

Bias Point Selection in the Importance Sampling Monte Carlo Simulation of Systems

EDICS: 2-DETC, 2-ESTM, 2-PERF, 2-SYSM

J.A. Bucklew and J.A. Gubner

Abstract— We consider the issue of whether it is better to bias the random variables at the input, at the output, or at some intermediate point of a system. We show that in a very general setting, the closer to the output that we can bias our system simulation variables, the better off we will be. We do show that surprisingly in some important special cases, the performance can be equal no matter where the bias point is selected.

In the second part of the paper we present a very general large deviation type theorem on the variance rates of importance sampling estimators. We then use this theorem to consider in a quantitative fashion what the difference in the variance rates can be for input versus output formulations. We present several examples illustrating the developed theory.

Keywords— Importance Sampling, Simulation, Monte Carlo.

I. INTRODUCTION

LARGE and/or nonlinear stochastic systems, due to analytic intractability, must often be simulated in order to obtain estimates of the key performance parameters. Typical situations of interest could be a buffer overload in a queuing network or an error event in a digital communication system. In many system designs or analyses an event of rare probability is a key parameter of the system's efficacy. To estimate such a parameter by a brute force direct simulation would require that a very large number of independent random numbers be generated from the computer's random number generator.

One way out of this quandary is to utilize a technique called *importance sampling*. Importance sampling has in the last few years established itself as the main method of variance reduction for the simulation of rare events. For an excellent review article over this methodology in the field of network simulation, see [7]. Another highly recommended review article in the field of communications systems is [12].

The main idea of the methodology is simple to present. Suppose we wish to estimate $\rho = E\{\phi(Z)\}$ where Z is a random variable describing some observation on a random system. Usually ϕ is the indicator function of some set implying that ρ is the probability of the set. Suppose that the observation random variable Z has probability density

function $p(\cdot)$. The direct (Monte Carlo) simulation method would be to generate a sequence of i.i.d. random numbers $Z^{(1)}, Z^{(2)}, \dots, Z^{(k)}$ from the density $p(\cdot)$ and form the estimate

$$\hat{\rho}_p \doteq \frac{1}{k} \sum_{i=1}^k \phi(Z^{(i)}).$$

By the law of large numbers $\hat{\rho}_p \rightarrow \rho$ as $k \rightarrow \infty$. Thus as the number of observations approaches infinity, we converge to the true value. Suppose instead, we generate a sequence of i.i.d. random numbers $Z^{(1)'}, Z^{(2)'}, \dots, Z^{(k)'}$ with a possibly different density $q(\cdot)$. We call these random variables the “biased” random variables and $q(\cdot)$, the “biased” distribution. We then form the estimate

$$\hat{\rho}_q \doteq \frac{1}{k} \sum_{i=1}^k \frac{p(Z^{(i)'})}{q(Z^{(i)'})} \phi(Z^{(i)'}).$$

The ratio $p(\cdot)/q(\cdot)$ will be called the *weight function* of the importance sampling estimator. (This ratio can be thought of as the Radon-Nikodym derivative of the probability measure associated with $p(\cdot)$ with respect to the probability measure associated with $q(\cdot)$.) It is simple to verify that the expected value of $\hat{\rho}_q$ under the density $q(\cdot)$ is precisely ρ . Therefore, the estimate $\hat{\rho}_q$ is unbiased and as $k \rightarrow \infty$, we also expect it to converge (by the law of large numbers) to its mean value ρ . The obvious question is, “Are there better choices for $q(\cdot)$ than just $p(\cdot)$?” The answer is that by making a good choice for $q(\cdot)$, orders of magnitude decrease in the estimator variance can be achieved over a direct Monte Carlo simulation. It is this fact that has spurred most if not all the recent interest in importance sampling techniques.

In this paper we consider a very general question in the field of system simulation: “Should we bias the random variables at the input, at the output, or at some intermediate point of a system?”. To be a bit more specific, consider the following example:

Example 1: Suppose we are interested in estimating

$$\rho = P(S + N > a),$$

where S and N are two independent random variables with densities $p_s(\cdot)$ and $p_n(\cdot)$ respectively. Denote the sum of these two random variables by R , with density denoted by $p_r(\cdot)$. We consider two types of estimators, an input

estimator and an output estimator. The “input” estimator is explicitly given as

$$\hat{\rho}_i = \frac{1}{k} \sum_{j=1}^k 1_{\{S^{(j)'} + N^{(j)'} > a\}} \frac{p_s(S^{(j)'})p_n(N^{(j)'})}{p_{s'}(S^{(j)'})p_{n'}(N^{(j)'})}$$

and the “output” estimator as

$$\hat{\rho}_o = \frac{1}{k} \sum_{j=1}^k 1_{\{R^{(j)} > a\}} \frac{p_r(R^{(j)})}{p_{r'}(R^{(j)})},$$

where we assume that the simulation random variables satisfy $R^{(j)'} = S^{(j)'} + N^{(j)'}$, and the original random variables similarly satisfy $R^{(j)} = S^{(j)} + N^{(j)}$. Both of these estimators are unbiased. Which has lower variance?

We should note that in many situations, an output formulation of the bias distribution is impossible. If the system is very complicated, it may very well be virtually impossible to calculate the biasing distributions at the output of the system. However, it may very well be possible to calculate the distributions at some intermediate point of the system. In this paper we attempt to show how much there is to gain or lose by using an input over an output formulation. It is essential to the theory of importance sampling in system simulation that we gain understanding of the role of the bias point in a Monte Carlo simulation. We note that many researchers have mentioned or considered the pro and cons of input versus output analysis [6],[10],[13] to name just a few. Indeed our terminology for this concept comes from [6]. For further examples, we recommend the encyclopedic text of simulation methodologies for communication systems [8].

