Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

ADAPTIVE MONTE CARLO METHODS FOR RARE EVENT SIMULATIONS

Ming-hua Hsieh

Department of Management Information Systems
National Chengchi University
Wenshan, Taipei 11623, TAIWAN

ABSTRACT

We review two types of adaptive Monte Carlo methods for rare event simulations. These methods are based on importance sampling. The first approach selects importance sampling distributions by minimizing the variance of importance sampling estimator. The second approach selects importance sampling distributions by minimizing the cross entropy to the optimal importance sampling distribution. We also review the basic concepts of importance sampling in the rare event simulation context. To make the basic concepts concrete, we introduce these ideas via the study of rare events of M/M/1 queues.

1 INTRODUCTION

Rare events, although are seldom happened as its name suggests, are important when they do happen in many application areas. For example, the buffer overflow event is rare in a high quality telecommunication network, but is significant when it happens; A system break down event is rare in a fault-tolerant computing system, but has a consequential effect when it happens. Therefore, accurate estimation of the probabilities of such rare events is important. However, if the probabilities of rare events are really small, estimation of these probabilities is often computationally intractable when studied using conventional Monte Carlo simulation. Therefore, powerful efficiency improvement techniques (see, e.g. (Glynn (1994)) and (Bratley, Fox, and Schrage (1987))) are needed. The most suitable technique for rare event simulations is importance sampling (see, e.g. Hammersley and Handscomb (1965), and we will review the basic concepts in Section 3.) When applied appropriately, importance sampling can improve the efficiency many orders of magnitude. Unfortunately, it is not a simple task to apply importance sampling appropriately on rare event simulations. The main difficulty lies in the selection of an effective importance sampling distribution. The selection usually requires simulationists have a good understanding of the structure of the system being simulated.

To overcome this difficulty, researchers have developed adaptive procedures for selecting effective importance sampling distributions. This paper's main purpose is to review these procedures. We reviewed two adaptive importance sampling methods in this paper. The first method selects effective importance sampling distributions by minimizing the variance of importance sampling estimator. The second method selects effective importance sampling distributions by minimizing the cross entropy to the optimal importance sampling distribution (see Section 3.2). For applications of these methods in selecting effective importance sampling distributions for queueing models and financial derivative models, see the citations of Section 4.

The rest of the paper is organized as follows. In Section 2, we describe a simple M/M/1 queue model and the rare event of interest. In Section 3, we review the basic concepts of importance sampling via the study of rare events of M/M/1 queues. Several criteria of evaluating the goodness of importance sampling estimators are given. In section 4, we reviewed two types of adaptive importance sampling methods for rare event simulations. The first method selects importance sampling distributions by minimizing the variance of importance sampling estimator. The second one selects importance sampling distributions by minimizing the cross entropy to the optimal importance sampling distribution. Finally, the paper is summarized in Section 5. This section remarks some properties of these two types of methods.

2 A SIMPLE RARE EVENT SIMULATION PROBLEM

Let us consider a stable M/M/1 queue with arrival rate $\lambda = 1$, service rate $\mu > 1$, and the buffer limit of the system $K > 1$, see Figure 1. Let $X(t)$ be the number of customers in the system at time t. Then, $X = (X(t) : t \geq 0)$ is a continuous time Markov chain with state space $S = \{0, 1, 2, \ldots K\}$.

Suppose that $X(0) = 1$ and we are interested in using importance sampling to estimate

$$\gamma_K = P(X(T) = K), \tag{1}$$

where

$$T = \inf\{t > 0 : X(t) = 0 \text{ or } X(t) = K\}.$$

It is well known that X is an irreducible, positive recurrent Markov chain and T is a stopping time, also known as Markov time (see p. 255 & p. 318 of Karlin and Taylor (1975)).

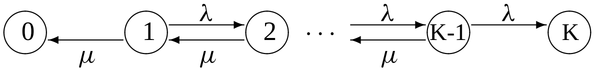

Figure 1: Transition Diagram of the M/M/1 Queue Example

If K is large, then γ_K is a small number. Thus, estimating γ_K in (1) is a rare event simulation problem.

