RICE UNIVERSITY

Performance Evaluation and Optimization of Stochastic Systems via Importance Sampling

by

Geoffrey Charles Orsak

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Behnaam Aazhang, Director
Assistant Professor of Electrical and Computer Engineering

Don H. Johnson
Professor of Electrical and Computer Engineering

David Scott
Professor of Statistics

Houston, Texas
April, 1990
Performance Evaluation and Optimization of Stochastic Systems via Importance Sampling

Geoffrey Charles Orsak

Abstract

Analytic solutions to determining the optimal set of system parameters and the associated performance of random input systems are typically intractable. One often only has access to the system's output under a variety of inputs, thereby requiring the optimization routine to be insensitive to the noise inherent in estimating the performance. The well known algorithms of Robbins-Monro and Kiefer-Wolfowitz asymptotically eliminate the effects of noise by effectively averaging the estimated performance over a sequence of system parameters. The convergence rate of the parameters is shown to depend on the performance estimator's variance. Consequently, implementing variance reduction techniques will greatly enhance the convergence properties of these algorithms. The Importance Sampling technique is employed to minimize the variance in estimating the system performance. A class of Importance Sampling biasing distributions is derived in this thesis for the specific use in analyzing single-user communication systems. This class is extended for use in estimating the performance of multiple-access systems. Additionally, a more general method of determining biasing
densities for estimating arbitrary functionals of random vectors is obtained by using ideas from robust statistics. All of the above mentioned methods render substantial improvements over standard Monte Carlo simulations when estimating system performance. By incorporating these techniques into both the Robbins-Monro and Kiefer-Wolfowitz algorithms, we show that the stopping times for these algorithms can be significantly reduced. Moreover, the computational savings over the traditional implementation of these algorithms are unbounded when optimizing systems whose performance criteria are diminishing probabilities of a set of events. These techniques can be directly applied to enhance more specific optimization routines such as training algorithms in neural networks and recursive algorithms in system identification.
Acknowledgments

I would like to thank my advisor and friend Prof. Behnaam Aazhang for his absolute support and encouragement during the past four years. The freedom and latitude he allowed me was truly invaluable. I hope that I will maintain the same patience and enthusiasm with my students.

Additionally, I gratefully thank Prof. Don Johnson for all that he has taught me about the value of education and the elegance of science throughout the eight years that I have known him.

I especially would like to thank my friends and family for sustaining me through the odyssey of graduate school, in particular, Peter and Srini for adding humor and lightness to the whole affair and Kate for giving me the strength for the home stretch.

Finally, I gratefully acknowledge the financial support from Amoco during the course of this research.
Contents

Abstract i
Acknowledgments iii
List of Illustrations v

1 Introduction 1

2 Detection Systems 8

2.1 Analysis of Detection Systems 8

2.2 Problem Formulation 11

2.3 Importance Sampling 13

2.4 Selection of Biasing Density 15

2.5 Applications of the Linear Shift 23

2.5.1 N Dimensional Gaussian Noise 25

2.5.2 One Dimensional Laplacian Noise 27

2.5.3 Numerical Comparison 28

2.6 Summary 33

3 Multiple-Access Communications 38
3.1 Performance Analysis of Multiple-Access Systems .................................. 38
3.2 Multiple-Access Communications ............................................................. 39
3.3 Importance Sampling for CDMA Systems .................................................... 43
    3.3.1 Parametric Biasing Densities ............................................................ 45
    3.3.2 Linear Detection ................................................................................. 47
    3.3.3 Nonlinear Detection ............................................................................. 55

4 Constrained Solutions in Importance Sampling .............................................. 60
    4.1 Overview ............................................................................................... 60
    4.2 Importance Sampling for General Expectations ....................................... 64
    4.3 Distance Measures .................................................................................. 67
    4.4 Robust Statistics .................................................................................... 70
    4.5 Constrained Solutions ............................................................................. 75
    4.6 Examples of Distribution Classes ............................................................ 78
    4.7 Conclusions ........................................................................................... 92

5 Stochastic System Design ............................................................................ 96
    5.1 Introduction ............................................................................................ 96
    5.2 Development .......................................................................................... 97
    5.3 System Design ....................................................................................... 100
        5.3.1 Monte Carlo Estimates .................................................................. 101
    5.4 The Robbins-Monro Algorithm ............................................................... 103
5.4.1 Stopping Times for the Robbins-Monro Algorithm .... 110

5.5 The Kiefer-Wolfowitz Algorithm ..................... 119

6 Conclusions ........................................ 131

A Appendix ............................................ 135

A.1 Extra Conditions for Restricting D to be a Distance Function .... 135

A.2 Proof of Proposition 3 ................................ 136

B Appendix ............................................ 138

B.1 Evaluation of $W$ in the Multiple-Access Environment ........ 138

B.2 Upper and Lower Bounds on $\Gamma$ in the Multiple-Access Environment . 139

Bibliography ........................................... 141
## Illustrations

<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Figure for Theorem 1</td>
</tr>
<tr>
<td>2.2 Gain versus Error Rate</td>
</tr>
<tr>
<td>2.3 Error Rate, Gain product in 2-D Gaussian Systems</td>
</tr>
<tr>
<td>2.4 Length of Simulations using Importance Sampling</td>
</tr>
<tr>
<td>2.5 Comparison of Theoretical versus Estimated Gain</td>
</tr>
<tr>
<td>2.6 Gain versus Error Rate for N-Dimensional Laplacian Systems</td>
</tr>
<tr>
<td>3.1 Graphical Interpretation of Conditional Linear Shift</td>
</tr>
<tr>
<td>3.2 Upper and Lower Bounds on the Gain</td>
</tr>
<tr>
<td>3.3 Gain versus Error Rate in Linear Detection</td>
</tr>
<tr>
<td>3.4 Gain versus Error Rate in Hardlimiting Detection</td>
</tr>
<tr>
<td>3.5 Gain versus Error Rate in Smooth-Limiting Detection</td>
</tr>
<tr>
<td>4.1 Comparison of Least Favorable Biasing Densities</td>
</tr>
<tr>
<td>5.1 Two Dimensional Signal Detection with Constant Energy Constraint</td>
</tr>
</tbody>
</table>
5.2 Error Rate versus $\theta$ for Various SNR ................. 128
5.3 $\theta^*$ versus SNR for 2-Dimensional Laplacian Detection Systems. .... 130
Chapter 1

Introduction

System design and performance evaluation are fundamental to all areas of engineering. In almost all instances one is faced with the task of designing a system such that some performance criterion is met. Typically, the system is parameterized to simplify the design procedure. Therefore, most design problems can be formulated in one of the following two ways: determine the set of system parameters such the performance attains some prescribed level or alternately determine the set of parameters such that the system achieves its optimal performance. If the input to the system is modeled as a random process or if some of the system parameters are themselves random processes, then the performance measure frequently takes the form of the expected value of a cost function.

If we are fortunate enough to have access to a reliable value of the performance measure for every set of system parameters, we may then utilize many well known techniques to affect a solution to the design problem. (For a good review of these methods see [16] and references found within.) In non-linear systems, these optimization methods typically take the form of numerical algorithms. These algorithms may be separated into two classes: those which require the evaluation of the gradient
of the performance and those which do not. Of the algorithms which require the
knowledge of the gradient, the most widely used is the Newton-Raphson method [16].
This iterative algorithm proportionally updates the systems parameters with respect
to the product of the inverse of the gradient and the difference between the iterated
performance and the desired performance. It is well known that under quite general
conditions this algorithm will obtain the desired set of parameters.

However, it is often the case in system analysis that the gradient of the perfor­
mance measure can not be evaluated. This situation typically arises when the func­
tionality of the system is unknown or ill-defined or the statistics of the underlying
random processes make the analysis too difficult.

There are many modern algorithms in nonlinear optimization theory which do not
require the explicit knowledge of the gradient and are therefore applicable to these
situations. Some of these are the Newton-Raphson method with finite difference
gradient approximations [16], pure random search [10], simulated annealing [25, 38],
genetic based algorithms [15, 22], and the Nelder-Mead simplex algorithm [48]. This
short list is certainly not comprehensive, but these above algorithms are some of the
more current and popular.

As is frequently the case in practical systems, we unfortunately do not have direct
access to either the gradient of the performance or the performance itself. At best,
we are able to estimate these values for a given set of parameters from repeated ob-
servations of the output of the system. This situation severely limits the optimization algorithms available for the determination of the desired system parameters.

In the 1950's, the Robbins-Monro [59] and the Kiefer-Wolfowitz [37] algorithms were developed to address just such a situation. These algorithms were based on extensions of the Newton-Raphson method. The estimation noise is asymptotically eliminated by effectively averaging over the performance estimates. Iterates of the system parameters derived from the Robbins-Monro algorithm were shown to converge to the desired set of system parameters in the presence of observation noise for the design problem of attaining a specified performance level. Alternately, iterates derived from the Kiefer-Wolfowitz algorithm were shown to converge to the set of parameters which attained the extremum of the performance measure.

Both of these stochastic approximation algorithms precipitated a large body of research centering principally on determining both less restrictive conditions for convergence and the limiting distribution of the iterates about the desired set of system parameters. Moreover, it can be shown that many well known algorithms are special cases of these stochastic approximation algorithms. As two examples, the Least Mean Square (LMS) algorithm [31] can be viewed as an adaptive implementation of the Robbins-Monro algorithm and the Backpropagation algorithm for "training" neural networks is a simplified implementation of the Robbins-Monro algorithm [74].

One of the advantages of these algorithms is that they only require estimated values of the system performance at an iterated set of system parameters. Their rel-
ative simplicity and theoretical foundation has made these attractive design tools for problems where only sample outputs of the system are available. It is not surprising that the convergence properties of these algorithms are directly related to the quality and quantity of the observations of the output. Heuristically speaking, the better we are able to estimate the system performance from the output for a given set of system parameters, the better solutions we will be able to obtain in the design problem for a fixed number of iterations. Clearly, all optimization algorithms which require performance estimates would benefit from variance reduction techniques, however this thesis only addresses these techniques in the context the two stochastic approximation algorithms mentioned above.

In particular, this thesis addresses the more theoretical aspects of variance reductions techniques for the estimation of system performance and the effects of these techniques on the convergence rate of the Robbins-Monro and Kiefer-Wolfowitz algorithms. Moreover, as a prime objective of this research we seek to develop general methods for analyzing and designing decision systems. As such, the majority of examples presented in following chapters will be generic communication systems with the associated cost function being the probability of some error set.

Principal to this objective is the development of some specific tools for improving the performance estimates of single and multi-user detection systems as well as developing general tools for the improvement in the performance estimates of arbitrary
stochastic systems. The method utilized in this thesis to reduce the variance of the estimated system performance is the technique known as Importance Sampling.

This technique was first introduced by Kahn in 1956 [33]. It is a rather clever and simple modification of traditional Monte Carlo estimation methods. It is based on the idea of modifying the underlying statistics of the system such that the outputs which contribute heavily to the performance are generated with greater frequency. The estimate is then appropriately weighted such that an unbiased estimate of the system performance is obtained. It can be shown that if we intelligently choose the way we modify the statistics, a reduced variance estimate of the performance will be obtained for the same amount of computation.

The primary application of Importance Sampling has been to the analysis of communication systems [49, 52, 62, 30]. However, this technique has also been applied to the problem of estimating the probability of overflow in queuing networks [53], in estimating rare events in stochastic algorithms [12], and in estimating functionals of Markov chains [28]. The improvements obtained in these applications have varied from modest to exponential.

The fundamental issue in Importance Sampling is to determine a method for modifying the statistics of the system. It can be shown that the optimal modification of the statistics requires the knowledge of the unknown performance, therefore solutions must be obtained from constraint classes of density functions. Typically, a parametric family of density functions is chosen as the class of "biasing" densities and
then optimized such that the maximum reduction in simulation trials is obtained.

This methodology has produce the Increased Variance [62, 43], the Exponential Shift [12, 61] and the Linear Shift [49, 44] constraint classes of biasing densities.

The Linear Shift class of biasing densities is a novel strategy for modifying the underlying statistics in optimal detection systems such that under somewhat general conditions, a reduced variance estimate of the performance is guaranteed. The results presented in Chapter 2 pertaining to this biasing strategy will demonstrate that in the most common detection systems, we obtain exponential improvements over standard Monte Carlo techniques with minimal modification to the simulation. Moreover, it will be shown that the length of simulations required to estimate the performance is approximately invariant to the true value of the error rate. This is significant because in using standard techniques, the required number of simulation trials is unbounded when attempting to estimate error rates which diminish.

In Chapter 3, we extend the Linear Shift class of biasing densities for analyzing multi-user communication systems. The results obtained in this chapter are consistent with those from Chapter 2 in that the number of estimator trials are approximately invariant to the error rate. Moreover, these improvements are again achieved with an almost trivial modification to the simulation.

As discussed above, all known techniques in Importance Sampling arise from parametric classes of biasing densities. This inherently leads to ad hoc solutions. In Chapter 4 we abandon this approach and alternately determine the optimal biasing
density from an arbitrary constraint class. The solution is achieved by reformulating the problem of determining the biasing density which minimizes the estimator variance to determining the density which is nearest the optimal biasing density in a statistical sense. This enables us to apply results from robust statistics to determine sufficient conditions for a density from the constraint set to minimize the estimator variance. These conditions are shown to be equivalent to finding the density from the constraint class which is least favorable against the optimal biasing density. Moreover, the generality of this approach enables us to determine constrained optimal biasing densities for estimating arbitrary functionals of random vectors. Utilizing this theory, we then proceed to derive the constrained optimal density function from the $\epsilon$-mixture class. Results will demonstrate that in applying this biasing strategy, the performance can be made arbitrarily close to the optimal performance for every choice of cost function.

Finally, in Chapter 5 we quantify the effects of the reduction in estimation noise obtained from utilizing Importance Sampling on the Robbins-Monro and Kiefer-Wolfowitz algorithms. Results will demonstrate that in optimizing systems with respect to diminishing probabilities of events, the computational savings are unbounded. Moreover, in a signal design example supplied, the computational savings varied between $10^3$ and $10^{15}$ depending on the Importance Sampling strategy implemented and the comparison scenario.
Chapter 2

Detection Systems

2.1 Analysis of Detection Systems

In recent years, much research has centered on nonlinear communication systems. In almost all cases, analytical solutions of the performance of these systems have been very difficult to obtain. Therefore efforts have been concentrated on obtaining asymptotic approximations and bounds for estimating system error probability [56, 36, 2, 67]. The Asymptotic Relative Efficiency (ARE) and the Central Limit Theorem (CLT) have been the two most powerful tools for approximating and simplifying the performance analysis of nonlinear and non-Gaussian communication systems. Both of these approaches require infinitely large sample sizes for analytically exact solutions. This is certainly not a realistic assumption in modern communication systems. Therefore, the use of the ARE and CLT inherently yield estimates of system performance. The accuracy of these estimates have been investigated for finite sample sizes. Results showed that for certain distributions, the use of the ARE as a performance measure could be misleading [46] and recently, the Central Limit Theorem has been shown to deviate exponentially in percent error from the true system performance for a wide class of detection systems [32].
Another, often more tractable and versatile method for obtaining performance estimates is Monte Carlo simulations. This method can produce very accurate estimates of system performance if the number of simulation trials is sufficiently large. However, the drawback of this approach is that today many optimum or nearly optimum nonlinear detectors operate with a very small probability of error, $P_e$. This puts a tremendous computational burden on the computer system. Moreover, it may require that the number of random variables approach or exceed the period of the random number generator. As an example, for a 95% confidence interval of $[2P_e/5, 8P_e/5]$ there is a requirement of approximately $10/P_e$ simulation trials [75, 3]. This can be quite large for $P_e \leq 10^{-7}$, thus making conventional Monte Carlo methods presently impractical for simulating rare event systems.

A technique known as Importance Sampling has been employed to greatly reduce the number of simulation trials [33, 7, 62, 50]. This technique is based on biasing the noise distribution such that more samples are taken from the important regions (regions that cause errors) and then unbiasing the estimate of system performance with a weighting function. Within certain restrictions, it can be shown that this biasing technique will reduce the variance of the estimate, thereby reducing the number of simulation trials for the same level of accuracy. Work by Shanmugam and Balaban [62] has demonstrated the effectiveness of Importance Sampling applied to a communication system problem, but there has been no general technique developed for utilizing this powerful method. Spaulding [65] attempted to apply Importance
Sampling to the analysis of Locally Optimum Bayes Decision systems, but found no good method of biasing the noise and therefore used standard Monte Carlo simulations. Lank [43] examined a specific type of biasing (scaling), but most of the previous work on scaling has been limited to Gaussian noise in a particular system [24, 14].

In this chapter, we introduce an alternate method of biasing the noise for a reduced variance estimate of $P_e$ which will apply to a large class of detection systems and noise disturbances. Due to the scope of this problem, we restrict our investigations to minimum error rate systems. However, as will be seen in later chapters, applications of this proposed method are not limited to optimum detection systems. Moreover, it will be shown that this technique or biasing scheme allows for very easy implementation in computer simulations. We then proceed to analyze the "gain" using this biasing scheme over standard Monte Carlo simulations by comparing the number of simulation trials required by each method to obtain a fixed estimator accuracy. Results will show that in cases examined, the gain is proportional to $1/P_e$, thus leading to simulations whose lengths are invariant to the probability of interest.

The chapter is organized as follows. Section 2.2 discusses Monte Carlo simulations and formulates the signal detection problem. Section 2.3 describes the Importance Sampling method of Monte Carlo simulations in this context. The problem of choosing an appropriate biasing scheme is examined for optimal detection systems in section 2.4. We then conclude with applications of the theory of Importance Sampling to
some examples of detection systems and demonstrate the dramatic improvement over standard Monte Carlo methods.

2.2 Problem Formulation

In a coherent signal detection problem, we wish to determine between two hypotheses:

\[ H_0 : \mathbf{R} = \mathbf{S}_0 + \mathbf{n} \]  
\[ H_1 : \mathbf{R} = \mathbf{S}_1 + \mathbf{n} \]

where \( \mathbf{R} \) is the N-dimensional received vector, \( \mathbf{S}_0 \) and \( \mathbf{S}_1 \) are N-dimensional signal vectors, and \( \mathbf{n} \) is the additive noise vector. Define \( Z_0 \subseteq \mathbb{R}^N \) to be the region that \( H_0 \) is chosen and \( Z_1 = Z_0^c \subseteq \mathbb{R}^N \) to be the region that \( H_1 \) is chosen. The average probability of a wrong decision (Bayes Risk) can be written as

\[ P_e = \pi_0 \int_{Z_1} p_{\mathbf{R}|H_0}(\mathbf{r})d\mathbf{r} + \pi_1 \int_{Z_0} p_{\mathbf{R}|H_1}(\mathbf{r})d\mathbf{r} \]  

(2.3)

where \( p_{\mathbf{R}|H_0}(\mathbf{r}) \) and \( p_{\mathbf{R}|H_1}(\mathbf{r}) \) are the density functions of the received vector conditioned on the hypotheses \( H_0 \) and \( H_1 \), respectively. In (2.3), \( \pi_0 \) and \( \pi_1 \) are the a priori probabilities that \( H_0 \) and \( H_1 \) are true.

When \( \pi_0 = \pi_1 = 1/2 \) and the detector is symmetric \(^1\), the Bayes Risk can be written as

\[ P_e = \int_{\mathbb{R}^N} I(\mathbf{r})p_{\mathbf{R}|H_0}(\mathbf{r})d\mathbf{r} \]  

(2.4)

\(^1\)The extension of these analyses to asymmetric detection systems is straightforward. In asymmetric systems, \( P_e \) in (2.4) represents the false alarm probability.
where $I(r)$ is the indicator function over $Z_1$, i.e., $I(r) = 1 \forall r \in Z_1$ and $I(r) = 0$ otherwise.

The Maximum Likelihood Estimator for $P_e$ using the binomial model in the binary symmetric channel is

$$\hat{P}_e = \frac{1}{M_{MC}} \sum_{i=1}^{M_{MC}} I(R_i)$$

(2.5)

where $M_{MC}$ is the number of trials of the Monte Carlo simulation and $R_i$ is the sample vector generated from $p_{R|H_0}$. It is easily shown that when the $R_i$ are independent and identically distributed, (i.i.d.), $\hat{P}_e$ is an unbiased estimator for $P_e$ with associated variance

$$\text{var} [\hat{P}_e] = \frac{P_e(1 - P_e)}{M_{MC}}.$$  

(2.6)

Given that the estimator variance tends to zero as the length of the simulation tends to infinity, we have that $\hat{P}_e$ converges in the mean square sense to the true error rate. Additionally, this above expression shows that the variance of the Monte Carlo estimate for small $P_e$ is approximately $\frac{P_e(1 - P_e)}{M_{MC}}$. It is well known that for unbiased estimators, the variance is not only a good measure of performance but it also determines the confidence interval of the estimator. Therefore, if we could reduce the variance, then we would obtain better confidence intervals with the same level of computation. In the following sections, we will demonstrate a method to substantially reduce the variance of this estimate by biasing the received vector.
2.3 Importance Sampling

In this section the Importance Sampling method in Monte Carlo simulations is discussed. This technique reduces the variance of the estimate of $P_e$ by generating data from the biased density function $p_{R^*|H_0}$. This joint density is chosen such that the random vector $R^*$ is more likely to come from the important regions in $\mathcal{R}^N$ (regions that cause more errors). Since we will generate more errors than expected, we must weight each error such that an unbiased estimate is obtained. This biasing and weighting of events in the simulation will enable us to more accurately estimate $P_e$, i.e., obtain a reduced variance estimate of system performance.

We obtain the Importance Sampling estimator by rewriting the error rate in (2.4) as

$$P_e = \int_{\mathcal{R}^N} I(x) \frac{p_{\overline{R}|H_0}(x)}{p_{R^*|H_0}(x)} p_{R^*|H_0}(x) \, dx,$$

(2.7)

where $p_{R^*|H_0}$ is such that the above expression in finite.

The resulting sample estimate of the above expectation is given by

$$P_e^* = \frac{1}{M_{IS}} \sum_{i=1}^{M_{IS}} W(R_i^*) I(R_i^*)$$

(2.8)

where $I(R_i^*)$ is as before the indicator function over $Z_1$ with the weighting function $W$ having the form

$$W(x) = \frac{p_{\overline{R}|H_0}(x)}{p_{R^*|H_0}(x)}.$$ 

(2.9)

In the above Importance Sampling estimator, the random vectors $R_i^*$ are now generated from the biasing density $p_{R^*|H_0}$.
Note that this estimator still counts errors based upon the original decision rule of choosing $H_0$ if $R^*_i \in Z_0$ and choosing $H_1$ if $R^*_i \in Z_1$, but each error is now weighted by $W(R^*_i)$ to compensate for the biasing effect.

The variance of the importance sampling estimate of the probability of a wrong decision can be shown to be

$$\text{var} [P_e^v] = \frac{W - P_e^2}{M_{IS}}$$

(2.10)

where $W$ is defined as

$$W = \int_{R^N} p_{B|H_0}(r) p_{B|H_0}(L) dL = E_{p_{B|H_0}} [W(R)I(R)].$$

(2.11)

Thus we have a method for estimating the error rate which counts errors based on the original decision rule, however we generate data using density function $p_{E^*|H_0}$ rather than $p_{E|H_0}$. The estimate is then unbiased by scaling the errors with the weighting function $W(R)$. We must note that the variance of $P_e^*$ is dependent upon $W, P_e$, and $M$, whereas the variance of $\hat{P}_e$ is dependent upon only $P_e$ and $M$ (see the expression for variance in (2.10)). The parameter $W$ will allow for the reduction of the variance of the estimate for a fixed $M$ and for a given system as compared to standard Monte Carlo simulations. While it has been argued that in cases with finite support noise densities, the variance is sometimes a questionable quantity for comparing the accuracy of the estimate in the context of Importance Sampling [71], it will not be an issue here because we will be restricting our investigation to densities with infinite support.
2.4 Selection of Biasing Density

The fundamental problem in Importance Sampling is to determine the biasing density $p_{R^*|H_0}$ such that $\text{var} [P_e]$ is minimized for all values of $P_e$. From (2.11), we see that $W$ must be minimized with respect to $p_{R^*|H_0}$ since $Z_1$ and $Z_0$ are determined by the form of the detector. Therefore, using the Calculus of Variations, we wish to minimize the quantity

$$
\int_{Z_1} \frac{p_{R^*|H_0}(r)}{p_{R^*|H_0}(r)} p_{R^*|H_0}(r) dr + \lambda \int_{R^N} p_{R^*|H_0}(r) dr,
$$

with the condition that $p_{R^*|H_0}$ be a valid density function and where $\lambda$ is the Lagrange multiplier. The solution can be shown to be

$$
p_{R^*|H_0}^{\text{opt}}(r) = \frac{I(r)p_{R|H_0}(r)}{\int_{R^N} I(r)p_{R|H_0}(r) dr} = \frac{I(r)p_{R|H_0}(r)}{P_e}. \tag{2.13}
$$

Observe that this is a degenerative solution to the optimization problem since it requires the knowledge of the value in question. However, it is interesting to note that this density function always generates errors by restricting the random data vector $R^*$ to be an element of $Z_1$ and then weights each error by $P_e$. The variance of $P^*_e$ using $p_{R^*|H_0}^{\text{opt}}$ is easily shown to be zero, by first considering the corresponding value for $W$ via (2.11) which is given as

$$
W = \int_{Z_1} \frac{P_e}{I(r)} p_{R|H_0}(r) dr = P_e \int_{Z_1} p_{R|H_0}(r) dr = P_e^2. \tag{2.14}
$$

This demonstrates that in using Importance Sampling, there exists an $p_{R^*|H_0}^{\text{opt}}$ which enables us to have a zero variance estimate, however to generate data with this density function we need to know $P_e$. Therefore we must then restrict ourselves to determining
a suboptimal biasing density which yields a reduced variance estimate compared with conventional Monte Carlo techniques.

