Conditional Importance Sampling Estimators
James A. Bucklew

Abstract—We give a unified presentation of the conditional importance sampling estimators. We show that they are always better than their non-conditional counterparts. We then present the large deviation theory associated with these estimators. In particular we give conditional simulation distributions that are optimal in the sense that they are efficient. Interestingly enough, these distributions will not in general be the usual exponential shifts. We give examples showing how to use the theory developed.

Keywords—Importance Sampling, Monte Carlo Simulation, Large Deviation Theory, Conditional Importance Sampling

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 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 $10^{-5}$ to $10^{-8}$. The rule of thumb for a straightforward Monte Carlo simulation is that we need $100/p$ number of simulations to reasonably estimate (20 % error with 95 % probability) an event of probability $p$. 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. Therefore, it is crucial to find efficient variance reduction techniques, such as those derived from importance sampling that lead to simulation speedup.

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 [10]. Another highly recommended review article in the field of communications systems is [22]. An encyclopedic text concerned only with the issues present in communication system simulation is [11]. Another recent very fine text dealing with rare event simulation in communication systems is [24]. There are many fine works using importance sampling to simulate rare events in communication systems. The references cited above will lead the reader to a vast collection of important results from that literature.

The main idea of the methodology is simple to present. Suppose we wish to estimate $p = E[\phi(Z)]$ where $Z$ is a random variable describing some observation on a random system. Usually $\phi$ is the indicator function of some set implying that $p$ 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 $Z^{(1)}, Z^{(2)}, \ldots, Z^{(k)}$ from the density $p(\cdot)$ and form the estimate

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

By the law of large numbers $\hat{p}_p \to p$ as $k \to \infty$. 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 $\tilde{Z}^{(1)}, \tilde{Z}^{(2)}, \ldots, \tilde{Z}^{(k)}$ 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

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

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 $\hat{p}_q$ under the density $q(\cdot)$ is precisely $p$. Therefore, the estimate $\hat{p}_q$ is unbiased and as $k \to \infty$, we also expect it to be converging (by the law of large numbers) to its mean value $p$. 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 new class of importance sampling estimators that we call conditional importance sampling estimators. This class was introduced in [23] in the context of the simulation of i.i.d. sums and called there the $g$-method. Here, we amplify the definition of these estimators and derive several new results. We follow a philosophy based upon maximizing the asymptotic rate to zero of the estimator variance. Simulation strategies that achieve this maximum rate are said to be efficient. One of our principal results is that we are able to give an efficient simulation class for the conditional importance sampling estimators.

In Section II, we present a very general form of the estimator and demonstrate that it always has smaller variance than the usual importance sampling estimator. This result is generalization of the important initial result of [23] in the setting of i.i.d. sums. In Section III, we consider the variance rate of these estimators. We present a large deviation theory analysis which explicitly defines the rate in terms of a rate function. Section IV is devoted to giving an efficient class of simulation distributions for the conditional
estimators. In section V, we present several examples using the theory developed. In section VI we discuss the results and their place in the theory of importance sampling and its relationship to rare event simulation. In Appendix A, we collect the relevant large deviation theorems that we employ in the variance rate and efficiency analysis.

II. VARIANCE REDUCTION BY CONDITIONAL IMPORTANCE SAMPLING

Suppose we are interested in

\[ \rho = E[f(Z_1, Z_2)] \]  

where \( Z_i \) is an \( \mathbb{R}^{n_i} \) valued random variable for \( i = 1, 2 \) and \( f : \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \to \mathbb{R}^d \). Denote the probability measure on \( \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \) associated with the random variables \( (Z_1, Z_2) \) as \( P \). We suppose that we wish to use importance sampling to estimate \( \rho \) and thus we use a biasing probability measure on \( \mathbb{R}^{n_1} \times \mathbb{R}^{n_2} \) which we denote as \( Q \). The usual importance sampling estimator is given by

\[ \hat{\rho}_{IS} = \frac{1}{k} \sum_{j=1}^{k} f(\tilde{Z}_1^{(j)}, \tilde{Z}_2^{(j)}) \frac{dP}{dQ}(\tilde{Z}_1^{(j)}, \tilde{Z}_2^{(j)}). \]  

Note that by the smoothing property of conditional expectation,

\[ \rho = E_P[f(Z_1, Z_2)] = E[E_P[f(Z_1, Z_2)|Z_3]] = E[g(Z_3)] \]  

where we denote \( E_P[f(Z_1, Z_2)|Z_3] \) which is a deterministic function of \( Z_3 \), as \( g(Z_3) \). It could very well be in certain situations that this conditional expectation \( g \) is known or is easily computable. Of course \( \rho \) is the just the expectation of the \( g \) function. Thus, we can use importance sampling to estimate the expectation of the \( g \) function. This leads us to the so-called conditional importance sampling estimate (we employ the subscript \( g \) in honor of the \( g \)-method of [23]),

\[ \hat{\rho}_g = \frac{1}{k} \sum_{j=1}^{k} E_P[f(Z_1^{(j)}, Z_2^{(j)})|Z_3^{(j)}] \frac{dP}{dQ}(Z_3^{(j)}), \]

where \( \frac{dP}{dQ}(Z_3) \) is just the Radon-Nikodym derivative of the marginal distribution of \( Z_3 \) under \( P \) with respect to the marginal distribution of \( Z_3 \) under \( Q \). For example it is easy to verify that

\[ Eq\left( \frac{dP}{dQ}(\tilde{Z}_1, \tilde{Z}_2)|\tilde{Z}_2 \right) = \frac{dP}{dQ}(\tilde{Z}_2). \]  

The principal result of this section is,

**Theorem II.1:**

\[ \text{Var}(\hat{\rho}_{g,i}) \leq \text{Var}(\hat{\rho}_{IS,i}) \quad i = 1, 2, \ldots, d. \]

**Proof:** The proof is given in Appendix B.

III. VARIANCE RATE OF CONDITIONAL IMPORTANCE SAMPLING ESTIMATORS

We consider a sequence of simulation problems indexed by \( n \). The idea here is that we are interested in estimating \( \rho_n = P(Z_n \in nE) \) for large \( n \) and for some set \( E \). We arrange our framework typically so that \( \rho_n \) is going to zero exponentially fast. For example, the classical setting is that we are interested in

\[ \rho_n = P\left( \frac{1}{n} \sum_{i=1}^{n} X_i > T \right), \]

where \( \{X_i\} \) are i.i.d. finite variance random variables and \( T \) is greater than the mean value. From the law of large numbers, we know that \( \rho_n \) is going to zero as \( n \) gets large. Large deviation theory tells us that this convergence is actually exponentially fast with a certain rate constant. Theorem A.1 of Appendix A shows that this exponential convergence to zero, holds in much more general settings than for just i.i.d. sums.

