A New and General Importance Sampling Technique for the Estimation of Bit Error Rates in Digital Communication Systems

Mihail Devetsikiotis

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

CCSP TR-90/3
June 1990
Abstract

DEVETSIKIOTIS-HADJINICOLAOU, MIHAIL. A New and General Importance Sampling Technique for the Estimation of Bit Error Rates in Digital Communication Systems. (Under the supervision of Dr. J. Keith Townsend)

Importance Sampling (IS) is recognized as a potentially powerful method for reducing runtimes when estimating the Bit Error Rate (BER) of digital communication systems using Monte Carlo simulation. The key to its effective implementation, however, is the choice of appropriate biasing parameters. In the past, analytical minimization of the variance of the IS estimator with respect to the biasing parameters has only led to solutions for systems which the BER could be found analytically. We present here a new technique for finding a near-optimal set of biasing parameters for the translation biasing scheme. The near-optimal translation values can be determined from repetitive, very short simulation runs by exploiting a theoretically justifiable relationship between the BER estimate and the amount of translation. Only mild assumptions are required of the noise distribution and system. Moreover, we extended the standard translation techniques to cover single-sided noise distributions by proposing a class of “quasi-linear” biasing schemes. Experimental results indicate that, using the technique presented here, improvement factors of up to eight orders of magnitude can be obtained for the Gaussian, Rayleigh, Asymmetrical Exponential and Avalanche Photodiode noise distributions, and linear or nonlinear systems.
Acknowledgements

I would like to thank my advisor, Dr. J. Keith Townsend for his guidance and warm support during the research effort that led to this thesis. I would also like to thank Dr. Sasan Ardalan and Dr. Yahya Fathi for serving on my advisory committee.
## Contents

### List of Figures

vi

### List of Tables

vii

### 1 Introduction

1.1 Statement of the Problem ........................................ 1
1.2 Background .................................................. 2
1.3 Summary and Perspective ........................................ 6

### 2 Model Description and Definitions

2.1 Model Description ................................................ 8
2.2 Probability of Error ............................................. 10
2.3 MC and IS Estimators ............................................ 12

### 3 Locating Near-Optimal IS Parameters

3.1 Overview ....................................................... 17
3.2 Memoryless System .............................................. 17
3.3 System with Memory: Direction of Biasing .................... 23
3.4 Nonlinear Systems .............................................. 28
3.5 System with Memory: Effects of Over- and Undertranslation . 35
3.6 Proposed method ............................................... 41
3.7 Summary ....................................................... 44

### 4 Importance Sampling for Single-Sided Distributions

4.1 Background ...................................................... 46
4.2 Importance Sampling Schemes .................................. 49
4.3 Summary ....................................................... 54
# Experimental Results

5.1 Simulation Set-up ........................................... 57
5.2 Linear Systems .............................................. 58
5.3 Nonlinear Systems .......................................... 64
5.4 Discussion .................................................. 67

# Conclusions ..................................................... 69

# Bibliography ................................................... 71
List of Figures

2.1 Model for baseband digital communications system .............. 8
2.2 Probability distributions of $Y_h$ under $H_0$ and $H_1$ hypotheses. The shaded area beyond the threshold $T$ corresponds to the probability of error under $H_0$ .................................................. 11
2.3 Original and biased pdf's, important region $\Omega_X$, and weight function $w_X(x)$ .......................................................... 14

3.1 Memoryless case: schematic illustration of underestimation proof, part 1 20
3.2 Memoryless case: schematic illustration of underestimation proof, part 2 21
3.3 A 3-dimensional input noise space with an impulse response $h$, a biasing direction $d$ and a decision surface $S$ ................................. 24
3.4 Linear system: a 2-dimensional illustration of an impulse response $h$, a biasing direction $d$ and a decision surface (hyperplane perpendicular to $h$) .................................................. 26
3.5 Discrete nonlinear system model based on a cascade of a linear system with memory and an instantaneous nonlinearity ........................ 29
3.6 An example of the mapping induced by an instantaneous nonlinearity 31
3.7 A 2-dimensional illustration of the mapping of important regions for a nonlinear system .................................................. 32
3.8 Mapping of the important region when the nonlinearity is monotonic (increasing) .................................................. 32
3.9 Input space and important regions when $\Omega_W = \{w : w < \tilde{T}_1 \text{ or } w > \tilde{T}_2\}$ 33
3.10 Input space and important regions when $\Omega_W = \{w : \tilde{T}_1 > w > \tilde{T}_2\}$ 34
3.11 System with memory: schematic illustration of underestimation proof, part 1 .................................................. 38
3.12 System with memory: schematic illustration of underestimation proof, part 2 .................................................. 39
3.13 Typical curve of $\hat{P}_e$ vs. $C$, for $0 \leq C \leq C_{\text{max}}$ .................................................. 41
3.14 Typical curves for various choices of number of decisions $N$ ........ 43
4.1 The WMC density distribution for three input levels (optical power) $P_1 < P_2 < P_3$: $P_1 = -52\text{dBm}$, $P_2 = -46\text{dBm}$ and $P_3 = -43\text{dBm}$.
The APD characteristics were: $G = 25$, $k = 0.8$, $\eta = 0.8$ and $\Delta t = 8.92 \times 10^{-10}$.

4.2 Three different quasi-linear biasing schemes: (a) Translation with uniform correction (b) Translation with "mirror-image" correction (c) Biasing by increasing the input level.

4.3 Nonlinear transformation ("quasi-linear" biasing) of a random variable.

4.4 Quasi-linear biasing applied to a Rayleigh distribution, for $b = 0.3$, $x_{\text{min}} = -2.0$, $\alpha = 0.5$ and $C = 0, 1, 2$.

4.5 Quasi-linear biasing applied to a WMC distribution, for $X = 696.59$, $\sigma = 595.92$, $\delta = 1.23$, $\alpha = 60$ and $C = 0, 1500, 3500$.

5.1 Impulse responses of the two linear, time-invariant filters that were used in the experiments.

5.2 Linear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the Gaussian distribution, where $P_e \approx 10^{-10}$, memory length $M = 24$ and realization number $j = 0$.

5.3 Linear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the Rayleigh distribution, where $P_e \approx 10^{-10}$, memory length $M = 40$ and realization number $j = 4$.

5.4 Linear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the ATE distribution, where $P_e \approx 10^{-8}$, memory length $M = 40$ and realization number $j = 9$.

5.5 Linear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the WMC distribution, where $P_e \approx 10^{-7}$, memory length $M = 40$ and realization number $j = 3$.

5.6 Impulse responses of the two linear filters that were used as parts of the nonlinear models.

5.7 Instantaneous nonlinearity used in the simulations.

5.8 Nonlinear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the Gaussian distribution, where $P_e \approx 10^{-7}$, memory length $M = 40$ and realization number $j = 5$.

5.9 Nonlinear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the Rayleigh distribution, where $P_e \approx 10^{-10}$, memory length $M = 24$ and realization number $j = 1$. 
List of Tables

5.1 Estimated probabilities of error and improvement factors over MC simulation: linear system .................................................. 62
5.2 Approximate runtimes in seconds, for the experimental simulations:
linear system ........................................................................ 62
5.3 Estimated probabilities of error and improvement factors over MC simulation: nonlinear system .............................................. 67
5.4 Approximate runtimes in seconds, for the experimental simulations:
nonlinear system .................................................................... 68
1

Introduction

1.1 Statement of the Problem

The probability of an error (or Bit Error Rate, BER) is the most commonly used measure of performance of digital communication systems. Various studies have addressed the problem of calculating the BER based on specific assumptions about the system and the random disturbances (noise or interference) involved [1, 2, 3]. Very often, however, the complexity of digital communications systems makes the analytical evaluation of the error probability very difficult. In fact, for systems other than the standard linear system with additive white Gaussian noise, analytical evaluation of the BER is virtually impossible unless restrictive assumptions are used.

Both numerical and simulation-based methods have been proposed to overcome this difficulty. One of the methods often used to obtain an estimate of the BER of a system is Monte Carlo (MC) simulation. However, the number of MC runs needed to estimate very low probabilities of error is extremely large, making the use of MC simulation impractical.

Among the various techniques that have been proposed for reducing run lengths of MC simulation, Importance Sampling (IS) techniques have attracted significant attention. The basic idea behind IS is to artificially increase the number of “important
events”, i.e., errors, that occur per simulation run by modifying the statistical properties of the noise process involved. This is generally referred to as “biasing” of the noise process. The artificially increased error count is then corrected or “unbiased”. Under the proper conditions, the above can result in a significant reduction of the estimator variance or, equivalently, in significant reduction of the required simulation length.

We provide in this work a new technique for finding near-optimal parameter settings for IS that can be used for the efficient estimation of the BER of digital communication systems. The significance of this technique is that it is simple and general, that is, it does not require most of the assumptions used in previous work (e.g., linear system and/or normally distributed noise).

1.2 Background

IS is a statistical technique known by simulation practitioners for quite a while [4, 5]. However, its potential usefulness in the analysis of communication systems has been explored only recently.

A first attempt to apply IS to the estimation of the BER of a digital communication system was presented by P. Balaban in [6]. The author proposed using IS to reduce the runtime of MC simulation of Shot-noise-limited fiberguide repeaters with an Avalanche Photodiode Detector (APD). Some fundamental concepts of IS like “important regions” and relationships between input and output weights were discussed. However, the practical significance of that work was limited by the fact that the problem of handling Intersymbol Interference (ISI) due to system memory was not addressed.
The first detailed examination of biasing and unbiasing procedures for IS and the effects of system memory and ISI was provided by K. S. Shanmugan and P. Balaban in [7]. In this paper the authors introduced the fundamental theoretical aspects of IS with respect to a baseband digital communication system with memory. Assuming additive Gaussian noise, they proposed a biasing scheme that was based on multiplying each noise random deviate by an appropriately chosen scalar constant, therefore increasing the noise variance by the same factor. The scalar constant used was optimal for each case, in the sense that it was chosen to maximize the runtime savings over MC simulation. The effects of system memory length and its approximate identification on the mean and the variance of the IS estimator were also discussed. Finally, simulation results showing substantial savings over MC simulation were given.

Two of the earliest papers examining the theoretical aspects of IS and its application to the analysis of communication and radar systems were [8] and [9]. R. L. Mitchell in [8] explored some of the theoretical aspects of MC simulation and IS, and presented an extensive series of examples involving one- and multi-dimensional probability distributions, where IS was shown to offer significant advantages over conventional MC simulation. G. W. Lank in [9] presented a detailed and mathematically rigorous description of IS and its application to the estimation of the probability of rare events. He developed conditions for the IS approach to be valid and, more importantly, to improve the efficiency of the simulation. Moreover, he presented bounds on the statistics of the IS estimator and studied the asymptotic behavior of the improvement factor. Experimental results showed the great potential for runtime savings offered by IS. Both authors above, emphasized mostly the usefulness of IS in radar applications and, like the authors in [7], adopted the IS scheme where the noise variance is multiplied by a scalar.
M. C. Jeruchim in [10] gave, among others, a complete description of the statistical properties of the IS estimator and outlined the biasing/unbiasing procedures used in IS.

The same author in [11] extended the analysis of IS to include multihop links and addressed the effects of phase and timing estimation on the performance of IS techniques.

Q. W. Wang and V. K. Bhargava in [12] gave a more general theoretical description of IS in a different notation, provided useful insight in some practical aspects of IS, and, based on the — trivial but impractical — global optimal biasing strategy, proposed a heuristic scheme that dealt with the difficulties induced by the insufficient knowledge about the system.

A basic disadvantage of biasing by increasing the variance of the noise [7, 9, 11] is the significant reduction of its efficiency as the memory of the system increases. In fact, it seems that increased system memory generally tends to reduce the improvement induced by any IS scheme. B. R. Davis in [13], proposed a method to reduce the dimensionality of the problem to unity so that the negative effects of memory would be eliminated. Although theoretically this method could be used for a large variety of systems and noise distributions, it seems that, for practical purposes, its application is limited to linear systems with additive Gaussian noise.