For Importance Sampling to be a useful method for estimating $P_e$, it is necessary that we have $\text{var} [P_e^*] \leq \text{var} [\hat{P}_e]$ or equivalently $W \leq P_e$. Conditions which satisfy this inequality are easily recognized by comparing the integral forms of $W$ and $P_e$; i.e.,

$$W = \int_{Z_1} \frac{p_{E|\mathcal{H}_0}(r)}{p_{E|\mathcal{H}_1}(r)} p_{E|\mathcal{H}_0}(r) dr \leq \int_{Z_1} p_{E|\mathcal{H}_0}(r) dr = P_e. \tag{2.15}$$

From the inequality in (2.15) it is clear that

$$p_{E^*|\mathcal{H}_0}(r) \geq p_{E|\mathcal{H}_0}(r) \forall r \in Z_1, \tag{2.16}$$

is a sufficient condition for $W \leq P_e$. Since the regions $Z_1$ and $Z_0$ are not readily available, we must develop some technique where the choice of $p_{E^*|\mathcal{H}_0}$ can be made without explicit knowledge of $Z_1$. Investigation of the sufficient condition requires further specification of the detection strategy. As mentioned earlier, in this paper we only consider biasing schemes for optimal detection systems. In this regard, the following theorem is supplied for choosing a $p_{E^*|\mathcal{H}_0}$ which guarantees a reduced variance estimate of $P_e$ for a class of minimum error rate systems.

**Theorem 2.1** Let $Z_1$ be the minimum $P_e$ decision region with $H_0 : \mathcal{H} = S_0 + \eta$ versus $H_1 : \mathcal{H} = S_1 + \eta$. Let $\tilde{Z}_1$ be the minimum $P_e$ decision region for $H_1$ with $H_0 : \mathcal{H} = S_0 + \eta$ versus $H_1 : \mathcal{H} = \alpha S_1 + (1 - \alpha)S_0 + \eta$. Then under some mild conditions on the noise density $p_\eta$, we have $\tilde{Z}_1 \supseteq Z_1$. 
To prove this theorem, we will need to explore the problem more closely. We start by restating the detection problem in (2.1) and then elaborate on the maximum likelihood decision rule.

Recall that we wish to choose between two equally likely hypotheses, $H_0$ and $H_1$. It is well established that the minimum Bayes Risk decision rule for the two above hypotheses is given by

$$L(R) = \frac{p_{R|H_1}(r)}{p_{R|H_0}(r)} \overset{H_1}{\gtrless} \frac{L(B_i)}{L(B_{i-1})} \overset{H_0}{\lesssim} 1.$$  

(2.17)

where $L(R)$ is the likelihood ratio. Note that since $\eta$ is additive noise, the likelihood ratio can be written as

$$L(R) = \frac{p_{\eta}(R - S_1)}{p_{\eta}(R - S_0)} \gtrless \frac{1}{1}.$$  

(2.18)

Under the restriction that $p_\eta(0)$ be greater than zero, we can rewrite $L(R)$ as

$$L(R) = \frac{p_\eta(R - S_1)/p_\eta(0)}{p_\eta(R - S_0)/p_\eta(0)} \gtrless 1.$$  

(2.19)

Assuming that the noise density has infinite support, the decision rule can be written as follows by taking the natural log and reversing the sign of the above inequality yielding

$$-\ln \left[ p_\eta(R - S_1)/p_\eta(0) \right] \overset{H_0}{\gtrless} -\ln \left[ p_\eta(R - S_0)/p_\eta(0) \right].$$  

(2.20)

We continue our considerations by defining a functional, based on the likelihood ratio in (2.18), as

$$D(X - Y) = -\ln \left[ p_\eta(X - Y)/p_\eta(0) \right] \forall X, Y \in \mathcal{R}^N.$$  

(2.21)
Using this functional, we can then restate the decision rule as

$$D(R - S_1) \begin{array}{c} H_0 \geq \cr H_1 \end{array} D(R - S_0). \tag{2.22}$$

In order to prove the theorem, we will need to constrain the joint density $p_\pi$ so that $D$ be a convex function. Note that from (2.21), this is equivalent to constraining $\ln [p_\pi(X)]$ to be concave over $\mathcal{R}^N$. Recall that a real-valued function $D$ is said to be convex over a convex region $C$ if, for all $X$ and $Y$ in $C$ and $\alpha \in [0, 1]$, the function satisfies

$$D(\alpha X + (1 - \alpha)Y) \leq \alpha D(X) + (1 - \alpha)D(Y). \tag{2.23}$$

In cases where the elements of $R$ are independent and identically distributed with marginal density $p_\eta$, the following condition on the marginal density can be shown to be sufficient for the convexity of $D(X)$,

$$p_\eta(\alpha X + (1 - \alpha)Y) \geq p_\eta^\alpha(X)p_\eta^{1-\alpha}(Y) \forall X, Y \in \mathcal{R}^N. \tag{2.24}$$

Now with the restriction that $D(X)$ be convex over $\mathcal{R}^N$ or that for i.i.d. noise, the marginal density of the noise satisfy the above condition, we are able to supply a proof for the theorem.

**Proof** The region that $H_1$ is chosen can be expressed in terms of the functional $D(X)$ as $Z_1 = \{R : D(R - S_1) \leq D(R - S_0)\}$. Similarly we write $\hat{Z}_1$ as $\{R :$
$D(R - (\alpha S_1 + (1 - \alpha) S_0) \leq D(R - S_0))$. Under the assumption that $D(X)$ is convex, we show that $\tilde{Z}_1 \supseteq Z_1$ or equivalently that if $R \in Z_1$, then $D(R - (\alpha S_1 + (1 - \alpha) S_0)) \leq D(R - S_0)$.

We start by writing the identity which holds for all vectors $R$:

$$D(R - (\alpha S_1 + (1 - \alpha) S_0)) = D((1 - \alpha)(R - S_0) + \alpha(R - S_1)).$$

Since we assumed that $D$ is convex and that $R \in Z_1$, we have

$$D(R - (\alpha S_1 + (1 - \alpha) S_0)) \leq (1 - \alpha)D(R - S_0) + \alpha D(R - S_1) \leq D(R - S_0). \quad (2.25)$$

We now apply this theorem to the problem of selecting $p_{R|H_0}$ such that the sufficient condition in (2.16) for a reduced variance estimate of $P_e$ is satisfied. From the theorem, we have the following relation

$$\tilde{Z}_1 = \left\{ \tau : \frac{p_2(\tau - (\alpha S_1 + (1 - \alpha) S_0))}{p_{R|H_0}(\tau)} \geq 1 \right\}$$

$$\supseteq \left\{ \tau : \frac{p_{R|H_1}(\tau)}{p_{R|H_0}(\tau)} \geq 1 \right\} = Z_1 \quad (2.26)$$

for joint density functions rendering $D$ to be a convex function. Therefore, we can observe that selecting

$$p_{R|H_0}(\tau) = p_2(\tau - (\alpha S_1 + (1 - \alpha) S_0)) \geq p_{R|H_0}(\tau) \quad (2.27)$$

fulfills the sufficient condition in (2.16) for a reduced variance estimate of $P_e$ for every $\alpha \in [0, 1]$. 

The choice of $p_{R|H_0}$ is then simply a shifted version of the original noise density; i.e., the random vector under $H_0$ is $R^* = \eta + \alpha S_1 + (1 - \alpha) S_0$. This allows for an almost trivial modification to the Monte Carlo simulations because only a simple offset is required to bias the input density. See Figure 2.1 for a graphical interpretation of this theorem.

To determine how well this choice of $p_{R|H_0}$ works for estimating $P_e$ when utilizing Importance Sampling, we will need some criterion which can be used to compare estimates based upon this method to the ones obtained from standard Monte Carlo simulations. Since Importance Sampling is generally employed to reduce the number of simulation trials, we choose to compare the number of trials required for the same estimate variance using the two methods. More precisely, we will use the performance measure [62]

$$\Gamma = \frac{M_{MC}}{M_{IS}}$$

where $M_{MC}$ is the number of samples required for a fixed variance using standard Monte Carlo simulations and $M_{IS}$ is the number of samples required for the same variance using Importance Sampling. From (2.6) and (2.10) we have

$$\Gamma = \frac{P_e - P_e^2}{W - P_e^2}.$$  \hspace{1cm} (2.29)

Clearly we would like for $\Gamma$ to be as large as possible, particularly for small $P_e$. Therefore, we wish to choose an $\alpha \in [0, 1]$ in (2.27) such that $\Gamma$ is maximized. This choice of $\alpha$ is specific to every noise density and signal to noise ratio, but $\alpha = 1/2$ will be seen to be robust for small $P_e$. We have the following proposition in support of this
claim which shows that as the error probability decreases to zero for one dimensional systems, \( \alpha = 1/2 \) yields an unbounded gain over standard Monte Carlo simulations.

**Proposition 1** When estimating the probability of the tail of a random variable \( R \) (e.g., \( \Psi_T = \Pr[R \geq T] \)) utilizing the Importance Sampling technique in Monte Carlo simulations, \( \lim_{\Psi_T \to 0} \Gamma = \infty \) for the choice of \( p_{R^*}(r) = p_R(r - T) \), assuming that \( p_R \) has infinite support.

**Proof** Consider the limiting behavior of \( \Gamma \) more closely,

\[
\lim_{\Psi_T \to 0} \Gamma = \lim_{\Psi_T \to 0} \frac{W/\Psi_T - \Psi_T}{1 - \Psi_T} \leq \lim_{\Psi_T \to 0} \frac{W}{\Psi_T}.
\]  

(2.30)
Using the integral forms for $W$ and $\psi_T$, the limit can then be written as

$$\lim_{\psi_T \to 1} \lim_{T \to 0} \frac{\int_T^\infty \frac{p_R^2(r)}{p_R(r-T)} dr}{\int_T^\infty \frac{p_R(r)}{p_R(r)} dr} = (2.31)$$

Given the added restrictions that for sufficiently large $T$

$$\frac{p_R(r + T)}{p_R(r)p_R(T)} \leq K,$$

and the derivative of the density function be absolutely integrable, then we can obtain the desired result via L'Hospital's rule.

Estimating the probability of the tail of a random variable is a common problem in statistical analyses. We apply Proposition 1 to estimating $P_e$ for a communication system of interest. In the one dimensional symmetric detection problem, the threshold is usually set at $(S_1 - S_0)/2$. This proposition, therefore, demonstrates that choosing $p_R^*(r) = \psi_\alpha(r - (S_1 - S_0)/2)$, which corresponds to the choice of $\alpha = 1/2$, is asymptotically optimum; i.e. the gain over the standard Monte Carlo estimate is unbounded.

In summary, we recall that direct minimization of the $\text{var} \left[ P_e^* \right]$ resulted in an untenable choice since it required the knowledge of $P_e$. By the use of the theorem, we were able to choose a suboptimal $p_{R^*|H_0}$ such that a reduced variance estimate of $P_e$ was obtained for a class of optimal detection systems. Finally, Proposition 1 demonstrated that asymptotically, the choice of $\alpha = 1/2$ for a one dimensional system, will dramatically reduce the number of simulation trials for a fixed variance over the standard Monte Carlo method for a broad class of marginal noise density
functions. With the understanding of the theory of Importance Sampling, we now apply this method to a few detection problems.

2.5 Applications of the Linear Shift

In this section we apply the above theory of Importance Sampling to a class of optimal detection problems. In particular, we consider signal detection in presence of noise from a family of marginal densities. The generalized Gaussian family of density functions covers a wide range of practical and much examined densities. This family has been used as a noise model in the design of many communication systems (see [47] and references found within). For efficiently simulating the performance of these detection systems, we find the value of shift $\alpha_{opt}$ that results in the minimum variance. We then proceed to analyze the gain over the standard Monte Carlo method by computing the corresponding $\Gamma$ for the most common members of the generalized Gaussian family. We start with the following proposition.

**Proposition 2** Density functions in the generalized Gaussian family with decay rate $p \geq 1$ satisfy the sufficient condition required by the theorem on the marginal density of the background noise.

**Proof** The density function for generalized Gaussian noise is given as

$$p_n(X) = \frac{p}{2\Gamma(1/p)A(p)} \exp \left\{ - \left( \frac{|X|}{A(p)} \right)^p \right\} \forall X \in \mathcal{R}^1$$  \hspace{1cm} (2.33)
where $A(p) = \left[ \sigma^2 \Gamma(1/p)/\Gamma(3/p) \right]^{1/2}$, $\sigma^2$ is the variance and $p$ is the decay rate. Recall that the condition in (2.24) is equivalent to having $\ln[p(X)]$ be concave. This requirement is satisfied for this family of density functions since $|X|^p$ is a convex function of $X$ for $p \geq 1$.

Therefore, this proposition guarantees, via the theorem, that shifting the noise under $H_0$ by $\alpha S_1 + (1 - \alpha) S_0$ will reduce the number of samples for a fixed estimator variance in optimal detection systems designed for i.i.d. noise samples using a generalized Gaussian noise model. We can also show by the following proposition that, in optimal systems based on N-dimensional Gaussian noise (possibly colored), we can again reduce the estimator variance by the application of the theorem.

**Proposition 3** Let $p_{\mathcal{E}|H_0}$ be the joint density of colored Gaussian noise, then $D(X)$ is convex over $\mathcal{R}^N$.

(For proof, see Appendix A)

By the application of these two propositions, we can easily implement a simple biasing scheme that will guarantee a reduced variance estimate in some of the most common detection systems. The question now is to determine $\alpha$ such that this variance is minimized. To solve for $\alpha_{opt}$ and the resulting $\Gamma$, we must derive an expression for $\mathcal{W}$. Since $Z_1$ is generally difficult to obtain, we can only determine $\alpha_{opt}$ and $\Gamma$ in specific examples. Therefore, we will derive these quantities for two generalized Gaussian densities, N-dimensional Gaussian and one dimensional Laplacian noise.
Results obtained on the performance of Importance Sampling estimators in these examples indicate the potential of this technique in estimating error rates of more general systems. Without loss of generality, in the following examples we use \( S_1 = S \) and \( S_0 = 0 \), where \( S \) denotes a constant vector.

### 2.5.1 N Dimensional Gaussian Noise

In this example, the additive noise in (2.1) is assumed to be a zero-mean Gaussian vector; i.e., \( \eta \) is distributed \( N(0, \Sigma) \), where \( 0 \) is the zero vector and \( \Sigma \) is the covariance matrix. It is well known that conditioned on \( H_0 \), the minimum error rate test statistic for the noise vector described above, is a zero mean Gaussian random variable with variance \( \sigma^2 = \| S \|^2_{\Sigma^{-1}} \) where \( \| \cdot \|_{\Sigma^{-1}} \) is the \( L_2 \) norm with respect to the kernel \( \Sigma^{-1} \).

Note that \( \sigma^2 \) can also be written as \( 2D(S) \) (see Appendix A). Conditioned on \( H_1 \), the test statistic is Gaussian with mean \( \bar{S} = 2D(S) \) and variance \( \sigma^2 \). The problem of solving for \( \alpha_{\text{opt}} \) then simplifies to a one dimensional minimization with respect to \( \alpha \), of the following expression for \( W \):

\[
W = \int_{\mathcal{S}/2}^{\infty} \frac{p_n(r)}{p_n(r - \alpha \bar{S})} dr
\]

where \( p_n(r) \) is \( N(0, \sigma^2) \). A simple expression can be obtained for \( W \) in terms of \( \alpha, \bar{S}, \) and \( \sigma \) as

\[
W = \int_{\mathcal{S}/2}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp \left[ -\frac{r^2}{2\sigma^2} + \frac{(r - \alpha \bar{S})^2}{2\sigma^2} \right] dr
\]
Recall that we are most concerned with the cases having small error probabilities; i.e., $P_e = Q(S/2\sigma) \ll 1/2$ or equivalently large signal to noise ratios. Therefore, using the approximation that $Q(x) \approx 1/x \rho(x)$ for large $x$, yields

$$W \approx \frac{\sigma}{\sqrt{2\pi}} (S/2 + aS)^{-1} \exp \left[ \frac{1}{2\sigma^2} (S^2 - aS^2 - S^2/4) \right].$$

Solving for $\alpha_{opt}$ over $[0,1]$ results in the following optimum value for $\alpha$:

$$\alpha_{opt} = \begin{cases} \sqrt{\frac{1}{4} + \frac{\sigma^2}{S^2}} \text{ for } \frac{\sigma^2}{S^2} \leq \frac{3}{4} \\ 1 \text{ otherwise} \end{cases}$$

Note that for large signal to noise ratios (SNR = $S^2/\sigma^2$) we have $\sigma^2/S^2 \ll 1/4$, thus making $\alpha_{opt} \approx 1/2$. It can be shown that the corresponding optimum N-dimensional shift is $\alpha_{opt}S$, (i.e., $\mathbf{R}^* = \eta + \alpha_{opt}S$, given $H_0$ is true). We can now calculate $\Gamma$ by substituting $\alpha_{opt}$ into (2.35). This yields

$$\Gamma = \frac{Q(S/2\sigma) - Q^2(S/2\sigma)}{\exp \left[ S^2/4\sigma^2 \right] Q(S/2\sigma + \sqrt{S^2/4\sigma^2 + 1}) - Q^2(S/2\sigma)}.$$

This expression for $\Gamma$ can be closely approximated as

$$\Gamma_{\text{Gaussian}} \approx \frac{2\sigma}{P_e S}$$

since in this example $P_e$ is assumed to be small.
2.5.2 One Dimensional Laplacian Noise

In this example, \( \alpha_{\text{opt}} \) and \( \Gamma \) are derived for a system with noise vector \( \eta \) of dimension one which has Laplacian probability density function and \( Z_1 \) that is determined by the optimum decision rule in (2.17). The Laplacian density function of the noise is given by

\[
p_\eta(r) = \frac{1}{\sigma \sqrt{2}} \exp \left\{ -\frac{\sqrt{2}}{\sigma} |r| \right\},
\]

where \( \sigma^2 \) is the variance of the noise sample. The optimum decision rule results in the following form for \( W \):

\[
W = \frac{1}{\sqrt{2} \sigma} \int_{-\infty}^{\infty} \exp \left\{ -\frac{2|\eta/S\sigma|}{\sqrt{2}} \right\} \exp \left\{ -\frac{2|\eta - \alpha S/S\sigma|}{\sqrt{2}} \right\} \, dr.
\]

This integral can be evaluated directly without the use of approximations. Minimizing with respect to \( \alpha \) yields

\[
\alpha_{\text{opt}} = \begin{cases} 
\frac{\ln(4)\sigma}{2\sqrt{2}S} + \frac{1}{2} & \text{for } \frac{\sigma^2}{S^2} \leq \frac{9}{2 \ln^2(4)} \\
1 & \text{otherwise}
\end{cases}
\]

with the corresponding

\[
W_{\text{min}} = 0.396 \exp \left\{ -\sqrt{2}S/\sigma \right\} = 1.584 P_e^2.
\]

Again, in examples with large values of SNR, \( \alpha_{\text{opt}} \approx 1/2 \). Calculating \( \Gamma \) for Laplacian noise using \( W_{\text{min}} \) yields

\[
\Gamma_{\text{Laplacean}} = \frac{(1 - P_e)}{0.584 P_e^2} \approx \frac{1.71}{P_e}.
\]
This expression for the optimum gain $\Gamma$ indicates that, in the Laplacian example, the improvement in performance over the standard Monte Carlo simulation is inversely proportional to $P_e$.

2.5.3 Numerical Comparison

In this subsection we numerically compare the performance of Importance Sampling estimators to that of the standard Monte Carlo estimator for the two examples considered. Figure 2.2 contains plots of the gain $\Gamma$ versus $P_e$ for both 10-dimensional i.i.d. Gaussian (i.e., $\Sigma = \sigma^2 I$) and 1-dimensional Laplacian examples. Note that $\Gamma$ is on the order of $1/P_e$ in both systems, demonstrating the effectiveness of this biasing technique in cases that standard Monte Carlo simulations require extensive computer time. In the analysis of the gain for correlated 2-dimensional Gaussian noise, we found no noticeable difference in $\Gamma$ versus $P_e$ for all values of correlation coefficient $\rho \geq 0$. However, in Figure 2.3, we illustrate one effect of $\rho$ by plotting $\Gamma \times P_e$ versus signal strength $\mathcal{S}$ for three different values of correlation coefficient. This product of gain and error rate is in one sense the proportionality constant $K(\rho)$ if we choose to write the gain as $\Gamma = K(\rho)/P_e$. We can see in this figure that as $\rho$ increases, the constant $K(\rho)$ slightly increases for all values of signal strength.

The gain $\Gamma$ is generally the most important quantity when measuring the usefulness of Importance Sampling since it indicates the reduction in the number of simulation trials. However in practice it is also useful to know the number of trials required
for a particular accuracy of the estimate. This accuracy is often expressed in terms of a confidence interval for the error probability. Recall that in standard Monte Carlo simulations, \( \frac{10}{P_e} \) trials are required for a 95% confidence interval of \([2P_e/5, 8P_e/5]\).

We can use \( \Gamma \) to find the equivalent number of trials needed in estimating \( P_e \) using Importance Sampling. This yields

\[
M_{IS} = \frac{M_{MC}}{\Gamma} = \frac{10}{P_e \Gamma}.
\]  

\[ (2.45) \]
Figure 2.3 $K(\rho)$ versus $S$ for signal detection in two-dimensional Gaussian noise, $S = [S \; S]^T$.

For the $N$-dimensional Gaussian example, using $\Gamma$ in (2.39) yields $M_{IS} = 5S\sqrt{N}/\sigma$ and for the one dimensional Laplacian example using $\Gamma$ in (2.44) results in $M_{IS} = 5.88$. Note that for Laplacian noise, the number of trails required to achieve the above 95% confidence interval is independent of $P_s$. Plots of $M_{IS}$ versus $P_s$ for the above confidence interval in the ten dimensional Gaussian and one dimensional Laplacian examples are depicted in Figure 2.4. Note that the maximum number of simulation
trials needed to achieve an estimate in the above confidence interval for $P_e \geq 10^{-20}$ is 100 in the Gaussian example and is 6 in the Laplacian example.

However, there are situations when one cannot evaluate $\Gamma$ analytically but would still like some indication of how well Importance Sampling works. In these cases, it is desired to estimate $\Gamma$ along with estimating $P_e$. We start by rewriting $W$ as

$$W = \int_{Z_1} W^2(\xi)p_{R^*|H_0}(\xi)d\xi = E_{p_{R^*|H_0}}\left[W^2(R^*)I(R^*)\right]. \tag{2.46}$$

It is direct to see that an unbiased estimator for $W$ is given by

$$W^* = \frac{1}{M} \sum_{i=1}^{M} W^2(B_i)I(B_i^*) \tag{2.47}$$

where $B_i^*$'s are generated from the biased density $p_{R^*|H_0}$. Moreover, if the original density function satisfies the condition in (2.24) and the set $Z_1$ is determined from the likelihood ratio then it can be shown that the variance of the estimator in (2.47) is less than an estimator for $W$ derived from standard Monte Carlo simulations. Utilizing this yields an Importance Sampling estimate of $\Gamma$ as

$$\Gamma^* = \frac{P_e^*(1 - P_e^*)}{W^* - P_e^{*2}}. \tag{2.48}$$

Although this is certainly not the optimal estimator of $\Gamma$, it is an intuitive heuristic since it uses unbiased estimates for each variable in the expression for $\Gamma$. The accuracy of the estimate of the gain $\Gamma$ is accessed by comparing plots of $\Gamma$ and $\Gamma^*$ as a function of error probability in Figure 2.5 for ten dimensional Gaussian and one dimensional Laplacian examples using 100 samples per estimate. From this figure we can see that
\( r \) in (2.48) is quite accurate for all error probabilities of interest.

Since a "good" estimate of \( r \) is available, we are better able to evaluate the effectiveness of biasing the noise by shifting the original density for more complicated detection systems. In these cases, we will need to use a robust value for \( \alpha \) since analytical expressions for \( \mathcal{W} \) may not be available for deriving \( \alpha_{opt} \). Recall from Proposition 1 that the choice of \( \alpha = 1/2 \) resulted in an asymptotically ideal gain in one dimensional systems. Moreover, in the two examples in this section, we determined that \( \alpha_{opt} \) approaches 1/2 as \( P_e \) decreases to zero. These observations seem to imply that the choice \( \alpha = 1/2 \) is somewhat robust. We will therefore use this value in the following illustrations.

Here we examine the gain in using Importance Sampling over the standard Monte Carlo technique for higher order detection systems in Laplacian noise. Figure 6 contains plots of \( r^* \) versus \( P_e^* \) for one, five, and ten dimensional optimal systems. Note that the gain decreases with an increase in system order, however, there is a substantial improvement over the standard Monte Carlo method even in the 10-dimensional example. Considering these results, Figures 2.2, 2.4, and 2.6 show that choosing \( p_{R^*|H_0} \) as an appropriately shifted version of \( p_{R|H_0} \) will dramatically reduce the number of simulation trials for systems having marginal densities satisfying the sufficient conditions in (2.24).
2.6 Summary

In this chapter, an Importance Sampling method in Monte Carlo simulations was introduced. This method was shown to dramatically reduce the number of simulation trials for a given fixed accuracy if the noise density is appropriately biased. It was shown that shifting the noise density by an appropriate amount will guarantee a reduced variance estimate of $P_e$ for optimal detection systems under some conditions. It is only required that the function $D(\cdot)$ defined in (2.21) be convex, or that the natural log of the marginal density be a concave function in the i.i.d. case. This biasing technique allows for very easy implementation on the computer since only a shifted version of the original data vector is required. In general, the optimum shift value cannot be determined. However, in the cases considered, it appears that shifting by half the signal is robust for small $P_e$.

