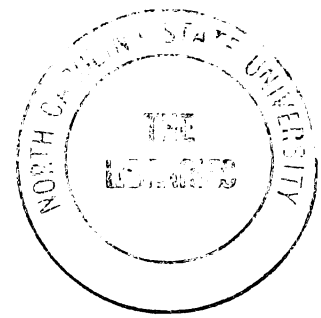
An Algorithmic Approach to the Optimization of Importance Sampling Parameters in Digital Communication System Simulation

Michael Devetsikiotis

J. Keith Townsend



Center for Communications and Signal Processing
Department of Electrical and Computer Engineering
North Carolina State University

TR-93/7
April 1993

To appear in the *IEEE Transactions on Communications*

An Algorithmic Approach to the Optimization of Importance Sampling Parameters in Digital Communication System Simulation

Michael Devetsikiotis, *Student Member IEEE*
J. Keith Townsend, *Member IEEE*
Center for Communications and Signal Processing,
Department of Electrical & Computer Engineering,
North Carolina State University, Raleigh, NC 27695-7914

Abstract

Importance sampling is recognized as a potentially powerful method for reducing simulation runtimes when estimating the bit error rate (BER) of communications systems using Monte Carlo simulation. Analytically minimizing the variance of the importance sampling (IS) estimator with respect to the biasing parameters has typically yielded solutions for systems for which the BER could be found analytically, e.g., linear system with additive Gaussian noise.

We present in this paper a new technique for finding an asymptotically-optimal set of biasing parameter values, in the sense that as the resolution of the search and the number of runs used both approach infinity, the algorithm converges to the true optimum. A key feature is that repetitive, very short simulation runs are used to determine asymptotically-optimal IS parameter values. Thus, knowledge of the system — required by any effective importance sampling scheme — is obtained from a controlled set of simulation runs. Our algorithm determines the amount of biasing that minimizes a statistical measure of the variance of the BER estimate and exploits a theoretically justifiable relationship, for small sample sizes, between the BER estimate, $\hat{P}_e$, and the amount of biasing. In this paper we consider the translation biasing scheme, although the algorithm is applicable to other parametric IS techniques. Only mild assumptions are required of the noise distribution and system.

Experimentally, improvement factors ranging from two to eight orders of magnitude are obtained for a number of distributions for both linear and nonlinear systems with memory.

1. This work was supported by the Center for Communications & Signal Processing as a core project.

2. Portions of this paper have been presented at the IEEE Global Telecommunications Conference, GLOBECOM'90, San Diego, December 1990.

1 Introduction

Importance Sampling (IS) is recognized as a potentially powerful method for reducing simulation runtimes when estimating the bit error rate (BER) of communications systems using Monte Carlo (MC) simulation [1, 2, 3, 4, 5, 6]. Parametric biasing schemes based on increasing the noise variance (*variance modification*) [2] and shifting (*translation*) of the noise probability density function (pdf) [6] have been presented.

However, to realize large improvement factors in simulation runtime with IS, good biasing parameter values for the noise distribution(s) must be found. One method researchers have used to attack this problem is to analytically minimize the variance of the IS estimator with respect to the biasing parameters [6, 7]. This approach has typically yielded solutions for systems for which the BER could be found analytically, or with minimal computational effort, e.g., linear system with additive Gaussian noise. For more general systems, finding optimal biasing parameter values analytically may require an amount of effort comparable to solving the BER problem itself [7, 8].

We present in this paper a new technique for finding an asymptotically-optimal set of biasing parameter values, in the sense that as the resolution of the search and the number of runs used both approach infinity, the algorithm converges to the true optimum. As will be further explained later, a *simulation run* or simply *run* consists of a fixed number of i.i.d. observations called here *decisions*, each one of which corresponds to a fixed number (equal to the system memory) of *samples*. A key feature of our approach is that repetitive, very short simulation runs are used to determine the asymptotically-optimal IS parameter values [9]. Information on the performance of the biasing scheme for each biasing value is extracted from this series of short simulation runs. Our method also exploits a theoretically justifiable relationship, for small sample sizes, between the BER estimate, $\hat{P}_e$, and the amount of bias.

The fundamental difference between our method and earlier techniques is that information about the system (required by any IS scheme) is extracted from simulation runs using a theoretically justifiable, statistical algorithm. Therefore, the applicability of our method is not restricted to systems in which analytical methods are used to provide the optimal biasing parameter values.

The promise that earlier techniques would soon be applicable to non-analytically tractable systems has always been impeded by fundamentals. Namely, that the estimator variance expression, which these techniques attempted to minimize analytically, is of the same form as the original problem to be solved (by simulation). The technique presented in this paper circumvents this problem and relaxes the assumptions about the system.

The algorithm we present assumes that the IS technique used is *parametric* with parameters C_i, $i = 1, \ldots, M$, defined in $[0, +\infty)$, such that $C_i = 0$, $i = 1, \ldots, M$, corresponds to standard MC simulation, while increasing C_i's imply an increased "amount of biasing". Our algorithm determines the amount of biasing that minimizes a statistical measure of the variance of the BER estimate.

An exhaustive, brute-force minimization of the estimator variance using simulation would not be practical without additional information. However, the search space of these parameter values is reduced to a manageable size by exploiting a relationship (rigorously described in Appendix A) between the behavior of the BER estimate and amount of biasing. For a class of biasing schemes including but not restricted to the translation technique [6], increas-

ing C_i's induce a monotonic increase of the means and/or a monotonic "probability mass transfer" of the noise distributions involved. We show that for such biasing schemes and for very small sample sizes (i.e., very short runs) and large biasing amounts (over-biasing), the BER will be *underestimated* in a given simulation run with probability arbitrarily close to unity. Proof of this important result is given in Appendix A. Improvement in simulation efficiency cannot occur in the "over-biasing region." This confines the search to a finite region in the given direction. Our algorithm then iteratively seeks to minimize the statistical measures of the estimator variance within this "improvement region."

The statistical measures of performance optimized by the algorithm are all asymptotically equivalent to maximum likelihood expressions for estimator variance, as shown in Section 3. Thus, as explained in Section 3, the algorithm produces an asymptotically optimal estimate of the amount of biasing.

In this paper we consider the translation biasing scheme, although the algorithm is applicable to other parametric IS techniques, like the "quasi-translation" scheme addressed in [10]. A topic of concurrent research has also utilized this method for the variance modification scheme, in a case where the distributions were single-sided exponentials [11].

For the translation technique, as in [6], the search can be confined along a given direction (a line) in the M-dimensional parameter space (where M is the number of samples of memory in the system), with the obvious advantage that the dimensionality of the optimization problem is reduced to unity. Here, the optimal translation magnitude (optimal point on this line) is found using simulation results.

After model description in Section 2, the algorithm is developed in Section 3. In Section 4 we briefly discuss the issues involved in choosing a favorable translation direction, although characteristics of the algorithm developed in Section 3 are invariant to direction.

Section 5 presents experimental results using the method to estimate the BER of a binary digital communications system. Experimentally, the near-optimal translation determined by the algorithm is in excellent agreement with the optimal translation derived analytically for a linear system with additive Gaussian noise. Large improvement factors for other distributions as well as nonlinear systems are also demonstrated. Concluding remarks are given in Section 6. In Appendix B we show that the distributions used in this paper satisfy the conditions of the theorem in Appendix A.

2 Definitions and Model Description

2.1 System Model Description

Most of our notation is based on that in [5] and in [6]. We consider the problem of estimating the BER of a binary communications system with sampled output, $Y_k = g(\mathbf{X})$, where $g(\cdot)$ is the system transfer characteristic (system response) and vector $\mathbf{X}$ is defined as $\mathbf{X} = [X_k, X_{k-1}, \cdots, X_{k-M+1}] = \mathcal{Q}(\mathbf{A}, \mathbf{N})$. $\mathbf{A}$ and $\mathbf{N}$ are, respectively, the M-dimensional input data and noise vectors, and M is the system memory length in samples. $\mathcal{Q}$ is a transformation, e.g., addition, that combines the signal and the noise into a composite random process. A_k takes values A (under hypothesis H_1) and $-A$ (under H_0). The H_0 hypothesis will be assumed in this paper, without loss of generality. The actual values that these

random vectors can take will be denoted by $\mathbf{x}$, $\mathbf{a}$ and $\mathbf{n}$.

2.2 Probability of Error

Let $f_X(\mathbf{x})$, $f_A(\mathbf{a})$ and $f_N(\mathbf{n})$ be the pdfs of $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{N}$ respectively. As in [6, 7] the superscript $(*)$ denotes the corresponding pdf under IS. In the general case, the characteristics of the noise depend on the transmitted symbol, thus we have a conditional pdf $f_{N|A}(\mathbf{n})$.

The output samples $\{Y_k\}$ are compared with a threshold T to determine whether a 0 or a 1 was sent. The BER of the system is given by

$$P_e = \int_{-\infty}^{\infty} I_T(g(\mathbf{a}\,,\,\mathbf{n}))f_A(\mathbf{a})f_{N|A}(\mathbf{n})\,d\mathbf{a}d\mathbf{n} \tag{1}$$

where $I_T(y)$ is an indicator function equal to 1 for $y \geq T$ and to 0 for $y < T$.