In [14] P. M. Hahn and M. C. Jeruchim examined the application of IS to systems with more than one input noise processes, determined the effects of the truncation of system memory on the IS estimator and analysed extensively the case where the noise is Gaussian, based on new concepts like the STC (System Threshold Characteristic) and the SBC (System BER Characteristic). They showed that, for a linear system with Gaussian noise, the problem dimensionality can be reduced to unity. Moreover,
they proposed using statistical regression to find the “equivalent impulse response” of a nonlinear system.

More recently a technique based on translating the original noise probability density function (pdf) has been proposed by D. Lu and K. Yao [15]. This technique has been shown to result in much larger improvement than conventional IS, at least for Gaussian noise. The authors based their development on a different framework that included conditioning on the possible realizations of the input data vector. This led to an alternate implementation of IS, commonly referred to as the “block approach”.

Wolfe et al. in [16], analyzed IS from a new perspective (using the STC notion from [14]), rederived the (impractical) global optimal biasing rule and provided useful insight in sub-optimal biasing procedures. They considered again the translation technique of [15] for a Gaussian pdf, showed that it was close, in a certain way, to the global optimal and demonstrated the independence of the improvement from system memory, always for the case of a linear system with Gaussian noise.

Importance Sampling has also attracted significant attention in contexts other than communication link performance analysis, like simulation of Markov chains [17], simulation of digital decoders [18] and simulation of queueing systems [19]. The last application is of very special interest since there exists today a growing need for analysis and design tools in the booming field of information networks.

Although in this work we focus our attention to IS techniques, other simulation-based methods for the efficient estimation of the BER of communication systems (or the probability of rare events, in general) also exist. A comprehensive survey of these methods was presented in [10]. Moreover, numerical methods have been also used successfully. Representative work in this field can be found in [20, 21, 22, 23].
1.3 Summary and Perspective

Choosing the parameters for IS in an optimal (or nearly optimal) way is the most basic problem in actually applying any technique similar to the above. This is usually done by formulating the variance of the estimator and finding the parameter settings that minimize it. The analytical calculations involved in finding the expression for the variance and minimizing it are far from trivial for almost any realistic noise pdf except the Gaussian. In general, for a system memory higher than unity, the problem is no more tractable than the original problem of finding the BER analytically [14, 24].

A way to overcome this difficulty, based on using bounds on the estimator statistics, was proposed in [24]. However this method still requires analytical calculations that can potentially become intractable.

Clearly, methods that estimate the proper parameters for IS without relying on the usual, analytically intractable, estimator variance minimisation approach are needed. One such method that is general, in the sense that the noise need not be Gaussian, and relatively simple to implement is presented here. The method is based on repetitive very short simulation runs and can determine parameter settings that are close to optimal, resulting in large savings in simulation length.

It is easy to show that the "translation" or "linear shift" technique, although potentially very powerful, cannot be applied, in general, to input noise processes that have single-sided pdf's. In this case, a class of nonlinear transformations that resemble translation is shown here to result in significant run time improvements. These "quasi-translation" techniques, combined with the proposed method to choose the parameters, can be very useful for the estimation of BER in certain communications systems like Shot-noise-limited lightwave communication links.
In Chapter 2 we describe the general model of the system under study. We also provide the basic definitions and assumptions used. In Chapter 3 we present and analyze an experiment-oriented method to determine the optimal parameters for IS based on translating the noise pdf. Both linear and nonlinear systems are discussed. In Chapter 4 we introduce a "quasi-translation"-based IS technique for noise processes with single-sided pdf's. Finally, in Chapter 5 we present complete simulation results and tables demonstrating the applicability and usefulness of the proposed method.
2

Model Description and Definitions

2.1 Model Description

Our goal is to estimate the probability of an error in the baseband digital communications system shown in Figure 2.1.

![Diagram of baseband digital communications system](image)

Figure 2.1: Model for baseband digital communications system

Most of our notation will be based on that in [10] and in [15]. We will assume
that the uncorrupted input signal has the form

\[ S(t) = \sum_{i} a_i p(t - iT_i) \]

where \( \{a_i\} \) can be \( A \) or \( -A \), \( p(t) \) is a rectangular pulse shape of unit amplitude and duration \( T_i \). Although we restrict here our attention to binary communications systems only, extension to \( M \)-ary systems is also possible. If \( n(t) \) is a noise process then, the input \( z(t) \) to the system can be described in general by

\[ z(t) = Q(S(t), n(t)) \]

where \( Q \) is some transformation that combines the signal and the noise into a composite random process. Many times, the noise is assumed to be i.i.d. and additive, that is,

\[ z(t) = S(t) + n(t) \]

This assumption is somewhat relaxed in the following development, in the sense that, although the final effect of noise is assumed to be additive, the noise characteristics can depend on the signal.

The output of the system is sampled every \( T \) units of time. The resulting output sequence \( \{Y_k\} \) will be given by \( Y_k = g(X) \), where \( g(\cdot) \) is the system transfer characteristic (system response) and vector \( X \) is defined as

\[ X = [X_k, X_{k-1}, \ldots, X_{k-M+1}] = Q(A, N) \]

where

\[ A = [A_k, A_{k-1}, \ldots, A_{k-M+1}] \]

and

\[ N = [N_k, N_{k-1}, \ldots, N_{k-M+1}] \]
A is the input data vector, N is the noise vector and M is the system memory length in samples. \( A_k \) takes values \( A \) (under hypothesis \( H_1 \)) and \(-A\) (under \( H_0 \)). The actual values that these random vectors can take will be denoted by \( x, a \) and \( n \).

### 2.2 Probability of Error

Throughout this paper, we will restrict our attention to the \( H_0 \) hypothesis. In many cases the probabilities of error \( P_{e_0} \), under \( H_0 \), and \( P_{e_1} \), under \( H_1 \), are equal: \( P_{e_0} = P_{e_1} = P_e \). Even if this is not true, the extension of the present development to cover \( H_1 \) is straightforward and is omitted. In the following, we will refer to the "probability of error under \( H_0 \)", \( P_{e_0} \), simply as the "probability of error" \( P_e \) or BER.

Let \( f_X(x) \), \( f_A(a) \) and \( f_N(n) \) be the pdf's of \( X \), \( A \) and \( N \) respectively. In order to account for the more general case where the characteristics of the noise depend on the transmitted symbol, let \( f_{X|A}(x) \) and \( f_{N|A}(n) \) be the pdf's of \( X \) and \( N \), respectively, conditioned on the data vector \( A \).

We assume that \( \{Y_k\} \) is compared with a threshold \( T \) in order to make a decision on whether a 0 or a 1 was sent. Then, the BER of the system is given by

\[
P_e = \int_{-\infty}^{\infty} I_T(y) f_Y(y) \, dy
\]

where \( I_T(y) \) is an indicator function equal to 1 for \( y \geq T \) and to 0 for \( y < T \). Figure 2.2 illustrates the above.

In the simulation context it is very important that we can write Eq.(2.1) as

\[
P_e = \int_{-\infty}^{\infty} I_T(g(x)) f_X(x) \, dx
\]

since it is the input distributions that we know and control.
We can also write (2.2) as

\[ P_e = \int_{-\infty}^{\infty} I_T(g(a,n)) f_{A,N}(a,n) \, da \, dn \]

\[ = \int_{-\infty}^{\infty} I_T(g(a,n)) f_{A}(a) f_{N|A}(n) \, da \, dn \]  \hspace{1cm} (2.3)

A "realization" is a block of input bits (symbols) with length equal to the system memory in bits, \( K \) (\( M = K \times \text{samples/bit} \)). It represents a specific combination of these \( K \) bits (recall that each bit can take values \( A \) or \(-A\)). Letting all possible realizations \( a(j), j = 0, 1, \ldots, 2^{K-1} - 1 = J - 1 \), of \( A \) be equiprobable, (2.3) becomes

\[ P_e = \frac{1}{J} \sum_{j=0}^{J-1} P_e(j) \]  \hspace{1cm} (2.4)

where

\[ P_e(j) = E[I_T(g(a(j) + n))] \]
\[ = \int_{-\infty}^{\infty} I_T(g(a(j), n)) f_{N|A}(n) dn \] (2.5)

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

### 2.3 MC and IS Estimators

In MC simulation, (2.1) is estimated by

\[
\hat{P}_e = \frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} I_T(y_i) 
\] (2.6)

where \( y_i = g(x_i) \) and \( \bar{N} \) is the number of decisions used in the simulation. In this “sequential” simulation model, a decision is made once for every sequence of samples with length equal to the system memory, that is, one decision corresponds to \( M = K \times \text{samples/bit} \) samples — spacing the decisions at least \( M \) samples from one another is necessary so that individual decisions are mutually independent.

An estimator for (2.3), corresponding to the so-called “block” approach, is from [15]:

\[
\hat{P}_e = \frac{1}{\bar{N}} \sum_{j=0}^{J-1} \sum_{i=1}^{N/J} I_T(g(a(j), n(j,i))) 
\] (2.7)

where \( j \) represents the conditioning on every realization \( a(j) \) of \( A \), and \( N/J \) decisions are used for each realization. In the block simulation model, a decision is made once in each realization or “block”, again to ensure independence of individual observations. The number of samples per realization equals the length of memory in samples.

When applying Importance Sampling, we modify either \( f_X(x) \) or \( f_N(n) \). Let the new random vectors be \( X^* \) and \( N^* \), and the modified (biased) pdf’s be \( f_X^*(x) \) and
Then, $P_e$ can be written as

$$P_e = \int_{-\infty}^{\infty} I_T(g(x^*)) w_X(x^*) f_X^*(x^*) dx^* \quad (2.8)$$

or as

$$P_e = \int_{-\infty}^{\infty} I_T(g(a, n^*)) f_A(a) w_{N|A}(n^*) f_{N|A}^*(n^*) da dn^*$$

$$= \frac{1}{J} \sum_{j=1}^{J} \int_{-\infty}^{\infty} I_T(g(a(j), n^*)) w_{N|j}(n^*) f_{N|j}^*(n^*) dn^* \quad (2.9)$$

where

$$w_X(x^*) = \frac{f_X(x^*)}{f_X^*(x^*)}$$

$$w_{N|A}(n^*) = \frac{f_{N|A}(n^*)}{f_{N|A}^*(n^*)}$$

and

$$w_{N|j}(n^*) = \frac{f_{N|j}(n^*)}{f_{N|j}^*(n^*)}$$

and $j$ denotes the conditioning on realization $a(j)$. The functions $w_{N|A}(n^*)$, $w_{N|j}(n^*)$ and $w_X(x^*)$ are commonly referred to as the weight functions. These notions are illustrated in Figure 2.3.

Note that $f_X^*(x)$ must be nonzero in $\Omega_X = \{x \in R^M | I_T(x) = 1\}$. Also, $f_N^*(n)$ must be nonzero in $\Omega_N = \{n \in R^M | I_T(a, n) = 1\}$. This constraint follows from the definition of the weight functions and the form of equations (2.8) and (2.9). From a practical standpoint, violation of this condition will lead to a biased estimate, that is, the expected value of the estimate will not any more be equal to the true $P_e$.

The sets $\Omega_X$ and $\Omega_N$ are usually called the "importance regions" in $X$- and $N$-space, respectively. As their name suggests, elements in these sets are the ones that
Figure 2.3: Original and biased pdf's, important region $\Omega_X$, and weight function $w_X(x)$

attract our attention, since they correspond to the occurrence of errors. In attempting to increase the efficiency of MC simulation by using IS, one must increase the probability that the random variable of interest will "fall" in the importance regions. Although the importance regions are usually known in $Y$-space (the output space), they are more difficult to describe analytically in $X$- or $N$-space (the input spaces).