Comparisons made between the Importance Sampling and the standard Monte Carlo simulation techniques were based on the ratio of the number of simulation trials required for a fixed accuracy. The gain in biasing the noise density was computed for a few optimal detection systems. For cases where an expression for the gain was not available, a sample estimate was obtained. The accuracy of the estimate of the gain was verified for a few examples. In all systems considered, the gain in the number of simulation trials required for a fixed accuracy over the standard Monte Carlo was approximately inversely proportional to the error probability. This indicates that the savings in the number of trials is quite substantial for cases where standard Monte
Carlo simulations require extensive computer time. Moreover, this result implies that the length of the simulations can be made invariant to the error rate. This can be advantageous when one has no insight into the required number of simulation trials.

In the following chapter, we will extend these results to the analysis of multi-user communication systems. These results also indicate the effectiveness of the linear shift biasing density as a biasing strategy.
Figure 2.4 Number of simulation trials versus $P_e$ for a 95 percent confidence interval of $[2/5 P_e, 8/5 P_e]$ for one-dimensional Laplacian and 10-dimensional i.i.d. Gaussian systems.
Figure 2.5 Gain and estimated gain for one-dimensional Laplacian and 10-dimensional i.i.d. Gaussian systems.
Figure 2.6 Estimated gain versus $P_e^*$ from simulations for 1, 5, and 10-dimensional i.i.d. Laplacian systems.
Chapter 3

Multiple-Access Communications

3.1 Performance Analysis of Multiple-Access Systems

Having developed a general approach for efficiently estimating performance in single-user systems, we look now to extend these results to more general communication systems. Specifically, in this chapter we consider the problem of determining the average bit error rate in code-division multiple-access (CDMA) communication systems. Due to their inherent complexity, the analysis of these systems is typically intractable. Presently, analytic solutions for the error rates exists only under the simplest detector structures and noise models.

The linear detector (correlator) has been widely applied to CDMA communications. Exact error rates have been determined for this detector [4, 58]. The analysis involved in evaluating these probabilities is computationally expensive. As a consequence of the linear detector's relatively poor performance, more sophisticated detection schemes are required for reliable communications. Exact error rates exist for only a few of these detectors [2, 69].

In general, complex receivers require the use of upper and lower bounds or asymptotic approximations to obtain some measure of performance. Bounds have been de-
rived for the optimal detector [70] and the near-optimum multistage detector [69]. Moment Space techniques [77] have been successfully used to bound error rates of conventional and optical CDMA systems. Statistical approximations have also been extensively used in lieu of exact analysis [3, 20]. As always, in applying these bounds and approximations, the issue of accuracy must be examined.

Invariably in practical situations, simulations are employed for estimating the probability of error in CDMA systems [3, 1, 69]. However, as new and more complicated detection schemes are derived, there is an increased computational burden placed on the simulation. This burden can inhibit the use of simulations altogether. Thus, variance reduction techniques are required to facilitate the use of simulations when the error rates are quite small. Therefore, it is natural that we extend our results in Importance Sampling from Chapter 2 [49] to the problem of estimating performance of multiple-access systems. As such, we supply examples of both non-linear systems and non-Gaussian background noise. It will be shown that for the systems investigated, the length of the simulation can be made approximately invariant to the error rate, thus rendering exponential improvement over Monte Carlo simulations.

3.2 Multiple-Access Communications

In the multiple-access environment, the received signal is modeled as

$$r(t) = n_t + \sum_{k=1}^{K} \sqrt{2P_k b_k(t - \tau_k)} a_k(t - \tau_k) \cos(\omega_c t + \phi_k), \quad (3.1)$$
where \( n_t \) represents the channel noise, \( P_k \) is the power received from the \( k \)th user, \( \omega_c \) is the carrier frequency common to all \( K \) signals, and where \( \phi_k \) and \( \tau_k \) are the phase and delay, respectively, of the \( k \)th signal at the receiver. The continuous signal representing the \( k \)th user's binary data sequence is \( b_k(t) \), which is a sequence of unit-amplitude, positive and negative, rectangular pulses of duration \( T \). This data waveform is modulated by a phase-coded carrier. The code waveform \( a_k(t) \) is generated by the spreading sequence assigned to the \( k \)th user. This signal can be written as

\[
a_k(t) = \sum_{j=-\infty}^{\infty} a_j^{(k)} P_{T_c}(t-jT_c), \quad k = 1, 2, \ldots, K
\]

where \( a_j^{(k)} \in \{-1, +1\} \) and also \( a_j^{(k)} = a_{j+N}^{(k)} \) for all \( j \) and \( k \) and for some integer \( N \). It is also assumed that \( N \) is the minimum period of the spreading sequences. The parameter \( T_c \) is the chip length, and we assume that \( T = NT_c \) so that there is one code period \( a_k = (a_1^{(k)}, a_2^{(k)}, \ldots, a_N^{(k)})' \) per data symbol.

For each signal in (3.1), there is an associated delay \( \tau_k \) for a given receiver. This delay accounts for the propagation delay and the lack of synchronism between transmitters. Without loss of generality, we assume that the detection system is synchronous with the first user, i.e., \( \tau_1 = 0 \) and \( \phi_1 = 0 \) and all phases and delays are relative to user 1. Furthermore, there is no loss of generality in assuming \( \phi_k \in [0, 2\pi) \) and \( \tau_k \in [0, T) \) for \( 2 \leq k \leq K \) since we are concerned only with the time delays modulo \( T \) and phase shifts modulo \( 2\pi \).
Conversion from a continuous-time waveform to a length $N$ discrete-time signal is accomplished via downconversion by $\cos(\omega_c t)$ followed by an integrate and dump sampled every $T_c$ seconds. If the noise process is white with two sided spectral density $N_0/2$, the elements of the corresponding discrete vector have the following form:

$$R_j = \eta_j + \sqrt{\frac{P_1}{2} T_c} b^{(1)}(1) + I_j, \quad j = 1, 2 \cdots N$$  \hspace{1cm} (3.3)

where the samples of the filtered noise process $\eta_j$ are assumed to be independent and identically distributed with zero mean and variance $\sigma^2 = \frac{N_0 T_c}{4}$, and $b^{(1)}$ is the first user's data bit, and where the samples of the accumulated multiple-access interference are defined as

$$I_j = \sum_{k=1}^{K} \sqrt{\frac{P_k}{2}} \cos(\phi_k) \int_{jT_c}^{(j+1)T_c} b_k(t - \tau_k) a_k(t - \tau_k) dt. \hspace{1cm} (3.4)$$

Defining $\theta = \sqrt{\frac{P_k}{2} T_c}$ we arrive at the discrete-time hypotheses testing problem for determining the parity of the first user's bit as

$$H_0 : \quad R = \eta - \theta a_1 + I$$  \hspace{1cm} (3.5)

$$H_1 : \quad R = \eta + \theta a_1 + I$$  \hspace{1cm} (3.6)

where $I$ is a function of the delays $\tau = (\tau_2, \tau_3, \cdots, \tau_K)'$, phases $\phi = (\phi_2, \phi_3, \cdots, \phi_K)'$ and the interfering user's bits $\vec{b} = (b_2, b_3, b_4, \cdots, b_K)'$. It is assumed that these random vectors are statistically independent from one another. In this context, any "one shot" detector (i.e., detectors which make decisions based entirely on the vector at hand)\cite{57} can be described as a functional which estimates the value
of the first user's bit, i.e., \( \hat{b}^{(1)} = G(R) \) where the functional \( G : \mathcal{R}^N \rightarrow \{-1, +1\} \) represents the detector. Using this formalism and incorporating the a priori statistics of the bit \( b^{(1)} \) into \( R \), the average bit error probability is written as

\[
P_e = Pr \left[ G(R) \neq b^{(1)} \right] = \int_{\mathcal{R}^N} I(\|G(r) - b^{(1)}\|) P_R(r) d\tau \tag{3.7}
\]

where \( I(x) = 1 \) if \( x > 0 \) and \( I(x) = 0 \) if \( x \leq 0 \). The error rate can also be formulated in terms of the underlying random vectors \( \eta, \xi, \phi \) and \( \hat{b} \) where \( \hat{b} = (b^{(1)}, \hat{b})' \). We begin by defining both \( \rho = (\xi, \phi, \hat{b}) \) and \( \tilde{\rho} = (\xi, \phi, \tilde{b}) \) with the notation that \( p_{\rho} \) and \( p_{\tilde{\rho}} \) be equivalent to \( p_{\xi, \phi, \hat{b}} \) and \( p_{\xi, \phi, \tilde{b}} \), respectively. With this notation we can rewrite (3.7) as

\[
P_e = \int_{\mathcal{R}^N} \int_{\mathcal{R}^N} I(\|G(r) - b^{(1)}\|) p_{\eta, \xi, \phi}(\eta, \xi, \phi) d\rho d\eta. \tag{3.8}
\]

In most practical situations, the integrals in (3.7) and (3.8) cannot be evaluated analytically. Only the simplest functional forms of \( G \) render a tractable integral. Consequently, Monte Carlo simulations are often employed as a convergent method of estimating \( P_e \). Recall that Monte Carlo is a technique which is based upon averaging the functional over the input space via simulation. More specifically, the Monte Carlo estimate of the error rate in the multiple-access environment is given by

\[
\bar{P}_e = \frac{1}{M_{MC}} \sum_{i=1}^{M_{MC}} I(\|G(R_i) - b^{(1)}_i\|) \tag{3.9}
\]

where it is required that \( R_i \) are independent and identically distributed incidences of the random vector \( R \) generated from distribution function \( P_R \). \( M \) is the number of simulation trials, and \( b^{(1)}_i \) is the \( i \)th realization of the first user's data bit. It is easily
shown that $\overline{P}_e$ is an unbiased estimator whose variance is given by

$$\text{var} [\overline{P}_e] = \frac{P_e \cdot P_e^2}{M_{MC}}. \quad (3.10)$$

It has been shown [49] that $M = \frac{10}{P_e}$ trials are required for a confidence interval defined by $\text{Pr} (\overline{P}_e \in \left[\frac{2}{5} P_e, \frac{6}{5} P_e\right]) = 0.95$. For multiple-access systems which operate at low error rates, $P_e \leq 10^{-6}$, the number of simulation trials $M$ can be on the order of the period of the random number generator. In order to circumvent this problem, we employ the technique of Importance Sampling to reduce the length of the simulation.

### 3.3 Importance Sampling for CDMA Systems

Recall from the previous chapter that Importance Sampling is a simulation technique which, if appropriately applied, can reduce the length of the simulation for a given accuracy. This approach is based upon modifying the simulation such that vectors which cause errors are more frequently generated, and then the estimate is accordingly weighted such that an unbiased estimate is obtained. This technique was first applied to the analysis of communication systems in [62] and subsequently investigated in [24, 43, 44, 49, 51] and references therein.

To obtain the multiple-access Importance Sampling estimator, we begin by rewriting the analytic expression for $P_e$ in (3.7) as

$$P_e = \int_{\mathbb{R}^N} I(|G(\tau) - b^{(1)}|) \frac{p_{R}(\tau)}{p_{R^*}(\tau)} d\tau, \quad (3.11)$$
where \( p_{R^*} \) is again known as the "biasing density" and is chosen such that the likelihood ratio between \( R \) and \( R^* \) is well defined. From this formulation, we obtain the unbiased Importance Sampling or weighted estimate of the error rate as

\[
P^*_e = \frac{1}{M_{IS}} \sum_{i=1}^{M_{IS}} \mathcal{I}(|G(R^*_i) - b^{(i)}|)W(R^*_i),
\]

(3.12)

where the so called weighting function \( W(\cdot) \) is written as

\[
W(R^*_i) = \frac{p_R(R^*_i)}{p_{R^*}(R^*_i)}
\]

(3.13)

and where the random vector \( R^*_i \) is generated from the distribution function \( P_{R^*} \).

As was previously shown in (2.10) and (2.11) [49], the variance of the Importance Sampling estimator in (3.12) is given by

\[
\text{var} [P^*_e] = \frac{W - P^2_e}{M_{IS}},
\]

(3.14)

where the average weight \( \bar{W} \) is

\[
\bar{W} = \int_{R^*} \mathcal{I}(|G(\tau) - b^{(i)}|)W_\tau p_R(\tau)d\tau.
\]

(3.15)

Again, we may formulate the Importance Sampling estimator in (3.12) in terms of the underlying random vectors. By applying the technique used in (3.11) to the expression for the error rate in (3.8) we obtain the general version of the multiple-access Importance Sampling estimate as

\[
P^*_e = \frac{1}{M_{IS}} \sum_{i=1}^{M_{IS}} \mathcal{I}(|G(R^*_i) - b^{(i)}|)W''(y^*_i, \mathcal{E}^*_i)
\]

(3.16)
where

\[ W'(\eta^*, \xi^*) = \frac{P_{\eta^*} \xi^*}{P_{\eta^*} \xi^*} \quad \text{(3.17)} \]

and where \( \eta^* \) and \( \xi^* \) are generated from the joint distribution function \( P_{\eta^*} \xi^* \).

The fundamental issue in Importance Sampling is to determine the biasing distribution \( P_{R^*} \) or equivalently the joint distribution \( P_{\eta^*} \xi^* \) such that the variance of the Importance Sampling estimator is minimized. It is well known [51] that the unconstrained optimal distribution is dependent upon the error rate, therefore constrained solutions must be derived.

### 3.3.1 Parametric Biasing Densities

The linear shift class of biasing densities introduced in the previous chapter has been employed as a density constraint class for estimating error rates in communication systems [44, 51, 49]. These density functions are of the form

\[ p_{R^*}(\xi^*) = P_{R}(\xi^* - \nu) \quad \nu \in \Upsilon, \quad \text{(3.18)} \]

where \( \Upsilon \) is some constraint set in \( \mathcal{R}^N \). As was shown [49], under mild restrictions on the background noise, this biasing scheme guarantees a reduced variance estimate of \( P_e \) for optimal single user systems. The asymptotically optimum value of \( \Upsilon \) corresponds to the location on the decision boundary which is the midpoint between the two signals, thereby generating errors with a probability of 1/2 [49]. Moreover, it was shown that when the underlying joint distributions are Gaussian or Laplacian, the reduction in simulation trials is exponential. No more than 100 simulation trials
were required to obtain the above confidence interval when \( P_e \geq 10^{-20} \) in the most common communication systems. Therefore, we attempt to extend the linear shift class to the multi-user environment.

In order to apply the existing theory for the linear shift class of biasing densities we must impose some restrictions on the statistics of the multiple-access interference. Typically, the multiple-access parameters are assigned uniform density functions, however in our analysis, this will not be required. In order to make the problem of determining the biasing density tractable, we impose the restriction that \( p_{e^*} = p_e \). Applying this condition and recalling the \( \eta \) and \( \rho \) are assumed to be statistically independent it is easily shown that the weighting function in (3.17) can be viewed as a conditional weighting function which is given by

\[
W'(\eta^*, \rho^*) = \frac{p_{a'}(\eta^*)}{p_{a'\eta^*}(\eta^*, \rho^*},
\]

where \( p_{a'\eta^*} \) is the conditional density function of the biased random vector \( \eta^* \) given the multiple-access parameters.

From this we are left with only having to determine the above conditional density function to fully specify an Importance Sampling scheme. As discussed above, we attempt to extend our results from the linear shift class of density functions to the selection of \( p_{a'\eta^*} \), therefore we choose to constrain the conditional density to be of the form

\[
p_{a'\eta^*}(\eta^* | \rho^*) = p_a(\eta^* - m(\rho^*) | \rho^*),
\]
where \( m(\rho^*) \) is the conditional mean of \( \eta^* \). In this way, the shift of the mean of background noise vector will be dependent upon the multiple-access parameters. The motivation for this can be seen by viewing Figure 3.1. For two users \((K = 2)\) operating synchronously we obtain a signal constellation with four possible signal locations corresponding to the four possible values for \( \textbf{b} \). In order to shift the signals to the decision boundary, the value of the shift \( m(\rho^*) \) must be dependent upon the location of the signal in the constellation, i.e., the multiple-access parameters.

The value of the shift will also depend upon the structure of the detector or equivalently the decision boundary in the signal space. Therefore it will be necessary to investigate a number of different detectors to gain insight into this problem. As that the linear detector in Gaussian background noise is the most common multiple-access scenario, we choose first to investigate it thoroughly.

### 3.3.2 Linear Detection

The form of the general linear detector is given by

\[
Y = \langle \mathbf{R}, \mathbf{a} \rangle_{\Sigma} \begin{cases} H_1 & \gamma, \\ H_0 & \leq \end{cases}
\]

(3.21)

where \( Y \) is the test statistic, \( \Sigma \) is any positive definite symmetric kernel and \( \gamma \) is the threshold. In the duration of this paper, we will omit the kernel from the notation of the inner product, therefore it is to be understood that \( \langle \cdot, \cdot \rangle_{\Sigma} = \langle \cdot, \cdot \rangle \). Also, in order
to facilitate analysis, we set $\gamma = 0$. From (3.21), the test statistic can be written as

$$Y = \langle \eta, a_1 \rangle + \langle L, a_1 \rangle + \theta b^{(1)}(a_1, a_1).$$  \hspace{1cm} (3.22)

It is clear that $Y$ is conditionally Gaussian, i.e., $Y$ conditioned on the multiple-access parameters $\rho$ is $\text{Normal}(\langle L, a_1 \rangle + \theta b^{(1)}||a_1||^2, \sigma^2||a_1||^2)$. It was previously shown [49] that for small error rates, the conditional optimal shift for modifying the simulation based on the output statistic $Y$ is given by $E_n[-Y|\rho]$, that is if

$$Y^* = \langle \eta, a_1 \rangle + \langle L, a_1 \rangle + \theta b^{(1)}(a_1, a_1) + m(\rho^*),$$  \hspace{1cm} (3.23)

then the optimal value of the scalar shift $m(\rho^*)$ is $-\langle L, a_1 \rangle - \theta b^{(1)}(a_1, a_1)$, thus rendering the biased test statistic as $Y^* = \langle \eta, a_1 \rangle$. From this it is readily seen that $Y^*$ is distributed $\text{Normal}(0, \sigma^2||a_1||^2)$ and is independent of $\rho^*$.

We must now determine the corresponding vector valued shift $m(\rho^*)$ in (3.20). It is well known that there is a continuum of vectors $m(\rho^*)$ such that $m(\rho^*) = \langle m(\rho^*), a_1 \rangle$. However, only one such vector $m(\rho^*)$ renders a consistent simulation, i.e., only one vector exists such that error rates estimated from the input version ($R$ and $R^*$) of the simulations are equivalent to error rates estimated from the output version ($Y$ and $Y^*$) of the simulations. To determine this vector, we find $m(\rho^*)$ such that the corresponding conditional weighting functions are equivalent.

---

1The extension to the case with non-zero threshold is straightforward.

2$\text{Normal}(\mu, \sigma^2)$ indicates that it is a Gaussian random variable with mean $\mu$ and variance $\sigma^2$. 
The conditional weighting function based on the output test statistic is given by

$$W(Y^*, \rho^*) = \frac{p_{Y^*|\rho^*}(Y^*)}{p_{Y^*|\rho^*}(\rho^*)} = \exp \left[ -\frac{1}{2\sigma^2} \left\{ (Y^* - \theta b^{(1)}(a_1, a_1) - (L, a_1))^2 - Y^*^2 \right\} \right]. \quad (3.24)$$

Using the fact that $Y^* = (\eta, a_1)$ and simplifying we obtain

$$W(Y^*, \rho^*) = \exp \left[ \frac{2(\eta, a_1) \left( \theta b^{(1)} + \frac{L(a_1)}{a_1, a_1} \right) - \theta^2(a_1, a_1) - 2\theta b^{(1)}(L, a_1) - \frac{(L, a_1)^2}{a_1, a_1}}{2\sigma^2} \right]. \quad (3.25)$$

In order to determine the corresponding input version of the weight function we note that the biased received vector is given by

$$R^* = \eta + \theta b^{(1)} a_1 + L + m(\rho^*). \quad (3.26)$$

This can be written in terms of a biased version of the background noise as $R^* = \eta^* + \theta b^{(1)} a_1 + L$ where $\eta^* = \eta + m(\rho^*)$. From this we can compute the input version of the conditional weighting function as

$$W'(\eta^*, \rho^*) = \frac{p_{\eta}(\eta^*)}{p_{\eta^*|\rho^*}(\eta^*|\rho^*)} = \exp \left[ -\frac{1}{2\sigma^2} (\eta^*, \eta^*) + \frac{1}{2\sigma^2} (\eta^* - m(\rho^*), \eta^* - m(\rho^*)) \right]. \quad (3.27)$$

Again simplifying, we obtain that

$$W'(\eta^*, \rho^*) = \exp \left[ \frac{-2(\eta^*, m(\rho^*)) - (m(\rho^*), m(\rho^*))}{2\sigma^2} \right]. \quad (3.28)$$

Applying the simulation consistency requirement that

$$W'(\eta^*, \rho^*) = W(Y^*, \rho^*), \quad (3.29)$$
it can be shown that the corresponding vector valued shift \( m(\rho^*) \) is given by

\[
m(\rho^*) = -\theta b(1) \bar{a}_1 - \frac{\langle L, \bar{a}_1 \rangle}{\langle \bar{a}_1, \bar{a}_1 \rangle} \bar{a}_1. \tag{3.30}
\]

It is easily verified that \( m(\rho^*) \) is mapped to \( m(\rho^*) \) via the inner product with \( \bar{a}_1 \). This conditional vector valued shift is depicted in Figure 3.1. It should be noted that any \( m(\rho^*) \) which shifts the signal constellation to the decision boundary will render \( Y^* \) distributed as Normal(0, \( \sigma^2 ||\bar{a}_1||^2 \)), however the unique value of \( m(\rho^*) \) which satisfies the consistency condition is given in (3.29). Therefore the optimal conditional biasing density for the linear shift class in (3.20) under linear detection is given by

\[
p_{\rho^*}(\rho^* | L^*) = p_{\rho}(\rho^* + \theta b(1) \bar{a}_1 + \frac{\langle L, \bar{a}_1 \rangle}{\langle \bar{a}_1, \bar{a}_1 \rangle} \bar{a}_1). \tag{3.31}
\]

It is necessary now to determine the performance of this biasing scheme. A previously used measure of performance [49] is the so called gain, \( \Gamma \) given by

\[
\Gamma = \frac{M_{MC}}{M_{IS}} = \frac{P_e - P_e^2}{W - P_e^2} \approx \frac{P_e}{W^*}, \tag{3.32}
\]

where \( M_{MC} \) and \( M_{IS} \) are the required number of simulation trials for a given variance using standard Monte Carlo and Importance Sampling simulations, respectively. As can be seen, in order to evaluate \( \Gamma \) we must determine both \( P_e \) from (3.11) and \( W^* \) from (3.15). It can be shown that when the eye open condition is satisfied, that is \( \theta(\bar{a}_1, \bar{a}_1) + \langle L, \bar{a}_1 \rangle \geq 0 \) for all values of \( \tilde{\rho} \), the error rate for the linear correlator is given by

\[
P_e = E_{\tilde{\rho}} \left[ Q \left( \frac{\theta(\bar{a}_1, \bar{a}_1) + \langle L, \bar{a}_1 \rangle}{\sigma ||\bar{a}_1||} \right) \right] \tag{3.33}
\]
where the expectation is with respect to the multiple-access parameters. As to determining $\mathcal{W}$, it can be shown (see Appendix B) that

$$\mathcal{W} = E_{\mathcal{E}} \left\{ \exp \left[ \frac{(\theta(a_1, a_1) + (L, a_1))^2}{\sigma^2 ||a_1||^2} \right] Q \left( \frac{2(\theta(a_1, a_1) + (L, a_1))}{\sigma ||a_1||} \right) \right\}. \quad (3.34)$$

Therefore the gain $\Gamma$ is closely approximated by

$$\Gamma \approx \frac{P_e}{\mathcal{W}} = \frac{E_{\mathcal{E}}[Q(X)]}{E_{\mathcal{E}}[\exp(X^2)Q(2X)]}, \quad (3.35)$$
where $X = \frac{\theta(\alpha, \beta) + \theta(\beta, \alpha)}{\|\alpha\| \|\beta\|}$ with the condition that $X \geq 0$ for all $\alpha$. Since the analytic evaluation of $\Gamma$ is intractable, upper and lower bounds for $\Gamma$ are obtained in Appendix B as

$$\sqrt{\frac{2}{\pi}} \frac{1}{\lambda X} \leq \Gamma \leq \sqrt{\frac{2}{\pi}} \frac{1}{\min_X \{P_e|\rho\}} \left( \max_{\lambda} \{P_e|\rho\} \right), \quad (3.36)$$

where $X$ is defined as above and $P_e|\rho$ is the conditional error rate for given values for $\rho = (\lambda, \phi, b)$. These bounds are depicted in Figure 3.2 for the asynchronous case with $K = 2$, $N = 31$. It can be seen that in this example, the lower bound can be considered to be tight. It should be noted that for the single user case, the upper and lower bounds given above are equivalent and are consistent with the results obtained in the previous chapter.

