \ldots, S_{n+1})$ equals $J^*(S_0)$ a.s., where

$$L(S_0, S_1, \ldots, S_{n+1}) = \prod_{m=0}^{n} \frac{p_{S_m,S_{m+1}}}{p^*_{S_m,S_{m+1}}} = \prod_{m=0}^{n} \frac{J^*(S_m)}{g(S_m, S_{m+1}) + J^*(S_{m+1})}$$

(see Booth, 1985; Kollman et al., 1999; Desai and Glynn, 2001). We show this via induction. First consider $T = 1$. Then

$$K = g(S_0, S_1)\frac{J^*(S_0)}{g(S_0, S_1) + J^*(S_1)}.$$

Since $J^*(S_1) = J^*(S_T) = 0$, the result follows. Now suppose that the result is correct for all paths of length less than or equal to n. Suppose that $T = n + 1$. Then, K equals

$$\begin{aligned} &g(S_0, S_1)\frac{J^*(S_0)}{g(S_0, S_1) + J^*(S_1)} \\ &\quad + \frac{J^*(S_0)}{g(S_0, S_1) + J^*(S_1)}\left(\sum_{m=1}^{T-1} g(S_m, S_{m+1})\right) \\ &\quad \times \prod_{j=1}^{m} \frac{J^*(S_j)}{g(S_j, S_{j+1}) + J^*(S_{j+1})}. \end{aligned}$$

By the induction hypothesis, $\sum_{m=1}^{T-1} g(S_m, S_{m+1}) \prod_{j=1}^{m} \frac{J^*(S_j)}{g(S_j,S_{j+1})+J^*(S_{j+1})}$ equals $J^*(S_1)$ and the result follows.

Adaptive importance sampling techniques described in the following subsections attempt to learn this change of measure via simulation using an iterative scheme that updates the change of measure (while also updating the value function) so that eventually it converges to the zero-variance change of measure.

5.2 *The adaptive Monte Carlo method*

We describe here the basic AMC algorithm and refer the reader to Kollman et al. (1999) and Desai and Glynn (2001) for detailed analysis and further enhancements.

The AMC algorithm proceeds iteratively as follows: Initially make a reasonable guess $J^{(0)} > 0$ for J^*, where $J^{(0)} = (J^{(0)}(x)\colon x \in \mathcal{I})$ and

$J^* = (J^*(x)\colon x \in \mathcal{I})$. Suppose that $J^{(n)} = (J^{(n)}(x)\colon x \in \mathcal{I})$ denotes the best guess of the solution J^* at an iteration n (since $J^*(x) = 0$ for $x \in \mathcal{T}$, we also have $J^{(n)}(x) = 0$ for such x for all n). This $J^{(n)}$ is used to construct a new importance sampling change of measure that will then drive the sampling in the next iteration. The transition probabilities $P^{(n)} = (p_{xy}^{(n)}\colon x \in \mathcal{I}, y \in \mathcal{S})$ associated with $J^{(n)}$ are given as

$$p_{xy}^{(n)} = \frac{p_{xy}(g(x,y) + J^{(n)}(y))}{\sum_{y \in \mathcal{S}} p_{xy}(g(x,y) + J^{(n)}(y))}. \tag{16}$$

Then for each state $x \in \mathcal{S}$, the Markov chain is simulated until time T using the transition matrix $P^{(n)}$ and the simulation output is adjusted by using the appropriate likelihood ratio. The average of many, say r, such independent samples gives a new estimate $J^{(n+1)}(x)$. This is repeated independently for all $x \in \mathcal{I}$ and the resultant estimates of $(J^{(n+1)}(x)\colon x \in \mathcal{I})$ determine the transition matrix $P^{(n+1)}$ used in the next iteration. Since at any iteration, i.i.d. samples are generated, an approximate confidence interval can be constructed in the usual way (see, e.g., Fishman, 2001) and this may be used in a stopping rule.

Kollman et al. (1999) prove the remarkable result that if r in the algorithm is chosen to be sufficiently large, then there exists a $\theta > 0$ such that

$$\exp(\theta n)\|J^{(n)} - J^*\| \to 0,$$

a.s. for some norm in $\Re^{|\mathcal{I}|}$.

The proof involves showing the two broad steps:

- For any $\varepsilon > 0$, $\mathbf{P}(\|J^{(n)} - J^*\| < \varepsilon$ infinitely often) equals 1.
- Given that $\|J^{(0)} - J^*\| < \varepsilon$ there exists a $0 \leqslant c < 1$ and a positive constant ν such that the conditional probability

$$\mathbf{P}\big(\|J^{(n)} - J^*\| < c^n\|J^{(0)} - J^*\|, \forall n \,\big|\, \|J^{(0)} - J^*\| < \varepsilon\big) \geqslant \nu,$$

which makes the result easier to fathom.

5.3 *The cross-entropy method*

The Cross Entropy (CE) method was originally proposed in Rubinstein (1997) and Rubinstein (1999). See De Boer et al. (2005) for a tutorial. The essential idea is to select an importance sampling distribution from a specified set of probability distributions that minimizes the Kullback–Leibler distance from the zero-variance change of measure. To illustrate this idea, again consider the problem of estimating the rare-event probability $\mathbf{P}(\mathcal{E})$ for $\mathcal{E} \subset \Omega$. To simplify the description suppose that Ω consists of a finite or countable number of elements (the discussion carries through more generally in a straightforward manner). Recall that $\mathbf{P}^*$ such that

$$\mathbf{P}^*(\omega) = \frac{I(\mathcal{E})}{\mathbf{P}(\mathcal{E})}\mathbf{P}(\omega) \tag{17}$$

is a zero-variance estimator for $\mathbf{P}(\mathcal{E})$.

The CE method considers a class of distributions ($\mathbf{P}_\nu$: $\nu \in \mathcal{N}$) where $\mathbf{P}$ is absolutely continuous w.r.t. $\mathbf{P}_\nu$ on the set $\mathcal{E}$ for all ν. This class is chosen so that it is easy to generate samples of $I(\mathcal{E})$ under distributions in this class. Among this class, the CE method suggests that a distribution that minimizes the Kullback–Leibler distance from the zero variance change of measure be selected. The Kullback–Leibler distance of distribution $\mathbf{P}_1$ from distribution $\mathbf{P}_2$ equals

$$\sum_{\omega\in\Omega} \log\left[\frac{\mathbf{P}_2(\omega)}{\mathbf{P}_1(\omega)}\right]\mathbf{P}_2(\omega)$$

(note that this equals zero iff $\mathbf{P}_1 = \mathbf{P}_2$ a.s.). Thus, we search for a $\mathbf{P}_\nu$ that minimizes

$$\sum_{\omega\in\Omega} \log\left[\frac{\mathbf{P}^*(\omega)}{\mathbf{P}_\nu(\omega)}\right]\mathbf{P}^*(\omega),$$

where $\mathbf{P}^*$ corresponds to the zero-variance change of measure. From (17) and the fact that $\sum_{\omega\in\Omega}\log[\mathbf{P}^*(\omega)]\mathbf{P}^*(\omega)$ is a constant, this can be seen to be equivalent to finding

$$\arg\max_{\nu\in\mathcal{N}} \sum_{\omega\in\mathcal{E}} \log\bigl[\mathbf{P}_\nu(\omega)\bigr]\mathbf{P}(\omega). \tag{18}$$

Let $\widetilde{\mathbf{P}}$ be another distribution such that $\mathbf{P}$ is absolutely continuous w.r.t. it. Let $\widetilde{L}(\omega) = \frac{\mathbf{P}(\omega)}{\widetilde{\mathbf{P}}(\omega)}$. Then solving (18) is equivalent to finding

$$\begin{aligned}&\arg\max_{\nu\in\mathcal{N}} \sum_{\omega\in\mathcal{E}} \log\bigl[\mathbf{P}_\nu(\omega)\bigr]\widetilde{L}(\omega)\widetilde{\mathbf{P}}(\omega)\\ &\quad= \arg\max_{\nu\in\mathcal{N}} \widetilde{\mathrm{E}}\log(\mathbf{P}_\nu)\widetilde{L}I(\mathcal{E}).\end{aligned} \tag{19}$$

Rubinstein (1997, 1999) (also see Rubinstein and Kroese, 2004) propose to approximately solve this iteratively by replacing the expectation by the observed sample average as follows: Select an initial $\nu_0 \in \mathcal{N}$ in iteration 0. Suppose that $\nu_n \in \mathcal{N}$ is selected at iteration n. Generate i.i.d. samples $(\omega_1, \dots, \omega_m)$ using $\mathbf{P}_{\nu_n}$, let $L_\nu(\omega) = \frac{\mathbf{P}(\omega)}{\mathbf{P}_\nu(\omega)}$ and select ν_{n+1} as the

$$\arg\max_{\nu\in\mathcal{N}} \frac{1}{m}\sum_{i=1}^{m} \log\bigl(\mathbf{P}_\nu(\omega_i)\bigr)L_{\nu_n}(\omega_i)I(\omega_i \in \mathcal{E}). \tag{20}$$

The advantage in this approach is that often it is easy to explicitly identify $\mathbf{P}_{\nu_n}$. Often the rare event considered corresponds to an event $\{f(\mathbf{X}) > x\}$, where $\mathbf{X}$ is a random vector, and $f(\cdot)$ is a function such that the event $\{f(\mathbf{X}) > x\}$ becomes rarer as $x \to \infty$. In such settings Rubinstein (1999) also proposes that the level x be set to a small value initially so that the event $\{f(\mathbf{X}) > x\}$ is not rare under the original probability. The iterations start with the original measure. Iteratively, as the probability measure is updated, this level may also be adaptively increased to its correct value.

In De Boer et al. (2000) and De Boer (2001) a more specialized Markov chain than the framework described in the beginning of this section is considered. They consider $\mathcal{T} = \mathcal{A} \cup \mathcal{R}$ ($\mathcal{A}$ and $\mathcal{R}$ are disjoint) and $g(x, y) = I(y \in \mathcal{R})$, so that $J^*(x)$ equals the probability that starting from state x, $\mathcal{R}$ is visited before $\mathcal{A}$. The set $\mathcal{A}$ corresponds to an attractor set, i.e., a set visited frequently by the Markov chain, and $\mathcal{R}$ corresponds to a rare set. Specifically, they consider a stable Jackson queueing network with a common buffer shared by all queues. The set $\mathcal{A}$ corresponds to the single state where all the queues are empty and $\mathcal{R}$ corresponds to the set of states where the buffer is full. The probability of interest is the probability that starting from a single arrival to an empty network, the buffer becomes full before the network re-empties (let $\mathcal{E}$ denote this event). Such probabilities are important in determining the steady state loss probabilities in networks with common finite buffer (see Parekh and Walrand, 1989; Heidelberger, 1995).

In this setting, under the CE algorithm, De Boer et al. (2000) and De Boer (2001) consider the search space that includes all probability measures under which the stochastic process remains a Markov chain so that $\mathbf{P}$ is absolutely continuous w.r.t. them. The resultant CE algorithm is iterative.

Initial transition probabilities are selected so that the rare event is no longer rare under these probabilities. Suppose that at iteration n the transition probabilities of the importance sampling distribution are $\mathrm{P}^{(n)} = (p_{xy}^{(n)}\colon x \in \mathcal{I}, y \in \mathcal{S})$. Using these transition probabilities a large number of paths are generated that originate from the attractor set of states and terminate when either the attractor or the rare set is hit. Let k denote the number of paths generated. Let $I_i(\mathcal{E})$ denote the indicator function of path i that takes value one if the rare set is hit and zero otherwise. The new $p_{xy}^{(n+1)}$ corresponding to the optimal solution to (20) is shown in De Boer (2001) to equal the ratio

$$\frac{\sum_{i=1}^{k} L_i N_{xy}(i) I_i(\mathcal{E})}{\sum_{i=1}^{k} L_i N_x(i) I_i(\mathcal{E})}, \tag{21}$$

where $N_{xy}(i)$ denotes the number of transitions from state x to state y and $N_x(i)$ denotes the total number of transitions from state x along the generated path i, L_i denotes the likelihood ratio of the path i, i.e., the ratio of the original probability of the path (corresponding to transition matrix P) and the new probability of the path (corresponding to transition matrix $P^{(n)}$). It is easy to see that as $k \to \infty$, the probabilities converge to the transition probabilities of the zero-variance change of measure (interestingly, this is not true if k is fixed and n increases to infinity).

The problem with the algorithm above is that when the state space is large, for many transitions (x, y), $N_{xy}(i)$ may be zero for all $i \leqslant k$. For such cases, the references above propose a number of modifications that exploit the fact that queues in Jackson networks behave like reflected random walks. Thus, consider the set of states where a subset of queues is nonempty in a network. For all these states, the probabilistic jump structure is independent of the state. This allows for clever state aggregation techniques proposed in the references above for updating the transition probabilities in each iteration of the CE method.

