EFFICIENT YIELD ESTIMATION USING RARE EVENT SIMULATION TECHNIQUES ON ANALOG DESIGN AUTOMATION TOOLS

by

Alphan Çamlı

B.S., Electronics Engineering, Istanbul Technical University, 2013

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of the
requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering
Bogazici University

2016

EFFICIENT YIELD ESTIMATION USING RARE EVENT SIMULATION TECHNIQUES ON ANALOG DESIGN AUTOMATION TOOLS

APPROVED BY:

Prof. Günhan Dündar ..... (Thesis Supervisor)

Assoc. Prof. Ali Emre Pusane .....

Assist. Prof. Berke Yelten .....

DATE OF APPROVAL: 23.08.2016

ACKNOWLEDGEMENTS

I would like to express my profound sense of gratitude to my thesis advisor Professor Günhan Dündar, for his continuous support, patience, encouragement, motivation, and admirable knowledge. His guidance encouraged me in all the time of research during my master thesis. Also, I would like to express my profound gratitude to Assoc. Prof. Ali Emre Pusane, Assist. Prof. Mustafa Berke Yelten, Assist. Prof. İsmail Faik Başkaya, and Engin Afacan for their constant support and deep expertise throught my research work.

I also would like to thank my family and friends for their constant love and support through out all my undergraduate and graduate education.

Finally, I would like to express my appreciation to TÜBİTAK and their scholarship programme named “2210-A Genel Yurt İçi Yüksek Lisans Burs Programı”. With their continous financial aid I was able to complete my master thesis in the first place.

ABSTRACT

EFFICIENT YIELD ESTIMATION USING RARE EVENT SIMULATION TECHNIQUES ON ANALOG DESIGN AUTOMATION TOOLS

With the improvements in fabrication processes, electronic circuit designers have begun to design complex circuits which consist of multibillion transistors. But, as circuit complexity increases, the silicon complexity also increases, leading to process variations having a profound effect on the circuit performance especially in sub micron technologies. Therefore, even if a circuit was designed to achieve a certain design specification, there will be a discrepancy between the simulated and the measured performances. This difference can lead to a decrease in the yield. Circuit designers tend to handle this problem by leaving a safety margin; however, this leads to overdesign and loss of precious chip area. Therefore, there is an undeniable need to have efficient design automation tools for reducing design time without compromising performance. Normally, a typical approach for analyzing a circuit would be running a Monte Carlo simulation with a small sample size and then fitting a standard analytical distribution to the data. Such an approach can be accurate for the main part of the distribution, however it will be heavily inaccurate in the tail of the distribution. Since, the distribution of design specifications with respect to process variation effects tends to have a long tail by nature, a classic Monte Carlo simulation can not be used. In this case, a rare event sampling method can be utilized for increasing number of samples corresponding to tail of the original distribution. Cross entropy minimization based importance sampling (IS) method is chosen as rare event sampling method for the scope of this thesis due to its efficiency, although there are lots of different Monte Carlo based proposals. Also, a hybrid Quasi-Monte Carlo (QMC) method has been utilized in order to both select rare event threshold that is needed for cross entropy based IS algorithm and performance comparison with the proposed algorithm.

ÖZET

ANALOG TASARIM OTOMASYONUNDA ETKİN VERİM HESABI İÇİN NADİR OLAY BENZETİM TEKNİKLERİ

Gelişen fabrikasyon prosesleri ile birlikte elektronik devre tasarımcıları günümüzde milyarlarca transistör içeren karmaşık devreler tasarlayabilmektedir. Ancak devrelerin karmaşıklığının artmasıyla birlikte silikon karmaşıklığı da artmakta, dolayısıyla proses varyasyonlarının devre üzerindeki etkisi gittikçe artmaktadır. Özellikle mikron-altı proseslerde, proses varyasyonlarının etkisi oldukça hissedilmektedir. Bu durum tasarlanan devrenin simülasyon performansı ile üretimden sonra ölçülen performansının farklı olmasına neden olmaktadır. Bu nedenle üretilen bazı devreler istenen spesifikasyon aralığının dışına çıkabilmekte, verim bu sebeple düşebilmektedir. Tasarımcılar genellikle istenen performans kriterleri için belli oranlarda pay bırakarak bu sorunlardan kaçınmaya çalışmaktadır. Ancak bu da aşırı tasarım ve de günümüzde artık çok değerli olan çip boyutunun artmasına neden olmaktadır. Dolayısıyla performanstan ödün vermeden tasarım sürecini kısaltacak tasarım otomasyon araçlarına büyük bir ihtiyaç vardır. Normalde, tasarıma dayalı verim hesaplaması için analiz yaparken klasik Monte Carlo yöntemi kullanılabilir. Ancak bu analiz, olasılık dağılım grafiğinin sadece gövde kısmı için gerçeğe yakın olacaktır, kuyruk kısmı varsa elde edilen sonuç gerçek değerden oldukça uzak olacaktır. Tasarım spesifikasyonlarının proses varyasyonlarına bağlı değişimi doğal olarak uzun bir kuyruğa sahip olduğundan simülasyon için klasik Monte Carlo yöntemini kullanmamız mümkün değildir. Bu durumda yapılacak olan grafiğin kuyruk kısmındaki nadir örneklerin sayısını arttırmaktır. Literatürde bu amaçla nadir benzetim teknikleri kullanılmaktadır. Tezin kapsamında çapraz entropiye dayalı önem örneklemesi etkinliğinden ötürü önerilmiştir. Bunun yanısıra hem nadir örnek eşiğini tespit edebilmek hem de önerdiğimiz algoritma ile kıyaslanması amacıyla hibrit QMC methodu da uygulanmıştır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZET v
LIST OF FIGURES viii
LIST OF TABLES xi
LIST OF ACRONYMS/ABBREVIATIONS xii
1. INTRODUCTION 1
1.1. Motivation 1
1.2. Overview of Analog IC Design Flow. 4
1.3. Overview of Rare Event Simulation. 6
1.4. Main Contributions and Outline 7
2. BACKGROUND 9
3. YIELD-AWARE CIRCUIT SYNTHESIS WITH A HYBRID QMC TECHNIQUE. 12
3.1. Background 13
3.2. Efficient Yield Estimation Techniques 14
3.3. Hybrid QMC Technique. 16
3.3.1. Yield Estimation Method. 17
3.3.2. Algorithm Implementation. 19
3.3.3. Simulation Results. 20
4. EFFICIENT YIELD ESTIMATION AND ENHANCEMENT USING RARE EVENT SAMPLING METHODS 24
4.1. Background. 24
4.2. Importance Sampling. 26
4.2.1. Mathematical Approach. 27
4.2.2. Algorithm Implementation. 29
4.2.3. Simulation Results. 30
4.3. Cross Entropy Minimization 36

4.3.1. Background. 36
4.3.2. Algorithm Details. 37
5. MULTI LEVEL CROSS ENTROPY MINIMIZATION BASED IMPORTANCE SAMPLING METHOD 41
5.1. Background 42
5.2. Algorithm Implementation 45
5.3. Simulation Results 48
6. YIELD CALCULATION 50
6.1. Rare Event Modelling 50
6.2. Yield Estimation Results 51
7. CONCLUSION 56
REFERENCES 58

LIST OF FIGURES

Figure 1.1. Circuit performance change after variation. 3

Figure 1.2. VLSI IC design flow diagram (from [7]). 5

Figure 3.1. Optimum point moves towards to infeasible region after variation. 11

Figure 3.2. Uniform scatter of N=500 process variation samples (W1 & L1) with QMC technique 16

Figure 3.3. Flow diagram of the optimizer. 19

Figure 3.4. Schematic of BTS OpAmp. 21

Figure 3.5. Schematic of FC Amplifier. 21

Figure 3.6. Bandwidth distribution for BTS OpAmp (QMC with 500 samples) 22

Figure 3.7. Gain distribution for BTS OpAmp (QMC with 500 samples) 22

Figure 3.8. Phase margin distribution for BTS OpAmp (QMC with 500 samples) . . . 23

Figure 4.1. Input normal distribution (mean=5, sigma=1) for N=10000 samples 32

Figure 4.2. Output distribution $f_1(x) = x^2 + \log(x)$ for given input distribution 32

Figure 4.3. Shifted input normal distribution (mean=8, sigma=1) for N=10000 samples. 33

Figure 4.4. Shifted output distribution $f_1(x) = x^2 + \log(x)$ for shifted input distribution. 33

Figure 4.5. Input normal distribution (mean=5, sigma=1) for N=10000 samples 34

Figure 4.6. Output distribution $f_2(x) = x^2 - e^{-x} + \log(x)$ for given input distribution. . . 34

Figure 4.7. Shifted input normal distribution (mean=8, sigma=1) for N=10000 samples.. 35

Figure 4.8. Shifted output distribution $f_2(x) = x^2 - e^{-x} + \log(x)$ for shifted input distribution 35

Figure 4.9. Main Cross Entropy Minimization Algorithm.. 40

Figure 5.1. Minimization for Inital Vector Selection Algorithm.. 46

Figure 5.2. Cross Entropy Minimization Based IS Algorithm.. 47

Figure 5.3. Bandwidth distribution after QMC with 500 samples.. 48

Figure 5.4. Bandwidth distribution after cross entropy based IS with 1500 samples 49

Figure 6.1. Yield Estimation Algorithm. 51

Figure 6.2. Conventional MC with 100k samples for bandwidth in BTS Op-Amp. 53

Figure 6.3. QMC with 1k samples for bandwidth in BTS OpAmp. 53

Figure 6.4. MCE based IS with 2k samples for bandwidth in BTS OpAmp (Rare event threshold is chosen as 8.15 kHz). 54

Figure 6.5. MCE based IS with 2k samples for bandwidth in BTS OpAmp (Rare event threshold is chosen as 8.2 kHz). 54

Figure 6.6. MCE based IS with 2k samples for bandwidth in BTS OpAmp (Rare event threshold is chosen as 8.3 kHz). 55

LIST OF TABLES

Table 6.1. Simulation results for T=8150 Hz for BTS Opamp. 52

Table 6.2. Simulation results for T=8200 Hz for BTS Opamp. 52

Table 6.3. Simulation results for T=8300 Hz for BTS Opamp. 52

LIST OF ACRONYMS/ABBREVIATIONS

BTS	Basic Two Stage
CAD	Computer Aided Design
CalTech	California Technology
FC	Folded Cascode
HDL	Hardware Description Language
IC	Integrated Circuit
IEEE	Institute of Electrical and Electronics Engineers
IP	Intellectual Property
IS	Importance Sampling
LDS	Low Discrepancy Sequences
LHS	Latin Hypercube Sampling
MATLAB	Matrix Laboratory
MC	Monte Carlo
MCE	Cross Entropy Minimization
OpAmp	Operational Amplifier
QMC	Quasi Monte Carlo
SoC	System on Chip
SPICE	Simulation Program Integrated Circuit Emphasis

1. INTRODUCTION

1.1. Motivation

The invention of the transistor in 1947 can be considered as the beginnning of a new era which is called the information age. In this age, smart electronic devices have become indispensable for our every day life including communication, health, energy, security, entertainment, and education. And all of these inventions are made possible with the microelectronic revolution. In the early days of electronics before 1950s, vacuum tubes also called electron tubes, were the main electronic devices that were used in just a few areas such as radio and television. Limitations were due to the nature of vacuum tubes being quite big and unreliable devices. They were also problematic due to heat dissipation problems. Therefore, by the end of the 1950s, the transistor replaced the hot, unreliable electron tube in nearly every existing type of electronic system. It also made electronic devices smaller, cooler in terms of heat, and less expensive. Later, the invention of the transistor led to research in the integrated circuit field. After the invention of the first integrated circuit in 1958, microelectronic industry has made a huge leap impacting the society in every way possible. With the realization of first IC, we have crammed more and more electronic components into a single chip year by year.