A "realization" is a block of input bits (symbols) with length equal to the system memory in bits, K ($M = K \times$ *samples/bit*). Assuming equiprobable realizations $\mathbf{a}(j)$, $j = 0, 1, \ldots, 2^{K-1} - 1 = J - 1$ of $\mathbf{A}$, Eq.(1) can be written $P_e = 1/J \sum_{j=0}^{J-1} P_e(j)$, where

$$P_e(j) = \int_{-\infty}^{\infty} I_T(g(\mathbf{a}(j)\,,\,\mathbf{n}))f_{N|j}(\mathbf{n})d\mathbf{n} \tag{2}$$

and $f_{N|j}(\mathbf{n})$ is the distribution of $\mathbf{n}$ conditioned on realization $\mathbf{a}(j)$. Clearly, when the data vector $\mathbf{A}$ and the noise vector $\mathbf{N}$ are independent $f_{N|A}(\mathbf{n}) = f_{N|j}(\mathbf{n}) = f_N(\mathbf{n})$.

2.3 MC and IS Estimators

During each simulation run, one decision (i.e., comparison of Y_k with the threshold) is made in every M samples, so that decisions correspond to i.i.d. observations. Under IS, an estimator of the BER using the so-called block approach is given by [6]:

$$\hat{P}_e^* = \frac{1}{N^*} \sum_{j=0}^{J-1} \sum_{i=1}^{N^*/J} I_T\left(g\left(\mathbf{a}(j)\,,\,\mathbf{n}^*(j,i)\right)\right) w\left(\mathbf{x}_i^*\right) \tag{3}$$

where N^* decisions are used for the simulation run, j represents the conditioning on every realization $\mathbf{a}(j)$ of $\mathbf{A}$, and $N^* \,/\, J$ decisions correspond to each realization. We say that such an estimate corresponds to a simulation run with *length* N^*. In this paper we use the terms "simulation run" and "run" interchangeably. Assuming only the noise pdf is biased and the noise samples are mutually independent, the weight function is:

$$w\left(\mathbf{x}_k^*\right) = \prod_{i=0}^{M-1} f_{N|j}(n_{k-i}^*)/f_{N|j}^*(n_{k-i}^*) = w_{N|j}(\mathbf{n}_k^*)\ \text{, for all } k \tag{4}$$

In the block simulation model, a decision is made once in each realization or "block".

It is easy to show that the estimator in (3) is unbiased [2, 5]: $E\left[\hat{P}_e^*\right] = P_e$. The typical approach to finding the optimal IS parameter values is to attempt to minimize *analytically* the estimator variance σ_{IS}^2 [6], or the *time-reliability product* $\sigma_W^2 = N\sigma_{IS}^2$ [7], where

$$\sigma_{IS}^2 = \frac{1}{JN^*} \sum_{j=0}^{J-1} \int_{-\infty}^{\infty} I_T(g(\mathbf{a}(j), \mathbf{n}^*))\left(w_{N|j}(\mathbf{n}^*) - P_e(j)\right) f_{N|j}(\mathbf{n}^*)d\mathbf{n}^* \tag{5}$$

with respect to those parameters, although this approach usually yields solutions for systems not requiring simulation.

The MC estimator, $\hat{P}_e$, and corresponding variance, σ^2_{MC} are obtained by replacing densities f^* with f, and the number of decisions N^* with N in (3) and (5).

3 Optimization Algorithm Development

3.1 Behavior of the IS Estimate

Under parametric IS, the biased pdf is of the same basic shape as the original pdf, with modifications completely described by a vector of IS parameters, **c**:

$$f^*_N(\mathbf{n} \mid j) = f_N(\mathbf{n}, \mathbf{c}(j)) \tag{6}$$

for each realization $j = 0, \ldots, J-1$. Note that this parametric formulation is very general and will accommodate variance modification [2], translation [6] or "quasi-translation" [10]. The vector of IS parameters $\mathbf{c}(j)$ can be written as $\mathbf{c}(j) = C(j)\mathbf{d}$, where **d** is a unit vector describing the *direction* of biasing. We assume here that $C(j) \in [0, +\infty)$, with $C(j) = 0$ corresponding to standard MC, and that the conditions of Theorem 1 in Appendix A are satisfied. We focus our attention on the problem of estimating, for a given direction **d**, optimal parameter values $C(j)$ for each realization $j = 0, \ldots, J-1$. Define the *improvement ratio* as $r_{IS}(C(j)) = \sigma^2_{MC}(C(j))/\sigma^2_{IS}(C(j))$ when the number of decisions used is the same, or equivalently, as $r_{IS}(C(j)) = N/N^*$ when the MC and IS variances are equal. We observe that the optimization space $[0, +\infty)$ for $C(j)$ can be divided into three regions: under-biasing region, improvement region, and over-biasing region.

In the under-biasing region, $C(j)$ is relatively small, and the biased pdfs f^* are very similar to the original pdfs f. Assuming a low P_e and a very small number of decisions, N, $\hat{P}_e$ will typically be $\hat{P}_e = 0$.

As the biasing amount $C(j)$ is increased towards more favorable settings, the variance in the estimate will decrease as the improvement $r_{IS}(C(j))$ of the IS estimator over the standard MC method increases. For values of $C(j)$ in the neighborhood of the optimal, the estimates will be closest to the true P_e. This region, the *improvement* region, includes $C_{opt}(j)$. Furthermore, in this region, statistical performance measures such as *estimates* of $\sigma^2_{IS}(j)$ can be used to locate the optimal biasing value, with increased statistical confidence. Also in a neighborhood around $C(j) = C$, the local "scatter" or "noisiness" of the estimate, observed as a function of the parameter $C(j)$, is an additional indication of the improvement $r_{IS}(C(j))$ at $C(j) = C$.

As the value $C(j)$ increases beyond the improvement region, estimator variance will increase and $r_{IS}(C(j))$ will deteriorate. A key observation here is that, under certain conditions, as $C(j)$ increases, the increasing estimate variance does *not* manifest itself as statistical "scatter" centered around the true value $P_e(j)$. Instead it appears as a severe and consistent *underestimation* of P_e which we call *apparent bias* or *sample bias* to distinguish it from the conventional statistical bias, and which becomes more severe as $C(j)$ increases. This may go against simple intuition based on the fact that "the IS estimator is *unbiased*". Note, however, that determining whether an estimator is "unbiased" involves an *ensemble expectation*, whereas a simulation estimate is based on a *finite* number of samples N, on a *sample path*.

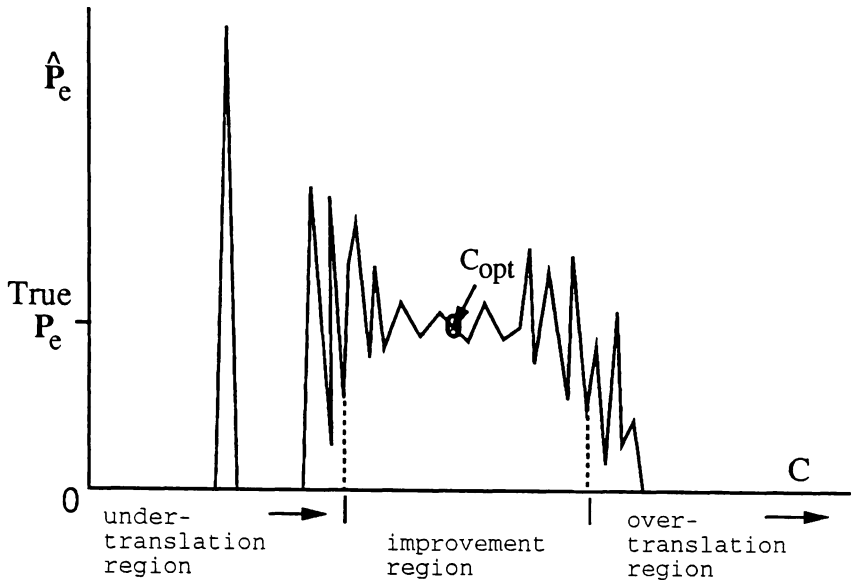


Figure 1: A typical curve of $\hat{P}_e$ vs. biasing parameter C. For C too low most estimates are equal to 0. For C too high underestimation occurs with high probability. Near-optimal C's are chosen from the flattest region in the middle (the improvement region).

These observations are substantiated by Theorem 1, (proved in Appendix A) which in essence states that BER estimates according to (3) (for any fixed number of samples N) can be made arbitrarily small with probability arbitrarily close to unity, as $C(j) \to \infty$. The significance of this theorem is that it confines the interval of $C(j)$ to be searched by the algorithm. Without this theorem, the search would be a hopeless trial-and-error.

Based on the behavior of the BER estimate in the three regions, a typical plot of the resulting estimates as a function of C will look something like the one in Figure 1. Clearly, in order to obtain a plot similar to Figure 1, the chosen biasing scheme must provide significant improvement for the system being analyzed.