II. TECHNICAL SETUP

We are given two (Borel measurable) functions $g : \mathcal{R}^N \rightarrow \mathcal{R}^M$ and $h : \mathcal{R}^M \rightarrow \mathcal{R}^L$, which define our system as shown in Fig. 1. Let (X_1, X_2, \dots, X_N) be an arbitrary random vector. We consider these to be our “input random variables”. We denote their joint probability measure as P_x . Define $(Y_1, Y_2, \dots, Y_M) = g(X_1, X_2, \dots, X_N)$ which we consider to be our “intermediate random variables” with joint probability measure P_y and lastly $Z_1, Z_2, \dots, Z_L = h(Y_1, Y_2, \dots, Y_M)$ our “output random variables” with joint measure P_z .

Let f be a (Borel measurable) function mapping \mathcal{R}^L to \mathcal{R}^d . Suppose we are interested in the quantity

$$\begin{aligned} \rho &= E(f(Z_1, \dots, Z_L)), \\ &= E(f(h(Y_1, Y_2, \dots, Y_M))), \\ &= E(f(h(g(X_1, X_2, \dots, X_N)))). \end{aligned}$$

$\rho = (\rho_1, \rho_2, \dots, \rho_d)$ is of course a d -dimensional vector. The bias probability measures will always be denoted with the symbol Q with a subscript to indicate which random variables are being biased, e.g. Q_x, Q_y, Q_z . We will assume that the original probability measures are absolutely

continuous with respect to these measures. It is enough to assume that $P_x \ll Q_x$ since this automatically implies $P_y \ll Q_y$ and $P_z \ll Q_z$. This of course guarantees the existence of the Radon-Nikodym derivatives $dP_x/dQ_x, dP_y/dQ_y, dP_z/dQ_z$ needed for our importance sampling estimators. We assume that biasing measures have the same relationship between them as do the actual measures. (We will denote the biased random variables as the original random variable written with a tilde over it.) Thus, if $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_N$ are generated to have measure Q_x , then $g(\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_N)$ will have measure Q_y and $h(g(\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_N))$ will have measure Q_z .

Depending on at which point of the system we wish to bias, we can define various estimators of ρ . The possibilities are

$$\hat{\rho}_i = \frac{1}{k} \sum_{j=1}^k f(h(g(\tilde{X}_1^{(j)}, \tilde{X}_2^{(j)}, \dots, \tilde{X}_N^{(j)}))) \frac{dP_x}{dQ_x}(\tilde{X}_1^{(j)}, \dots, \tilde{X}_N^{(j)})$$

$$\hat{\rho}_m = \frac{1}{k} \sum_{j=1}^k f(h(\tilde{Y}_1^{(j)}, \tilde{Y}_2^{(j)}, \dots, \tilde{Y}_M^{(j)}))) \frac{dP_y}{dQ_y}(\tilde{Y}_1^{(j)}, \dots, \tilde{Y}_M^{(j)})$$

and

$$\hat{\rho}_o = \frac{1}{k} \sum_{j=1}^k f(\tilde{Z}_1^{(j)}, \dots, \tilde{Z}_L^{(j)}) \frac{dP_z}{dQ_z}(\tilde{Z}_1^{(j)}, \dots, \tilde{Z}_L^{(j)})$$

as the input, intermediate, and output estimators respectively, and where the superscript on a random variable indicates which one of k independent simulation runs is under consideration. Each of these estimates are d -dimensional vectors; $\hat{\rho}_i = (\hat{\rho}_{i,1}, \hat{\rho}_{i,2}, \dots, \hat{\rho}_{i,d})$, $\hat{\rho}_m = (\hat{\rho}_{m,1}, \hat{\rho}_{m,2}, \dots, \hat{\rho}_{m,d})$, and $\hat{\rho}_o = (\hat{\rho}_{o,1}, \hat{\rho}_{o,2}, \dots, \hat{\rho}_{o,d})$.

We now state the following fundamental theorem of importance sampling Monte Carlo system simulation:

Theorem 1:

$$\text{Var}(\hat{\rho}_{i,r}) \geq \text{Var}(\hat{\rho}_{m,r}) \geq \text{Var}(\hat{\rho}_{o,r}) \quad r = 1, 2, \dots, d$$

with equality for the first inequality if and only if

$$\frac{dP_x}{dQ_x}(\tilde{X}_1^{(j)}, \dots, \tilde{X}_N^{(j)}) = s_i(\tilde{Y}_1^{(j)}, \dots, \tilde{Y}_M^{(j)})$$

for some function s_i , and with equality for the second inequality if and only if

$$\frac{dP_y}{dQ_y}(\tilde{Y}_1^{(j)}, \dots, \tilde{Y}_M^{(j)}) = s_o(\tilde{Z}_1^{(j)}, \dots, \tilde{Z}_L^{(j)})$$

for some function s_o .

We give the proof of the theorem in the appendix.

Remark 1: It is possible that the inequalities be met with equality. For example, consider the case of $h(g(x_1, \dots, x_N)) = \sum_{i=1}^N r(x_i)$, for some arbitrary function $r : \mathcal{R} \rightarrow \mathcal{R}$. We can suppose that the output estimator and the intermediate estimator are the same. Suppose

$dP_x(x_1, x_2, \dots, x_N) = \prod_{i=1}^N p(x_i)$, where $p(\cdot)$ is the input probability density (or mass function if we are dealing with discrete random variables). Suppose we choose the biasing distributions to be *exponential shifts*:

$$dQ_{x,\theta}(x_1, \dots, x_N) = \prod_{i=1}^N q_\theta(x_i) = \frac{\prod_{i=1}^N p(x_i) \exp(\theta r(x_i))}{M(\theta)^N}$$

where $M(\theta) = \int p(x) \exp(\theta r(x)) dx$ is the moment generating function of the scalar random variable $r(X)$. Now note that

$$\begin{aligned} \frac{dP_x}{dQ_x}(\tilde{X}_1, \dots, \tilde{X}_N) &= \frac{\prod_{i=1}^N p(\tilde{X}_i)}{\prod_{i=1}^N p(\tilde{X}_i) \exp(\theta r(\tilde{X}_i))} M(\theta)^{-N} \\ &= \exp(-\theta \sum_{i=1}^N r(\tilde{X}_i)) M(\theta)^N \\ &= \exp(-\theta \tilde{Y}) M(\theta)^N \\ &= s_i(\tilde{Y}). \end{aligned}$$

Thus, in this sum of i.i.d. random variables setting with exponential shift bias distributions, no performance loss is incurred by using the simpler input formulation.