The need for high performance and cost effective electronic products led electronic engineers towards system on chip (SoC). With the improvements in the fabrication processes which enabled decreasing feature sizes, currently multi-billion transistors can be combined into a single chip [1]. In early 1960s, only 30 transistors could be integrated into the single chip for primitive ICs. In his notable article which was published in 35th anniversary edition of “Electronics Magazine”, Gordon Moore correctly predicted the trend in components per integrated circuit [2]. He stated that the number of components per integrated circuit would be doubled every year. His prediction was amazingly accurate so that one of his friends, Dr. Carver Mead who was a professor at Cal Tech dubbed this

as Moore's Law. So, the original Moore's Law was doubling every year in complexity hence in computing power. As Moore stated at 1975 IEEE International Electronic Devices Meeting, advances in photolithography, wafer size, process technology, circuit and device cleverness allowed to his prediction to be realized. However, in 1975, he revisited his prediction and made a correction. He slowed the future rate of increase in complexity. According to his second prediction, the total number of transistors put on a chip will be doubled every 2 years [2]. Moore's second prediction is still accurate today. However, due to some limitations especially physical limitations, Moore's Law threatens to come to a halt unless a new integration technology isn't found [3].

Ever since, increasing the number of transistors per IC has become the motivating force for electronic engineers. However, developments in the scaling process and cramming many circuits into a single IC has a cost. As stated in [4], problems may be categorized in two subgroups. The first problem is due to silicon complexity, which refers to the effects of process variations on the circuit performance. Normally, it is widely assumed that process parameters are similar for all devices on the same wafer, but it may vary from wafer to wafer. In deep sub-micron technologies, the amount of process variation becomes particularly pronounced and process tolerances worsen along with transistor dimensions [5]. This is even more critical in analog circuits, because process variation effects lead to mismatches due to local changes on the same chip. Analog circuits heavily depend on the close matching of a set of devices, and variations will degrade the performance of the circuit. Therefore, even if a circuit was designed to achieve a certain set of design specifications, differences between the simulated and measured performances most likely to occur in a population of fabricated ICs due to process variations and mismatches [6], as shown in the Figure 1.1. If the variance causes the measured or simulated performance of a particular output metric such as bandwidth, gain, phase margin, rise time, etc. to fall below or rise above the certain set of specifications for the particular circuit or device, it reduces the overall yield for that set of devices. Furthermore, because circuit specifications are correlated with each other, adjusting one parameter may

affect overall performance which requires simultaneous optimization between all specifications. Hence, manual circuit sizing is an unfeasible and time exhaustive process for a human. In order to shorten the design time, analog design automation becomes a vital and significant alternative. The second problem is due to increased system complexity, which can be explained such that exponentially increased number of transistor counts leads to increased functionality and complexity. Hence, floorplan and power management of the ICs together with various trade-offs between circuit performances alongside the shorter time to market demand become critical and immense concerns.

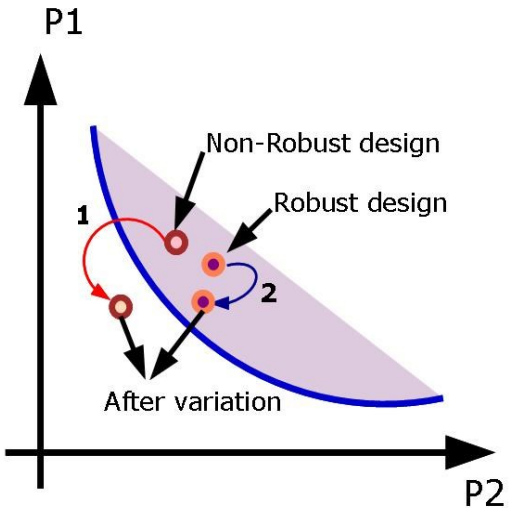

Figure 1.1. Circuit performance change after variation.

An efficient way for dealing with design challenges without decreasing productivity of designers is to benefit from computer aided design (CAD) tools [1]. CAD tools can provide assistance during analysis and verification of the system. Typically, from transistor level to system level, the designer benefits from CAD tools in order to determine the performance of the design. In the literature, efficient CAD tools are available for digital circuits. Digital systems are more suitable for design automation contrary to analog systems; hence, digital systems can be defined using Boolean algebra unlike analog systems. Today's advanced CAD tools are capable of synthesizing a transistor level design

that was described in a so called hardware description language (HDL) by either using Verilog or VHDL [1]. From the point of analog systems, developing a CAD tool is more challenging and dauntling task since analog circuits cannot be represented as digital circuits and have more complex trade-offs between circuit performances and physical parameters. In addition, taking into account effects resulted by device scaling, most of the design time of analog designers is spent by fine-tuning the system by utilizing a simulator through trial and error. Since trial and error is a brute force technique, it consumes a lot of design time.

As already mentioned, manual device sizing is an unfeasible and time exhaustive process for a designer. If certain optimization algorithms can be combined with circuit simulators, the overall time spent for the design could be substantially reduced. Also, such a synthesis tool would enable fine-tuning of analog circuits simultaneously reducing the design time. Therefore, the main focus of this thesis is to develop a yield aware analog circuit synthesis tool with extended rare event simuation capabilities that addresses these problems.

1.2. Overview of Analog IC Design Flow

In this section, analog IC design methodologies will be briefly introduced. As explained in the previous section, increased complexity of analog ICs results in growing design productivity gap for SoCs considering the shorter time to market constraint [4]. In order to improve the design process, some design methodologies are proposed for the circuit designers. As stated in [7], design methodologies can be divided into two groups. The first one is the top-down design methodology, the second is the bottom-up design methodology. The flow diagram as shown in the Figure 1.2 is same for both methodologies, the only difference is the direction of the flow as their names suggest.

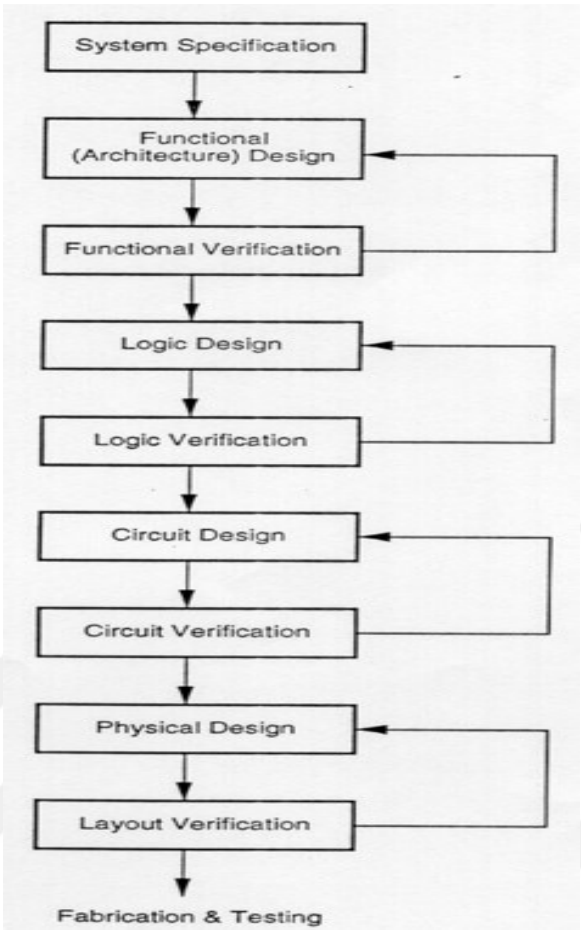

Figure 1.2. VLSI IC design flow diagram (from [7]).

In the top-down design methodology, the flow starts with system design, in which overall system specifications are provided. This step can be seen as general overview of the whole system. General blocks having dedicated tasks are designed and partitioned into sub-blocks. Typically, mathematical simulation tools such as MATLAB/SIMULINK are preferred for the system level design. The next step is called architecture or functional design where digital and analog blocks are separated and requirements of functional blocks are defined. The following step is called topology selection in which topologies for functional blocks and sub-blocks are determined. For example, if an operational amplifier (OpAmp) is required, the designer has to determine to use either a basic two-stage or a folded cascode topology. In the cell design, specific blocks are designed at transistor level and sizing is performed in order to achieve pre-defined performances in a certain techno-

logy. Finally, layouts of cells and general blocks are drawn and post-layout simulations are performed to validate circuit and system specifications. This methodology is advantageous because systematic design is suitable for capturing and fixing problems, since it allows interaction of blocks during the design process.

The other design methodology starts with the designer using previously designed cells [8]. However, using the library of analog cells may be inefficient considering the technology dependency and variety of analog circuits. However, if some form of soft intellectual property (IP) is used, design knowledge and optimization techniques could be embedded such that technology dependency is removed and a wide range of performance choices is provided for designers.

1.3. Overview of Rare Event Simulation

Rare event simulation or rare event sampling is a coin term for a group of computer simulation methods intended to selectively sample special regions of the dynamic space of systems which are very unlikely to be visited by using brute force simulation techniques [9].. A rare event is an event whose occurence is rare with probability less than 10^{-3}. However, typical probabilities of interest are between 10^{-8} and 10^{-10}. Although it seems like these probabilities are incredibly small, rare events occur when dealing with performance evaluation in many different areas such as telecommunication networks, dependability analysis, air control systems, particle transport, biology, insurance, finance. In most of the rare event problems, the mathematical model is too complicated to be solved by either analytical or numerical methods because the assumptions are not stringent enough, the mathematical dimension of the problem is large, or the state space is too large to get a result in a reasonable time [10].