3.2 Statistical Measures of IS Performance

This section presents the functions that the algorithm will optimize. Assume biasing according to (6). Let C_{max} be a value of the biasing parameter chosen in the over-biasing region (discussed below). Let $C_k = k \times C_{max}/K$, $k = 0, \ldots, K$, where K is a "resolution" number. Assume that $K+1$ simulation runs are performed, one at each C_k, $k = 0, \ldots, K$. Let N be the number of decisions used for each simulation run. The following statistical quantities will be used in our algorithm as *statistical measures of performance* of the IS scheme, expressed as functions of the parameter C_k and the number of decisions N:

- The BER estimate $\hat{P}(C_k, N)$
- The "intra-simulation variation coefficient" estimate $\hat{V}(C_k, N)$,

$$\hat{V}(C_k, N) = \begin{cases} \frac{\sum_{i=1}^{N} I(g(\mathbf{x}_i))w^2(C_k, \mathbf{x}_i) - N\hat{P}^2(C_k, N)}{(N-1)\hat{P}^2(C_k, N)} & \text{for } \hat{P} > 0 \\ N & \text{for } \hat{P} = 0 \end{cases}$$

- The "inter-simulation" variation coefficient estimate or "local scatter" measure $\hat{S}(C_k, N)$,

$$\hat{S}(C_k, N) = \frac{1/2W \sum_{l=k-W}^{k+W} \left(\hat{P}(C_l, N) - \overline{P}_W(C_k, N)\right)^2}{\overline{P}_W^2(C_k, N)}$$

where

$$\overline{P}_W(C_k, N) = 1/(2W+1) \sum_{m=k-W}^{k+W} \hat{P}(C_m, N)$$

and W is the window size, typically small ($W = 2$ or 3).

The usefulness of $\hat{P}(C_k, N)$ in evaluating the performance of IS at C_k lies in the fact that, under the conditions of Appendix A, $\hat{P}(C_k, N)$ has a predictable behavior for values of C_k outside the improvement region, i.e., is identically zero for C_k in the under-biasing region and decreases towards zero in the over-biasing region, both with very high probability.

The numerator of $\hat{V}$ is clearly an unbiased and consistent estimator of the *time-reliability product* $\sigma_W^2 = N\sigma_{IS}^2$ defined in Section 2, at C_k. Although $\hat{V}$ is not an unbiased estimator of the variation coefficient $\sigma_W^2(C_k)/P_e^2$ it is, however, a consistent estimator, $\hat{V}(C_k, N) \rightarrow \sigma_W^2(C_k)/P_e^2$ as $N \rightarrow \infty$. This "internal" variance estimate is based on the same statistical observations as the BER estimate, i.e., values of $I(g(\mathbf{x}_i))w(\mathbf{x}_i)$, therefore, not only $\hat{V}$ but also its *variance* is lower in the neighborhood of the optimal C. This turns out to be very important for our approach, since the number of decisions N it uses is very small.

The "inter-simulation" variation coefficient estimate or "local scatter" measure $\hat{S}$ extracts information on the variance at C_k from the scatter or variability of estimates obtained *in the neighborhood* of C_k, that is at $C = C_{k-W}, \ldots, C_{k+W}$. Our "inter-simulation" estimate takes advantage of the fact that the behavior at neighboring biasing amounts is strongly correlated. Thus, in a small neighborhood around C_k, $\sigma_{IS}^2(C)$ can be approximated by $\sigma_{IS}^2(C_k)$. $\hat{S}(C_k, N)$ is an approximate estimate of $\sigma_{IS}^2(C_k)/P_e^2$, in the sense of an "ensemble" estimate.

For K large ($K \rightarrow \infty$) and W constant so that $C_{k+W} - C_{k-W} = 2(W/K)C_{max} \rightarrow 0$, $\hat{S}$ would asymptotically approach an ensemble estimate of the variation coefficient $\sigma_{IS}^2(C_k)/P_e^2$. Then, for $W \rightarrow \infty$ while $W/K \rightarrow 0$, $\hat{S}$ would go to the exact value $\sigma_{IS}^2(C_k)/P_e^2$. In [9] $\hat{S}$ was minimized by the human user.

3.3 Optimization Algorithm

The above observations and measure definitions lead to the following algorithm that locates optimal settings for C, for a given direction $\mathbf{d}$. For each realization $\mathbf{a}(j)$:

1. Find a parameter value $C_{max}(j)$ in the over-biasing region as follows:

   (a) Run a series of simulations with length N, where biasing is done as in (6), with $C(j) = C_k(j)$, $k = 0, \ldots$, where $C_0(j) = C_{min}$ (C_{min} is an arbitrarily small number), and $C_k(j) = 10\, C_{k-1}(j)$, until $0 < \hat{P}(C_l(j), N) < \tau$ for some $k = l$. τ is a threshold value below which the estimate is assumed to severely underestimate the true P_e, allowing us to exclude the under-biasing region and part of the over-biasing region from consideration.

(b) Set $C(j) = C_l(j)/2$ and run another simulation. If $\hat{P}(C(j), N) < \tau$ set $C_{max}(j) = C_l(j)$ and repeat the step, otherwise stop.

2. Find the optimal parameter value as follows:

(a) Choose the resolution number K. Set $C_0(j) = 0$. Run $K + 1$ simulations with length N, using values $C_k(j)$, $k = 0, \ldots, K$ with $C_k(j) = k\, C_{max}(j)/K$. Store the estimates $\hat{P}(C_k(j), N)$ and $\hat{V}(C_k(j), N)$ from each simulation run.

(b) From the estimates $\hat{P}(C_k(j), N)$ generate $\hat{S}(C_k(j), N)$ and calculate the *cost function*

$$\mathcal{D}(C_k(j), N) = \begin{cases} N\hat{V}(C_k(j), N) \times \hat{S}(C_k(j), N), & \hat{P}(C_k(j), N) > \tau \\ BIG, & \hat{P}(C_k(j), N) \leq \tau \end{cases}, k = 0, \ldots, K$$

BIG is a very large number (say, the largest real on the computer).

(c) Find $k_{opt} \in \{0, \ldots, K\}$ such that the above cost function is minimum. Return $C_{k_{opt}}(j)$ as the optimum or "minimum estimated variance" value on the given search direction **d**.

The accuracy of any statistical method is proportional to the sample size of the experiments involved. Under IS, the "effective sample size" of each simulation run is given by the "time-improvement product" $N\, r_{IS}(C(j))$. Thus, within the improvement region, sample sizes N significantly smaller than those needed by standard MC methods are required to obtain useful statistical measures with acceptable accuracy. This, together with the fact that only the *relative* and not the absolute values of such statistical measures determine the outcome of the algorithm, allows our algorithm to operate successfully with extremely small sample sizes N. In our experiments we started with the smallest N possible and repeated parts 1-(a), 1-(b), and 2-(a) of the algorithm, increasing N by a factor of 10 until a well defined minimum could be detected in a filtered (smoothed) version of the intra-simulation variance estimate $\hat{V}(C_k(j), N)$, $k = 0, \ldots, K$.

The choice of K is not crucial for the performance of the algorithm since K merely determines the resolution or the numerical accuracy of our search along the chosen direction. A value of $K = 100$ has worked well in practice.

The threshold value τ can be chosen as a very small fraction of a very conservative lower bound $\hat{P}_{e,min}$ on P_e (e.g., $\tau = \hat{P}_{e,min}/10^{10} \ll \hat{P}_{e,min} \ll P_e$).

The window size W for $\hat{S}$ has to be relatively small to preserve the local character of $\hat{S}$, but also large enough to carry statistical significance. Values of W between 2 and 5 seem to work well in practice. In the worst case, the effect of both extremely small and extremely large selections for W will be for $\hat{S}$ to exhibit no statistically significant minima most of the time. In this case the burden of the minimization procedure would be on $\hat{V}$, which can still provide a favorable parameter value $C_{k_{opt}}(j)$.

3.4 Characteristics of the Algorithm

$\mathcal{D}(C(j), N)$ approaches asymptotically $N^2\sigma_{IS}^4/P_e^4$ as the number of decisions N, the resolution K, and the window size W increase towards infinity with $W/K \to 0$. In this sense, the

minima $C_{k_{opt}}(j)$ of $\mathcal{D}(C(j), N)$ detected by the algorithm for $j = 0, \ldots, J-1$, are *asymptotically optimal* IS biasing parameters.

Both $\hat{V}$ and $\hat{S}$ were shown earlier to be statistical measures of performance for the IS scheme at $C(j)$, in the sense that they are estimates of the coefficient of variation $\sigma_{IS}^2(C(j))/P_e^2$. Furthermore, they both approach $\sigma_{IS}^2(C(j))/P_e^2$ asymptotically. In fact, experimental results indicate that either of them could be used separately to find a near-optimum $C(j)$. Letting $\hat{P}$ define the improvement region and combining $\hat{V}$ with $\hat{S}$ leads to an objective function which has a very well-defined, steep valley around the optimal values of $C(j)$. The advantage of this composite objective function is that, by using all the information available, it allows for a more reliable and accurate minimization since the randomness of the simulation data can be dealt with much more effectively.

Statistical estimates of the IS variance would have been of much less practical value without some additional knowledge about the parameter search search or if an unrealistically large N had to be used. However, in this case the BER estimate $\hat{P}(C(j), N)$ helps us identify the improvement region, thus drastically reducing the search space. In using very small N, we rely on the predictable behavior of $\hat{P}(C(j), N)$ vs. $C(j)$ and on the fact that, although N is small, the increased improvement factor $r_{IS}(C(j))$ will lead not only to values of $\hat{P}$ close to the true P_e but also to estimates of variance $\hat{V}$ reliable enough so that a relative minimum (w.r.t. $C(j)$) is easy to locate. The advantage of combining (through $\hat{P}$) the under-estimation behavior of Appendix A with "internal" or "within-simulation" measures $\hat{V}$ and "inter-simulation" measures $\hat{S}$, is that a near-optimal value for C can be efficiently obtained with a rather small number of simulation runs K, each having a small number of decisions N. Thus, with a total number of observations equal to $K \times N$ and with only negligible additional processing (calculating $\hat{S}$ for K values of $C(j)$ and finding the minimum $\mathcal{D}$ again out of K values) we can obtain a value for C arbitrarily close to $C_{opt}(j)$.