Recall from (3.32) that $M_{IS} = M_{MC}/\Gamma = 10/(P_e \Gamma)$ for the 95% confidence interval previously supplied. Using the above lower bound on the gain we obtain as a maximum number of simulation trials for estimating the performance of CDMA systems under linear detection as

$$M_{IS} \leq \frac{10 \max_{\lambda} \{P_e|\rho\} \max_{\lambda} \{X\}}{\sqrt{\frac{2}{\pi} P_e}} \approx \frac{C \max_{\lambda} \{P_e|\rho\}}{P_e}, \quad (3.37)$$

where $C$ is some constant. It can be inferred from (3.37) that as the spreading codes approach orthogonality, that is the ratio of the maximum error rate to the average error rate approaches unity, we then have $\Gamma \approx C'/P_e$. Therefore, regardless of the error rate, the number of simulations is upper bounded by a fixed constant. This is verified in Figure 3.3 where $\log(\Gamma)$ is plotted versus $\log(P_e)$ for the linear detector.
operating in Gaussian background noise. In this example the length of the spreading codes \( N \) is 31. The slope of the curve of \( \log \Gamma \) versus \( \log P_e \) is closely approximated by \(-1\), verifying that the gain is inversely proportional to the error rate even in the finite \( N \) case.

It should be noted that this is as well as any suboptimum Importance Sampling scheme can work since it implies that \( W = C''P_e^2 \) where \( C'' \) is a bounded constant greater than 1. Any superior suboptimum scheme would only reduce the value of the constant given by the error rate - gain product.

\[ \begin{align*}
\log(\Gamma) & \quad 40 \\
35 & \quad 30 \\
25 & \quad 20 \\
15 & \quad 10 \\
10 & \quad 5 \\
0 & \quad 0 \\
\end{align*} \]

\[ \begin{align*}
\text{SNR(dB)} & \quad 8 \quad 10 \quad 12 \quad 14 \quad 16 \\
\end{align*} \]

**Figure 3.2** Upper and lower bounds for \( \Gamma \) for the case of two user's operating asynchronously \((N = 31)\). The estimated gain for this case is presented for comparison.
In the case that the background noise is non-Gaussian, analytic solutions for the gain $\Gamma$ are intractable. Moreover, obtaining reasonable bounds on $\Gamma$ is still an open issue in the Importance Sampling literature. Therefore, we must estimate the gain from the simulation itself. Recall that a previously established estimate for $\Gamma$ is given as follows:

$$
\Gamma^* = \frac{P_e^*(1 - P_e^*)}{W^* - P_e^2},
$$

(3.38)

where $P_e^*$ is the Importance Sampling estimate of the error rate and $W^*$ is the Importance Sampling estimate of the average weight given by

$$
W^* = \frac{1}{M_{IS}} \sum_{i=1}^{M_{IS}} I(|R_i^* - b_i^{*\text{(1)}}|)W^{*2}(q_i^*, l_i^*).
$$

(3.39)

Using this we can determine the performance of the biasing scheme given in (3.31) while estimating the performance of linear detectors in the presence of non-Gaussian noise via simulations. We choose in this paper to investigate the situation where $\eta$ is jointly Laplacian, i.e.,

$$
p_{\eta_j}(\eta) = \frac{1}{\sqrt{2\sigma^2}} e^{-\frac{1}{2\sqrt{2}\sigma^2}|\eta|} \quad \forall \eta.
$$

(3.40)

The simulations were performed using 5000 trials irrespective of the error rate. The results are presented in Figure 3.3. Again it is apparent that the gain is inversely proportional to the error rate as the slope of $\log(\Gamma^*)$ versus $\log(P_e^*)$ is closely approximated by $-1$. Therefore, in this non-Gaussian example, the length of the simulation is again approximately independent of $P_e$. 
3.3.3 Nonlinear Detection

It is well known that the performance of the linear detector is severely degraded by an increase in the number of users ($K$) and the presence of impulsive noise. Therefore, nonlinear correlation has been employed to combat these interferences [3]. The test statistic for the general nonlinear correlator is of the form

$$Y = \sum_{i=1}^{N} g(Z_i) a_i^{(t)}$$

(3.41)

where $g(\cdot)$ is typically an odd-symmetric bounded function. Previously investigated choices for $g(\cdot)$ are

**Hardlimiter [2]:**

$$g(x) = \begin{cases} +1 & x \geq 0 \\ -1 & x < 0 \end{cases}$$

(3.42)

**Smooth-limiter [3]:**

$$g(x) = \frac{2}{k} \left[ \frac{1}{2} - Q(\frac{x}{\xi}) \right]$$

In a previous study [3], Monte Carlo simulations were required for a direct comparison of linear detection to both the hardlimiter and smooth-limiter. Therefore we would like to apply results obtained in prior sections of this chapter to the problem of simulating these nonlinear detectors in both Gaussian and Laplacian noise. We again choose to employ the conditional shift from (3.30) as our means of biasing the simulation. The results are presented in Figures 3.4 and 3.5. Again, the slope of \(\log(\Gamma^*)\) versus \(\log(P^*_e)\) is closely approximated by $-1$ for both nonlinearities in the presence of Gaussian and non-Gaussian background noise. We can then infer from
this that the conditional shift appears to be somewhat robust to the function $g(\cdot)$ and the type of background noise in that under this biasing scheme, the length of the simulation is approximately constant regardless of the error rate. It is also believed that the conditional shift developed here will find applications in other types of multiple-access demodulation [70, 69] and to the analysis of $M$-ary systems.
Figure 3.3 $\log(\Gamma^*)$ versus $\log(P_e^*)$ for the linear detector in the presence of both Gaussian and Laplacian background noise. Curves are supplied for one, two and three users. (N=31)
Figure 3.4 $\log(\Gamma^\ast)$ versus $\log(P_e^\ast)$ for the hardlimiting correlator in the presence of both Gaussian and Laplacian background noise. Curves are supplied for one, two and three users. (N=31)
Smooth-limiter Correlator

Gaussian Noise

![Diagram showing log(Γ*) versus log(P* c) for different values of K (1, 2, 3) with Gaussian noise.]

Laplacian Noise

![Diagram showing log(Γ*) versus log(P* c) for different values of K (1, 2, 3) with Laplacian noise.]

Figure 3.5  $\log(\Gamma^*)$ versus $\log(P^*_c)$ for the smooth-limiting correlator in the presence of both Gaussian and Laplacian background noise. Curves are supplied for one, two and three users. (N=31)
Chapter 4

Constrained Solutions in Importance Sampling

4.1 Overview

Our interest in the preceding chapters of this thesis was to determine the performance of single and multi-user detection systems. This performance was characterized by a probability measure of some error set in the observation space. However in the analysis of general stochastic systems, it is often the case that one is interested in determining the value of an arbitrary performance measure. Equivalently one may wish to evaluate the expected value of a functional of the output of a stochastic system. These problems arise in the areas of detection theory [56] and stochastic signal processing [6]. Expectations of functions of random vectors also arise in situations where one is interested in the moments of the joint density function.

The problems specifically addressed in the previous chapters dealt with the special case that the functional of the random vector was given by an indicator function. As was elaborated on in Chapters 2 and 3, the direct evaluation of the performance of these systems is typically an analytically intractable problem for systems with nonlinearities or non-Gaussian inputs. The nonlinearities give rise to error sets in the observation space which are often ill-defined and the non-Gaussian inputs make it very
difficult to determine the exact distribution of the output of the detection system. Before the advent of digital computing machines, the only recourse was to either use an appropriate bound on the set probability or to use a Gaussian approximation (Central Limit Theorem) to the output statistic. There are pitfalls in both of these approaches. For one, it is not always clear as to how tight the bounds may be for non-asymptotic cases and recently [32], the Central Limit Theorem approximation was shown to render a percent error which increases exponentially in the number of observations, assuming the coefficient of variation was fixed or decreasing.

As a result, it is often necessary to estimate these expectations via Monte Carlo simulations. These estimates will converge to the actual value if the number of simulation trials is sufficiently large. This can be a problem in situations which require a large number of sample runs. In estimating probabilities of rare events, the set probability can be quite small. Typically, one needs approximately 10 over the set probability trials to achieve an acceptable level of accuracy [49, 3]. Not only will this amount of computation take an inordinate length of time, it may require that the number of random variables approach or exceed the period of the random number generator.

As was demonstrated, these problems can be circumvented by use of a simulation technique known as Importance Sampling. This technique was first introduced by Kahn [33] and subsequently investigated for applications in estimating error rates in fiberguide repeaters [7], estimating false alarm rates [43], estimating the probability of
rare events in stochastic algorithms [12], probability of overflow in queuing networks [53], expectations of functionals of Markov chains [28], and in estimating error rates in communication systems [62, 29, 44, 49]. Improvements over standard Monte Carlo estimation have varied from modest to exponential.

Recall from the previous chapters that Importance Sampling is based on the notion of modifying the underlying distribution of the random vector in such a way as to generate more vectors from the "important" regions of the input space. The output of the simulation is then weighted such that the final estimate is unbiased. The hope is that if the modification is done in an intelligent fashion, then the number of simulation runs, for a given accuracy, can be dramatically reduced as compared to the number required with standard Monte Carlo simulations.

In all the above mentioned works and in the previous chapters, a parameterized Importance Sampling scheme was chosen and then optimized for the particular systems investigated. One of the advantages of this approach is that it reduced the optimization problem from minimizing over an arbitrary class of functionals to minimizing over a parameterized family of functionals. Employing any specific parameterized family of density functions renders an ad hoc simulation scheme. This restriction clearly results in suboptimal solutions, however it does make the problem more tractable.

In this chapter, a more systematic approach is taken. Since the optimal density function frequently requires the knowledge of the expectation of interest, a con-
strained solution is often necessary. By use of robust statistics, we determine a solution from an arbitrary constraint set which minimizes the variance of the Importance Sampling estimator. The derivation involves showing that minimizing the variance of the Importance Sampling estimator is equivalent to determining the joint density function from the constraint set which is closest to the unconstrained optimal density in the Ali-Silvey sense [5]. By extending the existing robust detection theory, we show that the least favorable density function from the restricted class of density functions achieves the minimum distance to the optimal density, thereby minimizing the Importance Sampling variance. Not all of these solutions can be implemented, but we will be able to show that some parameterized solutions can resemble the constrained optimal solutions investigated. In this way, we will have some theoretical justification for choosing a specific technique.

In the following section, we begin by formulating the problem and deriving the global optimal density function. We continue in Section 4.3 by determining the relationship between the estimator variance and the Ali-Silvey distance measures. Sections 4.4 and 4.5 relate robust solutions in hypotheses testing to the Importance Sampling minimization problem. Finally, we complete this chapter by investigating the ε-mixture constraint set and apply these results to the problem of evaluating set probabilities.
4.2 Importance Sampling for General Expectations

Often in statistical analysis, one is interested in determining the value $\psi = E_P[g(X)]$, where $X$ is a $N$-dimensional random vector with distribution function $P$ and $g(\cdot)$ is a Borel measurable function. In many practical cases, this expectation can not be evaluated directly. Either the underlying probability measure is unknown or the expectation may be too difficult to analytically compute. Consequently, we are left with attempting to bound this value from above and below or to estimate it from some set of sample data. In the event that $p(z)$ (the density function of $X$) is known, we can rely on Monte Carlo simulations to estimate $E_P[g(X)]$. The sample estimate of $\psi$ is constructed by simple averaging, i.e.,

$$\hat{\psi} = \frac{1}{M_{MC}} \sum_{i=1}^{M_{MC}} g(X_i), \text{ where } X \sim P$$  (4.1)

and where $M$ denotes the number of simulation trials. It is easily shown that $\hat{\psi}$ is unbiased and the variance of the estimator is

$$\text{var}_P[\hat{\psi}] = \frac{\text{var}_P[g(X)]}{M_{MC}} = \frac{\int_{\mathbb{R}^N} g^2(z)p(z)dz - \psi^2}{M_{MC}}$$  (4.2)

where $\mathbb{R}^N$ is the observation space. To apply Importance Sampling to this general problem, we first modify the simulation by rewriting $\psi$ as $E_{P_1}[W(X)g(X)]$ where $X \sim P_1, W = \frac{p(z)}{p_1(z)}$ and $P$ is absolutely continuous with respect to the new probability distribution $P_1$, (i.e., $P \ll P_1$). An alternate unbiased estimate of $\psi$ is then given by

$$\psi^* = \frac{1}{M_{IS}} \sum_{i=1}^{M_{IS}} W(X_i)g(X_i), \text{ where } X_i \sim P_1.$$  (4.3)
The variance of \( \psi^* \) can be shown to be

\[
\text{var}_{P_1} [\psi^*] = \frac{\text{var}_{P_1} [W(X)g(X)]}{M_{IS}} = \int_{\Omega} \frac{p^2(z)g^2(z)}{p_1(z)} \, d\mathcal{Z} - \psi^2.
\]

For Importance Sampling to be useful, we must have \( \text{var}_{P_1} [\psi^*] \leq \text{var}_P [\psi] \). To this end, we seek to minimize \( \text{var}_{P_1} [\psi^*] \) with respect to \( P_1 \).

**Theorem 4.1** The choice of \( p_1(z) = \frac{p(z)g(z)}{E_P[g(z)]} \) achieves the minimum variance for the Importance Sampling estimator, \( \psi^* \).

**Proof** The solution can be obtained by use of Jensen's Inequality [56]. The appropriate quantity to minimize is \( E_{P_1} [W^2(X)g^2(X)] \):

\[
E_{P_1} [W^2(X)g^2(X)] = E_{P_1} [W^2(X)|g(X)|^2] \geq E_{P_1} [W(X)|g(X)|] = E_P [g(X)].
\]

(4.5)

The lower bound can be achieved if and only if \( W(X)|g(X)| \) is almost surely a constant with respect to \( P_1 \). This implies that

\[
\frac{p(z)}{p_1(z)} |g(z)| = K \quad a.s.
\]

(4.6)

Solving for \( P_1 \), we have the result that

\[
p_{opt}(z) = \arg \min_{P_1 \in \mathcal{P}} \text{var}_{P_1} [\psi^*] = \frac{p(z)|g(z)|}{E_P[|g(z)|]} \quad a.s.,
\]

(4.7)

where \( \mathcal{P} \) is the set of all valid probability measures.

We note that

\[
K = E_P[|g(z)|] \geq E_P[g(z)] = \psi,
\]

(4.8)
with equality if and only if \( g(\cdot) \) is non-negative almost surely with respect to \( P \). This implies that the minimum variance achievable using Importance Sampling is

\[
\min \ var_{P_1}[\psi^*] = \frac{K^2 - \psi^2}{M_{IS}}.
\]  

(4.9)

From (4.8) we see that a zero variance, unbiased estimate of \( \psi \) exists only for functionals \( g(\cdot) \) which are non-negative with respect to \( P \). While this unconstrained solution requires the evaluation of \( K \), it does lend insight into the type of solution we should be approaching. Heuristically, we should derive solutions which are as “close” to the optimal solution as possible. Intuitively, this solution should then minimize the variance of the Importance Sampling estimator. In this paper, we determine \( P_1 \) from a constraint set \( p_1 \) which is “closest” to the \( P_{opt} \). We show that this choice achieves the minimum Importance Sampling variance among all possible distributions from \( p_1 \).

In previous work on Importance Sampling, there have been only three predominant constraint classes of biasing densities, \( p_1 \). These constraint classes arise from parametric families of density functions. These approaches have demonstrated varying degrees of efficacy. As discussed in the earlier chapters, applications have primarily been in determining error rates for communication systems. These classes have the following forms:

1. **Increased Variance** [62, 43]  
   \[ p_1(z) = \frac{1}{A} p(z/A), \ A > 0. \]

2. **Linear Shift** [44, 49]  
   \[ p_1(z) = p(z - \ell), \ \ell \in \mathcal{R}^N. \]

3. **Exponential Shift** [12, 53, 28, 61]  
   \[ p_1(z) = \frac{e^{\langle \ell, z \rangle} p(z)}{M(\ell)}, \ \ell \in \mathcal{R}^N. \]
These techniques only require the optimization with respect to a set of parameters, thus allowing for a tractable method of determining a specific biasing scheme tailored for the individual problem. In this chapter, we take a different approach by not restricting ourselves to parametric families of density functions. Our solution will be determined from any arbitrary closed set of density functions. Therefore, these previous solutions will be special cases of a more general theory based upon distance measures.

4.3 Distance Measures

The notion of distances between two probability measures has been widely used in statistical analysis [23, 34, 55]. A well known class of distances was independently derived by Ali and Silvey [5] and Csiszar [13]. These distances are often referred to as Ali-Silvey Distances or f-Divergences. Many well known measures of discrimination are elements of this class. Examples are Kullback-Leibler distance [40], Kolmogorov's measure of variational distance, J-divergence and the Battacharyya distance measure [34].

We show in this paper that minimizing the variance of the Importance Sampling estimator is equivalent to minimizing an Ali-Silvey distance between two probability measures. As stated in Theorem 4.1, the critical quantity to minimize in Importance Sampling is $E_{\pi_1} [W^2(\mathcal{X}) g^2(\mathcal{X})]$. This is equivalent to minimizing

$$\int_{\mathbb{R}^n} \left[ \frac{g(z)}{p_1(z)} \right]^2 p_1(z) dz = K^2 \int_{\mathbb{R}^n} \left[ \frac{|g(z)| p(\mathcal{X})}{K p_1(z)} \right]^2 p_1(z) dz$$

(4.10)
with respect to $P_1(x)$. Substituting in $p_{opt}$ from (4.7) we have

\[
\text{var}_{P_1} [L(X)g(X)] = K^2 \int_{\mathbb{R}^N} \left[ \frac{p_{opt}(x)}{p_1(x)} \right]^2 p_1(x) dx - \psi^2. \tag{4.11}
\]

Since $K$ and $\psi$ are fixed constants, we can reduce the problem to minimizing the integral in (4.11). We relate this integral to a distance between $P_1$ and $P_{opt}$.

The Ali-Silvey distance between two probability measures is defined as

\[
d(P_1, P_2) = h \left( \int_{\mathbb{R}^N} C \left[ \frac{p_2(x)}{p_1(x)} \right] p_1(x) dx \right), \tag{4.12}
\]

where $C(\cdot)$ is any continuous convex function and $h(\cdot)$ is an increasing real function of a real variable. Using this definition, we have that

\[
\text{var}_{P_1} [\psi^*] = \frac{K^2 d_{IS}(P_1, P_{opt}) - \psi^2}{M}, \tag{4.13}
\]

where $d_{IS}(\cdot, \cdot)$ denotes the Ali-Silvey distance with $h(x) = x$ and $C(x) = x^2$. This result yields the mathematical framework for determining a constrained optimal solution for estimating $\psi$ via Importance Sampling. We look now for $P_1 \in p_1$ which is closest to $P_{opt}$ in the Ali-Silvey sense. Formally, we search for a solution to the problem

\[
P_1^* = \arg \min_{P_1 \in p_1} d_{IS}(P_1, P_{opt}), \tag{4.14}
\]

where $p_1$ is a constraint set of valid probability measures. It is well known [5] that if $P_{opt} \in p_1$, then $P_1^* = P_{opt}$ with $\min_{P_1 \in p_1} d_{IS}(P_1, P_{opt}) = 1$. In order to avoid this normally degenerate solution, throughout the duration of the paper we impose the restriction that $p_1$ not contain $P_{opt}$. Therefore we are limited to deriving biasing
densities which are at best "near-optimum". We formalize the notion of a near-optimum biasing strategy through the following definition.

**Definition 4.1** A biasing strategy is said to be near-optimum if

\[ 1 < d_{IS}(P^*_1, P_{opt}) \leq \Delta < \infty, \]  

(4.15)

for every \( P_{opt} \), where \( P^*_1 \) is the biasing distribution derived from this strategy.

Utilizing this definition, we have that for every functional \( g(\cdot) \) and distribution function \( P \), if \( P^*_1 \) is derived from a near-optimum biasing strategy then we may upper bound the Importance Sampling variance by

\[ \text{var}_{P^*_1}[\psi^*] \leq \frac{\Delta K^2 - \psi^2}{M_{IS}} \]  

(4.16)

where \( \Delta \) is finite.

One might well ask if traditional Monte Carlo simulations are near-optimum. As such, we address this question in the following proposition where we demonstrate via a counter example that standard Monte Carlo simulations are not a near-optimum biasing strategy for every functional \( g(\cdot) \).

**Proposition 4** The strategy of choosing the original Monte Carlo distribution \( P \) as our biasing distribution \( P^*_1 \) is not near-optimum.
Proof Let $g$ be an indicator function (implying that $K = \psi$). We therefore have that

$$\text{var}_{P_{rpe}}[\psi^*] = \frac{\left[\frac{1}{\psi} - 1\right] \psi^2}{M_{MC}}. \quad (4.17)$$

Thus we can exhibit a sequence of $P_{opt}$ such that no finite $\Delta$ bounds $\frac{1}{\psi}$. □

However, it can be shown that employing the linearly shifted version of the original density as our biasing strategy for the special case of estimating minimum error rate performance is approximately near-optimum. This result is derived from the fact that empirically, the gain $\Gamma$ is approximately proportional to $1/\psi$ implying that the Importance Sampling variance for these systems is proportional to the square of the performance for all values of the error rate.

We proceed in the following subsection to determine a strategy for deriving the minimum distance biasing distribution by appealing to robust statistics.

4.4 Robust Statistics

It is well known that there exists a relationship between risk robustness and distance measures between two probability distributions [54, 35]. We can apply the theory of robust solutions in hypotheses testing problems to minimizing $d_{IS}(P_1, P_{opt})$ with respect to $P_1$. We begin with preliminaries from robust detection theory.
The classical binary hypotheses testing problem is to decide between two hypotheses based upon an observation vector $\mathbf{X}$:

$$H_1 : \mathbf{X} \sim P_1$$

$$H_2 : \mathbf{X} \sim P_2$$

(4.18)

where $P_i$ are arbitrary joint probability distributions with $P_2 \ll P_1$. The problem is to design a test $\phi(\mathbf{X})$ constrained by some measure of optimality (e.g., Neyman-Pearson, Bayes). Typically, optimality is based on minimizing a risk function subject to some constraint. Mathematically, we let $\phi(\mathbf{X})$ be any decision rule between $H_1$ and $H_2$ accepting $H_2$ with conditional probability $\phi(\mathbf{X})$ when vector $\mathbf{X}$ is observed.

Assume that a loss $\ell_i > 0$ is incurred if $H_i$ is falsely rejected, then the expected risks are given by

$$R(P_i, \phi) = \ell_i E_{P_i} [\phi(\mathbf{X})] = \ell_i Pr [\phi \text{ accepts } H_2 | H_1]$$

(4.19)

and

$$R(P_2, \phi) = \ell_2 [1 - E_{P_2} [\phi(\mathbf{X})]] = \ell_2 Pr [\phi \text{ accepts } H_1 | H_2].$$

(4.20)

Optimality is frequently based upon one of the following minimization problems:

- minimize $\max_{i=1,2} R(P_i, \phi)$

- minimize $R(P_2, \phi)$ such that $R(P_1, \phi) \leq \alpha$

- minimize $[\pi_1 R(P_1, \phi) + \pi_2 R(P_2, \phi)]$
where $\alpha$ is the false alarm rate and $\pi_i$ are the a priori probabilities of $H_i$ occurring.

It is well known that the solution to these problems is the Probability Ratio Test between $P_1$ and $P_2$ [54],

$$\phi_{PRT}(X) = \begin{cases} 
1, & \frac{P_2}{P_1}(X) > \gamma \\
\xi, & \frac{P_2}{P_1}(X) = \gamma \\
0, & \frac{P_2}{P_1}(X) < \gamma
\end{cases} \quad (4.21)$$

where the threshold $\gamma$ and the randomization parameter $\xi$ are chosen to meet the appropriate constraint.

If, however, there is some uncertainty as to the underlying distributions $P_1$ and $P_2$, the performance of $\phi_{PRT}$ can be severely degraded [54]. We then would like to formulate the robust version of the previous testing problems so that $P_i$ are elements of some set of distribution functions, $\mathcal{P}_i$, rather than known exactly. The robust analog to the minimization problems stated above are as follows:

- minimize $\max_{i=1,2} \sup_{P_i \in \mathcal{P}_i} R(P_i, \phi)$

- minimize $\sup_{P_2 \in \mathcal{P}_2} R(P_2, \phi)$ such that $\sup_{P_1 \in \mathcal{P}_1} R(P_1, \phi) \leq \alpha$

- minimize $\sup_{P_i \in \mathcal{P}_i} [\pi_1 R(P_1, \phi) + \pi_2 R(P_2, \phi)]$

The solution to these problems depends on the notion of a least favorable pair of distributions from two sets of distribution functions [26].
Definition 4.2 A pair of distribution functions \((P'_1, P'_2) \in \mathcal{P}_1 \times \mathcal{P}_2\) is said to be least favorable in terms of risk if, for every probability ratio test \(\phi'\) between \(P'_1\) and \(P'_2\) as defined in (4.21), we have

\[
R(P'_i, \phi') \geq R(P_i, \phi') \quad \forall \ P_i \in \mathcal{P}_i, \ i = 1, 2 \quad (4.22)
\]

where \(R(\cdot, \cdot)\) is as defined in (4.19) and (4.20).

