An Experimental Investigation of Conventional and Efficient Importance Sampling

MICHEL C. JERUCHIM, FELLOW, IEEE, PETER M. HAHN, SENIOR MEMBER, IEEE, KEVIN P. SMYNTEK, MEMBER, IEEE, AND ROBERT T. RAY

Abstract—**Importance sampling is a technique that can significantly reduce computer run-time in the estimation of bit error rate (BER). However, in the conventional implementation (CIS) the improvement reduces markedly for systems with long memory. An approach to recover the full improvement for such systems has been previously suggested, and is called "efficient" importance sampling (EIS). This paper reports on an extensive series of simulation-based experiments with CIS and EIS, both to compare theoretical predictions to experimental observations, as well as to gain insight into the conditions of applicability, especially for EIS.**

I. Introduction

IMPORTANCE sampling (IS) has received attention as a promising method for reducing run-time in Monte Carlo (MC) simulation of digital transmission systems [1]–[7]. The seminal paper by Shanmugan and Balaban [1] indicated the possibility of enormous run-time improvement. However, it was pointed out in [1], and shown more generally in [4], [6], that the improvement decreases with increasing system memory M where memory is an indicator of the inverse BT product. For highly bandlimited systems, in fact, the IS improvement can become virtually nil.

In an attempt to recover the potentially high IS improvement associated with small memory, a method called "efficient" importance sampling (EIS) has been proposed [6]. (Henceforth, we will identify the original IS formulation as "conventional" importance sampling, or CIS, to distinguish it from the "efficient" version.) EIS rests on a linear approximation of the system, which is obtained empirically within the simulation itself. We will briefly review EIS in Section III.

Our purpose here is to report on a series of simulation experiments designed to develop insight into the behavior of CIS and EIS, and to compare observed results to theoretical predictions of the run-time improvement. Our primary goal has been to validate and to explore conditions of IS applicability, especially for EIS.

In spite of the theoretical activity concerning IS, there seems to be little supporting empirical verification. This is perhaps not too surprising, since this would imply a large expenditure of computer time. Indeed, the results reported here required hundreds of CPU hours. To our knowledge, the only previously reported empirical results are those in [1], and these apply to a relatively simple configuration. Our basic system context here is a bandlimited nonlinear satellite system with noise input on both uplink and downlink, which is representative of a practical situation where it might be desirable to apply importance sampling (for low BER specification).

This paper is organized as follows. First, the conventional theory is very briefly reviewed to set the stage for the experimental results, which are then presented. Next, we consider EIS. First, we outline the basic theory. Then we discuss errors associated with EIS and show some experimentally observed distributions of errors. Finally, we present the BER-related experimental results.

The experiments, for CIS and EIS, have been performed both for the basic nonlinear system mentioned above, and for a linearized version where the satellite nonlinearity is removed but the rest of the system is otherwise unchanged. This linear system is important as part of the verification procedure since we have independent, reliable means of estimating low BER performance for this case.

In the final section, we summarize the results and indicate some possible directions for additional work.

Paper approved by the Editor for Signal Design, Modulation, and Detection of the IEEE Communications Society. Manuscript received October 14, 1987; revised May 4, 1988. This paper was presented at the 1988 Conference on Information Sciences and Systems, Princeton, NJ, March 16–18, 1988.

The authors are with General Electric Aerospace, Philadelphia, PA 19101.

IEEE Log Number 8927603.

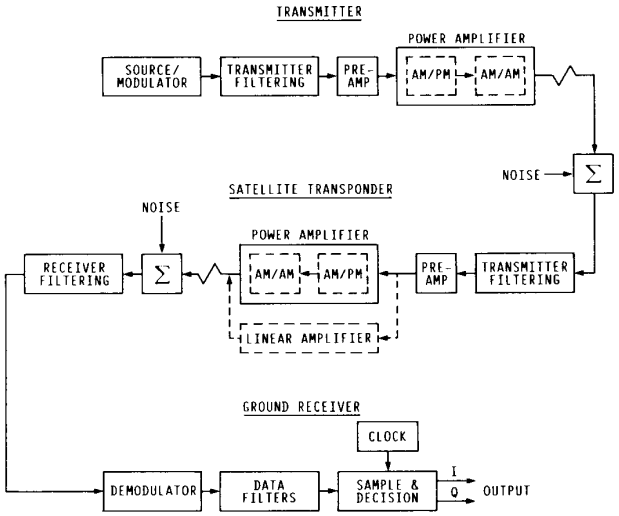


Fig. 1. Satellite system simulation block diagram for importance sampling experiments. The "linear" case is the situation where the satellite nonlinear power amplifier is replaced by the linear amplifier.

II. Conventional Importance Sampling (CIS)

The configuration we deal with is shown in Fig. 1. In the linearized version of the system, we replace the satellite nonlinearity by an ideal linear amplifier.[1] The input noise sources $n_1(t)$ and $n_2(t)$ are assumed Gaussian, with respective standard deviations σ_1 and σ_2. It is important to establish that we are dealing with a bandpass system, which is handled through the usual complex envelope technique. Hence, we use the representation

$$n_k(t) = n_{k,c}(t) \cos(\omega_0 t) - n_{k,s}(t) \sin(\omega_0 t) \quad k = 1, 2 \quad (1)$$

[1] Notice that this linearizes the system with respect to the noise. In the experiments, we did not linearize the transmitter nonlinearity, which processes only signal.

where ω_0 is the (radian) center frequency and $n_{k,c}(t)$ and $n_{k,s}(t)$ are, respectively, the in-phase and quadrature (I and Q) components. For symmetrical spectra, these components are independent, and in the simulation we enforce this independence. Furthermore, $\sigma_k^2 = \sigma_{k,c}^2 = \sigma_{k,s}^2$. We thus have four independent input (equivalent) low-pass Gaussian processes, although the in-phase and quadrature components of each source have the same power. We should reiterate that these are the *input* processes to the system (simulation), and these are under the control of the simulator. Of course, once these are sent through the system, their properties are no longer so simple. Each of the four processes induces a response in the output I and Q baseband channels. Thus, for a quadrature system there are generally eight responses. For purposes of discussion it is sufficient to consider only the output I channel since identical arguments apply to the output Q channel. Thus, for the output I channel, each of the four input processes has associated with it a memory (or dimensionality) M_{1C} and M_{1S} for the I and Q components of $n_1(t)$, and M_{2C} and M_{2S} for the I and Q components of $n_2(t)$. The dimensionality is defined to be the number of past samples, counting from the decision sampling instant, that have a measurable effect on the value of the decision voltage. Since this number depends on the simulation sampling interval, Δ, we prefer to call it dimensionality.[2]

