Proceedings of the 1989 Winter Simulation Conference
E.A. MacNair, K.J. Musselman, P. Heidelberger (eds.)

LARGE DEVIATIONS THEORY TECHNIQUES IN MONTE CARLO SIMULATION

JOHN S. SADOWSKY
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.

JAMES A. BUCKLEW
Department of Electrical and Computer Engineering
University of Wisconsin - Madison
Madison, WI 53706, U.S.A.

ABSTRACT

This paper considers the estimation via importance sampling simulation of *large deviation probabilities*. These are probabilities P_n which vanish with an exponential rate as $n \to \infty$. Let L_n denote the number of simulations required to obtain a specified relative precision. Since P_n vanishes exponentially fast, it turns out that for most practical importance sampling simulation schemes L_n will grow exponentially as $n \to \infty$. We say that a simulation scheme is asymptotically *efficient* if L_n grows less than exponentially fast. Identification of efficient importance sampling simulation schemes is the primary goal of this paper.

There are several problems of interest which can be couched in the framework of large deviations theory. For the purpose of developing intuition and presenting of our results with minimal mathematical preliminaries we shall concentrate on the problem of i.i.d. sums crossing a threshold. In this setting we shall show that there is precisely one i.i.d. simulation scheme that is efficient. Generalization and related problems are also overviewed. We assume the reader has no previous knowledge of large deviations theory.

1. INTRODUCTION

Importance sampling is a Monte Carlo simulation technique in which simulation data is generated using a probability distribution different from the true underlying distribution. To form an unbiased Monte Carlo estimator, one must weight the observed events by an appropriate likelihood ratio. The method is called "importance sampling" because simulation distributions which minimize the estimator variance also tend to increase the relative frequency of the "important" events. The efficiency of an importance sampling simulation is usually expressed in terms of computational requirement. Let $\hat{P}_n$ denote the estimator for a probability P_n, and let L_n be the total number of simulations runs required to obtain a specified relative precision ε; that is, $\varepsilon = \text{var}(\hat{P}_n)^{1/2}/P_n$. The goal is to select a simulation distribution which tends to minimize L_n. The unconstrained optimal solution is known, and in fact, this solution yields a perfect estimator. However, the unconstrained solution is not a practical solution because it assumes knowledge of P_n. A "practical" simulation distribution is one which can be efficiently implemented with computer random number generators. Thus, the practical problem of importance sampling is to obtain the most efficient simulation distribution from a suitably large class of candidate distributions determined by implementation constraints.

In this paper we consider the importance sampling design problem from an asymptotic point of view. We adopt the framework of *large deviations theory* which considers a sequence of probabilities $\{P_n\}$ vanishing exponentially fast. As P_n vanishes, it seems quite reasonable that the computational requirement L_n will generally grow as $n \uparrow \infty$. In fact, L_n may grow exponentially fast and in this case we say the simulation scheme is inefficient. Conversely, we say that *an importance sampling simulation scheme is asymptotically efficient if L_n grows less than exponentially fast.* In general, there are many efficient simulation schemes, but most of these will not satisfy practical implementation constraints. In various settings several previous works, including Siegmund (1976), Cottrell, Fort and Malgouyres (1983), Dupuis and Kushner (1987), Parekh and Walrand (1989), and Hunkel and Bucklew (1990), have considered the family of "exponential shifts" as a *parametric family* of candidate simulation distributions. These works show that the particular "optimized" exponential shift which occurs in large deviations proofs also asymptotically minimizes estimator variance within the this parametric family of candidates. In fact, this solution is "efficient" by our definition, and all other exponential shifts are inefficient.

This paper is intended to be an introduction to the theory of large deviations in Monte Carlo simulation. We assume no previous exposure to large deviations theory. In order to develop intuition with minimal mathematical preliminaries, the bulk of this paper is devoted to the simplest large deviations problem; the problem of threshold crossings of i.i.d. sums. Our main result is that from among candidate class of all i.i.d. simulation distributions only the optimized exponential shift is efficient. This is a stronger result then obtained in the references cited above because we consider any i.i.d. simulation distribution, not just an embedded parametric family. We shall discuss generalizations to Markov chains and systems with Gaussian inputs, but we do not attempt to fully develop these.

2. IMPORTANCE SAMPLING BACKGROUND

Let $(X_1,..,X_n)$ be n random variables with joint density $f_n(x_1,..,x_n)$. We use the term "density" in a generic sense. $f_n(x_1,..,x_n)$ may be either a joint probability density function or a probability mass function. For some set $E_n \subset \mathbf{R}^n$, we wish to estimate the probability

$$P_n = \mathcal{P}((X_1,..,X_n) \in E_n). \qquad (1)$$

To do this, we generate L independent realizations of $(X_1,..,X_n)$ using the *simulation density* $f_n^*(x_1,..,x_n)$ instead of the true joint density $f_n(x_1,..,x_n)$. The importance sampling Monte Carlo estimator for P_n is

$$\hat{P}_n = \frac{1}{L} \sum_{\ell=1}^{L} 1_{E_n}(X_1^{(\ell)},...,X_n^{(\ell)})\, w_n(X_1^{(\ell)},...,X_n^{(\ell)}) \qquad (2)$$

where $(X_1^{(\ell)},...,X_n^{(\ell)})$ is the ℓ'th independent sample from the simulation density $f_n^*(\cdot)$,