4 Direction of Biasing: The Translation Technique

The algorithm described in Section 3 finds asymptotically optimal biasing parameter values in a given direction. This section addresses the issue of how to choose this direction for the translation biasing scheme [6]. This biasing scheme is discussed here for two reasons: (1) it provides large improvement factors for a wide range of systems, and (2) it is easy to theoretically justify a good direction choice. Under the translation technique $f_N^*(\mathbf{n} \mid j) = f_N(\mathbf{n} - \mathbf{c}(j))$, where $\mathbf{c}(j) = C(j)\mathbf{d}$, and $\mathbf{d}$ is a unit vector describing the direction of biasing. Fixing the direction of translation $\mathbf{d}$ has the effect of reducing the dimensionality of the underlying optimization problem to unity. Although the characteristics of the described algorithm do not directly depend on the direction, clearly this choice can greatly affect the speed-up factor achieved by IS. The following discusses the choice of a favorable direction for the translation technique.

4.1 Linear Systems

For linear systems, the input-output relationship $g(\cdot)$ is described by $y = g(\mathbf{x}) = \mathbf{h}^T\mathbf{x}$ (we assume $\|\mathbf{h}\| = 1$). Let $y^*(C, \mathbf{d}) = \mathbf{h}^T\,\mathbf{n}^* = \mathbf{h}^T\,(\mathbf{n} + C\,\mathbf{d}) = y + C\,\mathbf{h}^T\mathbf{d}$ be the biased output

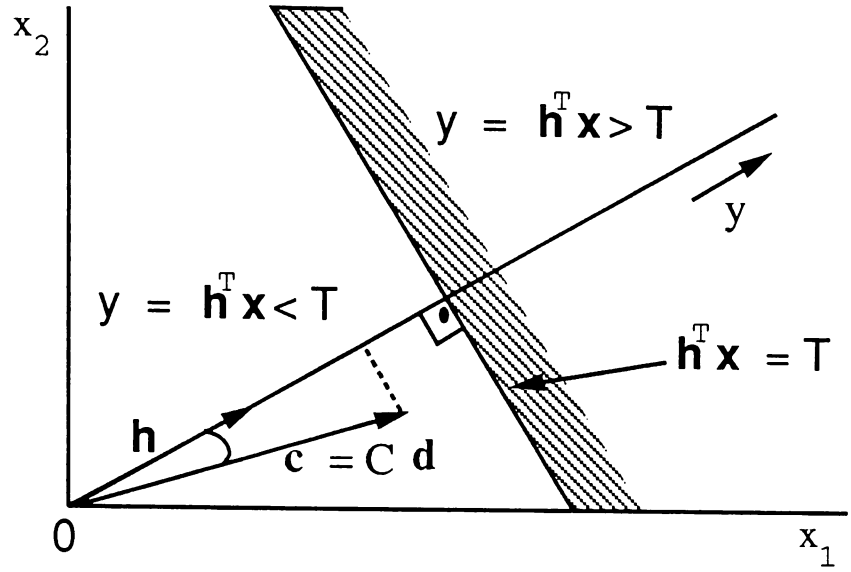


Figure 2: Two-dimensional illustration of an impulse response $\mathbf{h}$, a biasing direction $\mathbf{d}$, and a decision surface for a linear system.

resulting from translating the input $\mathbf{n}$ by $C\,\mathbf{d}$. Out of all possible directions, the entire half-space $\mathbf{h}^T\,\mathbf{d} \leq 0$ cannot be favorable since the resulting net translation at the output would be non-positive. It was implied in [6] and proven in [7] that, for a linear system with additive Gaussian noise, the optimal $\mathbf{c}(j)$ is $\mathbf{c}_{opt}(j) = C(j)\,\mathbf{h}$, where $\mathbf{h}$ is the impulse response. Furthermore, the direction of $\mathbf{h}$ is also optimal, for linear systems, in the sense that the dimensionality is reduced to unity [7].

As a last and most important argument in favor of biasing in the direction of $\mathbf{h}$, note that, in the M-dimensional input space, the important region boundary is a hyperplane perpendicular to the impulse response vector. The output y of the linear system is the projection of the input vector on the direction of the impulse response, with the decision threshold located T units of distance from the origin (Figure 2). Therefore, among all directions $\mathbf{d}$, an additive bias $C\,\mathbf{h}$ *maximizes the effect of translation at the output of a linear system* since $\max_{\mathbf{d}} \{y^*(C, \mathbf{d})\} = y^*(C, \mathbf{h}) = y + C\,\|\mathbf{h}\|^2 = y + C$. Although the optimal direction cannot be determined analytically, these arguments suggest using $\mathbf{d} = \mathbf{h}$ when the system is linear.

4.2 Nonlinear Systems

In general, the choice of good translation direction(s) for nonlinear systems will depend on the characteristics of the particular nonlinearity. The model we consider here is a cascade of a *linear system with memory* and a *memoryless nonlinearity*. According to this model, at time sample k,

$$y_k = g(w_k) = \sum_{l=0}^{\infty} a_l\, w_k^l \tag{7}$$

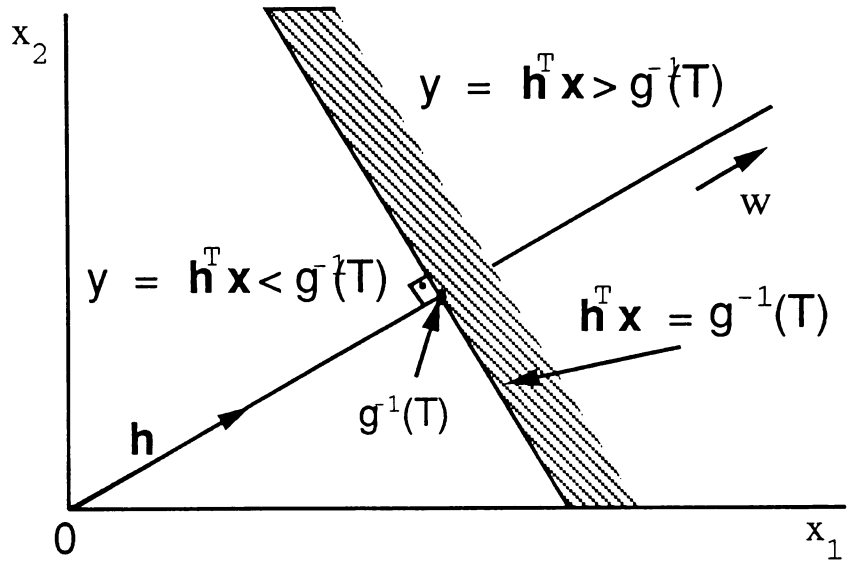


Figure 3: Mapping of the important region when the nonlinearity is monotonic (increasing).

and

$$w_k = \sum_{i=0}^{\infty} h_i \, x_{k-i} \tag{8}$$

Under IS, increasing the number of error events $y \in \Omega_Y$, is accomplished by increasing the probability of $\mathbf{x} \in \Omega_X$, where $y = g(\mathbf{x})$. Assuming monotonicity, $\Omega_Y = \{y : y > T\}$ and $y = g(w)$ imply $\Omega_W = \{w : w > g^{-1}(T)\}$ and $\Omega_X = \{\mathbf{x} : \mathbf{h}^T \mathbf{x} > g^{-1}(T)\}$ (an increasing function was assumed, without loss of generality). Thus biasing the linear case can be used, where T will be substituted by $\tilde{T} = g^{-1}(T)$ (Figure 3).

If $g(\cdot)$ cannot be assumed monotonic in the entire operating range the problem becomes more complicated. We can distinguish two interesting cases:
(i) Ω_W is of the type $\Omega_W = \{w : w < \tilde{T}_1 \text{ or } w > \tilde{T}_2\}$. Then, a split-and-translate scheme could be applied [12], where the pdf is split in two identical parts (with probability $\frac{1}{2}$ each) that are translated in the two opposite directions. Due to symmetry, and for all practical purposes, this scheme will have effects identical to simple translation and can be analyzed in the same way.
(ii) Ω_W is of the type $\Omega_W = \{w : \tilde{T}_1 > w > \tilde{T}_2\}$, where $|\tilde{T}_1 - \tilde{T}_2|$ is assumed large, i.e., many standard deviations of the pdf involved. Then, the scheme proposed earlier can still be applied, assuming that over-translation will never cause the modified pdf to have significant values beyond $\tilde{T}_1$, i.e., "spill-over" beyond the important region.

If $g(\cdot)$ is non-monotonic *and* does not fall in these two categories then, in general, biasing by translation may not be advantageous.

4.3 Applicability of Algorithm to Nonlinear Systems

The observed effects of under- and over-translation, as well as the behavior of our statistical measures are not restricted to the case of a linear system. For a given translation direction

d, our algorithm can be used to determine the *optimal amount* of translation C along that direction. For nonlinear systems, the problem of identifying favorable biasing directions is less simple in general. However, our technique works well for a nonlinear system model characterized by the cascade of a linear system followed by a memoryless nonlinearity.