The IS experiment is performed by "biasing" the noise sources, that is, increasing the standard deviations to σ_{1*} and σ_{2*}, respectively.[3] These new values may be thought of as characterizing the IS experiment.[4] The measurand of the experiment itself is the BER estimator $\hat{p}$, given by

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} H(v_i) w(\boldsymbol{n}_i) \tag{2}$$

where N is the number of transmitted symbols; $v_i = v(t_i)$ is the ith of the symbol-spaced decision variables observed at $t = t_i$; $H(\cdot)$ is an error counter such that $H(v_i) = 1$ if an error is made, and zero otherwise; and $w(\boldsymbol{n}_i)$ is the weighting (or unbiasing) factor at time t_i, defined as

$$\begin{aligned} w(\boldsymbol{n}_i) = & \prod_{j=0}^{M_{1C}-1} f_{N_1}(n_{1,i-j\Delta}) / f_{\tilde{N}_1}^{*}(n_{1,i-j\Delta}) \\ & \cdot \prod_{j=0}^{M_{1S}-1} f_{\tilde{N}_1}(\tilde{n}_{1,i-j\Delta}) / f_{\tilde{N}_1}^{*}(\tilde{n}_{1,i-j\Delta}) \\ & \cdot \prod_{j=0}^{M_{2C}-1} f_{N_2}(n_{2,i-j\Delta}) / f_{\tilde{N}_2}^{*}(n_{2,i-j\Delta}) \\ & \cdot \prod_{j=0}^{M_{2S}-1} f_{\tilde{N}_2}(\tilde{n}_{2,i-j\Delta}) / f_{\tilde{N}_2}^{*}(\tilde{n}_{2,i-j\Delta}) \end{aligned} \tag{3}$$

where $f_{N_1}(\cdot)$ and $f_{\tilde{N}_1}(\cdot)$ are the probability density functions, assumed normal, of the I and Q components of $n_1(t)$; the functions subscripted with 2 are the analogous quantities for $n_2(t)$; and the asterisk superscript denotes the corresponding biased densities. The sequence of noise values, e.g., $\{n_{1,i-j\Delta}\}$ for the I component of $n_1(t)$, are those actually generated during the experiment. The form of (3) implies that these values are independent,[5] and indeed, it is straightforward to ensure that the random number generator does produce independent normal variables at the succession of simulation sample times separated by Δ.

The purpose of the experiment is to investigate the mean and variance of $\hat{p}$, and in particular to compare empirical and theoretical results. The latter can be summarized as follows. First, we define the following parameters:

$$\gamma_k = \sigma_{k*}/\sigma_k, \quad k = 1, 2 \tag{4a}$$

$$R_k = \gamma_k^2/\sqrt{2\gamma_k^2 - 1}, \quad k = 1, 2 \tag{4b}$$

$$\sigma_{ke}^2 = [\gamma_k^2/(2\gamma_k^2 - 1)]\sigma_k^2, \quad k = 1, 2. \tag{4c}$$

It can then be shown [6] that the time-reliability[6] product $N\sigma^2(\hat{p})$ where $\sigma^2(\hat{p})$ is the variance of the BER estimator $\hat{p}$, is given by

$$N\sigma^2(\hat{p}) = (R_1^{M_1} \; R_2^{M_2}) P(\sigma_{1e}^2, \sigma_{2e}^2) - p^2 \tag{5}$$

where p is the true BER, $M_1 = M_{1c} + M_{1s}$, $M_2 = M_{2c} + M_{2s}$ and $P(\cdot, \cdot)$ is the true BER characteristic expressed here as a function of "equivalent" noise levels σ_{1e}, σ_{2e}. As can be seen from (4c), these reduce to the unbiased values σ_1, σ_2, when $\gamma_k = 1$. Implicit in (5) is a carrier power C_k, such that $C_k/\sigma_{ke}^2 = \rho_k$ is the equivalent CNR on each link (uplink and downlink).

It can be seen from (4) and (5) that the really fundamental parameters, insofar as the time-reliability product is concerned, are the γ_k, which we refer to as the "σ-multipliers." It may also be seen that there is an optimum value of the γ_k which minimizes $N\sigma^2(\hat{p})$. However, the two-dimensional variation that we have been dealing with would create an inordinate complication. So, to keep things manageable, we henceforth assume $\gamma_1 = \gamma_2 = \gamma$. Then, (5) simplifies to

$$N\sigma^2(\hat{p}) = R^{(M_1+M_2)} P(\rho) - p^2 \tag{6}$$

where

$$R = \gamma^2/\sqrt{2\gamma^2 - 1} \tag{7a}$$

$$\rho = [\rho_1^{-1} + \rho_2^{-1}]^{-1} \tag{7b}$$

$$\rho_k = C_k/\sigma_{ek}^2, \quad k = 1, 2 \tag{7c}$$

$$\sigma_{ke} = [\gamma^2/(2\gamma^2 - 1)]\sigma_k^2 \tag{7d}$$

and now, the BER characteristic $P(\cdot)$ has been written as a function of the composite CNR, which is the traditional way of expressing it.

A major difficulty with (5) or (6) is that $N\sigma^2(\hat{p})$ depends upon the very unknown we are trying to estimate. In the experimental investigation we do, in fact, estimate $P(\cdot)$ independently, through MC simulation. But we come up against run-time limitations for sufficiently low BER. For this reason, we also experimented with the linearized version of the system, for which we can estimate BER in just a few seconds using the quasianalytical (QA) method[7] [2]. These considerations are even more important for EIS because,

[2] If the actual system memory is, say, m symbols long, and the simulation uses K samples/symbol, then $M = mK$ is the dimensionality. Generally, the in-phase process has significantly greater dimensionality than the quadrature process (implying different effective values of m).

[3] That is, the processes remain Gaussian, with modified variances. This is the simplest and most practical approach. Attempts at optimizing the input distribution for IS purposes are emerging but were not considered in this investigation.

[4] It is understood that σ_{1*} applies to the I and Q components of $n_1(t)$, and similarly for σ_{2*}. Hence, as far as these quantities are concerned there is no need to distinguish between the I and Q components.

[5] We are speaking here of independence between elements of the same sequence. Independence between I and Q channels was discussed above, and independence between $n_1(t)$ and $n_2(t)$ is assured by physical independence.

[6] The usual measure of quality of an estimator is its variance. However, time-reliability product is even more basic, as it contains directly the trade between variance and number of observations.

[7] This is also referred to as the semianalytical method by some workers.