For the unbiased conditional importance sampling estimator of \( \rho_n \), the variance satisfies \( k \text{Var}(\hat{\rho}_n) = F_n - \rho_n^2 \). \( \rho_n \) is going to zero with some exponential rate, \( \rho_n^2 \) of course is going to zero at twice that rate. We might thus suspect that \( F_n \) is also going to zero at some exponential rate. In this section we identify the rate constant for \( F_n \), for a general sequence of simulation distributions. Unless \( F_n \) is efficient, this must be the variance rate of the conditional importance sampling estimator.

The variance rate for the usual importance sampling estimators is not hard to derive directly from the “associated” large deviation theorem [4]. This is due to the fact that \( F_{IS} \) is given by the following expectation

\[ F_{IS} = EQ\left[ \left( \frac{dP}{dQ}(\tilde{Z}_n) \right)^2 \right] \]

which is just the measure of the set \( nE \) under the measure given by \((dP/dQ)dQ\). This is not a probability measure but we can scale it so that it is and directly invoke a large deviation theorem. All this works because the square of an indicator function is the same indicator function. The current theory of large deviation theory is aimed at computing probabilities, i.e. the expectations of indicator functions. Interestingly enough, the techniques developed to date do not work for the conditional estimators. Instead of an indicator, we have a conditional expectation. The square of this conditional expectation does not give us back the original conditional expectation. This simple observation lies at the heart of the difficulty of proving a variance rate theorem for the conditional estimators.
For every integer \( n \), let \( Z_{p,n} \) be a random variable taking values in some complete separable metric space \( S_n \). Let \( P_n \) be the probability measure induced by \( Z_{p,n} \) on \( S_n \). We suppose that for each random variable \( Z_{p,n} \), we have access to some information about it contained in an information random variable \( Z_{p,i,n} \) taking values in some other probability space \( T_n \). Let \( P_{i,n} \) denote the probability measure induced by \( Z_{p,i,n} \) on \( T_n \). Let \( f_n \) be an \( \mathcal{R}^d \) valued measurable function on the space \( S_n \), i.e., \( f_n : S_n \to \mathcal{R}^d \) and we suppose we are interested in \( \rho_n = P(f_n(Z_{p,n}) \in nE) \), for some Borel set \( E \subset \mathcal{R}^d \).

We also assume that the assumptions A1, A2, and A3 (see Appendix A) hold for the sequence of convex functions \( \phi_n(\theta) = \frac{1}{n} \log E[\exp((\theta, f_n(Z_{p,n})))] \). This allows us to invoke the large deviation result, Theorem A.1. In particular, we have \( \lim_n \phi_n(\theta) = \phi(\theta) \), and \( \lim_n \frac{1}{n} \log \rho_n = -I(\theta) \).

Instead of directly simulating the information random variable \( Z_{p,i,n} \), we choose to simulate with another \( T_n \) valued random variable \( Z_{q,i,n} \) with associated probability measure on \( T_n, Q_{i,n} \). For \( z \in T_n \), let
\[
g(z) = P(f_n(Z_{p,n}) \in nE | Z_{p,i,n} = z). \tag{11}
\]
The conditional importance sampling estimator is
\[
\hat{\rho}_{g,n} = \frac{1}{k} \sum_{j=1}^{k} g(Z_{q,i,n}^{(j)}) \frac{dP_{i,n}}{dQ_{i,n}}(Z_{q,i,n}^{(j)}). \tag{12}
\]
Define,
\[
\phi_n(\theta, z) = \frac{1}{n} \log \int \exp((\theta, f_n(z_{p,n}))) dP(z_{p,n} | Z_{p,i,n} = z). \tag{13}
\]
Also let us define,
\[
c_{g,n}(\theta) = \frac{1}{n} \log \int \exp(2n\phi_n(\theta, z)) dP_{i,n}(z). \tag{14}
\]
We now define the conditional importance sampling rate function as
\[
R_g(x) = \sup_{\theta} 2(\theta, x) - c_{g}(\theta). \tag{15}
\]
For any Borel set \( E \subset \mathcal{R}^d \), define \( R_g(E) = \inf_{x \in E} R_g(x) \). We define the following assumptions,

**Assumption C1.** \( c_g(\theta) = \lim_n c_{g,n}(\theta) \) exists for all \( \theta \in \mathcal{R}^d \), where we allow \( \infty \) both as a limit value and as an element of the sequence \( \{c_{g,n}(\theta)\} \).

**Assumption C2.** The origin belongs to the interior of the domain of \( c_g \), i.e. \( 0 \in \mathcal{D}_0 \), and \( c_g \) itself is a lower semi-continuous, convex function

It turns out that in order to prove the lower bound, we need a stronger version of C1 plus some additional assumptions.

First let us define, for all \( \theta \) and \( \eta \) in \( \mathcal{R}^d \),
\[
c_{g,n}(\theta, \eta) = \int \exp(n[\phi_n(\theta + \eta, z) + \phi_n(\eta, z)]) \frac{dP_{i,n}}{dQ_{i,n}}(z) dP_{i,n}(z). \tag{16}
\]

**Assumption C1**. \( c_g^*(\theta, \eta) = \lim_n c_{g,n}(\theta, \eta) \) exists for all \( \theta, \eta \in \mathcal{R}^d \), where we allow \( \infty \) both as a limit value and as an element of the sequence \( \{c_{g,n}(\theta, \eta)\} \).

**Assumption C3.** We assume \( c_g^*(\theta, \eta) \) and \( c_g(\theta, \eta) \) are essentially smooth as functions of \( \theta \) for all \( \eta \in \mathcal{R}^d \).