For this simple problem, the analytical solution for γ_K is known. In particular,

$$\gamma_K = \frac{\mu - 1}{\mu^K - 1}. \tag{2}$$

Of course with a known analytical result there is no need to simulate. We just use this problem as a vehicle for the basic ideas to follow. Throughout this paper, the concepts of importance sampling and adaptive importance sampling methods will be introduced via this simple problem.

3 THE BASIC CONCEPTS OF IMPORTANCE SAMPLING

3.1 Importance Sampling

Since X is an irreducible, positive recurrent chain and T is a stopping time, it is well known that $P(\{\omega : T(\omega) < \infty\}) = 1$. Thus, without loss of generality, we can choose $\Omega = \{\omega : T(\omega) < \infty\}$ being the sample space. Let $f(\cdot)$ denote the (original) density function of X given $X(0) = 1$ and let

$$A = \{\omega : X(T(\omega)) = K\}.$$

Then, we can represent γ_K as

$$\int I(\omega \in A) f(\omega) d\omega = E_f[I(A)],$$

where $I(\cdot)$ is the indicator function.

To estimate γ_K via simulation, the direct approach would be to generate n independent sample paths, $\omega_1, \ldots, \omega_n$, from the probability density function $f(\cdot)$ and form the estimator

$$\hat{\gamma}_K = \frac{1}{n} \sum_{k=1}^{n} I(\omega_k \in A).$$

By the central limit theorem,

$$\sqrt{n}(\hat{\gamma}_K - \gamma_K) \Rightarrow \sqrt{\gamma_K(1-\gamma_K)} N(0, 1),$$

as $n \to \infty$, where $N(0, 1)$ denotes a normal random variable with mean 0 and variance 1. Thus, to construct a 95% confidence interval for γ_K with relative half-length of 1%, we need sample size $n \approx 1/(0.01^2) \times 1.96^2 \times (1-\gamma_K)/\gamma_K$. Therefore, if $\gamma_K = 10^{-9}$, then $n \approx 3.84 \times 10^{13}$. This demonstrates the main problem of rare event simulations.

Let $g(\cdot)$ be a density function such that $\omega \in A$ and $f(\omega) > 0$ implies $g(\omega) > 0$. Then we have another representation for γ_K:

$$\begin{aligned}\gamma_K &= \int I(\omega \in A) \frac{f(\omega)}{g(\omega)} g(\omega) d\omega \\ &= \int I(\omega \in A) L_{f,g}(\omega) g(\omega) d\omega \\ &= E_g[I(A) L_{f,g}],\end{aligned} \tag{3}$$

where $L_{f,g}(\omega) = f(\omega)/g(\omega)$ is called the likelihood ratio and $E_g(\cdot)$ denotes the expectation under $g(\cdot)$.

Identity (3) suggests an alternative estimation scheme: generate n samples, $\omega_1, \ldots, \omega_n$, from $g(\cdot)$. By (3),

$$\hat{\gamma}_{Kg} = \frac{1}{n} \sum_{k=1}^{n} L_{f,g}(\omega_k) I(\omega_k \in A) \tag{4}$$

is an unbiased estimate of γ_K. This alternative estimation scheme is called *importance sampling*. To apply importance sampling to more general stochastic systems including discrete-time Markov chains (DTMC's), continuous-time Markov chains (CTMC's), and generalized semi-Markov processes (a mathematical formalization of discrete-event simulations), consult Glynn and Iglehart (1989).

Before proceeding to next subsection, let us derive the explicit formula of the likelihood ratio for our M/M/1 queue example. A typical sample path ω of X is

$$((X_0, h_0), (X_1, h_1), \ldots, (X_{N-1}, h_{N-1}), X_N)$$

where N is the number of jumps before the stochastic process X hits 0 or K, $X_0, X_1, \cdots, X_N$ is the sequence of states of the embedded discrete time Markov chain, and h_n is the holding time in state X_n for $n = 0, 1, \ldots, N-1$. Thus, for the probability density function $f(\omega)$ of the sample path ω of the process X for which $P_f(\cdot, \cdot)$ denotes the transition probabilities of the embedded discrete time Markov chain, we have

$$f(\omega) = \prod_{n=0}^{N-1} (1+\mu) e^{-(1+\mu)h_n} P_f(X_n, X_{n+1}). \qquad (5)$$