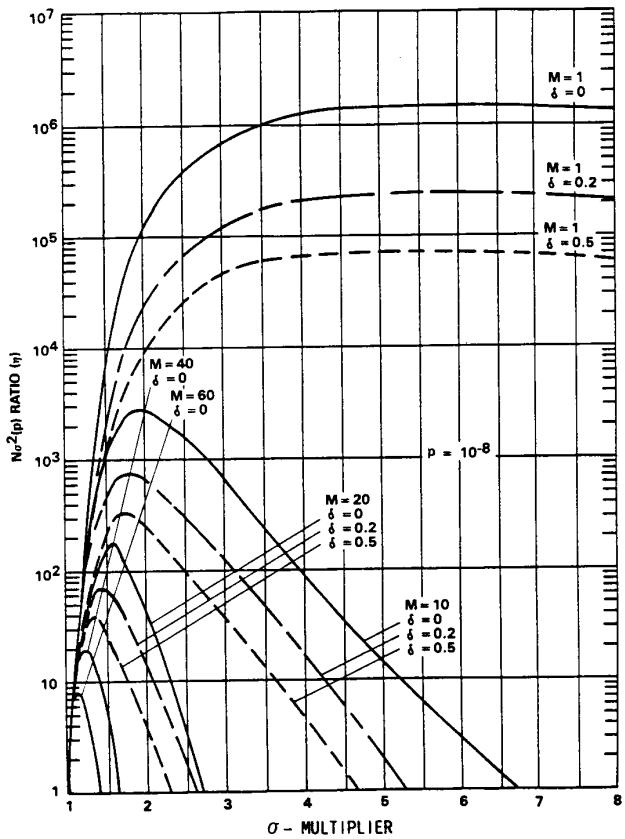


Fig. 2. Ratio of time-reliability product as a function of bias for different memories γ and fixed BER ($p = 10^{-8}$).

there, the potential improvement is much greater.

A general method of analyzing a linear system, without the necessity of dealing with the detailed characteristics of the transfer function, is given in [6] where it is shown that

$$N\sigma^2(\hat{p}) = R^{(M_1+M_2)}\left\{\frac{1}{b-a}\left[\phi^{(0)}(a)-\phi^{(0)}(b)\right] - \frac{a}{b-a}\left[\phi^{(-1)}(b)-\phi^{(-1)}(a)\right]+Q(b)\right\}-p^2 \quad (8)$$

where $a = \alpha/\sigma_e$, $b = \beta/\sigma_e$. In this model, the system is characterized by the spread of the ISI distribution, $|\beta - \alpha|$. That is, if z is the noise portion of the receiver output prior to bit decision and $p(z)$ denotes the probability of a bit error conditioned upon the value of z, then, $p(z) = 0$ for $z < \alpha$, $p(z) = 1$ for $z > \beta$ and $p(z)$ increases monotonically with z for $\alpha \le z \le \beta$. In (8), $\phi^{(0)}(x)$ is the standardized normal pdf, $\phi^{(-1)}(x)$ is the integral of $\phi^{(0)}$ from 0 to x, and $Q(x) = 0.5 - \phi^{(-1)}(x)$. Equation (8) assumes specifically that the ISI is uniformly distributed over $|\beta - \alpha|$. It is thus an approximation for other ISI distributions, but it gives us a reasonable benchmark against which to compare our observations. The utility of (8) is that it allows us to explicitly compute $N\sigma^2(\hat{p})$ as a function of γ, which is a necessity in order to have any quantitative theoretical results at all, and in order to be able to design an experiment. A typical set of results is shown in Fig. 2, which displays:

$$\eta = [N\sigma^2(\hat{p})]_{MC}/[N\sigma^2(\hat{p})]_{IS} \quad (9)$$

as a function of γ with $M = M_1 + M_2$ and $\delta = |\beta - \alpha|/|\beta + \alpha|$ as parameters. The latter is indicative of the severity of ISI, hence, the degradation. Note that the figure applies to a specific value of p. We have developed a set of such figures, for different p, to assist us in experimentation [8].

Before proceeding, we need to consider the bias[8] of $\hat{p}$ due to truncation [1], [8]. In practice, one must truncate the dimensionalities if the impulse response is infinite. Even if it were finite, it would still be desirable to reduce the dimensionalities to minimize $N\sigma^2(\hat{p})$. The tradeoff between $N\sigma^2(\hat{p})$ and bias is discussed further in Section II.

A. CIS Experiments

1) System Identification by Regression: As was noted, we need specific values for the dimensionalities in order to implement CIS. Since we often deal with IIR filters, we need a procedure to determine how to truncate the dimensionalities. This implies that we need to "identify" the system, i.e., determine the impulse response[9] for each noise source. This is not so simple because the filtering may consist of a cascade of elements, some of which may be specified by measured points. The response to quadrature noise is also not simple to obtain because it depends on the alignment of the reference I/Q axes at the receiver. Therefore, we have implemented a regression procedure at the "front-end" of the simulation. Although higher order, or nonlinear, regressions are possible, at this stage we have limited ourselves to a linear regression. That is, we assume that the following expression is a satisfactory representation for z_i, the noise portion of the output voltage at the bit-decision instant:

$$\hat{z}_i = \sum_{j=0}^{M_{1c}-1} C_{1j} n_{1,i-j\Delta} + \sum_{j=0}^{M_{1s}-1} \tilde{C}_{1j}\tilde{n}_{1,i-j\Delta} + \sum_{j=0}^{M_{2c}-1} C_{2j} n_{2,i-j\Delta} + \sum_{j=0}^{M_{2s}-1} \tilde{C}_{2j}\tilde{n}_{2,i-j\Delta} \quad (10)$$

where the $\{n_{k,i-j\Delta}, \tilde{n}_{k,i-j\Delta}\}$ are defined as in (3) and are directly observed in the simulation. To develop the coefficients in the above equation, it would be required that z_i as well as the $\{n_{k,i-j\Delta}, \tilde{n}_{k,i-j\Delta}\}$ be observable. This is not easily done. We have therefore chosen to perform an equivalent process whereby we make observations at key points in the simulation with both signal and noise present. Fig. 3 shows the important observable points in the simulation.

While it was not convenient to observe the output noise component z_i, we do observe signal-plus-noise v_i at that point. Furthermore, we observe complex samples $\{r, \tilde{r}\}$ of signal-plus-noise 1 at the same point that noise 1 enters the system. Thus, an equation similar to (10) may be written to characterize the output of the simulation as a function of these observable random inputs, simply by replacing $\hat{z}_i$ by the output $\hat{v}_i$, and replacing $(n_1, \tilde{n}_1)$ by $(r_1, \tilde{r}_1)$.

We rely on regression theory [9] to provide us with a formal procedure for estimating $\{C_{kj}, \tilde{C}_{kj}\}$. To proceed, we require a sufficiently complete set of independent observations of $\{v_i\}$ and the $\{r_1, \tilde{r}_1, n_2, \tilde{n}_2\}$ which gave rise to the $\{v_i\}$. Regression analysis requires further that VAR (v_i) is the same for all i, a

[8] Unfortunately, we use the term bias in two different ways, each correct in its context. It is used in the IS context in the sense of "biasing" a noise source, which here means increasing its variance. It is also used as a statistical term, as above, meaning the degree to which the expected value of an estimator agrees with the true population value. These meanings are sufficiently different, however, that no confusion should arise.