Under IS, the estimators in (2.6) and (2.7) become respectively

$$\hat{P}_e^* = \frac{1}{N_*} \sum_{i=1}^{N_*} I_T(g(x_i^*)) \, w_X(x_i^*)$$ \hspace{1cm} (2.10)

and

$$\hat{P}_e^* = \frac{1}{N_*} \sum_{j=0}^{J-1} \sum_{i=1}^{N_*/J} I_T(g(a(j), n^*(j,i))) \, w_{X|j}(x_i^*)$$ \hspace{1cm} (2.11)

when $N_*$ decisions are used for the simulation.

14
Usually, only the noise pdf is biased. This implies that \( w_{X|i}(x_k^i) = w_{N|i}(n_k^i) \), for any \( k \). Furthermore, assuming the noise samples to be mutually independent:

\[
w_{X|i}(x_k^i) = \prod_{i=0}^{M-1} \frac{f_{N|i}(n_{k-1}^i)}{f_{N|i}(n_k^i)} = w_{N|i}(n_k^i)
\]

(2.12)

for all \( k \). Equations (2.11) and (2.12) are the basis for the block implementation of IS.

It is easy to show that the estimator in (2.11) is unbiased (if and only if the condition imposed earlier on the modified pdf is satisfied):

\[
E [\hat{P}_e^*] = P_e
\]

(2.13)

For the variance \( \sigma_{IS}^2 = E \left[ (\hat{P}_e^*)^2 \right] - (P_e)^2 \) of this estimator, it was shown in [15] that

\[
\sigma_{IS}^2 = \frac{1}{JN_s} \sum_{j=0}^{J-1} \int_{-\infty}^{\infty} I_T(g(a(j), n)) f_{N|i}(n) \left[ w_{N|i}(n) - P_e(j) \right] dn
\]

\[
= \frac{1}{JN_s} \sum_{j=0}^{J-1} I_{IS}(j)
\]

(2.14)

where the conditioning on realization \( j \) was added here in order to account for the dependence of the noise on the input sequence.

The customary approach to finding the optimal IS parameters, is to attempt to minimize analytically the estimator variance or, equivalently its components \( I_{IS}(j) \) in (2.14), with respect to those parameters. Note that, although “block” simulation does not require more runtime (in both the “block” and the “sequential” approaches one decision per \( K \) bits is taken), it does require more optimization work, since \( J \) instead of one IS parameters are involved. However, the additional degrees of freedom
due to the conditioning on realizations, result in significantly larger reduction of the estimator variance.
3

Locating Near-Optimal IS Parameters

3.1 Overview

The technique we propose here is based on the observation that, when IS is applied to the problem of estimating $P_e$ and the input pdf is over-translated, the resulting estimate $\hat{P}_e$ is lower than the true $P_e$ with high probability. This may seem to be contradicting the well known fact that IS gives an unbiased estimate of $P_e$ (Eq. 2.13). However, the fact that the estimator is unbiased only means that the average over an infinite number of runs will always be equal to the true $P_e$. Intuitively, in the case of over-translation, the mode of the modified pdf will be located far away from the mode of the original pdf and in a region where the weight values are orders of magnitude smaller than $P_e$. The occurrence of a random deviate corresponding to a weight large enough to “correct” the average is, under these conditions, a rare event. Therefore, when the number of decisions is small, most of the time only very small weights will be generated, resulting in an estimate much lower than the true $P_e$.

3.2 Memoryless System

Recall now the system in Figure 2.1 and the corresponding description in the previous chapter. Consider first the case of a system with no memory and no signal (i.e.,
\( M = 1, S(t) = 0 \) and \( X(t) = Y(t) \). When the technique of "translation" or "linear shift" is used, \( f_Y \) is

\[
f_Y(y) = f_Y(y - C)
\]

Our purpose is to determine the optimal value for \( C \), without resorting to analytical minimization of the estimator variance in (2.14). We will show, under mild conditions, that for a given length of simulation, there exists an amount of translation \( C \) beyond which IS-based simulation underestimates the \( P_e \) by an arbitrary amount, with arbitrarily high probability.

Theorem 1. Let the true probability of error, under the \( H_0 \) hypothesis, be \( P_e \) and the IS estimator (when translation is used) be \( \hat{P}_e \). For any given \( w_{\text{max}} = \sigma P_e \), where \( 0 < \sigma < 1 \), any number of decisions \( N \) and any arbitrary \( P_{\text{min}} \), a translation amount \( C_{\text{min}} \) exists such that, for \( C = C_{\text{min}} \):

\[
\Pr[\hat{P}_e \leq w_{\text{max}}] = P_{\text{min}}
\]

Furthermore, for any \( C \geq C_{\text{min}} \), \( P_e \) will be underestimated by an even larger amount, i.e., \( \hat{P}_e \leq w'_{\text{max}} \leq w_{\text{max}} \), with probability \( P_{\text{min}} \).

**Proof:** Let \( w(y, C) \) represent the weight function

\[
w(y, C) = f_Y(y)/f_Y^*(y) = f_Y(y)/f_Y(y - C)
\]

and \( I_{T,k} \) be the indicator or counting function (equal to 1 if and only if \( y_k > T \) and equal to 0 otherwise). Since \( w(y_k, C) > 0 \) and \( 0 \leq I_{T,k} \leq 1 \),

\[
\hat{P}_e = \frac{1}{N} \sum_{k=1}^{N} I_{T,k} w(y_k, C) \leq \frac{1}{N} \sum_{k=1}^{N} \max \{ w(y_k, C) \} = \max \{ w(y_k, C) \} \quad (3.1)
\]
for any translation amount \( C \). In order to prove the first part of the theorem above, we only need to prove that there exists a \( C_{\min} \) such that

$$
\Pr[\max \{ w(y_k, C) \} \leq w_{\max}] = P_{\min}
$$

for \( C = C_{\min} \).

Let \( y = w^{-1}(W, C) \) be the value of \( Y \) such that

$$
w(y, C) = f_Y(y)/f_Y(y - C) = W
$$

for some weight \( W \). That is, \( w^{-1}(W, C) \) is the inverse weight function parameterized by the translation parameter \( C \). Furthermore, when the translation parameter is equal to \( C \), let \( P_A(C, W) \) be the probability that a single decision will correspond to a weight greater than \( W \), and \( P_B(C, W) \) be the probability that all \( N \) decisions will correspond to weights less than \( W \). Since decisions are independent,

$$
P_B(C, W) = (1 - P_A(C, W))^N
$$

Then, assuming that \( w(y, C) \) is non-increasing with \( y \), \( C_{\min} \) above exists and can be found as the minimum \( C \) that satisfies the tail probability equation

$$
P_A(C, w_{\max}) = 1 - \sqrt[5]{P_{\min}} \quad \quad (3.2)
$$

or, equivalently,

$$
\int_{-\infty}^{w^{-1}(w_{\max}, C)} f_Y(y - C) \, dy = 1 - \sqrt[5]{P_{\min}} \quad \quad (3.3)
$$

Clearly, when \( C_{\min} \) is chosen this way,

$$
P_B(C_{\min}) = (1 - P_A(C_{\min}))^N = P_{\min}
$$
and max \{ w(y_k, C_{\text{min}}) \} \leq w_{\text{max}} \text{ with probability equal to } P_{\text{min}}. \text{ Then,}

\[
\hat{P}_e = \frac{1}{N} \sum_{k=1}^{N} I_{T,k} w(y_k, C_{\text{min}}) \leq \frac{1}{N} \sum_{k=1}^{N} \max \{ w(y_k, C_{\text{min}}) \} \leq w_{\text{max}} \quad (3.4)
\]

with probability equal to \( P_{\text{min}} \) and, thus, we have proved the first part of the theorem. Figure 3.1 gives a schematic illustration of the above.

Figure 3.1: Memoryless case: schematic illustration of underestimation proof, part 1

To prove the second part of the theorem let \( y_{\text{wmax}} = w^{-1}(w_{\text{max}}, C_{\text{min}}) \) and \( b \) be any positive number. According to the first part, for \( C = C_{\text{min}} + b \geq C_{\text{min}} \),

\[
\max \{ w(y_k, C_{\text{min}} + b) \} \leq w'_{\text{max}} = w(y_{\text{wmax}} + b, C_{\text{min}} + b)
\]

with probability equal to \( P_{\text{min}} \). Our only assumption about the pdf is that \( f_Y(y+b) \leq f_Y(y) \) (i.e., the pdf is non-increasing in the tails), which is not restrictive for most realistic pdf's. It follows that

\[
w'_{\text{max}} = w(y_{\text{wmax}} + b, C_{\text{min}} + b)
\]
This implies that, for \( C \geq C_{\text{min}} \) and with probability \( P_{\text{min}} \), the probability of error will be underestimated by an even larger amount than in (3.4). Figure 3.2 illustrates the above.

\[
\begin{align*}
\frac{f_Y(y_{w_{\text{max}}})}{f_Y(y_{w_{\text{max}}})} &= \frac{f_Y(y_{w_{\text{max}}})}{f_Y(y_{w_{\text{max}}})} \\
&= \frac{f_Y(y_{w_{\text{max}}})}{f_Y(y_{w_{\text{max}}})} \\
&\leq \frac{f_Y(y_{w_{\text{max}}})}{f_Y(y_{w_{\text{min}}})} \\
&= w(y_{w_{\text{max}}}, C_{\text{min}}) = w_{\text{max}}
\end{align*}
\]

Figure 3.2: Memoryless case: schematic illustration of underestimation proof, part 2

Note that equation (3.1) is always true for the maximum weight occurring in each simulation. That is, the estimate \( \hat{P}_e \) is always bounded from above by the maximum value the weight function takes: \( \hat{P}_e \leq \max \{ w(y_k, C) \} \). In fact, Lu and Yao in [24] have used similar arguments when they proposed analytically minimizing a bound on the estimator variance instead of the variance itself. In our case, the crux of the
presented arguments is that over-translation causes the bound $w_{\text{max}}$ on the maximum occurring weight to be smaller than the true $P_e$ most of the time, therefore resulting in underestimation of $P_e$. The slight complication of the proof stems from the fact that not only the probability of a single random weight being greater than a preset maximum $w_{\text{max}}$ (i.e., the tail probability $P_A(C, w_{\text{max}})$) must be low, but also the probability $P_B(C, w_{\text{max}})$ of all weights being less than $w_{\text{max}}$ must be higher than a given $P_{\text{min}}$.

We have shown that over-translation will result in underestimation with high probability by showing that, under mild conditions, given an amount of underestimation and a probability (close to 1) that this underestimation should occur, one can find an amount of translation that will underestimate at least as much, with a probability at least as high. The arguments that were used constitute a rather conservative proof and therefore, the resulting bounds on parameter $C$ are accordingly loose. In practice, underestimation will be observed consistently for smaller values of $C$, as verified by experiments.

As an example, consider the case where the noise pdf is Gaussian with mean zero and $\sigma = 1$. Then, for a decision threshold $T = 5.6$, the true $P_e$ will be $P_e \approx 10^{-8}$. Assume translation by $C$. Let $N = 50,000$. We set $w_{\text{max}} = 10^{-10}$ and $P_{\text{min}} = 1 - 10^{-6}$. Solving (numerically) (3.3) for $C_{\text{min}}$ we find $C_{\text{min}} \approx 16$. This means that, for $C \geq 16$ our estimate will be less than $w_{\text{max}} = 10^{-10}$ with a probability of at least $P_{\text{min}} = 1 - 10^{-6}$.
3.3 System with Memory: Direction of Biasing

We now focus our attention to the general multidimensional problem (i.e., when memory and signal are present and where the input distributions are biased) which is the one of practical importance. We assume that the "block" approach of (2.7) will be used. That is, we assume that the problem is to obtain the optimal $c(j)$ for each realization $a(j), j = 0, \ldots, J - 1$ of $A$. The modified noise pdf will be

$$f^*_N(n) = f_N(n - c(j)) \quad (3.5)$$

This implies that the IS estimator is given by (2.11). This approach has been used successfully in [15] for the case of additive Gaussian noise and has been shown to be superior to the conventional approach where no explicit conditioning on "realizations" is taken into account.