III. THE VARIANCE RATE

In this section we first prove a very general theorem regarding the variance rate of importance sampling estimators. We then apply this theorem to the particular problem of the variance rates of input and output estimators.

For every integer n , let $Z_{p,n}$ be a random variable taking values in some complete separable metric space \mathcal{S}_n . Let P_n be the probability measure induced by $Z_{p,n}$ on \mathcal{S}_n . Instead of simulating $Z_{p,n}$, we choose to simulate with another \mathcal{S}_n valued random variable $Z_{q,n}$ which in turn induces a probability measure on \mathcal{S}_n , Q_n . Let f_n be an \mathcal{R}^d valued measurable function on the space \mathcal{S}_n , i.e., $f_n : \mathcal{S}_n \rightarrow \mathcal{R}^d$.

To create the importance sampling estimators, we must assume that P_n is absolutely continuous with respect to Q_n for all n (and hence the Radon-Nikodym derivative dP_n/dQ_n exists). We suppose we are interested in $\rho_n = P(\frac{f_n(Z_{p,n})}{n} \in E)$, for some Borel set $E \subset \mathcal{R}^d$. The importance sampling estimator of this probability is given as

$$\hat{\rho}_n = \frac{1}{k} \sum_{j=1}^k \frac{dP_n}{dQ_n}(Z_{q,n}^{(j)}) \mathbf{1}_{\{f_n(Z_{q,n}^{(j)}) \in E\}}.$$

We propose to study the variance of $\hat{\rho}_n$ as a function of n .

We first need to make a few definitions. For $\theta \in \mathcal{R}^d$, define

$$c_n(\theta) = \frac{1}{n} \log \left(\int \frac{dP_n}{dQ_n}(z) \exp(\langle \theta, f_n(z) \rangle) dP_n(z) \right),$$

which is a convex function in θ .

Assumption A1. $c(\theta) = \lim_n c_n(\theta)$ exists for all $\theta \in \mathcal{R}^d$, where we allow ∞ both as a limit value and as an element of the sequence $\{c_n(\theta)\}$.

In other words, the limit exists and is defined to be $c(\theta)$ for all $\theta \in \mathcal{R}^d$. We define the *domain* of c , as $D_c = \{\theta \in \mathcal{R}^d : c(\theta) < \infty\}$. Note that D_c is convex and thus c itself is convex since it is the limit of convex functions on a convex set.

Assumption A2. The origin belongs to the interior of the domain of c , i.e. $0 \in \overset{\circ}{D}_c$.

Before, we define our third assumption, we need to make a couple of additional definitions. A function $c : \mathcal{R}^d \rightarrow \mathcal{R}$ differentiable on the interior of its domain $\overset{\circ}{D}_c$ is *steep* if $\{x_n\} \subset \overset{\circ}{D}_c$ and $x_n \rightarrow x_0 \in \partial D_c$ (the boundary of the domain), implies that $\|\nabla c(x_n)\| \rightarrow \infty$

A convex function c is *essentially smooth* if three conditions hold, a) the set $\overset{\circ}{D}_c$ is nonempty, b) c is differentiable everywhere in $\overset{\circ}{D}_c$, and c) c is steep.

Assumption A3. c is essentially smooth.

We now define the Legendre-Fenchel transform of $c(\cdot)$, $R(\cdot)$, as

$$R(x) = \sup_{\theta \in \mathcal{R}^d} [\langle \theta, x \rangle - c(\theta)] \quad (\text{for } x \in \mathcal{R}^d).$$

We will call R , the *variance rate function*. We also define the *Cramér transform* for a Borel set E as

$$R(E) = \inf_{x \in E} R(x).$$

As stated above, we are interested in determining the rate of the variance expression as a function of n . Recall the variance expression is $k \text{Var} \hat{\rho}_n = F_n - \rho_n^2$. The first term (which usually determines the rate of the variance) of that expression may be written as

$$F_n = \int \left(\frac{dP_n}{dQ_n}(z) \right)^2 \mathbf{1}_{\{f_n(z) \in E\}} dQ_n(z).$$

We now have the following theorem,

Theorem 2: Assume A1, A2, and A3. Let E be any Borel set such that $E^o \neq \emptyset$, $\bar{E} = \bar{E}^o$ and $0 < R(E) < \infty$. Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log(F_n) = -R(E).$$

To make the paper self-contained, we give the proof of this theorem in the appendix. To our knowledge, the theorem is new even though the techniques of the proof are straightforward applications of now standard techniques.

Remark 2: It is the purpose of large deviation theory to be able to compute the exponential rate with which ρ_n tends to zero. From Theorem 3 of the appendix we have that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \rho_n = -I(E),$$

where $I(\cdot)$ is a set-valued function, called the *large deviation rate function*. An explicit representation for it is given

in the theorem statement. We note that since $F_n \geq \rho_n^2$ for all n , it must be true that $R(E) \leq 2I(T)$. If an importance sampling estimator can be found such that this inequality is met with equality ($R(E) = 2I(E)$), we say that this estimator is *efficient*. An efficient estimator has the fastest possible rate to zero of the square term in its variance expression. In general there are many efficient estimators for a given problem.