There are a variety of assumptions that we can make in order to bring derivative with respect to \( \theta \) inside of the defining integrals for \( c_{g,n}(\theta) \) and \( c_g^*(\theta, \eta) \). If this were possible, then we would have
\[
\frac{\partial}{\partial \theta} c_{g,n}(\theta, \eta) \bigg|_{\theta = 0} = \frac{1}{n} \int \exp(2n\phi_n(\theta, z)) dP_{i,n}(z) \frac{dP_{i,n}}{dQ_{i,n}}(z) dP_{i,n}(z) \tag{17}
\]

Operating on the other function, we would have
\[
\frac{\partial}{\partial \theta} c_g(\theta) \bigg|_{\theta = 0} = \frac{1}{n} \int \exp(2n\phi_n(\theta, z)) dP_{i,n}(z) \frac{dP_{i,n}}{dQ_{i,n}}(z) dP_{i,n}(z) \tag{18}
\]

Remarkably (see Theorem 25.7 [18]), if convex functions converge, then so do their derivatives. Thus taking limits in the above relationships we find, \( c_g^*(0, \theta) = c_g^*(\theta)/2 \).

Hence, we capture all these arguments in the following assumption

**Assumption C4.** For all \( \psi \in \mathcal{R}^d \),
\[
c_g^*(0, \theta) = \frac{1}{2} c_g^*(\theta). \tag{19}
\]

**Theorem III.1.** Let \( E \) be any Borel set such that \( E^c \neq \emptyset \), \( E = E^c, 0 < R_g(E) < \infty \). Assume C1, C2, hold, then
\[
\limsup_{n \to \infty} \frac{1}{n} \log(F_n) \leq -R_g(E) \tag{20}
\]

If C1*, C3, and C4 hold, then
\[
\liminf_{n \to \infty} \frac{1}{n} \log(F_n) \geq -R_g(E) \tag{21}
\]

**Proof.** The proof is given in Appendix C. \[\blacksquare\]
IV. Efficient Estimators

Again, we recall that the variance satisfies $k \text{Var} (\rho_n) = F_n - \rho_n^2$. Hence, $F_n \geq \rho_n^2$ which implies that $F_n$ can go to zero no faster than $\rho_n^2$. If $F_n$ goes to zero at exactly this rate, we say that the sequence of simulation distributions that allowed this to happen are efficient. Our philosophy is that efficient sequences of simulation distributions are good simulation distributions. In general there is no unique sequence of efficient estimators. In this section, we will present a sequence of simulation distributions and show directly that it is efficient. Somewhat surprisingly, it is not in general in the class of exponential shifts.

We consider the framework as the previous section: i.e. for every integer $n$, let $Z_{p,n}$, $S_n$, $P_n$, $Z_{pi,n}$, $T_n$, $P_{i,n}$, and $f_n$ be as before. Again we are interested in $\rho_n = P(\sum_{i=1}^{n} Y_i + X_i \in E)$, for some Borel set $E \subset \mathbb{R}^d$, such that $E^c \neq \emptyset$ and $E = E^c$.

Our principal assumption here is that this set can be covered by finitely many hyperplanes; in particular for the set $E$, we suppose that there exists $m < \infty$ points $\nu_1, \nu_2, \ldots, \nu_m$ such that

1) For each $\nu_i$ there exists a unique $\theta_{\nu_i} \in \mathbb{R}^d$ such that $\nabla \phi(\nu_{\nu_i}) = \nu_i$

2) $I(\nu_i) \geq I(E)$ for each $i = 1, \ldots, m$. 

3) $E \subset \bigcup_{i=1}^{m} H(\nu_i)$, where we recall our definition of a half space $H(\nu) = \{ x : \langle \theta_{\nu}, x - \nu \rangle \geq 0 \}$.

We also assume assumptions A1, A2, and A3 (of Appendix A) hold which are precisely those needed for Theorem A1: in particular defining $\phi_n(\theta) = \frac{1}{n} \log E[\exp(\langle \theta, f_n(Z_{p,n}) \rangle)]$, we have $\lim_n \phi_n(\theta) = \phi(\theta)$ and $\lim_n \frac{1}{n} \log \rho_n = -I(E)$.

The conditional importance sampling estimator is

$$\hat{\rho}_{\theta,n} = \frac{1}{k} \sum_{j=1}^{k} g(Z_{qj,n}) \frac{dP_{i,n}}{dQ_{i,n}}(Z_{qj,n})$$

Recall that,

$$\phi_n(\theta, z) = \frac{1}{n} \log \int \exp(\langle \theta, f_n(z_{p,n}) \rangle) dP_{p,n}(z_{p,n} = z).$$

Theorem IV.1: With the above assumptions, the sequence of probability distributions given by

$$dQ_{i,n}(z) = \left( \sum_{i=1}^{m} p_i \exp(\phi_n(\theta_{\nu_i}, z)) \exp(-n\phi_n(\theta_{\nu_i})) \right) dP_{i,n}(z).$$

is efficient, where $(p_1, p_2, \ldots, p_m)$ is a probability vector with strictly positive components.

Proof: The proof is given in Appendix D.

Remark IV.1: Note that this is a valid probability distribution due to (82).

Remark IV.2: We note that we now have a “free” probability vector $(p_1, \ldots, p_m)$ to choose. Parts of the set $E$ that have higher probability should be covered with hyperplanes associated with larger values of $p$. The best way to choose this vector remains an open research problem.

Remark IV.3: Consider the i.i.d. sum of $\mathbb{R}^d$ valued random variables setting. Let $E$ be a set as in the theorem statement with $m$ important points $\nu_1, \nu_2, \ldots, \nu_m$. Suppose we are interested in

$$\rho_n = P(\sum_{i=1}^{n} Y_i + X_i \in nE)$$

where $\{X_i\}$ is an i.i.d. sequence of $\mathbb{R}^d$ valued random variables independent of $\{Y_i\}$ another i.i.d. sequence of $\mathbb{R}^d$ valued random variables. We take

$$Z_{p,n} = (X_1, Y_1, X_2, Y_2, \ldots, X_n, Y_n)$$

$$Z_{pi,n} = (X_1, X_2, \ldots, X_n)$$

$$f(Z_{p,n}) = \sum_{i=1}^{n} Y_i + X_i.$$ 

Thus we can compute

$$\exp(n\phi_n(\theta, z)) = \int \exp(\langle \theta, f_n(z_{p,n}) \rangle) dP_{p,n}(z_{p,n} = z)$$

$$= \int \exp(\langle \theta, \sum_{i=1}^{n} (x_i + y_i) \rangle)$$

$$\times dP(x_1, y_1, x_2, y_2, \ldots, x_n, y_n | X_1 = z_1, X_2 = z_2, \ldots, X_n = z_n)$$

$$= E[\exp(\langle \theta, \sum_{i=1}^{n} z_i \rangle)]$$

$$= \exp(\langle \theta, \sum_{i=1}^{n} z_i \rangle) M_n(\theta)^n$$

Thus we have

$$\exp(n\phi_n(\theta)) = E[\exp(\langle \theta, f_n(Z_{p,n}) \rangle)] = M_n(\theta)^n M_{\theta}(\theta)^n.$$