$$1_{E_n}(x_1,..,x_n) = \begin{cases} 1 & \text{if } (x_1,..,x_n) \in E_n \\ 0 & \text{if } (x_1,..,x_n) \notin E_n \end{cases}, \qquad (3)$$

and

$$w_n(x_1,..,x_n) = \frac{f_n(x_1,..,x_n)}{f_n^*(x_1,..,x_n)}. \qquad (4)$$

The function $1_{E_n}(\cdot)$ is called the *indicator function* of the set E_n and the likelihood ratio $w_n(\cdot)$ is called the importance *sampling weighting function*. Notice that case $f_n^*(\cdot) = f_n(\cdot)$ is ordinary Monte Carlo estimation; in this case $w_n(\cdot) \equiv 1$ and estimator (2) reduces to a sum of indicator functions that simply counts the number of occurrences of the event $\{(X_1^{(\ell)},...,X_n^{(\ell)}) \in E_n\}$.

The purpose of the importance sampling weighting function in (2) is that it results in an unbiased estimator. Consider $E^*[\hat{P}_n]$ where $E^*[\cdot]$ denotes the $f_n^*(\cdot)$ expectation operator. The summands in (2) are independent, and hence, using (3) and (4) we have

$$\begin{aligned} E^*[\hat{P}_n] &= E^*[1_{E_n}(X_1,..,X_n)w_n(X_1,..,X_n)] \\ &= \int_{E_n} \frac{f_n(x_1,..,x_n)}{f_n^*(x_1,..,x_n)} f_n^*(x_1,..,x_n)\, dx_1..dx_n \\ &= E[1_{E_n}(X_1,..,X_n)] \\ &= P_n \end{aligned} \qquad (5)$$

which proves unbiasedness. In order for this computation to be meaningful, we require only that $f_n^*(x_1,..,x_n) > 0$ whenever $(x_1,..,x_n) \in E_n$ and $f_n(x_1,..,x_n) > 0$.

Next we consider the estimator's variance. Since $\hat{P}_n$ is unbiased, and since the summands in (2) are independent, we have

$$\begin{aligned} \mathrm{var}^*[\hat{P}_n] &= \frac{1}{L} \mathrm{var}^*[1_{E_n}(X_1,..,X_n)w_n(X_1,..,X_n)]. \\ &= \frac{1}{L}\left\{ \eta_n(f_n^*) - P_n^2 \right\} \end{aligned} \qquad (6)$$

where

$$\eta_n(f_n^*) = E^*[\,(1_{E_n}(X_1,..,X_n)w_n(X_1,..,X_n))^2\,]. \qquad (7)$$

Thus, minimizing $\mathrm{var}^*[\hat{P}_n]$ is equivalent to minimizing the functional $\eta_n(f_n^*)$. From (7) it is apparent that a good simulation density $f_n^*(\cdot)$ will tend to minimize the values of $w_n(x_1,..,x_n)$ on the set E_n. Referring back to (4), we see that this will be accomplished if $f_n^*(\cdot)$ puts a large percentage of its probability mass on E_n, concentrating its probability mass where $f_n(\cdot)$ is relatively large. This reasoning is formalized by following derivation of the unconstrained optimal simulation density (c.f. Hammersley and Handscomb (1964)). First we recall an elementary form of Jensen's inequality: $E^*[Z^2] \geq E^*[Z]^2$ with equality if and only if Z is constant with $f_n^*(\cdot)$-probability 1. Applying this to (7) we see that $\eta_n(f_n^*)$ is uniquely minimized if when $1_{E_n}(X_1,..,X_n)\, w_n(X_1,..,X_n)$ is constant with $f_n^*(\cdot)$-probability 1. Using definition (4), this is equivalent to

$$f_n^*(x_1,..,x_n) \propto 1_{E_n}(x_1,..,x_n)\, f_n(x_1,..,x_n). \quad (8)$$

This solution possesses precisely the attributes identified above; it puts all of its probability mass on E_n and in direct proportion to the values of $f_n(x_1,..,x_n)$. In addition, when (8) is used we have $\mathrm{var}^*[\hat{P}_n] = 0$! However, (8) is not a practical solution. Notice that we must normalize (8) to a probability density, but the resulting constant of proportionality is P_n^{-1}. Furthermore, one finds that the importance sampling weight is equal to P_n (with probability 1). Of course, P_n is precisely the parameter which we desire to estimate and it is inappropriate for P_n to appear explicitly in its own estimator.

The practical problem of importance sampling simulation design must take into account implementation constraints. For example, it can be difficult to generate random samples from a arbitrary joint distribution (such as (8)) on a digital computer. Computer generated random quantities must be computed as functions of i.i.d. uniform random variables. In addition, we must also be able to efficiently compute the importance sampling weight. Nonetheless, the unconstrained solution (8) has provided some valuable insight.