[9] We use the term impulse response somewhat loosely here to apply to either a linear or nonlinear system.

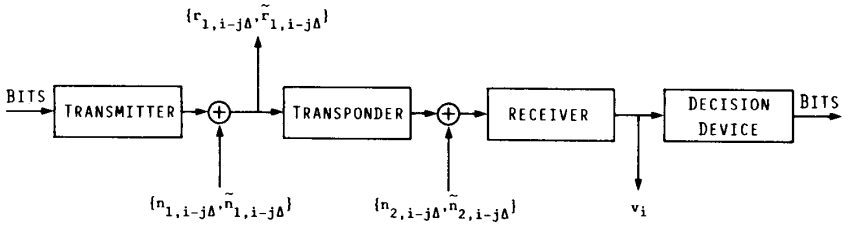


Fig. 3. Simplified sketch of system showing important quantities for the regression.

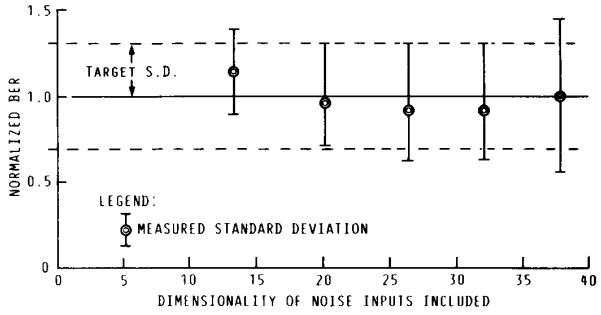


Fig. 4. Optimization of M (number of noise samples included in unbiasing calculating). CIS linear case. $\delta = 1.2$, $E_b/N_o = 14.1$ dB. Each experiment composed of 7 runs with 100 650 bits per run. The target standard deviation corresponds approximately to 95 percent confidence that BER estimate is within 2:1 of true BER.

condition that is easily satisfied. The solution is performed through available standard statistical library routines.

If the simulated system were linear and the dimensionalities sufficient to fully represent the system, then in theory only M independent observations would be required. However, unavoidable roundoff makes such perfect results unachievable. See Section III-A for further discussion.

Consequently, a larger number of observations (16 500 typically, while M is on the order of 100 or less) are used to generate regression data. It was determined experimentally that increasing the number of observations beyond this number produced no useful increase in model accuracy.

2) Experimental Results:

a) Determination of Dimensionality: The regression procedure described above was applied, and the four impulse responses for $n_{1c}(t)$, $n_{1s}(t)$, $n_{2c}(t)$, and $n_{2s}(t)$, were obtained. Since in CIS it is essential to minimize the dimensionality, we performed a short side experiment wherein we successively removed from consideration the smaller values of the impulse responses. At each value of $M = M_1 + M_2$, seven runs were made in order to provide a relationship between $E(\hat{p})$, which is the average of these seven runs, and $\sigma^2(\hat{p})$. In order to keep these runs relatively short, the target BER was set at 10^{-5} and the number of bits in each of the 7 runs was set at $N =$ 100,650 bits per (I or Q) channel. The results are shown in Fig. 4, which indicates that for M as low as 20, the BER is not measurably biased but the variance is considerably less than at the higher values of M. Consequently, $M = 20$ was chosen as a reasonable value for M. Note that this does not imply that the impulse responses are finite, but only that this many values have a significant effect on the decision voltage. M is the dimensionality *only* in the weighting procedure that yields the BER estimate. The memory of the system in the actual simulation run has not been tampered with.

b) Determination of δ and γ: In running IS, it is necessary to choose a value for γ. It is desirable to choose that

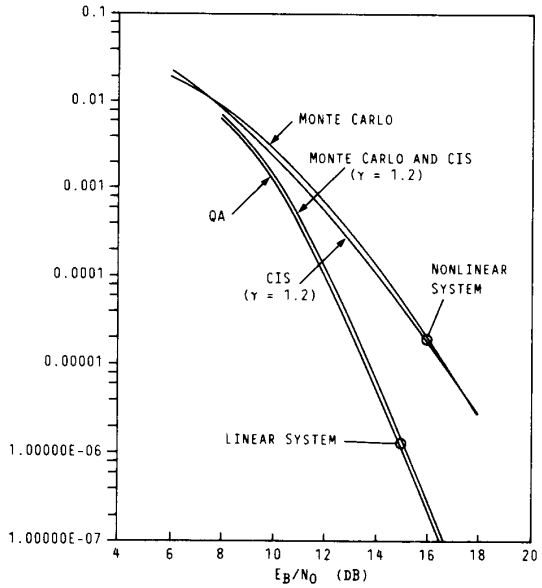


Fig. 5. CIS results—linear and nonlinear.

γ which maximizes η. Hence, we use curves like those of Fig. 2, and deduce that for $M = 20$, $\gamma = 1.2$ yields close to the optimum η. Notice that η also depends on δ. The latter, as indicative of the degradation, cannot be ascertained until after the fact although it can be estimated initially with a short run. In any case, for present purposes, it is not essential that the optimum γ was chosen, but only that we properly estimate the correct value of η for the chosen γ and δ. The value of η is used to size the length of run needed for the IS experiments. (Additional discussion on how to set up an IS simulation run is given in [5]).

c) Experimental Results: For the system of Fig. 1, runs were made both for the nonlinear and linearized system. The resulting BER curves are shown in Fig. 5. The lower set of curves applies to the linearized system, for which MC, CIS, QA estimates were made. One purpose of these runs was to calibrate the QA curve with respect to the MC result. This was necessary since we use the QA method to compute the BER in a range where it cannot easily be corroborated by other means. Fig. 5 shows that for the linearized system the MC and QA results track very closely, while the MC and CIS results were indistinguishable. For the nonlinear system only slight differences were observed between MC and CIS results.

Table I shows the runtime improvement observed on this series of CIS experiments. These compare favorably to the analytical estimates using the curves derived from (8). The values of γ used for the experiment are shown in the table.