5 Applications and Experimental Results

5.1 Set-Up

In order to verify the effectiveness of this new technique in locating favorable IS parameter values, extensive simulation experiments were performed. The algorithm was applied in conjunction with the translation biasing technique to systems with Gaussian and "asymmetrical two-sided exponential" (ATE) pdf's. The ATE used was defined as $f_N(n) = \exp(-n/0.7)$, when $n \geq 0$ and $f_N(n) = \exp(n/0.3)$ when $n < 0$. In Appendix B we show that these distributions satisfy the conditions of the theorem in Appendix A. The Gaussian simulations were performed because analytical results for the optimal translation and variance reduction are available. In no way did the algorithm require or make use of any *a priori* knowledge of the fact that the distribution was Gaussian.

The algorithm was also used for systems with a Rayleigh and a distribution describing the statistics of the shot noise of an avalanche photodiode (the WMC distribution). Since these distributions have semi-infinite support, pure translation cannot be used for biasing. Instead, we used a modification of the translation technique ("quasi-translation", see [10]) so that the IS estimators for the Rayleigh and WMC pdf's are unbiased.

Examples of both linear and nonlinear systems were used in the simulations. Two linear low-pass filters (based on a Hamming window design, normalized cut-off frequency 0.09) were used, one with $M = 24$ taps and one with $M = 40$ taps. The number of realizations (under H_0) was respectively, $J = 4$ and $J = 16$. The nonlinear systems were built by cascading linear low-pass filters (one with $M = 24$ taps, normalized cut-off $= 0.15$, and one with $M = 40$ taps, normalized cut-off $= 0.08$). The instantaneous nonlinearity is shown in Figure 4. Eight samples per bit was used in all cases.

5.2 Results

For each pdf and for each of the four systems, appropriate decision thresholds T were chosen to yield probabilities of error in the ranges of 10^{-7} to 10^{-8}, and 10^{-9} to 10^{-10}. Then, for each case and for every realization $\mathbf{a}(j)$, we used the proposed algorithm to locate favorable IS parameter values. The algorithm required an amount of runtime dominated by the term $K \sum_{i=1}^{MAX} t(N_i)$, where $t(N)$ was the runtime per simulation estimate $\hat{P}_e^*$ when N decisions were used, N_1 was the smallest number of decisions possible ($N_1 = 4$ when $M = 24$, $N_1 = 16$ when $M = 40$), $N_{i+1} = 10 \times N_i$, and N_{MAX} was the smallest number of decisions yielding a well-defined minimum in $\hat{V}(C_k(j), N_i)$, $k = 0, \ldots, K$. Some additional overhead for the estimation of $C_{max}(j)$ and some negligible post-processing for calculating and minimizing $\mathcal{D}(C(j), N)$ was also needed. In our experiments, N_{MAX} was between $N_{MAX} = 40$ and $N_{MAX} = 4,800$. Runtimes $t(N)$ ranged from 2.44 seconds ($N = 4$) to 73.1 seconds ($N =$

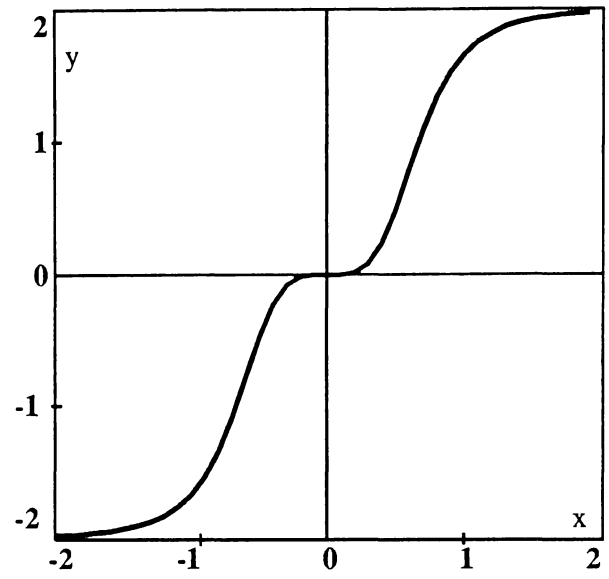


Figure 4: Instantaneous nonlinearity used in the simulations.

4,800), on a DECstation 3100. A resolution of $K = 100$ was used, combined with $C_{min} = 0.01$, $\tau = 10^{-20}$, and $W = 3$. Figures 5 and 6 show examples curves of $\hat{P}$, $\hat{V}$, $\hat{S}$ and $\mathcal{D}$ and the corresponding choices for $C(j)$. Curves for all other cases displayed similar behavior. Note that when the potential variance reduction is very large, as for the Gaussian and the Rayleigh pdfs, the corresponding curves are less noisy. A steeper valley in the cost function $\mathcal{D}$, as in the Gaussian case, indicates that variance reduction is more sensitive to the choice of C.

Variance reduction was computed by running complete block simulations using the values of $C(j)$ chosen by the algorithm. For each pdf, 100 IS estimates $\hat{P}_e^*$ of the total P_e (under H_0) were obtained according to (3). The BER and the estimator variance were estimated respectively as the sample mean and the sample variance over these 100 runs. The improvement ratio over MC simulation was then calculated as $r_{IS} = \sigma_{MC}^2/\sigma_{IS}^2$, where the approximation $\sigma_{MC}^2 \approx \hat{P}_e/N$ was used.

The estimated BER's and improvement ratios for each pdf and memory length appear in Table 1 (linear systems) and Table 2 (nonlinear systems). The improvement in all cases was inversely related to $\hat{P}_e$, i.e., $\hat{P}_e r_{IS} \approx 10^{-a}$, where larger values for a indicate less improvement. Thus, the lower $\hat{P}_e$, the better the improvement factor. From Table 1, we see that a ranges from 2 (Gaussian with $M = 24$) to 7 (ATE with $M = 40$). This variation is indicative of the effectiveness of the underlying biasing scheme (e.g., translation) and is not a deficiency of the algorithm.

5.3 Algorithm Performance

The optimal biasing amount predicted by our algorithm was compared with analytical results for Gaussian noise available in [6] for a linear system with a memory of $M = 24$ and $P_e \approx 10^{-10}$. $N = 40$ decisions were observed in each simulation run. The estimated translation amounts were tightly clustered around the analytically known true $C_{opt}(j)$=1.0. After 100

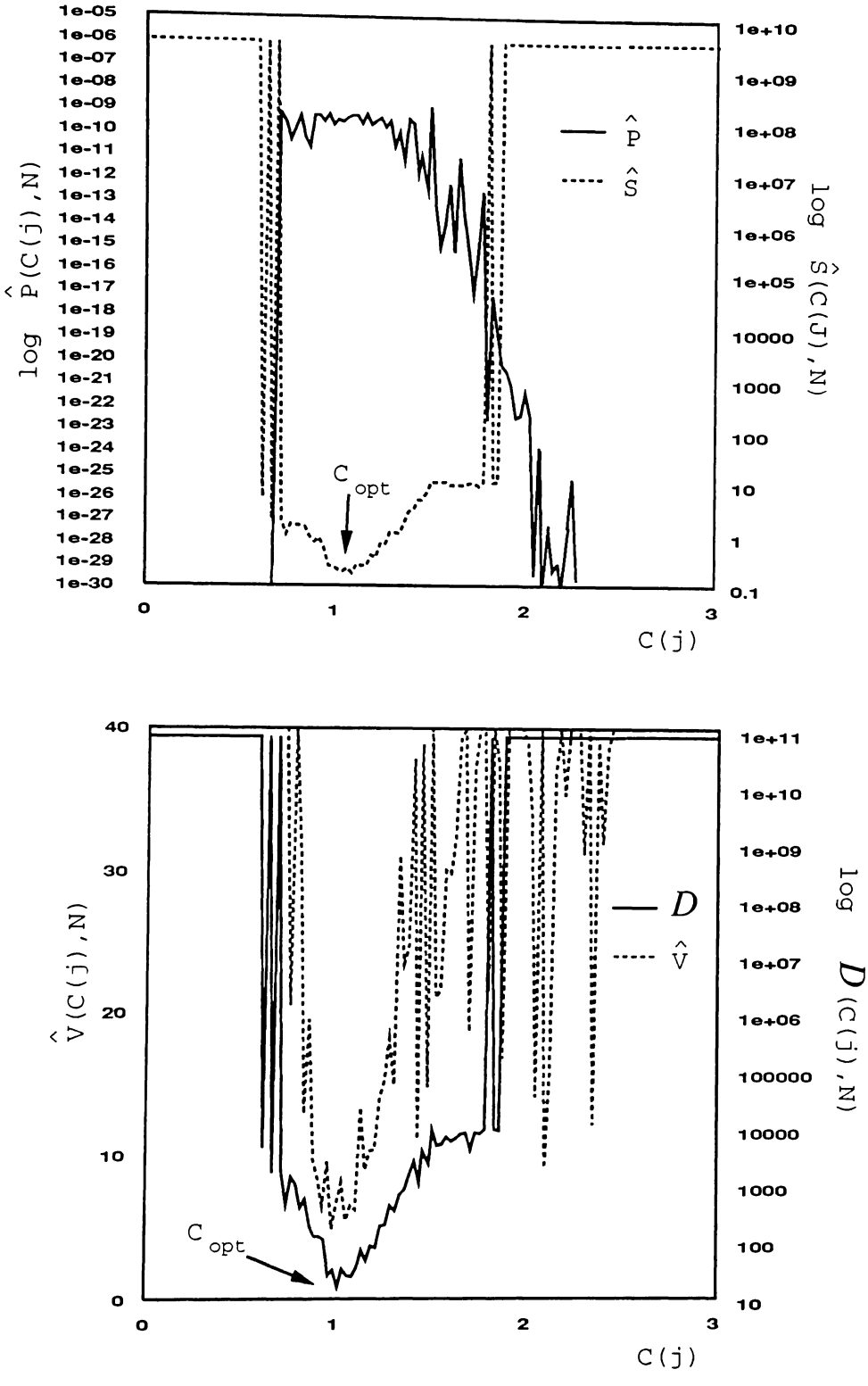


Figure 5: $\hat{P}$, $\hat{V}$, $\hat{S}$ and $\mathcal{D}$ vs. $C(j)$ for a Gaussian distribution, where $P_e \approx 10^{-10}$, memory length $M = 24$, realization number $j = 0$ and $N = 40$. Note that $C(j) = 1.0$ corresponds to the optimal translation found in [6].