3. I.I.D. SUM THRESHOLD CROSSING

3.1 The i.i.d. Simulation Problem

In this section we assume that $(X_1,..,X_n)$ is an i.i.d. sequence with marginal probability density $p(x)$. The joint density has the product form

$$f_n(x_1,..,x_n) = \prod_{k=1}^{n} p(x_k). \quad (9)$$

Consider the statistic

$$T_n = \frac{1}{n} \sum_{k=1}^{n} g(X_k). \quad (10)$$

We wish to estimate the probability $P_n = \mathcal{P}(T_n \geq \gamma)$ for some fixed threshold γ. This problem can be reformulated in terms of the framework of the previous section. Simply define

$$E_n = \{ (x_1,..,x_n) : \frac{1}{n} \sum_{k=1}^{n} g(x_k) \in E \}. \quad (11)$$

Before proceeding with the large deviations, let us consider what the unconstrained optimal simulation density looks like. Applying the product form (9) to (8) we find that unconstrained optimal joint simulation density is

$$f_n^*(x_1,..,x_n) \propto 1_{E_n}(x_1,..,x_n) \prod_{k=1}^{n} p(x_k). \quad (8')$$

Notice that even though $f_n(\cdot)$ has product, the indicator $1_{E_n}(\cdot)$ does not have product form, and hence, neither does $f_n^*(\cdot)$. In (8'), $f_n^*(\cdot)$ is not even a stationary sequence distribution. The main goal of this section is to consider optimization of i.i.d. simulation densities, that is,

$$f_n^*(x_1,..,x_n) = \prod_{k=1}^{n} q(x_k). \quad (12)$$

We shall assume throughout that $q(x) > 0$ whenever $p(x) > 0$ to ensure that the computation (5) is valid.

3.2 Large Deviations for I.I.D. Sums

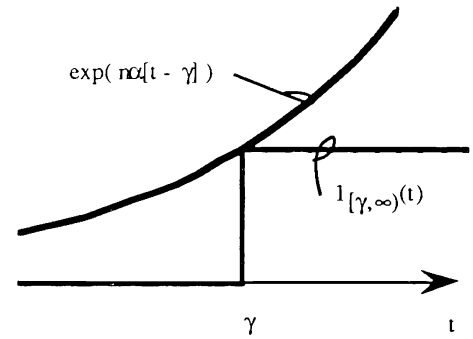

Figure 1: Illustration of the basic exponential bound.

Consider the simple exponential upper bound

$$P_n = \mathcal{P}(T_n \geq \gamma) = E[\, 1_{[\gamma,\infty)}(T_n) \,]$$

$$\leq E[\, \exp(n\alpha[T_n - \gamma]) \,] \quad (13)$$

which holds for any $\alpha \geq 0$. The bound (13) is based on the indicator function bound $1_{[\gamma,\infty)}(t) \leq \exp(n\alpha[t-\gamma])$, as illustrated in Figure 1. Recalling (11), this bound can also be expressed in terms of the variables $(x_1,..,x_n)$ by replacing t with $\sum g(x_k)$:

$$1_{E_n}(x_1,..,x_n) \leq \exp\Big(\alpha \sum_{k=1}^{n} [\, g(x_k) - \gamma \,] \Big)$$

$$= \prod_{k=1}^{n} \exp(\alpha [\, g(x_k) - \gamma \,]). \quad (14)$$

From (14) we obtain

$$P_n \leq E\Big[\prod_{k=1}^{n} \exp(\alpha [g(X_k) - \gamma]) \Big]$$

$$= \big(E[\exp(\alpha g(X) - \gamma]) \big]^n$$

$$= \exp(- n [\alpha\gamma - \Lambda(\alpha)]) \qquad (15)$$

where X is a random variable with density p(x), and

$$\Lambda(\alpha) = \log(E[\exp(\alpha g(X))]). \qquad (16)$$

We shall proceed under the following assumptions

(i) $\Lambda(\alpha) < \infty$ for all $\alpha \in (-\delta,\delta)$ for some $\delta > 0$,
(ii) $\mathrm{var}[g(X)] > 0$, and
(iii) $P_n = \mathcal{P}(T_n \geq \gamma)$ with $\gamma \geq E[g(X)]$.

$\Lambda(\alpha)$ is a cumulant function (that is, a log - moment generating function). It is known that $\Lambda(\alpha)$ is strictly convex and analytic on the interior of the set $\{\alpha: \Lambda(\alpha) < \infty\}$. We also have

$$\Lambda'(0) = E[g(X)] = E[T_n] \qquad (17)$$

and

$$\Lambda''(0) = \mathrm{var}[g(X)] = n\,\mathrm{var}[T_n]. \qquad (18)$$

See Ellis (1985).

In order to obtain the tightest upper bound in (15), we should minimimize the exponent with respect to the parameter $\alpha \geq 0$. This is equivalent to maximizing the quantity in square brackets in the exponent of (15). The results is

$$P_n \leq \exp(- I(\gamma)\, n) \qquad (19)$$

where

$$I(\gamma) = \sup_{\alpha \in \mathbf{R}} \{ \alpha\gamma - \Lambda(\alpha) \}. \qquad (20)$$

Actually, the maximization in (20) should be restricted to just $\alpha \geq 0$. However, under assumption (iii), it is easy to show (by differentiation) that the global maximization in (20) occurs either at some point $\alpha \geq 0$ or in the limit as $\alpha \to +\infty$. We could replace (iii) with

(iii') $P_n = \mathcal{P}(T_n \leq \gamma)$ with $\gamma \leq E[g(X)]$.

In this case, the upper bound (15) would be valid only for $\alpha \leq 0$, but then the maximization in (20) occurs either at some point $\alpha \leq 0$ or in the limit as $\alpha \to -\infty$. So, definition (20) handles both cases (iii) and (iii').