Now, let us consider a generalized M/M/1 queue, whose arrival rates and service rates vary. In particular, its arrival rate is λ_k and service rate is μ_k when it is at state k, $k = 1, 2, \cdots, K-1$; see Figure 2. For such a generalized M/M/1 queue, the density $g(\omega)$ is equal to

$$\prod_{n=0}^{N-1} (\lambda_{X_n} + \mu_{X_n}) e^{-(\lambda_{X_n}+\mu_{X_n})h_n} P_g(X_n, X_{n+1}), \qquad (6)$$

where $P_g(\cdot,\cdot)$ is the transition probabilities of the embedded DTMC of this generalized M/M/1 queue.

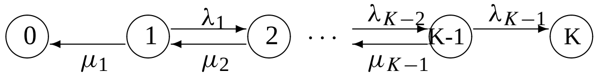

Figure 2: Transition Diagram of the Generalized M/M/1 Queue Example

Thus, if we choose such a generalized M/M/1 to do importance sampling, then the likelihood ratio of ω

$$L_{f,g}(\omega) = \frac{\prod_{n=0}^{N-1} (1+\mu) e^{-(1+\mu)h_n} P_f(X_n, X_{n+1})}{\prod_{n=0}^{N-1} (\lambda_{X_n} + \mu_{X_n}) e^{-(\lambda_{X_n}+\mu_{X_n})h_n} P_g(X_n, X_{n+1})}. \qquad (7)$$

See Glynn and Iglehart (1989) for explicit formulas of likelihood ratios for a variety of more general stochastic processes.

3.2 The Optimal Importance Sampling Distribution

Is importance sampling always better than direct simulation? This answer depends on the choice of importance sampling distribution g. An ideal g is a distribution which has the property $\mathrm{Var}_g[I(A)L_{f,g}] \ll \mathrm{Var}_f[I(A)]$, where $\mathrm{Var}_g(\cdot)$ and $\mathrm{Var}_f(\cdot)$ denote the variance under distributions $g(\cdot)$ and $f(\cdot)$, respectively. And the best g is a distribution which has the property $\mathrm{Var}_g[I(A)L_{f,g}] = 0$.

Definition 1 *A distribution g is called* the optimal importance sampling distribution *of distribution f if*

$$\mathrm{Var}_g[I(A)L_{f,g}] = 0,$$

where A is the rare event of interest.

Does such g exist? If we define the following probability density function on the sample path ω,

$$g^*(\omega) = \frac{f(\omega) I(\omega \in A)}{\gamma_K}, \qquad (8)$$

then we have

$$\begin{aligned} E_{g^*}[(I(A)L_{f,g^*})^2] &= E_f[I(A)L_{f,g^*}] \\ &= \int I(\omega \in A) \frac{f(\omega)}{g^*(\omega)} f(\omega) d\omega \\ &= \gamma_K^2. \end{aligned}$$

Thus,

$$\begin{aligned} &\mathrm{Var}_{g^*}[I(A)L_{f,g^*}] = \\ &\qquad E_{g^*}[(I(A)L_{f,g^*})^2] - E^2_{g^*}[I(A)L_{f,g^*}] = 0. \end{aligned}$$

Therefore, g^* is the optimal importance sampling distribution.

In most practical problems, the optimal importance sampling distribution is not achievable, since it contains the quantity γ_K, which is unknown. However, it is computable in our simple problem. We will demonstrate how to compute the optimal importance sampling distribution in this simple M/M/1 queue example.

Example 1 [The Optimal Importance Sampling Distribution] From (8), it is easy to see that $L_{f,g^*}(\omega) = \gamma_K$ (a constant) for $\omega \in A$. Now, consider two sample paths $\omega_1, \omega_2 \in A$

$$\begin{aligned} \omega_1 = ((1,1), (2,1), \ldots, \\ (k,1), (k+1,1), \cdots, (K-1,1), K) \end{aligned}$$

and

$$\begin{aligned} \omega_2 = ((1,1), \ldots, (k,1), (k+1,1), \\ (k,1), (k+1,1) \cdots, (K-1,1), K) \end{aligned}$$