Hence, our efficient simulation distribution becomes,

$$dQ_{i,n}(z) = \left( \sum_{i=1}^{m} p_i \exp(\langle \theta_{\nu_i}, \sum_{j=1}^{n} z_j \rangle) M_\theta(\theta_{\nu_i})^{-n} \right) dP_{i,n}(z).$$

Since, the $\{X_i\}$ are independent, this means that $P_{i,n}$ is a product measure which implies from the above equation that $Q_{i,n}$ is a mixture of product measures.

$$dQ_{i,n}(z = (z_1, z_2, \ldots, z_n))$$

$$= \sum_{i=1}^{m} p_i \left( \prod_{j=1}^{n} \exp(\langle \theta_{\nu_i}, z_j \rangle) M_\theta(\theta_{\nu_i})^{-1} dP_{i,1}(z_j) \right)$$

This is exactly the efficient exponential shift of the $\{X_i\}$ random variables that we would use if we were interested in simulating

$$\rho_n = P(\sum_{i=1}^{n} X_i \in nE).$$

This result is worked out in detail in [19]. Note that counterintuitively, the $\{Y_i\}$ don’t “help” or have any influence.
in the optimal biasing of the \{X_i\} random variables. This
starting property of the efficient conditional estimator is
noted in the scalar i.i.d. sum case (with \(m = 1\)) in [23],
[24].

Example IV.1: Suppose that
\[ Z_{p,n} = \sum_{i=1}^{n} Y_i + X_i \]
where the sequence \(\{Y_i\}\) is an i.i.d. real valued standard normal
sequence and is independent of \(\{X_i\}\) which is i.i.d. exponen-
tial with parameter one. We suppose that we are interested in
\[ \rho_n = P(Z_{p,n} > nT), \]
where \(T > 1\). The large deviation rate of the probability
is simple to compute. The rate function is given by
\[ I(T) = \sup_{\theta} \left[ \theta T - \log M_Y(\theta)M_X(\theta) \right] \]
\[ = \sup_{\theta} \left[ \theta T - \log \left( \exp \left( \frac{\theta^2}{2} \right) \right) \right] \]
\[ = \theta T - \frac{\theta^2}{2} + \log(1 - \theta T) \]
where \(\theta_T = (T + 1 - \sqrt{T^2 - 2T + 1})/2\). We decide to
use a conditional importance sampling procedure with the choice
of the the information random variables as \(Z_{i,n} = (X_1, X_2, \ldots, X_n)\). Note that
\[ P(Z_{p,n} > nT|Z_{i,n}) = \Phi_c(\sqrt{n}T - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i) \]
(34)
where \(\Phi_c(x) = P(Y_1 > x)\) is the complement of the stan-
> 

ard normal cumulative distribution function. Conditioned
on knowledge of \(Z_{i,n}, Z_{p,n}\) is normal with mean \(\sum_{i=1}^{n} X_i\)
and variance \(n\). Thus its moment generating function is
\[ \exp(n\phi_n(\theta, Z_{p,n})) = \exp(\theta \sum_{i=1}^{n} X_i + \frac{\theta^2 n}{2}). \]
(35)

Therefore, an efficient choice for biasing the information
random variables is
\[
dQ_{i,n}(x) = \exp(n\phi_n(\theta_T, x)) \exp(-n\phi_n(\theta_T))dP_{i,n}(x)
\]
\[ = \exp(\theta \sum_{i=1}^{n} x_i + \frac{\theta^2 n}{2}) \exp(-n\phi_n(\theta_T))dP_{i,n}(x)
\]
\[ = \exp(\theta \sum_{i=1}^{n} x_i + \frac{\theta^2 n}{2}) \exp(-n\phi_n(\theta_T))
\]
\[ \times \prod_{i=1}^{n} \exp(-x_i) \]
\[ = C \prod_{i=1}^{n} \exp(-(1 - \theta_T)x_i) \]
(36)

where \(C\) denotes the necessary constant so that the above
product of exponentials is a valid probability density. Per-
foming that simple computation reveals that \(C = (1 - \theta_T)^n\).
Thus an efficient choice is to bias the information
random variables to be i.i.d. exponential with parameter
\(1 - \theta_T\).

Example IV.2: A simplified model of a optical wave-
length division multiplexing (WDM) network is given in
[17]. The reader should consult this reference for an un-
derstanding of how the model comes about. We will just
concentrate on the simulation problem in this example.

The photocurrent generated by a photodiode is approxi-
minated by
\[ i_d = \frac{a_1 E^2}{2} + \sum_{m=2}^{M} \sqrt{\epsilon} a_1 E^2 \cos(\phi_m)
\]
\[ + \sum_{m,n=2, m > n}^{M} \epsilon E^2 \cos(\phi_m - \phi_n) + (M - 1) \epsilon \frac{E^2}{2} + n_G \]
(37)

where \(a_1 \in \{0, 1\}\) is the information bit, \(E\) is the pulse
amplitude, and \(n_G\) is the Gaussian zero mean, variance \(\sigma_G^2\), re-
ceiver thermal noise which is independent of the signal and
the crosstalk components. The “worst case” crosstalk ef-
fects of the \(M - 1\) interfering channels is represented by the
terms involving \(M\). The parameter \(\epsilon\) controls the crosstalk
intensity. The \(\{\phi_m\}\) are modeled as independent uniform
\([0, 2\pi]\) phase angles.