Example 2: Suppose the inputs to the system are i.i.d. random variables, $\{X_i\}$ which have the scalar density function $p(\cdot)$. We are interested in $\rho = P(\sum_{i=1}^n X_i > na)$. We simulate with i.i.d. $\{S_i\}$ whose individual density functions are $q(\cdot)$. In the light of our theorem let us compute the variance rates for the input and output estimators respectively.

For the output estimator we have $\mathcal{S}_n = \mathcal{R} \quad \forall n$. Also $d = 1$ and $f_n(x) = x \quad \forall n$. $dP_n/dQ_n = p * p * \dots * p / q * q * \dots * q$, where $*$ denotes the convolution operator and there are n convolutions each in the numerator and denominator respectively. For the output estimator we define

$$c_{n,o}(\theta) = \frac{1}{n} \log \left(\int \frac{(p * \dots * p)(z)^2}{(q * \dots * q)(z)} \exp(\theta z) dz \right)$$

and $c_o(\theta) = \lim_{n \rightarrow \infty} c_{n,o}(\theta)$.

For the input estimator we have $\mathcal{S}_n = \mathcal{R}^n \quad \forall n$. Also $d = 1$ and $f_n(x_1, x_2, \dots, x_n) = \sum_{i=1}^n x_i \quad \forall n$. $(dP_n/dQ_n)(x_1, x_2, \dots, x_n) = \prod_{i=1}^n (p(x_i)/q(x_i))$. For the input estimator we define

$$\begin{aligned} c_{n,i}(\theta) &= \frac{1}{n} \log \left(\int \prod_{i=1}^n \frac{p(x_i)^2}{q(x_i)} \exp(\theta x_i) dx_1 dx_2 \dots dx_n \right) \\ &= \log \left(\int \frac{p^2(x)}{q(x)} \exp(\theta x) dx \right) \end{aligned}$$

and hence $c_i(\theta) = c_{1,i}(\theta)$.

The variance rate of the input estimator is already known [3] and matches up with our calculation,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log(\text{Var}(\hat{\rho}_i)) = -\sup_{\theta} [\theta x - c_i(\theta)],$$

Example 3: For some $1/2 < a < 1$, we wish to estimate via simulation the probability

$$P\left(\sum_{k=1}^n B_k \geq na\right)$$

where the $\{B_k\}$ are i.i.d. symmetric Bernoulli random variables. An output simulation would have us simulate the binomial random variable with some other distribution which we choose to be Poisson with mean parameter na . Thus,

$$p_n(k) = b(k; n, 1/2)$$

and we simulate with

$$q_n(k) = p(k; na).$$

In the following derivation we make use of the following equality for approximating the binomial distributions with

a Poisson, [5, p. 172, Eq.10.3]

$$p(k; \lambda) e^{k\lambda/n} > b(k; n, p) > p(k; \lambda) e^{-k^2/(n-k) - \lambda^2/(n-\lambda)},$$

where $\lambda = np$. Thus we can show that

$$\exp(nc_{n,o}(\theta)) = \exp(na) 4^{-n} \sum_{k=0}^n \frac{\exp(\theta k)}{(na)^k} \frac{(n!)^2}{(k!)[(n-k)!]^2}$$

has the following asymptotic logarithmic limit behavior:

$$\lim_{n \rightarrow \infty} c_{n,o}(\theta) = c_o(\theta) = a - \log(4) + \log\left(1 + \frac{\exp(\theta)}{a}\right).$$

The corresponding input bias scheme for this problem is to simulate the symmetric Bernoulli random variables with Poisson random variables of mean a . For this scheme, $c_i(\theta) = c_1(\theta)$ and thus

$$\begin{aligned} c_i(\theta) &= \log \left(\sum_{k=0}^1 \frac{p_1(k)^2}{q_1(k)} \right) \\ &= a - \log(4) + \log\left(1 + \frac{\exp(\theta)}{a}\right). \end{aligned}$$

In other words the $c_i(\theta) = c_o(\theta)$, and thus the variance rates are equal!

So far, we have seen for the case of bias distributions of the form of exponential shifts and in the above example that the variance rate for the two schemes has been equal. This is not always the case, as the following example shows.

Example 4: Suppose we wish to estimate via simulation the probability

$$P\left(\sum_{k=1}^n X_k \geq n\right)$$

where the $\{X_k\}$ are i.i.d. exponential random variables with parameter $\lambda > 1$ (thus $E[X_1] = 1/\lambda$). An output simulation would have us simulate the sum of exponential random variables (which has an Erlang distribution) with some other distribution. In this example we choose the bias distribution to be that of a sum of standard Gaussian squared random variables (which has a χ^2 distribution). In other words, we select

$$p_n(x) = \frac{\lambda^n x^{n-1} \exp(-\lambda x)}{\Gamma(n)}$$

and we simulate with

$$q_n(x) = \frac{x^{\frac{n}{2}-1} \exp(-x/2)}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})}.$$

Therefore the variance rate is dependent on

$$\begin{aligned} \exp(nc_{n,o}(\theta)) &= \frac{\lambda^{2n} 2^{\frac{n}{2}} \Gamma(n/2)}{\Gamma(n)^2} \int_0^\infty x^{2n - \frac{n}{2} - 1} \exp\left(-\left(2\lambda - \theta - \frac{1}{2}\right)x\right) dx \\ &= \frac{\lambda^{2n} 2^{\frac{n}{2}} \Gamma(\frac{n}{2})}{\Gamma(n)^2} \frac{\Gamma(\frac{3n}{2})}{\left(2\lambda - \theta - \frac{1}{2}\right)^{\frac{3n}{2}}}. \end{aligned}$$