From a practical standpoint, the amount of effort required to obtain the $J \times M$ IS parameters $c(j,i), i = 1, \ldots, M$ implied by (3.5) seems unrealistically large — especially since analytical methods would not be helpful in the general case. Instead, letting the bias be along some direction $d$ that remains fixed for all realizations, that is, letting $c(j) = C(j) d$ for all $j$, reduces the degrees of freedom by a factor of $M$, thus reducing the problem to that of finding $J$ parameters $C(j)$, one for each realization. Clearly, unless the optimal biasing direction is known somehow in advance, there is a trade-off between parameter estimation effort and estimator efficiency. Our experimental results indicate that by biasing in a single direction, large improvement factors are still achievable, provided that this direction that is chosen beforehand is meaningful in some way. A 3-dimensional input noise space with the impulse response direction $h$, a generic biasing direction $d$ and a decision surface are shown in Figure 3.3.
Figure 3.3: A 3-dimensional input noise space with an impulse response $h$, a biasing direction $d$ and a decision surface $S$

It was shown in [15] and verified in [16] that, for a linear system with additive Gaussian noise, the optimal $c(j)$ is $c_{opt}(j) = C(j)h$, where $h$ is the impulse response of the linear system. The idea of biasing along the direction of $h$ seems intuitively appealing, even when the noise is not Gaussian. Since the optimal direction cannot be determined analytically, at least in general, we suggest using $d = h$ as a reasonable heuristic choice when the system is linear.

Furthermore, the direction of $h$ also seems to be optimal, for linear systems, in a different sense: Let $U$ be an orthogonal transformation in $\mathbb{R}^M$, where $U = [h, b_2, \ldots, b_M]$, and $b_i$, $i = 2, \ldots, M$ are column vectors in $\mathbb{R}^M$. Let $n$ and $n^*$ be the noise vectors before and after biasing. Let $v$ and $v^*$ be random vectors such that $n = Uv$ and $v^*$ is the biased version of $v$. Then

$$n^* = n + C(j)h$$

$$= Uv + C(j)U[1, 0, \ldots, 0]^T$$
\[ = \mathcal{U} \left[ \mathbf{v} + C(j) [1, 0, \ldots, 0]^T \right] \]

But \( \mathbf{n}^* \) is also \( \mathbf{n}^* = \mathcal{U} \mathbf{v}^* \). Therefore,

\[ \mathbf{v}^* = [v_1 + C(j), v_2, \ldots, v_M]^T \]

The biased output \( y^* \) of the linear system will be

\[ y^* = h^T \mathbf{n}^* = h^T \mathcal{U} \mathbf{v}^* = v_1 + C(j) \]

assuming \( \|h\| = 1 \). This implies that, in the case of a linear system, translating in the direction of \( h \) has the potential of reducing the dimensionality of the problem to unity, with all the obvious benefits of reduced memory. Davis in [13], first suggested using an orthogonal Householder transformation to reduce the dimensionality to unity, and presented results based on the IS technique where the variance of the noise distribution is increased. This approach is also discussed in [16] for the Gaussian case in conjunction with the translation technique. Although it might not always be possible to actually implement the Householder transformation method for any noise distribution, this approach provides additional arguments in favor of the direction of the impulse response, at least at the intuitive level.

As a last and most important argument in favor of biasing in the direction of \( h \), note that among all directions \( \mathbf{d} \), an additive bias \( C \mathbf{h} \) maximizes the effect of translation at the output of a linear system since, as shown in Figure 3.4,

\[ y^* = h^T \mathbf{n}^* = h^T (\mathbf{n} + C \mathbf{h}) = h^T \mathbf{n} + C \|h\|^2 = y + C \|h\|^2 \]

where \( y^* \) is the biased output resulting from translating the input \( \mathbf{n} \) by \( C \mathbf{h} \). In the \( M \)-dimensional input space the important region boundary is a hyperplane perpendicular to the impulse response vector. The output of the linear system \( y \) is the projection
Figure 3.4: Linear system: a 2-dimensional illustration of an impulse response $h$, a biasing direction $d$ and a decision surface (hyperplane perpendicular to $h$)

of the input vector on the direction of the impulse response. The direction $h$ can thus be thought of as a mapping of the $y$-axis, with the decision threshold located $T$ units of distance from the origin (Figure 3.4). Therefore, biasing along this direction clearly maximizes the effect of the translation, in the sense that out of all translations with the same magnitude $C \|d\|$, translation in the direction $d = h$ incurs the largest increase of error count.

For a nonlinear system, this suggests that a good biasing direction would be the one in which the effects of translation are maximized at the output. In general, the choice of good translation direction(s) for nonlinear systems will depend on the characteristics of the particular nonlinearity. We will present in the next section a model for nonlinear systems that greatly facilitates the search for a favorable biasing direction.

There is another interesting way of looking at the optimal direction problem: Let
\[ y = h^T x \text{ and } n^* = n + C(j)d. \] Then

\[
y^* = h^T n^*
= h^T n + C(j) h^T d
= h^T n + C_y
= y + C_y
\] (3.6)

where \( C_y = C(j) h^T d \). Then (temporarily dropping the explicit reference to \( j \))

\[
f_Y^*(y) = f_Y \left( y - C(j) h^T d \right) = f_Y(y - C_y)
\]

and we conclude that for a linear system, translation of the joint input distribution by \( Cd \) always results in pure translation of the one-dimensional output distribution by \( C_y = C h^T d \). Now, if one wants to specify \( C_y \) at the output, say to achieve a certain raw error count, there are infinite ways that this can be accomplished by translating the input pdf's (the infinite solutions \((C, d)\) to the equation \( C_y = C h^T d \)). However, each one of these solutions would result in a different estimator variance (Eq. 2.14) and only one of them would correspond to the optimal biasing of the input distribution (i.e., to minimum estimator variance).

From a non-geometric point of view, and assuming \( ||d|| = 1 \), the direction of bias can be thought of as the way of distributing the amount of output translation \( C_y \) over the input pdf's. It should be clear that, although the amount of translation \( C_y \) is important, the way this is distributed among the input pdf's is crucial for the resulting estimator variance.

It is relatively easy to give examples of unfavorable directions: \( d = -h \) would be obviously bad because, for \( C > 0 \), an additive bias \( Cd \) would translate the output distribution away from the important region, thus causing less errors to occur than in
the unbiased case. Also, \( d \) such that \( d^T h = 0 \) would imply \( C_y = 0 \), thus incurring no difference on the output distribution \( (y^* = y) \). In fact, the whole half-space defined by \( h^T d \leq 0 \) can be immediately rejected, since the resulting translation at the output is non-positive.

Clearly, \( d = h \) seems to be the most natural choice, although, to this point, no general, rigorous proof exists for its optimality.

### 3.4 Nonlinear Systems

Nonlinear systems are, in general, much more difficult to analyze from the IS perspective than linear systems. Although they are the very systems for which simulation is most commonly needed, describing and modeling them, even for simulation purposes, may be far from trivial. Attempting to apply IS techniques in a general and practical way appears to be a formidable task.

At a first glance, one might suggest attacking the problem on a case-by-case basis, taking advantage of special conditions each time. However, such an approach is not sufficient (clearly, simulation practitioners would not like having to approach every nonlinearity as a totally new problem) and provides no further insight into the problem.

A first step towards a solution for the problem is an attempt to approximate the nonlinear system with a linear system. The authors in [14] proposed using regression to estimate the coefficients of an equivalent impulse response that approximated the original system as close as possible. The obvious advantage of such a method is that, after a "linear equivalent" has been found, all the results from linear systems can be used directly (including, for example, techniques to reduce the dimensionality, as in
Unfortunately, and apart from being intuitively displeasing, the idea of a linear approximation to a nonlinearity is meaningful and useful only for very mild nonlinearities.

We suggest going a step further and using a more complex model, by cascading a linear, time-invariant, causal system to a memoryless nonlinear system. Figure 3.5 shows such a discrete nonlinear system.

\[
\begin{align*}
    y_n &= g(w_n) = \sum_{i=0}^{\infty} a_i w_i \\
    w_n &= \sum_{i=0}^{\infty} h_i x_{n-i}
\end{align*}
\]
As shown in [3], pages 9-10, such a representation of a discrete nonlinear system is equivalent to a **discrete Volterra series** model. In general, the memory of the linear part and the order of the nonlinear part can be infinite. In practice, however, we truncate the memory to $M$ and the order (highest power) of the nonlinearity to $K$. This results in a Volterra series representation with a finite number of summations of finite terms each. Note that such a model (Figure 3.5) is very general and can describe accurately any nonlinearity, assuming we allow $M$ and $K$ above to be sufficiently large. It is also a compact, relatively simple model and is used extensively by researchers and simulation practitioners. Several techniques have been proposed to estimate the coefficients of such a model, given a nonlinear system. In this work we will assume that this **system identification** problem can be solved with acceptable accuracy.

Consider initially the case where the system consists of an instantaneous (memoryless) nonlinearity: $y(t) = g(x(t))$ or $y_k = g(x_k)$, for the discrete case. Then, $y_k > T$ is equivalent to $g(x_k) > T$ or $x_k \in \Omega_X$, where $\Omega_X = \{x : g(x) > T\}$. In general, an instantaneous nonlinearity is merely equivalent to a transformation (mapping) of the one-dimensional output space important region to the one-dimensional input space. For example, if $g(x) = x^2$, $\Omega_Y = \{y : y > T\}$ maps to $\Omega_X = \{x : x > \sqrt{T} \text{ or } x < -\sqrt{T}\}$ as Figure 3.6 shows.

Therefore, in the IS context, in order to increase the number of occurrences of important events $y \in \Omega_Y$, we should somehow increase the probability of $x \in \Omega_X$ that is described by the mapping $y = g(x)$ or, equivalently, $x = g^{-1}(y)$.

Assume now that a linear system with memory $M$ and impulse response $h$ precedes the instantaneous nonlinearity $g(\cdot)$, as predicted by the general model we described earlier (Eq. 3.7,3.8). Referring again to Figure 3.5, $w = h^T x$ and $y = g(w)$. Then, $w$ is the projection of $x$ on the direction of $h$ and, as discussed earlier, the direction of
Figure 3.6: An example of the mapping induced by an instantaneous nonlinearity

\[ y = g(x) = x^2 \]

\[ x = g^{-1}(y) = \pm \sqrt{y} \]

\[ y \rightarrow \mathcal{O}_y \]
\[ x \rightarrow \mathcal{O}_x \]

\[ 0 \rightarrow T \]
\[ -\sqrt{T} \rightarrow 0 \]
\[ 0 \rightarrow \sqrt{T} \]

\( h \) becomes a mapping of the \( w \)-axis. The important region \( \Omega_Y \) (typically a threshold-type region, \( y > T \)) is mapped on one or more regions of the \( w \)-axis or, equivalently, of the direction of \( h \). For example, if \( \Omega_Y = \{ y : y > T \} \) and \( y = g(w) = w^3 \) then \( \Omega_W = \{ w : w > \sqrt[3]{T} \} \) and \( \Omega_X = \{ x : h^T x > \sqrt[3]{T} \} \). If \( y = g(w) = w^2 \) then \( \Omega_W = \{ w : w > \sqrt{T} \text{ or } w < -\sqrt{T} \} \) and \( \Omega_X = \{ x : h^T x > \sqrt{T} \text{ or } h^T x < -\sqrt{T} \} \).

Figure 3.7 illustrates these examples for \( M = 2 \).