We are interested in the probability of error when \(a_1 = 1\).
To proceed with our analysis, we assume that the third and
fourth terms in the equation for \(i_d\) can be neglected (small
\(\epsilon\)). Thus, we have
\[ i_d = \frac{a_1 E^2}{2} + \sum_{m=2}^{M} \sqrt{\epsilon} a_1 E^2 \cos(\phi_m) + n_G \]
(38)

When the decision threshold \(\tau\) is set at one half the ON-
signal output current (symmetric setting, i.e., \(\tau = E^2\)) and
under hypothesis that \(a_1 = 1\), then
\[
\rho = P(\text{Error}|a_1 = 1)
\]
\[ = P(i_d < \tau | a_1 = 1) \]
\[ = P \left( \frac{a_1 E^2}{2} + \sum_{m=2}^{M} \sqrt{\epsilon} a_1 E^2 \cos(\phi_m) + n_G < \frac{E^2}{4} \right| a_1 = 1 \)
\[ = P \left( \sum_{m=2}^{M} \sqrt{\epsilon} E^2 \cos(\phi_m) + n_G < \frac{E^2}{4} \right) \]
\[ = P \left( \sum_{m=2}^{M} (-\cos(\phi_m)) + \frac{(-n_G)}{\sqrt{\epsilon} E^2} > \frac{1}{4 \sqrt{\epsilon}} \right) \]
\[ = P \left( \sum_{m=2}^{M} \cos(\phi_m) + n_G' > \frac{1}{4 \sqrt{\epsilon}} \right) \]
(39)

where \(\{\phi_m\}\) are uniformly distributed on \([0, 2\pi]\) and \(n_G' \sim
N(0, \frac{\sigma_G^2}{\sqrt{\epsilon}})\).

We need to embed this probability into a sequence of prob-
abilities that satisfy a large deviation principle. Con-
Consider the sequence of probabilities

\[ \gamma_N = P(\frac{1}{N} \sum_{m=1}^{N} (\cos(\phi_m) + n_{G''_m}) > T). \]  

(40)

Let us take \( n_{G''_m} \) to be i.i.d. Gaussian zero mean, variance \( \sigma^2_G/(e^4(M-1)) \), and \( T = 1/(4\sqrt{\gamma}(M-1)) \) This is a large deviation sequence in \( N \) and furthermore when \( N = M - 1 \), we match up with our original problem, i.e. \( \gamma_{M-1} = \rho \). The large deviations of \( \gamma_N \) are easily worked out. It is an i.i.d. sum, so we could use Cramer’s theorem [3] or (with a bit more overkill) Theorem A.1 of Appendix A.

To use Theorem A.1, we must first compute

\[ \phi_N(\theta) = \frac{1}{N} \log E[e^{\theta \sum_{m=1}^{N} (\cos(\phi_m + n_{G''_m})} \]  

(41)

where \( M_e(\theta) \) and \( M_g(\theta) \) are the moment generating functions for the cosines and the Gaussians respectively and are obtained as

\[ M_e(\theta) = \int_0^{2\pi} \frac{1}{2\pi} e^{\theta \cos x} dx = I_0(\theta) \]  

\[ M_g(\theta) = \int_{-\infty}^{\infty} e^{\theta y} \frac{1}{\sqrt{2\pi} \sigma^2} e^{-\frac{y^2}{2\sigma^2}} dy = e^{\frac{\theta^2}{2}\sigma^2} \]  

(42)

\( I_0(\theta) \) is modified Bessel function of the first kind.

Thus,

\[ \lim_{N \to \infty} \phi_N(\theta) = \phi(\theta) = \log(M_e(\theta)M_g(\theta)) \]  

(43)

The large deviation rate of the probability can be computed as

\[ I(T) = \sup_{\theta} [\theta T - \phi(\theta)] \]  

\[ = \sup_{\theta} [\theta T - \log M_e(\theta)M_g(\theta)] \]  

\[ = \sup_{\theta} [\theta T - \log (I_0(\theta)e^{\frac{\theta^2}{2}\sigma^2})] \]  

\[ = \theta_T T - \log I_0(\theta_T) - \frac{1}{2}\theta_T^2 \sigma^2 \]  

(44)

To determine \( \theta_T \), we set

\[ \frac{\partial}{\partial \theta} [\theta T - \log (I_0(\theta)e^{\frac{\theta^2}{2}\sigma^2})] = 0 \]  

\[ \Rightarrow T - \frac{I_0'(\theta)}{I_0(\theta)} - \theta \sigma^2 = 0 \]  

(45)

Noting that

\[ I_n'(x) = \frac{n}{x} I_n(x) + I_{n+1}(x) \]  

(46)

we have

\[ T - \frac{I_1(\theta)}{I_0(\theta)} - \theta \sigma^2 \bigg|_{\theta = \theta_T} = 0 \]  

(47)

This equation can be solved to obtain \( \theta_T \)

To find the efficient conditional importance sampling distributions, we first compute

\[ \exp(N \phi_N(\theta, z)) \]  

\[ = E_{p,n} \left[ e^{\theta (z_{p,n} - z)} \right] \mid z \]  

\[ = E \left[ e^{\theta \sum_{i=1}^{N} (\cos(\phi_i + n_{x_i}))} \mid \phi_i = z_i, i = 1, 2, \ldots, N \right] \]  

\[ = \exp(\theta \sum_{i=1}^{N} \cos z_i + N \sigma^2)^2 \]  

(48)

Hence, an asymptotically efficient biasing p.d.f. for the information random variables \( Z_{p,n} = \{ z_i \} \) is

\[ q_{i,n}(x) \]  

\[ = e^{N \phi_N(\theta_T, z_{p,n})} e^{-N \phi_N(\theta_T)} \cdot 1_{[x_i \in [0,2\pi]]} \]  

\[ = e^{\theta_T \sum_{i=1}^{N} \cos x_i + \frac{1}{2}\sigma^2} \cdot e^{-\frac{1}{2}\sigma^2} \]  

\[ = e^{\theta_T \cos x_i} \cdot \frac{1}{2\pi I_0(\theta_T)} \]  

(49)

Since the \( \{ z_i \} \) are i.i.d. random variables, the biasing p.d.f.’s are also the same and thus, the p.d.f. for a single biased random variable \( \phi_i \) can be written as

\[ q_{\phi_i}(x) = \frac{e^{\theta_T \cos x}}{2\pi I_0(\theta_T)} \cdot 1_{x_i \in [0,2\pi]} \]  

(50)

Hence, the conditional importance sampling estimator of \( \rho \) is

\[ \hat{\rho}(k) \]  

\[ = \frac{1}{k} \sum_{j=1}^{k} \left( N_T - \sum_{i=1}^{N} \cos x_i(j) \right) \]  

\[ \cdot \frac{1}{\sqrt{2\pi} \sigma^2} \]  

(51)

To simulate, we choose the following system parameters:

\[ M = 4 \]  

(52)

\[ N = M - 1 = 3 \]  

(53)

\[ E^4 = 1024 \]  

(54)

In the simulation, the crosstalk-to-signal ratio \( XSR \) varies from -55 dB to -15 dB in increments of 1 dB.