The main idea behind the thesis is to develop accurate yield estimation method for a yield aware analog circuit synthesis tool. So, the performance evaluation metric is yield in

our case. In order to guarantee a certain yield for the design, some variability analysis to estimate the yield is required. With a crude Monte Carlo method such as Quasi Monte Carlo method, we are able to achieve relatively accurate yield estimates of up to 99%. However, this yield may not be sufficient for manufacturing, where 6-sigma is required. Hence, CAD tool must be extended to focus on infeasible region where rare events occur after variability analysis. This is achievable by utilizing rare event sampling techniques as variance reduction method. In the scope of the thesis, a hybrid rare event sampling algorithm is proposed for this context.

1.4. Main Contributions and Outline

This thesis presents the following key features and contributions:

- Yield estimation techniques are developed.
- In order to achieve higher yields up to 99.6%, different rare event sampling techniques are implemented and analyzed with synthetic data.
- After a brief analysis, a commonly known hybrid rare event simulation technique is chosen to be used in our CAD tool. The technique is a hybrid of Importance Sampling and Cross Entropy Minimization techniques, and is used frequently in statistical analysis.
- The algorithm used for yield estimation is proposed as a novel design technique.

The organization of thesis as follows: Chapter 2 presents the background of the thesis by emphasizing key concepts. Chapter 3 gives details and implementation of yield aware design CAD tool by applying Quasi Monte Carlo (QMC) technique which is required for a yield estimation. Chapter 4 explains one of the most commonly used rare event simulation technique known as Importance Sampling (IS) and discusses why this technique is not suitable for our analog circuit synthesis tool. Also another commonly used technique dubbed as Cross Entropy Minimization (MCE) is borrowed from information theory in order to improve efficiency of IS. Chapter 5 gives the details of the our proposed

algorithm called Cross Entropy Minimization Based Importance Sampling in multiple levels with integrated SPICE circuit simulator. Chapter 6 gives details about yield estimation. The final chapter states the conclusion and results.

2. BACKGROUND

Analog circuit synthesis consists of 3 main phases in analog design processes. These phases are correctly biasing the whole circuit in terms of node voltages and electrical currents flowing in circuit branches, calculating the passive element components, and transistor sizing. Analog design as already mentioned is a very hard and time consuming process. Therefore, in recent years, analog design automation has attracted researchers' attention and become a very hot search topic. Because novel methodologies and CAD tools are needed for shortening analog design time. In the literature, analog design automation algorithms can be classified into three categories. These categories are mainly knowledge based approaches, equation based approaches, and simulation based approaches.

Knowledge based algorithms have been developed at first and depend upon the designer solely, because optimization depends on the designer's knowledge, experience, and know-how. In knowledge based algorithms, design strategies for a given circuit topology are utilized during circuit synthesis. Design plans consisting of design equations and design strategies reduce the computation time required for obtaining solutions. Since knowledge based algorithms are based on predefined design plans, the result of these knowledge based algorithms may be inaccurate. Furthermore, preparing design plans for each topology requires excessive human effort and therefore is not feasible [1]. These knowledge based algorithms can not have place in commercial CAD tools considering the disadvantages of human interaction in the optimization process and creation time of design plans for various circuit topologies. In the literature, there are some known knowledge based analog synthesis computer programs: OASYS [11], IDAC [12], and BLADES [13].

Equation based algorithms are quite fast compared to knowledge based algorihtms due to using analytical equations for circuit evaluation. Therefore, if these analytical equations get complex, equation based algorithms lose their efficiency. The optimal solution can be obtained by solving equations using mathematical equation solving tools with poly-

nomial models of circuit properties and transistor parameters. Polynomial models have been already developed for CMOS OpAmps [14, 15], multi-stage amplifiers [16] and oscillators [17]. But large prediction errors often occur in some transistor parameters because of the short channel effect of CMOS technologies. In order to eliminate prediction errors, this approach requires several iterations. In addition to this, time required for model development and topology dependence limit the usefulness of this approach. Therefore, although equation based algorithms provide fast convergence rate and flexibility for carrying out various search algorithms, design equations still have to be derived by hand which means human interaction is still needed for optimization. Another disadvantage is the loss of accuracy since it is not easy to derive all design equations without making simplifications since the equations are most often complicated. In the literature, commonly known equation based CAD tools are OPASYN [18], OPTIMAN [19], and AMGIE [20].

Simulation based algorithms are based on a circuit simulator which evaluates the circuit performances. There are lots of commercially available simulation based tools. In the context of the thesis, HSPICE will be used for accurate circuit simulation. With the usage of simulation based algorithms, human interaction during circuit synthesis is eliminated. In addition, design automation tools overcome the loss of accuracy and become comparable with manual designs. Also, topology and technology dependency is no longer valid since circuit element values, types, and input parameters can be easily manipulated at the input of SPICE. For example, SPICE netlist includes all transistor dimensions, types, passive component values which is required for simulation. Therefore, it is easy to manipulate device dimensions; hence, technology dependency is minimized. The downside is that total synthesis time is increased, because optimization algorithms require excessive number of long simulations to find the optimal solution for the given circuit. In the literature, Anaconda [22] and FRIDGE [23] can be given as examples of simulation based computer programs.

The constant downscaling of the technology has led process tolerances to worsen along with transistor dimensions [5]. Process variations are natural occuring variations due

to gate oxide thickness, random dopant fluctuations, and device geometry. Therefore, it is a challenging problem to cope with process variations which cause worsened reliability issues in CMOS circuits during the fabrication process. The circuit designer must take into account these variations because they cause a difference between simulated and measured performance. If the measured performance of a particular output metric such as bandwidth, gain, phase margin, rise time, etc. goes out of the range of a certain set of spesifications for the particular circuit, it reduces the overall yield of that circuit. Therefore, in order to prevent this discrepancy, some additional steps should be included in the design procedure. The aim here is to achieve a certain performance after fabrication which corresponds to the yield in our case. Thus, in order to guarantee a certain yield for the design, some variability analyses to estimate the yield are needed. In the previous research, hybrid Quasi Monte Carlo technique is proposed as yield estimation technique for the CAD tool [24]. In the scope of the thesis, hybrid Quasi Monte Carlo sampling will be also used for the CAD tool.

In the previous research [24], QMC is utilized for the accurate yield estimation. Yield estimates of up to 99% is achieved by using QMC. As mentioned before, this yield may not be sufficient for manufacturing, where 6-sigma is required. Therefore, obtained yield is seen to be unfit for the manufacturing purposes. In order to enhance the yield estimation technique, rare event sampling methods are researched in the scope of the thesis. The most common rare event sampling techniques are importance sampling (IS) and cross entropy minimization based techniques. Though importance sampling can be used alongside QMC for enhancing yield estimation, it has some disadvantages. In order to find the optimal solution, IS could be strengthened with cross entropy minimization. In fact, this hybrid technique is applied to other areas such as estimating probability of failure rate of SRAM cells [25]. In the context of thesis, cross entropy minimization based IS is applied for a more enhanced and accurate yield estimation.

3. YIELD-AWARE CIRCUIT SYNTHESIS WITH A HYBRID HYBRID QMC TECHNIQUE

The continous downscaling of device geometries resulted in downgrading of process tolerances along with the device sizes. Especially in smaller process nodes, process variations significantly affect the performance of the manufactured devices as the variation becomes a larger percentage of the full length or width of the device. Therefore, it is a challenging problem to deal with process variations during fabrication process, which reduces the overall yield and hence reliability of CMOS circuits. For example, it becomes a headache to handle variations in the fabrication steps, such as line-edge roughness that is induced by gate etching and the lithography process [26], oxide thickness fluctuations that cause the fluctuation of the voltage drop across the oxide layer, affecting V_{th} , and random dopant fluctuations that significantly alter V_{th} [27]. Thus, circuit design without considering variation, leads to discrepancy between the simulated and the measured performance as shown in Figure 3.1. To prevent this discrepancy, some additional steps should be included in the conventional design procedure and analog design automation algorithms to achieve a certain performance after fabrication process [28]. As a result, in order to guarantee a certain yield for the design, some variability analysis to estimate the yield is required.

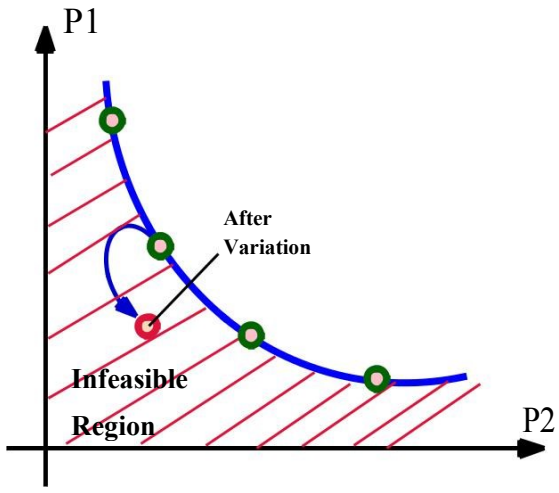

Figure 3.1. Optimum point moves towards to infeasible region after variation.

3.1. Background

In the literature, there are different methods for variability analysis including sensitivity analysis, corner analysis, regression based models, and Monte Carlo (MC) based analysis [28]. Among them, Monte Carlo analysis is the most popular method to estimate the yield of a design, because it is simply based on simulating randomly selected points in the uncertain parameter space and observing the variation effects on the output. MC based approaches generally utilize variance reduction techniques because variance reduction techniques increase the precision of the estimates that can be obtained for a given number of iterations. Therefore, effects of variations to the yield can be found most accurately by using MC based approaches. However, the disadvantage of using classical MC approach is that it requires a very large number of simulations to provide a certain accuracy. Hence, the total simulation time and computational effort is significantly high for conventional MC approaches. Therefore, conventional MC method is not suitable for a yield-aware circuit synthesis CAD tool due to inefficiency.