where $2 \le k \le K-2$. If we choose a generalized M/M/1 queue to do importance sampling, then by (7), we have

$$L_{f,g}(\omega_1) = \frac{e^{-(K-1)(1+\mu)}}{\prod_{n=1}^{K-1} \lambda_n e^{-(\lambda_n+\mu_n)}}$$

and

$$L_{f,g}(\omega_2) = \frac{\mu e^{-(K+1)(1+\mu)}}{\lambda_k e^{-(\lambda_k+\mu_k)} \mu_{k+1} e^{-(\lambda_k+\mu_k)} \prod_{n=1}^{K-1} \lambda_n e^{-(\lambda_n+\mu_n)}}.$$

We can substantially simplify our expressions for $L_{f,g}(\omega_2)$ and $L_{f,g}(\omega_2)$ by setting

$$\lambda_k + \mu_k = 1 + \mu, \quad 1 \le k \le K - 1;$$

and this yields

$$\begin{aligned} L_{f,g}(\omega_1) &= \frac{1}{\prod_{n=1}^{K-1} \lambda_n} \\ L_{f,g}(\omega_2) &= \frac{\mu}{\lambda_k \mu_{k+1} \prod_{n=1}^{K-1} \lambda_n}. \end{aligned}$$

It is easy to see $L_{f,g}(\omega_1) = L_{f,g}(\omega_2)$ if $\mu_{k+1} = \mu/\lambda_k$. Therefore, the recursion

$$\mu_k = \frac{\mu}{\lambda_{k-1}}, \quad \lambda_k = 1 + \mu - \mu_k, \quad 2 \le k \le K - 1$$

is necessary for a generalized M/M/1 queue being the optimal importance sampling distribution. With suitably chosen boundary condition, above recursion does define the optimal importance sampling distribution:

$$\begin{aligned} \mu_1 &= 0, \quad \lambda_1 = 1 + \mu \\ \mu_k &= \frac{\mu}{\lambda_{k-1}}, \quad \lambda_k = 1 + \mu - \mu_k, \quad 2 \le k \le K - 1. \end{aligned} \tag{9}$$

Note that $\mu_1 = 0$ is necessary for the generalized M/M/1 queue to serve as the optimal importance sampling distribution. Since, otherwise there exists $\omega \notin A$ such that $P(\omega) > 0$.

3.3 Asymptotically Optimal Importance Sampling Distributions

In the queueing and random walks literature, there is a notion called asymptotically optimal for measuring the effectiveness of an importance sampling distribution in the rare event simulation context. See, e.g. Siegmund (1976), Lehtonen and Nyrhinen (1992), and Heidelberger (1995).

Definition 2 *Let $A_K = \{\omega : X(T(\omega)) = K\}$. If*

$$\lim_{K\to\infty} \frac{\log E_g[I(A_K)L_{f,g}^2]}{\log \gamma_K} = 2, \tag{10}$$

we call g an asymptotically optimal importance sampling distribution.

Note that

$$\begin{aligned} E_g[I(A_K)L_{f,g}^2] &= \mathrm{Var}[I(A_K)L_{f,g}] + (E_g[I(A_K)L_{f,g}])^2 \\ &= \mathrm{Var}[I(A_K)L_{f,g}] + \gamma_K^2 \\ &\ge \gamma_K^2. \end{aligned}$$

In view of the preceding development, we obtain the following inequality,

$$\log E_g[I(A_K)L_{f,g}^2] \ge 2 \log \gamma_K,$$

for all valid distribution g. Letting $K \to \infty$ yields

$$\liminf_{K\to\infty} \frac{\log E_g[I(A_K)L_{f,g}^2]}{\log \gamma_K} \ge 2.$$

Thus, asymptotically optimal importance sampling distributions are optimal on the logarithmic scale.