In most realistic cases nonlinear systems are operated in the range where \( g(\cdot) \) is monotonic. Assuming monotonicity, \( \Omega_Y = \{ y : y > T \} \) and \( y = g(w) \) imply \( \Omega_W = \{ w : w > g^{-1}(T) \} \) and \( \Omega_X = \{ x : h^T x > g^{-1}(T) \} \) (an increasing function was assumed, without loss of generality). In the IS context this implies that an approach similar to that taken in the linear case can be used, where \( T \) will be substituted by \( \tilde{T} = g^{-1}(T) \) (Figure 3.8). Using this model and assuming monotonicity of \( g(\cdot) \), it seems that biasing by translation in the direction of \( h \) would result in substantial
Figure 3.7: A 2-dimensional illustration of the mapping of important regions for a nonlinear system

Figure 3.8: Mapping of the important region when the nonlinearity is monotonic (increasing)
runtime savings, comparable to those achieved when the system is linear.

If \( g(\cdot) \) cannot be assumed monotonic in the entire operating range the problem becomes more complicated. We can distinguish two interesting cases:

(i) \( \Omega_w \) 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 [25], 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) \( \Omega_w \) 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).

Figure 3.10: Input space and important regions when $\Omega_W = \{w : \tilde{T}_1 > w > \tilde{T}_2\}$

Figures 3.9 and 3.10 illustrate, respectively, cases (i) and (ii).

If $g(\cdot)$ is non-monotonic and does not fall in these two categories then, in general, biasing by translation may not be advantageous and the problem should be solved on a case-by-case basis.
3.5 System with Memory: Effects of Over- and Undertranslation

Assuming now that the modified pdf is given by (3.5), where \( e(j) = C(j)d \) and \( d = [d_1, \ldots, d_M] \) has been chosen appropriately, we focus our attention to the problem of estimating near-optimal parameters \( C(j) \) for each realization. Arguments similar to those used for the memoryless case can be used to justify the observation that, when \( C(j) \) is too large, underestimation occurs with high probability.

Theorem 2. Let the true probability of error, under the \( H_0 \) hypothesis and realization \( j \), be \( P_e(j) \), and the IS estimator (when translation is used) be \( \hat{P}_e(j) \). Assume that the system under consideration can be represented by the equation \( y = g(n) \), where \( n \) is an \( M \)-dimensional input vector. For any given \( w_{o,max} = aP_e(j) \), where \( 0 < a < 1 \), any number of decisions \( N \) and any arbitrary \( P_{min} \), a translation amount \( C_{min}(j) \) exists such that, for \( C(j) = C_{min}(j) \):

\[
\Pr[\hat{P}_e(j) \leq w_{o,\text{max}}] = P_{min}
\]

Furthermore, for any \( C(j) \geq C_{min}(j) \), \( P_e(j) \) will be underestimated by an even larger amount, i.e., \( \hat{P}_e(j) \leq w'_{o,\text{max}} \leq w_{o,\text{max}} \), with probability \( P_{min} \).

**Proof:** Let \( w_o(y, C(j)) \) denote the output weight function and \( w_{i,j}(n, C(j)) \) the weight function of the \( i \)-th input pdf, under realization \( j \). Let \( y_k \) denote the \( k \)-th decision sample in a simulation and \( n_{ki} \) denote the \( i \)-th element of the corresponding input vector. Also, let \( I_{T,k} \) be an indicator function (equal to 1 if and only if \( y_k > T \).
and equal to 0 otherwise). Since \( w_o(y, C(j)) > 0 \) and \( 0 \leq I_{T,k} \leq 1 \),

\[
\hat{P}_e(j) = \frac{1}{N} \sum_{k=1}^{N} I_{T,k} w_o(y_k, C(j)) \leq \frac{1}{N} \sum_{k=1}^{N} \max\{w_o(y_k, C(j))\} = \max\{w_o(y_k, C(j))\}
\]

(3.9)

for any translation amount \( C(j) \). In order to prove the first part of the theorem, we only need to prove that there exists a \( C_{\text{min}}(j) \) such that

\[
\Pr[\max\{w_o(y_k, C(j))\} \leq w_{o,\text{max}}] = P_{\text{min}}
\]

for \( C(j) = C_{\text{min}}(j) \).

Let \( n = w_{i,j}^{-1}(W, C(j)) \) be the value such that \( w_{i,j}(n, C(j)) = W \) for some weight \( W \). That is, \( w_{i,j}^{-1}(W, C(j)) \) is the inverse weight function (for the \( i \)-th input pdf) parameterized by the translation parameter \( C(j) \). Furthermore, when the translation parameter is equal to \( C(j) \), let \( P_A(C(j), W) \) be the probability that a single decision will correspond only to input weights greater than \( W \), and \( P_B(C(j), W) \) be the probability that all \( N \) decisions will correspond only to input weights less than \( W \). Since decisions are independent,

\[
P_B(C(j), W) = (1 - P_A(C(j), W))^N
\]

The input weight functions are assumed to be non-increasing with \( n \). Assuming mutual independence of the input samples, the total weight for each output sample is given by the product of the corresponding input weights (Eq. 2.12). A way to force output weights to be less than \( w_{o,\text{max}} \) is to make input weights be less than \( w_{i,max} = \sqrt[2]{w_{o,\text{max}}} \). Then, \( C_{\text{min}}(j) \) above exists and can be found as the minimum \( C(j) \) that satisfies the equation

\[
1 - P_A(C(j), w_{i,\text{max}}) = \sqrt[N]{P_{\text{min}}}
\]

(3.10)
or, equivalently,

\[
\prod_{i=1}^{M} \Pr \left[ w_{i,j}(n, C(j)) \leq w_{\text{in,max}} \right] = \prod_{i=1}^{M} I(C(j), d_i) = \sqrt{P_{\text{min}}} \quad (3.11)
\]

where \( C_{\text{min}}(j) > 0 \), \( I(C(j), d_i) \) are integrals (i.e., tail probabilities) of the form

\[
\int_{\infty}^{w_{i,j}^{-1}(w_{\text{in,max}}, C(j))} f_{N, j}(n - C(j) d_i) \, dn \quad \text{, if } d_i > 0
\]

\[
\int_{-\infty}^{w_{i,j}^{-1}(w_{\text{in,max}}, C(j))} f_{N, j}(n - C(j) d_i) \, dn \quad \text{, if } d_i < 0
\]

Clearly, when \( C_{\text{min}}(j) \) is chosen in such a way,

\[
P_{B}(C_{\text{min}}(j), w_{\text{in,max}}) = (1 - P_{A}(C_{\text{min}}(j), w_{\text{in,max}}))^N = P_{\text{min}} ,
\]

in which case, both

\[
\max \{ w_{i,j}(n_{ki}, C_{\text{min}}(j)) \} \leq w_{\text{in,max}}
\]

and

\[
\max \{ w_{o}(y_{k}, C_{\text{min}}(j)) \} \leq (w_{\text{in,max}})^M = w_{o,\text{max}}
\]

with will hold with probability equal to \( P_{\text{min}} \). Then,

\[
\hat{P}_{e}(j) = \frac{1}{N} \sum_{k=1}^{N} I_{T,k} w_{o}(y_{k}, C_{\text{min}}(j)) \leq \frac{1}{N} \sum_{k=1}^{N} \max \{ w_{o}(y_{k}, C_{\text{min}}(j)) \} \leq w_{o,\text{max}}
\]

(3.12)

with probability equal to \( P_{\text{min}} \), which proves the first part of the theorem. Figure 3.11 illustrates schematically these facts along one of the input dimensions.

To prove the second part of the theorem let \( n_{w_{\text{in,max}}, i} = w_{i}^{-1}(w_{\text{in,max}}, C_{\text{min}}(j)) \) and \( b \) be any positive number. According to the first part, for \( C(j) = C_{\text{min}}(j) + b \geq C_{\text{min}}(j) \),

\[
\max \{ w_{i,j}(n_{ki}, C_{\text{min}}(j) + b) \} \leq w_{\text{in,max}}' = w_{i,j}(n_{w_{\text{in,max}}, i} + b, C_{\text{min}}(j) + b)
\]

37
Figure 3.11: System with memory: schematic illustration of underestimation proof, part 1

and

$$\max\{w_o(y_k, C_{\min}(j) + b)\} \leq w_{o,\text{max}}'$$

$$= \left( w_{\text{in, max}}' \right)^M$$

$$= \left( w_{i,j}(n_{\text{win, max}}, i + b, C_{\min}(j) + b) \right)^M$$

both with probability equal to $P_{\min}$. Invoking the non restrictive assumption that $f_{N\mid j}(n + b) \leq f_{N\mid j}(n)$, results in

$$w_{\text{in, max}}' = w_{i,j}(n_{\text{win, max}}, i + b, C_{\min}(j) + b) \leq w_{i,j}(n_{\text{win, max}}, i, C_{\min}(j)) = w_{\text{in, max}}$$

and, therefore

$$w_{o,\text{max}}' = \left( w_{\text{in, max}}' \right)^M \leq \left( w_{\text{in, max}} \right)^M = w_{o,\text{max}}$$

This implies that, for $C(j) \geq C_{\min}(j)$ and with probability $P_{\min}$, the probability of
error will be underestimated by an even larger amount than in (3.12). Figure 3.12 illustrates the above.

As in the memoryless case, equation (3.9) is always true, that is, $\hat{P}_e(j)$ is less than the maximum output weight occurring in the simulation: $\hat{P}_e(j) \leq \max\{w_o(y_k, C(j))\}$. Our purpose was to show that this maximum weight will eventually become very small with high probability when the input distributions are over-translated, thus resulting in underestimation of $P_e(j)$. Some added complexity stems from the higher dimensionality of this case with respect to the memoryless case. Equation (3.11) is indicative of the fact that the probability of an input weight being greater than a preset maximum is now the product of $M$ tail probabilities. The fact that all output weights must be less than the prespecified bound $w_{o,\text{max}}$ is taken into account by using the $N$th root at the extreme right hand side of (3.11).
As an illustrative example, consider the case where the input pdf's are Gaussian with mean zero and \( \sigma = 1 \). Let the system be linear with \( M = 3 \) and \( h = [-0.5, 2, -0.5] \). The output will be Gaussian \( N(0,4.5) \) and for a threshold \( T = 12, P_e \approx 10^{-8} \). Assume translation by \( c = C h \). Let \( N = 500 \). We set \( w_{o,\text{max}} = 10^{-10} \) and \( P_{\min} = 1 - 10^{-2} \). Then \( w_{\text{in, max}} = \sqrt[w_{o,\text{max}}] = 4.641 \times 10^{-4} \). Solving (3.11) for \( C_{\min} \) we find \( C_{\min} \approx 20 \). This means that, for \( C > 20 \) our estimate will be less than \( w_{o,\text{max}} = 10^{-10} \) with a probability of at least \( P_{\min} = 1 - 10^{-2} \).

As pointed out earlier, the bounds on \( C \) and \( C(j) \) implied by (3.3) and (3.11) are very conservative, and are used here only in order to rigorously prove the high probability of underestimation. In actual experiments, underestimation occurs consistently for values of \( C \) considerably less than those predicted above.

Turning our attention to the other extreme case of interest, when \( C \) is relatively small, we observe that the biased input pdf's \( f^* \) are still very similar to the original pdf's \( f \) and their behavior resembles strongly that of the (unmodified) MC case. Assuming a low \( P_e \) and a very small number of decisions, \( N \), no errors will be generated in most simulation runs and the corresponding estimates will be \( \hat{P}_e = 0 \), thus pathologically underestimating \( P_e \). In the rare case that one or more errors occur, the \( P_e \) will be over-estimated because of the small number of decisions used — consider, for example, the case where \( P_e = 10^{-8} \), \( N = 1000 \) and one error is detected. If \( C \) is close to zero the weight corresponding to this error will probably be large, say \( w \geq 0.001 \). This will result in an over-estimation: \( \hat{P}_e = \frac{w}{N} \geq 10^{-6} \). However, since this will be an isolated, rare event, the overall behavior of the estimator in this range of \( C \)'s will be that of underestimation.

