

On the Monte Carlo Simulation of Digital Communication Systems in Gaussian Noise

J. A. Bucklew, *Senior Member, IEEE*, and R. Radeke

Abstract—In this paper, we consider the general problem of the simulation of highly reliable systems operating in the presence of Gaussian noise. Our methodology uses importance sampling which has been shown to be a particularly effective method in the general discipline of rare-event simulation. The methods we propose are optimal in a certain sense, i.e., they are efficient. We also give a new class of simulation distributions that are universally efficient in the sense that they depend *only* on a single scalar parameter, regardless of the dimensionality of the underlying system or of the error sets to be simulated.

Index Terms—Gaussian noise, importance sampling, Monte Carlo simulation.

I. INTRODUCTION

MODERN communication systems are complex, nonlinear, highly reliable systems designed to operate in noisy environments. Because of their complexity, they are very difficult, if not impossible, to analyze mathematically in closed form. Due to this analytic intractability, they must often be simulated in order to obtain estimates of the key performance parameters.

In this paper, we present a general simulation methodology to use for complex, highly reliable stochastic systems operating in the presence of Gaussian noise. Gaussian noise is the most common assumption used for the ambient thermal background in which most communications systems have to operate. Our method can be extended to other noise models (most notably, certain types of shot noise) but we will not further discuss this issue here.

In communication system design, an event of rare probability is usually the key parameter of the system's efficacy. The events we have in mind here would be something like an error in the transmission of a bit. Typically, this number is on the order of 10^{-5} – 10^{-8} . The rule of thumb for a straightforward Monte Carlo simulation is that we need $100/\rho$ number of simulations to reasonably estimate (20% error with 95% probability) an event of probability ρ . Hence, to estimate the small probabilities of error found in digital communications via brute-force direct simulation requires that a very large number (an impractical number, even) of independent random numbers be generated from the computer's random number generator.

Paper approved by A. Ahlen, the Editor for Modulation and Signal Design of the IEEE Communications Society. Manuscript received September 15, 2001; revised April 15, 2002 and July 15, 2002.

J. A. Bucklew is with the Department of Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI 53706-1691 USA.

R. Radeke is with the Department of Electrical Engineering and Information Technology, Dresden University of Technology, 01069 Dresden, Germany.

Digital Object Identifier 10.1109/TCOMM.2003.809280

One way out of this quandary is to utilize the 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 of this methodology in the field of network simulation, see [9]. Another highly recommended review article in the field of communications systems is [15]. An encyclopedic text concerned only with the issues present in communication system simulation is [10].

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 independent, identically distributed (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. 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 be converging (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, of the recent interest in importance sampling techniques.

In this paper, we consider a certain important special case, that of rare-event simulation for systems operating in the presence of Gaussian noise. We present a philosophy based upon minimizing (more or less) the asymptotic rate to zero of the

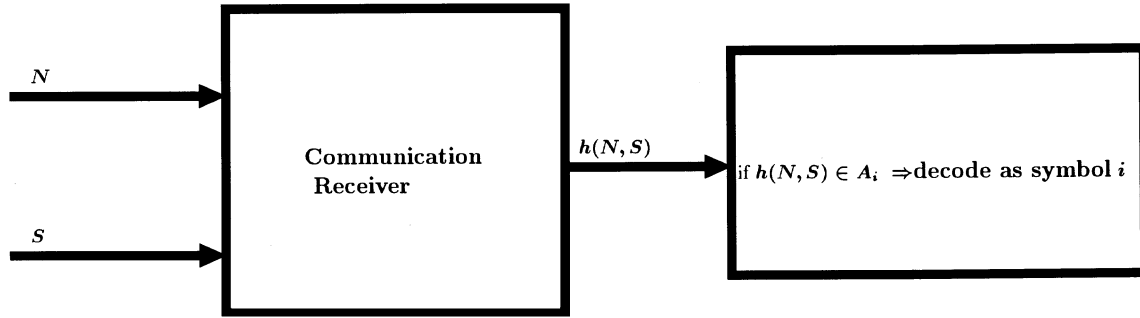


Fig. 1. Generic digital communication model.

estimator variance. Simulation strategies that achieve this minimum rate are said to be efficient. This philosophy of efficient simulation allows us to choose biasing strategies based upon a very simple geometric interpretation of the error sets of the system under consideration. For example, in the white noise case, the methodology indicates that we shift the mean of our original zero-mean noise to the point of the error set closest to the origin (if that point is unique). In the nonwhite case, we use a generalized distance, but the ideas remain the same. We show that if there is more than one point at minimum distance from the origin, we should use a biasing density that is a mixture. A new result in this paper is that in the setting of an infinite number (countable or uncountable) of minimum distance points, we prove that a particular mixture distribution strategy is efficient.

The geometry of these error sets, especially in high-dimensional spaces, can be very complicated. It may be very hard to specify (or even find) the minimum distance points. A key advantage of this mixture distribution is that it is parameterized by a single scalar parameter, the generalized distance. This means that the search for a good biasing distribution can take place over a single scalar parameter, regardless of the dimensionality of the problem or the shape of the error sets!

In Section II, we discuss a result from the theory of large deviations and how it can be used to suggest a simulation methodology. In Section III, we discuss the problem of efficient estimator construction and present our main theorem, giving an explicit universal family of Gaussian simulation distributions. Section IV is devoted to several examples using the theory developed. Section V is a final discussion over the philosophy of importance sampling and its place in rare-event simulation. The Appendix is devoted to a proof of the main theorem.