5.4 The adaptive stochastic approximation based algorithm

We now discuss the adaptive stochastic approximation algorithm proposed in Ahamed et al. (2006). It involves generating a trajectory via simulation where at each transition along the generated trajectory the estimate of the value function of the state visited is updated, and along with this at every transition the change of measure used to generate the trajectory is also updated. It is shown that as the number of transitions increases to infinity, the estimate of the value function converges to the true value and the transition probabilities of the Markov chain converge to the zero-variance change of measure.

Now we describe the algorithm precisely. Let $(a_n(x)\colon n \geqslant 0, x \in \mathcal{I})$ denote a sequence of nonnegative step-sizes that satisfy the conditions $\sum_{n=1}^{\infty} a_n(x) = \infty$ and $\sum_{n=1}^{\infty} a_n^2(x) < \infty$, a.s. for each $x \in \mathcal{I}$. Each $a_n(x)$ may depend upon the history of the algorithm until iteration n. This algorithm involves generating a path via simulation as follows:

- Select an arbitrary state $s_0 \in \mathcal{I}$. A reasonable positive initial guess $(J^{(0)}(x)\colon x \in \mathcal{I})$ for $(J^*(x)\colon x \in \mathcal{I})$ is made. Similarly, the initial transition probabilities $(p_{xy}^{(0)}\colon x \in \mathcal{I}, y \in \mathcal{S})$ are selected (e.g., these may equal the original transition probabilities). These probabilities are used to generate the next state s_1 in the simulation.
- At transition n, state s_{n+1} is generated using $(p_{xy}^{(n)}\colon x \in \mathcal{I}, y \in \mathcal{S})$. The updated values $(J^{(n+1)}(x)\colon x \in \mathcal{I})$ and $(p_{xy}^{(n+1)}\colon x \in \mathcal{I}, y \in \mathcal{S})$ are determined as follows:

$$\begin{aligned} &J^{(n+1)}(s_n) \\ &\quad = \big(1 - a_n(s_n)\big)J^{(n)}(s_n) \\ &\qquad + a_n(s_n)\big(g(s_n, s_{n+1}) + J^{(n)}(s_{n+1})\big)\left(\frac{p_{s_n s_{n+1}}}{p_{s_n s_{n+1}}^{(n)}}\right) \end{aligned} \tag{22}$$

and $J^{(n+1)}(x) = J^{(n)}(x)$ for $x \neq s_n$. Also, let

$$\tilde{p}_{s_n s_{n+1}}^{(n+1)} = p_{s_n s_{n+1}}\left(\frac{g(s_n, s_{n+1}) + J^{(n+1)}(s_{n+1})}{J^{(n+1)}(s_n)}\right). \tag{23}$$

This is normalized by setting $p_{s_n y}^{(n+1)} = (\tilde{p}_{s_n y}^{(n+1)})/(\sum_{z\in S} \tilde{p}_{s_n z}^{(n+1)})$ for all y (here $\tilde{p}_{s_n z}^{(n+1)} = p_{s_n z}^{(n)}$ for all $z \neq s_{n+1}$). Again for $x \neq s_n$, $p_{xy}^{(n+1)} = p_{xy}^{(n)}$ for all y.

- If $s_{n+1} \in \mathcal{T}$, the simulation is resumed by selecting s_{n+2} in $\mathcal{I}$ according to a probability distribution μ with the property that the Markov chain with transition probabilities that are the same as the original for all transitions from states in $\mathcal{I}$ and transition probabilities that are identically given by μ for transitions out of $\mathcal{T}$, is irreducible. In that case $(J^{(n+2)}(x)\colon x \in \mathcal{I})$ and $(p_{xy}^{(n+2)}\colon x \in \mathcal{I}, y \in \mathcal{S})$ are set to $(J^{(n+1)}(x)\colon x \in \mathcal{I})$ and $(p_{xy}^{(n+1)}\colon x \in \mathcal{I}, y \in \mathcal{S})$.

Ahamed et al. (2006) show that the algorithm above has the standard Robbins–Monro stochastic approximation form

$$J^{n+1} = (1 - a_n)J^n + a_n\bigl(HJ^n + w_n\bigr),$$

where each $J^n \in \Re^{|\mathcal{I}|}$, H is a mapping $\Re^{|\mathcal{I}|} \to \Re^{|\mathcal{I}|}$, w_n take values in $\Re^{|\mathcal{I}|}$ and are zero mean random vectors (see Kushner and Yin, 1997), and a_n are the step sizes. Under mild regularity conditions the mapping H can be seen to be a contraction under a suitable norm with a unique fixed point $J^* = (J^*(x)\colon x \in \mathcal{I})$ for any set of transition probabilities used to generate transitions (as long as the requirement of absolute continuity is met). Further, it can be shown that the second moment of w_n conditioned on the history of the algorithm up till time $n-1$ is well behaved as required for convergence of the Robbins–Monro algorithm. From this it becomes easy to show that $J^{(n)} \to J^*$ and $P^{(n)} \to \mathrm{P}^*$. If, each step size a_n is set equal to a constant $a > 0$, then it is further shown that $\limsup_{n\to\infty} \mathrm{E}[\|J^{(n)} - J^*\|^2] = \mathrm{O}(a)$ and $\limsup_{n\to\infty} \mathrm{E}[\|P^{(n)} - \mathrm{P}^*\|^2] = \mathrm{O}(a)$.

Ahamed et al. (2006) report that empirically on representative examples, the ASA algorithm performs better than the AMC and the CE algorithm in estimating rare event probabilities when the state spaces involved become large or even moderately large. They further empirically show that for large state spaces, it may perform better than numerical procedures such as value iteration in developing estimates within a reasonable degree of accuracy.

5.5 Multiplicative Poisson equation and conditional measures

Many asymptotically optimal static importance sampling techniques often involve solving a complex set of equations to determine a provably effective static importance sampling distribution (see, e.g., Heidelberger, 1995). This could become particularly difficult when the underlying variables are Markov chains or more general Markov additive processes (see, e.g., Chang et al., 1994; Beck et al., 1999). We illustrate this through an example. Again, consider an irreducible Markov chain $(S_i\colon i \geqslant 0)$ on a large finite state space $\mathcal{S}$ with transition probabilities $(p(x, y)\colon x, y \in \mathcal{S})$. Let $g\colon \mathcal{S} \to R$. Let $\mathrm{E}[\cdot]$ denote the expectation under the associated invariant measure (i.e., steady-state measure) and let $\alpha > \mathrm{E}[g(S_n)]$. Now consider the problem of estimating the probability

$$\mathrm{P}_x\left(\frac{1}{n}\sum_{i=0}^{n-1} g(S_i) \geqslant \alpha\right) \tag{24}$$

for large values of n, where the subscript x denotes the condition $S_0 = x$ (note that this generalizes the i.i.d. case considered in Section 3.1). For such a probability, the static asymptotically optimal importance sampling measure (as $n \to \infty$) is well known (see, e.g., Bucklew, 1990). Under it, the transition probabilities are given by $(p^{\zeta^*}(x, y)\colon\ x, y \in \mathcal{S})$ where

$$p^{\zeta}(x, y) = \frac{\mathrm{e}^{\zeta g(x)} p(x, y) V_{\zeta}(y)}{\rho_{\zeta} V_{\zeta}(x)},$$

where $(V_{\zeta}(x)\colon\ x \in S)$ (resp., ρ_{ζ}) are the Perron–Frobenius eigenvector (resp., eigenvalue) of the positive operator

$$f(\cdot) \to \mathrm{e}^{\zeta g(\cdot)} \sum_{y} p(\cdot, y) f(y), \tag{25}$$

i.e., they solve the multiplicative Poisson equation

$$V_{\zeta}(x) = \frac{\mathrm{e}^{\zeta g(x)}}{\rho_{\zeta}} \sum_{y} p(x, y) V_{\zeta}(y), \quad x \in \mathcal{S}, \tag{26}$$

for $\zeta > 0$ and

$$\zeta^* \stackrel{\Delta}{=} \arg\max_{\zeta \geqslant 0}\bigl(\zeta\alpha - \log(\rho_{\zeta})\bigr).$$

It can further be shown that $\log(\rho_{\zeta})$ is convex and that

$$\zeta^* = \arg\max\bigl(\zeta\alpha - \log(\rho_{\zeta})\bigr) \tag{27}$$

so that $\alpha = \frac{\rho'_{\zeta^*}}{\rho_{\zeta^*}}$ (the superscript "$\prime$" denotes the derivative). Furthermore, let $\mathrm{E}^{\zeta}[\cdot]$ denote the expectation under the invariant measure associated with transition probabilities $(p^{\zeta}_{xy}\colon\ x, y \in \mathcal{S})$. Then $\frac{\rho'_{\zeta}}{\rho_{\zeta}} = \mathrm{E}^{\zeta}[g(S_n)]$ (see, e.g., Bucklew, 1990).

Kontoyiannis and Meyn (2003) develop exact asymptotics for the probability (24) that again requires the knowledge of $(V_{\zeta^*}(x)\colon\ x \in \mathcal{S})$ and ρ_{ζ^*}. Borkar et al. (2004) use these exact asymptotics and observe the following asymptotic conditional law

$$\lim_{n\to\infty} \mathrm{P}\left(S_m = s_m \middle| S_k = s_k, 0 \leqslant k < m, \frac{1}{n}\sum_{i=0}^{n-1} g(S_i) \geqslant \alpha\right)$$
$$\to p^{\zeta^*}(s_{m-1}, s_m).$$

(The discussion in Section 3.3 can be generalized to include the probability in (24). The left-hand side above can be associated with the zero-variance measure that is shown to be asymptotically similar to the exponentially twisted distribution in the right-hand side.) Thus, the knowledge of these transition probabilities is useful in evaluating performance measures conditioned on occurrence of certain rare events, e.g., expected behavior of one queue conditioned on abnormal behavior of another queue in a Markovian finite state space network.

As mentioned earlier, in queueing set-ups when the constituent input or service processes are modeled as Markov additive processes (see, e.g., Chang et al., 1994; Beck et al., 1999) related Perron–Frobenius eigenvectors and eigenvalues need to be determined. Thus, the adaptive methodology discussed below becomes useful in such settings as well.

Evaluating $(V_{\zeta^*}(x)\colon\ x \in \mathcal{S})$ and ρ_{ζ^*} is especially difficult when the state space $\mathcal{S}$ is large. A deterministic iterative method to do this may involve first fixing ζ and evaluating ρ_ζ and ρ'_ζ (by re-evaluating the eigenvalue at a perturbed value of ζ). Borkar and Meyn (2002) develop deterministic numerical iterative schemes to solve this (however, such numerical schemes may not be computationally viable when large state spaces are involved). Once ρ_ζ and ρ'_ζ have been ascertained, ζ may be varied by adding to it a suitable step-size times $\alpha - \frac{\rho'_\zeta}{\rho_\zeta}$, the gradient of (27).

Borkar et al. (2004) develop an adaptive scheme (outlined below) that emulates this using stochastic observations. The adaptive scheme is based on a single simulation run of $(S_i\colon\ i \geqslant 0)$. Fix a distinguished state $s_0 \in \mathcal{S}$. Let $\{a(n)\}$, $\{b(n)\}$ be positive scalar sequences satisfying

$$\sum_n a(n) = \sum_n b(n) = \infty,$$
$$\sum_n \bigl(a(n)^2 + b(n)^2\bigr) < \infty, \qquad \frac{b(n)}{a(n)} \to 0. \tag{28}$$

These serve as step-sizes or 'learning parameters' for the iterative scheme. At each iteration n of the algorithm,

(1) simulate a transition from $S_n = s_n$ to $S_{n+1} = s_{n+1}$ (say) according to the current 'guess' of the transition probability $p^{\zeta^*}(s_n, \cdot)$ given by

$$p_n(s_n, y) \triangleq \frac{\mathrm{e}^{\zeta_n g(s_n)}}{V_n(s_n)V_n(s_0)} p(s_n, y)V_n(y), \tag{29}$$

normalized suitably to render it a probability vector, and

(2) update current guesses for $(V_{\zeta^*}(s_n), \zeta^*)$, denoted by $(V_n(s_n), \zeta_n)$, according to the iterative scheme that sets $V_{n+1}(s_n)$ to equal

$$V_n(s_n) + a(n)\left(\frac{e^{\zeta_n g(s_n)}}{V_n(s_0)} V_n(s_{n+1})\left(\frac{p(s_n, s_{n+1})}{p_n(s_n, s_{n+1})}\right) - V_n(s_n)\right), \tag{30}$$

and

$$\zeta_{n+1} = \zeta_n + b(n)\bigl(\alpha - g(s_{n+1})\bigr). \tag{31}$$

Note that the two iterations proceed on different time-scales as $b(n) = o(a(n))$ so that from the viewpoint of (30), ζ_n is more-or-less a constant function of n, while from the viewpoint of (31), (30) has reached the equilibrium associated with ζ_n.