Note that by Stirling's formula, we have

$$\Gamma(z) \approx \exp(-z) z^{z-\frac{1}{2}} \sqrt{2\pi}.$$

So, $\frac{\Gamma(\frac{n}{2})\Gamma(\frac{3n}{2})}{\Gamma(n)^2} \approx (\frac{1}{3})^{\frac{n}{2}} (\frac{3}{2})^{2n} \frac{2}{\sqrt{3}}$ and therefore, as $n \rightarrow \infty$

$$\exp(nc_{n,o}(\theta)) \approx \frac{(\lambda^{2n})(2^{\frac{n}{2}})}{(2\lambda - \theta - \frac{1}{2})^{\frac{3n}{2}} (\frac{2}{\sqrt{3}})} (\frac{1}{3})^{\frac{n}{2}} (\frac{3}{2})^{2n}.$$

Hence,

$$\lim_{n \rightarrow \infty} c_{n,o}(\theta) = c_o(\theta) = \log\left(\frac{\lambda^2(\sqrt{2})}{(2\lambda - \theta - \frac{1}{2})^{\frac{3}{2}} \frac{3}{2}}\right).$$

For the input bias scheme we have

$$c_1(\theta) = c_i(\theta) = \log\left(\frac{\lambda^2(\sqrt{2})}{(2\lambda - \theta - \frac{1}{2})^{\frac{3}{2}} \frac{\pi}{2}}\right).$$

Hence the rates are indeed slightly different. Remember that any difference in the rate will eventually translate into huge differences in the actual estimator variance as n grows large.

Example 5: A common non-coherent communication receiver is the energy detector. In the simplest binary setting, we must choose between two hypotheses; that of $H_1 = \{\text{signal one present}\}$ and $H_0 = \{\text{signal zero present}\}$. We suppose that signal one has more power than signal zero. When signal zero is being transmitted (i.e. H_0 is true), we suppose that during the signaling time period of n samples, we receive the m -periodic signal sequence $\{s_i\}$ in the presence of an additive white Gaussian noise $\{N_i\}$ (mean zero, variance one). Denote the average power of the sequence as $P_s = (\sum_{i=1}^m s_i^2)/m$ and the d.c. value as $m_s = (\sum_{i=1}^m s_i)/m$. Also, we assume, for simplicity, that n is a multiple of m . Denote the received sequence as $\{R_i\}$. Of course the purpose of the communications receiver is to determine which signal is present. An energy detector would compute the following test statistic,

$$\sum_{i=1}^n R_i^2 \underset{H_0}{\overset{H_1}{>}} Tn,$$

where the symbol $\underset{H_0}{\overset{H_1}{>}}$ signifies that we choose hypothesis H_1 (H_0) if the left hand (right hand) side of the equation is greater than the other side.

To compute the false alarm probability (the miss probability is computed similarly), we need to compute

$$\begin{aligned} & P\left(\sum_{i=1}^n R_i^2 > Tn \mid H_0 \text{ is true}\right) \\ &= P\left(\sum_{i=1}^n (s_i + N_i)^2 > Tn\right). \end{aligned}$$

The exponential rate with which this probability goes to zero can be computed easily from Theorem 3. Note that

the moment generating function of a general term in the summand is given by,

$$\begin{aligned} M_i(\theta) &= E[\exp(\theta(s_i + N_i)^2)] \\ &= \frac{1}{\sqrt{1-2\theta}} \exp\left(\frac{\theta^2 s_i^2}{\frac{1}{2}-\theta} + \theta s_i^2\right). \end{aligned}$$

Thus

$$\begin{aligned} \phi(\theta) &= \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log M_i(\theta) \\ &= P_s \frac{\theta}{1-2\theta} - \frac{1}{2} \log(1-2\theta). \end{aligned}$$

The rate of the probability to zero is thus,

$$\begin{aligned} I(T) &= \sup_{\theta} [\theta T - \phi(\theta)] \\ &= \frac{2T - \alpha}{4} - P_s \frac{2T - \alpha}{2\alpha} + \frac{1}{2} \log \frac{\alpha}{2T}, \end{aligned}$$

where $\alpha = 1 + \sqrt{1 + 4TP_s}$.

Suppose we choose to simulate with an input strategy where we compute the sum of independent random variables $\sum_{i=1}^n (Y_{i,\nu} + s_i)^2$ where $Y_{i,\nu}$ is the (functional) exponential shift of N_i , i.e. the probability density of Y_i is given by

$$q_{i,\nu}(x) = \frac{\exp(\nu(x + s_i)^2) \exp(-\frac{x^2}{2})}{M_i(\nu) \sqrt{2\pi}}.$$

Completing the square, we find that $q_{i,\nu}$ is a Gaussian probability density function with mean $2\nu s_i / (1 - 2\nu)$ and variance $1/(1 - 2\nu)$. If we choose $\nu = \nu_o = (2T - \alpha)/4T$, the resulting simulation distributions $\{q_{i,\nu_o}\}$ are Gaussian with means $s_i(2T - \alpha)/\alpha$ and common variance $2T/\alpha$. We can easily compute $R(T)$ for this choice to find that $R(T) = 2I(T)$ and thus this is an *efficient* simulation strategy. Since the input strategy is efficient, the output strategy must also be efficient (by Theorem 1).