$I(\gamma)$ is called the *large deviations rate function*. The relationship (20) is the *Legredre-Fenchel transform* (also known as the *convex conjugate*) of $\Lambda(\alpha)$. For a thorough exposition of the role of convex function theory in large deviations analysis, the reader is directed to Chapter VI in Ellis (1985).

Now, the upper bound (19) decays exponentially as a function of n. The fact that P_n vanishes follows directly from conditions (iii) or (iii') and the law of large numbers; $T_n \to E[g(X)] = \Lambda'(0)$ in probability. So we see that large deviations theory deals with an "exponentially fast" form of convergence in probability. At this point we only have an upper bound. It may happen that P_n vanishes faster than $\exp(-I(\gamma)n)$. However, this is not the case. In essence, the main theorem of large deviations theory states (19) is exponentially tight in the sense that the exponential rate of decay of P_n is precisely $I(\gamma)$. The following theorem is a modern version of the first large deviations theorem originally published by H. Cramér in 1938.

Theorem 1: Assume $\{X_k\}$ is i.i.d. and conditions (i), (ii) and (iii) or (iii') hold. Then $I(\gamma)$ is a convex function, $I(\gamma) \geq 0$ and $I(\gamma) = 0$ if and only if $\gamma = \Lambda'(0) = E[T_n]$. Define $I = \{\gamma: I(\gamma) < \infty\}$ = an interval (by convexity). For all $\gamma \in I^o$ (= the interior of I), we have

$$\lim_{n\to\infty} \frac{1}{n} \log(P_n) = - I(\gamma), \qquad (21)$$

and furthermore, there exists a unique α_γ such that $\Lambda'(\alpha_\gamma) = \gamma$ and we have

$$I(\gamma) = \alpha_\gamma\gamma - \Lambda(\alpha_\gamma). \qquad (22)$$

If $\gamma \notin$ closure of I then $P_n \equiv 0$. This happens when g(X), and hence also T_n, it bounded. Notice that (21) still predicts the true asymptotic rate as $I(\gamma) = +\infty$ for $\gamma \notin I$. The case $\gamma \in I^o$ is the case of practical interest.

There is one more important device of i.i.d. large deviations theory which we need to discuss. This is the notion of the α-conjugate distribution. For each value of α such that $\Lambda(\alpha) < \infty$, we define the twisted probability density function

$$q^{(\alpha)}(x) = \exp(\alpha g(x) - \Lambda(\alpha))\, p(x). \qquad (23)$$

To see that this is indeed a probability density, recall that $\exp(\Lambda(\alpha))$ is the moment generating function of g(X). Thus, $\exp(-\Lambda(\alpha))$ is the appropriate normalizing constant in (23). The corresponding i.i.d. sequence distribution for $\{X_k\}$ is called the *α-conjugate distribution*. Notice that we actually have an entire

family of twisted densities parameterized by $\alpha \in \{\alpha: \Lambda(\alpha) < \infty\}$. The utility of the α-conjugate distribution is that we can write (after a some manipulation)

$$P_n = E^{(\alpha\gamma)}[\exp(-\alpha\gamma n[T_n - \gamma])\, 1_{[\gamma,\infty)}(T_n)] \times \exp(-I(\gamma)n) \quad (24)$$

where $E^{(\alpha)}[\cdot]$ is the α-conjugate expectation operation. Notice that the exponential factor on the right side of (24) is just the basic exponential upper bound (19). Thus, to prove Theorem 1 it is sufficient to show that the expectation in (24) either grows or decays less than exponentially fast. (If fact, it is known that the expectation in (24) behaves like $O(1/\sqrt{n})$.) To further develop this intuition, we note that that

$$E^{(\alpha\gamma)}[T_n] = E^{(\alpha\gamma)}[g(X)] = \Lambda'(\alpha\gamma) = \gamma.$$

To derive this last expression, first show that the α-conjugate cumulant function as a function of β is $\Lambda^{(\alpha)}(\beta) = \Lambda(\alpha+\beta) - \Lambda(\alpha)$. Recalling (17) and (18), we have

$$E^{(\alpha)}[g(X)] = \Lambda^{(\alpha)\prime}(0) = \Lambda'(\alpha) \quad (17')$$

and

$$\mathrm{var}^{(\alpha)}[g(X)] = \Lambda^{(\alpha)\prime}(0) = \Lambda'(\alpha). \quad (18')$$

Thus we see that α_γ-conjugate distribution shifts the mean of T_n to γ. This suggests that α_γ-conjugate distribution might be a good candidate for importance sampling simulation.