For a particular \( XSR \), we obtain

\[ \epsilon = 10^{XSR/10} \]  

(55)
where \[ \sigma^2 = \frac{1}{cN} \frac{E}{\sigma^2} \] \[ T = \frac{1}{4\sqrt{cN}} \] (56) (57)

Once these values are obtained, we compute \( \theta_T \) above using the MATLAB symbolic toolbox. We then implemented the conditional estimator in MATLAB.

A plot of \( \hat{\rho}(k) \) vs. \( k \) for XSR = -50 dB is shown in Figure 1. A plot of \( \hat{\rho} \) vs. \( XSR \) is shown in Figure 2.

V. DISCUSSION

In previous works [2], [3], [4], [5], [10], [12], [13], [14], [15], [20], [19], [21] many authors have expounded and amplified the large deviation theory philosophy of how to intelligently choose importance sampling simulation distributions for rare event simulation problems. These researchers argue that for highly reliable systems, 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, 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 can’t be done.

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. This paper is about significantly expanding this large deviation framework to now include the conditional importance sampling estimators which always perform better than conventional ones.

APPENDIX

I. APPENDIX A

In this appendix, we collect some of the principal theorems of large deviation theory that we use above.

Suppose that we have some infinite sequence of \( R^d \)-valued random variables \{\( Y_n \)\}. Define for \( \theta \in R^d \),
\[
\phi_n(\theta) = \frac{1}{n} \log E[\exp((\theta, Y_n))].
\] (58)

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

Assumption A2. The origin belongs to the interior of the domain of \( \phi \), i.e. \( 0 \in \tilde{D}_\phi \), and \( \phi \) itself is a lower semi-continuous, convex function.

A convex function \( \phi \) is essentially smooth if three conditions hold, a) the set \( \tilde{D}_\phi \) is nonempty, b) \( \phi \) is differentiable everywhere in \( \tilde{D}_\phi \), and c) \( \phi \) is steep.

Assumption A3. \( \phi \) is essentially smooth.

Now we may define
\[
I(x) = \sup_{\theta} \{[\theta, x] - \phi(\theta)\}.
\] (59)

For any measurable set \( E \), define inf\(_{x \in E} \) \( I(x) = I(E) \). The following theorem may be found in [6],

Theorem A.1 (Ellis) Assume A1 and A2. For every closed subset \( F \subset R^d \),
\[
\limsup_{n \to \infty} \frac{1}{n} \log P\left( Y_n \in F \right) \leq -I(F).
\] (60)

Assume A1 and A3. For every open subset \( G \subset R \)
\[
\liminf_{n \to \infty} \frac{1}{n} \log P\left( Y_n \in G \right) \geq -I(G).
\] (61)

Lemma A.1: Let \( F_n \) denote the distribution function of \( Y_n \) and define the distribution function of the random variable \( Y_n(\theta) \) as
\[
dF_n(\theta)(x) = \frac{dF_n(x) \exp((\theta, x))}{\int \exp((\theta, x))dF_n(x)} = \frac{dF_n(x) \exp((\theta, x))}{\exp(n\phi_n(\theta))}.
\] (62)

Let \( B_\delta(v) = \{x : ||x - v|| < \delta\} \). Suppose \( v \in \nabla \phi(D_\phi) \), i.e. \( \nabla \phi(\theta) = v \) has a solution \( \theta_v \). Then
\[
\lim_{n \to \infty} P\left( \frac{Y_n(\theta_v)}{n} \in B_\delta(v) \right) = 0.
\] (63)

Given a nonzero point \( \theta \in R^d \) and a real number \( \alpha \), define the closed half space \( H_+ (\theta, \alpha) = \{z \in R^d : (\theta, z) - c(\theta) \geq \alpha \} \). The following theorem is due to Ellis [7].

Theorem A.2: Assume A1 and A2. Let \( K \) be a nonempty closed set in \( R^d \). Then if \( 0 < I(K) < \infty \), then for any positive \( \epsilon \), there exists finitely many nonzero points \( \theta_1, \ldots, \theta_k \) such that \( K \subset \cup_{i=1}^k H_+ (\theta_i, I(K) - \epsilon) \).

II. APPENDIX B

Proof: [Proof of Theorem II.1:] For simplicity, we just take \( d = 1 \), otherwise without loss of generality, we can just consider the 1th component of the estimator in isolation. As always \( k \text{Var}(\rho_\theta) = F_n - \rho^2 \) and of course \( k \text{Var}(\rho_{15}) = F_{15} - \rho^2 \), where \( F_n \) and \( F_{15} \) are the expected square value of the summands in the respective estimators. Thus we have
\[
F_n = \int g(z_2)^2 \left( \frac{dP}{dQ}(z_2) \right)^2 dQ(z_2)
\]
\[
= \int g(z_2)^2 \frac{dP}{dQ}(z_2) dP(z_2).
\] (64)
Consider the first term in the integrand above,
\[ g(z) = \int f(z_1, z_2) dP(z_1|z_2) \]
\[ = \int f(z_1, z_2) dP(z_1|z_2) dQ(z_1|z_2) dP(z_2) \]
\[ \leq \int f^2(z_1, z_2) dQ(z_1|z_2) dP(z_1, z_2) dP(z_2) \]
\[ = \int f^2(z_1, z_2) dQ(z_1|z_2) dP(z_1, z_2) \]
\[ = \frac{dP_{i,n}}{dQ_{i,n}}(z) dP_{i,n}(z) \]