The iteration (30) is motivated by the following considerations: To solve the equation

$$V_\zeta(x) = \frac{e^{\zeta g(x)}}{\rho_\zeta} \sum_y p(x, y) V_\zeta(y), \quad x \in \mathcal{S}, \tag{32}$$

the following 'value iteration' scheme has been justified in Borkar and Meyn (2002) and Borkar (2002)

$$V^{n+1}(x) = \frac{e^{\zeta g(x)}}{V^n(s_0)} \sum_y p(x, y) V^n(y), \quad x \in \mathcal{S}.$$

The conditional average on the right-hand side may be replaced by an actual evaluation at a simulated transition, i.e., by

$$\frac{e^{\zeta g(x)}}{V^n(s_0)} V^n(y)$$

when $S_n = x$ and $S_{n+1} = y$. However, since this transition is conducted under the probability $p_n(x, y)$, the sample is unbiased by multiplying it by the likelihood ratio

$$\frac{p(x, y)}{p_n(x, y)}.$$

Then the averaging property of stochastic approximation is used to get (30) with $\zeta_n \equiv \zeta$. The iteration (31) corresponds to the stochastic gradient scheme applied to solve (27). In Borkar et al. (2004) convergence conditions of this algorithm are also discussed.

5.6 Brief review of adaptive schemes in other contexts

In a recent work Dupuis and Wang (2004) show that an adaptive importance sampling scheme can be devised to asymptotically optimally estimate

$P(S_n/n \in \mathcal{R})$ for general sets $\mathcal{R}$ (recall that static importance sampling techniques for this probability were discussed in Section 3.1). Under this scheme each X_i is generated using a probability distribution that depends on the previously generated sum $\sum_{j=1}^{i-1} X_i$ (although, they do not learn the associated zero-variance measure).

Stochastic-approximation based adaptive approaches to importance sampling in the specific context of option pricing have been developed in Vázquez-Abad and Dufresne (1998), Su and Fu (2000, 2002). These take a 'stochastic-gradient' based approach using an approximate gradient search. They search for an optimal importance sampling change of measure from among a class of change of measures that is easy to implement. However this class does not include the zero-variance change of measure.

We conclude this section by noting that development of adaptive importance sampling techniques is an exciting, evolving area of research as many problems that were difficult to efficiently solve under naive Monte Carlo simulation or under static importance sampling can be efficiently solved under adaptive importance sampling techniques. The existing work on adaptive techniques has focused primarily on learning the zero-variance change of measure in discrete time discrete state Markov chains. Further research is needed to generalize this to continuous state spaces (see Bolia et al., 2004, for initial attempts in this direction for pricing American options).

6 Queueing systems

Heidelberger (1995) provides a survey of the literature on queueing systems. In this section we briefly mention some recent developments and review the earlier ones to motivate these. As discussed in Section 5, the recent research in adaptive importance sampling techniques shows great promise for estimation of rare-event probabilities associated with queues and queueing networks modeled as finite state Markov chains with relatively small state space. However, as the following discussion indicates, this problem remains open for a large variety of rare events associated with queueing networks when large state spaces or non-Markovian distributions are involved.

6.1 *Single queues*

Szechtman and Glynn (2002) develop large deviations asymptotics and asymptotically optimal importance sampling techniques for the probability of large queue lengths at a fixed time t for $GI/GI/\infty$ systems in heavy traffic, i.e., when the arrival rates are large. Kroese and Nicola (1999) and Juneja (2001) develop asymptotically optimal importance sampling techniques for estimation of buffer overflow probabilities for queues subject to server breakdowns.

6.2 Queueing networks

The key issue in rare-event estimation is illustrated by viewing rare events associated with suitably scaled queueing networks. For example, consider a Jackson network having K queues. Let $Q(t) = (Q_1(t), \ldots, Q_K(t))$ for $t \geqslant 0$ denote the vector of the queue-length process. Consider the scaled process $\widetilde{Q}_n(t) = (1/n)Q(nt)$. Note that in a stable Jackson network $\widetilde{Q}_n(t)$ converges to zero for any t as $n \to \infty$. Significant large deviations literature has focused on identifying the most likely paths along which the process $(\widetilde{Q}_n(t)\colon\ t \geqslant 0)$ hits a set not containing the origin in the nonnegative orthant $\Re_+^K$ (see, e.g., Ignatiouk-Robert, 2000; Ignatyuk et al., 1994). In particular, it is shown that the most likely paths to such rare events are piece-wise linear. Avram et al. (2001) show this in the setting of semimartingale reflected Brownian motion, which typically provides a good approximation for heavily loaded queueing networks.

This is illustrated via a simple two queue tandem Jackson network example, with arrival rate to the first queue equal to λ and service rates at the first and second queues equal to μ_1 and μ_2, respectively, such that $\lambda < \mu_1 < \mu_2$. From the analysis in Ignatyuk et al. (1994) specialized to this network it can be inferred that, for the scaled network, the most likely path to reach the state (x_1, x_2), $x_1 \geqslant 0$, $x_2 > 0$, from the origin involves two piece-wise linear paths. Along the first path, the arrival rate to queue 1 equals μ_1, the service rate at queue 1 equals λ and the service rate at queue 2 remains the same at μ_2. Queue 1 builds up along this path until it reaches the level

$$\frac{\mu_2 - \mu_1}{\mu_2 - \lambda} x_2 + x_1.$$

Thereafter, along the second path, the arrival rate to queue 1 equals μ_1, the service rate at queue 1 equals μ_2 and the service rate at queue 2 equals λ, so that now queue 1 empties as queue 2 builds up until the state (x_1, x_2) is reached. This can also be inferred from the fact that starting from an empty network, the most likely paths to the rare event associated with queue lengths in Jackson networks correspond to the most likely path (in reverse direction) followed by the reversed Jackson network starting from the rare set until it empties; see Frater et al. (1991), Anantharam et al. (1990) and Heidelberger (1995). Figure 4 shows these paths for $x_1 = 0$. (Figure 4 also shows the case when $\mu_1 > \mu_2$, where there is a single path leading to $(0, x_2)$ along which the arrival rate equals μ_2, the service rate at the second queue equals λ, and the service rate at queue 1 remains unchanged.) This suggests that a change of measure should emphasize these two piecewise linear paths to efficiently estimate the buffer overflow probability in the second queue. More generally, this suggests that in general Jackson and other networks, a change of measure that is appropriately piecewise constant ought to be used. However, no such successful implementation has so far been designed. The problem is that in importance sampling it is not sufficient that the new measure emphasizes the

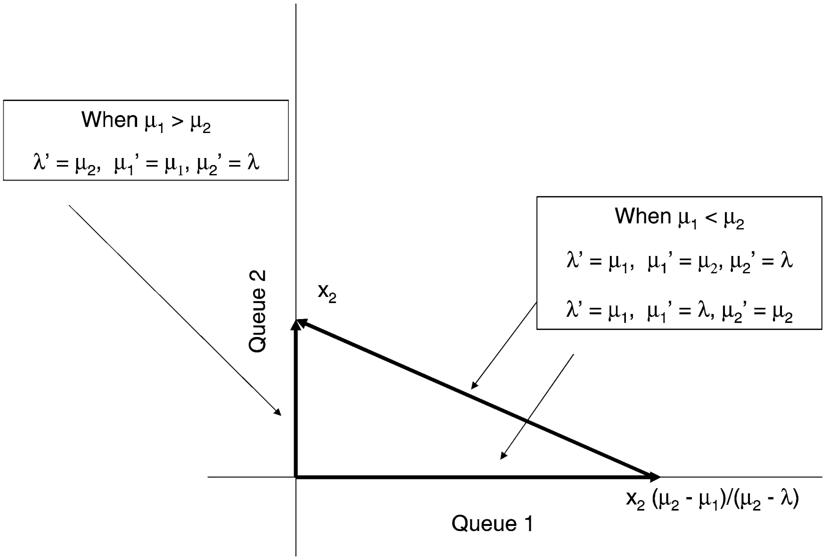
When μ1 > μ2
λ' = μ2, μ1' = μ1, μ2' = λ
When μ1 < μ2
λ' = μ1, μ1' = μ2, μ2' = λ
λ' = μ1, μ1' = λ, μ2' = μ2
Queue 2
x2
Queue 1
x2 (μ2 - μ1)/(μ2 - λ)

Fig. 4. The most likely paths along which the second queue builds up to level x_2 in a scaled two queue tandem Jackson network when $\mu_1 < \mu_2$ and when $\mu_1 > \mu_2$.

most likely paths to the rare event. It must not significantly reduce the original probability for *any* path to the rare event (to avoid build-up of the square of the likelihood ratio that can potentially blow up the second moment of the estimate).

The existing literature has focused primarily on two types of rare events. The first type focuses on estimating the probability that starting from an empty Jackson network, the total network population exceeds a given threshold before the network re-empties. For such a probability, when there exists a unique bottleneck queue, the most likely path to the rare event is linear and corresponds to this queue building up (this queue is unstable while the others remain stable along the most likely paths). Based on heuristic arguments using insights from large deviations theory Parekh and Walrand (1989) present a nonlinear program whose solution identifies these most likely paths (they consider generalized Jackson networks) and the associated importance sampling change of measure to simulate the queueing network. For Jackson networks Frater et al. (1991) obtain an explicit solution to this nonlinear program leading to an explicit form for this change of measure. However, for the simple two-node tandem Jackson network Glasserman and Kou (1995) show that this change of measure does not give good simulation results when the service rates at the two queues are close to each other. In a two-node Jackson network Randhawa and Juneja (2004) show that if feedback is allowed, then for certain traffic parameters the suggested change of measure leads to an estimator with infinite variance. They also discuss how the sensitivity of the second moment under importance sampling may be dampened by combining importance sampling with temporal difference control variates. Frater (1993) illustrate the difficulties associated with the case where the service rates in queues of the networks are close to each other.

The second type of rare-event probability in queueing network settings was considered by Chang et al. (1994) (also see Beck et al., 1999; L'Ecuyer and Champoux, 2001; Kroese and Nicola, 2002). They consider a special class of queueing networks referred to as in-tree networks. These in-tree networks are feed-forward networks (with no feedback). They consist of a 'target' queue at the root and other 'feeder' queues which are like leafs of a tree feeding the target queue at the root. They focus on estimating the probability that the buffer at the target queue overflows during its busy period (a busy period of the target queue is initiated when an arrival to it finds it empty, and it ends when subsequently the target queue re-empties). The remaining network is assumed to have the steady-state distribution at the instants of busy period initiation. The problem of efficient estimation of this probability is closely related to the problem of estimating the steady-state loss probability, i.e., the fraction of customers lost due to buffer overflow at the target queue in the steady-state (see, e.g., Chang et al., 1994; Heidelberger, 1995). They propose a change of measure that typically gives a large amount of variance reduction compared to naive simulation. Based on empirical observations they further suggest that the proposed change of measure is asymptotically optimal or near optimal when the queue-lengths at the feeder queues are assumed to be bounded at the initiation of a busy period of the target queue.

Juneja and Nicola (2005) consider this second type of rare-event probability in a more general setting that allows probabilistic routing with feedback. They generalize the change of measure proposed by Chang et al. (1994) to these settings (however, their analysis is restricted to Jackson networks). Here, they prove the asymptotic optimality of the proposed change of measure when the queue lengths at the feeder queues are assumed to be bounded at the initiation of a busy period of the target queue. Under the condition that the service rates at each feeder queue exceed a specified threshold, they prove that the proposed change of measure is asymptotically optimal, even when the feeder queue-lengths have steady state distributions at the instants of initiation of target queue busy periods. The condition on the feeder queue service rates ensures the large deviations path along which the target queue builds up has a single linear component. For example, in the simple two-queue tandem Jackson network example discussed earlier, the second queue builds up along a single linear path when $\mu_1 > \mu_2$.

7 Heavy-tailed simulations

A recent area of research is investigating rare-event simulation techniques when the random variables in the stochastic system are heavy-tailed. For the purposes of this paper we may define heavy-tailed to mean that the moment generating function of the distribution is infinite for any positive value of the argument. One important consequence of this is that the framework of exponential twisting that is used widely in the light-tailed area can no longer be used here. Also, as explained later, the manner in which rare events occur is very different in the light-tailed and the heavy-tailed settings.