TABLE I
RUNTIME IMPROVEMENT FOR CONVENTIONAL IMPORTANCE SAMPLING
(M = 20)

SYSTEM L = LINEAR NL = NONLIN	E_b/N_0 (dB)	BER [a]	δ	γ	RUNTIME IMPROVEMENT, η ANALYTICAL ESTIMATE	EXPERIMENTAL
L	10.0	1.4E - 3	0.2	1.20	~ 2	2.3
L	12.0	1.0E - 4	0.2	1.20	~ 2	2.0
L	14.1	1.0E - 5	0.2	1.20	~ 3	1.9
L	15.1	7.0E - 7	0.2	1.20	~ 4	2.9
NL	14.0	1.4E - 4	0.65	1.20	~ 2	2.1
NL	19.0	7.5E - 7	0.65	1.35	~ 4.5	3.65
NL	21.0	3.5E - 8	0.65	1.35	~ 15	12.7

[a] Experimental BER and Target (MC/QA). BER in such close agreement that common value was considered correct for both measurements.

III. Efficient Importance Sampling

As was seen in the previous section, even with a very diligent effort to reduce dimensionality, we still have M on the order of 20. And for such an M, the IS Improvement η is very modest, especially when compared to the maximum possible improvement shown in Fig. 2, which occurs for $M = 1$. In fact, for a linear system it is possible to reformulate IS in such a way that, theoretically, an improvement corresponding to $M = 1$ is obtained [6]. It was also postulated [6] that the same approach could be reasonably applied to "mildly" nonlinear systems. We shall se that the system we are dealing with is in fact more than "mildly" nonlinear. Nevertheless, we were able to obtain substantial improvements with respect to CIS. We shall later discuss the limitations of EIS as currently implemented and suggest some possibilities for improving the process.

If we have a linear system, with input Gaussian noises, it is clear that their combined effect could be replaced by a single equivalent output noise sample at the decision instant. Hence, if we can characterize this equivalent output noise sample, the IS problem reduces to one of unit dimensionality.

The output of a linear system at the ith decision instant t_i is given by

$$v(t_i) \triangleq v_i = s(t_i) + z(t_i). \tag{11}$$

Namely, the sum of signal plus noise, and we are specifically interested in characterizing the output noise portion $z(t_i) = z_i$. But the estimator for z_i is the expression developed earlier in (10).

We may now use the empirical estimator given by

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} H(v_i) w(\hat{z}_i) \tag{12}$$

where

$$w(\hat{z}_i) = f_n(\hat{z}_i)/f_n^*(\hat{z}_i) \tag{13}$$

and $\hat{z}_i$ is the value computed from (10). The densities $f_n(\cdot)$, $f_n^*(\cdot)$ are zero-mean normal, with variances σ^2, σ_*^2, respectively, given by

$$\sigma^2 = \sum_{k=1}^{2} \sigma_k^2 \left\{ \sum_{j=0}^{M_{k,c}-1} C_{kj}^2 + \sum_{j=0}^{M_{k,s}-1} \tilde{C}_{kj}^2 \right\} \tag{14}$$

with an identical definition for σ_*^2, but using σ_{k*} instead of σ_k.

It is clear from (12) and (13) that the IS estimate corresponds to one of unit dimensionality. Notice that we still need to keep track of the number of noise samples corresponding to the true dimensionalities in order to compute $\hat{z}_i$, but they do not affect the "effective" dimensionality of the IS process, which remains at unity. For this reason, we can afford to take account of a much longer span of the impulse response, than in CIS, and thus reduce errors due to truncation.

It can also be seen that it is virtually mandatory to estimate the impulse responses within the simulation itself, for two reasons. First, this will be considerably simpler, and probably more accurate, than to analytically determine individual filter responses (especially if they are given as measured frequency domain data), and then convolve those responses. Second, impulse responses are not invariant; they depend upon the phase alignment of the I and Q phasors at the receiver. It is therefore sensible and practical to have a simulation-based "measurement" procedure as a "front-end" to an IS run, to produce a set of estimated coefficients $\{C_{kj}\}$, $\{\tilde{C}_{kj}\}$. This we do through the regression procedure mentioned earlier. This procedure is even more important in EIS because the estimated coefficients are central to EIS, while in CIS the procedure is used only to estimate the memory. Hence, it is essential to have a sense of the variability of these estimated coefficients. This we shall review in Section III-A.

Implicit in our discussion of EIS has been the assumption of a linear system. We maintain that for a mildly nonlinear system (admittedly a fuzzy concept) the procedure described ought to be reasonably good. As is evident in Fig. 5, the nonlinear system we deal with is more than mildly nonlinear. As we show later, this means that we must limit γ to relatively small values, hence limiting ourselves to a smaller improvement.

A. Errors in EIS

There are three main sources of errors in EIS, all of which may apply to the nonlinear system, but only two of which could apply to the linearized system. The first comes from truncation of the impulse response. The second arises from imperfect estimation of the impulse response coefficients. In general, this second type is coupled to the first. We call this second type "representation" error. The third, which we call "lack-of-fit" error, stems from the degree to which the true decision-instant noise voltage pdf departs from Gaussian. The principal effect of these errors is to induce estimator bias.

It can be shown [8] that the bias error due to truncation is given by the same formulation as that for CIS. However, this error can be made much smaller in EIS because we can afford to include much more of the "tail" of an impulse response since this will not have an adverse effect on runtime improvement. Truncation error implies that while impulse

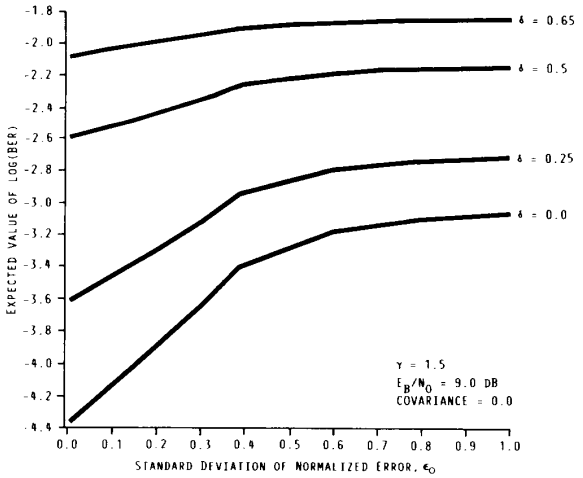


Fig. 6. Expected value of BER estimate as a function of the standard deviation of the normalized error.