II. PROBLEM SETUP AND PRELIMINARIES

Let N be a d -dimensional Gaussian random vector, with mean m and covariance matrix K . Define the so-called rate function $R(\cdot)$ as

$$R(x) = \frac{1}{2} x^T K^{-1} x.$$

From the probabilistic theory of large deviations, it is known that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log P \left(\frac{1}{\sqrt{n}} N \in F \right) \leq - \inf_{x \in F} R(x) \quad (1)$$

for each closed set $F \subset \mathbb{R}^d$, and

$$\liminf_{n \rightarrow \infty} \frac{1}{n} \log P \left(\frac{1}{\sqrt{n}} N \in O \right) \geq - \inf_{x \in O} R(x)$$

for each open set $O \subset \mathbb{R}^d$. The term “rate function” is appropriate, since it gives the asymptotic rate to zero of these probabilities as $n \rightarrow \infty$. In other words, if $\inf_{x \in A} R(x) = r$, then (in some sense) $P((1/\sqrt{n})N \in A) \approx \exp(-nr)$. In applications, in the vast majority of cases, we can take our set of interest to be closed (or open) by just including (or not) the boundary of where a threshold decision is to be made without affecting the overall symbol probability of error. Hence, we will assume that for the sets of interest A in our applications, that $\bar{A} = A^\circ$ where \bar{A} denotes the closure of the set A and A° denotes the interior of the set A . Hence

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \log P \left(\frac{1}{\sqrt{n}} N \in A \right) &= - \inf_{x \in A} R(x) = R(A) \\ &= - R(x_A). \end{aligned}$$

These probabilities give us a way to approximate the probabilities of “small” Gaussian random variables lying in sets. We will use these approximations as a way to formulate good simulation strategies for small Gaussian noise digital communication problems. A point x_A where the minimization of the rate function actually occurs is called a *minimum rate point* of the set A . Note that there may be many minimum rate points, depending on the shape of the set A . The set of points $B_a = \{x : R(x) = a\}$ is, of course, the boundary of a hyperellipsoid in \mathbb{R}^d . If $\inf_{x \in A} R(x) = a$, all of the points where the boundary of A intersects B_a are minimum rate points. The number of such points could be one, or several, or countably infinite, or even uncountably infinite. To say more, we typically need to know something about the structure or shape of the set. For example, if the set A is convex and does not contain the origin, then there can only be one such minimum rate point.

Consider the system simulation problem depicted in Fig. 1. Here, we suppose that a Gaussian noise vector N is input to a communication receiver along with an independent message symbol S . The receiver performs some operations on these inputs and outputs the function $h(N, S) \in \mathbb{R}^{d'}$. Depending on where in $\mathbb{R}^{d'}$ this d' -dimensional vector lies, we decode the symbol. In other words, the receiver partitions the space $\mathbb{R}^{d'}$ into disjoint sets $\{D_i\}$. If $h(N, S) \in D_i$, the receiver announces S_i as the transmitted symbol (or, equivalently, announces i , the subscript of the symbol). Note that without loss of generality

(due to our very general system model) we can just assume that the noise N is zero mean. (If not, just modify the system model so as to add the mean value in as part of its operation.)

An error occurs whenever symbol j is decoded as symbol i where $i \neq j$. To evaluate the average bit-error rate (BER), we need to estimate $p_{ji} = P(\text{announce symbol } j | \text{symbol } i \text{ transmitted})$ for all pairs $i \neq j$. One of our underlying assumptions is that we are dealing with a highly reliable system, and hence, the probabilities $\{p_{ji}\}$ are very small. Hence, we assume that our zero-mean noise vector N can be written as

$$N = \frac{G}{\sqrt{n}} \quad (2)$$

where G is some fixed zero-mean Gaussian vector with covariance matrix K , multidimensional probability density $p(\cdot)$, and n is a large integer. (Note that the probability density of N is, thus, $p(\cdot\sqrt{n})\sqrt{n}^d$.)

Given that the symbol transmitted is independent of the noise, it is true that there exists a set $A_{ji} \subset \mathbb{R}^d$ such that

$$\rho_{ji,n} = P(N \in A_{ji}).$$

The significance of the set A_{ji} is that, if the noise vector is in this set and if symbol i was transmitted, then the receiver will announce j , and thus, create a certain type of symbol error.

Instead of directly simulating the noise vector G with density p and scaling it by dividing by \sqrt{n} , we shall use importance sampling and simulate another noise vector Y_n with density q_n . The form of our estimator is

$$\hat{\rho}_{ji,n} = \frac{1}{k} \sum_{l=1}^k 1_{\{Y_n^{(l)} \in A_{ji}\}} \frac{p(\sqrt{n}Y_n^{(l)})\sqrt{n}^d}{q_n(Y_n^{(l)})}$$

where $Y_n^{(l)}$ is the l th independent sample drawn from the q_n density.

The variance of this estimate goes to zero like $1/k$ as the number of sample runs k increases (since it is just an average of k independent random variables). The behavior in n is a little bit more difficult to see, however, from [3], we know that (usually) the variance is decreasing to zero exponentially fast in n .

Since it is an unbiased estimate, we can write

$$k \text{Var}(\hat{\rho}_{ji,n}) = F_{q_n} - \rho_{ji,n}^2$$

where

$$F_{q_n} = E_{q_n} \left[1_{\{Y_n \in A_{ji}\}} \left(\frac{p(\sqrt{n}Y_n)\sqrt{n}^d}{q_n(Y_n)} \right)^2 \right]$$

where we note that F_{q_n} is the just the mean-square value of one of the summand terms in the estimator.