Work in this area began with the problem of simulating tail probabilities of a sum of n i.i.d., nonnegative, heavy-tailed random variables, where n is either fixed or random. In the latter case n is denoted by N which is assumed to be independent of the i.i.d. random sequence. Estimation of tail probabilities of some simple functions (instead of just sums) of a fixed number of random variables that appear in the financial engineering setting and PERT (Project Evaluation and Review Technique) setting have also been developed recently. In this section we focus on fixed and geometric sums, and deal with the financial engineering setting in Section 8; for the PERT setting the reader is referred to Juneja et al. (2005). Recently, advances have been made in efficiently estimating level crossing probabilities of random walks. Recall that application of these probabilities include queueing and insurance settings (as discussed in Examples 5 and 6). We briefly review these developments later in this section.

The case of geometric N has applications in estimating the probability of large steady-state delays in the $M/\mathrm{GI}/1$ queue with heavy-tailed service times, or equivalently the probability of ultimate ruin in the insurance risk process with heavy-tailed claims (see Example 6). Consider Example 5, where the A_i's are exponentially distributed with rate λ and the B_i's are heavy-tailed. Let $\rho = \lambda \mathrm{E}(B_1)$. Let $(H_1, H_2, \ldots)$ denote the 'ladder-heights' of the random walk described in Example 5. The ladder-height H_i is the difference between the ith maximum and $(i-1)$st maximum of the random walk on a sample path where it achieves at least i new maxima; a maximum is achieved at j, if $S_i < S_j$ for all $i < j$ (see, e.g., Asmussen, 2003). These ladder heights are i.i.d., and when the A_i's are exponentially distributed, they have the distribution of the integrated-tail of B_i; note that if F_B denotes the distribution function of B_i then the distribution function of its integrated tail is given by $F_{B,I}(x) = \frac{1}{\mathrm{E}(B_1)} \int_0^x (1 - F_B(y))\,\mathrm{d}y$. Using the Pollaczeck–Khinchine formula, $\mathrm{P}(W_\infty > u)$ may be represented as $\mathrm{P}(\sum_{i=1}^N H_i > u)$, where N is independent of the H_i's, and geometrically distributed with parameter ρ (see, e.g., Asmussen and Binswanger, 1997). Typically H_i is heavy-tailed if B_i is.

The definition of heavy-tailed random variables as given above is almost equivalent to random variables belonging to the *subexponential family*. A nonnegative random variable X is said to be subexponential iff

$$\lim_{u\to\infty} \frac{\mathrm{P}(S_n > u)}{\mathrm{P}(X_1 > u)} = n$$

for all n (Chistyakov, 1964; Sigman, 1999), where $X_1, X_2, \ldots, X_n$ are i.i.d. copies of the random variable X and $S_n = \sum_{i=1}^n X_i$. This definition can be seen to be equivalent to the requirement that

$$\lim_{u\to\infty} \mathrm{P}\Big(\max_{i\leqslant n} X_i > u \big| S_n > u\Big) = 1.$$

This provides the main intuition that is used to investigate stochastic models with heavy tails, i.e., the most likely way for a sum of heavy-tailed random variables to become large is by one of the random variables becoming large. This is different from the light-tailed case, where all the random variables in the sum contribute to the sum becoming large. Common examples of subexponential random variables are the Pareto that has a tail that decays at a polynomial rate (e.g., $1/x^\alpha$ for $\alpha > 0$), the log-normal whose tail decays at the rate $\mathrm{e}^{-(\ln x)^2/(2\sigma^2)}$ (σ^2 is the variance of the associated normal random variable), and the heavy-tailed Weibull whose tail decays at the rate $\mathrm{e}^{-\lambda x^\alpha}$ for $0 < \alpha < 1$ and $\lambda > 0$. A comprehensive reference for subexponential distributions is Embrechts et al. (1997).

Consider the estimation of $\mathrm{P}(S_n > u)$ when $X_1, X_2, \ldots, X_n$ are nonnegative, subexponential random variables, and u is large. The first algorithm for their fast simulation was given by Asmussen and Binswanger (1997). It made direct use of the intuition mentioned above. Let $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ denote the order statistics of $X_1, X_2, \ldots, X_n$. Since the most likely way u is exceeded is by one of the X_i's becoming large, it is $X_{(n)}$ that is the main cause of the variance of $I(S_n > u)$. Asmussen and Binswanger (1997) propose a sampling technique that involves generating samples of $X_1, X_2, \ldots, X_n$, then discarding sample of $X_{(n)}$, and using the conditional probability $\mathrm{P}(S_n > u | X_{(1)}, X_{(2)}, \ldots, X_{(n-1)})$ as an estimator of $\mathrm{P}(S_n > u)$. Asmussen and Binswanger (1997) show that this estimator is asymptotically optimal when the X_i's belong to the class of subexponential distributions that have regularly varying tails, e.g., Pareto-type tails (see Embrechts et al., 1997, for a precise definition) both for fixed n and geometric N. Asmussen and Kroese (2004) gives a related algorithm that has bounded relative error (instead of just asymptotic optimality) for all regularly varying distributions with Pareto-type tails.

Asmussen et al. (2000) gave an importance sampling algorithm for the fixed n and geometric N case. It involves doing importance sampling with another distribution with p.d.f. h, that has the following properties: the distribution K with density $k(x) = f^2(x)/(ch(x))$ where $c = \int f^2/h$ is subexponential and has the property $\ln \overline{K}(x) \leqslant \ln(\overline{F}(x))^2$ (here f denotes the original p.d.f. of X_i, $\overline{F}$ and $\overline{K}$ denote the tail d.f. associated with f and k, respectively). The main motivation for this condition is that terms similar to $k(X_i)$ occur in the square of the second moment of the importance sampling estimator. The condition is accomplished by selecting a p.d.f. h that has a tail that is much heavier than f. For example, for the case of Pareto, Weibull and lognormal f, the choice

$$h(x) = \frac{1}{(x+\mathrm{e})\ln(x+\mathrm{e})^2}, \quad x > 0,$$

works. For the geometric N case, an additional condition $c\rho < 1$ is needed for asymptotic optimality of h (recall that ρ is the parameter of the geometric distribution). Asmussen et al. (2000) also propose modifications to h to achieve this condition (similar modifications were independently proposed in Juneja et al., 1999; Juneja and Shahabuddin, 2002).

Juneja and Shahabuddin (2002) propose an importance sampling distribution obtained by *hazard rate twisting* the original distribution. Define the hazard rate function of X as $\Lambda(x) = -\ln \overline{F}(x)$ where $F(x)$ is the d.f. of X. The hazard rate twisted distribution with twisting parameter θ is given by

$$\mathrm{d}F_\theta(x) = \frac{\mathrm{e}^{\theta\Lambda(x)}\,\mathrm{d}F(x)}{\int_0^\infty \mathrm{e}^{\theta\Lambda(s)}\,\mathrm{d}F(s)} \tag{33}$$

for $0 < \theta < 1$. This is similar to exponential twisting, except that the twisting rate is $\theta\Lambda(x)$ instead of θx. Juneja and Shahabuddin (2002) show that the distributions associated with $\theta \equiv \theta_u = 1 - b/\Lambda(u)$ where b is any positive constant, asymptotically optimally estimate $\mathrm{P}(S_n > u)$, for a large class of subexponential distributions. The proof is easily seen if we assume that $\Lambda(\cdot)$ is concave (this is true for the common subexponential distributions like Weibull and Pareto). Assuming that X has a p.d.f.,

$$\begin{aligned}\int_0^\infty \mathrm{e}^{\theta\Lambda(s)}\,\mathrm{d}F(s) &= \int_0^\infty \mathrm{e}^{\theta\Lambda(s)} f(s)\,\mathrm{d}s\\ &= \int_0^\infty \mathrm{e}^{\theta\Lambda(s)}\bigl(\Lambda'(s)\mathrm{e}^{-\Lambda(s)}\bigr)\,\mathrm{d}s\\ &= \frac{1}{1-\theta}.\end{aligned}$$

Hence the likelihood ratio of the original distribution w.r.t. the hazard rate twisted distribution is given by

$$\prod_{i=1}^{n}\biggl(\frac{1}{1-\theta}\mathrm{e}^{-\Lambda(X_i)\theta}\biggr).$$

Since Λ is concave,

$$\sum_{i=1}^{n}\Lambda(X_i) \geqslant \Lambda\Biggl(\sum_{i=1}^{n} X_i\Biggr) = \Lambda(S_n).$$

Hence one can upper bound the second moment

$$\begin{aligned}&\mathrm{E}_\theta\Biggl(I(S_n > u)\biggl(\frac{1}{1-\theta}\biggr)^{2n}\exp\Biggl\{-\sum_{i=1}^{n}2\theta\Lambda(X_i)\Biggr\}\Biggr)\\ &\quad\leqslant \biggl(\frac{1}{1-\theta}\biggr)^{2n}\mathrm{e}^{-2\theta\Lambda(u)}.\end{aligned}$$

(Here E_θ denotes the expectation operator under the hazard rate twisted distribution.) It is easily verified that selecting $\theta = 1 - b/\Lambda(u)$ gives asymptotic optimality. For fixed n, the choice $b = n$ minimizes the above bound on the second moment (this choice also minimizes the cross entropy as mentioned in Asmussen et al., 2005). Delayed hazard rate twisting and weighted delayed hazard rate twisting are modifications of hazard rate twisting, that, for suitably chosen parameters, asymptotically optimally estimate $\mathrm{P}(S_N > u)$ when N has a geometrically decaying tail. In cases where the hazard function is not available in closed form (e.g. log-normal), 'asymptotic' hazard rate twisting may also be effective (Juneja et al., 2005). In this case the $\Lambda(x)$ in (33) is replaced by $\widetilde{\Lambda}(x)$ where $\lim_{x\to\infty} \widetilde{\Lambda}(x)/\Lambda(x) = 1$ and $\widetilde{\Lambda}(x)$ is available in a simple closed form.

Kroese and Rubinstein (2004) and Huang and Shahabuddin (2004) consider the problem of estimating small tail probabilities of general functions of a finite number of heavy-tailed random variables. Kroese and Rubinstein (2004) propose approaches based on various parameterizations of the input distributions (that determine allowable changes of measure for the importance sampling), one of them being hazard rate twisting. The specific parameters for doing the importance sampling for a given problem are obtained by adaptively minimizing the cross-entropy. No asymptotic optimality results are shown. Huang and Shahabuddin (2004) propose applying hazard rate twisting by an amount θ to all the input distributions, and give an analytical way of determining a θ that yields asymptotic optimality in the simulation.

Huang and Shahabuddin (2003) and Kroese and Rubinstein (2004) make the observation that transforming a heavy-tailed random variable via its hazard function converts it to an exponential random variable with rate one. Hence one of the approaches suggested in Kroese and Rubinstein (2004) is to first use this transformation on each of the input random variables, and then apply exponential twisting to each of these resulting light-tailed random variables. This approach and the one mentioned earlier are equivalent, since it is shown in Huang and Shahabuddin (2003) that applying an exponential change of measure to the transformed input random variables by amount θ, $0 < \theta < 1$, is equivalent to hazard rate twisting the original random variable by amount θ.

We now consider two basic probabilities associated with random walks and queues: The level crossing probability and the level crossing probability during the busy cycle. Specifically, let S_0 be the initial state and for $n \geqslant 1$, let $S_n = S_0 + \sum_{i=1}^n X_i$ where the X_i's are random variables with negative mean and with right tails decaying at a subexponential rate. Let $\tau = \inf\{n \geqslant 1\colon S_n < 0\}$. Consider the problems of estimating $\mathrm{P}_0(\max_n S_n > u)$ and $\mathrm{P}_0(\max_{n\leqslant\tau} S_n > u)$ for large u, where the subscript x in P_x denotes that $S_0 = x$.

Boots and Shahabuddin (2000) propose a hazard rate twisting based algorithm for efficiently estimating $\mathrm{P}_0(\max_n S_n > u)$. They propose that the random walk be truncated after a specified number of transitions and they develop an asymptotically valid upper bound on the bias resulting from this truncation. Though this approach is asymptotically optimal for Weibull-type tails, it did not work well for Pareto-type tails. Bassamboo et al. (2005b) show that no distribution among the class of importance sampling distributions under which X_i's remain i.i.d., can asymptotically optimally estimate $\mathrm{P}_0(\max_{n\leqslant\tau} S_n > u)$ when the X_i's have Pareto-type polynomially decaying tails. They also develop explicit upper bounds on the improvement possible under such importance sampling distributions. This motivates the development of *state-dependent* importance sampling changes of measure that we now briefly discuss.