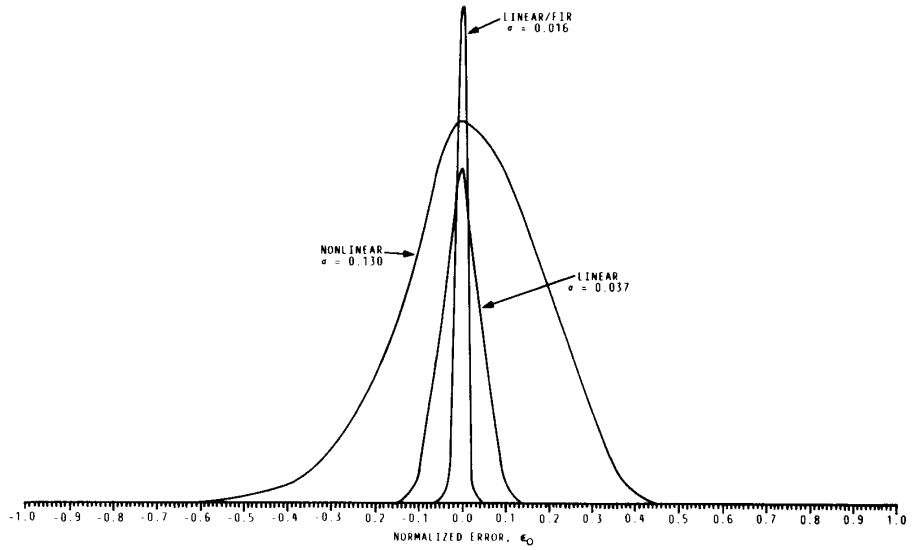


Fig. 7. Empirically obtained distributions of normalized regression model error, ϵ_0, E_b/N_0 = 9.0 (dB).

response coefficients beyond a certain duration are truncated, those that are not are known accurately. In actuality, this is not the case, as all coefficients that are retained must be estimated. In EIS we *represent* the system by the estimated coefficients. Erroneously ignoring coefficients (truncation) or misestimating them produces the same effect. There is no separately identifiable truncation error. Rather, the consequence of truncation is manifested in what we call the representation error.

The representation error is $z_i - \hat{z}_i$ where z_i is the true decision-instant noise voltage and $\hat{z}_i$ is the value obtained by applying the linear equivalent model using coefficient estimated by the regression. This error is thus attributable to the degree to which the regression coefficients are unable to account for the true output value. This may be due to roundoff, statistical factors, or, inherently, the limitations in approximating nonlinear system response to noise. The ultimate effect of representation error depends on its distribution, as well as on the nature of the system itself, i.e, the way a given error biases the BER.

Actually, since it is only the noise that is being biased and unbiased, the representation error of interest is $\epsilon_i = z_i - \hat{z}_i$, namely, the difference between the true value of noise at the bit-decision instant and the value computed via the regression. As a rough way of trying to understand the effect of this error, we have developed expressions for $E(\hat{p})$ based on the assumption that ϵ_i and z_i are jointly normally distributed. An example calculation is shown in Fig. 6 for the set of conditions indicated, namely, unbiased SNR = 9 dB, cov $(\epsilon_i, z_i) = 0$, and several values of δ [defined following (8)].

The abscissa is the standard deviation of ϵ_0, the error normalized to the mean sampled voltage. It can be seen that $E(\hat{p}) \geq p$, hence, the bias is always in a conservative direction.

It is instructive to look at the actual (empirical) distribution of ϵ_0. Fig. 7 shows three empirically obtained distributions of

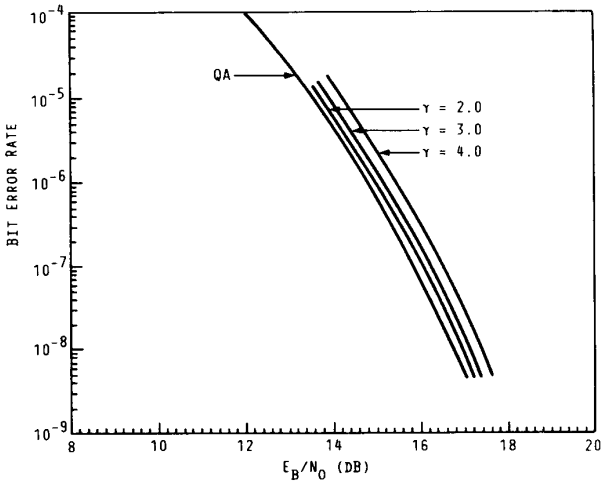


Fig. 8. Results of Efficient importance sampling experiments for linear case.

ϵ_0, using a run of 33 000 bits for each. The narrowest of these distributions applies to a linear system in which we imposed a finite impulse response. The dimension of the regression was chosen large enough not to incur truncation. Hence, in principle, one could have obtained a perfect representation of the system. The fact that we did not reflects roundoff and quantization errors in the process. While these errors are reasonably small, they are not negligible. For the linear system, the additional spread in the distribution stems from truncation of the impulse response. Still, the standard deviation is small enough that it leads to manageable estimator bias, except perhaps for large γ. Finally, the distribution for the nonlinear system is seen to be considerably more spread out, which accounts for the observed larger estimator bias than for the other two systems.

Fig. 6 may be applied to a nonlinear system to the extent that we use a value of δ that matches approximately that in the system. However, also embedded in this figure is the assumption that the output noise is Gaussian, which is generally not the case. This latter assumption leads to what we called "lack-of-fit" error in EIS, which would add to the representation error reflected in Fig. 6.

B. EIS Experiments

1) Linearized System: Fig. 8 shows BER curves for the linear system. The "true" performance curve was obtained using the QA method. This curve is identical to the corresponding curve shown in Fig. 5.

The three other curves on Fig. 8 are EIS results for $\gamma = 2$, 3, and 4, as indicated. These cases correspond to extrapolations (along the SNR axis) of 6, 9.5, and 12 dB, respectively. It can be seen that there is a bias in the BER, although even for an extrapolation as great as that corresponding to $\gamma = 4$, the bias along the horizontal axis is still a modest 0.5 dB. At $\gamma = 2$, the bias reduces to less than 0.2 dB. As discussed previously, this bias is due to representation error and is reasonably well explained by curves such as those in Fig. 4. The increasing bias with γ is a consequence of the fact that representation errors become magnified with increasing γ. As γ becomes small, the ratio of the densities that constitutes the weighting is less affected by errors in the value of the variable. In fact when $\gamma = 1$, the weighting is unaffected, regardless of the size of the error. It should also be noted that the bias is predictably to the right of the BER curve, and it may thus be possible in practice to partially compensate for it.

The goal of EIS, of course, is to reduce runtime further than is possible with CIS. The results relevant to this are shown in Table II, along with other relevant conditions. The runtime improvement η is defined by (9). The variance corresponding to the IS runs was experimentally computed from the number of runs shown in the rightmost column. The "analytical estimate" is the value of η obtained from figures such as Fig. 2, based on the model of (8), using the value $\delta = 0.20$. These analytical estimates are intended only to give an idea of the actual improvement, but are not expected to be precise predictions because the model on which they are based will not conform exactly with the system under study. Table II basically confirms the large anticipated EIS improvements for the linear case.