In order to reduce the computational effort and simulation time, several speed-up techniques have been proposed in the literature. The main idea behind these techniques is minimizing the number of samples by using either variance reduction techniques such as Importance Sampling (IS) or utilizing some mechanics for the use case. For example, Quasi Monte Carlo (QMC) technique that utilizes Low Discrepancy Sequences (LDS) can be used instead of classical MC technique which is based on sequences of pseudo-random numbers. Among these techniques, QMC has been the most efficient approach in terms of computational effort or CPU memory. The main advantage of QMC is scattering the samples on the space homogeneously rather than randomly. Other important advantage of QMC is that it exhibits itself for applications that require iterative sampling, such as yield-aware optimization. Since QMC is a deterministic approach, the sample size can be increased iteratively by pre-determined sample steps. This feature is highly crucial during the optimization process to enhance the efficiency. Using the QMC approach for yield estimation, which promises adaptive sample size determination and automated stopping

criterion mechanism, results in keeping the sample size to a minimum to avoid the redundant simulations.

3.2. Efficient Yield Estimation Techniques

There is a trade-off between the yield estimation accuracy and computational cost. Computational cost includes both total simulation time and required resources such as CPU memory. For a more accurate yield estimation, computational effort increases drastically. However, MC based techniques are reliable and independent from the problem dimensionality, which makes them popular for yield estimation. The efficiency problem can also be handled by introducing some speed enhancement techniques. Classical MC approach is based on random sampling of the uncertain parameter space [29]. However, random sampling can cause sample clusters and empty spaces in the MC distribution over the sampling region and therefore requires a large number of samples for spreading out in the space. In the optimization process, there are many candidate individuals, for which yield analysis will be carried out, hence the total synthesis time would increase drastically. If, somehow, samples were spread out in the space more uniformly, then the number of samples required for simulation would decrease dramatically. To reduce the synthesis time, one solution can be Infeasible Solution Elimination method (ISE), which is based on performing yield analysis only for the candidates that satisfy the user defined specifications [8]. On the yield estimation side, the efficiency of MC based techniques can be enhanced by changing sequences of pseudo-random numbers to LDS. The main idea behind such approaches is to spread out the samples as homogeneously as possible to cover the whole design space with a minimum number of samples.

Classical MC approach has an estimation error rate of $O(n^{-0.5})$ [29]. This error can be seperated into the factor related to the function itself and the factor related to generated set of random points according to the Koksma-Hlawka theorem [30], where the error is given as in the following equation.

$$|\hat{y} \;\; y| \leq D_n^*(x_1, x_2, x_3, \ldots, x_n) V_{HK}(f) \tag{3.1}$$

In Equation 3.1, $\hat{y}$ and y are the estimated and real values of the yield, respectively, and D^* is Star Discrepancy which is used for measuring the uniformity of the generated points, where uniform distributions provide a smaller D^*. V_{HK} (f) is the total variance of the underlying integrand in the yield formula. As can be seen from the formula, the estimation error can be decreased via two methods: increasing the uniformity of samples and decreasing the variance of the function f.

Although there are different types of variance reduction techniques, Quasi-MC technique is the most suitable one for our yield-aware CAD tool. For example, in one of the previous research [24], another common variance reduction method called Latin Hypercube Sampling (LHS) was applied to the yield-aware CAD tool. LHS method based on stratification [19], in which the term V_{HK} (f) is reduced. The variance for one dimensional projections is highly reduced via LHS sampling, as in [19]. However, for higher order projections, the behaviour is similar to conventional MC. QMC is based on lowering D^* via uniformly generated samples as showed in Figure 3.2., which provides enhanced efficiency, therefore it is a better fit for the CAD tool. According to the Koksma-Hlawka theorem, homogeneous sample sets correspond to having lower discrepancy, which reduce MC estimation errors. The discrepancy of the conventional MC for n samples is given in [30] as

$$D_n^*{}_{|MC} = O(n^{-0.5}\log(\log(n))^{-0.5}) \tag{3.2}$$

where the estimation error of the conventional MC is $O(n^{-0.5})$. On the other hand, considering QMC, the discrepancy is given as

$$D_n^*{}_{|QMC} = O(n^{-0.5}(\log^8(n)) \tag{3.3}$$

Low Discrepancy sequences provide an asymptotic integration error rate, which is much faster than the error rate of conventional MC method.

Several LDS strategies for generating sequences have been proposed such as Halton, Sobol, and Faure sequences [31]. All of these LDS based strategies are deterministic, contrary to the random sampling performed in the conventional MC and LHS approaches. This deterministic behaviour becomes a superiority when an iterative variability analysis is required. Variability analysis is carried out many times during the optimization process and this results in longer synthesis times. Using constant sample sizes during yield estimation can still be problematic: keeping the size too small may lead to non-accurate estimations, whereas oversampling may cause inefficiency in terms of total simulation time.

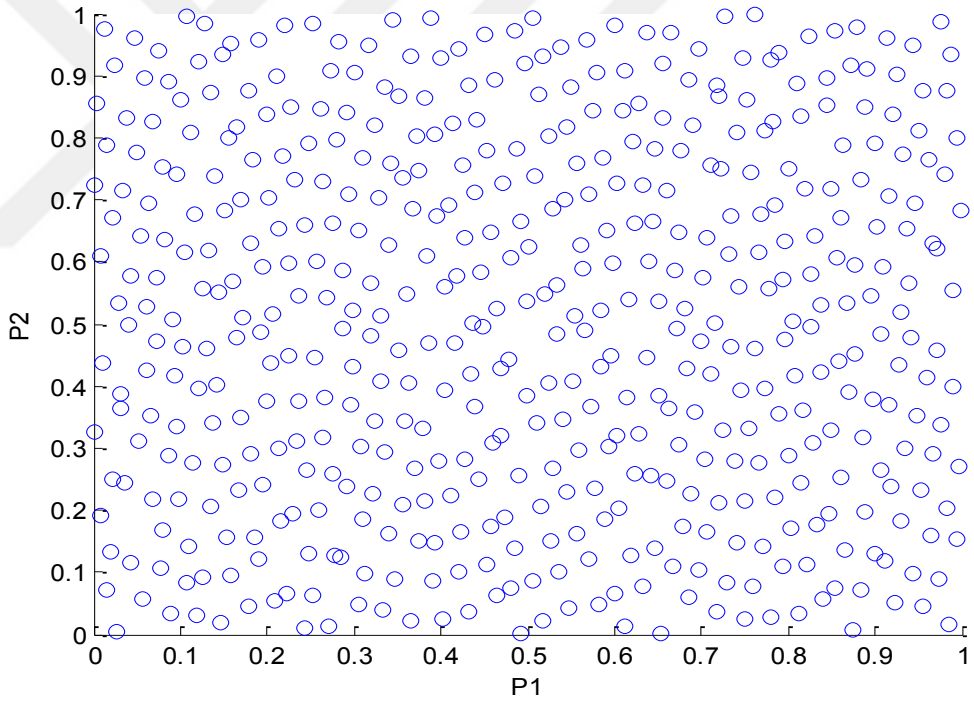

Figure 3.2. Uniform scatter of N=500 process variation samples (W1 & L1) with QMC technique.

3.3. Hybrid QMC Technique

As seen in the previous section, QMC benefits to reduce computational effort for yieldestimation by utilizing uniformly generated samples. However, the major disadvan-

tage of the QMC approach is that there is no practical way to know the error in the estimated yield, since it is impossible to calculate the total variation V_{HK} (f) in the Koksma-Hlawka in-equality. As a result, a confidence interval for the estimation can not be obtained. To overcome this issue, LDSs are randomly permuted by scrambling [32], which simply reorders the sequence of values independently. Therefore, a few differently scrambled QMC runs provide a standard deviation that can be used as a probabilistic measure of the estimation error [33]. However, the requirement of multiple runs result in increased synthesis time. The proposed approach described in [28] promises a solution to estimate the error bounds for the estimated yield while preserving the time efficiency of the optimizer, in which QMC and scrambled QMC are combined together.

3.3.1. Yield Estimation Method

In the previous section, QMC approach is proposed for the yield estimation part of the yield aware analog design automation tool, and it is seen that typically a few hundred LDS points are sufficient to make quite accurate estimation. On the other hand, the drawback of the QMC approach is that the statistical error in the estimation cannot be calculated because deterministic samples from LDS which have no natural variance are used in the QMC technique. Also, in higher dimensions, it is hard to calculate the exact values of the error ($Y - Y_N$), because the variance of integrand $V_{HK}(f)$ becomes intractable. Even if V_{HK} (f) is calculated and an upper bound is obtained, the estimated and exact value of integration would be very different. To overcome this bottleneck, and obtain a confidence interval of the estimated yield, scrambled-QMC technique [32], which is based on permuting the order of the sample set within a random manner, is exploited to obtain artificial statistical variance, as described in [28]. A scrambled-QMC sample set, $\{x_i^{(j)}\}^{N}_{i}$; j = 1,2,…,M, is selected and the yield is estimated for each sample set as

$$y^{(j)} = (1/N) * \sum_{i=1}^{N} f(x_i^{(j)}) \text{ , } j=1,2,\ldots,M \tag{3.4}$$

Then, the mean of the yield is calculated as

$$\hat{y} = (1/M) * \sum_{j=1}^{M} y^{(j)} \tag{3.5}$$

The error of numerical integration is estimated using the variance of the evaluated yield values, which is calculated as

$$\hat{\sigma}^2 = (1/M(M-1)) * \sum_{j=1}^{M} (y^{(i)} - \hat{y})^2 \tag{3.6}$$

Finally, the magnitude of the QMC error is calculated as

$$|E_{QMC}| = \hat{\sigma}.\phi^{-1}((1+p)/2) \tag{3.7}$$

with user defined probability p, where ϕ is the standard normal cumulative function. As a result, thanks to the randomness property of scrambled QMC, the minimum and the maximum bounds of yield with probability p can be obtained. In Figure 3.3 from the previous work [24], it can be seen that both conventional QMC and scrambled QMC are combined for a relatively accurate yield estimation.

It is observed that yield estimates up to 99% can be achieved by using QMC method. As already mentioned, this yield may not be sufficient for manufacturing purposes where 6-sigma is required. Therefore, we propose a new method for yield estimation part of the optimizer in order to further enhance the yield. Hence, only yield estimation part of the optimizer of the yield-aware analog CAD tool will be changed. Instead of using QMC, cross entropy minimization based importance sampling method will be utilized in the yield estimation part of the tool. Other parts of the optimizer will be unchanged.