We recall that if $\{U_i\}$ are i.i.d. standard normal random variables (denote the standard normal density as p) and $\{\delta_i\}$ are constants, then $\sum_{i=1}^n (U_i + \delta_i)^2$ has a non-central chi-square distribution with n degrees of freedom and non-centrality parameter $\lambda = \sum_{i=1}^n \delta_i^2$. An explicit expression for the density is given by [9][Eq. 28.3.5 (note typographical error)],

$$f_n(\lambda, x) = \frac{1}{2} \left(\frac{x}{\lambda}\right)^{\frac{1}{4}(n-2)} I_{\frac{1}{2}(n-2)}(\sqrt{\lambda x}) \exp\left(-\frac{1}{2}(\lambda + x)\right).$$

Thus the distribution of $\sum_{i=1}^n (Y_{i,\nu_o} + s_i)^2$ has probability density $f_n(\lambda_q, \frac{\alpha}{2T} x) \frac{\alpha}{2T}$, where

$$\lambda_q = \frac{2\alpha}{T} \sum_{i=1}^n s_i^2.$$

Thus, we can conclude, by Theorem 1, that the output simulation strategy with this as the biasing probability density is efficient ($R(T) = 2I(T)$). (The authors note that to try

to show this directly from the definition of efficiency appears to be very difficult.)

Now let us consider a simpler input simulation strategy. We simulate $\sum_{i=1}^n (Y_i' + s_i)^2$ where $\{Y_i'\}$ are i.i.d. Gaussian with mean m_q and variance σ_q^2 . This is a simpler biasing distribution since the mean is held constant and does not vary with the index i . Let us choose m_q, σ_q^2 so that the associated output simulation distribution also has density $f_n(\lambda_q, \frac{\alpha}{2T}x) \frac{\alpha}{2T}$. Thus the associated output simulation is efficient. We can indeed do this if we choose $\sigma_q^2 = 2T/\alpha$ and

$$m_q = -m_s + \sqrt{m_s - (1 - \frac{2T}{\alpha}P_s)}.$$

We denote the normal density with these choices for mean and variance as q' . We still need to compute $R(T)$ for this input strategy. We thus compute

$$\begin{aligned} c(\theta) &= \lim_{n \rightarrow \infty} \frac{1}{n} \log \int \exp(\theta(x + s_i)^2) \frac{p(x)^2}{q'(x)} dx \\ &= \frac{1}{2} \log\left(\frac{\sigma_q^2}{2}\right) - \frac{1}{2} \log\left(1 - \theta - \frac{1}{2\sigma_q^2}\right) + \theta P_s + \frac{m_q^2}{2\sigma_q^2} \\ &\quad + \frac{\theta P_s}{1 - \theta - \frac{1}{2\sigma_q^2}} - \frac{\theta m_s m_q}{\sigma_q^2(1 - \theta - \frac{1}{2\sigma_q^2})} \\ &\quad + \frac{m_q^2}{4\sigma_q^4(1 - \theta - \frac{1}{2\sigma_q^2})}. \end{aligned}$$

Now we must compute $R(T) = \sup_{\theta} [\theta T - c(\theta)] = \theta_T T - c(\theta_T)$. The optimizing value of θ is given by

$$\begin{aligned} \theta_T &= 1 - \frac{1}{2\sigma_q^2} - \frac{1}{4T} \\ &\quad + \frac{\sqrt{\frac{1}{4} + 4T[P_s(1 - \frac{1}{2\sigma_q^2})^2 - \frac{m_s m_q}{\sigma_q^2}(1 - \frac{1}{2\sigma_q^2}) + \frac{m_q^2}{4\sigma_q^4}]}{2T}. \end{aligned}$$

We can compute the variance rate for the constant mean estimator and compare it to an efficient estimator for a variety of signal sequences. In Fig. 2 we see that for values of T near $1 + P_s$, the constant mean estimator works quite well for sawtooth waveforms but has increasingly bad performance as T gets large. In Fig. 3 we see that for T values near $1 + P_s$, we actually have a *negative* variance rate, indicating that we are doing far worse than just a direct Monte Carlo simulation. This isn't too surprising given that for this sinusoidal signal, one would think that a constant mean estimator should have mean near zero (which our choice doesn't have).

In Fig. 4, we plot a typical simulation of the two input estimators and the associated output estimator as a function of the number of simulation runs. We choose $T = 48.5$ and $s(j) = j$, $j = 1, \dots, m$, $m = 10$. The two efficient estimators are giving (after 7000 runs) a value of 8.65×10^{-5} which has an error of at most 5% with 95% confidence. To get this same level of accuracy and confidence with the constant mean estimator, we estimate that 1.8×10^7 simulation runs would be required.

IV. SUMMARY AND CONCLUSIONS

We consider the performance loss or gain to be realized from applying an intermediate or output biasing scheme versus an input biasing scheme in the Monte Carlo simulation of rare events in probabilistic systems. We show that the output scheme is optimal but not uniquely optimal. Of course the drawback is that for a complicated system it may be nearly impossible to compute the weight functions for the output method since in general the weight functions are the ratio of the probability densities of the output variables which are in general very difficult to compute. However, in almost any real simulation problem, the practitioner will be able to combine analytically some of the front end random variables (up to some point). The question is whether he should bother to do this or not. Theorem 1 tells the practitioner that indeed this is (in general) a good practice and even provides him some mathematical guidelines and tools to decide if the variance reduction to be gained is worth the effort. In our opinion, that this is indeed a fundamental result to the science of importance sampling in systems.

APPENDIX

Proof of Theorem 1: Without loss of generality, we can just consider the relationship between the input and intermediate estimators. Also without loss of generality, we just suppose that that $d = 1$, otherwise we could just work with the r th component of the estimators and have the same supposition.