As the translation amount \( C \) is increased from zero, the variance in the estimate will decrease, and for values of \( C \) around the optimal the estimates will be reasonably
close to the true $P_e$. Eventually, when the translation amount grows well beyond the "optimal" range of $C$'s and reaches the over-translation region, the estimator will consistently underestimate the probability of error.

Combining these theoretical observations, we conclude that if the number of decisions is very small with respect to the true $P_e$ expected, a typical plot of the resulting estimates as a function of $C$ for $0 \leq C \leq C_{\text{max}}$ will look like the one in Figure 3.13.

![Diagram](image)

Figure 3.13: Typical curve of $\hat{P}_e$ vs. $C$, for $0 \leq C \leq C_{\text{max}}$

### 3.6 Proposed method

The above theoretical observations suggest the following heuristic scheme to locate optimal or near-optimal settings for $C$, for a given direction $d$. For each realization $a(j), j = 0, 1, \ldots, J - 1$:

- Choose $N$ to be extremely small with respect to the true $P_e$ — a rough estimate
of the BER is assumed to be always available; even if this is not true, $N$ could be later modified, as we explain in the following. The choice of $N$ depends on the actual $P_e$ and the variance reduction that is expected to be obtained by IS.

— Run a series of simulations with length $N$, where $f_N(y(n))$ is translated by $C(j)d$, that is

$$n_i^*(j) = n_i(j) + C(j)d, \quad i = 1, \ldots, M$$

for every block of $M$ samples. Plot the curve of estimates $\hat{P}_e$ as a function of $C(j)$ for $0 \leq C(j) \leq C_{\max}$, where $C_{\max}$ is chosen to result in significant underestimation of $P_e$.

— If $N$ is not too small, the curve will consist of three distinguishable regions corresponding to "C too low", "C reasonable" and "C too high". Then, a near-optimal $C(j)$ can be picked from the range where the curve is at its flattest, as shown in Figure 3.13.

The choice of $N$ together with the actual improvement for good $C$'s determines the shape of the curve in a predictable way: When $N$ is too large, the flat region will be relatively wide making the choice of a good $C$ more difficult (actually, this wide flat region implies that, for the $N$ used, the estimator variance is less sensitive to the choice of $C$ in this range). When $N$ is too small and/or the variance reduction is small, behavior similar to the case where $C$ is too small will dominate for the whole range of $C$'s; that is, most estimates will be zeros with occasional over-estimation spikes. The curve in this case will never rise to a flat region before underestimation starts occurring, indicating that the $N$ chosen is too small to give reasonable estimates, even for $C$'s close to optimal. In other words, the variance reduction of the estimator, induced by the translation of the pdf will never be enough to compensate for the very
small $N$ used. Figure 3.14 shows typical curves expected for various choices of $N$.

![Figure 3.14: Typical curves for various choices of number of decisions $N$](image)

To maximize computational efficiency, it is better to start experimenting with an $N$ that might be too small and later increase it, if needed. Since the number $N$ for each run will typically be much smaller than the number that will eventually be used for an accurate estimate of $P_e$, the overhead involved in using this method to find good $C$'s will be low. A useful byproduct of this trial-and-error process is a rough estimate of the variance reduction expected for good $C$'s (the ratio of the number of decisions that would have been used for MC simulation over the number of decisions $N$ that is actually used).

An alternate, more efficient scheme, in the sense that less set-up simulation runs are required, could exploit the prior knowledge of the shape of the curve and use a type of "binary search" or "variable stepsise" approach to zero-in on the middle, flat region, without actually obtaining estimates for all $C$'s between 0 and $C_{max}$.
Ongoing research is attempting to answer the question of locating the optimal \( C \) more accurately within the flat region, and the robustness of the improvement factor around this optimal. Moreover, the relationship between the raw error count (i.e., number of errors occurring at the output) and the optimal \( C \) is still under investigation. Our experiments indicate that very large improvement factors can be obtained even when we choose a value of \( C \) in this region almost arbitrarily, although the point where the curve exhibits maximum flatness seems to be the best choice. On this last subject note that, since consecutive estimates \( \hat{P}_e \) on such a curve differ only slightly in translation amount \( C \), they can be considered (approximately) as estimates at the same value of \( C \). Therefore, it appears that the local flatness of the curve (i.e., region where \( d\hat{P}_e(C)/dC \approx 0 \)) is a good indicator of reduced estimator variance.

Clearly, a second statistical indicator of the goodness of the estimate as a function of \( C \), would be very useful. Such a statistic, together with the one we propose and maybe a more efficient search procedure would allow near-optimal translation parameters to be found within the actual simulation itself and not by previously run, exploratory simulations. A natural candidate for secondary statistic appears to be the sample variance (or coefficient of variation) of the BER estimates. At this stage, the practicality and reliability of such a statistic as well as the sensitivity of the improvement factor to changes of \( C \) in the flat region are still under investigation.

3.7 Summary

Since the effects of under-translation and over-translation are predictable in the way described for a very broad class of pdf's, we conclude that the heuristic scheme we propose would be a simple, practical and efficient method to locate at least a range of
near-optimal values for the translation parameters $C(j)$, without requiring intractable analytical derivations.

In order to justify the observed effects of under- and over-translation, the arguments presented above are not restricted to the case of a linear system. Therefore, once a good translation direction $d$ has been chosen (e.g., $d = h$ for a linear system), similar observations can be used to determine whether the amount of translation $C$ along that direction is too large or too small. For nonlinear systems, the problem of identifying near-optimal translation directions appears to be less simple in general; however, these difficulties can be surpassed in many cases by using the nonlinear system model (i.e., cascade of linear system and memoryless nonlinearity) that we proposed earlier. Our experimental results indicate that this approach can lead to significant improvement over MC simulation.
Importance Sampling for Single-Sided Distributions

4.1 Background

A class of noise distributions that has not attracted significant attention so far in the Importance Sampling context, is the class of *single-sided* distributions. Let $X$ be a random variable with pdf $f_X(x)$. We say that $X$ has a single-sided distribution when there exists an $x_{\text{min}}$ such that

$$f_X(x) = 0, \text{ for } x < x_{\text{min}}$$

Suppose that we want to estimate the probability of the rare event $\{x \in \Omega_X\}$, where $\Omega_X$ is what we called earlier the important region. The fundamental rule when applying IS techniques is that the modified pdf $f^*_X(x)$ must be non-zero everywhere in $\Omega_X$ [10, 12]. If $f^*_X(x) = 0$ somewhere in $\Omega_X$, the variance of the IS estimator in (2.10) becomes infinite, resulting in underestimation of the true probability of an important event.

When the important region is of the type $\Omega_X = \{x : x > T\}$ (the case of binary detection with memory $M = 1$), and $f^*_X(x) = f_X(x - C)$, it is sufficient that $C \leq T - x_{\text{min}}$ for this basic requirement to be met. However, in the case of a system with
memory, where the system output is given by $y = g(x)$ and we are trying to estimate $P_e = \Pr[y > T]$ by modifying the joint distribution of the input variables, $f_x(x)$ must be non-zero in $\Omega_X = \{x \text{ such that } y = g(x) > T\}$. In this case, it is less obvious what parts of the domain of $X$, $D_X$, contribute to important events, in other words what $\Omega_X$ is. Actually, for many systems, linear or nonlinear, $\Omega_X$ can very well cover the whole domain $D_X$. It follows that it is impossible, in general, to apply IS based on pure “linear shift” or “translation” of the input pdf’s, if these pdf’s are single-sided.

A good example of a single-sided distribution with practical importance that was a strong motivation for this discussion, is the pdf describing the output current of an Avalanche Photodiode Detector (APD) in lightwave communication systems. The widely used approximation obtained by Webb, McIntyre and Conradi in [26] (referred to in the following as the WMC pdf), is given by

$$f(x) = \frac{1}{\sqrt{2\pi \sigma}} \frac{\exp \left[ -\frac{(x - \bar{X})^2}{2\sigma^2 \left( 1 + \frac{\epsilon - \bar{X}}{\sigma \delta} \right)} \right]}{(1 + \frac{\epsilon - \bar{X}}{\sigma \delta})^{\frac{\epsilon}{\sigma \delta}}}$$ (4.1)

for $x > -\sqrt{n_p / F_e}$, where $\bar{X}$ is the mean output current, $\sigma^2$ is the variance of the current and $\delta$ is a shape parameter and

\[
\begin{align*}
\bar{X} &= (n_e q / \Delta t) G \\
q &= \text{charge of an electron} \\
\Delta t &= \text{time interval} \\
G &= \text{average avalanche gain} \\
n_e &= (\eta / \Delta t / \hbar \Omega) P_{opt} = \text{number of primary electrons at the input of the APD} \\
\eta &= \text{quantum efficiency of the diode} \\
P_{opt} &= \text{average optical power in time } \Delta t \\
\hbar \Omega &= \text{energy of a photon} \\
\sigma^2 &= n_e G^2 F_e (q / \Delta t)^2 = \text{variance of the diode output current} \\
\delta &= \sqrt{n_e F_e} / (F_e - 1) \\
F_e &= k G + [2 - (1/G)](1 - k) = \text{excess noise factor} \\
k &= \text{ionization ratio}
\end{align*}
\]
The properties of this pdf were studied in the context of efficient simulation in [6] and [27].

The Shot noise affecting an APD is a typical example of noise that is not i.i.d. and additive. In fact, the underlying noise process stems from a doubly stochastic Poisson process (Poisson arrivals of photons multiplied by a random gain) and is of multiplicative nature. However, using the WMC approximation the noise can be modeled as additive but signal dependent. The signal dependency is obvious from the fact that the variance $\sigma^2$ and the shape parameter $\delta$ depend on the level of the input signal, i.e., the number of electrons in time $\Delta t$: [6, 27]:

$$
\sigma^2 \propto n_e
$$

$$
\delta \propto \sqrt{n_e}
$$

For very small inputs the WMC distribution takes an exponential or Gamma-like behavior, while for large inputs the distribution resembles closely a Gaussian, as can be seen in Figure 4.1.

In order for the conventional IS techniques (and most importantly for our development) to accommodate cases like the APD noise, we extended the standard notation used in [7, 10, 15] in Chapter 2.3 to cover the case where the noise characteristics are depending on the input signal (Eq. 2.14). However, since this pdf is single-sided, it is not possible, in general, to use the standard translation-based IS technique to estimate the $P_e$ of an APD receiver in a lightwave link. Similar problems arise when dealing with distributions like the Rayleigh, the Gamma and many others that are single-sided but not necessarily signal-dependent like the WMC.
Figure 4.1: The WMC density distribution for three input levels (optical power) \( P_1 < P_2 < P_3 \): \( P_1 = -52\text{dBm}, P_1 = -46\text{dBm} \) and \( P_3 = -43\text{dBm} \). The APD characteristics were: \( G = 25, k = 0.8, \eta = 0.8 \) and \( \Delta t = 8.92 \times 10^{-10} \).

### 4.2 Importance Sampling Schemes

Motivated by the great power of the translation technique in the case of double-sided pdf's, we propose a class of nonlinear biasing transformations, or "quasi-linear" techniques, that retain most of the characteristics of pure translation, without violating the fundamental condition mentioned earlier.

One way of dealing with this problem is to shift only the a version of the original distribution that carries a fixed percentage of the total probability mass, while spreading the remaining probability mass uniformly over the range of the domain of \( z \) that would have otherwise remained "empty":

\[
f^*(z) = \begin{cases} 
Af(x - C) & , z > z_{\text{min}} + C \\
1 - A & , z_{\text{min}} < z < z_{\text{min}} + C 
\end{cases}
\]

where \( 0 < A \leq 1 \). \( A \) would normally be chosen close to 1, say \( A = 0.8 \) (Figure 4.2.a).
For large $z$ this would resemble pure translation.