Using this definition of least favorable pairs, the solution to the robust hypothesis testing problem stated above is the probability ratio test between \(P'_1\) and \(P'_2\). This choice both minimizes the maximum risk and maximizes the minimum risk, thus rendering a saddle point solution.

Intuitively, the two least favorable distributions should be the ones from \(\mathcal{P}_1\) and \(\mathcal{P}_2\) which are as close as possible to one another in terms of some distance function. In this regard, we supply the following theorem.

Theorem 4.2 If \((P'_1, P'_2) \in \mathcal{P}_1 \times \mathcal{P}_2\) are least favorable in terms of risk, then these distributions minimize a particular Ali-Silvey distance measure among all \((P_1, P_2) \in \mathcal{P}_1 \times \mathcal{P}_2\).

Proof We start with the direct consequence of \(P'_i\) being least favorable in terms of risk:

\[
\pi_1 R(P'_1, \phi'_B) + \pi_2 R(P'_2, \phi'_B) \geq \pi_1 R(P_1, \phi'_B) + \pi_2 R(P_2, \phi'_B) \quad \forall \ P_i \in \mathcal{P}_i \text{ and } \forall (\pi_1, \pi_2),
\]

(4.23)
where \( \phi_B \) is the Bayes decision rule between \( P_1' \) and \( P_2' \), i.e., \( \gamma = \frac{\pi_1}{\pi_2} \) and \( \xi = 0 \) or 1 in (4.21). It is well known that Bayes risk is minimized by using the appropriate probability ratio test derived from the true underlying distributions [56], therefore we have that

\[
\pi_1 R(P_1', \phi_B) + \pi_2 R(P_2', \phi_B) \geq \pi_1 R(P_1, \phi_B) + \pi_2 R(P_2, \phi_B) \quad \forall \ P_i \in \mathfrak{p}, \text{ and } \forall (\pi_1, \pi_2),
\]

(4.24)

where \( \phi_B \) is the Bayes rule for discriminating between \( P_1 \) and \( P_2 \). It can be shown that

\[
\pi_1 R(P_1, \phi_B) + \pi_2 R(P_2, \phi_B) = \int_{\mathcal{R}^N} \min [\pi_1 \ell_1 p_1(\mathfrak{z}), \pi_2 \ell_2 p_2(\mathfrak{z})] d\mathfrak{z}.
\]

(4.25)

Using the relationship that \( \min(t, s) = \frac{1}{2}(t + s - |t - s|) \), we have

\[
\pi_1 R(P_1, \phi_B) + \pi_2 R(P_2, \phi_B) = \frac{1}{2}(\pi_1 \ell_1 + \pi_2 \ell_2) - \frac{1}{2} \int_{\mathcal{R}^N} |\pi_1 \ell_1 p_1(\mathfrak{z}) - \pi_2 \ell_2 p_2(\mathfrak{z})| d\mathfrak{z}.
\]

(4.26)

The integral in (4.26) can be written as \( E_{\mathfrak{P}_1} \left[ |\pi_1 \ell_1 - \pi_2 \ell_2 \frac{p_2(x)}{p_1(x)}| \right] \). This yields the equality

\[
\pi_1 R(P_1, \phi_B) + \pi_2 R(P_2, \phi_B) = \frac{1}{2}(\pi_1 \ell_1 + \pi_2 \ell_2) - \frac{1}{2} d_R(P_1, P_2),
\]

(4.27)

where \( d_R \) is an Ali-Silvey distance with \( C(x) = |\pi_1 \ell_1 - \pi_2 \ell_2 x| \) and \( h(x) = x \). Applying (4.24), we have the result

\[
d_R(P_1', P_2') \leq d_R(P_1, P_2) \quad \forall \ P_i \in \mathfrak{p};
\]

(4.28)
Therefore the two least favorable distributions are the ones which are nearest one another in the $d_R$ sense. It is interesting to note that if $\pi_1 \ell_1 = \pi_2 \ell_2$, then

$$d_R(P_1, P_2) = 2\pi_1 \ell_1 \ d_K(P_1, P_2),$$

(4.29)

where $d_K(P_1, P_2)$ is Kolmogorov's measure of "variational distance" $^1$.

### 4.5 Constrained Solutions

Recall that to solve the general constrained Importance Sampling problem, we need to find distributions from two sets which are nearest each other in the Ali-Silvey sense defined in (4.13). We have from Theorem 4.2 that the least favorable pair is nearest one another in the Ali-Silvey sense defined in (4.27). It is well known that all Ali-Silvey distance measures share some common properties. By extending a theorem due to Kobayashi [39] we will be able to show that the least favorable pair is the minimum distance pair for every Ali-Silvey distance.

**Theorem 4.3** With $R(\cdot, \cdot)$ as defined in (4.19) and (4.20), then

$$[\pi_1 R(P_1, \phi_B) + \pi_2 R(P_2, \phi_B)] \leq [\pi_1 R(P_1', \phi_B') + \pi_2 R(P_2', \phi_B')]$$

(4.30)

for all sets of a priori probabilities $\pi = (\pi_1, \pi_2)$ if and only if we have

$$d(P_1, P_2) \geq d(P_1', P_2')$$

(4.31)

for all Ali-Silvey distance measures.

$^1_{d_K(P_1, P_2) = \int \int p_1(x) - p_2(x) |d_x}$
Proof  $\Rightarrow$ From (4.27), it is easily verified that the condition in (4.30) is equivalent to

$$E_{P_1} \left[ \pi_1 \ell_1 - \pi_2 \ell_2 \frac{P_2(X)}{P_1(X)} \right] \geq E_{P_1} \left[ \pi_1 \ell_1 - \pi_2 \ell_2 \frac{P_2'(X)}{P_1'(X)} \right]$$  (4.32)

for any pair of losses $\ell_i$ and every pair of priors $\pi_i$, where $E_{P_1}$ is the expectation with respect to the probability distribution $P_1$. We also have that

$$E_{P_1} \left[ \varphi \left( \frac{P_2(X)}{P_1(X)} \right) \right] = E_{P_1} \left[ \varphi \left( \frac{P_2'(X)}{P_1'(X)} \right) \right]$$  (4.33)

for every linear function $\varphi(\cdot)$. These two equations imply that

$$E_{P_1} \left[ C \left( \frac{P_2(X)}{P_1(X)} \right) \right] \geq E_{P_1} \left[ C \left( \frac{P_2'(X)}{P_1'(X)} \right) \right]$$  (4.34)

for every continuous convex function $C(\cdot)$. We complete the first part of the proof by applying $h(\cdot)$ to both sides of the inequality to obtain

$$d(P_1, P_2) \geq d(P_1', P_2')$$  (4.35)

for every Ali-Silvey distance measure. The proof of the converse statement is direct from (4.27).

We need now only find the pair $(P_1', P_2')$ such that they have greater total Bayes risk for every pair of priors $\pi_i$. In the following corollary, it is shown that the least favorable pair does satisfy the condition in (4.30) and is thus the minimum Ali-Silvey distance pair.
Corollary 4.1 If \((P_1', P_2')\) is least favorable in terms of Bayes risk for classes \(p_i\), then

\[
d(P_1, P_2) \geq d(P_1', P_2') \quad \forall P_i \in p_i
\]  

(4.36)

for all Ali-Silvey distance measures.

Proof The proof follows immediately from applying (4.24) to (4.30) in Theorem 4.3. \(\square\)

We can appeal to this corollary to solve the problem of minimizing the variance of the Importance Sampling estimator of \(\psi\). The result is summarized in the following corollary.

Corollary 4.2 If \((P_1', P_{opt})\) from \(p_1 \times p_2\) is the least favorable pair in terms of Bayes risk, then \(\text{var}_{P_1} [\psi^*] \leq \text{var}_{p_1} [\psi^*]\) for every \(P_1 \in p_1\).

Proof The proof follows from the application of Corollary 3.1 to equation (4.13). \(\square\)

If there exists a least favorable pair in terms of Bayes risk such that \(P_2' = P_{opt}\), then the optimal biasing distribution from \(p_1\) will be

\[
P_1^* = \arg \min_{A_1 \in p_1} d_{IS}(P_1, P_{opt}) = P_1'.
\]  

(4.37)

Thus, to obtain the probability distribution which minimizes the variance of \(\psi^*\), we search for \(P_1' \in p_1\) which is least favorable in terms of Bayes risk against \(P_{opt}\).
4.6 Examples of Distribution Classes

Finding least favorable pairs has been intensely investigated [26, 45, 35]. In order to find the least favorable pair, we must first specify the sets of probability measures. Huber [27] introduced the $\varepsilon$-mixture constraint class as a general convex class of probability measures

$$\mathcal{P}_1 = \{ P_1 | P_1 = (1 - \epsilon_1)P_{N_1} + \epsilon_1H_1, H_1 \in \mathcal{P} \}$$  \hspace{1cm} (4.38)

$$\mathcal{P}_2 = \{ P_2 | P_2 = (1 - \epsilon_2)P_{N_2} + \epsilon_2H_2, H_2 \in \mathcal{P} \}$$  \hspace{1cm} (4.39)

where $P_{N_1}$ and $P_{N_2}$ are the nominal distributions from classes $\mathcal{P}_1$ and $\mathcal{P}_2$ and $0 \leq \epsilon_i < 1$ are fixed constants.

Huber showed that the least favorable pair of density functions from these classes when the vector $X$ has independent and identically distributed components will be of the form

$$p_1'(z) = \begin{cases} 
(1 - \epsilon_1)p_{N_1}(z) & p_{N_2}(z)/p_{N_1}(z) < c_1 \\
\frac{1}{c_1}(1 - \epsilon_1)p_{N_2}(z) & p_{N_2}(z)/p_{N_1}(z) \geq c_1 
\end{cases}$$  \hspace{1cm} (4.40)

$$p_2'(z) = \begin{cases} 
(1 - \epsilon_2)p_{N_2}(z) & p_{N_2}(z)/p_{N_1}(z) > c_2 \\
c_2(1 - \epsilon_2)p_{N_1}(z) & p_{N_2}(z)/p_{N_1}(z) \leq c_2 
\end{cases}$$  \hspace{1cm} (4.41)

provided that $0 \leq c_2 < c_1 \leq \infty$. The constants $c_2$ and $c_1$ must be determined such that $p'_i(z)$ are valid density functions, i.e.,

$$\int_{\mathbb{R}^n} p_1'(z)dz = (1 - \epsilon_1)\int_{p_{N_2}/p_{N_1} < c_1} p_{N_1}(z)dz + \frac{1}{c_1}(1 - \epsilon_1)\int_{p_{N_2}/p_{N_1} \geq c_1} p_{N_2}(z)dz = 1$$  \hspace{1cm} (4.42)
and

\[ \int_{\mathbb{R}^N} p_1'(z) dz = (1 - \epsilon_2) \int_{\mathbb{P}_N/\mathbb{P}_N > c_2} p_N(z) dz + c_2(1 - \epsilon_2) \int_{\mathbb{P}_N/\mathbb{P}_N < c_2} p_N(z) dz = 1. \]

(4.43)

It was shown that these equations have solutions for all \( \epsilon_i \in [0, 1] \) and that \( c_2 \) and \( c_1 \) where unique for \( 0 < \epsilon_i < 1 \). Moreover, if the \( \epsilon_i \) are sufficiently small, then we also have that \( c_2 < c_1 \), provided that \( P_{N_1} \neq P_{N_2} \).

Since we desire that the least favorable pair be \( (P_1', P_{opt}) \) where \( P_1' \in \mathcal{P}_1 \), we define \( \varphi_2 \) as in (4.39) with \( \epsilon_2 = 0 \) and \( P_{N_2} = P_{opt} \). In this way, \( \varphi_2 = \{P_{opt}\} \) with \( c_2 = 0 \) as a solution to (4.43). It can then be verified from (4.41) that this definition of \( \varphi_2 \) results in \( p_2'(z) = p_{opt}(z) \). It is also easily verified that \( P_1' \neq P_2' \) for every \( \epsilon_1 < 1 \), thereby alleviating the normal requirement that \( \epsilon_1 \) be small.

Moreover, it is desired that this theory encompass the case where the elements of \( X \) are not necessarily independent and identically distributed. We can obtain the more general result by recalling that both Corollary 4.1 and Corollary 4.2 require only that \( P_1' \) and \( P_2' \) be least favorable in terms of Bayes risk. We do not have to impose the restriction that \( P_1' \) and \( P_2' \) be least favorable in terms of every probability ratio test, we only require that they satisfy the weaker condition that \( (P_1', P_2') \) be least favorable for every Bayes test. Therefore, by assigning \( \xi = 1 \) and \( \gamma = \frac{\pi_1 \eta_1}{\pi_2 \xi_2} \), we then have the following modification of Huber's main result [26]:

**Theorem 4.4** \( P_1' \) and \( P_2' \) as defined in (4.40) and (4.41) are the least favorable pair in terms of Bayes risk for \( \epsilon \)-mixture classes \( \mathcal{P}_1 \) and \( \mathcal{P}_2 \) where
the elements $P_1 \in \varrho_i$ are not necessarily independent and identically distributed.

Proof The proof follows the one given in [26]. We begin with the following lemma which holds for joint distribution functions and is proved in Huber [26].

Lemma 4.1 Let $\varrho_1$ and $\varrho_2$ be $\epsilon$-mixture classes, then for any $P_i \in \varrho_i$ and any real number $t$, we have

\[
P_1 \left[ \frac{P_2'}{P_1'}(X) < t \right] \geq P_1' \left[ \frac{P_2'}{P_1'}(X) < t \right] \geq P_2' \left[ \frac{P_2'}{P_1'}(X) < t \right] \geq P_2 \left[ \frac{P_2}{P_1}(X) < t \right]
\]

(4.44)

provided $c_2 < c_1$.

To show that $P_1'$ and $P_2'$ are least favorable in terms of Bayes risk, we must show that

\[
R(P_1', \phi_B') \geq R(P_i, \phi_B^i) \forall P_i \in \varrho_i \ i = 1, 2
\]

(4.45)

where $\phi_B'$ is the probability ratio test between $P_1'$ and $P_2'$ with $\xi = 1$, i.e.,

\[
\phi_B'(X) = \begin{cases} 
1 & \frac{P_2}{P_1} \geq \gamma \\
0 & \frac{P_2}{P_1} < \gamma 
\end{cases}
\]

(4.46)

for any $\gamma$.

We have the desired result via the following steps:

\[
R(P_1', \phi_B') = \ell_1 P_1' \left[ \frac{P_2'}{P_1'} \geq \gamma \right]
\]
Similarly, one can show that \( R(P_2', \phi_B') \geq R(P_2, \phi_B') \) \( \forall P_2 \in \varphi_2 \). □

Therefore, the optimal biasing density \( p_1' \) from the \( \epsilon \)-contaminated class is \( p_1' \) as defined in (4.40). No other density function from this class will be closer in the sense of the Ali-Silvey distance \( d_{IS} \) to the unconstrained optimal density function derived in (4.7). To determine how well this biasing density works we compute \( d_{IS}(P_1', P_{opt}) \) to determine the \( \text{var}_{P_1} [\psi^*] \). Comparing this result to \( d_{IS}(P, P_{opt}) \), we can derive the improvement over standard Monte Carlo simulations.

Using \( \varphi_1 \) and \( \varphi_2 \) as defined above (i.e., \( P_{N_2} = P_{opt} \) and \( \epsilon_2 = 0 \) in (4.39)), the least favorable density from \( \varphi_1 \) is written as

\[
p_1'(z) = \begin{cases} 
(1 - \epsilon_1)p_{N_1}(z) & \frac{P_{opt}}{p_{N_1}} < c_1 \\
\frac{1}{c_1}(1 - \epsilon_1)p_{opt}(z) & \frac{P_{opt}}{p_{N_1}} \geq c_1 
\end{cases}
\]  

(4.47)

where \( P_{N_1} \) is any nominal distribution function such \( P_{opt} \ll P_{N_1} \). To determine \( d_{IS}(P_1', P_{opt}) \), recall the definition

\[
d_{IS}(P_1', P_{opt}) = \int_{\mathbb{R}^N} \left[ \frac{p_{opt}(z)}{p_1'(z)} \right]^2 p_1'(z) d\bar{z} = \int_{\mathbb{R}^N} \frac{p_{opt}(z)}{p_1'(z)} p_{opt}(z) d\bar{z}.
\]  

(4.48)
Substituting \( p'_1(\bar{x}) \) from above into (4.48), we can arrive at the following upper bound on the distance:

\[
\int_{R^N} \frac{p_{opt}(\bar{x})}{p'_1(\bar{x})} p_{opt}(\bar{x}) d\bar{x} = \frac{1}{(1 - \epsilon_1)} \left[ \int_{p_{opt}/p_{N_1} < c_1} \frac{p_{opt}(\bar{x})}{p_{N_1}(\bar{x})} p_{opt}(\bar{x}) d\bar{x} + \int_{p_{opt}/p_{N_1} \geq c_1} c_1 p_{opt}(\bar{x}) d\bar{x} \right] \\
\leq \frac{1}{(1 - \epsilon_1)} \left[ \int p_{opt} \leq c_1 c_1 p_{opt}(\bar{x}) d\bar{x} + \int p_{opt} \geq c_1 c_1 p_{opt}(\bar{x}) d\bar{x} \right] = \frac{c_1}{(1 - \epsilon_1)} \tag{4.49}
\]

This inequality renders the bound

\[
1 \leq d_IS(P'_1, P_{opt}) \leq \frac{c_1}{(1 - \epsilon_1)}. \tag{4.51}
\]

Since \( c_1 \) depends on the nominal distribution \( P_{N_1} \), it would be helpful to determine an upper bound independent of the choice of the nominal distribution. To this end, we define the function

\[
f(c_1) = (1 - \epsilon_1) \int_{p_{N_1}/p_{N_2} < c_1} p_{N_1}(\bar{x}) d\bar{x} + \frac{1}{c_1}(1 - \epsilon_1) \int_{p_{N_1}/p_{N_2} \geq c_1} p_{N_2}(\bar{x}) d\bar{x}, \tag{4.52}
\]

which is simply a restatement of the integral constraint on \( p'_1 \) in (4.42). As stated before, there exists constant \( c_1 > 0 \) such that \( f(c_1) = 1. \) Using a straightforward upper bound for \( f(c_1) \), we have that

\[
1 \leq (1 - \epsilon_1) \left[ 1 + \frac{1}{c_1} \right]. \tag{4.53}
\]

Solving for \( c_1 \) gives the upper bound \( c_1 \leq \frac{1 - \epsilon_1}{\epsilon_1} \) which is dependent only upon \( \epsilon_1 \). Inserting into (4.51), renders the most general upper bound for \( 0 \leq \epsilon_1 < 1 \)

\[
1 \leq d_IS(P'_1, P_{opt}) \leq \frac{1}{\epsilon_1}. \tag{4.54}
\]
Therefore, we have as an upper bound on the variance of the Importance Sampling estimator with the \( \epsilon \)-mixture constraint class

\[
\text{var}_{P_1} [\psi^*] \leq \frac{K^2 / \epsilon_1 - \psi^2}{M_{IS}},
\]  

which holds for any functional \( g(\cdot) \) and any nominal distribution. As \( \epsilon_1 \to 1 \), the variance of \( \psi^* \) approaches the minimum achievable variance using Importance Sampling. Intuitively, this must be so, since \( \epsilon_1 \to 1 \) is equivalent to \( \rho_1 \to \rho \), the set of all probability distributions.

Thus the strategy of choosing the least favorable density function from the \( \epsilon \)-mixture constraint class is near-optimum for all values of \( \epsilon \in (0,1) \) and for every choice of functional \( g(\cdot) \).

For comparison with the standard Monte Carlo method, consider the example of \( g(\mathbf{x}) = I_A(\mathbf{x}) \), where \( I_A(\cdot) \) is an indicator function over the set \( A \). The variance of the Monte Carlo estimate can be shown to be

\[
\text{var}_P [\psi] = \frac{\psi - \psi^2}{M_{MC}}.
\]  

Since the functional \( g(\cdot) \) is non-negative with respect to \( P \), the upper bound on the variance of \( \psi^* \) can be simplified to be

\[
\text{var}_{P_1} [\psi^*] \leq \frac{\psi^2 \left( \frac{1-\epsilon_1}{\epsilon_1} \right)}{M_{IS}}.
\]  

It is important to have some measure of comparison between standard Monte Carlo simulations and Importance Sampling simulations. As demonstrated in the
previous chapters [49], the ratio of the variances of the two simulation techniques captures the essence in using Importance Sampling. The so called "gain" $\Gamma$, gives a direct relationship between the number of simulation samples required for a fixed estimator variance in using the two methods. The mathematical relationship is given by

$$M_{IS} = \frac{M_{MC}}{\Gamma} \quad (4.58)$$

where $M_{IS}$ and $M_{MC}$ are the number of trials for Importance Sampling and Monte Carlo simulations, respectively, and where $\Gamma$ is given by

$$\Gamma = \frac{\text{var}_P[\hat{\psi}]}{\text{var}_{P_1}[\psi^*]} \quad (4.59)$$

It is evident that any effective Importance Sampling scheme should render $\Gamma \gg 1$.

The exact simulation gain is dependent upon the functional $g(\cdot)$ and the choice of biasing density. However, we can determine a lower bound on $\Gamma$ for the $\epsilon$-contaminate constraint class by use of the upper bound on the variance of $\psi^*$ in (4.57). In keeping with our example that $g(\cdot)$ be an indicator function, we obtain

$$\Gamma \geq \frac{\epsilon_1(1 - \psi)}{(1 - \epsilon_1)\psi} \quad (4.60)$$

For small $\psi$, the gain is approximately inversely proportional to the parameter of interest. Using the relationship in (4.58), it can be shown that the number of trials required for $Pr(\psi^* \in \left[\frac{\xi}{2}\psi, \frac{\xi}{3}\psi\right]) = .95$ while using $P_1'$ as a biasing density is upper bounded by

$$M_{IS} \leq \frac{10(1 - \epsilon_1)}{\epsilon_1(1 - \psi)} \approx \frac{10(1 - \epsilon_1)}{\epsilon_1} \quad (4.61)$$
Note that this implies that the number of simulation trials required for a fixed variance of \( \psi^* \) is upper bounded by a constant which is approximately independent of the value of \( \psi \) regardless of the choice of nominal density.

To insure that we have a reduced variance estimate for every \( \epsilon_1 \), or equivalently that

\[
\text{d}_{I_S}(P_1', P_{\text{opt}}) \leq \text{d}_{I_S}(P, P_{\text{opt}}),
\]

we may define \( \varphi_1 \) as

\[
\varphi_1 = \{ P_1 | P_1 = (1 - \epsilon_1)P + \epsilon_1 H_1, H_1 \in \varphi \}.
\]

In this way, we have defined the nominal distribution to be the original distribution under which the Monte Carlo simulation is implemented. Therefore, we then choose the closest distribution to the \( P_{\text{opt}} \) from a \( \epsilon \)-ball centered about \( P \). However, if we already have an alternate biasing scheme which renders a reduced variance estimate of \( \psi^* \), it would be prudent to use that density function as the nominal density. We can then use any previously optimized density from one of the three classes discussed in Section 4.2 as nominal density for \( \varphi_1 \). This implies that the performance of all previous biasing schemes can be improved through the use of equation (4.47) since the least favorable density from a ball centered about the previous solution will be nearer the optimal biasing distribution.

To demonstrate the effect of different choices for the nominal density, we supply the following example of estimating the probability of the tail of a Laplacian random variable.
The density function of a zero mean Laplacian random variable with \( \sigma^2 = 2 \) is given by
\[
p(x) = \frac{1}{2} e^{-|x|},
\] (4.64)
The probability of a tail is denoted by \( \psi = \Pr[X \geq T] \) where \( X \sim \text{Laplacian} \). This probability can be evaluated directly without appealing to simulations, but for our purposes, it does give insight into the form of these constrained solutions.

We begin by recalling that the optimal biasing density from (4.7) is given by
\[
p_{\text{opt}}(x) = I_{[T, \infty)}(x) e^{T-x},
\] (4.65)
where \( I_{[T, \infty)}(\cdot) \) is the indicator function over \([T, \infty)\). We must first determine \( c_1 \) from the integral constraint in (4.42) for each choice of \( P_{N_1} \). Selecting the original density function as our choice for the nominal density function, (i.e., \( p_{N_1}(x) = \frac{1}{2} e^{-|x|} \)), it can be shown that
\[
c_1 = \frac{1 - \epsilon_1}{\epsilon_1 + (1 - \epsilon_1) \psi}.
\] (4.66)
It is easily verified that \( c_1 \leq \frac{(1-\epsilon_1)}{\epsilon_1} \) and that for this example, the upper bound for \( c_1 \) derived from (4.53) is asymptotically tight as \( \psi \to 0 \). From this, it can be shown that the constrained optimal biasing density is
\[
p_1'(x) = \begin{cases} 
\frac{(1-\epsilon_1)}{2} e^{-|x|} & x < T \\
[(1 - \epsilon_1) \psi + \epsilon_1] e^{T-x} & x \geq T
\end{cases}
\] (4.67)

We now investigate an alternate choice for the nominal density. The choice of
\[
p_{N_1}(x) = p(x - T),
\] (4.68)
from the class of linearly shifted densities was shown in Chapter 2 to be asymptotically optimum in estimating the probability of the tail of a random variable [49]. We conjecture that the best choice from an $\epsilon$-ball surrounding a previously optimized solution should not only improve the performance over the previously optimized solution but should also improve the performance over the choice of any biasing density from an $\epsilon$-ball surrounding the original density function. This is intuitive since the shifted version of $p(x)$ is closer than the original density to $p_{opt}(x)$.