Example 2 [Asymptotical Optimality]

Let us consider a M/M/1 queue with with arrival rate μ and service rate 1, i.e., by switching the service rate and arrival rate of the original M/M/1 queue. If we use such a M/M/1 queue to do importance sampling, then

$$L_{f,g}(\omega) = \frac{1}{\mu^{K-1}},$$

for all $\omega \in A$, i.e., $L_{f,g}(\omega)$ is constant for all $\omega \in A$. Now,

$$\begin{aligned} E_g[I(A_K)L_{f,g}^2] &= \frac{1}{\mu^{2(K-1)}} E_g[I(A_K)] = \frac{1}{\mu^{K-1}} \frac{\mu - 1}{\mu^K - 1} \\ \gamma_K &= \frac{\mu - 1}{\mu^K - 1}. \end{aligned}$$

Thus,

$$\lim_{K\to\infty} \frac{\log E_g[I(A_K)L_{f,g}^2]}{\log \gamma_K} = 2.$$

Such a change-of-measure is, therefore, an asymptotically optimal importance sampling distribution. This type of simple but effective change-of-measures exist in more general setting, see Heidelberger (1995) for a complete survey.

It is interesting to note that if the boundary condition of (9) is set to $\mu_1 = 1$ and $\lambda_1 = \mu$, the resulting M/M/1 queue is this asymptotically optimal one.

3.4 Bounded Relative Error

Definition 3 *Let $A_K = \{\omega : X(T(\omega)) = K\}$. If*

$$\limsup_{K\to\infty} \frac{\sqrt{\mathrm{Var}_g[I(A_K)L_{f,g}]}}{\gamma_K} < \infty \tag{11}$$

we call importance sampling distribution g has bounded relative error property.

Importance sampling distributions with bounded relative error property are desirable, since they require only a finite number of samples to construct a confidence interval with a given precision, no matter how small the γ_K is.

Example 3 [Bounded Relative Error]
The importance sampling distribution of Example 2 has bounded relative error property, since

$$\frac{\sqrt{\mathrm{Var}_g[I(A_K)L_{f,g}]}}{\gamma_K} = \sqrt{\frac{u^{K-1}-1}{\mu^{K-1}(\mu-1)}} < \frac{1}{\sqrt{\mu-1}}.$$

4 ADAPTIVE IMPORTANCE SAMPLING METHODS

To use importance sampling in rare event simulations, it is good to have an importance sampling distribution with asymptotical optimality or bounded relative error property. However, to have such an importance sampling distribution, it is usually required to have a good understanding of the large deviation behavior of the rare event of interest; this is the major obstacle to wide-spread application of importance sampling in rare event simulations. (Large deviations is a body of asymptotic theory which may be used to obtain the rare event asymptotics that we are interested in; cf. Bucklew (1990) and Dembo and Zeitouni (1993) for general background.) Thus, it is nice to have more automatical ways of finding good importance sampling distributions, regardless of what rare event problem we are interested. In this section, we review two types of adaptive importance sampling methods, which are to serve this need.

The effectiveness of the importance sampling estimator depends on the choice of importance sampling distribution g. To make the selection simpler, we usually only consider a family of importance sampling distributions parameterized by $\theta \in \Theta \subseteq \Re^d$; e.g., the family of exponential twisting distributions.

4.1 Approach via Minimizing Estimator's Variance

The most direct measure of effectiveness of an estimator is its variance. From Section 3, we know the variance of an importance sampling estimator is $1/m$ times $\mathrm{Var}_g[I(A)L_{f,g}]$ if the estimator is computed by m independent copies of $I(A)L_{f,g}$ sampling from the importance sampling distribution g. Let f_θ denote the family of distributions. Then, selecting the best importance sampling distribution from f_θ can be formulated as

$$\min_\theta \mathrm{Var}_{f_\theta}[I(A)L_\theta], \tag{12}$$

where $L_\theta(\omega) = f(\omega)/f_\theta(\omega)$. But

$$\begin{aligned}\mathrm{Var}_{f_\theta}[I(A)L_\theta] &= E_{f_\theta}[I(A)L_\theta^2] - (E_{f_\theta}[I(A)L_\theta])^2 \\ &= E_{f_\theta}[I(A)L_\theta^2] - \gamma_K^2.\end{aligned}$$

Thus, the variance-minimization problem (12) is is easily seen to be equivalent to

$$\min_\theta E_{f_\theta}[I(A)L_\theta^2]. \tag{13}$$

How does one compute an (approximate) minimizer of (13)? Since (13) is a stochastic optimization problem, traditional stochastic approximation algorithms, Robbins-Monro algorithm (Robbins and Monro (1951)) and Kiefer-Wofowitz algorithm (Kiefer and Wolfowitz (1952)) comes naturally for use.