It is easy to show that

$$E[X^*] = AE[X] + \frac{1 + A}{2} C$$

and that, when the system is linear,

$$E[Y^*] = AE[Y] + \frac{1 + A}{2} C_v$$

where $C_v = h^T C$, as opposed to

$$E[X^*] = E[X] + C$$

and

$$E[Y^*] = E[Y] + C_v$$

for pure translation (i.e., $A = 1$). However, if $f(z_{min}) = 0$, this scheme would still lead to infinite estimator variance since $f^*(z_{min} + C) = 0$.

Another way one could visualize is to use the mirror image of $f(z)$ (with probability mass $1 - A$) to "fill" the empty space left behind when $f(z)$ is shifted. Still, this method can lead to infinite estimator variance if $f(z_{min}) = 0$ (Figure 4.2.b).

A more natural and intuitively appealing way of achieving our goal would be, in cases like the WMC, to bias by simply "increasing the input level". Recall that the characteristics of the WMC distribution depend on the level of the signal (optical power) at the input of the APD. Also recall that, increasing the input would result in a distribution with increased mean but also with different shape. Furthermore, since the modified pdf would in this case remain nonzero for $z > z_{min}$ no fundamental rule of IS is violated. The problem with this method, however, is that it is not general enough to account for all cases of single-sided distributions (Figure 4.2.c).
Figure 4.2: Three different quasi-linear biasing schemes: (a) Translation with uniform correction (b) Translation with "mirror-image" correction (c) Biasing by increasing the input level

A more general idea is simply to let $x^*$, the biased random deviate, asymptotically approach $x + C$ as $x$ gets larger, while letting $x^*$ approach $x$ when $x$ is close to $x_{min}$. The convergence rate of the transformation $T(x)$ to $x + C$ as $x$ increases, can be proportional to $1/x$, $(1/x)^\nu$ (rational) or $e^{-x}$ (exponential). Figure 4.3 gives a graphical illustration of such nonlinear transformations.

In this work we have chosen, mainly for simplicity, the transformation:

$$x^* = T(x) = x + C \left[ 1 - \frac{\alpha}{\alpha + (x - x_{min})} \right]$$  \hspace{1cm} (4.6)

Figure 4.4 shows an example of using (4.6) for a Rayleigh distribution

$$f_x(x) = \frac{2}{b} (x - x_{min}) e^{-(x - x_{min})^2 / b^2} u(x - x_{min})$$

when $b = 0.3$, $x_{min} = -2.0$, $\alpha = 0.5$ and $C = 0, 1, 2$. Figure 4.5 shows another
Figure 4.3: Nonlinear transformation ("quasi-linear" biasing) of a random variable

Figure 4.4: Quasi-linear biasing applied to a Rayleigh distribution, for $b = 0.3$, $x_{\text{min}} = -2.0$, $\alpha = 0.5$ and $C = 0, 1, 2$
example of using (4.6) for a WMC distribution (Eq. 4.1) when \( \bar{X} = 696.59, \sigma = 595.92, \delta = 1.23 \) and \( \alpha = 60 \). As expected, in both cases, \( f_X \) approaches \( f_X(x - C) \)

for large \( z \) while remaining non-zero between \( x_{\text{min}} \) and \( x_{\text{min}} + C \).

It is easy to show that, under this nonlinear transformation,

\[
E[X^*] = E[X] + C - C\alpha \int_0^\infty \frac{f(x)}{\alpha + x} \, dx
\]

\[
= E[X] + C (1 - \alpha I)
\]  \hspace{1cm} (4.7)

where

\[
I = \int_0^\infty \frac{f(x)}{\alpha + x} \, dx
\]  \hspace{1cm} (4.8)

Clearly, when \( \alpha \to 0 \) we have pure translation, while for \( \alpha \to \infty \) no bias is applied. At the output of a linear system with impulse response \( h \)

\[
[Y^*] = E[Y] + C_y (1 - \alpha I)
\]
for $C_y = h^T C$.

So far we have been concerned with biasing in the positive direction (i.e., away from $x_{\text{min}}$). When quasi-translation in the negative direction is needed, we propose using the inverse transformation of $T$

$$x^* = T^{-1}(x)$$

It should be clear that, since $x^* = T(x)$ is shifting the pdf to one direction, $x^* = T^{-1}(x)$ should shift the pdf to the opposite direction. The reason for using $T^{-1}(x)$ instead of pure translation is that generating samples from $f^*(x)$ in a region where $f(x) = 0$ is also illegitimate (i.e., would result in a biased, incorrect estimator). Any scheme (say, pure translation in the negative direction towards $x_{\text{min}}$) that would generate such "illegal" samples and then ignore them, would be highly inefficient, especially for large system memory $M$. Using $T^{-1}(x)$ (Eq. 4.6) ensures that no illegal samples will be generated.

4.3 Summary

Summarizing, for the general case where $x_{\text{min}} \neq 0$, we suggest using the following quasi-linear biasing scheme:

Let

$$T(x) = x + C \left[ 1 - \frac{\alpha}{\alpha + x - x_{\text{min}}} \right]$$

(4.9)

and

$$T^{-1}(x) = \frac{x - \alpha + x_{\text{min}} - C + \sqrt{\left(\alpha + C - x_{\text{min}} - x\right)^2 + 4(\alpha - x_{\text{min}})x + 4x_{\text{min}}C}}{2}$$

(4.10)

where $\alpha > 1$ is chosen appropriately.
(i) For biasing in the positive direction, let

\[ x^* = T(x) \]  \hspace{1cm} (4.11)

Then,

\[ f^*(x) = f(T^{-1}(x)) \times \left[ \frac{1}{2} + \frac{1}{2} \frac{(x + \alpha - C - x_{\text{min}})}{\sqrt{\left(\alpha + C - x_{\text{min}} - x\right)^2 + 4(\alpha - x_{\text{min}})x + 4x_{\text{min}}C}} \right] \]  \hspace{1cm} (4.12)

(ii) For biasing in the negative direction, let

\[ x^* = T^{-1}(x) \]  \hspace{1cm} (4.13)

Then,

\[ f^*(x) = f(T(x)) \times \left[ 1 + \frac{C\alpha}{(\alpha + x - x_{\text{min}})^2} \right] \]  \hspace{1cm} (4.14)

In either case, the weight function will be \( w(x) = \frac{f(x)}{f^*(x)} \) where \( f^*(x) \) will be given by (4.12) or (4.14). Equation (4.13) above comes from the solution of (4.11) for \( x \), that is we are simply inverting the transformation \( T(x) \). Equations (4.12) and (4.14) are derived using the well known rule from probability theory (assuming \( T(x) \) to be monotonic):

\[ y = T(x) \Rightarrow f_Y(y) = f_X(T^{-1}(y)) \left| \frac{dT^{-1}(y)}{dy} \right| \]

where \( f_Y(y) \) and \( f_X(x) \) are the probability density functions of \( Y \) and \( X \) respectively.

Without doubt, other nonlinear transformations like

\[ x^* = x + C \left[ 1 - e^{-\alpha(x-x_{\text{min}})} \right] \]
could be used. On this subject, we note that, although we want $f^*$ to resemble $f(x - C)$ enough for the IS estimator to be efficient, too close a resemblance to pure translation is bound to be plagued by the same negative effects associated with that illegal transformation: pure translation leads to infinite variance which, from a practical standpoint, means that underestimation is certain to occur. Therefore, the choice of a particular transformation and its parameters is case-dependent and will require some experimentation each time. Our choice in (4.6) seemed to perform really well for the pdf's used and, in any case, could be tried as a good starting point.

As noticed in Figures 4.4 and 4.5, an obvious result of using such a transformation as opposed to pure translation, is that the resulting $f^*$ is scaled down, in order to account for the non-zero probability of values between $x_{\min}$ and $x_{\min} + C$. This tends to reduce the efficiency of the IS technique. However, experimental results show that the potential for great run-time savings over MC simulation still exists, when the parameters are chosen appropriately. Moreover, the strong similarities of the proposed "quasi-translation" scheme to pure translation suggest that, the heuristic method proposed previously can still be used to determine near-optimal settings for $C$. Thus, we are able, using such a method, to incorporate even single-sided distributions in the general techniques proposed in the previous chapter, as our experimental results also verify.
5

Experimental Results

5.1 Simulation Set-up

In order to verify the effectiveness of this new technique in locating a near-optimal translation parameter, $C$, extensive simulation experiments were performed. The noise distributions and the BER's used were chosen to demonstrate the generality of this approach while remaining compatible with results from previous researchers. In addition to the Gaussian pdf which was included for verification purposes (since analytical results for the optimal $C$ and the variance reduction are available), a Rayleigh and an "Asymmetric Two-sided Exponential" (ATE) distribution were used. 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 $n < 0$. An example of a noise process that is not purely additive is the Shot noise of an avalanche photodiode. As mentioned in the previous section, under the WMC (Webb-McIntyre-Conradi) approximation to this distribution [26, 6, 27], its shape and variance depend on the transmitted symbol (Eq. 4.1). For one-sided pdf's like the Rayleigh and the WMC pdf's a modification of the standard, translation-based biasing technique had to be used (Eqs. 4.9,4.10) for the IS estimator in (2.11) to be correct.
5.2 Linear Systems

Two linear low-pass filters (based on a Hamming window design, normalized cut-off frequency = 0.09), one with $M = 24$ taps and one with $M = 40$ taps, were used for the simulation experiments on linear systems. The impulse response of these filters is shown in Figure 5.1.

![Impulse responses of the two linear, time-invariant filters that were used in the experiments](image)

Figure 5.1: Impulse responses of the two linear, time-invariant filters that were used in the experiments

The number of realizations (under $H_0$) was respectively, $J = 4$ and $J = 16$. A sampling rate of 8 samples per bit was assumed.

For each pdf and for each of the two filters, appropriate decision thresholds $T$ were chosen to yield the desired probabilities of error, on the order of $10^{-7}$ to $10^{-8}$ and $10^{-9}$ to $10^{-10}$. Then, for each case and for every realization $\mathbf{a}(j)$, 100 IS estimates $\hat{P}_e(j)$ were plotted as a function of the translation parameter $C(j)$, for $0 \leq C(j) \leq C_{\text{max}}(j)$. $C_{\text{max}}(j)$ was selected as a value for which significant underestimation occurred. The
number of decisions in every case was chosen according to the guidelines described earlier, typically $N = 40$ or $N = 400$.

Figures 5.2, 5.3, 5.4 and 5.5 show examples of the generated curves and our corresponding choices for $C(j)$.

Figure 5.2: Linear system: Curve of $P_e$ vs. the translation amount $C$, for the Gaussian distribution, where $P_e \approx 10^{-10}$, memory length $M = 24$ and realization number $j = 0$.

For these examples, curves for all other cases displayed similar behavior, as predicted by the theoretical observations in the previous section. Note that when the potential variance reduction is very large, as for the Gaussian pdf, the corresponding curves are flatter. This behavior also indicates that variance reduction is less sensitive to the choice of $C$. When the variance reduction is smaller, as for the ATE pdf, the curves are less flat for the same number of decisions.

In order to estimate the variance reduction, complete block simulations were performed using the near-optimal values of $C(j)$ chosen from the curves. For each pdf,
Figure 5.3: Linear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the Rayleigh distribution, where $P_e \approx 10^{-10}$, memory length $M = 40$ and realization number $j = 4$.

Figure 5.4: Linear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the ATE distribution, where $P_e \approx 10^{-8}$, memory length $M = 40$ and realization number $j = 9$. 
10 IS estimates \( \hat{P}_e^* \) of the total \( P_e \) (under \( H_0 \)) were obtained according to (2.11). For every estimate, \( N/J = 400 \) decisions per realization were used. That is, a total number of \( N = 1600 \) decisions (38,400 samples) for the memory of \( M = 24 \) samples, and \( N = 6400 \) decisions (256,000 samples) for the memory of \( M = 40 \) samples. The BER and the estimator variance were estimated respectively as the sample mean and the sample variance over these 10 runs. The improvement factor over MC simulation was then calculated as \( r_{IS} = \sigma^2_{MC}/\sigma^2_{IS} \), where the approximation \( \sigma^2_{MC} \approx \hat{P}_e^*/N \) was used.