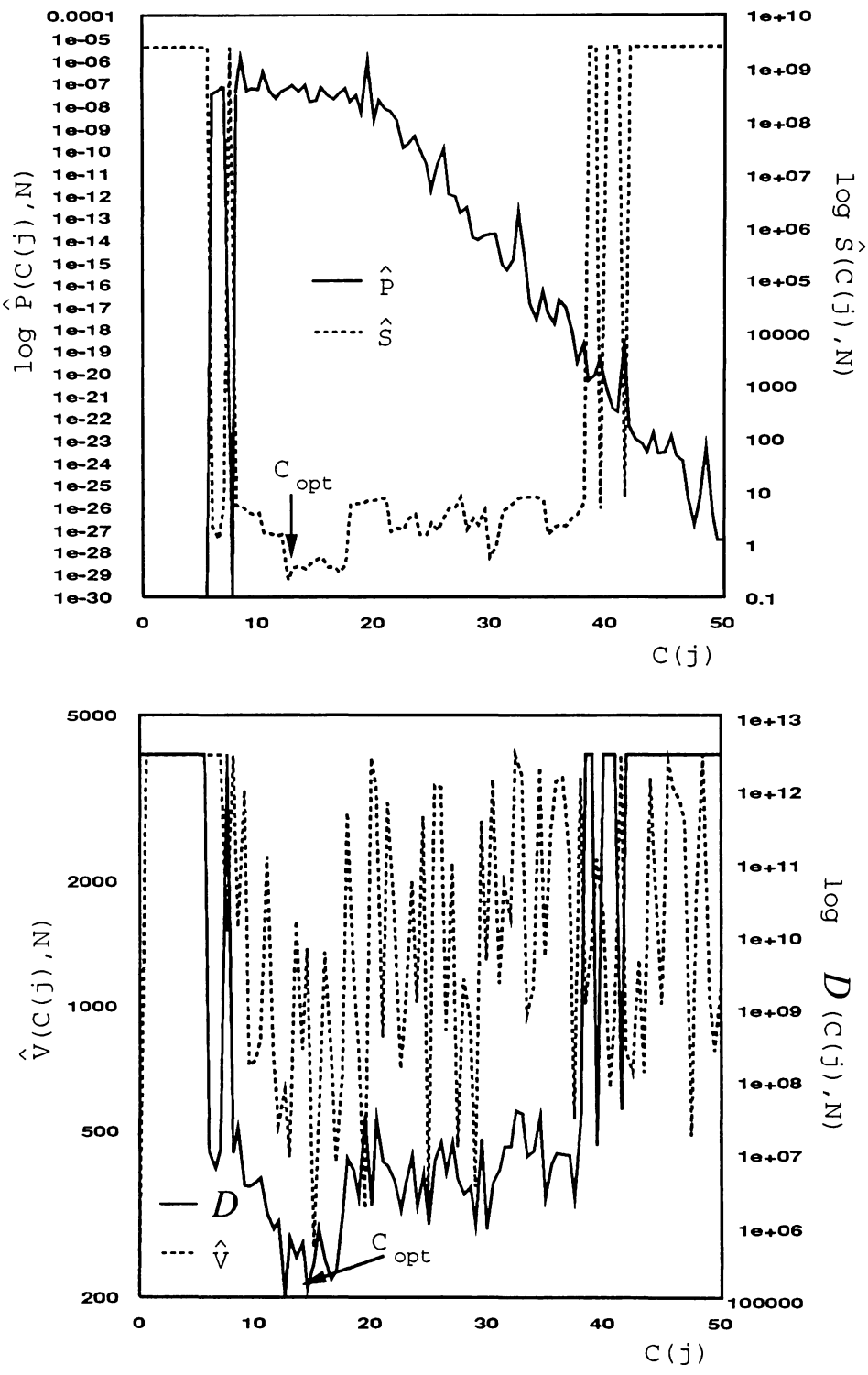


Figure 6: $\hat{P}$, $\hat{V}$, $\hat{S}$ and $\mathcal{D}$ vs. $C(j)$ for an ATE distribution, where $P_e \approx 10^{-8}$, memory length $M = 24$, realization number $j = 3$ and $N = 4,000$.

Distribution		M = 24		M = 40	
Gaussian	$\hat{P}_e$	1.1×10^{-8}	2.0×10^{-10}	7.0×10^{-8}	2.4×10^{-10}
(infinite support)	r_{IS}	1.1×10^{7}	5.8×10^{8}	4.5×10^{5}	7.4×10^{7}
ATE	$\hat{P}_e$	1.8×10^{-8}	1.3×10^{-10}	1.2×10^{-8}	2.4×10^{-10}
(infinite support)	r_{IS}	1.1×10^{2}	1.2×10^{4}	1.2×10^{2}	3.0×10^{3}
Rayleigh	$\hat{P}_e$	3.9×10^{-8}	1.6×10^{-10}	1.3×10^{-7}	9.2×10^{-10}
(semi-infinite support)	r_{IS}	2.7×10^{6}	3.9×10^{8}	1.8×10^{5}	2.0×10^{7}
WMC	$\hat{P}_e$	8.4×10^{-8}	3.0×10^{-9}	3.5×10^{-8}	1.3×10^{-9}
(semi-infinite support)	r_{IS}	1.4×10^{4}	1.1×10^{5}	1.2×10^{3}	6.8×10^{4}

Table 1: Estimated probabilities of error $\hat{P}_e$ and approximate improvement factors r_{IS} over Monte Carlo simulation, for the Gaussian, ATE, Rayleigh, and WMC distributions (linear system). The memory length in samples was $M = 24$ and $M = 40$.

Distribution		M = 24		M = 40	
Gaussian	$\hat{P}_e$	1.5×10^{-8}	1.1×10^{-10}	1.1×10^{-7}	1.5×10^{-9}
	r_{IS}	3.1×10^{6}	3.2×10^{8}	7.0×10^{5}	3.2×10^{7}
Rayleigh	$\hat{P}_e$	1.2×10^{-8}	3.7×10^{-10}	5.9×10^{-8}	5.4×10^{-10}
	r_{IS}	1.5×10^{6}	2.6×10^{7}	4.3×10^{5}	1.7×10^{7}

Table 2: Estimated probabilities of error $\hat{P}_e$ and approximate improvement factors r_{IS} over Monte Carlo simulation, for the Gaussian and Rayleigh distributions (nonlinear system). The memory length in samples was $M = 24$ and $M = 40$.

runs of the algorithm, the sample mean of the estimated optimal translation was 1.0575, with a sample standard deviation of 0.07636429, leading to a 99% confidence interval of (0.98607853,1.0254214). Thus the outstanding performance of the algorithm in terms of simulation variance reduction is combined with statistical evidence of agreement with existing analytical results.

6 Conclusions

We have presented a new method for finding an asymptotically-optimal set of IS biasing parameter values and demonstrated its applicability to the translation biasing scheme. The method differs from earlier techniques in that knowledge about the system is obtained from a series of short, repetitive, presimulation runs. This loosens the assumptions required about the system. Our method exploits a theoretically justifiable relationship, for small sample sizes, between the BER estimate and the amount of biasing. The algorithm optimizes statistical measures of performance asymptotically related to the sample variance of the BER estimate. The technique is mathematically justified, easy to implement, and is *not* subject to the usual restrictions, e.g., linear system with additive Gaussian noise.

As experimental validation, the optimal translation determined by the algorithm is in good agreement with the optimal derived analytically for a linear system with additive

Gaussian noise. In addition, simulation runtimes were reduced using this method by factors ranging from 10^2 to 10^8 for a variety of noise distributions and error probabilities.

It should be pointed out that the algorithmic approach presented in this paper is not restricted to the translation biasing scheme. A similar approach was applied to the "quasi-translation" scheme in [10]. A topic of concurrent research has also utilized this method for the variance modification scheme, in a case where the distributions were single-sided exponentials [11].

Acknowledgement

The authors would like to thank Dr. Heinz-Josef Schlebusch of Aachen University of Technology, Germany, for his insightful comments about this work, and in particular, on the formulation of the proof of Theorem 1. The authors also wish to thank the anonymous reviewers for their helpful comments and suggestions.