2) Nonlinear System: Results for the nonlinear system are shown on Fig. 9. The linear system BER curve is also shown, along with some associated EIS estimated points (from the previous figure). One reason for showing the two sets of results together is that it shows the effect of the satellite amplifier nonlinearity. Apart from this one amplifier, the systems are identical. Therefore, considerable degradation is induced by the nonlinearity, and there is some question as to whether one could regard it as "mild," a qualitative condition that, intuitively, appears to be necessary for the applicability of EIS. In fact, the nonlinearity is mild enough for EIS to yield substantial improvement with moderately small estimator bias, but strong enough to prevent the large improvements obtained with the linear system. Equivalently, we are restricted to lower values of γ, say $\gamma < 2$; higher values lead to unacceptably large estimator bias as detailed below. The reason for this bias appears to be the difficulty in faithfully representing a nonlinear characteristic with a linear equivalent, over a large operating range of the device. Recall that, given a target BER, a larger value of γ means that the IS (Monte Carlo) run takes place at a lower value of SNR, which implies that the noise sweeps a larger segment of the transfer characteristic. The reverse is true when γ is small and the target BER occurs at a moderately large SNR. Also, as γ becomes smaller the lack-of-fit error reduces.

The solid BER curve for the nonlinear system is the MC curve shown in Fig. 5. In the latter figure, we also saw that the MC curve was essentially reproduced by CIS, and as a remainder, the CIS label also appears. The last point on the curve is impractical to obtain by MC means and was therefore obtained using CIS only, with $\gamma = 1.35$. It can be seen that for $\gamma = 1.4$, EIS is only slightly biased, but for $\gamma \geq 2$, the estimator bias is too large to be generally acceptable.

TABLE II
EIS—LINEAR (BER PREDICTION ACCURACY—RUNTIME IMPROVEMENT)

SNR	γ	BER TARGET	BER MEASURED	RUNTIME IMPROVEMENT, η ANALYTICAL ESTIMATE	RUNTIME IMPROVEMENT, η MEASURED	NO. OF EXPERIMENTS
14.0 DB	2.0	0.6×10^{-5}	0.6×10^{-5}	160	233	60
14.0 DB	3.0	0.6×10^{-5}	1.1×10^{-5}	320	482	60
14.0 DB	4.0	0.6×10^{-5}	2.8×10^{-5}	350	757	60
15.0 DB	2.0	0.8×10^{-6}	1.0×10^{-6}	600	400	14
15.0 DB	3.0	0.8×10^{-6}	1.3×10^{-6}	1,500	1,490	14
15.0 DB	4.0	0.8×10^{-6}	2.7×10^{-6}	1,800	1,840	14
16.0 DB	2.0	0.8×10^{-7}	1.1×10^{-7}	2,800	819	20
16.0 DB	3.0	0.8×10^{-7}	1.9×10^{-7}	8,800	3,053	20
16.0 DB	4.0	0.8×10^{-7}	3.9×10^{-7}	11,500	4,055	24
16.8 DB	2.0	1.0×10^{-8}	1.5×10^{-8}	11,500	4,360	20
16.8 DB	3.0	1.0×10^{-8}	2.8×10^{-8}	46,000	17,030	20
16.8 DB	4.0	1.0×10^{-8}	6.0×10^{-8}	65,000	19,030	20

TABLE III
EFFICIENT IMPORTANCE SAMPLING, NONLINEAR (BER PREDICTION ACCURACY—RUNTIME IMPROVEMENT)

SNR	γ	BER TARGET	BER MEASURED	RUNTIME IMPROVEMENT, η ANALYTICAL ESTIMATE	RUNTIME IMPROVEMENT, η MEASURED	NO. OF EXPERIMENTS
16.0 dB	1.4	2.5×10^{-5}	3.6×10^{-5}	4	10	14
16.0 dB	2.0	1.5×10^{-5}	5.8×10^{-5}	46	140	14
19.0 dB	1.4	0.8×10^{-6}	1.1×10^{-6}	62	20	14
19.0 dB	2.0	0.8×10^{-6}	2.6×10^{-6}	420	133	18
21.0 dB	1.4	5.0×10^{-8}	6.4×10^{-8}	200	35	14
21.0 dB	2.0	5.0×10^{-8}	1.8×10^{-7}	2,500	67	10

Nevertheless, even for $\gamma = 1.4$, we can get very useful improvements, as can be seen from Table III, which shows the relevant experimental data.

IV. Concluding Remarks

We have performed an extensive series of experiments to look into the properties of "conventional" and "efficient" importance sampling. We verified that CIS virtually reproduces the Monte Carlo result when the dimensionality M is chosen so as to avoid estimator bias. That choice is facilitated by a built-in simulation measurement of the impulse response, which we opted to do via regression. Even with careful "trimming" M will typically be large enough to severely limit improvement. (See Table I.) We also observed CIS runtime improvements to be closely predicted by our approximate analytical model.

For very low BER, say 10^{-6} or less, much larger improvement than is generally afforded by CIS is highly desirable. This is the objective of EIS, which attempts to represent a system by a linearized equivalent, which in turn permits the system to be equated to one of unit sample memory for IS purposes. We have described the types of errors in this process, including the obvious one of approximating a nonlinearity by a linear equivalent. Nevertheless, we have seen dramatic improvement due to EIS when applied to a linear system, and still substantial improvement in the nonlinear case.

Our results, though encouraging, must still be considered of a preliminary nature in terms of our ultimate goal, which is to implement an "expert system" to assist a simulation user in the application of importance sampling in particular, and extrapolative methods in general. Regarding IS specifically, more work is needed in several areas, some of which is in progress. The system identification procedure must be refined so that the representation error is eliminated when there should be none. Higher order regressions should be considered so as to reduce representation error in the nonlinear case. Also in the latter case, work needs to be done to better understand the effect of the lack-of-fit error. Fig. 6 suggests that estimator bias may be predictable, hence, consideration should be given to finding means for automatically compensating for it. In the experiments reported, additive noise was the only external