Since the two estimators have the same mean, it suffices to compare the second moments of typical terms. For simplicity we will write $(\tilde{X}_1^{(j)}, \tilde{X}_2^{(j)}, \dots, \tilde{X}_N^{(j)}) = \tilde{X}$ and $(\tilde{Y}_1^{(j)}, \tilde{Y}_2^{(j)}, \dots, \tilde{Y}_M^{(j)}) = \tilde{Y}$. Thus, the typical term of the input estimator has second moment,

$$\begin{aligned} E[f(h(g(\tilde{X})))^2 \frac{dP_x}{dQ_x}(\tilde{X})^2] &= E[f(h(\tilde{Y}))^2 \frac{dP_x}{dQ_x}(\tilde{X})^2] \quad (1) \\ &= E[f(h(\tilde{Y}))^2 E[\frac{dP_x}{dQ_x}(\tilde{X})^2 | \tilde{Y}]] \end{aligned}$$

while the typical term for the intermediate estimator has second moment,

$$E[f(h(\tilde{Y}))^2 \frac{dP_y}{dQ_y}(\tilde{Y})^2] = E[f(h(\tilde{Y}))^2 E[\frac{dP_x}{dQ_x}(\tilde{X}) | \tilde{Y}]^2],$$

where we have used the easily verified fact that

$$\frac{dP_y}{dQ_y}(\tilde{Y}) = E[\frac{dP_x}{dQ_x}(\tilde{X}) | \tilde{Y}].$$

Now observe that by Jensen's inequality,

$$E[\frac{dP_x}{dQ_x}(\tilde{X}) | \tilde{Y}]^2 \leq E[\frac{dP_x}{dQ_x}(\tilde{X})^2 | \tilde{Y}].$$

Hence the general term for the input estimator has greater than equal second moment (and hence greater than or equal variance) than that of the intermediate estimator. We note

also that we have equality in the Jensen's inequality if and only if $\frac{dP_x}{dQ_x}(\tilde{X})$ conditioned on \tilde{Y} is almost surely a constant (dependent possibly on \tilde{Y}). This is equivalent to $\frac{dP_x}{dQ_x}(\tilde{X}) = s(\tilde{Y})$ for some deterministic function s . This completes the proof of the theorem.

Remark 3: In writing (2), we appealed to two elementary properties of conditional expectation: equations (34.6) and (34.4) of [2]. Eq. (34.6) requires that the right-hand side of (1) be finite. However, if (1) is infinite, the theorem is trivial. To use (34.4) further requires that $E[(dP_x/dQ_x(\tilde{X}))^2] < \infty$. If this is not the case, put $L_n(\cdot) = \min(dP_x/dQ_x(\cdot), n)$, and write

$$\begin{aligned} & E[f(h(\tilde{Y}))^2 \frac{dP_x}{dQ_x}(\tilde{X})^2] \\ &= \lim_{n \rightarrow \infty} E[f(h(\tilde{Y}))^2 L_n(\tilde{X})^2] \\ &= \lim_{n \rightarrow \infty} E[f(h(\tilde{Y}))^2 E[L_n(\tilde{X})^2 | \tilde{Y}]] \\ &\geq \lim_{n \rightarrow \infty} E[f(h(\tilde{Y}))^2 E[L_n(\tilde{X}) | \tilde{Y}]^2] \\ &= E[f(h(\tilde{Y}))^2 \lim_{n \rightarrow \infty} E[L_n(\tilde{X}) | \tilde{Y}]^2] \\ &= E[f(h(\tilde{Y}))^2 E[\frac{dP_x}{dQ_x}(\tilde{X}) | \tilde{Y}]^2] \end{aligned}$$

where we use the monotone convergence theorem along with a conditional dominated convergence theorem [2, Theorem 34.2(v)]. Note also that $dP_x/dQ_x(\tilde{X})$ is integrable since its expectation is one.

The proof of Theorem 2 requires the use of the following theorem a proof of which may be found in [4][Theorem II.2].

Theorem 3: Let $\{Y_n\}$ be a sequence of \mathcal{R}^d random variables and define for $\theta \in \mathcal{R}^d$,

$$\phi_n(\theta) = \frac{1}{n} \log E[\exp(\langle \theta, Y_n \rangle)].$$

Suppose that assumptions A1,A2,A3 hold for the sequence of convex functions ϕ_n (with limit ϕ). Define for $x \in \mathcal{R}^d$, $I(x) = \sup_{\theta} [\langle \theta, x \rangle - \phi(\theta)]$. Given a subset A of \mathcal{R}^d , define

$$I(A) = \inf\{I(x) : x \in A\}.$$

Let A be any Borel set such that $A^o \neq \emptyset$, $\bar{A} = \bar{A}^o$ and $0 < I(E) < \infty$. Then

$$\lim_{n \rightarrow \infty} P(Y_n \in nE) = -I(E).$$

Proof of Theorem 2: Note that we can define a probability measure, μ_n , (if $c_n(0) < \infty$) by defining for every Borel set $B \subset \mathcal{S}_n$,

$$\mu_n(B) = \int_B \exp(-c_n(0)) \frac{dP_n(z)}{dQ_n(z)} dP_n(z).$$

Let Y_n be a \mathcal{S}_n valued random variable with associated probability measure μ_n . Then $\{f_n(Y_n)\}$ is a sequence of

\mathcal{R}^d valued random variables. For $\theta \in \mathcal{R}^d$,

$$\begin{aligned} \phi_n(\theta) &= \frac{1}{n} \log \left(\int \frac{dP_n(z)}{dQ_n(z)} \exp(\langle \theta, f_n(z) \rangle) \exp(-c_n(0)) dP_n(z) \right) \\ &= c_n(\theta) - c_n(0). \end{aligned}$$

Under the stated assumptions (A1,A2,A3) we have made for the $\{c_n\}$ sequence and its limit c , we also have the same assumptions holding for the $\{\phi_n\}$ sequence and its limit ϕ . Hence Theorem 3 holds for the random variables $\{f_n(Y_n)\}$. This implies that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \int_{\frac{f_n(z)}{n} \in E} d\mu_n(z) = - \sup_{x \in E} [\langle \theta, x \rangle - c(\theta) - c(0)].$$