Again consider the problem of estimating $\mathrm{P}_0(\max_n S_n > u)$. From the discussion in Section 3.2 it can be seen that conditioned on $(X_1, X_2, \ldots, X_{n-1})$ and that the rare event has not occurred, the distribution of X_n under the zero-variance measure is given by

$$\mathrm{d}F_s^*(x) = \mathrm{d}F(x)\frac{\mathrm{P}_{s+x}(\max_n S_n > u)}{\mathrm{P}_s(\max_n S_n > u)},$$

when $\sum_{i=1}^{n-1} X_i = s$ (here F denotes the d.f. of X_1, and $\mathrm{P}_s(\max_n S_n > u) = 1$ for $s \geqslant u$). This is not implementable as $\mathrm{P}_s(\max_n S_n > u)$ is not known for $s < u$; if these were known, there would be no need for simulation. The idea then is to use asymptotic approximations to these probabilities as surrogates for them. Note that the same discussion also applies to $\mathrm{P}_0(\max_{n\leqslant\tau} S_n > u)$. Bassamboo et al. (2005b) develop asymptotic approximations for the latter probability in discrete settings. They then use the corresponding approximate zero variance measure and empirically demonstrate its asymptotic optimality in certain settings. Shahabuddin (2005) uses the approximate zero variance measure associated with the well-known asymptotic

$$\mathrm{P}_s\Bigl(\max_n S_n > u\Bigr) \sim \frac{1}{|\mathrm{E}(X)|}\int_{u-s}^{\infty} \overline{F}(t)\,\mathrm{d}t$$

derived in Pakes (1975) that holds when X is subexponentially distributed (under additional mild restrictions; see, e.g., Asmussen, 2003) to estimate the level crossing probability. However, they achieve limited success in this.

In a recent presentation, Blanchet and Glynn (2005) display a more refined approximation of the level crossing probability and claim that the change of measure corresponding to it leads to provably asymptotically optimal estimation of the level crossing probability.

8 Financial engineering applications

We first present some examples of rare-event simulation problems that arise in financial engineering.

Example 10 (Light-tailed value-at-risk). We first give a brief overview of the standard setting that has been given in Glasserman et al. (2000). Consider a portfolio consisting of several instruments (e.g., shares, options, bonds, etc.). The value of each instrument depends on one or more of m risk factors (e.g. stock price, price of gold, foreign exchange rate, etc.). Let $S(t) =$

$(S_1(t), \ldots, S_m(t))$ denote the values of the risk factors at time t and let $V(S(t), t)$ denote the value of the portfolio at time t (the values of several instruments, e.g., options, may depend directly on the time). Let t denote the current time, and let $\Delta S = [S(t + \Delta t) - S(t)]^{\mathrm{T}}$ (the notation A^{T} stands for the transpose of the matrix A) be the random change in risk factors over the future interval $(t, t + \Delta t)$. Hence the loss over the interval Δt is given by $\mathcal{L} = V(S(t), t) - V(S(t) + \Delta S, t + \Delta t)$. Note that the only random quantity in the expression for the loss is ΔS. The risk problem is to estimate $\mathrm{P}(\mathcal{L} > x)$ for a given x, and the value-at-risk problem is to estimate x such that $\mathrm{P}(\mathcal{L} > x) = p$ for a given p, $0 < p < 1$. Usually p is of the order 0.01 and Δt is either 1 day or 14 days. Techniques that are efficient for estimating $\mathrm{P}(\mathcal{L} > x)$ for a given x, can be adapted to estimate the value-at-risk. Hence the focus in most papers in this area is on efficient estimation of $\mathrm{P}(\mathcal{L} > x)$ for a given x.

A quadratic approximation to $\mathcal{L}$ is an approximation of the form

$$\mathcal{L} \approx a_0 + a^{\mathrm{T}}\Delta S + (\Delta S)^{\mathrm{T}} A \Delta S \equiv a_0 + Q, \tag{34}$$

where a_0 is a scalar, a is a vector and A is a matrix. The importance sampling approach given in Glasserman et al. (2000) involves determining an efficient change of measure on the ΔS for estimating $\mathrm{P}(Q + a_0 > x)$, and then using the same change of measure for estimating $\mathrm{P}(\mathcal{L} > x)$; since $\mathcal{L} \approx a_0 + Q$, it is likely that such an approach will be efficient for estimating the latter. The r.v. Q is more tractable and it is easier to come up with efficient changes of measure for estimating $\mathrm{P}(Q + a_0 > x)$ and proving their asymptotic optimality as $x \to \infty$. Glasserman et al. (2000) use the 'delta–gamma' approximation. This is simply the Taylor series expansion of the loss $\mathcal{L}$ with respect to ΔS and it uses the gradient and the Hessian of $\mathcal{L}$ with respect to ΔS to come up with Q. The gradient and Hessian may be computed analytically in cases where the portfolio consists of stocks and simple options.

Usually some probability model is assumed for the distribution of ΔS, and parameters of the model are estimated from historical data. A common assumption, that is also used in Glasserman et al. (2000), is that ΔS is distributed as $\mathrm{N}(0, \Sigma)$, i.e., it is multi-variate normal with mean zero and Σ is its covariance matrix. If we let C be such that $CC^{\mathrm{T}} = \Sigma$, then ΔS may be expressed as CZ where $Z \sim \mathrm{N}(0, I)$. Hence $Q = (Z^{\mathrm{T}}C^{\mathrm{T}}ACZ) + (a^{\mathrm{T}}CZ)$. For the case where Σ is positive definite, Glasserman et al. (2000) give a procedure to find such a C so that $C^{\mathrm{T}}AC$ is a diagonal matrix. In that case

$$Q = Z^{\mathrm{T}}\Lambda Z + b^{\mathrm{T}}Z = \sum_{i=1}^{m}\bigl(\lambda_i Z_i^2 + b_i Z_i\bigr), \tag{35}$$

where Λ is a diagonal matrix with λ_i's in the diagonal, and b is a vector with elements b_i. The problem is to find an efficient change of measure to estimate $\mathrm{P}(Q > y)$, for large $y := x + a_0$. Note that in this case Q is a sum of the independent random variables $X_i = (\lambda_i Z_i^2 + b_i Z_i)$.

Example 11 (Heavy-tailed value-at-risk). The multivariate normal is quite light-tailed and there is evidence from empirical finance that risk factors may have tails that are heavier than normal. Glasserman et al. (2002) consider the case where ΔS has a multivariate t distribution (i.e., the marginals have the univariate t distribution) with mean vector 0. The univariate t distribution with ν degrees of freedom has a tail that decays polynomially, i.e., similar to $x^{-\nu}$, as compared to $x^{-1}\exp(-x^2/2\sigma^2)$ which roughly describes the order of decay for the normal distribution. Glasserman et al. (2000) consider the version of the multivariate t as defined in Anderson (1984) and Tong (1990). This random variable may be expressed as

$$\frac{W}{\sqrt{\chi_\nu^2/\nu}},$$

where $W \sim \mathrm{N}(0, \Sigma)$ and χ_ν^2 is a chi-square random variable with ν degrees of freedom (see, e.g., Fang et al., 1987) that is independent of W. If we let $V = \chi_\nu^2/\nu$, then similar to (35), the diagonalized quadratic form becomes

$$Q = \sum_{i=1}^{m}\left(\frac{1}{V}\lambda_i Z_i^2 + \frac{1}{\sqrt{V}} b_i Z_i\right) \tag{36}$$

(as before, $Z \sim \mathrm{N}(0, I)$ and λ_i and b_i are constants). The problem is to determine an efficient change of measure for estimating $\mathrm{P}(Q > y)$ for large y, so that the same change of measure can be used for estimating the actual probability $\mathrm{P}(\mathcal{L} > y)$.

In this case the quadratic form is more complicated than the quadratic form for the normal case, due to two reasons.

- In this case, Q is heavy-tailed.
- We no longer have a sum of independent random variables; we now have dependence among the components in the sum through V. In this sense, this problem is more complex than the heavy-tailed problems considered in Asmussen and Binswanger (1997), Asmussen et al. (2000) and Juneja and Shahabuddin (2002), that dealt with sums of independent heavy-tailed random variables.

Example 12 (Credit risk). Consider a portfolio of loans that a lending institution makes to several obligors, say m. Obligors may default causing losses to the lending institution. There are several default models in the literature. In "static" default models, the interest is in the distribution of losses for the institution over a fixed horizon. More formally, corresponding to each obligor there is a default indicator Y_k, i.e., $Y_k = 1$ if the kth obligor defaults in the given time horizon, and it is zero otherwise. Let p_k be the probability of default of the kth obligor and c_k be the loss resulting from the default. The loss is then given by $\mathcal{L}_m = \sum_{k=1}^{m} c_k Y_k$. Efficient estimation of $\mathrm{P}(\mathcal{L}_m > x)$ when

m and x are large then becomes important. To study the asymptotics and rare-event simulation for this as well as more general performance measures, it is assumed that $x \equiv x_m = qm$ for fixed q, so that $\mathrm{P}(\mathcal{L}_m > x_m) \to 0$ as $m \to \infty$ (Glasserman and Li, 2005).

An important element that makes this problem different from the earlier random walk models is that in this case the Y_k's are dependent. One method to model this dependence is the normal copula model. (This methodology underlies essentially all models that descend from Merton's seminal firm-value work Merton (1974); also see Gupta et al. (1997).) In this case, with each Y_k a standard normal random variable X_k is associated. Let x_k be such that $\mathrm{P}(X_k > x_k) = p_k$, i.e., $x_k = \Phi^{-1}(1 - p_k)$ where Φ is the d.f. of standard normal distribution. Then, setting $Y_k = I(X_k > x_k)$, we get $\mathrm{P}(Y_k = 1) = p_k$ as required. Dependence can be introduced among the Y_k's by introducing dependence among the X_k's. This is done by assuming that each X_k depends on some "systemic risk factors" $Z_1, \ldots, Z_d$ that are standard normal and independent of one another, and an "idiosyncratic" risk factor ε_k that is also standard normal and independent of the Z_i's. Then each X_k is expressed as

$$X_k = \sum_{i=1}^{d} a_{ki} Z_i + b_k \varepsilon_k.$$

The a_{ki}'s are constants and represent the "factor-loadings", i.e., the effect of factor i on obligor k. The b_k is a constant that is set to $\sqrt{1 - \sum_{i=1}^{d} a_{ki}^2}$ so that X_k is standard normal.

8.1 Approaches for importance sampling

There are two basic approaches that have been used for determining asymptotically optimal changes of measures for problems of the type mentioned above. The first approach makes use of the light-tailed simulation framework of exponential twisting. As in Section 4, this is done with the aim of getting a uniform bound (see Section 2.4.1) on the likelihood ratio. For light-tailed problems like the one in Example 10, the framework can be applied directly. Heavy-tailed problems like the ones in Example 11, are transformed into light-tailed problems and then the framework is applied to them. All this is discussed in Sections 8.2–8.4. A general reference for this approach applied to several value-at-risk problems is Glasserman (2004); in the further discussion we attempt to bring out the essentials.

The second approach uses conditioning. Note that in Example 11, if we condition on V or B, then Q is reduced to the one in Example 10 (that is light-tailed) for which exponential twisting can be effectively used. Similarly in Example 12 in the normal copula model, conditioned on Z, the loss function is a sum of independent random variables for which the exponential twisting approach is well known. The question then arises as to what change of measure to use on the conditioning random variable, if any. This is discussed in Sections 8.5 and 8.6.

8.2 A light-tailed simulation framework

Consider the problem of estimating $\mathrm{P}(Y > y)$ where $Y = h(X_1, \ldots, X_m)$, h is some function from $\Re^m$ to $\Re$, and $X_1, \ldots, X_m$ are independent random variables, not necessarily i.i.d. For simplicity in presentation we assume that each X_i has a p.d.f. $f_i(x)$ and that the function h is sufficiently smooth so that Y also has a p.d.f. Let $F_i(x)$ be the d.f. of X_i, let $\overline{F_i}(x) = 1 - F_i(x)$, and define the hazard function as $\Lambda_i(x) = -\ln \overline{F_i}(x)$. Recall that for any two functions, say $g_1(x)$ and $g_2(x)$, $g_1(x) \sim g_2(x)$ means that $\lim_{x\to\infty} g_1(x)/g_2(x)$ exists and equals 1.

If we let $\tilde{f}_i(x)$ be a new p.d.f. for X_i, with the same support as X_i, then the importance sampling equation (2) specializes to

$$\mathrm{P}(Y > y) = \mathrm{E}\bigl(I(Y > y)\bigr) = \widetilde{\mathrm{E}}\bigl(I(Y > y)L(X_1, \ldots, X_m)\bigr), \tag{37}$$

where

$$L(x_1, \ldots, x_m) = \prod_{i=1}^{m} \frac{f_i(x_i)}{\tilde{f}_i(x_i)},$$

and $\widetilde{\mathrm{E}}(\cdot)$ denotes the expectation operator associated with the p.d.f.'s $\tilde{f}_i$. Once again, the attempt is to find $\tilde{f}_i$'s so that the associated change of measure is asymptotically optimal.