The estimated BER's and improvement factors appear in Table 5.1 for each pdf and memory length. Representative runtimes (approximate CPU times) for the experimental simulations are given in Table 5.2.

Before turning our attention to nonlinear systems, it is worth noting that the behavior of the WMC pdf appears somewhat different from the others. As seen in
Table 5.1: Estimated probabilities of error and improvement factors over MC simulation: linear system

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$P_e$</th>
<th>$r_{IS}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>$1.2 \times 10^{-8}$</td>
<td>$4.7 \times 10^{6}$</td>
</tr>
<tr>
<td></td>
<td>$2.1 \times 10^{-10}$</td>
<td>$8.0 \times 10^{8}$</td>
</tr>
<tr>
<td></td>
<td>$6.0 \times 10^{-8}$</td>
<td>$2.8 \times 10^{5}$</td>
</tr>
<tr>
<td></td>
<td>$2.4 \times 10^{-10}$</td>
<td>$1.0 \times 10^{8}$</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>$3.9 \times 10^{-8}$</td>
<td>$5.0 \times 10^{8}$</td>
</tr>
<tr>
<td></td>
<td>$1.5 \times 10^{-10}$</td>
<td>$2.3 \times 10^{8}$</td>
</tr>
<tr>
<td></td>
<td>$1.2 \times 10^{-7}$</td>
<td>$4.8 \times 10^{5}$</td>
</tr>
<tr>
<td></td>
<td>$9.0 \times 10^{-10}$</td>
<td>$1.9 \times 10^{7}$</td>
</tr>
<tr>
<td>ATE</td>
<td>$2.8 \times 10^{-8}$</td>
<td>$2.1 \times 10^{3}$</td>
</tr>
<tr>
<td></td>
<td>$2.6 \times 10^{-10}$</td>
<td>$3.5 \times 10^{5}$</td>
</tr>
<tr>
<td></td>
<td>$1.5 \times 10^{-8}$</td>
<td>$4.8 \times 10^{3}$</td>
</tr>
<tr>
<td></td>
<td>$5.4 \times 10^{-9}$</td>
<td>$1.3 \times 10^{4}$</td>
</tr>
<tr>
<td>WMC</td>
<td>$1.6 \times 10^{-7}$</td>
<td>$4.8 \times 10^{3}$</td>
</tr>
<tr>
<td></td>
<td>$6.8 \times 10^{-9}$</td>
<td>$1.2 \times 10^{4}$</td>
</tr>
<tr>
<td></td>
<td>$1.6 \times 10^{-7}$</td>
<td>$2.5 \times 10^{3}$</td>
</tr>
<tr>
<td></td>
<td>$1.2 \times 10^{-9}$</td>
<td>$3.0 \times 10^{5}$</td>
</tr>
</tbody>
</table>

Table 5.2: Approximate runtimes in seconds, for the experimental simulations: linear system

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$M = 24$</th>
<th>$M = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>400 decisions</td>
<td>496 decisions</td>
</tr>
<tr>
<td></td>
<td>9600 samples</td>
<td>19840 samples</td>
</tr>
<tr>
<td>Gaussian</td>
<td>3.0</td>
<td>8.0</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>3.5</td>
<td>9.0</td>
</tr>
<tr>
<td>ATE</td>
<td>2.0</td>
<td>6.5</td>
</tr>
<tr>
<td>WMC</td>
<td>5.0</td>
<td>11.0</td>
</tr>
</tbody>
</table>

Figure 5.5, transitions between predicted regions are more gradual: the underestimation region does not start as abruptly and, within that region, the amount of underestimation increases at a lower rate with respect to the translation amount $C$. Comparing the performance of the WMC with that of the Rayleigh pdf (the only other single-sided pdf used here), the disparities can be attributed to the radically different characteristics of the WMC: while the tail of Rayleigh behaves approximately like $e^{-x^2}$, the tail of WMC asymptotically approaches $e^{-x}$ behavior, thus decaying much slower. Furthermore, in contrast to the Rayleigh pdf, the WMC is signal dependent.
which results in different distribution shapes for 0's and 1's (Gamma-like for 0's and Gaussian-like for 1's, Figure 4.1).

It seems that these peculiarities make the outcome of the proposed simulation experiments more sensitive to the choice of parameter \( \alpha \) in (4.9) and (4.10). Recalling our remarks from Chapter 4, for single-sided distributions pure translation leads to a biased estimator while our quasi-translation scheme leads to an unbiased estimator. However, as parameter \( \alpha \) becomes smaller, quasi-translation becomes very similar to pure translation with the same negative effects. Such potential behavior warrants special caution when performing simulation experiments: a poor choice of the quasi-translation parameter \( \alpha \) can lead to significant and consistent underestimation of the BER. In such a case the statistical variance of the estimator will be very small but the estimator itself will be, for all practical purposes, biased. Fortunately, in such cases the curve of \( P_e \) vs. the translation amount \( C \) will be decaying from the beginning (i.e., no "flat" region will exist) allowing the experimenter to detect the problem and correct it by adjusting \( \alpha \).

In addition to the above and for any biasing scheme, the experimenter can convince himself that the IS results are correct (i.e., unbiased) by examining estimates based on the same biasing parameters but a varying number of decisions \( N \): If the results are not changing with \( N \) the estimator is unbiased, otherwise an error might exist in the theoretical or the experimental part. This sort of check for bias was included in all our experiments.
5.3 Nonlinear Systems

In order to include nonlinear systems in our experimentation, we used the model described by Figure 3.5 and built nonlinearities by cascading linear systems with memory and instantaneous nonlinearities. Two linear low-pass filters (based on a Hamming window design), one with \( M = 24 \) taps (normalized cut-off frequency = 0.15) and one with \( M = 40 \) taps (normalized cut-off = 0.08) were used as the linear part of the systems under consideration. The impulse response of these filters is shown in Figure 5.6. The instantaneous nonlinearity used after the linear filters is shown in Figure 5.7. It is described by the equation

\[
y = \begin{cases} 
  c \left( 1 - \frac{v}{\alpha + v} \right), & \text{if } x \geq 0 \\
  -c \left( 1 - \frac{v}{\alpha + v} \right), & \text{if } x < 0 
\end{cases}
\]  

where \( c = 2 \), \( v = 4 \) and \( \alpha = 0.2 \).
The number of realizations (under $H_0$) was respectively, $J = 4$ and $J = 16$. A sampling rate of 8 samples per bit was assumed.

For two different distributions, namely Gaussian and Rayleigh, and for each of the two filters, appropriate decision thresholds $T$ were chosen to yield the desired probabilities of error, on the order of $10^{-8}$ and $10^{-10}$. Then, for each case and for every realization $a(j)$, 100 IS estimates $\hat{P}_e(j)$ were plotted as a function of the translation parameter $C(j)$, for $0 \leq C(j) \leq C_{\text{max}}(j)$. $C_{\text{max}}(j)$ was selected as a value for which significant underestimation occurred. The number of decisions was set to $N = 400$.

Figures 5.8 and 5.9 show examples of the generated curves and our corresponding choices for $C(j)$.

As in the linear case, curves for all other cases displayed similar behavior, as predicted by the theoretical observations in the previous section.

In order to estimate the variance reduction, complete block simulations were per-
Figure 5.8: Nonlinear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the Gaussian distribution, where $P_e \approx 10^{-7}$, memory length $M = 40$ and realization number $j = 5$.

Figure 5.9: Nonlinear system: Curve of $\hat{P}_e$ vs. the translation amount $C$, for the Rayleigh distribution, where $P_e \approx 10^{-10}$, memory length $M = 24$ and realization number $j = 1$. 
formed using the near-optimal values of $C(j)$ chosen from the curves. As for the linear systems, 10 IS estimates $\hat{P}_e$ of the total $P_e$ (under $H_0$) were obtained for each pdf, according to (2.11). For every estimate, $N/J = 400$ decisions per realisation were used. That corresponded to a total number of $N = 1600$ decisions (38,400 samples) for the memory of $M = 24$ samples, and $N = 6400$ decisions (256,000 samples) for the memory of $M = 40$ samples. The BER and the estimator variance were estimated respectively as the sample mean and the sample variance over these 10 runs. As before, the improvement factor 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 factors appear in Table 5.3 for each pdf and memory length.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$M = 24$</th>
<th>$M = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\hat{P}_e$</td>
<td>$r_{IS}$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$1.4 \times 10^{-8}$</td>
<td>$2.5 \times 10^{8}$</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>$1.4 \times 10^{-8}$</td>
<td>$1.1 \times 10^{6}$</td>
</tr>
</tbody>
</table>

Table 5.3: Estimated probabilities of error and improvement factors over MC simulation: nonlinear system

Representative runtimes (approximate CPU times) for the experimental simulations are given in Table 5.4.

5.4 Discussion

From Tables 5.1 and 5.3 above, our results appear to be compatible with results in previous work, where the optimal parameters were found analytically. The improvement
Table 5.4: Approximate runtimes in seconds, for the experimental simulations: nonlinear system

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$M = 24$</th>
<th>$M = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>400 decisions</td>
<td>496 decisions</td>
</tr>
<tr>
<td></td>
<td>9600 samples</td>
<td>19840 samples</td>
</tr>
<tr>
<td>Gaussian</td>
<td>4.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>4.5</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Factors obtained are substantial and seem to justify the choices for the translation parameters $C(j)$. The disparities among improvement factors for different pdf's were anticipated because of their different characteristics. It is significant that the improvement factors obtained for nonlinear systems were comparable to those for linear systems, regardless of the noise distribution used. This, once again, demonstrates the generality and the effectiveness of our technique even for nonlinear systems.

The significance of the results we presented is that, good $C(j)$'s which result in considerable estimator variance reduction have been obtained without analytical calculations. For example, it would have been impossible to find analytically optimal or even near-optimal $C(j)$'s for the Rayleigh or the WMC distribution. Instead, using the proposed method, it is relatively simple to find $C(j)$'s that lead to large variance reduction in the estimator.
6

Conclusions

We have presented a new technique for finding a near-optimal set of biasing parameters for the translation-based Importance Sampling scheme. This method exploits a relationship, for small sample sizes, that exists between the BER estimate and the amount of translation. We have given mathematical justification for this technique and shown that it is general, in the sense that it is not subject to the usual restrictions of a linear system with additive Gaussian noise. Under only very mild conditions, namely, that the tails of the pdf be decreasing, this technique is simple to implement and applicable to both linear and nonlinear systems, limited by noise that is not necessarily Gaussian.

Moreover, we have introduced a class of quasi-linear biasing schemes that extend the applicability of our method to single-sided noise distributions, something that was impossible with standard translation methods.

The precise location of the optimal translation parameter within the flat region predicted by our theoretical and experimental observations is part of ongoing research. However, initial results indicate that the improvement factor is fairly robust within that range.

Our experimental results demonstrate that, using either the standard translation
or the quasi-translation scheme combined with the our technique for finding near-optimal parameters, leads to simulation runtime improvement factors up to $10^8$ for a variety of noise distributions, linear and nonlinear systems.

Based on our theoretical and experimental observations, we conclude that the technique we propose would be a general, practical and efficient method to locate a set of near-optimal values for the translation parameters that can lead to significant runtime savings without requiring intractable analytical derivations.
Bibliography


[16] R. J. Wolfe, M. C. Jeruchim, and P. M. Hahn. On optimum and suboptimum biasing procedures for importance sampling in communication simulation. Accepted for publication in *IEEE Transactions on Communications*.


[27] J. K. Townsend and K. S. Shanmugan. On improving the computational efficiency of digital lightwave link simulation. Accepted for publication in *IEEE Transactions on Communications*. 

73