3.3 The Asymptotics of the i.i.d. Simulation Problem

We consider all i.i.d. simulation densities, that is, all simulation densities of the form (12) with marginal density $q(\cdot)$ with $q(x) > 0$ whenever $p(x) > 0$. We now rewrite the variance formula (6) as

$$\mathrm{var}^*[\hat{P}_n] = \frac{1}{L}\left\{ \eta_n(q) - P_n^2 \right\}. \quad (6')$$

This is identical (6), except we have now replaced $\eta_n(f_n^*)$ by

$$\eta_n(q) = E^q[(1_{E_n}(X_1,..,X_n) w_n(X_1,..,X_n))^2]$$

$$= E^q\left[\left(1_{E_n}(X_1,..,X_n)\frac{\prod_{k=1}^n p(X_k)}{\prod_{k=1}^n q(X_k)}\right)^2\right] \quad (25)$$

where $E^q[\cdot]$ is the expectation operator for the i.i.d. distribution with marginal density $q(x)$. Recall that the exponential upper bound for P_n was obtained using the product form exponential bound

$$1_{E_n}(x_1,..,x_n) \le \prod_{k=1}^n \exp(\alpha[g(x_k) - \gamma]). \quad (14)$$

Using the same technique, we now apply (14) to (23). The result is

$$\eta_n(q) \le \exp(-[\alpha\gamma - \Lambda_q(\alpha)]\, n) \quad (26)$$

where

$$\Lambda_q(\alpha) = \log\left(E^q\left[\left(\exp(\alpha g(X))\frac{p(X)}{q(X)}\right)^2\right]\right). \quad (27)$$

Minimizing (26) with respect to the parameter $\alpha \ge 0$, we obtain

$$\eta_n(q) \le \exp(-I_q(\gamma)\, n) \quad (28)$$

where

$$I_q(\gamma) = \sup_{\alpha \in \mathbf{R}} \{\alpha\gamma - \Lambda_q(\alpha)\}. \quad (29)$$

Using precisely the same method as those used to prove Theorem 1, we can prove the following large deviations theorem for the asymptotics of the importance sampling variance.

Theorem 2: Assume the conditions of (i) - (iii). Then $I_q(\gamma)$ is convex and $I_q = \{\gamma: I_q(\gamma) < \infty\}$ = an interval. For any $\gamma \in I_q^o$ such that $\Lambda_q'(0) \le \gamma$ we have

$$\lim_{n\to\infty} \frac{1}{n}\log(\eta_n(q)) = -I_q(\gamma). \quad (30)$$

If $\Lambda_q'(0) > \gamma$ then $\eta_n(q)$ does not vanish. Similar statements also hold if we replace (iii) by (iii').

Recall (6'): $\mathrm{var}^*[\hat{P}_n] = 1/L\ \{\eta_n(q) - P_n^2\}$. Since $\mathrm{var}^*[\hat{P}_n] \ge 0$, clearly we have $\eta_n(q) \ge P_n^2$. This inequality must also hold in the limit as $n \to \infty$, in particular, the exponential rate of decay of $\eta_n(q)$ cannot exceed the exponential rate of decay of P_n^2. Replacing γ by the variable t, we have

$$I_q(t) \le 2I(t). \quad (31)$$

for all $t \in \mathbf{R}$.

Definition: Assume the conditions of Theorem 1. We say that an i.i.d. simulation distribution specified by the marginal density $q(x)$ is *asymptotically efficient* for

estimating P_n if $I_q(\gamma) = 2I(\gamma)$.

Our definition of "efficiency" is based on asymptotics in the summation limit n, not on the computational requirement $L_n = L_n(q)$. (As discussed in the introduction, we define $L_n(q)$ to be the minimum number required for $\text{var}^*(\hat{P}_n)^{1/2}/P_n \le \varepsilon$.) However, we can relate the above definition of efficiency to $L_n(q)$ in a very meaningful fashion. As $n \to \infty$, since P_n vanishes exponentially fast it seems reasonable that accurate estimation of P_n should require more and more simulations. A simple manipulation indicates that $L_n(q) \sim \exp(r(q)n)$. Specifically,

$$r(q) = \lim_{n\to\infty} \frac{1}{n} \log(L_n(q)) = 2I(\gamma) - I_q(\gamma) \ge 0$$

and *$r(q) = 0$ if and only if $q(x)$ is asymptotically efficient.*

Our goal now is to identify the efficient i.i.d. simulation densities (if there are any) and then select the optimal one by considering the resulting non-exponential behavior of $L_n(q)$. From (31) $I_q(t) \le 2I(t)$ for all $t \in \mathbf{R}$ and $q(\cdot)$ is efficient for estimating P_n if $I_q(\gamma) = 2I(\gamma)$. We now apply some convex function theory dealing with the *"convex duality"* in Legredre-Fenchel relationships (20) (which defines $I(\gamma)$) and (29) (which defines $I_q(\gamma)$). This results in a characterization of efficiency in terms of the Λ-functions instead of the I-functions. By Theorem 1 there exists a unique solution of $\Lambda'(\alpha_\gamma) = \gamma$ and we have $I(\gamma) = \alpha_\gamma\gamma - \Lambda(\alpha_\gamma)$. It turns out that the inequality (31) is equivalent to $\Lambda_q(2\alpha) \ge 2\Lambda(\alpha)$ for all $\alpha \in \mathbf{R}$ and

$$I_q(\gamma) = 2I(\gamma) \quad \text{if and only if} \quad \Lambda_q(2\alpha_\gamma) = 2\Lambda(\alpha_\gamma). \quad (32)$$

A full explanation of the convex duality in (32) is beyond the scope of this presentation. We shall simply ask the reader to accept (32) on faith. The key point is that (32) gives us the desired alternative characterization of efficiency: $\Lambda_q(2\alpha_\gamma) = 2\Lambda(\alpha_\gamma)$.