R-M algorithm basically is the following recursion

$$\theta_{n+1} = \Pi_\Theta(\theta_n - \frac{a}{n+1}\widehat{\nabla h}(\theta_n)), \tag{14}$$

where Π_Θ is the projection operator onto Θ, $h(\cdot)$ is the objective function ($E_{f_\theta}[I(A)L_\theta^2]$ in our example) and $\widehat{\nabla h}(\theta_n)$ is an estimate of ∇h at θ_n. There exist several different approaches for obtaining the gradient estimation $\widehat{\nabla h}(\theta_n)$: infinitesimal perturbation analysis (Glasserman (1991)), likelihood ratio methods (Glynn (1986), Glynn (1990)), Conditional Monte Carlo (Fu and Hu (1997)), and the "push-out" approach (Rubinstein (1992)).

K-W algorithm also uses recursion (14). The difference between these two algorithms is on the method of estimating $\nabla h(\cdot)$. K-W algorithm use finite differences to estimate $\nabla h(\cdot)$.

Using importance sampling for accelerating simulation by finding an approximate minimizer of (13) has been applied in various applications, especially in queueing and reliability models; see, e.g. Al-Qaq, Devetsikiotis, and Townsend (1995), Devetsikiotis and Townsend (1993a, 1993b), Rubinstein (1997, 1999), and Rubinstein and Melamed (1998). Similar idea has also been applied to speeding up the simulation for pricing financial derivative, such as (out-of-the-money) Asian options. See Su and Fu (2000) and Vazquez-Abad and Dufresne (1998).

It is sometimes advantageous to rewrite (13) as

$$\min_\theta E_f[I(A)L_\theta]. \tag{15}$$

See Su and Fu (2000) for a successful example of using (15).

4.2 Approach via Minimizing Cross Entropy

4.2.1 Cross Entropy

Given a probability density function f, cross entropy defines a measurement of "distance to f". Let g be a probability density function defined on the same sample space such that $f(\omega) > 0$ implies $g(\omega) > 0$. Then the *cross entropy*, also known as *relative entropy* or *Kullback Leibler distance*, of the probability density function g with respect to the probability density function f is

$$\begin{aligned} D(f,g) &= \int \log\left(\frac{f(\omega)}{g(\omega)}\right) f(\omega)d\omega \\ &= E_f[\log(L_{f,g})]. \end{aligned} \tag{16}$$

For more details on this definition, see Kapur and Kesavan (1992). However, beware that this definition of cross entropy is not universal. For example, Jelinek (1997) defines cross entropy as

$$H(f,g) = -\int \log(g(\omega)) f(\omega)d\omega.$$

There are some important properties of $D(\cdot)$:

1. $D(\cdot)$ is non-symmetric; i.e., $D(f,g) \neq D(g,f)$
2. $D(f,g) \geq 0$
3. $D(f,f) = 0$

Since $D(\cdot)$ measures the distance between distributions, it is reasonable to expect $D(\cdot)$ can be used to select importance sampling distributions. In particular, we want to find a distribution g such that g is the minimizer of

$$\min_g \int \log\left(\frac{g^*(\omega)}{g(\omega)}\right) g^*(\omega)d\omega.$$

Of course, g^* solves this problem. However, g^* is usually unattainable. So, the candidate distributions is again a family of importance sampling distributions parameterized by $\theta \in \Theta \subseteq \Re^d$. Thus, the problem to be solved becomes

$$\min_\theta \int \log\left(\frac{g^*(\omega)}{f_\theta(\omega)}\right) g^*(\omega)d\omega = E_{g^*}[\log L_{g^*,f_\theta}], \tag{17}$$

where $L_{g^*,f_\theta}(\omega) = g^*(\omega)/f_\theta(\omega)$ is the likelihood ratio.