In determining $c_1$ for this choice of nominal density, it can be shown that

$$c_1 = \frac{2(1 - \epsilon_1)}{(1 + \epsilon_1)}. \quad (4.69)$$

From this it can be shown that the constrained optimal density function is

$$p'_1(x) = \begin{cases} 
\frac{(1-\epsilon_1)}{2} e^{-|x-T|} & x < T \\
\frac{(1+\epsilon_1)}{2} e^{-|x-T|} & x \geq T 
\end{cases} \quad (4.70)$$

Important to note is that the choice of the shifted density as the nominal density renders a solution which is independent of $\psi$. This biasing density is a simple transformation of the shifted density function. It is constructed by scaling $p(x-T)$ by $(1-\epsilon_1)$ for all $x < T$ and correspondingly weighting $p(x-T)$ by $(1+\epsilon_1)$ for all $x \geq T$.

For a graphical comparison of the two biasing density functions given in (4.67) and (4.70) see Figure 4.1.
To determine the relative merits of these two nominal densities (i.e., $P$ and shifted $P$), we must evaluate $d_{IS}(P^*, P_{opt})$. For the case that $P_{N_1} = P$,

$$d_{IS}(P^*, P_{opt}) = \frac{1}{\epsilon_1 + (1 - \epsilon_1)\psi} \leq \frac{1}{\epsilon_1}. \quad (4.71)$$

As $\psi \to 0$, the distance between the biasing density and the optimal density tends to $\frac{1}{\epsilon_1}$, the upper bound given in (4.54).

In choosing the shifted version of the original biasing density as the nominal, the distance $d_{IS}$ is given as

$$d_{IS}(P^*_1, P_{opt}) = \frac{2}{1 + \epsilon_1} \leq \frac{1}{\epsilon_1}. \quad (4.72)$$

It is easily shown via the upper bound $\psi \leq \frac{1}{2}$ that

$$\frac{1}{\epsilon_1 + (1 - \epsilon_1)\psi} \geq \frac{2}{1 + \epsilon_1} \quad (4.73)$$

for all values of $T \geq 0$ and for every value of $\epsilon_1$, confirming our conjecture in this case that previously optimize biasing schemes, when available, should be used as the nominal density function.

Evaluating $\Gamma$ for these two nominal densities, we have that for $p_{N_1}(x) = p(x)$

$$\Gamma = \frac{\epsilon_1 + (1 - \epsilon_1)\psi}{\psi(1 - \epsilon_1)} \quad (4.74)$$

which yields the required number of simulation trails

$$M_{IS} = \frac{10(1 - \epsilon_1)}{\epsilon_1 + (1 - \epsilon_1)\psi} \approx \frac{10(1 - \epsilon_1)}{\epsilon_1}. \quad (4.75)$$

In choosing $p_{N_1}(x) = p(x - T)$, the gain is given by

$$\Gamma = \frac{(1 - \psi)(1 + \epsilon_1)}{\psi(1 - \epsilon_1)} \quad (4.76)$$
with
\[ M_{IS} = \frac{10(1 - \epsilon_1)}{(1 - \psi)(1 + \epsilon_1)} \approx \frac{10(1 - \epsilon_1)}{(1 + \epsilon_1)}. \] (4.77)

As is evident, the number of simulation trials is approximately independent of \( \psi \) for both choices of nominal distributions. However for every \( \epsilon_1 \), the number of simulation trials for the choice \( p_{N_1}(x) = p(x - T) \) is less than the number of trials necessary for the choice \( p_{N_1}(x) = p(x) \). As an example, for the case that \( \epsilon_1 = 0.1 \), \( M_{IS} = 90 \) when using the original density as the nominal, and \( M_{IS} = 8 \) when the shifted density is used.

Not every set of probability distributions has a least favorable pair. This, however, does not imply that there does not exist a distribution which minimizes the distance to \( P_{opt} \). The condition that \( P_1^* \) is least favorable in terms of Bayes risk against \( P_{opt} \) is a sufficient condition for \( P_1^* \) to minimize \( d_{IS}(P_1, P_{opt}) \), not a necessary one. Even though we can not be assured that every choice of \( P_1 \times P_2 \) has a least favorable pair of distributions, if it exists, we can say something about the least favorable pair.

To demonstrate this, let us again consider the linearly shifted class of density functions for \( P_1 \)
\[ \mathcal{P}_1 = \{ p_1(x) = p(x - T), T \in \mathcal{T} \} \] (4.78)

and the optimum distribution for \( P_2 \)
\[ \mathcal{P}_2 = \{ p_2(x) = p_{opt}(x) \} \] (4.79)
Again, using the familiar example of estimating the tail of a random variable, \( Pr[X \geq T] \), we have the following proposition.

**Proposition 5** Consider the problem of estimating \( \psi = Pr[X \geq T] \) for large \( T \) where \( p(x) \) has infinite support and is bounded above by \( M \), if \((P'_1, P'_2)\) from \( \rho_1 \times \rho_2 \) as defined in (4.78) and (4.79) with \( T = (-\infty, T] \) is the least favorable pair of distributions in terms of Bayes risk, then

\[
p'_1(x) = p(x - T) \quad \text{and} \quad p'_2(x) = p_{\text{opt}}(x). \tag{4.80}
\]

**Proof** If \((P'_1, P'_2)\) is the least favorable pair in terms of Bayes risk, then from (4.24) and (4.25), we have the following inequality

\[
\int_{\mathbb{R}_1} \min [\pi_1 \ell_1 p'_1(x), \pi_2 \ell_2 p'_2(x)] dx \geq \int_{\mathbb{R}_1} \min [\pi_1 \ell_1 p_1(x), \pi_2 \ell_2 p_2(x)] dx \tag{4.81}
\]

for every set of \((\pi_1, \pi_2)\) and every \((P_1, P_2) \in \rho_1 \times \rho_2\). From the definition of \( \rho_1 \times \rho_2 \), we have the equivalent relationship

\[
\int_{\mathbb{T}}^{\infty} \min \left[ \pi_1 \ell_1 p(x - T'), \frac{\pi_2 \ell_2}{\psi} p(x) \right] dx \geq \int_{\mathbb{T}}^{\infty} \min \left[ \pi_1 \ell_1 p(x - T), \frac{\pi_2 \ell_2}{\psi} p(x) \right] dx \tag{4.82}
\]

for every set of \((\pi_1, \pi_2)\) and for every \( T \in \mathbb{T} \), where \( T' \) is the shift corresponding to the least favorable density. We now show that there exists \((\pi_1, \pi_2)\) such that \( T' = T \).

Since \( p(x) \leq M \forall x \), there exists constants \( \pi \) such that \( \frac{\pi_2 \ell_2}{\psi} p(x) \geq \pi_1 \ell_1 p(x - T') \) for every \( x \in [T, 2T] \). With the condition that \( T \) be large, we can assume that \( p(x) \) is strictly convex for all \( x \geq T \). This implies that \( p(x - T') \) is strictly convex for
\[ x \geq 2T, \text{ therefore we conclude that there exists } (\pi_1, \pi_2) \text{ such that} \]
\[
\int_T^\infty \min \left[ \pi_1 \ell_1 p(x - T'), \frac{\pi_2 \ell_2}{\psi} p(x) \right] \, dx = \int_T^\infty \pi_1 \ell_1 p(x - T') \, dx. \tag{4.83}
\]

The choice \( T' = T \) clearly maximizes the integral for this pair of priors and set \( \Upsilon \).

\[ \square \]

It can be shown that if in addition \( p(x) \) is bounded and monotonically decreasing, (i.e., \( p(|x|) > p(|y|) \) when \( |y| > |x| \)), then a least favorable pair of distributions does not exist for the case that \( \Upsilon = (T, +\infty) \). The proof is constructed as follows:

Let \( T \) be any element of \( \Upsilon \), then there exists a set of prior probabilities such that
\[
\frac{\pi_2 \ell_2 p(x)}{\psi} \geq \pi_1 \ell_1 p(x - T) \forall x \geq T. \tag{4.84}
\]

Therefore,
\[
\int_T^\infty \min \left[ \pi_1 \ell_1 p(x - T'), \frac{\pi_2 \ell_2}{\psi} p(x) \right] \, dx = \int_T^\infty \pi_1 \ell_1 p(x - T) \, dx. \tag{4.85}
\]

This integral increases with an increase in \( T \) for this set of \((\pi_1, \pi_2)\).

There also exists the set \((\pi_1', \pi_2')\) and \( \tilde{T} < T \) such that
\[
\pi_1' \ell_1 p(x - T) > \frac{\pi_1' \ell_2 p(x)}{\psi} \forall x > \tilde{T} \tag{4.86}
\]

and
\[
\pi_1' \ell_1 p(x - T) < \frac{\pi_1' \ell_2 p(x)}{\psi} \forall x \in [T, \tilde{T}]. \tag{4.87}
\]

From this, it can be shown that the Bayes risk, i.e.,
\[
\int_T^\infty \min \left[ \pi_1' \ell_1 p(x - T), \frac{\pi_1' \ell_2}{\psi} p(x) \right] \, dx \tag{4.88}
\]
increases with a decrease in $T$. Thus, for any $T \in \mathcal{T} = (T, +\infty)$, the choice $T'$ which maximizes the Bayes risk is a function of the set of prior probabilities $(\pi_1, \pi_2)$. Therefore, no single shift $T'$ holds for every pair of priors, i.e., no least favorable pair of density functions exists for this set of $T$.

This, with monotonicity, implies that within the restrictions stated in the above proposition, if a least favorable pair of distributions in terms of Bayes risk exists for the linearly shifted class where $\mathcal{T} = (-\infty, +\infty)$, then the corresponding shift of $p_1'$ will be the threshold $\mathcal{T}$.

This result is consistent with an earlier result [49] that shifting the original density to the threshold is asymptotically optimum, i.e., $\mathcal{T} \to \infty$ implies $\Gamma \to \infty$, moreover, this shift generates values of the random variable $X$ under distribution $P_1'$ such that $Pr [X > \mathcal{T}] = \frac{1}{2}$.

4.7 Conclusions

In all results prior to those presented in this chapter, solutions in Importance Sampling arose from ad hoc approaches. This, however, does not imply that these solutions were ineffective. Rather, most of these solutions were shown to render significant improvement over classical Monte Carlo techniques. Additionally, we were able to empirically demonstrate in Chapter 2 that under certain conditions the linear shift biasing strategy was approximately near optimum in that the variance of the estimator was proportional to the square of the performance measure $P_e$. Nevertheless, there
was a lack of the development in the theory surrounding Importance Sampling. There existed a technical void between the well known optimal solution and the tailored solutions derived in previous works.

The fundamental requirement has always been to eliminate the global optimal solution from the class of allowable solutions since it typically depends on the expectation of interest. All previous approaches inherently imposed this constraint by choosing parametric families of density functions as the class of allowable biasing densities. This did eliminate the optimal solution, but it consequently restricted the performance of Importance Sampling.

Our approach differs in that we derived a general framework for determining a constrained optimal solution from any constraint set. This approach was based on showing that the "best" density function from a constraint set was the one which was closest to the optimal biasing density. Once we had determined the appropriate distance measure, we employed results from robust statistics to show that the least favorable density function from the constraint set was nearest the optimal density function and was therefore a constrained minimizer of the variance of the Importance Sampling estimator. This approach renders improved performance over all preceding techniques due to it's generality. Moreover, this strategy was shown to be near-optimum for every choice of functional $g(\cdot)$. We believe that practical biasing schemes will arise from this formulation when more constraint classes are investigated.
In the following chapter we will utilize the results and methodologies obtained so far in this thesis to improve the convergence properties of two stochastic approximation algorithms. As that we have demonstrated a strategy for determining near-optimum biasing densities, we will henceforth assume that all Importance Sampling implementations are near-optimum.
Least Favorable Densities

Figure 4.1 Least Favorable biasing densities for estimating the $\Pr [X \geq 5]$ with $\epsilon = .3$
Chapter 5

Stochastic System Design

5.1 Introduction

As stated in the Chapter 1, this thesis addresses the dual problem of both optimizing and estimating the performance of stochastic systems. Typically, the more complex the system, the less likely one can analytically determine the performance and the desired system parameters. Additionally, in most systems, one does not have access to the true value of the performance. Thus, if we wish to optimize the desired performance measure and not an analytically tractable heuristic, then the optimization algorithms we employ must rely solely on estimates of the performance.

Thus far in this thesis, we have developed a number of theoretical tools for efficiently estimating an arbitrary performance measure of a stochastic system. In this chapter, we exploit these results to improve the convergence properties of two well known stochastic optimization algorithms, the Robbins-Monro and Kiefer-Wolfowitz algorithms. These extremely versatile algorithms (or derivatives thereof) have been shown to be effective in applications ranging from system identification [73, 41], pattern recognition [68], process control [72] and even in determining the appropriate dosage levels in the design of medical treatments [42]. The primarily limitation
of these algorithms has been their relatively slow convergence rate. However, with the prevalence of faster computers, these algorithms are receiving renewed interest. Additionally, with the theory developed in the preceding chapters, we may effectively improve the convergence properties of these algorithms by reducing the sampling noise.

Even though there exists results pertaining to multi-dimensional stochastic approximation algorithms [68, 9], we will limit our investigation to the scalar design problem. It should be noted however that all results obtained in the duration of this chapter are directly extended to the vector parameter case.

5.2 Development

In almost all design problems involving systems of a stochastic nature, one is required to optimize some statistical measure of performance. This is typically a formatable and frequently intractable analytic problem. Therefore, it is often required to apply numerical techniques to affect a solution to the particular problem at hand.

Iterative techniques have been shown to be effective in solving a wide variety of numerical problems. A well known method of determining the root of a regression function is the so called Newton-Raphson method. It is easily described as follows:

let \( \psi(\theta) \), \( \theta \in \mathcal{R}^1 \) be an unknown function (performance measure) whose values can be determined at any point \( \theta \) and whose unique root \( \theta^* \) is sought while solving the
nonlinear equation

\[ \psi(\theta) = \alpha. \]  

(5.1)

If the error of computation of \( \psi(\theta) \) is negligible, then the root \( \theta^* \) can be obtained through the iterative procedure

\[ \theta_{n+1} = \theta_n - a_n [\psi(\theta_n) - \alpha] \]  

(5.2)

where \( a_n \) is the reciprocal of the first derivative of the regression function evaluated at \( \theta_n \). Under somewhat general conditions on \( \psi \) and \( \psi' \), it can be shown that \( \theta_n \to \theta^* \).

However, it may be and is often the case that \( \psi(\theta) \) is evaluated with significant error. Specifically, at each point \( \theta \), we observe \( \hat{\psi}(\theta) \) such that

\[ \psi(\theta) = E [\hat{\psi}(\theta)], \]  

(5.3)

where \( \hat{\psi}(\theta) \) is a random variable defined as

\[ \hat{\psi}(\theta) = \psi(\theta) + \eta(\theta). \]  

(5.4)

Robbins and Monro [59] in their seminal paper addressed just this problem. In the vein of Newton-Raphson, the Robbins-Monro stochastic approximation algorithm is defined as follows:

\[ \theta_{n+1} = \theta_n - a_n [\hat{\psi}(\theta_n) - \alpha] \]  

(5.5)

\[ = \theta_n - a_n [\psi(\theta_n) - \alpha] - a_n \eta(\theta_n). \]

Defining the deterministic sequence \( \{a_n\} \) such that \( a_n > 0 \) for all \( n \) and

\[ \sum_{n=1}^{\infty} a_n = \infty \quad \sum_{n=1}^{\infty} a_n^2 < \infty, \]  

(5.6)
Robbins and Monro determined conditions on $\psi$ and $\hat{\psi}$ under which their procedure converges in mean square, i.e.,

$$\lim_{n \to \infty} E (\theta_n - \theta^*)^2 = 0. \quad (5.7)$$

Arising from this initial work, there has been an extensive body of research to determine less restrictive conditions under which $\theta_n \to \theta^*$ [11, 76, 18, 8]. Additionally, there has been work to determine the best form of $\{a_n\}$ [72] for accelerated convergence and also to determine the limiting distribution at $\theta^*$ [19, 60].

As a parallel research initiative, Kiefer and Wolfowitz [37] introduced an alternate iterative algorithm for determining the extremum of the regression function $\psi$. Their algorithm is defined as follows:

$$\theta_{n+1} = \theta_n - \frac{a_n}{c_n} \left[ \hat{\psi}(\theta_n + c_n) - \hat{\psi}(\theta_n - c_n) \right] \quad (5.8)$$

$$= \theta_n - 2a_n \left[ \frac{\psi(\theta_n + c_n) - \psi(\theta_n - c_n)}{2c_n} \right] - \frac{a_n}{c_n} [\eta(\theta + c_n) - \eta(\theta - c_n)] \quad (5.9)$$

where $\{a_n\}$ and $\{c_n\}$ are such that $a_n > 0$, $c_n > 0$ and

$$\sum_{n=1}^{\infty} a_n = \infty \quad \sum_{n=1}^{\infty} \frac{a_n^2}{c_n^2} < \infty \quad c_n \to 0. \quad (5.10)$$

This algorithm shares many similarities with the Robbins-Monro optimization algorithm. Consider that under appropriate smoothness conditions,

$$\frac{\psi(\theta + c_n) - \psi(\theta - c_n)}{2c_n} \quad (5.11)$$
is a convergent central difference approximation to \( \psi' \). Therefore this algorithm attempts to determine the solution to

\[
\psi'(\theta) = 0
\]  

(5.12)

without explicit knowledge of the derivative. Kiefer and Wolfowitz set forth conditions under which their algorithm converged in mean square to \( \theta^* \). As with Robbins-Monro, these conditions have been softened and extended to demonstrate almost sure convergence \([8, 11, 18]\). Additionally, there has exists a large body of related research centering on determining both the limiting distribution and the optimum sequences \( \{a_n\} \) and \( \{c_n\} \)\([72] \) and references within).

These two algorithms have not received much attention recently because of their relatively slow convergence properties. However, with the theory developed in the previous chapters we will demonstrate improved convergence rates by effectively reducing the observation noise. These results will be quantified by determining the lengths of the iterative algorithms both with and without the implementation of Importance Sampling.

5.3 System Design

Consider the general problem of determining the real parameter \( \theta \) such that the performance measure

\[
\psi(\theta) = \mathbb{E}_{P}[g(X, \theta)]
\]  

(5.13)
achieves some level of performance, where \( X \) is a random vector with distribution function \( P \) and \( g(\cdot) \) is Borel measurable. There are two distinct but related problems in system design. We either wish to determine the system parameter such that some prescribed level of performance is achieved or alternately we wish to determine the parameter such that the optimal level of performance is achieved. We can formulate these criterion in the following two ways:

\[
\theta^* = \arg_{\theta \in \mathbb{R}^1} \psi(\theta) = \alpha
\]

(5.14)

\[
\theta^* = \arg_{\theta \in \mathbb{R}^1} \min \psi(\theta).
\]

(5.15)

If analytic expressions for \( \psi(\theta) \) and \( \psi'(\theta) \) exist, then solutions to the above problems can be obtained by inverting \( \psi \) for problem (5.14) and inverting \( \psi' \) for problem (5.15). However, this fortuitous situation does not generally exist. It is often the case that we can only estimate these functions at every level \( \theta \). Using these estimated values, we can then apply the techniques of Robbins-Monro and Kiefer-Wolfowitz to determine \( \theta^* \).

5.3.1 Monte Carlo Estimates

The universally applied method of Monte Carlo is typically used to estimate expectations of functions of random vectors. In the event that the distribution function of \( X \) is known, the sample estimate of \( \psi(\theta) \) is constructed by simple averaging over the
functional $g$ with respect to the underlying distribution function, i.e.,

$$\hat{\psi}(\theta) = \frac{1}{M_{MC}} \sum_{i=1}^{M_{MC}} g(X_i, \theta),$$  \hspace{1cm} (5.16)

where $X_i \in \mathcal{R}^N$ are independent and identically distributed realizations of the random vector $X$ with distribution function $P$, and where the integer $M_{MC}$ is the number of Monte Carlo simulation trials. As was shown previously in this thesis, the variance of the above estimator is given by

$$\text{var}_P[\hat{\psi}(\theta)] = \frac{\text{var}_P[g(X, \theta)]}{M_{MC}} = \frac{\int_{\mathcal{R}^N} g^2(x, \theta)p(x, \theta)dx - \psi(\theta)^2}{M_{MC}}. \hspace{1cm} (5.17)$$

It can be shown that if $g$ is both square integrable and absolutely integrable then $\hat{\psi} \rightarrow \psi$ in both the mean square and almost sure sense [63].

As before, the Importance Sampling estimator for the above performance measure is given by

$$\psi^*(\theta) = \frac{1}{M_{IS}} \sum_{i=1}^{M_{IS}} W(X_i, \theta)g(X_i, \theta), \hspace{1cm} (5.18)$$

where $X_i$ are generated from the biasing distribution function $P_1$ and where the weighting function $W$ is equal to the likelihood ratio $\frac{P(x, \theta)}{P_1(x, \theta)}$. Recall from Chapter 4 that the variance of $\psi^*$ (with its associated lower bound) is given by

$$\text{var}_{P_1}[\psi^*(\theta)] = \frac{\int_{\mathcal{R}^N} \frac{p^2(x, \theta)g^2(x, \theta)}{p_1(x, \theta)}dx - \psi^2(\theta)}{M_{IS}} \geq \frac{E_P^2|g(X, \theta)| - \psi^2(\theta)}{M_{MC}}. \hspace{1cm} (5.19)$$

Additionally, recall from Definition 4.1 that the biasing density $p_1(x, \theta)$ is said to be derived from a near-optimum biasing strategy if

$$\text{var}_{P_1}[\psi^*(\theta)] \leq \frac{\Delta E_P^2|g(X, \theta)| - \psi^2(\theta)}{M_{IS}} \text{ for } 1 < \Delta < \infty \hspace{1cm} (5.20)$$
for every functional $g(z, \theta)$ and density function $p(z, \theta)$. In this thesis, we have obtained two distinct solutions to the problem of determining the appropriate biasing density $p_1$ such that we reduce the variance of the system performance estimate at the operating point $\theta$. Moreover, the strategy of choosing the least favorable density from the $\epsilon$-mixture constraint class was shown to be near-optimum for any functional $g(\cdot)$. Additionally, in Chapters 2 and 3 we have empirically demonstrated that under certain restrictions, the conditional linear shift biasing density is near-optimum when the functional $g$ defines the error rate in a detection system. In the duration of this chapter we assume that all Importance Sampling implementations are near-optimum. Further, we quantify the effects of Importance Sampling on the Robbins-Monro and Kiefer-Wolfowitz algorithms and supply some asymptotic results for the case the performance measure is some set of events.

5.4 The Robbins-Monro Algorithm

In this section we analyze the Robbins-Monro stochastic optimization algorithm in greater detail for both the finite sample and asymptotic scenarios. In particular, we seek to quantify the effects of the inherent observation noise in the estimation process of the performance measure. From this analysis, we will be able to determine the improvement in the rate of convergence under the application of Importance Sampling.
To this end, we begin by establishing some results for the case where the regression function $\psi$ is linear in $\theta$, i.e.,

$$\psi(\theta) = E_P [g(X, \theta)] = \alpha_1 (\theta - \theta^*) + \alpha,$$  

(5.21)

with the added conditions that the derivative of the regression function evaluated at the fixed level $\theta^*$ is positive, i.e., $\alpha_1 > 0$ and where $\alpha$ is an offset.

Our interest is to determine $\theta$ such that $\psi(\theta) = \alpha$ given that we observe $\hat{\psi}(\theta) = \psi(\theta) + \eta(\theta)$, where $\eta$ is zero mean random variable with finite second moment. The iterative algorithm for determining $\theta^*$ in this case is given by

$$\theta_{n+1} = \theta_n - a_n (\hat{\psi}(\theta_n) - \alpha)$$

$$= \theta_n - a_n (\alpha_1 [\theta_n - \theta^*] + \eta(\theta_n)).$$

(5.22)

We can observe that the above algorithm decreases the iterate $\theta_n$ if the estimate is greater than the desired level $\alpha$ and increases if the converse is true. As stated before, without the presence of the observation noise, we can guarantee that $\theta_n \to \theta^*$. It is not obvious that we have convergence in a noisy environment. However, in the following analysis we will not only demonstrate convergence but will determine the rate as well.

Since we are interested in the convergence properties of $\theta_n \to \theta^*$ we choose to write the above equation as

$$\theta_{n+1} - \theta^* = \theta_n - \theta^* - a_n \alpha_1 [\theta_n - \theta^*] - a_n \eta(\theta_n)$$

$$= (\theta_n - \theta^*) [1 - a_n \alpha_1] - a_n \eta(\theta_n).$$

(5.23)
Recognizing that we can rewrite the recursive equation \( X_{n+1} = d_n X_n - b_n \) as

\[
X_n = X_1 \prod_{i=1}^{n-1} d_i - \sum_{i=1}^{n-1} b_i \prod_{j=i+1}^{n-1} d_j,
\]

we obtain the following form for the iterate \( \theta_n \) solely as a function of the sequence \( a_n \) and constant \( \alpha_1 \) as

\[
\theta_n - \theta^* = (\theta_1 - \theta^*) \prod_{i=1}^{n-1} [1 - a_i \alpha_1] - \sum_{i=1}^{n-1} a_i \eta(\theta_i) \prod_{j=i+1}^{n-1} [1 - a_j \alpha_1]
\]

(5.25)

where \( \theta_1 \) can be chosen to be a fixed number or a random variable with finite first and second moments.