Now, recall formula (25):

$$\Lambda_q(\alpha) = \log\Big(E^q\Big[\Big(\exp(\alpha g(X)) \frac{p(X)}{q(X)} \Big)^2 \Big] \Big). \quad (25)$$

To minimize $\Lambda_q(2\alpha_\gamma)$, we employ Jensen's inequality to get

$$\begin{aligned}
\Lambda_q(2\alpha_\gamma) &\ge \log\Big(E^q\Big[\exp(\alpha_\gamma g(X)) \frac{p(X)}{q(X)} \Big]^2 \Big) \\
&= 2 \log\Big(\int \exp(\alpha_\gamma g(x)) \frac{p(x)}{q(x)} q(x)\, dx \Big) \\
&= 2 \log\Big(\int \exp(\alpha_\gamma g(x))\, p(x)\, dx \Big) \\
&= 2 \log\big(E[\exp(\alpha_\gamma g(X))] \big) \\
&= 2\,\Lambda(2\alpha_\gamma)
\end{aligned}$$

where the last equality is just definition (16). Of course, we already knew that $\Lambda_q(2\alpha_\gamma) \ge 2\,\Lambda(2\alpha_\gamma)$. However, just as in the development of the unconstrained optimal simulation density (8) in Section II, Jensen's inequality gives us an if and only if condition for equality. In particular, $\Lambda_q(2\alpha_\gamma) = 2\,\Lambda(2\alpha_\gamma)$ if and only if

$$\exp(\alpha_\gamma g(X)) \frac{p(X)}{q(X)} = \text{constant}$$

with $q(\cdot)$-probability 1. Thus, *there is precisely one efficient i.i.d. simulation distribution* and it is $q(x) \propto \exp(\alpha_\gamma g(x))\, p(x)$. The constant of proportionality that normalizes this solution to a probability density is just $\exp(-\Lambda(\alpha_\gamma))$. Thus we have the following result.

Theorem 3: Assume the conditions of Theorem 1 and $\gamma \in I^o$. Then the α_γ-conjugate distribution is the unique asymptotically efficient i.i.d. simulation distribution.

Since $L_n(q)$ grows less than exponentially fast when $q(\cdot) = q^{(\alpha_\gamma)}(\cdot)$. In fact, it turns out that $L_n(q) \sim O(\sqrt{n})$. See Hunkel and Bucklew (1990).

3.4 A Numerical Example

Let $\{X_k\}$ be a sequence of i.i.d. random variables with Laplacian probability density function $p(x) = 1/2\, e^{-|x+1|}$, and take $g(x) = -1$ for $x \le -1$, $g(x) = x$ for $-1 < x < 1$, and $g(x) = +1$ for $x \ge 1$. We are interested in estimating the probability $P_n = \mathcal{P}(T_n \ge 0)$, that is, $\gamma = 0$. This calculation would arise in the computation of the error probabilities for an i.i.d. data Neyman-Pearson log-likelihood ratio test of $p(\cdot)$ vs. the alternative Laplacian density with mean +1 (instead of −1). Figure 2 illustrates $p(\cdot)$ and two simulation densities; $q(\cdot) = q^{(\alpha_\gamma)}(\cdot)$ is the asymptotically efficient simulation density and $\tilde{q}(\cdot) = 1/2\, e^{-|x|}$. Notice that the mean of

T_n for the both $q(\cdot)$ and $\tilde{q}(\cdot)$ distributions is $\gamma = 0$.

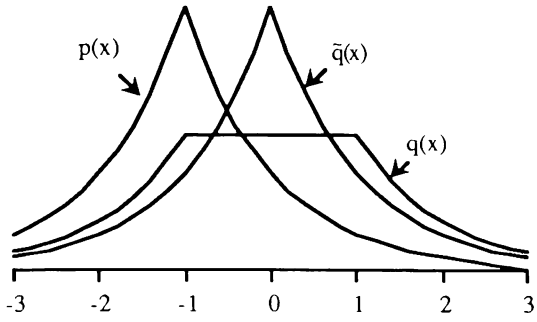

Figure 2: Comparison of simulation densities.

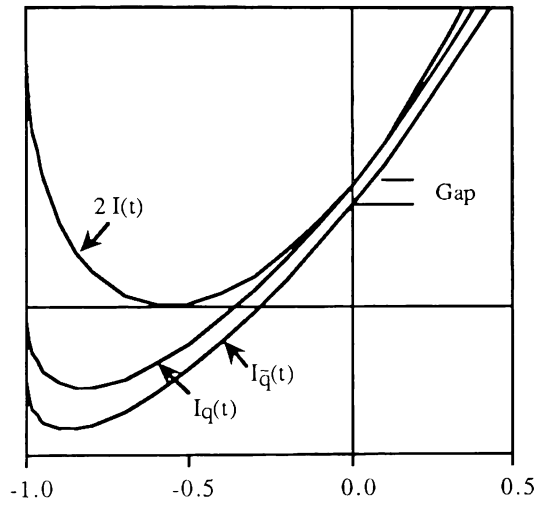

Figure 3: Comparison of variance rate functions.