Hence
\[ F_n = \int g(z)^2 dP dQ(z_2) \]
\[ \leq \int f^2(z_1, z_2) dQ(z_1|z_2) dP(z_1, z_2) \]
\[ = \int f^2(z_1, z_2) dQ(z_1|z_2) dP(z_1, z_2) \]
\[ = \int f^2(z_1, z_2) dP(z_1|z_2) dP(z_1, z_2) \]
\[ = \frac{m^2}{m} \left( \frac{1}{m} \sum_{i=1}^{m} \exp(2\phi_n(\theta_{vi}, z)) \exp(-n\phi_n(\theta_{vi}, z)) \right) \]
\[ = \frac{m^2}{m} \left( \frac{1}{m} \sum_{i=1}^{m} \exp(2\phi_n(\theta_{vi}, z)) \exp(-n\phi_n(\theta_{vi}, z)) \right) \]
\[ = \frac{m}{m} \exp(-n\phi_n(\theta_{vi}, z)) \exp(nc_{i,n}(\phi_n(\theta_{vi}))) \]

This completes the proof of the theorem.

III. APPENDIX C

Proof: [Proof of Theorem III.1] We first prove the upper bound. Due to Assumptions C1 and C2, from the covering theorem A.2, we have that there exists (for each $\varepsilon > 0$) $m < \infty$ points $\nu_1, \nu_2, \ldots, \nu_m$ such that
1) For each $\nu_i$, there exists a unique $\theta_{vi} \in \mathcal{R}^d$ such that $\nabla c_{\beta}(\theta_{vi}) = \nu_i$.
2) $R_\beta(\nu_i) \geq R_\beta(E) - \varepsilon$ for each $i = 1, \ldots, m$.
3) $E \subset \bigcup_{i=1}^{m} \mathcal{H}(\nu_i)$

where we recall our definition of a half space $\mathcal{H}(\nu) = \{ x : \langle \theta_{vi}, x - \nu \rangle \geq 0 \}$

Also note that
\[ g(z) = P(f_n(Z_{pi,n}) \in nE|Z_{pi,n} = z) \]
\[ \leq \int \left( \frac{1}{n} \sum_{i=1}^{m} \exp(n\langle \theta_{vi}, y_{ni} \rangle) \right) dP(z_{pi,n} | Z_{pi,n} = z) \]
\[ \leq \sum_{i=1}^{m} \exp(n\langle \theta_{vi}, y_{ni} \rangle) dP(z_{pi,n} | Z_{pi,n} = z) \]

We now use this result to lower bound $F_n$ as follows,
\[ F_n = \int g(z)^2 dQ_{i,n}(z) \]
\[ \leq \int \left( \frac{1}{m} \sum_{i=1}^{m} \exp(n\phi_n(\theta_{vi}, z)) \exp(-n\phi_n(\theta_{vi}, z)) \right)^2 \]
\[ \leq \sum_{i=1}^{m} \exp(n\phi_n(\theta_{vi}, z)) \exp(-n\phi_n(\theta_{vi}, z)) \]

Since $\varepsilon > 0$ is arbitrary, the upper bound part of the theorem is proved.

We now prove the lower bound. Let $\psi \in E$. Fix $\varepsilon > 0$.

Define
\[ G_n(\psi) = \{ z' : \left\| \frac{f_n(z')}{n} - \psi \right\| < \varepsilon \}. \]

Therefore for $z' \in G_n$,
\[ \left| \langle \theta_{vi}, \frac{f_n(z')}{n} - \psi \rangle \right| \leq \left\| \theta_{vi} \right\| \varepsilon \]
\[ \left| \langle \theta_{vi}, f_n(z') - \psi \rangle - n\psi \right| \leq n\left\| \theta_{vi} \right\| \varepsilon. \]

Hence, we may lower bound the conditional probability $g(z)$ as follows
\[ g(z) = P(f_n(Z_{pi,n}) \in nE|Z_{pi,n} = z) \]
\[ \geq P(f_n(Z_{pi,n}) \in G_n|Z_{pi,n} = z) \]
\[ \geq \exp(-n\left\| \theta_{vi} \right\| \varepsilon) \]
\[ \times \int_{G_n} \exp(\langle \theta_{vi}, f_n(z') - \psi \rangle) dP(z_{pi,n} | Z_{pi,n} = z). \]
Together, these measures define a sequence of probability measures for the $Z_{p_i,n}$ random variable, which we will denote as $N_{Q,n}$. Let us consider the moment generating function sequence,

$$E_{N_{Q,n}}[\exp(\langle \theta, f_n(Z_p) \rangle)] = \int \int \exp(\langle \theta, f_n(z) \rangle) \exp(-n\phi_n(\theta, z)) \cdot dP(z_{p,n}|Z_{p_i,n} = z) \frac{dP_{i,n}(z)}{dQ_{i,n}(z)} \exp(-nc_{g,n}(\theta, z)) \cdot dP_{i,n}(z) = \exp(-nc_{g,n}(\theta, z)) \cdot \int \exp(n[\phi_n(\theta + \psi, z) + \phi_n(\theta, z)]) \frac{dP_{i,n}(z)}{dQ_{i,n}(z)} \exp(nc_{g,n}^*(\theta, \psi)) \cdot dP_{i,n}(z)$$

(76)

Hence,

$$\lim_{n \to \infty} \frac{1}{n} \log E_{N_{Q,n}}[\exp(\langle \theta, f_n(Z_p) \rangle)] = c_g^*(\theta, \theta) - c_g(\theta, \theta).$$

(77)

Due to Assumption C4, we have

$$\frac{\partial}{\partial \theta} c_g^*(\theta, \theta) = c_g(\theta, \theta) \bigg|_{\theta = 0} = \frac{\partial}{\partial \theta} c_g^*(0, \theta) \bigg|_{\theta = 0} = c_g^*(0, \theta) = \frac{1}{2} c_g'(\theta) = \frac{1}{2} [\theta^2] = \psi.$$  

(78)

Thus, from Theorem A.1 we have that

$$N_{Q,n}(G_n) = \int \exp(n\phi_n(\theta, z)) \exp(\langle \theta, f_n(z) \rangle) \cdot dP(z_{p,n}|Z_{p_i,n} = z) \frac{dP_{i,n}(z)}{dQ_{i,n}(z)} \exp(-nc_{g,n}(\theta, z)) \to_{n \to \infty} 0.$$  

(79)

In fact from the proof of Theorem A.1, we know that this convergence is exponentially fast with a strictly positive rate constant.