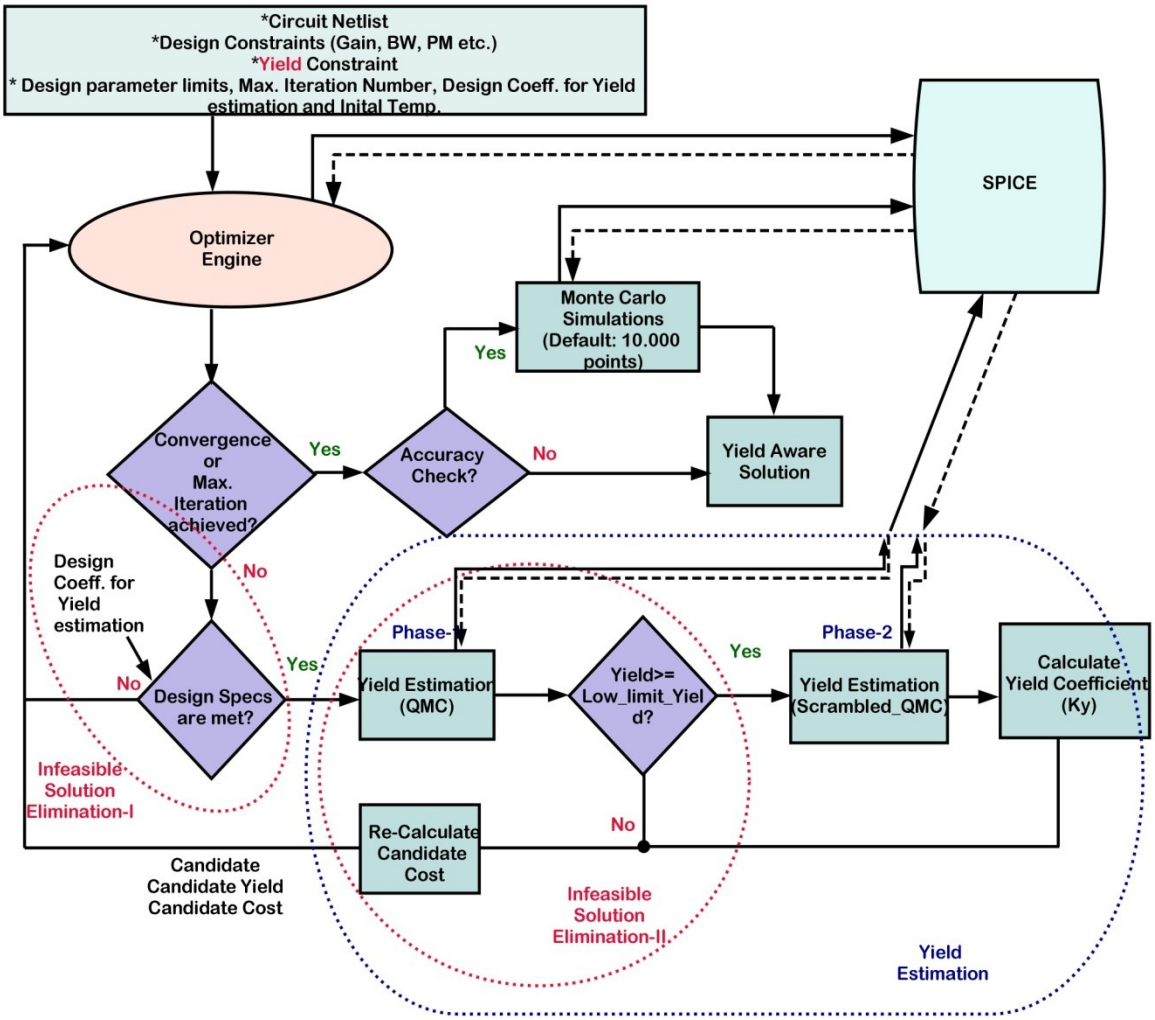

Figure 3.3. Flow diagram of the optimizer

3.3.2. Algorithm Implementation

As seen from Figure 3.3., conventional QMC is run for the first phase of the yield estimation. In QMC implementation, Sobol sequence is preferred among other LDS sets such as Halton, Faure etc., because it is empirically shown that it provides better results in higher dimensions as stated in [34]. Also, the first N points can be skipped in order to achieve more homogeneity, and thus, better sampling performance [33].

In the second phase, scrambled QMC simulations are run in order to obtain standard deviation. At the end of this phase, upper-lower bounds and standard deviation of the esti-

mation are obtained.

3.3.3. Simulation Results

In order to simulate and test QMC implementation, the basic two stage (BTS) OpAmp as shown in Figure 3.4 and folded cascode (FC) amplifier as shown in Figure 3.5 were chosen as test circuits. Deviations at threshold voltage, oxide thickness, and device geometries referring W and L, were considered as variation parameters during the yield estimation. The number of sample sizes for QMC can be determined as small as few hundred samples.

QMC simulation is run for exemplary BTS OpAmp circuit solution by using integrated circuit simulator SPICE. Sample size for QMC is selected as 500. Bandwidth, gain and phase margin are defined as circuit specifications for the BTS OpAmp. After running QMC simulation for given circuit solution, distribution histograms for bandwidth, gain and phase margin are obtained as shown in Figure 3.6, Figure 3.7, and Figure 3.8. respectively. We can obtain mean values and select rare event threshold points from these distribution histograms. For example, we can easily see that central frequency for the respective BTS OpAmp solution is 8.9 kHz. Similarly, mean of gain is found as 74.6 whereas mean of phase margin is found as 62.8 degrees.

QMC method will be essential for cross entropy minimization based importance sampling method, because IS needs a rare event threshold for an efficient estimation. Therefore, QMC will be run before the algorithm in order to select a good rare event threshold point for the desired set of specifications of the chosen circuit.

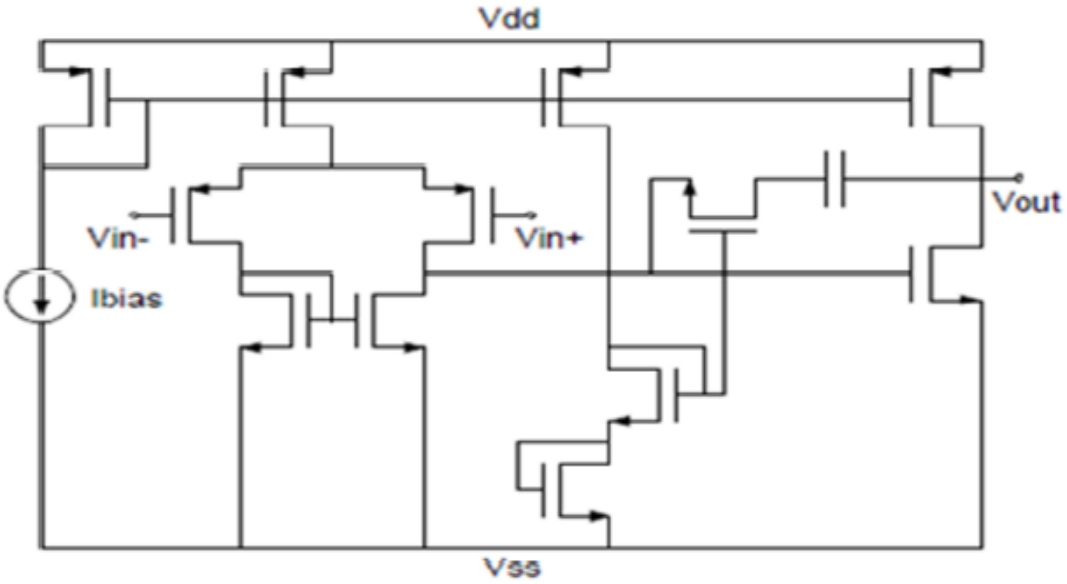

Figure 3.4. Schematic of BTS OpAmp

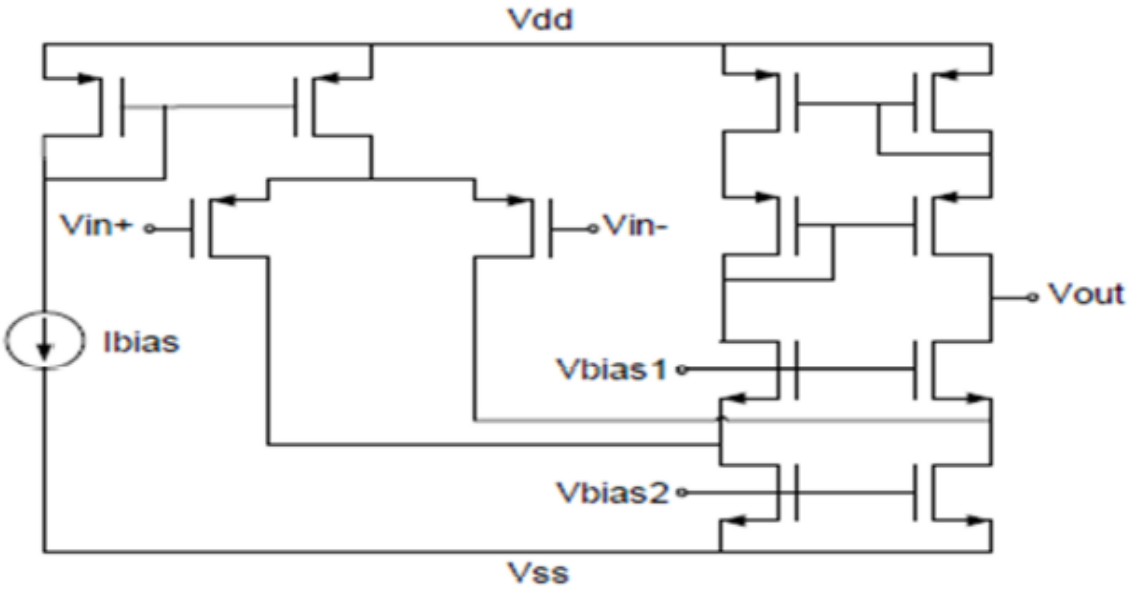

Figure 3.5. Schematic of FC Amplifier

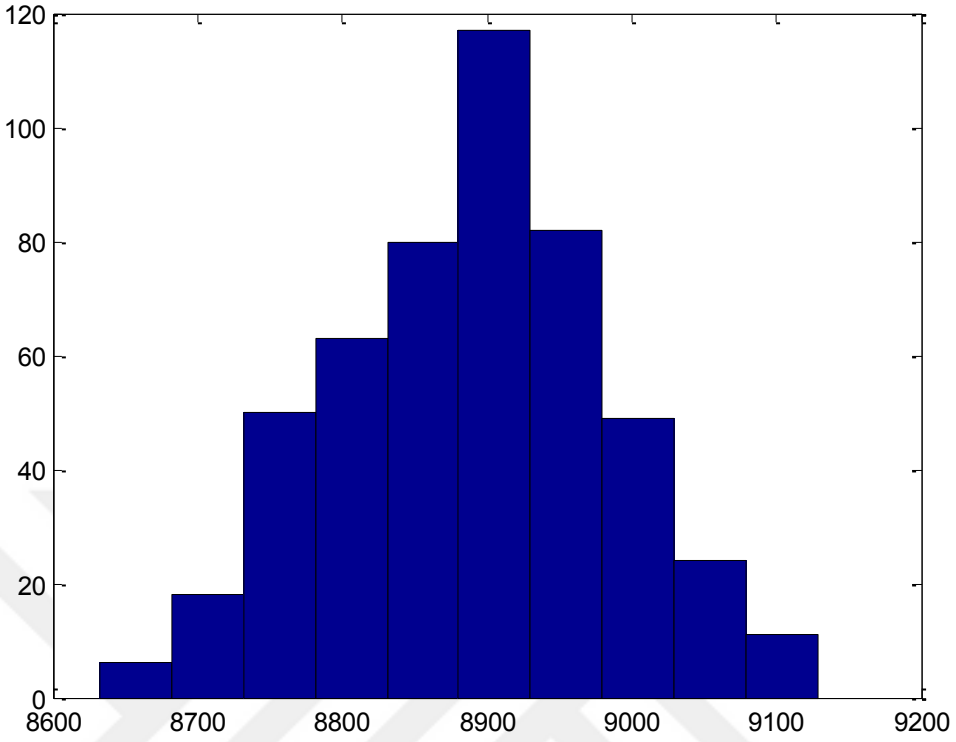

Figure 3.6. Bandwidth distribution for BTS OpAmp (QMC with 500 samples)