Figure 3 compares the variance rate functions, $I_q(t)$ and $I_{\tilde{q}}(t)$, along with the upper bound $2I(t)$. Notice that $I_q(0) = 2I(0)$, as predicted by Theorem 3, but $I_{\tilde{q}}(0) < 2I(0)$ which is indicated as a "performance gap." Finally, Table 1 lists estimates of $L_n(q)$, $L_n(\tilde{q})$ and $L_n(p)$ (the ordinary Monte Carlo computational requirement) for a 10% relative precision. These were obtained from sample variance estimates. Notice that the values of n in Table 1 are not particularly large. In fact, the asymptotically efficient simulation density clearly establishes it's superior efficiency for just n = 2!

Table 1: Comparison of computational efficiency.

n	$\hat{P}_n$	$L_n(q)$	$L_n(\tilde{q})$	$L_n(p)$
1	.189	121	95	529
2	.166	141	246	602
4	.058	203	411	1.7×10^3
6	.027	249	533	3.7×10^3
8	.013	265	751	7.7×10^3
10	.0066	302	1074	1.5×10^4
12	.0031	326	1217	3.2×10^4
14	.0016	348	1629	6.3×10^4
16	.00086	369	2012	1.2×10^5

4. GENERALIZATIONS

The i.i.d. sum threshold crossing problem can be generalized in a number of ways. The sequence $\{X_k\}$ may be a Markov chain, and we may redefine the statistic T_n as

$$T_n = \frac{1}{n} \sum_{k=1}^{n} g(X_{k-1}, X_k). \qquad (33)$$

Assume a finite state space and let P(i,j) denote the transition probability for the Markov chain $\{X_k\}$. We also assume that P(i,j) is irreducible This stochastic matrix is transformed into a non-negative matrix $K(i,j;\alpha) = \exp(\alpha g(i,j))\ P(i,j)$. In this setting

$$\Lambda(\alpha) = \log(\text{ spectral radius of } K(\alpha)\ ).$$

By the Perron-Frobenius Theorem, the spectral radius above is a unique positive eigenvalue. Let $r(i;\alpha)$ denote the associated right eigenvector of $K(\alpha)$. Then the α-conjugate distribution is the Markov chain distribution generated by the *twisted transition probability*

$$Q^{(\alpha)}(i,j) = \frac{r(j;\alpha)}{r(i;\alpha)} \exp(-\Lambda(\alpha))\ K(i,j;\alpha)$$

$$= \frac{r(j;\alpha)}{r(i;\alpha)} \exp(\alpha g(i,j) - \Lambda(\alpha))\ P(i,j). \qquad (34)$$

Theorems 1, 2 and 3 can be generalized to this setting.

In particular, note that if $\{X_k\}$ is i.i.d., expanding the class of candidate simulation distributions to the Markov chains (instead of just the i.i.d. distributions) does not produce any more efficient simulation distributions.

More generally, $g(X_{k-1},X_k)$ in (32) can be replaced by a random variable Z_k such that given the history of the Markov chain $\{X_k = x_k\}$, the random variables $\{Z_k\}$ are conditionally independent and for each k the conditional distribution of Z_k depends only on x_k and x_{k-1}. This is what is known as a *Markov-Additive Process*.

For a rigorous treatment of the case of Markov chains on an abstract state space, the reader is referred to Bucklew, Ney and Sadowsky (1990).

Another line of generalization is to consider multidimensional statistics. In this case, T_n is a random vector in $\mathbf{R}^d$ and we consider the probabilities $P_n = \mathcal{P}(T_n \in E)$ for some set $E \subset \mathbf{R}^d$. In Section 3, where $P_n = \mathcal{P}(T_n \geq \gamma)$, we have $E = [\gamma,\infty)$. The multi-dimensional version of Theorem 1 replaces $I(\gamma)$ by

$$I(E) = \inf_{t \in E} I(t).$$

In this multidimensional setting there is still a convex duality between $\Lambda(\alpha)$ and $I(t)$: $I(t) = \alpha_t \cdot t - \Lambda(\alpha_t)$ where α_t is the unique solution of $\nabla\Lambda(\alpha_t) = t$ and the $\cdot$ denotes the Euclidean dot product on $\mathbf{R}^d$. The rate function $I(t)$ is globally minimized at $t = \nabla\Lambda(0)$. We define a *minimum rate point* to be a point on the γ on the boundary of E such that $I(E) = I(\gamma)$. A key strategy in multidimensional large deviations theory is to reduce the multidimensional case to a one dimensional problem using half spaces. Suppose that γ is called a minimum rate point of E, and in addition, E is covered by the half space $\mathcal{H}(\gamma) = \{t: \alpha_\gamma(t-\gamma) \geq 0\}$. Then γ is a *dominating point*. (See Ney (1983).) A dominating point is illustrated in Figure 5. Notice that $\mathcal{H}(\gamma)$ (the slashed half plane) is tangent to the rate function "level set" $\{t: I(t) = I(E)\}$.

Theorem 3 (and its Markov generalizations) can be immediately extended to the case where E has a dominating point. The exponential twisting formulas (23) (or (34)) provide the unique i.i.d. (or Markov) efficient simulation distributions.