From our large deviation result in *Theorem 1*, we know that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \rho_{ji,n} = - \inf_{x \in A_{ji}} R(x) = -R(x_{A_{ji}})$$

where

$$R(x) = \frac{1}{2} x^T K^{-1} x.$$

Hence, $\rho_{ji,n}^2$ goes to zero exponentially in n with rate $2R(x_{A_{ji}})$. Since the variance is always nonnegative, we must have that the

rate with which F_{q_n} goes to zero (if it does at all) must be less than or equal to $2R(x_{A_{ji}})$. If F_{q_n} goes to zero at exactly the rate $R(x_{A_{ji}})$, we say that the estimator is *efficient*.

In the next section, we address the problem of finding efficient estimators in the setting of our problem.

III. EFFICIENT ESTIMATORS

In general, there is no unique sequence of efficient estimators. If we impose enough conditions on what sort of bias distributions will be allowed, sometimes we can specify some uniquely optimal efficient estimator. We will consider a class of efficient estimators based upon the so-called exponential shifts. In the Gaussian setting, this corresponds to shifting the mean vector of the original distribution. In [13], it was stated (in error) that these were the only possible efficient estimators in the Gaussian setting. In a subsequent paper, Schlebusch [14] corrected this error and showed that in the small parameter Gaussian setting (our current setting), there are several possibilities (other than mean shifts) for efficient biasing distributions.

Recall that we wish to estimate $\rho_{ji,n}$ using the estimator $\hat{\rho}_{ji,n}$. The situation is complicated much by the question of the number of minimum rate points. In the setting where there are only a finite number of minimum rate points $\nu_1, \nu_2, \dots, \nu_m$, both [13] and [14] give formulas for efficient sequences. Following the mean-shift strategy, [13] (under some mild technical assumptions which we do not repeat here) prove that the following choice is efficient:

$$q(x) = \sum_{l=1}^m p_l p((x - \nu_l)\sqrt{n})\sqrt{n}^d$$

where $\{p_l\}$ is an arbitrary probability vector with all positive components.

The simulation situation where there is an infinite number of minimum rate points has not been considered in the literature. The existence of efficient estimators has been conjectured in both [13] and [14]. Indeed, the obvious conjecture is that if we replace $\{p_l\}$ by some arbitrary probability distribution with support exactly equaling the set of minimum rate points, we will have an efficient biasing distribution. Unfortunately, we still cannot say whether even this particular choice is always efficient.

In this important small Gaussian setting, we can indeed demonstrate a universally optimal choice. Consider an arbitrary closed set A and denote $\min_{\{x \in A\}} R(x) = R(A) = r$. We make no further specifications on the set. Hence, the set could have an uncountable number of minimum rate points. In the Appendix, we give a proof of the following theorem.

Theorem 1: The sequence of bias probability densities given by

$$q_n(x) = p(x\sqrt{n})\sqrt{n}^d \exp(-nr) \frac{\Gamma(\frac{d}{2})\sqrt{2}^{(d/2)-1}}{\sqrt{r}^{(d/2)-1}} \times \frac{I_{(d/2)-1}(n\sqrt{2r}\|K^{-(1/2)}x\|)}{(n\|K^{-(1/2)}x\|)^{(d/2)-1}}$$

is efficient, where $I_\nu(\cdot)$ is a modified Bessel function.

Note that the only information about the set contained in the $\{q_n\}$ sequence is the scalar quantity r !

Remark 1: Despite its formidable appearance, obtaining samples for this distribution is actually quite simple. In the Appendix, we show that it is obtained as the result of randomly mean-shifting our original Gaussian noise vector over the surface of a hyperellipsoid. To obtain a sample, first generate U , a uniform random variable over the surface of the unit radius d -dimensional hypersphere centered at the origin. One way to generate U is to generate a d -dimensional Gaussian sample, Z , with covariance matrix id . Then return $U = Z/\|Z\|$. (Note that any random variable with a spherically symmetric distribution can be used instead of Z for this purpose.) We then generate a sample of the noise distribution N and return, $Y = N + \sqrt{2r}K^{1/2}U$. Y will have the density $q_n(\cdot)$.

IV. EXAMPLES

Example 1: Let $X = (X_1, X_2)$ be a two-dimensional Gaussian random variable, with independent standard Gaussian marginals. Suppose that we are interested in $\rho = P(X \in A)$ where A is the complement of a disk of radius 10 centered at the origin. Every point on the boundary of the circle of radius 10 is, thus, a minimum rate point. It is easy enough to analytically compute the probability in this case, giving

$$\begin{aligned} \rho &= \int_{10}^{\infty} z \exp\left(-\frac{z^2}{2}\right) dz \\ &= \exp(-50) \approx 1.93 \times 10^{-22}. \end{aligned}$$

We, of course, wish to simulate this probability as a test of our theory developed above. We use the density given in the previous theorem, with $n = 1$, $d = 2$, $K = id$, $r = 50$, or

$$q(x) = \exp\left(-\frac{\|x\|^2}{2}\right) \frac{\exp(-50)}{2\pi} I_0(10\|x\|).$$

We should remark that there is a degree of freedom here that we have not really discussed. We chose $n = 1$, $K = id$, but we could have chosen n to be anything and $K = n \cdot id$. We get exactly the same estimators, no matter what our choice of n is. What is happening is that we have an asymptotic theory based upon large n . However, any one simulation problem is a fixed noise variance that we need to emulate. We have only assumed that our noise vector N can be expressed as $N = G/\sqrt{n}$, where G has covariance K . Clearly, we can choose n to be whatever we desire, as long as we scale K appropriately also.