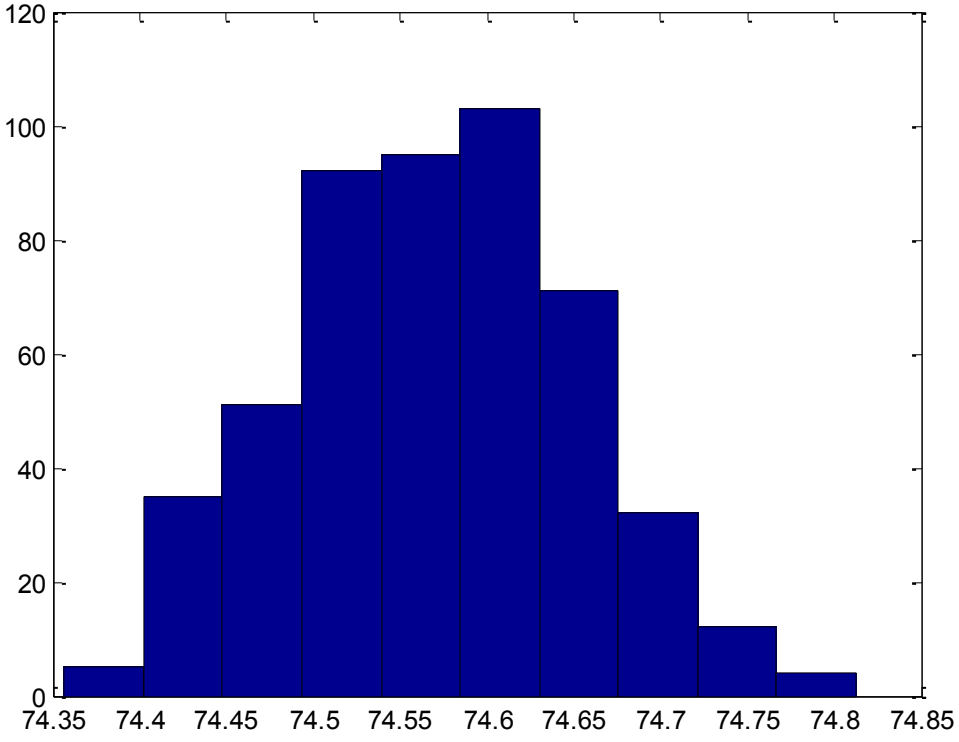

Figure 3.7. Gain distribution for BTS OpAmp (QMC with 500 samples)

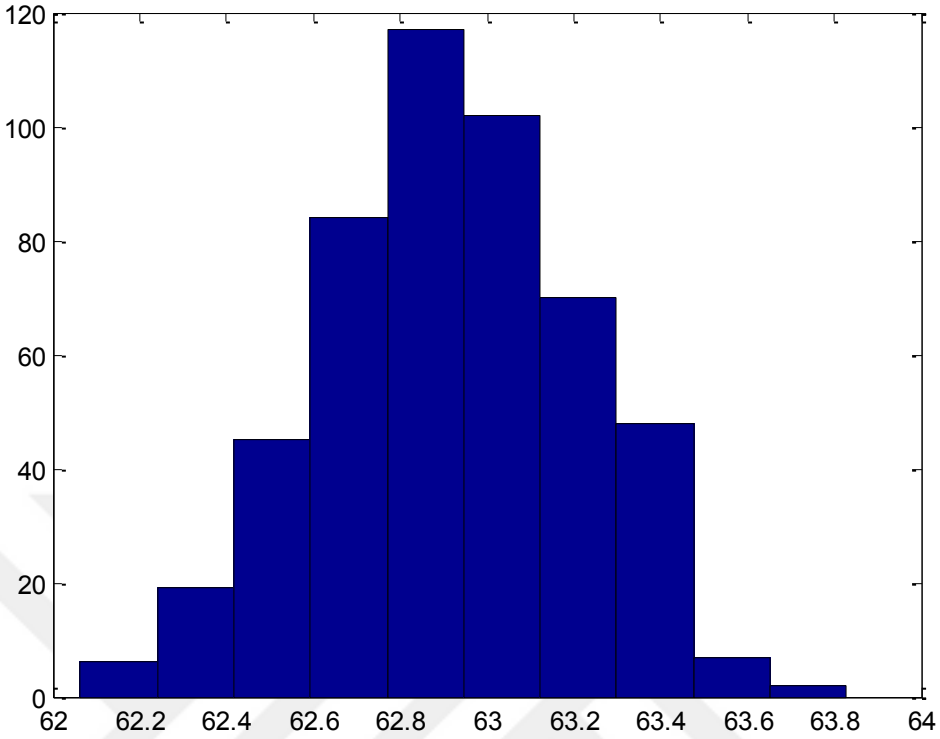

Figure 3.8. Phase margin distribution for BTS OpAmp (QMC with 500 samples)

4. EFFICIENT YIELD ESTIMATION AND ENHANCEMENT USING RARE EVENT SAMPLING METHODS

In the previous chapter, a hybrid QMC method is proposed for yield aware synthesis CAD tool. Since QMC method is deterministic and has no natural variance, there is no convenient way to obtain error bounds for the estimation. To determine the confidence interval of the estimated yield, scrambled QMC method and conventional QMC method were combined for an accurate yield estimation. However, with this hybrid QMC method, accurate yield estimations up to 99% can be achieved. This yield value is under the desired yield for manufacturing purposes. One solution for an enhanced yield estimation can be oversampling a special region where QMC fails to deliver an accurate estimation. This special region can be called as rare event or infeasible region. By using rare event sampling techniques, it is possible to get reliable estimation results from infeasible region. Therefore, rare event sampling methods will be utilized along with hybrid QMC method as variance reduction technique.

In the context of this chapter, rare event sampling methods will be analyzed and a hybrid method will be proposed in the next chapter to be introduced in CAD tool instead of QMC technique. Although there are many proposed rare event sampling methods in the literature, Importance Sampling (IS) and Cross Entropy Minimization based methods are the most common ones. These methods will be combined to create an efficient proposal for the scope of the thesis.

4.1. Background

Considering the effects of the process variations, the expected output specification distributions for the given circuit such as bandwidth, phase margin, gain, are expected to be skewed to the either left or right with a long right or left tail. Normally, a typical appro-

ach for simulating this circuit, would be running a Monte Carlo simulation with a small sample size (e.g. 1000) and fit a standard analytical distribution to data. Generally, either normal or lognormal distribution is chosen to be fit into simulation data in such a scenario. Such an approach can be accurate for the body part of the distribution, however it will be grossly inaccurate in the tail of the distribution. The skewness of the actual distribution or the heaviness of the tail will be difficult to overcome. As a result, any predictions of the statistics of rare events, lying far in the tail, will be very inaccurate. Therefore, it is unfeasible to use classic Monte Carlo approach with a small sample size in our yield aware CAD tool due to output distributions possessing a long tail.

The solution of this problem is somehow generating a large number of the samples in the tail. Generating a large number of samples in the tail is also theoretically possible by using Monte Carlo simulation with an extremely large sample size. But this approach is not practical considering heavy computational complexity and long simulation time. For our yield estimation, a straightforward Monte Carlo implementation would require hundreds of millions of samples in order to produce a handful of failures for obtaining the accurate data. Furthermore, the estimate of yield still can not be trusted because of the lack of statistical confidence because the estimate is computed using only one failing example. Such a large number of simulations is unfeasible and intractable. Thus, the best option is using rare event sampling methods for estimating the extreme statistics of rare events lying in the tail of the distribution.

In the literature, there are lots of different rare event simulation techniques that are proposed for estimating extreme rare event statistics most notable one being SRAM failure rate or SRAM yield. A handful of approaches based on statistical analysis have been proposed and investigated especially for verification of SRAM circuits and their rare failure event statistics. The most common approaches are based on Monte Carlo simulations ([36]-[39]). Most notable Monte Carlo based approaches in the literature can be given as statistical blockade [36], spherical sampling [37], mixture importance sampling [38], and scaled sigma sampling [39].

The other approaches are solely based on designing analytical performance measurement models in order to estimate parametric yield or probability of rare event circuit failure [40, 41]. However, these approaches suffer from approximations that are necessary to to make the problem tractable. There are also some proposals to combine both approaches [42].

In order to predict rare event failure probabilities in Monte Carlo simulation, Importance Sampling method can be applied [35]. Thus, sampling efficiency of Monte Carlo technique can be greatly improved with the aid of Importance Sampling. In Importance Sampling, the original distribution is shifted towards the rare event region also known as infeasible region. New shifted distribution can be called as practical Importance Sampling distribution. By using practical IS distribution, the infeasible region is now directly sampled. This approach is very trivial because essentially shifting the original distribution is all that is done. The nontrivial part is finding the optimum shift amount for fast and efficient estimation. For finding the optimum shift of the distribution, cross entropy can be used as a measure of distance between practical IS distribution and original distribution. In this context, minimizing cross entropy will lead to finding the optimum distribution shift [25].