But

$$\begin{aligned} D(g^*, f_\theta) &= \int \log\left(\frac{g^*(\omega)}{f_\theta(\omega)}\right) g^*(\omega)d\omega \\ &= \int g^*(\omega)\log g^*(\omega)d\omega - \int g^*(\omega)\log f_\theta(\omega)d\omega \\ &= \int g^*(\omega)\log g^*(\omega)d\omega + H(g^*, f_\theta). \end{aligned}$$

Since the first term is independent of θ, the minimizer of $H(g^*, f_\theta)$ is also a minimizer of (17).

We obtain the following alternative formulation of the function $H(g^*, f_\theta)$:

$$\begin{aligned} H(g^*, f_\theta) &= -\int \log(f_\theta(\omega))g^*(\omega)d\omega \\ &= -\int \log(f_\theta(\omega))f(\omega)I(\omega \in A)/\gamma_K d\omega \\ &= -\frac{1}{\gamma_K}\int \log(f_\theta(\omega))I(\omega \in A)f(\omega)d\omega \\ &= -\frac{1}{\gamma_K}E_f[I(A)\log(f_\theta)]. \end{aligned}$$

Thus, the maximizer of

$$\max_\theta E_f[I(A)\log(f_\theta)] \tag{18}$$

is also the minimizer of (17).

4.2.2 Algorithm

Based on (18), it is straightforward to derive an iterative procedure for computing an approximate minimizer of (17). The key idea is to express (17) as

$$\max_\theta E_{f_{\theta'}}[I(A)L_{\theta'}\log(f_\theta)], \tag{19}$$

for $\theta' \in \Theta$, where $L_{\theta'}(\omega) = f(\omega)/f_{\theta'}(\omega)$ for $\omega \in A$.

Combine (18) and (19), the iterative procedure is now clear:

1. Select an initial guess θ_0 of (18); set $n = 0$
2. Compute an approximate minimizer of (19) with $\theta' = \theta_n$
3. $\theta_{n+1} \leftarrow$ approximate minimizer computed in Step 2; $n \leftarrow n+1$
4. (convergence test) If $\|\theta_n - \theta_{n-1}\| < \epsilon$ (ϵ is a small positive number), stop; otherwise, goto Step 2

This adaptive approach for minimizing cross entropy has been adopted in Lieber, Rubinstein and Elmakis (1997), Rubinstein (1997, 1999), and de Boer, Nicola and Rubinstein (2000).

5 CONCLUDING REMARKS

We have reviewed the basic concepts of importance sampling and selection criteria of good importance sampling distributions in rare event simulation context. Also, we have reviewed two adaptive importance sampling methods in the literature. In this section, we will emphasize some properties of these two methods.

Both methods adaptively look for parameters which let an importance sampling distribution optimal in their settings.

But cross entropy method has an advantage on computing an approximate solution on each iteration for certain stochastic models. For example, the optimal transition probabilities of DTMC can be computed analytically because of the logarithm of likelihood ratio; see Section 3.1 of de Boer, Nicola and Rubinstein (2000) for details.

The key optimization problem (18) in cross entropy method is equivalent to

$$\min_{\theta} E_f[I(A)\log(L_\theta)].$$

Since

$$\begin{aligned} &\arg\max_{\theta} E_f[I(A)\log(f_\theta)] \\ =&\arg\min_{\theta} E_f[-I(A)\log(f_\theta)] \\ =&\arg\min_{\theta} E_f[I(A)\log(f) - I(A)\log(f_\theta)] \\ =&\arg\min_{\theta} E_f[I(A)\log(L_\theta)]. \end{aligned}$$

In other words, under original density f, minimizing estimator's variance is equivalent to minimize the expected *likelihood ratio* conditioned on the rare event happens; and minimizing cross entropy is equivalent to minimize the expected *logarithm of likelihood ratio* conditioned on the rare event happens. Therefore, minimizing cross entropy in some sense is close to, but definitely is different from minimizing estimator's variance.

In terms of estimator's variance, cross entropy method does not seek for the optimal solution. Although intuitively, the optimizers of both methods are close to each other. It would be beneficial to know how close they are.

REFERENCES

Al-Qaq, W. A., M. Devetsikiotis, and J. K. Townsend. 1995. Stochastic gradient optimization of importance sampling for the efficient simulation of digital communication systems. *IEEE Transactions on Communications*, 43:2975–2985.