To generate samples from this distribution, we first generate samples uniform on the boundary of the disk of radius 10 as $U = (10 \cos(\theta), 10 \sin(\theta))$, where θ is a uniform $[0, 2\pi]$ random variable. We then generate $Y = (Y_1, Y_2)$ as $Y = X - U$. Our estimate then appears as

$$\hat{\rho} = \frac{1}{k} \sum_{l=1}^k 1_{\{\|Y^{(l)}\| > 10\}} \frac{\exp(50)}{I_0(10\|Y^{(l)}\|)}.$$

Using the standard random number generators of Matlab, we chose $k = 20\,000$ and obtained a value $\hat{\rho} = 1.8962 \times 10^{-22}$, giving an error of less than 2%.

Example 2 (Simulation of Intersymbol Interference (ISI) and the Linear Problem): Suppose that we have a digital signal corrupted by an ISI channel and white Gaussian noise. The received sample appears as

$$R_m = B_m + \sum_{k=1}^l a_k B_{m-k} + N_m$$

where $B_m \in \{-1, 1\}$ is the information bit at time m , and N_m is the i.i.d. normal, mean zero, variance σ^2 , noise sample at time m . This signal is then passed through an inverting filter to give

$$Y_m = B_m + \sum_{k=0}^L i_k N_{m-k}.$$

We assume here that a perfect inverting finite impulse response (FIR) filter exists, which, of course, cannot mathematically be true. We assume that for large enough L , this is a good approximation.

Therefore, to simulate an error, we wish to bias the noise samples $N = (N_m, N_{m-1}, \dots, N_{m-L})$. Suppose for now that $B_m = +1$. We wish to mean shift N to the unique (hopefully) minimum rate point of the error set

$$A = \left\{ x = (x_{m-L}, x_{m-L+1}, \dots, x_m) : \sum_{k=0}^L i_k x_{m-k} < -1 \right\}.$$

In this setting, $R(x) = (1/2\sigma^2)\|x\|^2$. To find the minimum rate point, we need to minimize $\sum_{k=0}^L x_{m-k}^2$ subject to $\sum_{k=0}^L i_k x_{m-k} < -1$. We do this with Lagrange multipliers by finding the critical point of $J(x) = \sum_{k=0}^L x_{m-k}^2 + \lambda \sum_{k=0}^L i_k x_{m-k}$, which leads to

$$x_{m-j} = -\frac{\lambda}{2} i_j.$$

Solving for the constraint gives us

$$x_A = -\frac{i}{\|i\|^2}.$$

Therefore, we should bias our input random noises so that they have x_A as their mean value.

This problem (in a much more general setting) is considered in [16]. Here, our solution is presented from a large deviation theory perspective, as opposed to a direct minimization of the estimator variance. The end results match up well in the simple cases where they can be compared.

Example 3 (Decision Feedback Equalizer): We are interested in finding the error probability caused by a decision feedback equalizer operating in an ISI channel in the presence of white Gaussian noise. As in the previous section, we receive the following bit sample:

$$R_m = B_m + \sum_{k=1}^l a_k B_{m-k} + N_m$$

where $B_m \in \{-1, 1\}$ is the information bit sent at time m , the $\{a_k\}$ are the ISI channel coefficients, and $\{N_m\}$ are the i.i.d. Gaussian, zero mean, variance σ^2 , noise samples. We then use

our past bit decisions and our current channel model (which is updated using an adaptive filtering algorithm) to try to cancel the current samples ISI, giving us the output

$$\hat{R}_m = R_m - \sum_{k=1}^l \hat{a}_k B_{m-k}.$$

We then take the sign of this result to get our current bit decision

$$\hat{B}_m = \text{sign}(\hat{R}_m)$$

For simplicity in the following analysis, we will assume that the adaptive system is in steady state and locked to the true channel model (which, in particular, implies that $a_k = \hat{a}_k \forall k$).

The problem with these types of receiver structures is that they employ a type of nonlinear feedback. If an error is made in a previous bit decision, the receiver is far more likely to make another error, since the ISI instead of being perfectly cancelled is actually made worse. This causes a “burstiness” to the errors appearing in these systems. Simulating these systems is far more complicated due to this characteristic behavior. We should note that [2] considers a far more general model and proposes various other sophisticated strategies for simulating this system.

What we propose to do is to look at d transmitted bits at a time. We denote a given vector of d transmitted bits as $b = (b_1, b_2, \dots, b_d)$. We then consider the $2^d - 1$ possible error patterns that can occur for those d bits. Any one such error pattern will be denoted as $e = (e_1, e_2, \dots, e_d)$. We propose to compute the probability of each error pattern.

For a given transmitted b and any e , there is a certain set $A_{be} \subset \mathcal{R}^d$ such that if the noise vector $N = (N_1, N_2, \dots, N_d) \in A_{be}$, the error pattern e will be produced. We wish to find the minimum rate points of this set. Hence, we need to minimize

$$R(x) = \frac{1}{\sigma^2} \sum_{i=1}^d x_i^2 \text{ with respect to } \frac{|b - \hat{b}|}{2} = e$$

where again, b is the transmitted bit sequence, \hat{b} the estimated sample sequence at the receiver, and e is the error pattern of which we wish to simulate the probability.