Thus from (73), we have

$$\lim_{n \to \infty} \frac{1}{n} \log F_n \geq -2\|\theta_p\|_\epsilon - 2(\theta_p, \psi) + c_g(\theta_p) = -2\|\theta_p\|_\epsilon - R_g(\psi).$$  

(80)

Since, $\epsilon > 0$ is arbitrary, and $\psi$ can be varied over $E$, we have that

$$\lim_{n \to \infty} \frac{1}{n} \log F_n \geq -R_g(E).$$  

(81)
IV. APPENDIX D

Proof: [Proof of Theorem IV.1.] Note that
\[ E_{P_{
u,n}}[\exp(n\phi_n(\theta, Z_{\nu,i}, n))] = E[\exp(\theta, f_n(Z_{\nu,i}))] = \exp(n\phi_n(\theta)). \] (82)

Reproducing the arguments leading to (67) for this new set of points \( \nu_1, \nu_2, \ldots, \nu_m \) gives us again
\[ g(z) \leq \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \exp(-n(\theta_{\nu_i}, \nu_i)). \] (83)

Obviously, \( \kVar(\theta_{\nu,n}) = F_n - \rho_n^2 \), where
\[ F_n = \int g(z_{\nu,i,n}) \frac{dP_{\nu,n}(z_{\nu,i,n})}{dQ_{i,n}(z_{\nu,i,n})} dQ_{i,n}(z_{\nu,i,n}). \] (84)

Now suppose we make the theorem statement choice that
\[ dQ_{i,n}(z_{\nu,i,n}) = \sum_{i=1}^{m} p_i \exp(n\phi_n(\theta_{\nu_i}, z)) \exp(-n(\theta_{\nu_i}, \nu_i)) dP_{\nu,n}(z). \] (85)

where \((p_1, p_2, \ldots, p_m)\) is a probability vector with strictly positive components. Note that this is a valid probability distribution due to (82). Also let \( p^* = \min_{1 \leq i \leq m} p_i \).

Then for this choice of biasing distribution, and using (83), we have
\[ F_n \leq \int \left( \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \exp(-n(\theta_{\nu_i}, \nu_i)) \right)^2 \frac{dP_{\nu,n}(z)}{dQ_{i,n}(z)}. \]
\[ \leq \int \left( \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \exp(-n(\theta_{\nu_i}, \nu_i)) \right)^2 \frac{dP_{\nu,n}(z)}{p^* \left( \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \exp(-n(\theta_{\nu_i}, \nu_i)) \right)}. \] (86)

Fix \( \varepsilon > 0 \). There exists an \( n_0 \) such that for all \( n \geq n_0 \) we must have \( |\phi_n(\theta_{\nu_i}) - \phi(\theta_{\nu_i})| < \varepsilon \) for \( i = 1, 2, \ldots, m \). Thus since \( I(\nu_i) = (\theta_{\nu_i}, \nu_i) - \phi(\theta_{\nu_i}) \), we have \( \phi_n(\theta_{\nu_i}) \leq I(\nu_i) - (\theta_{\nu_i}, \nu_i) + \varepsilon \). For the purpose of displaying the next two equations, let us define
\[ d_1(z) = p^* \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \cdot \exp(-n(\theta_{\nu_i}, \nu_i) + nI(\nu_i) - n\varepsilon) \]
\[ d_2(z) = p^* \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \cdot \exp(-n(\theta_{\nu_i}, \nu_i) + nI(\nu_i) - n\varepsilon) \] (87)

Note that \( d_1(z) \geq d_2(z) \). Hence, for \( n \geq n_0 \),
\[ F_n \leq \int \left( \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \exp(-n(\theta_{\nu_i}, \nu_i)) \right)^2 \frac{d_1(z)}{d_2(z)} \frac{dP_{\nu,n}(z)}{dQ_{i,n}(z)} \]
\[ \leq \int \left( \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \exp(-n(\theta_{\nu_i}, \nu_i)) \right)^2 \frac{d_1(z)}{d_2(z)} \frac{dP_{\nu,n}(z)}{dQ_{i,n}(z)} \]
\[ \leq \frac{1}{p^*} \exp(-n(\varepsilon)) \]
\[ \times \int \left( \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \exp(-n(\theta_{\nu_i}, \nu_i)) \frac{dP_{\nu,n}(z)}{dQ_{i,n}(z)} \right) \]
\[ \leq \frac{1}{p^*} \exp(-n(\varepsilon)) \left[ \left( \sum_{i=1}^{m} \exp(n\phi_n(\theta_{\nu_i}, z)) \exp(-n(\theta_{\nu_i}, \nu_i)) \right) \right] \]
\[ \leq \frac{1}{p^*} \exp(-n(\varepsilon)) \left[ m \exp(-n(\varepsilon)) \right] \]
\[ \leq \frac{1}{p^*} \exp(-2n(\varepsilon)). \] (89)

Since \( \varepsilon > 0 \) is arbitrary, this simulation distribution sequence is efficient. ■

REFERENCES


[16] D. Lu and K. Yao, “Improved importance sampling techniques for efficient simulation of digital communication systems,” IEEE
KEYWORDS
Importance Sampling, Large Deviation Theory, Monte Carlo Simulation, Conditional Importance Sampling.
FIG. CAPTIONS

Fig. 1. Plot of $\rho(k)$ vs. $k$.
Fig. 2. Plot of Error vs. XSR.

BIOGRAPHIES

James A. Bucklew 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 in Mathematics at the University of Wisconsin, Madison. His research interests are in the application of probability and statistics to signal processing and communication problems. He has published over 100 articles in these fields. He has published several fly fishing articles (in Spanish) and also works as a translator. Prof. 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. He is the author of the books *Large Deviation Techniques in Decision, Simulation, and Estimation* (New York: Wiley-Interscience, 1990) and *Introduction to Rare Event Simulation* (New York: Springer-Verlag, 2004).
Fig. 1. Plot of $\hat{p}(k)$ vs. $k$.

Fig. 2. Plot of Error vs. XSR.