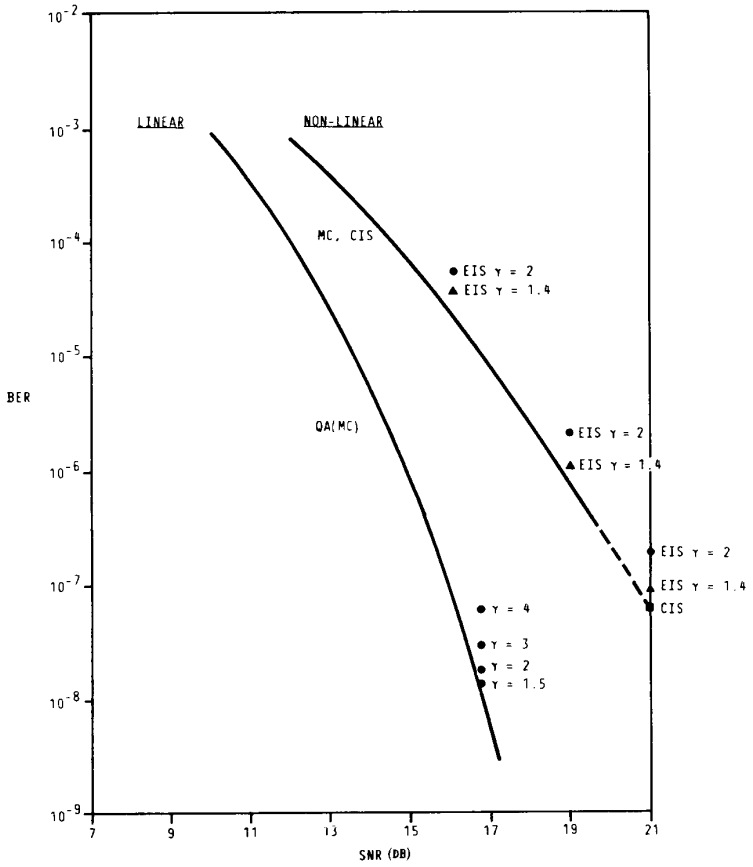


Fig. 9. Results of Efficient importance sampling experiments for nonlinear and linear cases.

source of errors. In actual digital systems, phase jitter in the carrier synchronization process and timing jitter in the clock recovery process are potentially important noise sources. There is thus a need to extend the biasing/unbiasing procedure to properly take account of these (and perhaps other) possible impairments.

References

[1] K. S. Shanmugan and P. Balaban, "A modified Monte Carlo Simulation technique for the evaluation of error rate in digital communications systems," *IEEE Trans. Commun.*, vol. COM-28, Nov. 1980.

[2] M. C. Jeruchim, "Techniques for estimating the bit error rate in the simulation of digital communications system," *IEEE J. Select. Areas Commun.*, vol. SAC-2, pp. 153-170, Jan. 1984.

[3] M. C. Jeruchim, "On the application of importance sampling to the simulation of digital satellite and multi-hop links," *IEEE Trans. Commun.*, vol. COM-32, Oct. 1984.

[4] B. R. Davis, "An improved importance sampling method for digital communication system simulations," *IEEE Trans. Commun.*, vol. COM-34, pp. 715-719, July 1986.

[5] P. M. Hahn, M. C. Jeruchim, and T. J. Klandrud, "Implementation of importance sampling in multi-hop communication simulation," in *Proc. GLOBECOM '86*, vol. 1, Houston, TX, Dec. 1-4, 1986, pp. 4.1.1-4.1.5.

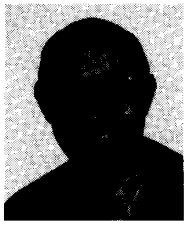
[6] ——, "Developments in the theory and application of importance sampling," *IEEE Trans. Commun.*, vol. COM-35, pp. 706-714, July 1987.

[7] D. Lu and K. Yao, "Improved importance sampling technique for efficient simulation of digital communication systems," *IEEE J. Selected Areas Commun.*, vol. SAC-6, Jan. 1988.

[8] M. C. Jeruchim *et al.*, "Technical final report for IR&D project: "Implementation of importance sampling in communication system simulation," General Electric Co. Tech. Inform. Series 875SDS020, Mar. 31, 1987.

[9] C. R. Rao, Advanced Statistical Methods in Biometrics. New York: Wiley, 1952.

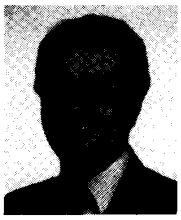
★



Michel C. Jeruchim (S'60-M'61-SM'81-F'86) was born in Paris, France, in 1937. He received the B.E.E. degree (cum laude) from the City College of New York, New York, NY, in 1961 and the M.S.E.E. and Ph.D. degrees from the University of Pennsylvania, Philadelphia, in 1963 and 1967, respectively.

Since 1961, he has been with General Electric Aerospace, King of Prussia, PA, working in a variety of communication-related disciplines, as applied to satellite communications. One of his main activities has been the development of simulation tools for the performance evaluation of digital satellite communication systems.

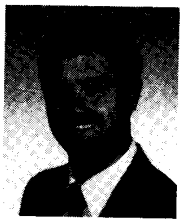
Dr. Jeruchim has published or presented more than two dozen papers on various topics in communications and is coauthor of the book *Communication Satellites in the Geostationary Orbit* (Artech, 2nd ed., 1987). He has served as a member of a number of U.S. delegations to international technical standards and radio regulatory conferences. He is currently Vice-Chairman of the Communications Society's Subcommittee on Computer-Aided Modeling, Analysis and Design of Communication Systems.



Peter M. Hahn (M'60–SM'71) was born in Vienna, Austria, in 1937. In 1958, he received the B.E.E. degree from the City College of New York, New York, NY, and the M.S. and Ph.D. degrees in electrical engineering from the University of Pennsylvania, Philadelphia, in 1962 and 1968, respectively.

He is Manager of Communication Analysis and Simulation at General Electric Aerospace, King of Prussia, PA. He is responsible for the development of computer models for performance analysis of satellite communication systems. His current interest is in the improvement of such models using statistical and artificial intelligence techniques. After graduation in 1958, he was employed by Philco Research for 3 years, attended graduate school and worked at RCA until 1967. From 1967 to 1976, he worked on communication and postal systems at Philco-Ford Corporation (later Ford Aerospace and Communications Corporation). His last position at Ford was an Engineering Section Head for packet and message switching systems. From 1976 to 1985, he worked at RCA Government Systems Division as an advanced programs manager and as a staff technical advisor. Since 1985, he has supervised communications systems analysis and simulation at General Electric. In addition, he is an Adjunct Assistant Professor at the University of Pennsylvania, and an Adjunct Professor at Drexel University, Philadelphia, PA.

Dr. Hahn has published 14 papers in IEEE Transactions or conference proceedings and has been appointed by the IEEE Educational Activities board as an ABET program accreditor. Also, he is a member of the IEEE Technology Transfer Committee.



Kevin P. Smyntek (S'83–M'85) received the B.S. in electrical engineering from the University of Rochester, Rochester, NY, in 1985 and is presently pursuing an advanced degree from Saint Joseph's University, Philadelphia, PA.

He is a Communications Engineer employed by General Electric Aerospace, King of Prussia, PA, where he is involved in the analysis and simulation of Communications Systems.

★

Robert T. Ray received the B.S. and M.S. degrees in electrical engineering from the University of Nebraska, Lincoln, in 1981 and 1983, respectively.

Since 1983, he has been with General Electric in Schenectady, NY, and King of Prussia, PA, where he has been involved in Receiver Design and Communications Link Simulation. Current interests include signal processing, tracking loops, and system modeling.