Note that, to achieve the error pattern e , we need

$$\hat{b}_i = b_i(1 - 2e_i).$$

The estimated received bit appears as

$$\hat{b}_i = \text{sign} \left(b_i + \sum_{k=1}^d a_k b_{i-k} - \sum_{k=1}^d \hat{a}_k \hat{b}_{i-k} + x_i \right)$$

and with $a_k = \hat{a}_k$, we get

$$\begin{aligned} \hat{b}_i &= \text{sign} \left(b_i + \sum_{k=1}^d a_k (b_{i-k} - \hat{b}_{i-k}) + x_i \right) \\ \hat{b}_i &= \text{sign} \left(b_i + 2 \sum_{k=1}^d a_k b_{i-k} e_{i-k} + x_i \right). \end{aligned}$$

We denote the channel offset at the receiver as $o = (o_1, o_2, \dots, o_d)$, where

$$o_i = b_i + 2 \sum_{k=1}^f a_k b_{i-k} e_{i-k}.$$

This implies

$$\hat{b}_i = \text{sign}(o_i + x_i).$$

To minimize $R(x)$ subject to the constraint, we can take $x_i = 0$ if

$$\hat{b}_i = b_i(1 - 2e_i) = \text{sign}(o_i)$$

otherwise, we must take

$$x_i = -o_i$$

which leads to

$$x_i = -o_i \frac{|\text{sign}(o_i) - \hat{b}_i|}{2}.$$

This choice of x is denoted as x_{be} being the (unique) minimum rate point of the set A_{be} . To simulate, we then would employ white Gaussian noise with mean x_{be} and individual component variance σ^2 to simulate the ISI channel.

After computing the minimum rate point x_{be} for each data and error pattern of b and e , we should first recognize that those patterns with the smallest norm for x_{be} will have the largest contributions to the overall symbol error rate. For a given data and error pattern, our estimator appears as

$$\begin{aligned} P(b, e) &= \frac{1}{k} \sum_{i=1}^k 1_{\{b^{(i)} = b^{(i)}(1 - 2e^{(i)})\}} \\ &\times \prod_{l=1}^d \exp \left(-\frac{(n_l^{(i)} - x_{be,l})^2 - (n_l^{(i)})^2}{2\sigma^2} \right). \end{aligned}$$

The overall BER is

$$\text{BER} = \sum_{\forall b \forall e} \frac{P(b, e) \sum_{j=1}^d e_j}{2^d d}.$$

To test our theory above, we choose a channel with ISI coefficients $a = (1/2, -1/4, 1/10)$ and white Gaussian noise with zero mean and variance $\sigma^2 = 0.04$. We are interested in the probability that a specific bit pattern produces a specific error pattern. We choose $b = (-1 \ 1 \ -1 \ 1)$ and $e = (1 \ 0 \ 1 \ 1)$. We compute the channel offset as

$$o = \left(-1, 0, -\frac{1}{2}, -\frac{1}{5} \right).$$

Knowing the distribution of the Gaussian noise, we can compute (in closed form) the probability that each decoded bit follows the specific error pattern, giving us the overall probability for this (b, e) combination as

$$\begin{aligned} P(b, e) &= P(\hat{b}_1 = 1)P(\hat{b}_2 = 1)P(\hat{b}_3 = 1)P(\hat{b}_4 = -1) \\ &= P(n_1 > 1)P(n_2 > 0)P(n_3 > 0.5)P(n_4 < 0.2) \\ &= (2.867 \times 10^{-7})(0.5)(6.21 \times 10^{-3})(0.8413) \\ &= 7.488 \times 10^{-10}. \end{aligned}$$

The minimum rate point of the set is calculated as $x_{be} = (1 \ 0 \ 0.5 \ 0)$. Now we simulate the ISI channel using Gaussian

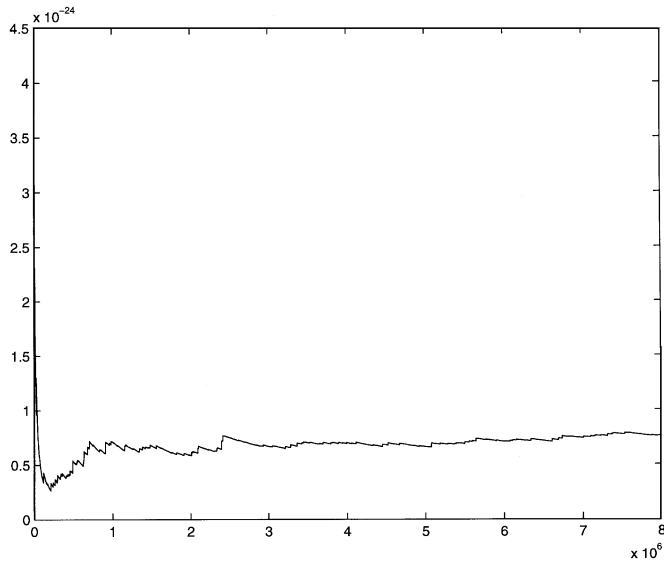


Fig. 2. Estimator output as a function of simulation runs.

noise with mean x_{be} and variance $\sigma^2 = 0.04$. Using $k = 25\,000$, we obtain a probability of $P(b, e) = 7.632 \times 10^{-10}$, giving an error of less than 2% from the true value.

To find the overall bit-error ratio, we use the same scheme to compute the probability of all combinations of bit and error patterns. Supposing that all data bit patterns occur with the same probability, we calculate that the actual BER is $\text{BER} = 3.6336 \times 10^{-7}$.