References

[1] P. Balaban. Statistical Evaluation of the Error Rate of the Fiberguide Repeater Using Importance Sampling. *Bell Systems Technical Journal*, 55(6):745–766, July-August 1976.

[2] K. S. Shanmugan and P. Balaban. A Modified Monte-Carlo Simulation Technique for the Evaluation of Error Rate in Digital Communication Systems. *IEEE Transactions on Communications*, COM-28(11):1916–1924, November 1980.

[3] R. L. Mitchell. Importance Sampling Applied to Simulation of False Alarm Statistics. *IEEE Transactions on Aerospace and Electronic Systems*, AES-17(1):15–24, January 1981.

[4] G. W. Lank. Theoretical Aspects of Importance Sampling Applied to False Alarms. *IEEE Transactions on Information Theory*, IT-29(1):73–82, January 1983.

[5] M. C. Jeruchim. Techniques for Estimating the Bit Error Rate in the Simulation of Digital Communication Systems. *IEEE Journal on Selected Areas in Communications*, SAC-2(1):153–170, January 1984.

[6] D. Lu and K. Yao. Improved Importance Sampling Technique for Efficient Simulation of Digital Communication Systems. *IEEE Journal on Selected Areas in Communications*, 6(1), January 1988.

[7] R. J. Wolfe, M. C. Jeruchim, and P. M. Hahn. On Optimum and Suboptimum Biasing Procedures for Importance Sampling in Communication Simulation. *IEEE Transactions on Communications*, COM-38(5):639–647, May 1990.

[8] D. Lu and K. Yao. On some New Importance Sampling Results for Simulation of Non-Linear Digital Communications Systems. In *Proceedings of the 1988 Conference on Information Sciences and Systems*, pages 336–340, 1988.

[9] M. Devetsikiotis and J. K. Townsend. A Useful and General Technique for Improving the Efficiency of Monte Carlo Simulation of Digital Communication Systems. In *Proceedings of IEEE GLOBECOM '90*, San Diego, CA, 1990.

[10] M. Devetsikiotis. A New and General Importance Sampling Technique for the Estimation of Bit Error Rates in Digital Communication Systems. Technical Report TR-90/3, Center for Communications and Signal Processing, North Carolina State University, June 1990.

[11] M. Devetsikiotis and J. K. Townsend. A Dynamic Importance Sampling Methodology for the Efficient Estimation of Rare Event Probabilities in Regenerative Simulations of Queueing Systems. In *Proceedings of IEEE International Conference on Communications, ICC '92*, Chicago, June 1992.

[12] H. J. Schlebusch. Nonlinear Importance Sampling Techniques for Efficient Simulation of Communication Systems. In *Proceedings of IEEE International Conference on Communications, ICC '90*, Atlanta, GA, 1990.

Appendix

A Underestimation Theorem

Let the random components X_i of the input random vector $\mathbf{X}$ be defined in the convex set S_i, with $\inf S_i = x_{iMIN} \in \mathbb{R} \cup \{-\infty\}$ and $\sup S_i = x_{iMAX} \in \mathbb{R} \cup \{+\infty\}$, $i = 1, \ldots, M$. If we denote the cumulative distribution function (CDF) of X_i by $F_{X_i}(x_i)$ then $F_{X_i}(x_{iMIN}) = 0$ and $F_{X_i}(x_{iMAX}) = 1$.

In forming $f_X^*(\mathbf{x})$, we assume that the pdf of each one of the random components X_i is biased *individually* through the *one-to-one* transformation $x_i^* = T_i(x_i, d_i)$, d_i a parameter. The value of the parameter $d_i \in [0, +\infty)$ specifies uniquely the IS transformation $T_i(\cdot, d_i)$, $i = 1, \ldots, M$, with *inverse* transformation given by $x_i = T_i^{-1}(x_i^*, d_i)$. The weight function $w(\mathbf{x})$ is substituted with $w_{\mathbf{d}}(\mathbf{x}) = \prod_{i=1}^{M} w_{d_i}(x_i)$, where $\mathbf{d}^T$ is the vector $[d_1, \ldots, d_M]$ and the random components $X_i, i = 1, \ldots, M$, of the random vector $\mathbf{X}$ are assumed independent.

Let W be a positive real number. For $i = 1, \ldots, M$, define the sets $S_{iW}(d_i) \subseteq S_i$ as $S_{iW}(d_i) \subseteq \{x_i : w_{d_i}(x_i) \leq W\}$ with $\inf S_{iW}(d_i) = x_{imin}(W, d_i) \in \mathbb{R} \cup \{-\infty\}$ and $\sup S_{iW}(d_i) = x_{imax}(W, d_i) \in \mathbb{R} \cup \{+\infty\}$. The IS estimate $\hat{P}^*(\mathbf{d}, N)$ is defined as $\hat{P}^*(\mathbf{d}, N) = \frac{1}{N} \sum_{j=1}^{N} I(g(\mathbf{x}_j)) w_{\mathbf{d}}(\mathbf{x}_j)$, with $\mathrm{E}[\hat{P}^*(\mathbf{d}, N)] = P$.

Empirical observations indicate consistently that, in many cases, employing IS parameter values that are much larger than the optimal (*over-biasing*) can lead to significant *apparent* underestimation of the expectation P. The following theorem provides formal (although conservative) justification for this behavior, under the given conditions. These conditions essentially define a class of biasing schemes where increasing d_i's induce a monotonic increase of the means and/or a monotonic "probability mass transfer" of the noise distributions involved. This class includes but is not restricted to the translation technique.

Theorem 1: If, $\forall\, i = 1, \ldots, M$, and for any $W > 0$, the following three conditions are met:

C1.1 $\exists\, d_W > 0$ such that we can find $S_W^i(d_i)$ that is non-empty and convex for $d_i > d_W$

C1.2 $\lim_{d_i \to +\infty} T_i^{-1}(x_{imin}(W, d_i), d_i) = x_{iMIN}$

C1.3 $\lim_{d_i \to +\infty} T_i^{-1}(x_{imax}(W, d_i), d_i) = x_{iMAX}$

Then, for any real $W > 0$ and any integer N, $\Pr[\hat{P}^*(\mathbf{d}, N) \leq W] \to 1$ as $d_i \to +\infty$, $i = 1, \ldots, M$. In other words, for any $W > 0$, any integer N and any $0 < P_{min} < 1$, there exists $\mathbf{d}_{min}$ with components $d_{imin} > 0$ such that, when $d_i > d_{imin}$, $i = 1, \ldots, M$, then $\Pr[\hat{P}^*(\mathbf{d}, N) \leq W] \geq P_{min}$.

PROOF: Let W be a real number with $W > 0$. Let P_{min} also be a real number with $0 < P_{min} < 1$. Clearly,

$$\begin{aligned} \Pr[\hat{P}^*(\mathbf{d}, N) \leq W] &= \Pr[\frac{1}{N} \sum_{l=1}^{N} I(g(\mathbf{x}_l)) w_{\mathbf{d}}(\mathbf{x}_l) \leq W] \\ &\geq \Pr[I(g(\mathbf{x}_l)) w_{\mathbf{d}}(\mathbf{x}_l) \leq W, \ l = 1, \ldots, N] \\ &= \Pr[I(g(\mathbf{x})) w_{\mathbf{d}}(\mathbf{x}) \leq W]^N \end{aligned} \tag{9}$$

where all probabilities are taken with respect to the modified pdf $f_X^*(\mathbf{x})$. In the last equation we have dropped index l, assuming instances of the random variable $I(g(\mathbf{x}))w_{\mathbf{d}}(\mathbf{x})$ are i.i.d. Since $I(g(\mathbf{x})) \leq 1$, $\Pr[I(g(\mathbf{x}))w_{\mathbf{d}}(\mathbf{x}) \leq W] \geq \Pr[w_{\mathbf{d}}(\mathbf{x}) \leq W]$. Thus,

$$\begin{aligned} \Pr[I(g(\mathbf{x}))w_{\mathbf{d}}(\mathbf{x}) \leq W] &\geq \Pr[w_{\mathbf{d}} \leq W] \\ = \Pr[\prod_{i=1}^{M} w_{d_i}(x_i) \leq W] &\geq \Pr[w_{d_i}(x_i) \leq W^{1/M}, i = 1, \ldots, M] \\ &= \prod_{i=1}^{M} \Pr[w_{d_i}(x_i) \leq W^{1/M}] \end{aligned} \tag{10}$$

For each $i \in \{1, \ldots, M\}$ we distinguish four cases based on conditions (C1.2), (C1.3):

CASE 1.1: $\exists d_{i1} > 0$ such that, when $d_i > d_{i1}$, $x_{imax}(W^{1/M}, d_i) < x_{iMAX}$ and $x_{imin}(W^{1/M}, d_i) > x_{iMIN}$. Then, $\exists \epsilon_i$ such that $0 < \epsilon_i \leq 1 - P_{min}^{1/NM}$, since $0 < P_{min} < 1$. Choose $X_{i1} \in S_i$ such that $F_{X_i}(X_{i1}) \geq P_{min}^{1/NM} + \epsilon_i$. Choose $X_{i2} \in S_i$ such that $F_{X_i}(X_{i2}) \leq \epsilon_i$. $\exists d_{i2} > 0$ such that $\forall\ d_i > d_{i2}$, $T_i^{-1}(x_{imax}(W^{1/M}, d_i), d_i) \geq X_{i1}$ (from (C1.2)). $\exists d_{i3} > 0$ such that $\forall\ d_i > d_{i3}$, $T_i^{-1}(x_{imin}(W^{1/M}, d_i), d_i) \leq X_{i2}$ (from (C1.3)).