As mentioned in Section 3.3, for light-tailed random variables one may use the change of measure obtained by exponentially twisting the original distributions. In our case, exponentially twisting $f_i(x)$ by amount θ, $\theta > 0$, gives the new density

$$f_{i,\theta}(x) = \frac{f_i(x)\mathrm{e}^{\theta x}}{M_{X_i}(\theta)},$$

where $M_{X_i}(\theta)$ denotes the moment generating function (m.g.f.) of the random variable X_i.

Consider the case when Y is light-tailed. In that case the attempt in the literature is to find $\tilde{f}_1, \ldots, \tilde{f}_m$, that translate into exponential twisting of Y by amount θ. This means that the new likelihood ratio, $L(X_1, \ldots, X_m)$, is of the form $M_Y(\theta)\mathrm{e}^{-\theta Y}$. For example, consider the simple case where $Y = \sum_{i=1}^m X_i$, and the X_i's are light-tailed random variables. Now consider doing exponential twisting by amount θ on X_i. Then one can easily see that

$$L(X_1, \ldots, X_m) = \prod_{i=1}^{m} \bigl(M_{X_i}(\theta)\mathrm{e}^{-\theta X_i}\bigr) = M_Y(\theta)\mathrm{e}^{-\theta Y}.$$

Hence, in this specific case, the exponential twisting of X_i's by θ translates into exponential twisting of Y by θ.

If such an exponential twist on the X_i's can be found, then the second moment can be bounded as follows:

$$\begin{aligned}\widetilde{\mathrm{E}}\big(I(Y > y)L^2(X_1, \ldots, X_m)\big) &= \widetilde{\mathrm{E}}\big(I(Y > y)M_Y^2(\theta)\mathrm{e}^{-2\theta Y}\big) \\ &\leqslant M_Y^2(\theta)\mathrm{e}^{-2\theta y}. \end{aligned} \tag{38}$$

Then $\theta = \theta_y^*$ may be selected that minimizes $M_Y^2(\theta)\mathrm{e}^{-2\theta y}$ or equivalently that minimizes $\ln M_Y(\theta) - \theta y$. Huang and Shahabuddin (2003) generalize earlier specific results and show that under fairly general conditions, this procedure yields asymptotically optimal estimation.

Hence, the main challenge in this approach is to find a change of measure on the X_i's that translates into exponential twisting of Y. We now see how this is done for the examples mentioned in the beginning of this section.

8.3 Light-tailed value-at-risk

Consider the problem of estimating $\mathrm{P}(Q > y)$, where Q is given by (35). In this case $Q = \sum_{i=1}^m V_i$, where $V_i = \lambda_i Z_i^2 + b_i Z_i$. Hence, as shown in Section 8.2, exponentially twisting each V_i by θ, will translate into exponential twisting of Q by θ. The question then is: What is the change of measure on the Z_i's that would achieve exponential twisting of the V_i's. Glasserman et al. (2000) show that this is achieved if the mean and variance of Z_i are changed to $\mu_i(\theta)$ and $\sigma_i^2(\theta)$, respectively, where

$$\sigma_i^2(\theta) = \frac{1}{1 - 2\theta\lambda_i}, \qquad \mu_i(\theta) = \theta b_i \sigma_i^2(\theta)$$

(the Z_i's remain independent).

Glasserman et al. (2000) perform a further enhancement to the simulation efficiency by using stratification on Q. Note that by completing squares, each V_i may be expressed as the sum of a noncentral chi-square r.v. and a constant. Hence its m.g.f. is known in closed form and thus the m.g.f. of Q can easily be obtained in closed form. This can be inverted to get the distribution of Q. This enables stratification on Q that further brings down the variance of the importance sampling estimator $I(Q > y)M_Q(\theta_y^*)\exp(-\theta_y^* Q)$. Glasserman et al. (2000) give a simple algorithm for generating $(Z_1, \ldots, Z_m)$ conditional on Q lying in given stratas.

8.4 Heavy-tailed value-at-risk: transformations to light tails

Consider estimating $\mathrm{P}(Q > y)$ where Q is given by (36). As mentioned before, Q is heavy-tailed and thus direct application of exponential twisting cannot be attempted here. Glasserman et al. (2002) transform this problem into a light-tailed problem before using exponential twisting. In particular, they define

$$Q_y = V(Q - y) = \sum_{i=1}^{m} \big(\lambda_i Z_i^2 + b_i Z_i \sqrt{V}\big) - yV.$$

It is easy to check that Q_y is light-tailed for each y, since all its components are light-tailed. Also $\mathrm{P}(Q > y) = \mathrm{P}(Q_y > 0)$ and hence a heavy-tailed simulation problem is transformed into a light-tailed one!

An exponential change of measure by amount $\theta \geqslant 0$ on Q_y can be attempted through selecting appropriate changes of measure for the Z_i's and V. In this case, we have the following simple bound on the second moment:

$$\mathrm{E}\big(I(Q_y > 0) M_{Q_y}^2(\theta) \mathrm{e}^{-2\theta Q_y}\big) \leqslant M_{Q_y}^2(\theta).$$

Adapting the same approach as in Section 8.2, a θ_y^* is selected that minimizes this bound. Indeed, as proved in Glasserman et al. (2002), for the case where $\lambda_i > 0$ for $i = 1, \ldots, m$, this selection gives bounded relative error. Glasserman et al. (2002) also give an explicit change of measure (in terms of θ) on V, and changes of measure (in terms of θ) on Z_i's conditional on V, that achieve exponential twisting of Q_y by amount θ.

Huang and Shahabuddin (2003) give another approach for transforming a heavy-tailed simulation problem into a light-tailed one. Note that the hazard function of any random variable whose p.d.f. is positive on $\Re^+$ (resp., $\Re$) is an increasing function on $\Re^+$ (resp., $\Re$). Let $\Lambda_Y(y)$ be the hazard function of Y, and let $\Lambda(y)$ be any monotonically increasing function such that $\Lambda(y) \sim \Lambda_Y(y)$. Then it is shown in Huang and Shahabuddin (2003) that $\Lambda(Y)$ is exponential-tailed with rate 1. Usually such a $\Lambda(y)$ may be determined through asymptotic results in heavy-tailed theory, or by clever application of the Laplace method for solving integrals. Then $\mathrm{P}(Y > y)$ may be re-expressed as $\mathrm{P}(\Lambda(Y) > \Lambda(y))$, and we again have a light-tailed simulation problem where y is replaced by its monotonic transformation $\Lambda(y)$ (note that $\Lambda(y) \to \infty$ as $y \to \infty$). In this case, since $\Lambda(Y)$ is usually not in the form of a sum of functions of the individual X_i's, it is difficult to find a change of measure on the X_i's that will achieve exponential twisting on the $\Lambda(Y)$. For the case where the changes in risk factors have the Laplace distribution, Huang and Shahabuddin (2003) find upper bounds on $\Lambda(Y)$ that are in this form, so that exponential twisting can easily be applied.

8.5 Conditional importance sampling and zero-variance distributions

As mentioned in Example 11, conditioned on V, Q has the same form as in Example 10, for which the asymptotically optimal change of measure is much simpler to determine. This motivates a conditioning approach for such problems.

Consider the more general problem of estimating $\mathrm{P}(Y_y > 0)$ where $Y_y = h_y(X_1, \ldots, X_m)$ and h_y is some function from $\Re^m$ to $\Re$ that also depends on y.

For the class of problems considered in Section 8.2, $Y_y = Y - y$. The Q_y described in Section 8.4 is also an example of this. Assume that $\mathrm{P}(Y_y > 0) \to 0$ as $y \to \infty$. Let $V = \tilde{h}(X_1, \dots, X_m)$ be a 'conditioning' random variable, where $\tilde{h}$ is some other function of the input random variables (usually V is a function of just one of the input random variables). As mentioned in the previous paragraph, it is important to select V such that, given $V = v$, it is easy to determine changes of measure on the X_i's that translate into exponential twisting of the Y_y. This implies that for any v, given $V = v$, the Y_y should be light-tailed.

For conditional importance sampling, we again use insights from the zero-variance change of measure. Note that

$$\mathrm{P}(Y_y > 0) = \int \mathrm{P}(Y_y > 0 | V = v) f_V(v)\,\mathrm{d}v. \tag{39}$$

Hence if $\mathrm{P}(Y_y > 0|V = v)$ were computable for each v, then the zero-variance change of measure on the V (for estimating $\mathrm{P}(Y_y > 0)$) would be

$$\frac{\mathrm{P}(Y_y > 0|V = v) f_V(v)}{\int \mathrm{P}(Y_y > 0|V = v) f_V(v)\,\mathrm{d}v}. \tag{40}$$

Recall that Y_y is a tractable approximation to the actual loss function. Usually, given $V = v$, Y_y is a sum of independent random variables and hence $\mathrm{P}(Y_y > 0|V = v)$ may be determined by numerical transform inversion techniques. Once one is able to compute $\mathrm{P}(Y_y > 0|V = v)$ for each v, then one can compute $\mathrm{P}(Y_y > 0)$ by numerical integration. One can then generate from the zero-variance change of measure on the V by first computing its cumulative distribution function (using numerical integration) and then using numerical inversion. All this, even though theoretically possible, is practically possible usually for the case of only discrete V. The asymptotic optimality proof is usually not possible for either continuous or discrete V. This approach has been proposed in Shahabuddin and Woo (2004) and applied to the estimation of the tail probability of the quadratic form in the value-at-risk problem, where the risk factors have the distribution of a finite mixture of multivariate normals. In this case, the conditioning random variable V is the random identifier of the multivariate normal that one samples from in each step. The multivariate mixture of normals has applications for the case where the asset prices obey the jump diffusion model (see Shahabuddin and Woo, 2004).

As another approach to this problem, Shahabuddin and Woo (2004) use the Markov inequality

$$\mathrm{P}(Y_y > 0|V = v) \leqslant \mathrm{E}\big(\mathrm{e}^{Y_y \theta^*_{y,v}} | V = v\big),$$

where $\theta^*_{y,v}$ is obtained from minimizing the Markov bound $\mathrm{E}(\mathrm{e}^{Y_y \theta} | V = v)$ over all $\theta \geqslant 0$. Then $\mathrm{E}(\mathrm{e}^{Y_y \theta^*_{y,v}} | V = v)$ may be used as a close surrogate to $\mathrm{P}(Y_y > 0|V = v)$ in (40). Usually, this inequality is somewhat tight for large y, and hence not much loss in performance may be expected due to this substitution. Also, $\mathrm{E}(\mathrm{e}^{Y_y\theta^*_{y,v}}|V = v)$, the conditional moment generating function, is usually computable in closed form. By using this surrogate, the approximate zero-variance change of measure for the V would be

$$\tilde{f}_V(v) = \frac{\mathrm{E}(\mathrm{e}^{Y_y\theta^*_{y,v}}|V = v) f_V(v)}{\int \mathrm{E}(\mathrm{e}^{Y_y\theta^*_{y,v}}|V = v) f_V(v)\,\mathrm{d}v}.$$

Once again, even though theoretically possible, there are difficulties with this approach both regarding the implementation and the asymptotic optimality proof. In addition to the numerical computational burden in the previous approach, we have the additional burden of determining $\theta^*_{y,v}$ for each v, and $\theta^*_{y,v}$ is rarely available in closed form. Hence this approach is again efficient only for the case where V is a discrete random variable taking a finite number of values. In Shahabuddin and Woo (2004) it has been applied to the mixture of normal problems mentioned in the previous paragraph. The asymptotic optimality for this case is also proved.