Simulating all (b, e) combinations by shifting to the minimum rate point and using $k = 25\,000$ simulation runs for each combination, gives us an estimated BER of $\hat{\text{BER}} = 3.6401 \times 10^{-7}$, giving an error of approximately 0.2% from the true value.

Example 4: To present a problem with more multidimensional character, try to estimate

$$\rho = P \left(\sum_{j=1}^3 \cos(N_j) + \sum_{k=1}^3 \sum_{l=1}^3 \cos(N_k - N_l) + N_4 > 13 \right)$$

where N_1, N_2, \dots, N_4 are mean zero, variance $1/100$ i.i.d. Gaussian random variables. Hence, we have a four-dimensional biasing problem. We choose the covariance matrix of the random vector G in (2) as $K = id$. To match up with our problem, then we must choose $n = 100$ to give the correct variance for the N vector. We should note that the point $(g_1, g_2, g_3, g_4) = (0, 0, 0, 1)$ gives $\sum_{j=1}^3 \cos(g_j) + \sum_{k=1}^3 \sum_{l=1}^3 \cos(g_k - g_l) + g_4 = 13$. Furthermore, $I(g) = g^T g / 2 = 1/2$. Thus, we know that r need not be chosen bigger than $1/2$. We choose (after a few trials) the value of $r = .45$.

In Fig. 2, we plot our estimate as a function of the number of runs for up to eight million runs. The final value 7.6155×10^{-25} is (empirically) accurate to $\pm 20\%$ with 95% probability. Of course, a direct Monte Carlo simulation of this probability would have required on the order of 10^{28} simulation runs.

V. DISCUSSION

This paper is more about presenting a simulation philosophy than it is about simulating any one particular type of communi-

cation system. We argue that for highly reliable systems operating in the presence of Gaussian noise, we should consider the large deviation theory of the problem; in particular, we should view our simulation problem from the point of view of trying to find efficient biasing distributions. The reason for following this philosophy is that by first embedding our problem as but one of a parametric sequence of problems (the parameter is n , the small noise control parameter), we can concern ourselves with maximizing the estimator variance rate to zero, instead of minimizing the actual estimator variance itself. The mathematics of maximizing the variance rate is often far simpler than trying to minimize the actual variance over some class of simulation distributions. This is intuitive, since our large deviation framework is only trying to maximize a rate parameter instead of the actual variance. Trying to minimize analytically the estimator variance directly almost always leads to a very complicated functional minimization problem. Of course, when this minimization can be carried out, it is very desirable to do so. In most practical situations, it really cannot be done.

Even in situations where our methodology is difficult to apply and perhaps more *ad hoc* techniques must be used, the theory of large deviations still gives deep theoretical insight into the fundamental problems facing the simulation of highly reliable systems. The situation here is very akin to the position occupied by information theory to the communication system designer. Information theory provides insight into the communication problem but many of its constructs are difficult to compute, e.g., channel capacity, rate distortion functions, etc. However, this fact does not negate its fundamental usefulness as a framework in which to view this problem. At the risk of a bit of hyperbole, it is the authors' opinion that the role of large deviation theory in rare-event simulation is analogous. It is now apparent that we have a more or less encompassing theory that seems to explain quite well exactly what is occurring in this search for good biasing distributions. The theory, and definitely the practice, of large deviation theory techniques is still very much in its infancy. We suspect that much work will have to be done in this area to convince communication systems designers that the necessary front effort of learning something of the mathematical framework will give large payoffs in the science of evaluating highly reliable systems.

During the review of this paper, one of the reviewers pointed out that identifying the error set A in terms of the regions visited by the input Gaussian vector will be a difficult exercise in all but the simplest cases. We agree wholeheartedly with this statement. Our argument is that we give a "universal" biasing strategy that only depends on one parameter. The simulation practitioner need only search (adaptively from the estimator variance would be one possibility) over a single parameter. The true complexity of the set A does NOT need to be investigated. To us, this is an amazing result.

The same reviewer felt compelled to comment on our claim that the mathematics of large deviation theory arguments is often simpler than direct estimator-variance minimization. He argued that in many difficult examples, the techniques of direct estimator-variance minimization require not much more ingenuity to implement than the large deviation theory-based examples of this paper. In particular, he argues that one can adap-

tively minimize an empirical estimator variance over a parametric class of simulation distributions. We reply first that in the present Gaussian setting, it seems to us almost compellingly simple that we only need to find the minimum distance points of the error set from the origin. Even if this is difficult to do in practice, it gives such insights into the problem (which led directly to this new “universal” class of efficient biasing distributions) that it seems obvious that the methodology of large deviation theory cannot be ignored. Second, we have no complaints with the use of empirical/adaptive searches to minimize estimator variance. We are using large deviation theory to propose families of (parametric) simulation distributions. One (often very good) way to choose that parameter is empirically/adaptively from the estimator variance.