Thus,

$$\begin{aligned} F_n &= \int \left(\frac{dP_n(z)}{dQ_n(z)} \right)^2 1_{\{\frac{f_n(z)}{n} \in E\}} dQ_n(z) \\ &= \exp(nc_n(0)) \int_{\frac{f_n(z)}{n} \in E} d\mu_n(z) \end{aligned}$$

and so,

$$\begin{aligned} \frac{1}{n} \log F_n &= c_n(0) + \frac{1}{n} \log \int_{\frac{f_n(z)}{n} \in E} d\mu_n(z) \\ &\rightarrow_{n \rightarrow \infty} c(0) - \sup_{x \in E} [\langle \theta, x \rangle - c(\theta) - c(0)] \\ &= - \sup_{x \in E} [\langle \theta, x \rangle - c(\theta)] \\ &= -R(E). \end{aligned}$$

This completes the proof of the theorem.

REFERENCES

- [1] W. A. Al-Qaq, M. Devetsikiotis, and J. K. Townsend, "Importance Sampling Methodologies for Simulation of Communication Systems with Time-Varying Channels and Adaptive Filters," *IEEE Journal on Selected Areas of Communications*, Vol. 11, No. 3, pp. 317-326, April 1993.
- [2] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.
- [3] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*, Wiley-Interscience, New York, 1990.
- [4] R. Ellis, "Large deviations for a general class of random vectors," *Ann. Probab.*, Vol. 12, No. 1, pp. 1-12, 1984.
- [5] W. Feller, *An Introduction to Probability Theory and its Applications, Vol. I*, third edition, Wiley, New York, 1968.
- [6] P. Hahn and M. Jeruchim, "Developments in the theory and application of importance sampling," *IEEE Transactions on Communications*, Vol. COM-35, No. 7, pp. 706-714, July 1987.
- [7] P. Heidelberger, "Fast simulation of rare events in queuing and reliability models," *ACM Trans. Modeling and Comput. Simulation*, Vol. 5, No. 1, 1995.
- [8] M.C. Jeruchim, P. Balaban, and K.S. Shanmugan, *Simulation of Communications Systems: Modeling, Methodology, and Techniques, 2nd Edition*, Kluwer Academic/Plenum Publishers, New York, 2000.
- [9] N. Johnson and S. Kotz, *Continuous Univariate Distributions-2*, Wiley-Interscience, New York, 1970.
- [10] D. Lu and K. Yao, "Improved importance sampling techniques for efficient simulation of digital communication systems," *IEEE Journal on Selected Areas in Communication*, Vol. 6, No. 1, pp. 67-75, January 1988.
- [11] B. Ripley, *Stochastic Simulation*, Wiley, New York, 1987.

- [12] P.J. Smith, M. Shafi, and H. Gao, "Quick Simulation: A Review of Importance Sampling Techniques in Communication Systems," *IEEE Journal on Selected Areas in Communications*, Vol. 15, No.4, pp. 597-613, May 1997.
- [13] R.J. Wolfe, M.C. Jeruchim, and P. Hahn, "On Optimum and Suboptimum Biasing Procedures for Importance Sampling in Communication Simulation," *IEEE Trans. Comm.*, Vol. 38, No. 5, pp. 639-646, May 1990.

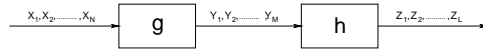


Fig. 1. A multi-input, multi-output system.

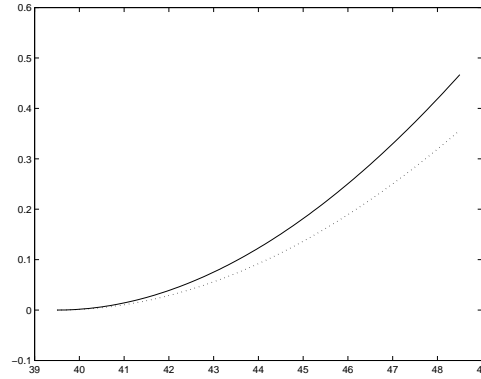


Fig. 2. The variance rate for the constant mean estimator (dotted line) compared to that of the efficient estimator (solid line) for a sawtooth input.

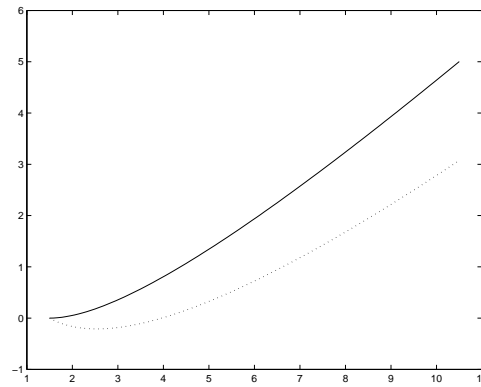


Fig. 3. The variance rate for the constant mean estimator (dotted line) compared to that of the efficient estimator (solid line) for a sinusoidal input.

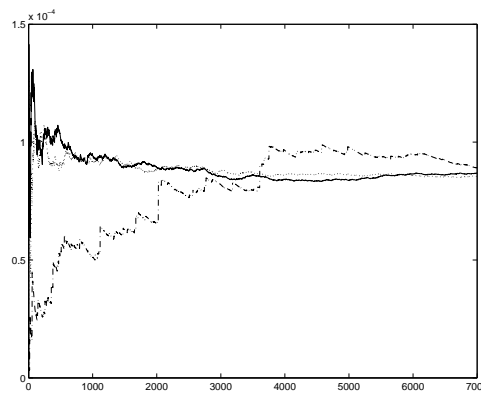


Fig. 4. Estimator values as a function of the number of simulation runs. The solid line is the efficient output estimator, dotted is the efficient input estimator, and dashed-dotted is the constant mean input estimator.