4.2. Importance Sampling

In many applications we want to compute $\mu = E(\mathrm{f}(X))$ where $f(x)$ is nearly zero outside a region A for which $P(X \in A)$ is small. In this context, the outside region can be named as rare event region or infeasible region. The set covering outside of A may have small volume, or it may be in the tail of the X distribution. A conventional Monte Carlo sampling from the distribution of X could fail to have even one point outside the region A, depending on the probability distribution. Some problems result in extreme rare event probabilities by their nature. Problems of this type arise in high energy physics, Bayesian inference, finance, insurance, and rendering in computer graphics among other areas. It is clear intuitively that we must somehow get some samples from the important region. One way to do this is sampling from a distribution that overweights the important region. Thus, this type of solution is named as importance sampling. Having oversampled the important

region, we have to correct our estimate somehow to account for having sampled from this other distribution.

Importance sampling can bring enormous gains, making an otherwise infeasible problem feasible compared to classical Monte Carlo. It can also backfire, yielding an estimate with infinite variance when simple Monte Carlo would have had a finite variance. It is the hardest variance reduction method to use as well [43]. Therefore, it is clear that importance sampling method should be applied carefully, otherwise it may backfire. Importance sampling is also more than just a variance reduction method. It can be used to study one distribution while sampling from another. Some probability density functions are hard to integrate or sample. In that case, alternatively, we can use another probability density function which can be easily sampled. Then, calculating error and correcting the solution, we can get the estimate effectively. As a result we can use importance sampling as an alternative to acceptance-rejection sampling, as a method for sensitivity analysis and as the foundation for some methods of computing normalizing constants of probability densities.

4.2.1. Mathematical Approach

Our problem is to find

$$\mu = E(f(X) = \int_D f(x)p(x)dx \tag{4.1}$$

where p is a probability density function on $D \subseteq R^d$ and f is the integrand. We take $p(x)=0$ for all $x \notin D$. If q is a positive probability density function on R^d, then

$$\mu = \int_D f(x)p(x)dx = \int_D \frac{f(x)p(x)}{q(x)}\, q(x)dx = E_q\left(\frac{f(x)p(x)}{q(x)}\right) dx \tag{4.2}$$

where $Eq(\cdot)$ denotes expected value for $X \sim q$. Our original goal then is to find $Ep(f(X))$. By doing some multiplicative adjustment to f, we compensate for sampling from q instead

of p. The adjustment factor $p(x)/q(x)$ is called the likelihood ratio. The distribution q is known as importance distribution and p is known as nominal distribution.

The importance distribution q does not have to be positive everywhere. It is enough to have $q(x) > 0$ whenever $f(x)p(x) \neq 0$. That is, for $Q = \{x \mid q(x) > 0\}$ we have $x \in Q$ whenever $f(x)p(x) \neq 0$. Therefore, if $x \in D \cap Q^c$ we know that $f(x) = 0$, while similarly if $x \in Q \cap D^c$ we have $p(x) = 0$.

$$E_q\left(\frac{f(x)p(x)}{q(x)}\right) = \int_D f(x)p(x)dx + \int_{Q\cap D^c} f(x)p(x)dx - \int_{D\cap Q^c} f(x)p(x)dx \qquad (4.3)$$

$$E_q\left(\frac{f(x)p(x)}{q(x)}\right) = \int_D f(x)p(x)dx \qquad (4.4)$$

But one can ask that what happens for x with $q(x) = 0$ in the denominator. The answer is that there are no such points $x \in Q$ and we will never see one when sampling $X \sim q$. The importance distribution $q(x)$ can be close to 0 which leads extreme difficulties, but $q(x) = 0$ is not a problem if $f(x)p(x) = 0$ too.

When we want q to work for many different functions f_j, then we need $q(x) > 0$ at every x where any $f_j(x)p(x) \neq 0$. Then, a density q with $q(x) > 0$ whenever $p(x) > 0$ will suffice, and will allow us to add new functions fj to our list after we've drawn the sample.

The importance sampling estimate of $\mu = E_p(f(X))$ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\frac{f(Xi)p(Xi)}{q(Xi)}, \; Xi \sim q \qquad (4.5)$$

To use equation in (4.5) we must be able to compute $f\,p/q$. Assuming that we can compute f, this estimate requires that we can compute $p(x) / q(x)$ at any x we might sample.

When *p* or *q* has an unknown normalization constant, then we will resort to a ratio estimate.

The very basic idea of importance sampling is to draw a distribution *q(x)* from a similar distribution *p(x)* and then modify the resulting equation to correct the error introduced by sampling from wrong distribution. In equation (4.6), we can see that error correction or importance weight *w*, can be precisely determined for a given *x*, since we assumed that we could evaluate *p(x)* at a given point.

$$E_q\left(\frac{f(x)p(x)}{q(x)}\right) = \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w\, f(Xi) \;,\; w = \frac{p(Xi)}{q(Xi)} \tag{4.6}$$

4.2.2. Algorithm Implementation

As previously stated, importance sampling method can backfire for particular situations. Especially, as the number of samples is increased the variance of the estimation will lower. This means selection of *q(x)* will have huge impact on the success of importance sampling method by affecting accuracy of the estimation [44]. Thus, poor selection of importance sampling distribution *q(x)* will lead to wrong answer without any implication. In fact, this is one of the biggest problems of the importance sampling method. Importance sampling distribution *q(x)* should be chosen as close as possible to nominal distribution *p(x)*. In the literature, there are various divergence definitions in order to define the distance between *p(x)* and *q(x)*, and minimize this distance for an effective estimation. For example, alpha divergence [45] and Kullback-Leibler divergence [46] can be used as kind of distance measurement of two functions. In fact, Kullback-Leibler divergence is a special case of alpha divergence. When alpha is zero, alpha divergence becomes KL divergence. And for a special scenario KL divergence and cross entropy collides with each other.

For IS implementation, we will choose *q(x)* such that it covers *p(x)*. We will be avoiding variance and estimation blow up by choosing *q(x)* accordingly. Importance sampling

implementation is actually very trivial without applying cross entropy which helps to find optimum, best suited *q(x)* distribution. In the next chapter, cross entropy minimization based importance sampling will be proposed.

4.2.3. Simulation Results

In order to verify the implementation, we will use synthetic data with already defined distribution functions. $f_1(x) = x^2 + \log(x)$ and $f_2(x) = x^2 - e^{-x} + \log(x)$ functions are chosen as example simulation functions. These functions are chosen such that they reflect the nonlinear relation of the circuit specifications with respect to process variation parameters. All three bandwith, gain, phase margin functions are nonlinear combinations of the process variation parameters such as transistor width, transistor length, oxide thickness, and threshold voltage. Thus, by choosing similar examplary functions, we are able to simulate importance sampling method without integrating SPICE circuit simulation and synthesis.

Input distributions in the simulation actually refers to process variations in real circuit synthesis. Therefore, *x* input distribution is actually mapped to transistor width, transistor length, oxide thickness, and threshold voltage in real life. Generally, all static process variation parameter distributions are Gaussian distributions. Therefore, input synthetic data is chosen as normal distribution similar to static process variation parameters. For simulation purposes, input distributions are chosen as Gaussian distributions with the mean of 5.

Simulation results for exemplary functions are presented in the figures below. In Figure 4.1., a random normal input distribution with the mean of five is chosen. Input Gaussian distribution is generated for 10,000 samples. Then, input distribution is applied to f_1 and the resulting output distribution is presented in Figure 4.2. As can be seen from Figure 4.2., output distribution possesses a long tail due to having a nonlinear relationship with the input distribution. Then, IS is applied and the shifted input distribution is determined. As can be seen from Figure 4.3., shifted distribution is determined as a normal distribution with the mean of eight. Thus, the original input distribution is shifted to right by

three. This shift amount is heuristically determined. In fact, this is the drawback of the IS method because shift amount should be optimally selected for an efficient estimation. In Figure 4.4, the shifted output distribution of f_1 with respect to shifted input distribution is presented. If we compare Figure 4.2. and Figure 4.4., we can see that the tail of the original output distribution is oversampled by the new shifted output distribution. Thus, number of samples from the tail of the original distribution is dramatically increased with the IS method.

Similarly, same steps can be applied to f_2. In Figure 4.5., normal distribution with the mean of five is chosen as input distribution. Again, input Gaussian distribution is generated for 10,000 samples. Then, input distribution is applied to f_2 and the resulting output distribution is presented in Figure 4.6. Then, IS method is applied and the shifted input distribution is determined. As shown in Figure 4.7., shifted distribution is determined as a normal distribution with the mean of eight. Hence, the original input distribution is shifted to right by three. The shifted output distribution of f_2 with respect to shifted input distribution is obtained which is shown in Figure 4.8. If we compare Figure 4.6. and Figure 4.8., again we see that the tail of the original output distribution is oversampled by the new shifted output distribution.

Our main goal for using IS method is increasing the number of samples in the rare event region as already mentioned. This is achieved by shifting the output distribution to rare event region so that these region is oversampled. And hence as its name suggest, we give importance to this rare event region, and try to get as many sample as possible. After simulation, we can clearly see that input distributions (process variations) should be shifted left or right in order to also shift the output distribution. Hence, the question arise, how can we choose optimum shift amount? This can be achieved with cross entropy minimization.