APPENDIX

Proof of Theorem 1

We suppose that $R(x) = x^T K^{-1} x / 2$. K^{-1} is symmetric, as is $K^{-(1/2)}$. We suppose that $A \subset \mathcal{R}^d$ is closed and $\inf_{x \in A} R(x) = R(A) = r$. Let U be a random variable uniform on the surface of the d -dimensional unit radius hypersphere S_d . Denote the area of the surface as A_d . Let us define a biasing distribution as

$$q_n(x) = \frac{1}{A_d} \int_{S_d} p\left((x - \sqrt{2r} K^{(1/2)} u) \sqrt{n}\right) \sqrt{n}^d du$$

which corresponds to randomly shifting the mean value of the distribution of N to points on the hyperellipsoid $\{x : R(x) = r\}$. Now for $\theta \in \mathcal{R}^d$, define $M(\theta) = E[\exp(\langle \theta, G \rangle)] = \exp((1/2)\theta^T K \theta)$ which is the moment generating function of the G random variable. Note that $\langle \cdot, \cdot \rangle$ is our notation for the standard Euclidean inner product on \mathcal{R}^d . It is known that $R(x) = \sup_{\theta} [\langle \theta, x \rangle - \log M(\theta)] = \langle \theta_x, x \rangle - \log M(\theta_x)$, where $\theta_x = K^{-1} x$ and $M(\theta_x) = \exp(x^T K^{-1} x / 2)$. It is easy to verify that

$$p((x - \sqrt{2r} K^{1/2} u) \sqrt{n}) \sqrt{n}^d = p(x \sqrt{n}) \sqrt{n}^d \exp(n[\langle \theta_u^*, x \rangle - \log(M(\theta_u^*))])$$

where $\theta_u^* = \sqrt{2r} K^{-1/2} u$ and $\log(M(\theta_u^*)) = r u^T u$. Hence

$$\begin{aligned} q_n(x) &= \frac{1}{A_d} \int_{S_d} p((x - \sqrt{2r} K^{1/2} u) \sqrt{n}) \sqrt{n}^d du \\ &= p(x \sqrt{n}) \sqrt{n}^d \frac{1}{A_d} \int_{S_d} \exp(n[\langle \theta_u, x \rangle - \log(M(\theta_u))]) du \\ &= p(x \sqrt{n}) \sqrt{n}^d \frac{\exp(-nr)}{A_d} \int_{S_d} \exp(n[\langle \sqrt{2r} K^{-(1/2)} u, x \rangle]) du. \end{aligned}$$

Note that $\langle \sqrt{2r} K^{-(1/2)} u, x \rangle = \sqrt{2r} \langle u, K^{-(1/2)} x \rangle$. Also, due to the spherical symmetry of U , we have that $n \langle U, K^{-(1/2)} x \rangle$ is equal in distribution to $n \|K^{-(1/2)} x\| U_1$, where U_1 is one of the marginal random variables of U , i.e., $U = (U_1, U_2, \dots, U_d)$. The probability density of U_i , $f_U(u)$ is known to be [5, Th. 3.1]

$$f_U(u) = \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})\sqrt{\pi}} (1 - u^2)^{(d-1/2)-1} - 1 \leq u \leq +1.$$

Therefore

$$\begin{aligned} & \frac{1}{A_d} \int_{S_d} \exp(n[\langle \sqrt{2r} K^{-(1/2)} u, x \rangle]) du \\ &= \frac{1}{A_d} \int_{S_d} \exp(n\sqrt{2r} \langle u, K^{-(1/2)} x \rangle) du \\ &= E \left[\exp(n\sqrt{2r} \langle U, K^{-(1/2)} x \rangle) \right] \\ &= E \left[\exp(n\sqrt{2r} U_1 \|K^{-(1/2)} x\|) \right] \\ &= \int_{-1}^1 \exp(n\sqrt{2r} u \|K^{-(1/2)} x\|) f_U(u) du \\ & \quad \text{now change variables using } u = \cos(\theta) \\ &= \frac{\Gamma(\frac{d}{2})}{\Gamma(\frac{d-1}{2})\Gamma(\frac{1}{2})} \\ & \quad \times \int_0^\pi \exp(n\sqrt{2r} \|K^{-(1/2)} x\| \cos(\theta)) \sin^{d-2}(\theta) d\theta \\ &= \frac{\Gamma(\frac{d}{2}) \sqrt{\pi} \sqrt{2}^{(d/2)-1}}{\Gamma(\frac{1}{2}) \sqrt{r}^{(d/2)-1}} \frac{I_{(d/2)-1}(n\sqrt{2r} \|K^{-(1/2)} x\|)}{(n \|K^{-(1/2)} x\|)^{(d/2)-1}} \end{aligned}$$

from [7, eq. 8.431 3]

$$= b \frac{I_{(d/2)-1}(n\sqrt{2r} \|K^{-(1/2)} x\|)}{(n \|K^{-(1/2)} x\|)^{(d/2)-1}}$$

where $I_\nu(\cdot)$ is a modified Bessel function. We now have the form of the biasing distribution that we stated in the theorem

$$q_n(x) = p(x \sqrt{n}) \sqrt{n}^d \exp(-nr) b \frac{I_{(d/2)-1}(n\sqrt{2r} \|K^{-(1/2)} x\|)}{(n \|K^{-(1/2)} x\|)^{(d/2)-1}}.$$

To verify that this choice is efficient, we have to compute

$$\begin{aligned} F_{q_n} &= \int_A \left(\frac{p(y \sqrt{n}) \sqrt{n}^d}{q_n(y)} \right)^2 q_n(y) dy \\ &= \frac{\exp(2nr)}{b^2} \int_A \frac{(n \|K^{-(1/2)} x\|)^{d-2}}{I_{(d/2)-1}^2(n\sqrt{2r} \|K^{-(1/2)} x\|)} q_n(x) dx. \end{aligned}$$