Not all sets have dominating points. For example, Figure 6 illustrates the case of a unique minimum rate point that is not a dominating point. There may also be more than one minimum rate point. In these situations, one can construct efficient schemes as convex combinations of α-conjugate distributions. A set $\Gamma = \{\gamma_1,...,\gamma_m\}$ and a strictly positive probability vector $\mathbf{p} = (p_1,...,p_m)$ defines such a convex combination. A sample is generated as follows: First, select a point γ_i from Γ using the distribution $\mathbf{p}$. Then sample $(X_1^{(\ell)},...,X_n^{(\ell)})$ from the α_{γ_i}-conjugate distribution. (Notice that this is a convex combination of the joint distribution, not of the marginal twisted densities (23) or (34).) In Sadowsky and Bucklew (1990), it is demonstrated that a necessary condition for efficiency is that Γ contain all minimum rate points, and a sufficient condition is that Γ define a covering of E by half spaces. The reader is directed to the reference for more detail.

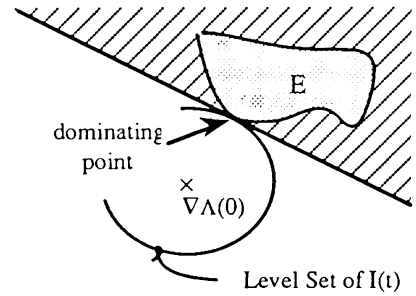

Figure 5: Illustration of a dominating point.

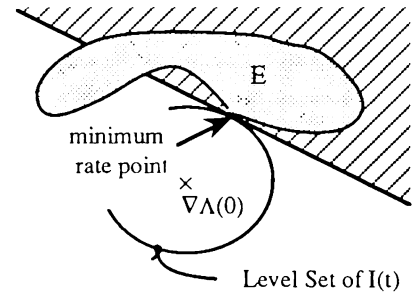

Figure 6: Illustration of a minimum rate point.

Finally, we consider one more setting of practical interest; the setting of systems with small Gaussian noise inputs. Consider $P_n = \mathcal{P}(T_n \in E)$ where T_n is a d-dimensional Gaussian random vector with mean vector μ and covariance matrix 1/n C. Actually, this situation can be addressed in terms of i.i.d. formulation. T_n is statistically equivalent to $1/n \sum_{k=1}^{n} X_k$ where $\{X_k\}$ is a sequence of i.i.d. Gaussian random vectors having mean vector μ and covariance matrix C. For a point γ the α_γ-conjugate distribution of T_n turns out to be the Gaussian distribution with mean γ and covariance matrix 1/n C. Thus, the asymptotically efficient way to simulate systems with Gaussian inputs is first find the minimum

rate points, and then use convex combinations of mean translations to these points. Further discussion and some examples are found in Sadowsky and Bucklew (1990).

REFERENCES

Bucklew, J. A., Ney, P. and Sadowsky, J. S. (1990), Monte Carlo simulation and large deviations theory for uniformly recurrent Markov chains. To appear in the *Journal of Applied Probability*.

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

Dupuis, P. and Kushner, J. (1987). Stochastic systems with small noise, analysis and simulation; A phase licked loop example. *SIAM Journal of Appplied Mathematics*, **47**, 643-661.

Ellis, R. S. (1984). Large deviations for a general class of random vectors. *Annals of Probability*, **12**, 1-12, 1984.

Ellis, R. S. (1985). *Entropy, Large Deviations and Statistical Mechanics*. Springer-Verlag, New York.

Hammersley, J. M. and Handscomb, D. C. (1964). *Monte Carlo Methods*. Chapman and Hall, New York.

Hunkel, V. M. F. and Bucklew, J. A. (1990). Fast simulation for functionals of Markov chains. To appear in *IEEE Transactions on Information Theory*.

Ney, P. (1983). Dominating points and the asymptotics of large deviations for random walks on R^d. *Annals of Probabability*, **11**, 158-167.

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

Sadowsky, J. S. and Bucklew, J. A. (1990). On large deviations theory and asymptotically efficient Monte Carlo estimation. To appear in *IEEE Transactions on Information Theory*.

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

AUTHORS' BIOGRAPHIES

JOHN S. SADOWSKY was born in Hammond, IN, in 1956. He received the BSEE and BSMa degrees from Rose-Hulman Institute of Technology in 1978, the MSEE degree from Iowa State University in 1981, and the MAMa degree in 1983 and the Ph.D in Electrical Engineering in 1984 from the University of Wisconsin - Madison. He is a member of the IEEE and Eta Kappa Nu. From 1978 - 1981 he was employed at Rockwell-Collins Avionics in Cedar Rapids, Iowa. He currently holds a faculty position with the School of Electrical Engineering, Purdue University - West Lafayette.

John S. Sadowsky
School of Electrical Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.
(317) 494-3483

JAMES A. BUCKLEW was born in Big Spring, TX, in 1954. He received the Ph.D. degree in Electrical Engineering from Purdue University, West Lafayette, in 1979. He has held visiting appointments with the Electrical Engineering departments of Purdue and the University of Texas at Austin, and the Department of Statistics (Center for Stochastic Processes) at the University of North Carolina, Chapel Hill. He is currently a Professor with the Departments of Electrical Engineering and Mathematics at the University of Wisconsin - Madison.

Dr. Bucklew has received the Presidential Young Investigator Award from the National Science Foundation in 1984. He is a member of the IEEE, Eta Kappa Nu, Tau Beta Pi, the Institute of Mathematical Statistics, and the Mathematical Association of America. Currently he is an Associate Editor for the *IEEE Transactions on Information Theory*.

James A. Bucklew
Electrical and Computer Engineering
University of Wisconsin - Madison
Madison, WI 53706, U.S.A.
(608) 238 - 5523