Bratley, P., B. Fox, and L. Schrage. 1987. *A Guide to Simulation*. Springer-Verlag, New York, second edition.

Bucklew, J. 1990. *Large Deviation Techniques in Decision, Simulation, and Estimation*. John Wiley and Sons, Inc., New York.

de Boer, P.T., V. F. Nicola and R. Y. Rubinstein, 2000. Adaptive importance sampling simulation of queueing networks. *Proceedings of the 2000 Winter Simulation Conference*, J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds., 646–655. Institute of Electrical and Electronics Engineers.

Dembo, A. and O. Zeitouni. 1993. *Large Deviation Techniques and Applications*. Jones and Bartlett, Boston.

Devetsikiotis, M. and J. K. Townsend. 1993a. An algorithmic approach to the optimization of importance sampling parameters in digital communication system simulation. *IEEE Transactions on Communications* 41:1464–1473.

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

Fu, M. C., and J. Q. Hu. 1997. *Conditional Monte Carlo: Gradient estimation and optimization applications*. Kluwer Academic Publisher, Boston.

Glasserman, P. 1991. *Gradient estimation via perturbation analysis*. Boston: Kluwer.

Glynn, P. W. 1986. Stochastic approximation for Monte Carlo optimization. In *Proceedings of the 1986 Winter Simulation Conference*, 356–365.

Glynn, P. W. 1990. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM* 33:75–84.

Glynn, P. W. 1994. Efficiency improvement techniques. *Annals of Oper. Res.*, 53:175–197.

Glynn, P. W. and D. L. Iglehart. 1989. Importance sampling for stochastic simulations. *Manage. Sci.*, 35:1367–1392.

Hammersley, J. M. and D. C. Handscomb. 1965. *Monte Carlo Methods*. Methuen & Co. Ltd., London.

Heidelberger, P. 1995. Fast simulation of rare events in queueing and reliability models. *ACM Trans. on Modeling and Computer Simulation*, 5:43–85.

Lehtonen, T. and H. Nyrhinen. 1992. Simulating level-crossing probabilities by importance sampling. *Adv. in Appl. Probab.*, 24(4):858–874.

Lieber, D., R.Y. Rubinstein and D. Elmakis. 1997. Quick estimation of rare events in stochastic networks. *IEEE Trans. Rel.*, 46(2):254–265.

Jelinek, F. 1997. *Statistical methods for Speech Recognition*, Massachusetts Institute of Technology Press.

Kapur, J.N. and H. K.Kesavan. 1992. *Entropy Optimization Principles with Applications*. Academic Press.

Karlin, S. and H. M. Taylor. 1975. *A First Course in Stochastic Processes*. Academic Press, New York-London, second edition.

Kiefer, J., and J. Wolfowitz. 1952. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23:462–466.

Robbins, H., and S. Monro. 1951. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Rubinstein, R. Y. 1992. Sensitivity analysis of discrete event systems by the "push-out" method. *Annals of Operations Research* 39:229–237.

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 crossentropy and importance sampling. In *Second International Workshop on Rare Event Simulation*, RESIM 99, 1–17.

Rubinstein, R. Y. and B. Melamed. 1998. *Modern Simulation and Modeling*. Wiley.

Siegmund, D. 1976. Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.*, 4(4):673–684.

Su, Y., and M. Fu. 2000. Importance sampling in derivative securities pricing. *Proceedings of the 2000 Winter Simulation Conference*, J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick, eds., 587–596. Institute of Electrical and Electronics Engineers.

Vazquez-Abad, F., and D. Dufresne. 1998. Accelerated simulation for pricing Asian options. *Proceedings of the 1998 Winter Simulation Conference*, D. J. Medeiros, E. F. Watson, J. S. Carson, and M. S. Manivannan, eds., 1493–1500. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

AUTHOR BIOGRAPHY

MING-HUA HSIEH is an assistant Professor of the Department of Management Information Systems at National Chengchi University, Taiwan. From 1997 to 1999, he was a software design engineer at Hewlett Packard company, California. He is a member of INFORMS. His research interests include simulation methodology and financial engineering. His e-mail address is `<mhsieh@mis.nccu.edu.tw>`.