Since $\inf_{x \in A} R(x) = r$, we have

$$\begin{aligned} A &\subset \left\{ x : \frac{1}{2} x^T K^{-1} x \geq r \right\} \\ &= \left\{ x : \frac{1}{2} \langle K^{-(1/2)} x, K^{-(1/2)} x \rangle \geq r \right\} \\ &= \{ x : \|K^{-(1/2)} x\|^2 \geq 2r \} \\ &= \{ x : \|K^{-(1/2)} x\| \geq \sqrt{2r} \} \\ &= C. \end{aligned}$$

Hence, since the Bessel function $I_\nu(x)$ is monotonically increasing in x , (this can be deduced from the fact that in [1, p. 375, eq. 9.6.10] is given a power series for this function with all positive coefficients)

$$\begin{aligned} F_{q_n} &\leq \frac{\exp(2nr)}{b^2} \\ & \quad \times \int_C \frac{(n \|K^{-(1/2)} x\|)^{d-2}}{I_{(d/2)-1}^2(n\sqrt{2r} \|K^{-(1/2)} x\|)} q_n(x) dx \\ &\leq \frac{\exp(2nr)}{b^2 I_{(d/2)-1}^2(n\sqrt{2r} \|K^{-(1/2)} x\|)} \\ & \quad \times \int_C (n \|K^{-(1/2)} x\|)^{d-2} q_n(x) dx. \end{aligned}$$

It is easy to verify that the integral has, at most, polynomial growth in n as n grows large. Also, for large arguments we have that the Bessel function

$$I_\nu(z) \approx \frac{\exp(z)}{\sqrt{2\pi z}}.$$

Hence, with a little algebra, we get

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log F_{q_n} \leq -2r.$$

Since the variance is greater than zero, we must have that $\liminf_{n \rightarrow \infty} (1/n) \log F_{q_n} \geq -2r$, and hence, we deduce

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log F_{q_n} = -2r.$$

Therefore, this choice of biasing distribution sequence is efficient in the small Gaussian setting, regardless of the number (one, several, countably infinite, or uncountably infinite) of minimum rate points of the set.

ACKNOWLEDGMENT

The authors would like to acknowledge the efforts of the two reviewers who substantially improved the presentation of the paper and provided a challenging intellectual discussion of the relevant issues of rare-event simulation.

REFERENCES

- [1] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*. New York: Dover, 1970.
- [2] 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 J. Select. Areas Commun.*, vol. 11, pp. 317–326, Apr. 1993.
- [3] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*. New York: Wiley-Interscience, 1990.
- [4] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*. New York: Springer-Verlag, 1985.
- [5] K.-T. Fang, S. Kotz, and K.-W. Ng, "Symmetric multivariate and related distributions," in *Monographs on Statistics and Applied Probability* no. 36. New York: Chapman & Hall, 1990.
- [6] W. Feller, *An Introduction to Probability Theory and its Applications*, 3rd ed. New York: Wiley, 1968, vol. I.
- [7] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 4th ed. New York: Academic, 1965.
- [8] P. Hahn and M. Jeruchim, "Developments in the theory and application of importance sampling," *IEEE Trans. Commun.*, vol. COM-35, pp. 706–714, July 1987.
- [9] P. Heidelberger, "Fast simulation of rare events in queuing and reliability models," *ACM Trans. Modeling Comput. Simulation*, vol. 5, no. 1, pp. 43–84, 1995.
- [10] M. C. Jeruchim, P. Balaban, and K. S. Shanmugan, *Simulation of Communications Systems: Modeling, Methodology, and Techniques*, 2nd ed. New York: Kluwer Academic/Plenum, 2000.
- [11] D. Lu and K. Yao, "Improved importance sampling techniques for efficient simulation of digital communication systems," *IEEE J. Select. Areas Commun.*, vol. 6, pp. 67–75, Jan. 1988.
- [12] B. Ripley, *Stochastic Simulation*. New York: Wiley, 1987.
- [13] J. S. Sadowsky and J. A. Bucklew, "On large deviations theory and asymptotically efficient Monte Carlo simulation," *IEEE Trans. Inform. Theory*, vol. 36, pp. 579–588, May 1990.
- [14] H.-J. Schlegelbusch, "On the asymptotic efficiency of importance sampling techniques," *IEEE Trans. Inform. Theory*, vol. 39, pp. 710–715, Mar. 1993.
- [15] P. J. Smith, M. Shafi, and H. Gao, "Quick simulation: A review of importance sampling techniques in communication systems," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 597–613, May 1997.
- [16] R. J. Wolfe, M. C. Jeruchim, and P. Hahn, "On optimum and suboptimum biasing procedures for importance sampling in communication simulation," *IEEE Trans. Commun.*, vol. 38, pp. 639–646, May 1990.

J. A. Bucklew (S'75–M'79–SM'96) received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, in 1979.

He is currently a Professor in the Departments of Electrical and Computer Engineering and Mathematics at the University of Wisconsin-Madison. His research interests are in the application of probability and statistics to signal processing and communication problems.

Dr. Bucklew has served as an Associate Editor (1990–1992) for the IEEE TRANSACTIONS ON INFORMATION THEORY and as Associate Editor (1997–1999) for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

R. Radeke is studying electrical engineering at the Dresden University of Technology, Dresden, Germany. In 2000, he received a scholarship from the German Academic Exchange Program for studying at the University of Wisconsin-Madison (2000–2001). His research interests are information technology, communications, and teletraffic theory.

He is a recipient of the Philips Preliminary Diploma Award (1999). In 2000, he was awarded the memberships for the German National Merit Foundation and the DaimlerChrysler Scholarship Program in Research and Technology.