For $d_i > d_{imin} = \max\{d_{i1}, d_{i2}, d_{i3}, d_W\}$, using (C1.1), $\Pr[w_{d_i}(x_i) \leq W^{1/M}] \geq \Pr[X_i \in S_{iW}(d_i)] = F_{X_i}^*(x_{imax}(W^{1/M}, d_i)) - F_{X_i}^*(x_{imin}(W^{1/M}, d_i)) = F_{X_i}(T_i^{-1}(x_{imax}(W^{1/M}, d_i), d_i)) - F_{X_i}(T_i^{-1}(x_{imin}(W^{1/M}, d_i), d_i))$. Then $F_{X_i}(T_i^{-1}(x_{imax}(W^{1/M}, d_i), d_i)) \geq F_{X_i}(X_{i1}) \geq P_{min}^{1/NM} + \epsilon_i$ and $-F_{X_i}(T_i^{-1}(x_{imin}(W^{1/M}, d_i), d_i)) \geq -F_{X_i}(X_{i2}) \geq -\epsilon_i$. It follows that $\Pr[w_{d_i}(x_i) \leq W^{1/M}] \geq P_{min}^{1/NM} + \epsilon_i - \epsilon_i = P_{min}^{1/NM}$.

CASE 1.2: $\exists d_{i1} > 0$ such that $x_{imax}(W^{1/M}, d_i) = x_{iMAX}$ and $x_{imin}(W^{1/M}, d_i) > x_{iMIN}$, for $d_i > d_{i1}$. Then, one only needs to choose $X_{i2} \in S_i$ such that $F_{X_i}(X_{i2}) \leq 1 - P_{min}^{1/NM}$. $\exists\ d_{i2} > 0$ (from (C1.3)) such that $\forall\ d_i > d_{i2}$, $T_i^{-1}(x_{imin}(W^{1/M}, d_i), d_i) \leq X_{i2}$. It follows that, for $d_i > d_{imin} = \max\{d_{i1}, d_{i2}, d_W\}$, $\Pr[w_{d_i}(x_i) \leq W^{1/M}] \geq \Pr[X_i \in S_{iW}(d_i)] = 1 - F_{X_i}(T_i^{-1}(x_{imin}(W^{1/M}, d_i), d_i)) \geq 1 - F_{X_i}(X_{i2}) \geq 1 - (1 - P_{min}^{1/NM}) = P_{min}^{1/NM}$.

CASE 1.3: $\exists d_{i1} > 0$ such that, when $d_i > d_{i1}$, $x_{imin}(W^{1/M}, d_i) = x_{iMIN}$ and $x_{imax}(W^{1/M}, d_i) < x_{iMAX}$. Then choose X_{i2} such that $F_{X_i}(X_{i2}) \geq P_{min}^{1/NM}$. $\exists$ (C1.2) $d_{i2} > 0$ such that, $\forall\ d_i > d_{i2}$,

$T_i^{-1}(x_{imax}(W^{1/M}, d_i), d_i) \geq X_{i2}$. For $d_i > d_{imin} = \max\{d_{i1}, d_{i2}, d_W\}$, $\Pr[w_{d_i}(x_i) \leq W^{1/M}] \geq \Pr[X_i \in S_{iW}(d_i)] = F_{X_i}(T_i^{-1}(x_{imax}(W^{1/M}, d_i), d_i)) \geq F_{X_i}(X_{i2}) \geq P_{min}^{1/NM}$.
CASE 1.4: $\exists d_{i1} > 0$ such that, when $d_i > d_{i1}$, $x_{imin}(W^{1/M}, d_i) = x_{iMIN}$ and $x_{imax}(W^{1/M}, d_i) = x_{iMAX}$. Then $\Pr[w_{d_i}(x_i) \leq W^{1/M}] = \Pr[X_i \in S_{iW}(d_i)] = 1 \geq P_{min}^{1/NM}$.
From the above four cases we conclude that, under the specified conditions (C1.1) - (C1.3), $\exists d_{imin} > 0$, $i = 1, \ldots, M$, such that $\Pr[w_{d_i}(x_i) \leq W^{1/M}] \geq P_{min}^{1/NM}$ when $d_i > d_{imin}$, $i = 1, \ldots, M$. For such d_i, $\prod_{i=1}^{M} \Pr[w_{d_i}(x_i) \leq W^{1/M}] \geq P_{min}^{1/N}$ and, from (9) and (10), $\Pr[\hat{P}^*(\mathbf{d}, N) \leq W] \geq \Pr[I(g(\mathbf{x}))w_{\mathbf{d}}(\mathbf{x}) \leq W]^N \geq (P_{min}^{1/N})^N = P_{min}$. **Q.E.D.**

B Proof that Pdfs Used Satisfy Conditions of Underestimation Theorem

B.1 Gaussian pdf

For the Gaussian i.i.d. case, the random component pdfs are given by: $f_{X_i}(x_i) = (2\pi\sigma_X^2)^{-1/2} \exp\left(\frac{-(x_i-\mu_X)^2}{2\sigma_X^2}\right)$ where σ_X^2 is the variance and μ_X is the mean value of the random variable X_i. Under translation, the biasing transformation $T_i(x_i, c_i)$ is given by $x_i^* = T_i(x_i, c_i) = x_i + c_i$ and the inverse transformation by $x_i = T_i^{-1}(x_i^*, c_i) = x_i^* - c_i$, where $c_i \in \mathbb{R}$ is the translation amount for the i-th component. The biased pdf is then $f_{X_i}^*(x_i) = f_{X_i}(x_i - c_i)$ and the weight function is

$$w_{c_i}(x_i) = \frac{f_{X_i}(x_i)}{f_{X_i}^*(x_i)} = \exp\left(\frac{c_i^2 - 2c_i(x_i + \mu_X)}{2\sigma_X^2}\right)$$

For $c_i > 0$, $w_{c_i}(x_i)$ is strictly decreasing, $S_{iW}(c_i) = \{x_i : w_{d_i}(x_i) \leq W\} = [x_{imin}(W, c_i), +\infty)$ for any $W > 0$, and $x_{imin}(W, c_i)$ is the solution of the equation $W = \exp\left(\frac{c_i^2 - 2c_i(x_i+\mu_X)}{2\sigma_X^2}\right)$. Therefore, $x_{imin}(W, c_i) = c_i^2 - 2\sigma_X^2 \ln W - 2c_i\mu_X/(2c_i)$ and $T_i^{-1}(x_{imin}(W, c_i), c_i) = x_{imin}(W, c_i) - c_i = -c_i^2 - 2\sigma_X^2 \ln W - 2c_i\mu_X/(2c_i)$. Finally, $\lim_{c_i \to +\infty} T_i^{-1}(x_{imin}(W, c_i), c_i) = -\infty = x_{iMIN}$, therefore the conditions of Theorem 1 are satisfied. The case $c_i < 0$ is similar and is omitted.

B.2 ATE pdf

For the ATE i.i.d. case, the random component pdfs are given by:

$$f_{X_i}(x_i) = \begin{cases} \exp\left(\frac{x_i}{a}\right), & x_i \leq 0 \\ \exp\left(\frac{-x_i}{b}\right), & x_i > 0 \end{cases}$$

where a, $b > 0$ and $a + b = 1$. Again, under translation, and for $c_i > 0$ the weight function is

$$w_{c_i}(x_i) = \begin{cases} \exp\left(\frac{c_i}{a}\right), & x_i \leq 0 \\ \exp\left(\frac{-x_i + bc_i}{ab}\right), & 0 < x_i \leq c_i \\ \exp\left(\frac{-c_i}{b}\right), & c_i < x_i \end{cases}$$

Clearly, since $w_{c_i}(x_i)$ is non-increasing, $S_{iW}(c_i) = \{x_i : w_{d_i}(x_i) \leq W\}$. When $W > 1$, $S_{iW}(c_i) = \mathbb{R}$ for $0 < c_i \leq a \ln W$ and $S_{iW}(c_i) = [x_{imin}(W, c_i), +\infty)$ for $c_i > a \ln W$. When

$W \leq 1$, $S_{iW}(c_i) = \emptyset$ for $0 < c_i \leq -b \ln W$ and $S_{iW}(c_i) = [x_{imin}(W, c_i), +\infty)$ for $c_i > -b \ln W$. Therefore, for any $W > 0$ and $c_i > \max\{a \ln W, -b \ln W\}$, $S_{iW}(c_i) = [x_{imin}(W, c_i), +\infty)$ is non-empty and $x_{imin}(W, c_i)$ is the solution of the equation $W = \exp\left(\frac{-x_i + bc_i}{ab}\right)$. Thus, for $c_i > \max\{a \ln W, -b \ln W\}$, $x_{imin}(W, c_i) = bc_i - ab \ln W$ and $T_i^{-1}(x_{imin}(W, c_i), c_i) = x_{imin} - c_i = (b-1)c_i - ab \ln W$. Therefore, since $b < 1$, $\lim_{c_i \to +\infty} T_i^{-1}(x_{imin}(W, c_i), c_i) = -\infty = x_{iMIN}$, and the conditions of Theorem 1 are satisfied. The case $c_i < 0$ is similar and is omitted.