In order to make the implementation simpler for a continuous conditioning random variable V, Shahabuddin and Woo (2004) consider relaxing the bound on $\mathrm{P}(Y_y > 0|V = v)$, by using the same θ for all $V = v$, and then determining the best θ to use. In that case, the approximate zero-variance distribution is given by

$$\tilde{f}_V(v) = \frac{\mathrm{E}(\mathrm{e}^{Y_y\theta}|V = v) f_V(v)}{\int \mathrm{E}(\mathrm{e}^{Y_y\theta}|V = v) f_V(v)\,\mathrm{d}v}.$$

In this case, if V is such that

$$\mathrm{E}\big(\mathrm{e}^{Y_y\theta}|V = v\big) = g_1(\theta, y)\mathrm{e}^{g_2(\theta,y)v} \tag{41}$$

(for any functions g_1 and g_2) then

$$\tilde{f}_V(v) = \frac{\mathrm{e}^{g_2(\theta,y)v} f_V(v)}{\int \mathrm{e}^{g_2(\theta,y)v} f_V(v)\,\mathrm{d}v},$$

i.e., the approximate zero-variance change of measure is then an exponential twisting by amount $g_2(\theta, y)$. Once V is sampled from the new measure, then one needs to do a change of measure on the X_i's given $V = v$, so that one achieves exponential twisting of Y_y. If this can be done, then it is easy to check that the likelihood ratio is $M_{Y_y}(\theta)\mathrm{e}^{-Y_y\theta}$, and thus we are in a framework similar to that in Section 8.4. As in that section, the second moment $\mathrm{E}(I(Y_y > 0)M^2_{Y_y}(\theta)\mathrm{e}^{-2\theta Y_y})$ may then be upper bounded by $M^2_{Y_y}(\theta)$, and a θ^*_y may be selected that minimizes this bound.

It is easy to check that for Q_y in the multivariate t case in Section 8.4, selecting V as the chi-square random variable achieves the condition given in (41). However, the choice of the conditioning variable may not always be obvious.

For example, consider the case where the risk factors have the Laplace distribution with mean vector 0, as considered in Huang and Shahabuddin (2003). In this case, the tails of the marginal distributions decay according to $\frac{1}{\sqrt{x}}\mathrm{e}^{-cx}$, for some constant $c > 0$. Justifications of this type of tail behavior may be found in Heyde and Kou (2004). The multivariate Laplace random-variable with mean vector 0 may be expressed as

$$\sqrt{B}W,$$

where $W \sim \mathrm{N}(0, \Sigma)$ and B is an exponentially distributed random variable with rate 1 (see, e.g., Kotz et al., 2001). In this case the Q_y becomes

$$Q_y = \sum_{i=1}^{m}\left(\lambda_i Z_i^2 + \frac{1}{\sqrt{B}}b_i Z_i - \frac{y}{B}\right). \tag{42}$$

However, taking B as the conditioning random variable does not work. In fact, a conditioning random variable V that satisfies (41) in this case is $V = -1/B$. This is indeed surprising since the V does not even take positive values and we are doing exponential twisting on this random variable! Shahabuddin and Woo (2004) thus generalize the exponential twisting idea in Glasserman et al. (2002) to make it more widely applicable. It also improves the earlier method in Huang and Shahabuddin (2003) for the case where the changes in risk factors are Laplace distributed (that was based on hazard rate twisting).

8.6 Credit risk models

Consider the model described in Example 12 where the problem is to estimate $\mathrm{P}(\mathcal{L}_m > x_m)$ where $\mathcal{L}_m = \sum_{k=1}^{m} c_k Y_k$ and $x_m = qm$ for some constant q. For the case of independent obligors, where the Y_k's are independent Bernoulli's, the basic procedure is the same as described in Section 8.2. For the case when the Y_k's are dependent, with the dependence structure specified in Example 12, Glasserman and Li (2005) first attempt doing importance sampling conditional on the realization of the normal random variable Z, but leaving the distribution of Z unchanged. A similar approach is also followed by Merino and Nyefeler (2004). Note that in this case, the probability of default for obligor k now becomes a function of Z, i.e.,

$$p_k(Z) = \Phi\left(\frac{\sum_{i=1}^{d} a_{ki} Z_i + \Phi^{-1}(p_k)}{b_k}\right).$$

In the simulation procedure, first Z is sampled and $p_k(Z)$'s are computed, then the importance sampling procedure mentioned above is applied by treating the $p_k(Z)$'s as fixed. In particular, let $\psi_i^{(m)}(\theta, z)$ be the log moment generating function of $\mathcal{L}_m$ given $Z = z$, and let $\theta_m(z)$ be the $\theta \geqslant 0$ that maximizes $-\theta qm + \psi_i^{(m)}(\theta, z)$. Then after sampling Z, exponential twisting is performed on the $c_i Y_i$'s by the amount $\theta_m(Z)$. Note that $\theta_m(Z) > 0$ only when Z is such that $\sum_{i=1}^{m} c_k p_k(Z) = \mathrm{E}(\mathcal{L}_m|Z) < qm$; for the other case $\theta_m(Z) = 0$, and we do not do importance sampling.

Glasserman and Li (2005) show that when the dependence among the Y_i's is sufficiently low, conducting importance sampling conditional on the Z is enough, i.e., the distribution of Z's need not be changed under importance sampling. However, when the dependence is higher, one also has to change the distribution of Z so that it has greater probability of falling in regions where the default events are likely to occur. Again, one approach is to select the importance sampling distribution of Z that is close to the zero-variance distribution.

In an earlier paper Glasserman et al. (1999) consider the problem of estimating $\mathrm{E}(\mathrm{e}^{G(Z)})$ where Z is $\mathrm{N}(0, I)$, and G is some function from $\Re^m$ to $\Re$. An importance sampling method proposed in Glasserman et al. (1999) was to find the point that maximizes $\mathrm{e}^{G(z)}\phi(z)$ (assuming it is unique) where $\phi(z)$ is the p.d.f. of $\mathrm{N}(0, I)$. If the maximum occurs at μ, then the new measure that is used for Z is $\mathrm{N}(\mu, I)$.

Once again, the intuition behind this procedure in Glasserman et al. (1999) is obtained from the zero-variance distribution. Note that the zero-variance distribution of Z is one that is proportional to $\mathrm{e}^{G(z)}\phi(z)$. The heuristic in Glasserman et al. (1999) is based on the idea that if one aligns the mode of the new normal distribution (i.e., μ) and the mode of $\mathrm{e}^{G(z)}\phi(z)$, then the two may also roughly have the same shape, thus approximately achieving the proportionality property. One can also see this if one approximates $G(z)$ by its first order Taylor series expansion around μ. Note that if $G(z)$ is exactly linear with slope a, then the zero-variance distribution can be easily derived as $\mathrm{N}(a, I)$. Also, in this case, it is easy to see that a minimizes $\mathrm{e}^{G(z)}\phi(z)$.

In the credit risk case, $G(z) = \ln \mathrm{P}(\mathcal{L}_m > qm|Z = z)$. Since $\mathrm{P}(\mathcal{L}_m > qm|Z = z)$ is usually not computable, one uses the upper bound obtained from the Markov inequality, i.e.,

$$G(z) = \ln \mathrm{P}(\mathcal{L}_m > qm|Z = z) \leqslant -\theta qm + \psi^{(m)}(\theta, z)$$

for all $\theta \geqslant 0$. As before, if we let $\theta_m(z)$ be the $\theta \geqslant 0$ that maximizes $-\theta qm + \psi^{(m)}(\theta, z)$ for a given z, and define $F_m(z) := -\theta_m(z)qm + \psi^{(m)}(\theta_m(z), z)$, then $G(z) = \ln \mathrm{P}(\mathcal{L}_m > qm|Z = z) \leqslant F_m(z)$. One can then use $F_m(z)$ as a close surrogate to $G(z)$, in order to determine the importance sampling change of measure for Z. Glasserman and Li (2005) develop some new asymptotic regimes and prove asymptotic optimality of the above procedure as $m \to \infty$, again for the homogeneous ($p_k = p$ and $c_k = 1$) single factor case.

Algorithms and asymptotic optimality results for the multi-factor, nonhomogeneous case have been analyzed in Glasserman et al. (2005). Another approach, but without any asymptotic optimality proof has been presented in Morokoff (2004). Algorithms for the "t-copula model" (in contrast to the Gaussian copula model) and related models, have been studied in Bassamboo et al. (2005c) and Kang and Shahabuddin (2005). Bassamboo et al. (2005c) develop sharp asymptotics for the probability of large losses and importance sampling techniques that have bounded relative error in estimating this probability. This analysis is extended to another related and popular performance measure, namely *expected shortfall* or the expected excess loss given that a large loss occurs, in Bassamboo et al. (2005a) (also see Merino and Nyefeler, 2004).

Acknowledgement

This work was partially supported by the National Science Foundation (USA) Grant DMI 03-00044.

References

Ahamed, T.P.I., Borkar, V.S., Juneja, S. (2006). Adaptive importance sampling technique for Markov chains using stochastic approximation. *Operations Research*, in press.

Anantharam, V., Heidelberger, P., Tsoucas, P. (1990). Analysis of rare events in continuous time Markov chains via time reversal and fluid approximation. Report RC 16280, IBM, Yorktown Heights, NY.

Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd edition. Wiley, New York.

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

Asmussen, S. (1985). Conjugate processes and the simulation of ruin problems. *Stochastic Process and Applications* 20, 213–229.

Asmussen, S. (1989). Risk theory in a Markovian environment. *Scandinavian Actuarial Journal* 1989, 69–100.

Asmussen, S. (2000). *Ruin Probabilities*. World Scientific, London.

Asmussen, S. (2003). *Applied Probability and Queues*, 2nd edition. Springer-Verlag, New York.

Asmussen, S., Binswanger, K. (1997). Simulation of ruin probabilities for subexponential claims. *ASTIN Bulletin* 27 (2), 297–318.

Asmussen, S., Kroese, D. (2004). Improved algorithms for rare-event simulation with heavy tails. Research report, Department of Mathematical Sciences, Aarhus University, Denmark.

Asmussen, S., Rubinstein, R.Y. (1995). Steady state rare-event simulation in queueing models and its complexity properties. In: Dshalalow, J.H. (Ed.), *Advances in Queueing: Theory, Methods and Open Problems*. CRC Press, New York, pp. 429–462.

Asmussen, S., Binswanger, K., Hojgaard, B. (2000). Rare-event simulation for heavy-tailed distributions. *Bernoulli* 6 (2), 303–322.

Asmussen, S., Kroese, D., Rubinstein, R. (2005). Heavy tails, importance sampling and cross entropy. *Stochastic Models* 21 (1), 57–76.

Avram, F., Dai, J., Hasenbein, J. (2001). Explicit solutions for variational problems in the quadrant. *Queueing Systems* 37, 261–291.

Bahadur, R., Rao, R.R. (1960). On deviations of the sample mean. *The Annals of Mathematical Statistics* 31, 1015–1027.

Bassamboo, A., Juneja, S., Zeevi, A. (2005a). Expected shortfall in credit porfolios with extremal dependence. In: Kuhl, M.E., Steiger, N.M., Armstrong, F.B., Joines, J.A. (Eds.), *Proceedings of the 2005 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 1850–1858.

Bassamboo, A., Juneja, S., Zeevi, A. (2005b). On the efficiency loss of state-dependent importance sampling in the presence of heavy tails. *Operations Research Letters*, in press.

Bassamboo, A., Juneja, S., Zeevi, A. (2005c). Portfolio credit risk with extremal dependence. Preprint.

Beck, B., Dabrowski, A., McDonald, D. (1999). A unified approach to fast teller queues and ATM. *Advances in Applied Probability* 31, 758–787.

Bertsekas, D., Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific, Athena, MA.

Blanchet, J., Glynn, P. (2005). Efficient simulation for the maximum of a random walk with heavy-tailed increments. Presentation at the 13th INFORMS Applied Probability Conference, July 6–8, 2005, Ottawa, Canada.

Bolia, N., Glasserman, P., Juneja, S. (2004). Function-approximation-based importance sampling for pricing American options. In: Ingalls, R.G., Rossetti, M.D., Smith, J.S., Peters, B.A. (Eds.), *Proceedings of the 2004 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 604–611.

Booth, T. (1985). Exponential convergence for Monte Carlo particle transport. *Transactions of the American Nuclear Society* 50, 267–268.

Boots, N., Shahabuddin, P. (2000). Simulating $GI/GI/1$ queues and insurance processes with subexponential distributions. In: Joines, J.A., Barton, R.R., Kang, K., Fishwick, P.A. (Eds.), *Proceedings of the 2000 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 656–665. Latest version: Research report, IEOR Department, Columbia University, New York.

Boots, N., Shahabuddin, P. (2001). Simulating ruin probabilities in insurance risk processes with subexponential claims. In: Peters, B.A., Smith, J.S., Medeiros, D.J., Rohrer, M.W. (Eds.), *Proceedings of the 2001 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 468–476.

Borkar, V. (2002). Q-learning for risk-sensitive control. *Mathematics of Operations Research* 27, 294–311.

Borkar, V.S., Meyn, S.P. (2002). Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research* 27, 192–209.

Borkar, V.S., Juneja, S., Kherani, A.A. (2004). Performance analysis conditioned on rare events: An adaptive simulation scheme. *Communications in Information* 3 (4), 259–278.

Bucklew, J.S. (1990). *Large Deviations Techniques in Decision, Simulation and Estimation*. Wiley, New York.