To proceed in the analysis, it is necessary to specify the form of the sequence \( a_n \). Recall that the sequence must satisfy the conditions that it be square summable but not summable (see equation (5.6)) to obtain mean square convergence. Results set forth by Robbins and Monro [59] specify the form as \( a_n = A/n \) with \( A > 0 \). Subsequent results pertaining to the limiting distribution of \( \theta_n \) have also restricted the sequence to be of \( n^{-1} \) type [60, 72]. Therefore, in the duration of this chapter we will similarly limit \( a_n \) to be of the above form.

Making use of the definition that \( A' = A \alpha_1 \) we can define the partial product in (5.25) as

\[
\beta_{i,n-1} = \prod_{i=1}^{n-1} [1 - A'/j]
\]

(5.26)

to obtain the following form of the difference of \( \theta_n - \theta^* \) as

\[
\theta_n - \theta^* = (\theta_1 - \theta^*) \beta_{0,n-1} - \frac{1}{\alpha_1} \sum_{i=1}^{n-1} A' \eta(\theta_i) \beta_{i,n-1}.
\]

(5.27)
From this it is direct to show that the expected deviation at the $n^{th}$ iteration is given by

$$E \left[ \theta_n - \theta^* \right] = E\left[ (\theta_1 - \theta^*) \right] \beta_{0,n-1}. \quad (5.28)$$

By applying well established bounds on $\beta$ [60] it can be shown that

$$k_1(n)n^{-A'} \leq E \left[ \theta_n - \theta^* \right] \leq k_2(n)n^{-A'} \quad (5.29)$$

where $k_1(n) \leq 1 \leq k_2(n)$ are bounded sequences with $\lim_{n \to \infty} k_i(n) = 1$ for $i = 1, 2$. We can infer from this result that the bias of $\theta_n$ tends to zero of order at least $O(n^{-A'})$. Therefore we can obtain mean square convergence of $\theta^*$ if we can show that the variance diminishes.

Under the assumption that the observation noise is independent we have the closed form expression for the mean square error at the $n^{th}$ iteration given by

$$E \left[ (\theta_n - \theta^*)^2 \right] = E \left[ (\theta_1 - \theta^*)^2 \right] \beta_{0,n-1}^2 + \frac{1}{\alpha_1^2} \sum_{i=1}^{n-1} \frac{A_i^2}{i^2} \text{var}[\eta(\theta_i)] \beta_{i,n-1}^2. \quad (5.30)$$

Appealing to a lemma due to Sacks [60], we can show that asymptotically

$$\lim_{n \to \infty} n \text{var} \left[ (\theta_n - \theta^*) \right] = \frac{A^2 \text{var}[\eta(\theta^*)]}{2A\alpha_1 - 1}. \quad (5.31)$$

From the above equation, it is evident that we obtain the mean square convergence of $\theta_n \to \theta^*$ with rate $O(n^{-1})$. This relatively slow rate of convergence has in the past inhibited the practical use of the Robbins-Monro algorithm. However as stated earlier, we may significantly reduce the numerator of the asymptotic variance by obtaining better estimates of $\psi$ via Importance Sampling. In this way we will be able to achieve
a desired confidence interval with much fewer iterations of \( n \), thereby effectively over-
coming the relatively slow convergence rate. Before proceeding to evaluating the
stopping times for Robbins-Monro, we supply convergence and asymptotic results for
more general non-linear performance measures.

Consider the following conditions on \( \psi(\theta) \) and \( \eta(\theta) \):

(R1) \( \psi(\theta) \) is a Borel measurable function.

(R2) For each \( \epsilon > 0 \), \( \inf_{\epsilon < \theta - \theta^* < \epsilon} \psi(\theta) - \alpha > 0 \) and \( \sup_{\epsilon < \theta - \theta^* < \epsilon} \psi(\theta) - 0. \)

(R3) For some constants \( K_1 \) and \( K_2 \), \( |\psi(\theta) - \psi(\theta^*)| \leq K_1 + K_2|\theta - \theta^*| \forall \theta. \)

(R4) \( \sup_{\theta} E[\eta^2(\theta)] = \varsigma < \infty. \)

(R5) \( \lim_{\theta \to \theta^*} E[\eta^2(\theta)] = E[\eta^2(\theta^*)] = \sigma^2 > 0. \)

(R6) \( \lim_{R \to \infty} \lim_{\epsilon \to 0^+} \sup_{|\theta - \theta^*| \leq \epsilon} \int_{|\eta(\theta)| > R} \eta^2(\theta) dP = 0. \)

(R7) For some positive constant \( l \) and \( \alpha_1 \), if \( (\theta - \theta^*) < l \)
then \( \psi(\theta) = \alpha_1(\theta - \theta^*) + \delta(\theta) \) where \( \delta(\theta) = o(|\theta - \theta^*|) \) as \( \theta \to \theta^*. \)
By \( \delta(x) = o(|x|^p) \) we mean that as \( x \to 0 \) we have that \( \delta(x) = \zeta(x)|x|^p \) where \( \zeta(x) \to 0 \). Additionally, we require via the above conditions that the absolute difference between the performance and the desired level be bounded by arbitrary linear functions. Moreover, for obvious reasons, it is required that \( \psi \) not cross the level \( \alpha \) more than once. The restrictions on the observation noise are only that its density function be well behaved (see (R5-R6)) and that it have a finite second moment. Additionally, it can be shown that a sufficient condition for (R7) to hold is the existence of the derivative of the performance at \( \theta^* \). Given these restrictions, we are able to address the issues of convergence and asymptotic distributions.

Under slightly weaker conditions than those given in (R1) through (R4), Blum [8] demonstrated the following convergence results (see also [18]).

**Theorem 5.1** Under conditions (R1) through (R4) we have that

\[
\lim_{n \to \infty} Pr[\theta_n = \theta^*] = 1
\]

and that

\[
\lim_{n \to \infty} E[(\theta_n - \theta^*)^2] = 0.
\]

What remained to be determined was the asymptotic distribution of the iterative process \( \{\theta_n\} \). This result was obtained by Sacks [60] under conditions (R1) through (R7) and subsequently sharpened by Gladyshev [21]. It is summarized as follows:

**Theorem 5.2** Suppose that (R1) through (R7) are satisfied. Let \( A \) be such that \( 2A\alpha_1 > 1 \), where \( \alpha_1 \) is defined in (R7). Then the random
variable \( \sqrt{n}(\theta_n - \theta^*) \) is asymptotically normally distributed with mean 0 and variance \( A^2\sigma^2(2A\alpha_1 - 1)^{-1} \).

Note that the asymptotic variance is consistent with the result obtained in (5.31) earlier in this chapter. This above result is significant in that it also supplies the asymptotic distribution function about the limit point. From this we can determine stopping times for the Robbins-Monro algorithm based upon \( \theta_n - \theta^* \).

However, in system design, one is typically more interested in obtaining a performance which is near the desired level as opposed to obtaining parameters which are near their desired values. As such, one would be more interested in knowing the limiting distribution of \( \psi(\theta_n) - \alpha \). In this regard, we may modify the above theorem by recognizing that

\[
\psi(\theta_n) - \alpha = \frac{\psi(\theta_n) - \alpha}{(\theta_n - \theta^*)} \cdot (\theta_n - \theta^*).
\]

Further, we note that condition (R7) along with (R1) through (R4) are sufficient to insure that

\[
\frac{\psi(\theta_n) - \alpha}{\theta_n - \theta^*} \rightarrow \alpha \quad a.s.
\]

From this it is easily verified that \( \psi(\theta_n) - \alpha \) is asymptotically normally distributed with mean 0 and variance \( A^2\alpha_1 \sigma^2(2A\alpha_1 - 1)^{-1}n^{-1} \). This extension will allow us to determine the stopping time based upon the performance level rather than the system parameter.
5.4.1 Stopping Times for the Robbins-Monro Algorithm

Limited results exist in the area of stopping times for the Robbins-Monro algorithm [64, 66]. Of these, results have centered on estimating parameters of the asymptotic variance so that the optimal numerator for the sequence $a_n$ can be approximated. This problem will not be addressed in this chapter. Rather, we seek to quantify the effects of Importance Sampling on the Robbins-Monro algorithm. More specifically, the reduction in iterations $n$ required for a fixed confidence interval will be derived. To this end, we begin by defining the stopping time for the Robbins-Monro algorithm.

**Definition 5.1** The stopping time $N_{RM}$ is defined as the smallest integer $n$ such that

$$Pr [ | \psi(\theta_n) - \alpha | > \varepsilon ] < \delta,$$

(5.36)

where $\varepsilon$ and $\delta$ are fixed constants which define the confidence interval about $\alpha$.

We can determine $N_{RM}$ by utilizing Theorem 5.2. We begin by rewriting the desired confidence interval in its complementary form:

$$Pr [-\varepsilon \leq \psi(\theta_n) - \alpha \leq \varepsilon] \geq 1 - \delta.$$  

(5.37)

Assuming that the theorem holds with equality for large $n$, it can be shown that $N_{RM}$ is the smallest integer such that

$$n \geq \frac{K^2 A^2 \alpha^2 \sigma^2}{\varepsilon^2 (2A\alpha_1 - 1)}$$

(5.38)
where $K_t = Q^{-1}(\delta/2)^1$. Observe that we can reduce the stopping time for a fixed confidence interval by reducing the effect of the estimation noise at $\theta^*$. We quantify this through the following two equations:

$$N_{RM} = \left[ \frac{K_t^2 A^2 \alpha^2 \text{var}[\hat{\psi}(\theta^*)]}{\varepsilon^2(2A\alpha_1 - 1)} \right]$$

(5.39)

$$N_{RM}^* = \left[ \frac{K_t^2 A^2 \alpha^2 \text{var}[\psi^*(\theta^*)]}{\varepsilon^2(2A\alpha_1 - 1)} \right]$$

(5.40)

where $[\cdot]$ is the ceiling function and where $N_{RM}$ and $N_{RM}^*$ are the stopping times of the Robbins-Monro algorithm under standard Monte Carlo and Importance Sampling simulations, respectively. Substituting in the expressions for the variance of the estimate we have that

$$N_{RM} = \left[ \frac{K_t^2 A^2 \alpha^2 \text{var}[g(X, \theta^*)]}{M_{MC}(\theta^*)\varepsilon^2(2A\alpha_1 - 1)} \right]$$

(5.41)

$$N_{RM}^* = \left[ \frac{K_t^2 A^2 \alpha^2 \text{var}[W(X, \theta^*)g(X, \theta^*)]}{M_{IS}(\theta^*)\varepsilon^2(2A\alpha_1 - 1)} \right]$$

(5.42)

where $M_{MC}(\theta^*)$ and $M_{IS}(\theta^*)$ are the lengths of the Monte Carlo and Importance Sampling estimators at level $\theta^*$.

As a measure of comparison, we define the computational complexity $\mu$ for both the Monte Carlo and Importance Sampling implementations as

$$\mu_{MC} = \sum_{n=1}^{N_{RM}} M_{MC}(\theta_n) \text{ and } \mu_{IS} = \sum_{n=1}^{N_{RM}^*} M_{IS}(\theta_n).$$

(5.43)

These measures can be viewed as the total computations required in the respective implementations of the optimization algorithm.

---

1Note that as $\varepsilon \to 0$, the stopping time given in (5.38) tends to the similarly derived stopping time for the case that $\psi(\theta_n) - \alpha$ is normally distributed for all $n$.
Using these definitions, we may now make some direct comparisons between standard Robbins-Monro and Robbins-Monro with Importance Sampling.

**Scenario 1** Consider the example where we choose the number of estimation trials, $M_{MC}$ and $M_{IS}$, such that at each level $\theta_n$ we have that

$$\text{var} \left[ \hat{\psi}(\theta_n) \right] = \text{var} \left[ \psi^*(\theta_n) \right].$$

In this way, the stopping times for the two optimization schemes will be the same (denoted by $N$). Moreover, from (2.28) in Chapter 2 we may relate $M_{MC}(\theta_n)$ to $M_{IS}(\theta_n)$ in this scenario through the gain evaluated at $\theta_n$ by

$$M_{IS}(\theta_n) = \frac{M_{MC}(\theta_n)}{\Gamma(\theta_n)}. \quad (5.45)$$

Therefore we have that the computational complexity using Importance Sampling is given by

$$\mu_{IS} = \sum_{n=1}^{N} \frac{M_{MC}(\theta_n)}{\Gamma(\theta_n)}. \quad (5.46)$$

Note that the computational savings under this comparison arise solely from the reduction of estimator trials at every level $\theta_n$. Applying the obvious bounds on the $\mu_{IS}$, it can be shown that

$$\min_{\theta_n} \Gamma(\theta_n) \leq \frac{\mu_{MC}}{\mu_{IS}} \leq \max_{\theta_n} \Gamma(\theta_n). \quad (5.47)$$

The above bounds demonstrate that with an effective implementation of Importance
Sampling at every level \( \theta_n \), we obtain a substantial reduction in total computations. In the following proposition, we supply the asymptotic result for the case that the performance measure is the probability of a set in the observation space \( \mathcal{R}^N \).

**Proposition 6** Consider the case where \( g(X, \theta) \) is an indicator function, \( \psi(\theta) \) satisfies (R1) through (R7) for every \( \alpha \neq 0 \), and the biasing density is derived from a near-optimum strategy. Let \( \epsilon = k\alpha, k > 0 \) and \( \delta > 0 \). Additionally, set \( \infty > M_U \geq M_{MC}(\theta) \geq M_L > 0 \). Then under complexity Scenario 1

\[
\lim_{\alpha \to 0^+} \frac{\mu_{MC}}{\mu_{IS}} = \infty \text{ a.s.} \tag{5.48}
\]

**Proof** Under the conditions of the proposition, we have that \( \lim_{\alpha \to 0^+} N_{RM} = N_{RM}^* = \infty \). Additionally, the conditions set forth in Theorem 5.1 guarantee that \( \theta_n \to \theta^* \) almost surely. Given that \( \Gamma(\theta_n) \geq \frac{\Delta'}{\psi(\theta_n)} \) for the prescribed choice of biasing density, we have that \( \lim_{\alpha \to 0^+} \Gamma(\theta_{N_{RM}}) = \infty \) almost surely, leading to the desired result.

This proposition demonstrates that as the desired performance level (given by a probability measure of some error set) tends to its minimum achievable value, we have that the computational savings in applying Importance Sampling are unbounded. Therefore in optimizing systems with respect to error rates or false alarm probabilities, we substantially reduce the total computations with an effective implementation of Importance Sampling.
Scenario 2 As an alternate scenario for comparison, consider the case where \( M_{MC}(\theta_n) = M_{IS}(\theta_n) = M(\theta_n) \geq M \) at all levels \( \theta_n \). It is easily shown that the stopping times are related by the following equality

\[
N_{RM}^* = \frac{N_{RM}}{\Gamma(\theta^*)}.
\] (5.49)

Note that under this comparison scenario, the computational savings arise strictly from the reduction in the stopping time. The two computational complexities can be written as

\[
\mu_{MC} = \sum_{n=1}^{N_{RM}} M(\theta_n) \quad \text{and} \quad \mu_{IS} = \sum_{n=1}^{N_{RM} / \Gamma(\theta^*)} M(\theta_n).
\] (5.50)

If we impose the additional condition that \( M(\theta_n) \) be invariant to \( \theta_n \) we then obtain the following equality

\[
\frac{\mu_{MC}}{\mu_{IS}} = \Gamma(\theta^*). \tag{5.51}
\]

Although the reductions in computations varies with respect to the form of \( g(X, \theta) \), we can again supply the following asymptotic results under this comparison scenario for the case where the performance is given by the probability measure of a set in \( \mathcal{R}^N \).

**Proposition 7** Consider the case where \( g(X, \theta) \) is given by an indicator function, \( \psi(\theta) \) satisfies (R1) through (R7), and the biasing density is derived from a near-optimum strategy. Let \( \epsilon = k\alpha, k > 0 \) and \( \delta > 0 \) with the stopping times given by (5.41) and (5.42). Then under complexity
Scenario 2 we have the following:

\[
\lim_{a \to 0^+} \frac{\mu_{MC}}{\mu_{IS}} = \infty
\]

\[
\lim_{a \to 0^+} \frac{N_{RM}^*}{N_{RM}} = \infty
\]

\[
\lim_{a \to 0^+} N_{RM}^* < \infty
\]

Proof. From the conditions set forth in the statement of the proposition, it can be shown that \(\lim_{a \to 0^+} N_{RM} = \infty\). We additionally have that \(\lim_{a \to 0^+} \Gamma(\theta^*) = \infty\). Therefore, by use of the upper bound on the variance of \(\psi^*\) given in (5.20) we obtain the desired results.

This proposition demonstrates that in systems where the performance criterion is the probability of some event of \(X\), the computational savings and the reduction in the stopping time are asymptotically unbounded while the Importance Sampling stopping time is always finite.

It is not unexpected that the reduction in computation given by (5.47) and (5.51) is dependent on the Importance Sampling scheme and performance measure to be optimized. As an example of particular interest, we can consider the problem of determining the system parameter of a detection system when the performance measure is the probability of error.

Determining signal parameters in the presence of non-Gaussian disturbances is typically an analytically intractable problem. To affect any analytic solution, either suboptimal performance measures must be employed [34] or we must appeal to
asymptotic approximations [32]. Thus our only recourse in obtaining exact solutions
for finite sample systems is to utilize numerical techniques. To illustrate the benefits
in employing Importance Sampling in the Robbins-Monro algorithm, we supply the
following example of determining a one dimensional signal parameter in the presence
of non-Gaussian noise.

Example 1: Signal Design

Consider the following binary hypotheses testing problem:

\[ H_0 : R_i \triangleq \eta_i \]
\[ i = 1, 2 \]

\[ H_1 : R_i \triangleq S_i(\theta) + \eta_i \]

where \( S_1(\theta) = \sqrt{E}\cos(\theta), S_2(\theta) = \sqrt{E}\sin(\theta) \) and \( \eta_i \) are independent and identically
distributed random variables with Laplacian density function. The design problem
is as follows: determine the real angle \( \theta \) such that the performance of the optimal
detection system for the above problem achieves an error rate of \( 1.5 \times 10^{-8} \) given that
the system is operating at \( 26dB \) signal to noise ratio conditioned on \( H_1 \). See Figure
5.1 for an illustration of the problem definition. It can be verified that the problem
meets the requirements (R1) through (R7) for convergence.

\[ ^2\text{Since the error rate is periodic in this example we restrict } \theta \in [0, 45^\circ]. \]
Determine \( \theta \) such that minimum error rate is \( \alpha = 1.5 \times 10^{-8} \).

While this problem can be solved through alternate numerical techniques (\( \theta^* = 30.8^\circ \)), it does illuminate the potential of this approach. To begin, we set the constant \( A \) equal to its optimal\(^3 \) value \( 1/\alpha_1 \) thereby rendering the respective stopping times as

\[
N_{RM} = \left[ \frac{K_2^2 \text{var}_{P}[g(X, \theta^*)]}{M_{MC}(\theta^*)\epsilon^2} \right] \quad \text{and} \quad N_{RM}^* = \left[ \frac{K_2^2 \text{var}_{P}[W(X, \theta^*)g(X, \theta^*)]}{M_{IS}(\theta^*)\epsilon^2} \right].
\]

Furthermore, we set \( \delta = .05 \) and \( \varepsilon = \frac{\sigma}{10} \). After computing the remaining terms in the above equation, it can be shown that under the standard Robbins-Monro implementation we require that

\[
N_{RM} M_{MC}(\theta^*) \geq 2.56 \times 10^{10}.
\]

\(^3\)This choice of \( A \) minimizes the asymptotic variance of \( \{\theta_n - \theta^*\} \).
However, if we choose to employ the conditional linear shift density derived in Chapter 3 as our biasing density we may reduce the above lower bound to

\[ N_{RM}^* M_{IS}(\theta^*) \geq 5.8 \times 10^6. \] (5.54)

Equations (5.53) and (5.54) quantify the trade-off between the length of the performance estimators and the number of iterations of the Robbins-Monro algorithm. Using (5.54), we can show that under comparison Scenario 1 we have \( 1.9 \times 10^3 \leq \frac{\mu_{MC}}{\mu_{IS}} \leq 1.0 \times 10^5 \) and under Scenario 2 we have \( \frac{\mu_{MC}}{\mu_{IS}} = 4.4 \times 10^4 \).

Evaluating the computational reduction by employing the solution from the \( \epsilon \)-mixture constraint class derived in Chapter 4 (\( \epsilon = .5 \)), we can further reduce the above lower bounds to

\[ N_{RM}^* M_{IS}(\theta^*) \geq 3.8 \times 10^2, \] (5.55)

yielding under Scenario 1 that \( 2.9 \times 10^6 \leq \frac{\mu_{MC}}{\mu_{IS}} \leq 1.0 \times 10^8 \) and under Scenario 2 \( \frac{\mu_{MC}}{\mu_{IS}} = 6.5 \times 10^7 \).

These above results imply that we may obtain a significant reduction in the total computations/stopping times in the Robbins-Monro algorithm with a successful implementation of Importance Sampling. Furthermore, it appears that this savings may be enough to warrant the reevaluation of Robbins-Monro as a viable algorithm for the design of decision systems.
5.5 The Kiefer-Wolfowitz Algorithm

In the previous section we determined the parameter $\theta$ such that the performance obtained some prescribed level. In this section we analyze the Kiefer-Wolfowitz algorithm for determining $\theta$ such that the performance measure achieves its extreme value [37]. As before, we seek to determine the effect of the observation noise on both the stopping time and the computational complexity.

We begin by establishing the following finite sample result for the instance that $\psi$ is quadratic. Consider the case where we wish to determine $\theta$ such that

$$\psi(\theta) = E[g(X, \theta)] = \alpha + \alpha_2(\theta - \theta^*)^2$$

for $\alpha_2 > 0$ (5.56)

achieves its minimum value given that we observe $\hat{\psi}(\theta) = \psi(\theta) + \eta(\theta)$, where $\eta$ is a zero mean random variable with finite second moment. As discussed before in this chapter, the Kiefer-Wolfowitz algorithm for this form of $\psi$ is given by

$$\theta_{n+1} = \theta_n - \frac{a_n}{c_n} \left[ \hat{\psi}(\theta_n + c_n) - \hat{\psi}(\theta_n - c_n) \right]$$

(5.57)

$$= \theta_n - \frac{a_n}{c_n} \left[ \alpha_2(\theta_n + c_n - \theta^*)^2 - \alpha_2(\theta_n - c_n - \theta^*)^2 + \eta_n \right]$$

(5.58)

where $\eta_n = \eta(\theta_n + c_n) - \eta(\theta_n - c_n)$ and $a_n, c_n$ satisfy the conditions set forth in (5.10).

Simplifying, we can show that

$$\theta_{n+1} - \theta^* = [\theta_n - \theta^*](1 - 4\alpha_2 a_n) - \frac{a_n}{c_n}\eta_n.$$  

(5.59)

Using the relationship in (5.24), we may rewrite the above recursive equation as

$$\theta_n - \theta^* = (\theta_1 - \theta^*) \prod_{i=1}^{n-1} [1 - 4\alpha_2 a_j] - \sum_{i=1}^{n-1} \frac{a_i}{c_i} \prod_{j=i+1}^{n-1} (1 - 4\alpha_2 a_j).$$

(5.60)
where $\theta_1$ is restricted to have finite first and second moments. Let $a_n = A/n$, $c_n$ be such that $a_n/c_n$ is square summable. Let $A' = 4A\alpha_2$ and $\beta_{i,n-1}$ be as defined in (5.26), then using arguments similar to those in Section 5.4 it can be shown that the bias of $\theta_n$ is given by

$$E[\theta_n - \theta^*] = E[(\theta_1 - \theta^*)] \beta_{0,n-1}$$

(5.61)

with the associated mean square deviation

$$E[(\theta_n - \theta^*)^2] = E[(\theta_n - \theta^*)^2] \beta_{0,n-1}^2 + \sum_{i=1}^{n-1} \frac{A^2c_i^{-2}}{i^2} \text{var}[\eta_i] \beta_{i,n-1}^2.$$  

(5.62)

Again appealing to the lemma due to Sacks [60] it can be shown that

$$k_1(n)n^{-4\alpha_2A} \leq E[\theta_n - \theta^*] \leq k_2(n)n^{-4\alpha_2A}$$

$$\lim_{n \to \infty} \sqrt{n}c_n \text{var}[(\theta_n - \theta^*)] = \frac{2A^2\text{var}[\eta(\theta^*)]}{8\alpha_2A - 1},$$

where $k_i(n)$ are bounded sequences tending to unity. Observe from these above results that $\theta_n$ converges in mean square to $\theta^*$ at least of order $O(n^{-1-c_n^{-2}})$ under the prescribed restrictions on $c_n$. Typically, $c_n$ takes the form of $C/n^\gamma$ with $\gamma \in (0,1/2)$. This constraint renders the convergence rate $O(n^{1-2\gamma})$, implying that at best we can approximate but not achieve $1/n$ convergence. However, as demonstrated in the previous section, we may effectively accelerate convergence by reducing the estimation noise at $\theta^*$. To this end, we begin by supplying asymptotic convergence properties for a more general class of performance measures.

Consider the following conditions on $\psi(\theta)$ and $\eta(\theta)$:
(K1) \( \psi(\theta) \) is a Borel measurable function with unique minimum at \( \theta^* \).