Chang, C.S., Heidelberger, P., Juneja, S., Shahabuddin, P. (1994). Effective bandwidth and fast simulation of ATM in tree networks. *Performance Evaluation* 20, 45–65.

Chistyakov, V.P. (1964). A theorem on sums of independent positive random variables and its applications to branching random processes. *Theory of Probabability and Applications* 9, 640–648.

Collamore, J.F. (1996). Hitting probabilities and large deviations. *The Annals of Probability* 24 (4), 2065–2078.

Collamore, J.F. (2002). Importance sampling techniques for the multidimensional ruin problem for general Markov additive sequences of random vectors. *The Annals of Applied Probability* 12 (1), 382–421.

De Boer, P. (2001). Analysis and efficient simulation of queueing models of telecommunication systems. PhD thesis, University of Twente.

De Boer, P., Nicola, V., Rubinstein, R. (2000). Adaptive importance sampling simulation of queueing networks. In: Joines, J.A., Barton, R.R., Kang, K., Fishwick, P.A. (Eds.), *Proceedings of the 2000 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 646–655.

De Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research* 134 (1), 19–67.

Dembo, A., Zeitouni, O. (1998). *Large Deviations Techniques and Applications*. Springer-Verlag, New York, NY.

Desai, P., Glynn, P. (2001). A Markov chain perspective on adaptive Monte Carlo algorithms. In: Peters, B.A., Smith, J.S., Medeiros, D.J., Rohrer, M.W. (Eds.), *Proceedings of the 2001 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 379–384.

Dupuis, P., Wang, H. (2004). Importance sampling, large deviations, and differential games. Preprint.

Embrechts, P., Kluppelberg, C., Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Springer-Verlag, Berlin, Heidelberg.

Fang, K.T., Kotz, S., Ng, K. (1987). *Symmetric Multivariate and Related Distributions*. Chapman and Hall, London.

Fishman, G. (2001). *Discrete-Event Simulation: Modeling, Programming, and Analysis*. Springer-Verlag, Berlin.

Frater, M. (1993). Fast simulation of buffer overflows in equally loaded networks. *Australian Telecommun. Res.* 27 (1), 13–18.

Frater, M., Lennon, T., Anderson, B. (1991). Optimally efficient estimation of the statistics of rare events in queueing networks. *IEEE Transactions on Automatic Control* 36, 1395–1405.

Glasserman, P. (2004). *Monte Carlo Methods in Financial Engineering*. Springer-Verlag, New York.

Glasserman, P., Kou, S. (1995). Analysis of an importance sampling estimator for tandem queues. *ACM TOMACS* 5, 22–42.

Glasserman, P., Li, J. (2005). Importance sampling for portfolio credit risk. *Management Science* 50 (11), 1643–1656.

Glasserman, P., Wang, Y. (1997). Counterexamples in importance sampling for large deviations probabilities. *The Annals of Applied Probability* 7, 731–746.

Glasserman, P., Heidelberger, P., Shahabuddin, P. (1999). Asymptotically optimal importance sampling and stratification for pricing path-dependent options. *Mathematical Finance* 9, 117–152.

Glasserman, P., Heidelberger, P., Shahabuddin, P. (2000). Variance reduction techniques for estimating value-at-risk. *Management Science* 46, 1349–1364.

Glasserman, P., Heidelberger, P., Shahabuddin, P. (2002). Portfolio value-at-risk with heavy-tailed risk factors. *Mathematical Finance* 9, 117–152.

Glasserman, P., Kang, W., Shahabuddin, P. (2005). Fast simulation of multifactor portfolio credit risk. Working paper, IEOR Department, Columbia University.

Glynn, P., Iglehart, D. (1989). Importance sampling for stochastic simulations. *Management Science* 35, 1367–1392.

Glynn, P., Whitt, W. (1992). The asymptotic efficiency of simulation estimators. *Operations Research* 40, 505–520.

Goyal, A., Lavenberg, S. (1987). Modeling and analysis of computer system availability. *IBM Journal of Research and Development* 31 (6), 651–664.

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

Grassberger, P. (2002). Go with the winners: A general Monte Carlo strategy. *Computer Physics Communications* 147 (1/2), 64–70.

Grassberger, P., Nadler, W. (2000). "Go with the winners" – Simulations. In: Hoffman, K.H., Schreiber, M. (Eds.), *Computational Statistical Physics: From Billards to Monte Carlo*. Springer-Verlag, Heidelberg, pp. 169–190.

Gupta, G., Finger, C., Bhatia, M. (1997). Credit metrics technical document. Technical report, J.P. Morgan and Co., New York.

Heidelberger, P. (1995). Fast simulation of rare events in queueing and reliability models. *ACM TOMACS* 5 (1), 43–85.

Heyde, C., Kou, S. (2004). On the controversy over tailweight of distributions. *Operations Research Letters* 32, 399–408.

Huang, Z., Shahabuddin, P. (2003). Rare-event, heavy-tailed simulations using hazard function transformations with applications to value-at-risk. In: Chick, S., Sanchez, P.J., Ferrin, D., Morrice, D. (Eds.), *Proceedings of the 2003 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 276–284. Latest version: Working paper, IEOR Department, Columbia University.

Huang, Z., Shahabuddin, P. (2004). A unified approach for finite dimensional, rare-event Monte Carlo simulation. In: Ignalls, R.G., Rossetti, M.D., Smith, J.S., Peters, B.A. (Eds.), *Proceedings of the 2004 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 1616–1624.

Ignatiouk-Robert, I. (2000). Large deviations of Jackson networks. *The Annals of Applied Probability* 10 (3), 962–1001.

Ignatyuk, I., Malyshev, V., Scherbakov, V. (1994). Boundary effects in large deviations problems. *Russian Mathematical Surveys* 49 (2), 41–99.

Juneja, S. (2001). Importance sampling and the cyclic approach. *Operations Research* 46 (4), 900–912.

Juneja, S. (2003). Efficient rare-event simulation using importance sampling: An introduction. In: Misra, J.C. (Ed.), *Computational Mathematics, Modelling and Algorithms*. Narosa Publishing House, New Delhi, pp. 357–396.

Juneja, S., Nicola, V. (2005). Efficient simulation of buffer overflow probabilities in Jackson networks with feedback. *ACM TOMACS* 15 (4), 281–315.

Juneja, S., Shahabuddin, P. (2001). Efficient simulation of Markov chains with small transition probabilities. *Management Science* 47 (4), 547–562.

Juneja, S., Shahabuddin, P. (2002). Simulating heavy-tailed processes using delayed hazard rate twisting. *ACM TOMACS* 12, 94–118.

Juneja, S., Shahabuddin, P., Chandra, A. (1999). Simulating heavy-tailed processes using delayed hazard-rate twisting. In: Farrington, P.A., Nembhard, H.B., Sturrock, D.T., Evans, G.W. (Eds.), *Proceedings of the 1999 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 420–427.

Juneja, S., Karandikar, R., Shahabuddin, P. (2005). Tail asymptotics and fast simulation of delay probabilities in stochastic PERT networks. *ACM TOMACS*, undergoing review.

Kang, W., Shahabuddin, P. (2005). Fast simulation for multifactor portfolio credit risk in the *t*-copula model. In: Kuhl, M.E., Steiger, N.M., Armstrong, F.B., Joines, J.A. (Eds.), *Proceedings of the 2005 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 1859–1868.

Kollman, C., Baggerly, K., Cox, D., Picard, R. (1999). Adaptive importance sampling on discrete Markov chains. *The Annals of Applied Probability* 9, 391–412.

Kontoyiannis, I., Meyn, S. (2003). Spectral theory and limit theorems for geometrically ergodic Markov processes. *The Annals of Applied Probability* 13, 304–362.

Kotz, S., Kozubowski, T., Podgorski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*. Birkhäuser, Boston.

Kroese, D., Nicola, V. (1999). Efficient estimation of overflow probabilities in queues with breakdowns. *Performance Evaluation* 36/37, 471–484.

Kroese, D., Nicola, V. (2002). Efficient simulation of a tandem Jackson network. *ACM TOMACS* 12, 119–141.

Kroese, D., Rubinstein, R. (2004). The transform likelihood ratio method for rare-event simulation with heavy tails. *Queueing Systems* 46, 317–351.

Kushner, H., Yin, G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, New York.

L'Ecuyer, P., Champoux, Y. (2001). Estimating small cell-loss ratios in atm switches via importance sampling. *ACM TOMACS* 11 (1), 76–105.

Lehtonen, T., Nyrhinen, H. (1992a). Simulating level-crossing probabilities by importance sampling. *Advances in Applied Probability* 24, 858–874.

Lehtonen, T., Nyrhinen, H. (1992b). On asymptotically efficient simulation of ruin probabilities in a Markovian environment. *Scandinavian Actuarial Journal* 1992, 60–75.

Luenberger, D. (1984). *Linear and Non-Linear Programming*, 2nd edition. Addison–Wesley, Reading, MA.

Merino, S., Nyefeler, M. (2004). Applying importance sampling for estimating coherent credit risk contributions. *Quantitative Finance* 4, 199–207.

Merton, R. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29, 449–470.

Morokoff, W. (2004). An importance sampling method for portfolios of credit risky assets. In: Ingalls, R.G., Rossetti, M.D., Smith, J.S., Peters, B.A. (Eds.), *Proceedings of the 2004 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 1668–1676.

Nakayama, M., Nicola, V., Shahabuddin, P. (2001). Techniques for fast simulation of models of highly dependable systems. *IEEE Transactions on Reliability* 50, 246–264.

Ney, P. (1983). Dominating points and the asymptotics of large deviations for random walks on $\Re^d$. *The Annals of Probability* 11, 158–167.

Pakes, A. (1975). On the tails of waiting time distribution. *Journal of Applied Probability* 12, 555–564.

Parekh, S., Walrand, J. (1989). A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control* 34 (1), 54–66.

Randhawa, R.S., Juneja, S. (2004). Combining importance sampling and temporal difference control variates to simulate Markov chains. *ACM TOMACS* 14, 1–30.

Royden, H.L. (1984). *Real Analysis*. Prentice Hall, New York.

Rubinstein, R.Y. (1997). Optimization of computer simulation models with rare events. *European Journal of Operations Research* 99, 89–112.

Rubinstein, R.Y. (1999). Rare-event simulation via cross-entropy and importance sampling. In: *Second Workshop on Rare Event Simulation, RESIM'99*, *Enshede*, *The Netherlands*, pp. 1–17.

Rubinstein, R.Y., Kroese, D.P. (2004). *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte Carlo Simulation and Machine Learning*. Springer-Verlag, New York.

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

Sadowsky, J. (1996). On Monte Carlo estimation of large deviations probabilities. *The Annals of Applied Probability* 6 (2), 399–422.

Sadowsky, J.S., Bucklew, J. (1990). On large deviation theory and asymptotically efficient Monte Carlo estimation. *IEEE Transactions on Information Theory* 36 (3), 579–588.

Sadowsky, J.S., Szpankowski, W. (1995). The probability of large queue lengths and waiting times in a heterogeneous multiserver queue part I: Tight limits. *Advances in Applied Probability* 27, 532–566.

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

Shahabuddin, P. (2005). An approach to simulation of random walks with heavy-tailed increments. Technical report, Department of IEOR, Columbia University.

Shahabuddin, P., Woo, B. (2004). Conditional importance sampling with applications to value-at-risk simulations. Working paper, IEOR Department, Columbia University.

Shwartz, A., Weiss, A. (1995). *Large Deviations for Performance Analysis*. Chapman and Hall, New York.

Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *The Annals of Statistics* 4, 673–684.

Sigman, K. (1999). A primer on heavy-tailed distributions. *Queueing Systems* 33, 261–275.

Su, Y., Fu, M. (2000). Importance sampling in derivatives security pricing. In: Joines, J.A., Barton, R.R., Jang, K., Fishwick, P.A. (Eds.), *Proceedings of the 2000 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 587–596.

Su, Y., Fu, M.C. (2002). Optimal importance sampling in securities pricing. *Journal of Computational Finance* 5, 27–50.

Szechtman, R., Glynn, P. (2002). Rare-event simulation for infinite server queues. In: Yucesan, E., Chen, H., Snowdon, J.L., Charnes, J.M. (Eds.), *Proceedings of the 2002 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 416–423.

Tong, Y. (1990). *The Multivariate Normal Distribution*. Springer-Verlag, New York.

Vázquez-Abad, F., Dufresne, D. (1998). Accelerated simulation for pricing Asian options. In: Medeiros, D.J., Watson, E.F., Carson, M.S., Manivannan, J.S. (Eds.), *Proceedings of the 1998 Winter Simulation Conference*. IEEE Press, Piscataway, NJ, pp. 1493–1500.