(K2) For \( 0 < t_0 < t_1 < t_2 < \infty \),

\[
\inf_{t_1 < \theta - \theta^* < t_2} \frac{(\theta - \theta^*)(\psi(\theta + \varepsilon) - \psi(\theta - \varepsilon))}{\varepsilon} > 0.
\]

where \( 0 < \varepsilon \leq t_0 \)

(K3) There exists a constant \( c_1 \) such that for all \( c \) with \( 0 < c < c_1 \)

and for some constants \( K_1 \) and \( K_2 \), \( |\psi(\theta + c) - \psi(\theta)| < K_1 + K_2|\theta| \) \( \forall \theta \).

(K4) For all \( \theta \), \( \psi(\theta) = \alpha + \alpha_2(\theta - \theta^*)^2 + \delta(\theta, \theta^*) \) where \( \alpha > 0 \)

and \( \delta = o(|\theta - \theta^*|^2) \) as \( \theta \to \theta^* \).

(K5) \( \sup_{\theta} E[\eta^2(\theta)] = \varsigma < \infty \).

(K6) \( \lim_{\theta \to \theta^*} E[\eta^2(\theta)] = E[\eta(\theta^*)] > 0 \).

(K7) \( \lim_{R \to \infty} \lim_{\varepsilon \to 0^+} \sup_{|\theta - \theta^*| \leq \varepsilon, |\eta(\theta)| > R} \int_{|\eta(\theta)| > R} \eta^2(\theta) dP = 0 \).
(K8) For some positive constants \( \epsilon, c_0, K_1 \), with \( \epsilon > c_0 \) and \( K_1 \leq 4\alpha_2 \) we have that for all \( c \leq c_0 \) and all \( \theta \) such that \( c < |\theta - \theta^*| < \epsilon \)

\[
(\theta - \theta^*)[\psi(\theta + c) - \psi(\theta - c)]c^{-1} > K_1(\theta - \theta^*)^2.
\]

These above conditions are consistent with usual regularity conditions required for the uniqueness of the global minimizer. Conditions (K4) and (K8) are both implied by the continuity of the second derivative of \( \psi \) in the neighborhood of \( \theta^* \). Additionally, we require that certain regularity conditions on the noise be satisfied. Given the above, we may now state the following convergence results.

**Theorem 5.3** Under conditions (K1)-(K6), we have that \( \theta_n \to \theta^* \) in the mean square and almost sure senses [8, 11, 18].

By further specifying the form of \( c_n \) as \( C/n^{1/4} \) we can obtain the limiting distribution through the following theorem 4.

**Theorem 5.4** Suppose that (K1) through (K8) are satisfied. Let \( A \) be such that \( 8A\alpha_2 > 1 \). Then the random variable \( n^{-1/4}(\theta_n - \theta^*) \) is asymptotically normally distributed with mean 0 and variance \( 2A^2\text{var}[\eta(\theta^*)](8\alpha_2A - 1)^{-1}C^{-1} \) [60].

---

4This form of \( c_n \) was shown to be optimal in the sense that it maximized the rate of convergence for every distribution function of the underlying random variable \( X \) [17].
Again, this result is consistent with the finite sample analysis presented earlier in this section. As before, we can determine the stopping times and computational complexity for the Kiefer-Wolfowitz algorithm by exploiting Theorem 5.4.

As a consequence of condition (K4), we are unable to determine the limiting distribution of $\psi(\theta_n) - \psi(\theta^*)$, therefore we must define our stopping time for the Kiefer-Wolfowitz algorithm with respect to a confidence interval about the desired operating point $\theta^*$.

**Definition 5.2** The stopping time $N_{KW}$ is defined as the smallest integer $n$ such that

$$Pr [|\theta_n - \theta^*| > \varepsilon] < \delta$$

(5.63)

where $\varepsilon$ and $\delta$ define the confidence interval about $\theta^*$.

Following arguments similar to the ones used in determining the stopping time for Robbins-Monro, it can be shown that

$$N_{KW} = \left[ \frac{2K_s^4 A^4 \text{var}^2 [g(X, \theta^*)]}{M_{MC}(\theta^*) \varepsilon^4 (8\alpha_2 A - 1)^2 C^2} \right]$$

(5.64)

$$N_{KW}^* = \left[ \frac{2K_s^4 A^4 \text{var}^2 [W(X, \theta^*)g(X, \theta^*)]}{M_{IS}(\theta^*) \varepsilon^4 (8\alpha_2 A - 1)^2 C^2} \right]$$

(5.65)

It should be noted that for large $N_{KW}$ and $N_{KW}^*$ we benefit by proportionally increasing the length of the Monte Carlo and Importance Sampling simulations as that the stopping times decrease with the square of these simulation lengths.

Under comparison Scenario 1, it can be shown that the ratio of total computations required in the Monte Carlo implementation of the Kiefer-Wolfowitz algorithm
to the computations required in the Importance Sampling implementation of the same algorithm satisfies the bounds given in (5.47). Furthermore, as was shown for the Robbins-Monro optimization algorithm, we supply the following asymptotic result under comparison Scenario 1 for the case that the performance measure is the probability of a set in the observation space.

**Proposition 8** Let \( g(X, \theta) \) be an indicator function, \( \psi(\theta) \) satisfy (K1) through (K8), \( M_{MC} \) and \( P_1 \) satisfy the conditions of Proposition 6. Moreover, choose the parameter \( \varepsilon \) in (5.63) such that \( \varepsilon^2 \) is \( o(|\alpha|) \). Then under comparison Scenario 1 we have that

\[
\lim_{a \to 0^+} \frac{\mu_{MC}}{\mu_{IS}} = \infty. \tag{5.66}
\]

**Proof** Under the above conditions set forth for choosing \( \varepsilon \), it can be shown that

\[
\lim_{a \to 0^+} N_{KW} = N'_{KW} = \infty. \quad \text{Additionally, } a \to 0^+ \text{ implies that } \psi(\theta^*) \to 0^+, \text{ therefore } 
\lim_{a \to 0^+} \Gamma(\theta_{N_{KW}}) = \infty \text{ leading to the desired result.} \]

Alternately, in investigating the computational reduction under comparison Scenario 2, we can show that the stopping times are related through the following equality:

\[
N_{KW}' = \frac{N_{KW}}{\Gamma^2(\theta^*)}. \tag{5.67}
\]

\(^5\)The condition that \( \varepsilon^2 \) be \( o(|\alpha|) \) implies that for any \( \psi(\theta^*) \) we have \( |\psi(\theta^* + \varepsilon) - \psi(\theta^*)| \leq \xi \psi(\theta^*), \xi > 0. \)
If in addition, we set the length of the performance estimators to be equal to a constant for all levels $\theta$, we can show that the total computations are related by

$$\mu_{IS} = \frac{\mu_{MC}}{\Gamma^2(\theta^*)}. \tag{5.68}$$

Therefore, under this comparison scenario, the stopping time and in the total number of computations are reduce by the square of the gain. This implies that the implementation of Importance Sampling has a greater effect on the Kiefer-Wolfowitz algorithm than on the Robbins-Monro algorithm. The reason for this is that the stopping time is proportional to the square of the estimator variance in the Kiefer-Wolfowitz algorithm.

As demonstrated for the Robbins-Monro optimization algorithm in Proposition 7, we supply the following asymptotic result under comparison Scenario 2 for the situation that the performance measure is again the probability of a set in the observation space.

**Proposition 9** Under the conditions stated in Proposition 8, we have

that under comparison Scenario 2

$$\lim_{\alpha \to 0^+} \frac{\mu_{MC}}{\mu_{IS}} = \infty$$

$$\lim_{\alpha \to 0^+} \frac{N_{KW}}{N_{KW}^*} = \infty$$

**Proof** The proof follows as that $\lim_{\alpha \to 0^+} N_{KW} = \infty$ and $\lim_{\alpha \to 0^+} \Gamma(\theta^*) = \infty$. □
Therefore, as the minimum achievable performance tends to zero, we have that the computational savings is unbounded under both comparison scenarios. Moreover, the reduction in the stopping time is also unbounded under comparison Scenario 2.

As stated before, the improvement in the non-asymptotic case varies from problem to problem. As an illustrative example, we reexamine the signal design problem in Example 1 of this chapter. However, appropriate to the Kiefer-Wolfowitz algorithm, our design criterion is to determine the angle $\theta$ such the the error rate is minimized. As before, conditioned on $H_1$ (signal present), the signal to noise ratio is $26dB$, and let $\varepsilon = \theta^*/10$, $\delta = .05$. While $\theta^*$ can be determined numerically ($\theta^* = 42.17^\circ$), this application will again illuminate the potential in applying efficient estimation schemes to the Kiefer-Wolfowitz algorithm.

Assigning the constant $A$ its optimal value $1/4\alpha_2$ and $C = 10$, it can be shown using the above values that the lower bound on the trade-off between $N_{KW}$ and $M_{MC}(\theta^*)$ is given by

$$N_{KW}M_{MC}^2(\theta^*) \geq 1.1 \times 10^{17}. \quad (5.69)$$

The significant increase over the equivalent lower bound given in (5.53) for the Robbins-Monro algorithm is a consequence of the $O(n^{-1/2})$ convergence rate. However, if we employ the biasing density derived from the conditional linear shift strategy, we reduce the lower-bound to

$$N_{KW}M_{IS}^2(\theta^*) \geq 1.15 \times 10^{10}. \quad (5.70)$$
Using this, we can show that under comparison Scenario 1 we have the \(1.9 \times 10^3 \leq \frac{\mu_{MC}}{\mu_{IS}} \leq 1.0 \times 10^5\) and under Scenario 2 with \(M_{MC}(\theta) = M\) we have \(\frac{\mu_{MC}}{\mu_{IS}} = \frac{N_{KW}}{N_{KW}} = 1.6 \times 10^9\). Therefore, we obtain substantial computational savings under both methods of comparison, however the savings is dramatically increased if we choose to equate the number of estimator trials in both the Monte Carlo and the Importance Sampling implementations. Again, this is because the reduction in the stopping times under Scenario 2 is equivalent to the square of the gain while under comparison Scenario 1, the reduction in the length of the performance estimator is only proportional to \(\Gamma\).

Furthermore, if we employ the least favorable density from the \(\epsilon\)-mixture constraint class \((\epsilon = .5)\), we can further reduce the lower bound in (5.70) to

\[N_{KW}M_{IS}^2(\theta^*) \geq 2.1 \times 10^3\]  

yielding under comparison Scenario 1 that \(2.9 \times 10^6 \leq \frac{\mu_{MC}}{\mu_{IS}} \leq 1.0 \times 10^8\) and under Scenario 2

\[\frac{\mu_{MC}}{\mu_{IS}} = \frac{N_{KW}}{N_{KW}} = 8.2 \times 10^{15}.\]  

We can extend this application to the more involved problem of determining the functional relationship between the minimizing \(\theta^*\) and the signal to noise ratio for the problem in Example 1. The probability of error versus the angle \(\theta\) for a variety of signal to noise ratios is presented in Figure 5.2 for this system.

The theoretical problem of determining the optimal signal location in the presence of additive non-Gaussian noise is still an open issue in the literature. The only known results to date are limited to the domain of small and large signal analysis ([32]...
and references within). In this work, it was shown that in the presence of additive background noise, the error rate was approximately independent of the angle $\theta$ for small signal strengths. If in addition, the noise is i.i.d. with Laplacian marginal density function, then it was shown that the value of $\theta$ which minimizes the error rate tends to $45^\circ$ as the signal strength tends to infinity. However, the rate at which $\theta$ tends to $45^\circ$ is not known. With the use of the Kiefer-Wolfowitz algorithm, we may determine the value of $\theta^*$ for more practical ranges of the signal to noise ratio. These estimated values are plotted against the actual values of $\theta^*$ attained from alternate numerical techniques in Figure 5.3. As can be seen, these results support the claim
that $\theta^* \to 45^\circ$ as the energy gets large. Moreover, we can see that even for moderate values of the signal to noise ratio ($SNR \approx 30$), the minimizing angle is approximately its asymptotic value.

While this example is not comprehensive, it does illuminate the potential for solving some difficult problems in stochastic optimization via the two algorithms discussed in this chapter. Moreover, the benefits attained from utilizing the efficient estimation schemes developed in Chapters 2 through 4 of this thesis seem to overcome the relatively slow convergence properties of the Robbins-Monro and Kiefer-Wolfowitz algorithms, thus making both of these algorithms a viable option for determining system parameters.
Figure 5.3 The angle $\theta^*$ which minimizes the probability of error plotted against the signal to noise ratio for the two dimensional detection system with constrained signal energy in the presence of i.i.d. Laplacian noise.
Chapter 6

Conclusions

We have considered in this thesis the problem of the design and performance evaluation of systems with random inputs. It is well known that in most realistic systems analytic results are not readily obtained, therefore numerical techniques must be utilized. Furthermore, as noted in the Introduction, optimization and performance evaluation are often intimately related in that many optimization algorithms require the value of the system performance at an iterated set of parameters. Unfortunately, one may only have access to the output of the system for a variety of inputs. In this case we are limited to estimating the performance by statistically averaging the cost function over the observations. Clearly, any optimization algorithm which depends on estimates of the performance would benefit from the utilization of variance reduction techniques.

As such, Importance Sampling was employed in this thesis to reduce the effects of the noise inherent in the estimation process. The fundamental problem in using this technique is to determine the appropriate biasing density such that a reduced variance estimate of the system performance is obtained. In Chapter 2, the linear shift class of biasing densities was developed as the biasing strategy for the analysis.
of optimal detection systems. This strategy biases the received vector by adding to it a convex combination of the signal vectors. Therefore, one immediate benefit of this biasing scheme is the relative ease in implementation.

Additionally, it was proved that utilizing this biasing strategy guarantees a reduced variance estimate of the probability of error for a wide class of optimal detection systems. Moreover, in the examples investigated, results demonstrated that the reduction of simulation trials required to obtain the equivalent confidence interval in using standard Monte Carlo simulations is approximately inversely proportional to the true value of the performance. Furthermore, no more than 100 trials were required to estimate error probabilities greater than $10^{-20}$. It was empirically demonstrated that this biasing density is approximately near-optimum when estimating the minimum error rate in the most common detection systems.

This parametric class of biasing densities was extended for analyzing multi-user communication systems. This extension required the development of the conditional linear shift. The additive offset for this biasing scheme is a function of desired signal vectors and the parameters of the interfering users. As in the single-user communication scenario, it was again demonstrated that the gain in implementing Importance Sampling is approximately proportional to the inverse of the error rate for a variety of detector structures and background noise models. Further, it is believed that this biasing strategy will find applications in more general systems such as $M$-ary decision systems and Neural Network classifiers.
It should be noted that prior to this thesis, all known biasing strategies in Importance Sampling arose from the utilization and subsequent optimization of parametric families of density functions. This methodology inherently lead to ad hoc biasing schemes. However, in Chapter 4 we were able to determined sufficient conditions for a density function from an arbitrary constraint class to minimize the variance of the Importance Sampling estimate of the system performance. The crux of the derivation involved showing that minimizing the Importance Sampling variance is equivalent to minimizing an Ali-Silvey distance to the optimal biasing density. From this we were able to utilize results from robust statistics to show that the least favorable density from the constraint set against the optimal biasing density is a constrained minimizer, i.e., no other density function from this set gives rise to an Importance Sampling scheme with less variance. The generality of this theory enables us to determine constrained optimal biasing densities for estimating arbitrary functionals of random vectors. Furthermore, in utilizing the well known \( \varepsilon \)-mixture constraint class, we were able to demonstrate a biasing scheme with associated variance which can be made arbitrarily close to the minimum achievable in Importance Sampling. Moreover, this general framework for determining biasing densities can be utilized to improve all existing biasing methods.

The Robbins-Monro and Kiefer-Wolfowitz algorithms were utilized to determine the desired scalar system parameter in the presence of estimation noise. The effects of this noise on the rate of convergence of the above algorithms was quantified. It
was shown that the stopping time for the Robbins-Monro algorithm is directly proportional to the variance of the performance estimator while in the Kiefer-Wolfowitz algorithm, it was shown that the stopping time was proportional to the square of the variance of the estimator. Clearly, the effective application of Importance Sampling to these design algorithms will have a dramatic effect on the total number of computations. As such, it was shown that in optimizing systems with respect to diminishing probabilities, the computational savings in both algorithms is unbounded. Moreover, depending on the comparison scenario, it was shown that the reduction in stopping times is directly related to the gain of the Importance Sampling implementation evaluated at the desired system parameter, thus leading to significantly shorter optimization runs.

In summary, it is believed that similar results will be obtained in the extension of the these methods to the multi-parameter design problem. Moreover, it is believed that every optimization algorithm which depends solely on estimates derived from the system output can benefit from the utilization of Importance Sampling.
Appendix A

A.1 Extra Conditions for Restricting D to be a Distance Function

Recall that $D(X)$ is defined as

$$D(X) = -\sum_{i=1}^{N} \ln \left[ \frac{p_{\eta}(X_i)}{p_{\eta}(0)} \right], \quad (A.1)$$

where $X$ is an $N$ dimensional vector. If the elements of $X$ are i.i.d. with marginal density $p_{\eta}(X)$, then we can write $D$ as

$$D(X) = -\sum_{i=1}^{N} \ln \left[ \frac{p_{\eta}(X_i)}{p_{\eta}(0)} \right]. \quad (A.2)$$

When $p_{\eta}$ is symmetric, decreasing and unimodal, then it is easily verified that $D(X)$ has the following properties:

a) $D(X) \geq 0 \forall X$

b) $D(X) = D(-X)$ \quad (A.3)

c) $D(X - Y) = 0$ if and only if $X = Y$.

These above conditions restrict $D(X)$ to be a classical distance function.
We can further restrict $D$ to be a metric by requiring that the triangle inequality hold. This imposes the added restriction that

$$- \ln [p_\eta(X + Y)] \leq - \ln [p_\eta(X)] - \ln [p_\eta(Y)], \quad (A.4)$$

which is equivalent to

$$p_\eta(X + Y)p_\eta(0) \geq p_\eta(X)p_\eta(Y) \forall X, Y \in \mathcal{R}^1. \quad (A.5)$$

Therefore $D$ is a metric for i.i.d. noise under the following conditions on the marginal density,

1) $p_\eta(X) = p_\eta(-X)$

2) $p_\eta(X) > p_\eta(Y) \forall |Y| > |X|$

3) $p_\eta(X + Y)p_\eta(0) \geq p_\eta(X)p_\eta(Y) \forall X, Y \in \mathcal{R}^1.$

We must note here that the conditions of $D$ being both a metric and convex are in opposition to one another. These two sufficient conditions only hold when $D$ is linear, implying that they only hold in the case that the marginal density of $\mathcal{R}$ is Laplacian.

A.2 Proof of Proposition 3

For Gaussian noise, it is easily shown that $D(X) = 1/2\langle X, X \rangle_{K^{-1}}$, where $K$ is the covariance matrix and $\langle \cdot, \cdot \rangle_{K^{-1}}$ is the standard inner product with respect to the kernel $K^{-1}$. We must show that

$$D(\alpha X + (1 - \alpha)Y) \leq \alpha D(X)(1 - \alpha)D(Y). \quad (A.6)$$
This is equivalent to showing
\[
\|\alpha X + (1 - \alpha) Y\|_{K^{-1}}^2 \leq \alpha \|X\|_{K^{-1}}^2 + (1 - \alpha) \|Y\|_{K^{-1}}^2.
\] (A.7)

Simplifying by use of the standard properties of norms, we need only show that
\[
\|X\|_{K^{-1}}^2 - 2\langle X, Y \rangle_{K^{-1}} + \|Y\|_{K^{-1}}^2 \geq 0.
\] (A.8)

This inequality will always hold since
\[
\|X\|_{K^{-1}}^2 - 2\langle X, Y \rangle_{K^{-1}} + \|Y\|_{K^{-1}}^2 = \|X - Y\|_{K^{-1}}^2.
\] (A.9)
Appendix B

B.1 Evaluation of $W$ in the Multiple-Access Environment

Recall that the average weight $W$ is given by

$$W = \int_{\mathbb{R}^N} I(|G(r) - b^{(1)}|)W(r)p_B(r)dr.$$  \hfill (B.1)

We can formulate this in terms of the test statistic $Y$ in (3.22) as

$$W = \pi_0 \int_0^\infty \frac{p^2_{y|b^{(1)}}(y|b^{(1)} = -1)}{p_{y*|b^{(1)}}(y|b^{(1)} = -1)} dy + \pi_1 \int_{-\infty}^0 \frac{p^2_{y|b^{(1)}}(y|b^{(1)} = +1)}{p_{y*|b^{(1)}}(y|b^{(1)} = +1)} dy,$$ \hfill (B.2)

where $\pi_i$ are the a priori probabilities of $H_i$ occurring. The above equation can be written as

$$W = \pi_0 E_{\tilde{y}} \left[ \int_0^\infty \frac{p^2_{y|b^{(1)}}(y|b^{(1)} = -1, \tilde{p})}{p_{y*|b^{(1)}}(y|b^{(1)} = -1, \tilde{p})} dy \right] + \pi_1 E_{\tilde{y}} \left[ \int_{-\infty}^0 \frac{p^2_{y|b^{(1)}}(y|b^{(1)} = +1, \tilde{p})}{p_{y*|b^{(1)}}(y|b^{(1)} = +1, \tilde{p})} dy \right].$$ \hfill (B.3)

Using symmetry we have that

$$W = E_{\tilde{y}} \left[ \int_0^\infty \frac{p^2_{y|b^{(1)}}(y|b^{(1)} = -1, \tilde{p})}{p_{y*|b^{(1)}}(y|b^{(1)} = -1, \tilde{p})} dy \right].$$ \hfill (B.4)

In linear detection, the above conditional densities are Gaussian whose means and variances are obtained from (3.22) and (3.23). The solution to the integral in (B.4)
is then given by
\[ \int_0^\infty \frac{p_{Y|b^{(1)}}(y|b^{(1)} = -1, \bar{v})}{p_{Y^*|b^{(1)}}(y|b^{(1)} = -1, \bar{v})} dy = \left( \frac{\exp \left[ \frac{(\theta(a_1, a_2) + \langle L, a_2 \rangle)^2}{\sigma^2 ||a_1||^2} \right] Q \left( \frac{2(\theta(a_1, a_2) + \langle L, a_2 \rangle)}{\sigma ||a_1||} \right)}{\sigma ||a_1||^2} \right), \] (B.5)
(see [49] for explicit derivation), rendering the final result as
\[ W = E_{\bar{X}} \left\{ \exp \left[ \frac{(\theta(a_1, a_2) + \langle L, a_2 \rangle)^2}{\sigma^2 ||a_1||^2} \right] Q \left( \frac{2(\theta(a_1, a_2) + \langle L, a_2 \rangle)}{\sigma ||a_1||} \right) \right\}. \] (B.7)

**B.2 Upper and Lower Bounds on \( \Gamma \) in the Multiple-Access Environment**

We have from (3.35) that under linear detection and Gaussian background noise
\[ \Gamma \approx \frac{E_{\bar{X}}[Q(X)]}{E_{\bar{X}}[\exp(X^2)Q(2X)]}, \] (B.8)
where \( X = \frac{\theta(a_1, a_2) + \langle L, a_2 \rangle}{\sigma ||a_1||} \). For small error rates, we can assume this holds with equality. Therefore we have that
\[ E_{\bar{X}} \left[ \Gamma \exp(X^2)Q(2X) - Q(X) \right] = 0. \] (B.9)

From this we can infer that either \( \Gamma = \frac{Q(X)}{\exp(X^2)Q(2X)} \) for every \( X \), or that there exists Borel sets \( A \) and \( B \) with nonzero measure such that \( \Gamma \exp(X^2)Q(2X) - Q(X) \) is non-negative for every \( X \in A \) and \( \Gamma \exp(X^2)Q(2X) - Q(X) \) is non-positive for every \( X \in B \).
We can eliminate the first possibility as that $\Gamma$ is a constant. Examining the second possibility we have that

$$\Gamma \geq \frac{Q(X)}{\exp(X^2)Q(2X)} \quad \forall X \in A$$

(B.10)

with the corresponding upper bound on $\Gamma$ for all $X$ in $B$. As that we are dealing with small error rates, we can employ the well known approximation that $Q(X) \approx \frac{1}{X\sqrt{2\pi}} \exp\left[-\frac{X^2}{2}\right]$ to show that

$$\frac{Q(X)}{\exp(X^2)Q(2X)} = 2\exp\left[\frac{X^2}{2}\right].$$

(B.11)

We then directly obtain the following bounds on the gain

$$2\exp\left[\frac{\min \frac{X^2}{2}}{2}\right] \leq \Gamma \leq 2\exp\left[\frac{\max \frac{X^2}{2}}{2}\right].$$

(B.12)

From this it is evident that the gain is exponential. Again, from the above approximation for the $Q$ function, we also have that

$$2\exp\left[\frac{X^2}{2}\right] \approx \sqrt{\frac{2}{\pi}} \frac{1}{XQ(X)}.$$

(B.13)

rendering the upper and lower bounds on $\Gamma$ as follows:

$$\sqrt{\frac{2}{\pi}} \min_{\frac{1}{X}} \left\{ \frac{1}{X} \right\} \frac{1}{\max_{\frac{1}{P_x|z}}} \leq \Gamma \leq \sqrt{\frac{2}{\pi}} \max_{\frac{1}{X}} \left\{ \frac{1}{X} \right\} \frac{1}{\min_{\frac{1}{P_x|z}}}. $$

(B.14)
Bibliography