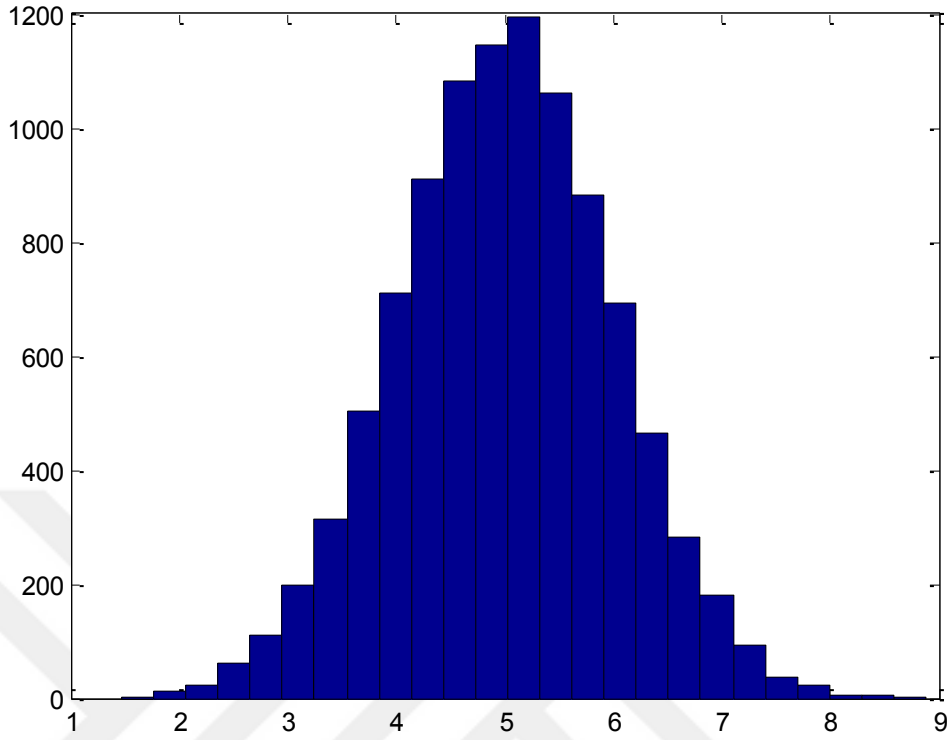

Figure 4.1. Input normal distribution (mean=5, sigma=1) for N=10000 samples

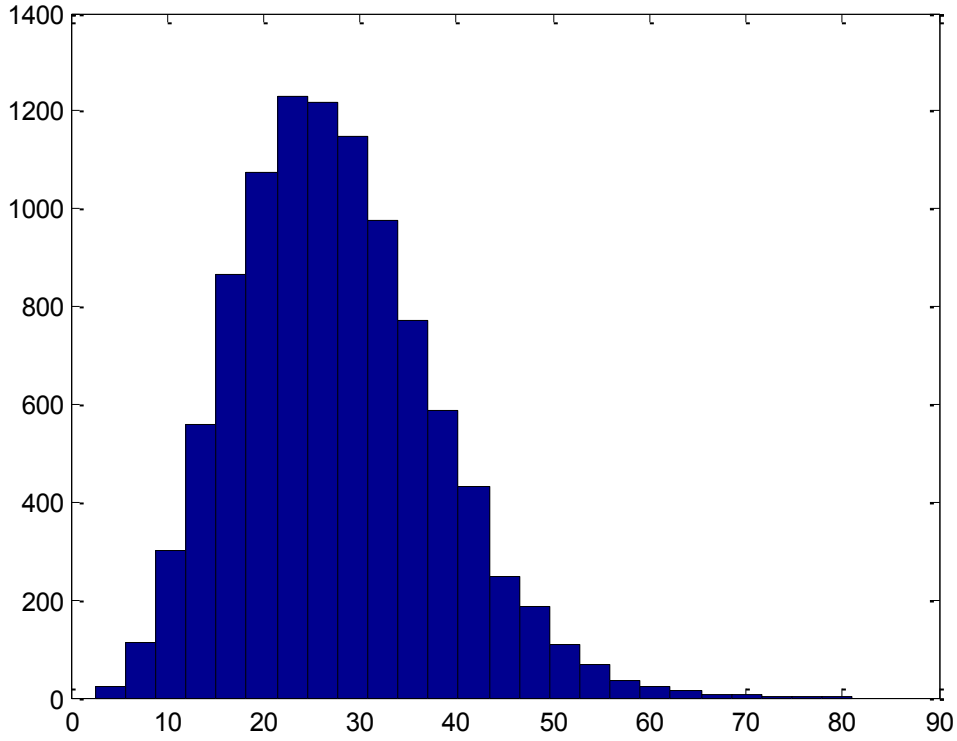

Figure 4.2. Output distribution $f_1(x) = x^2 + \log(x)$ for given input distribution

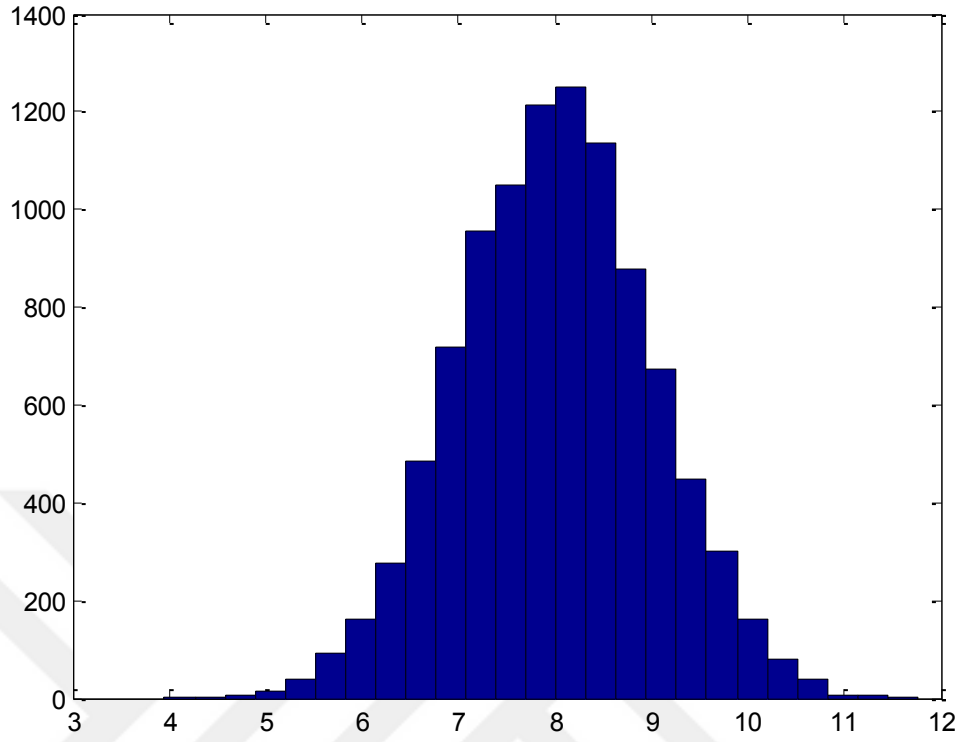

Figure 4.3. Shifted input normal distribution (mean=8, sigma=1) for N=10000 samples

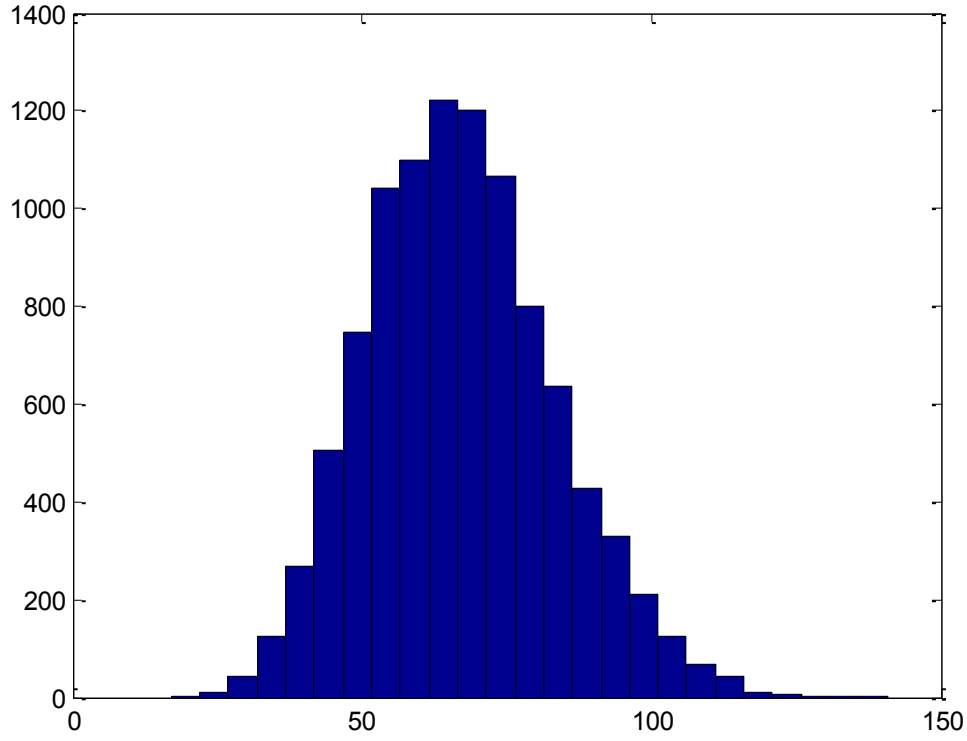

Figure 4.4. Shifted output distribution $f_1(x) = x^2 + \log(x)$ for shifted input distribution

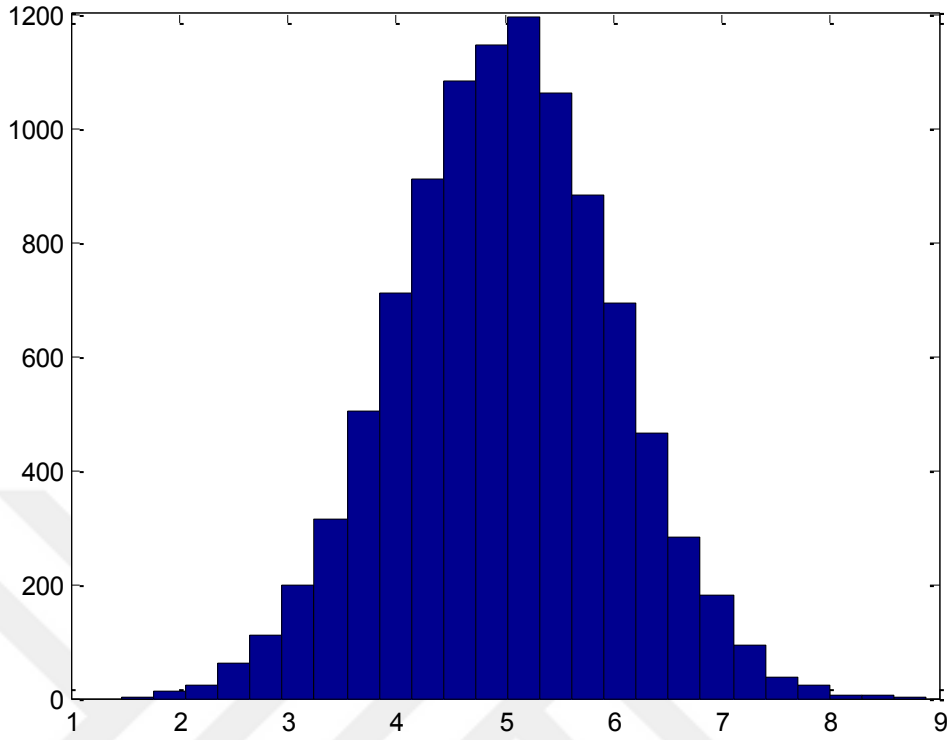

Figure 4.5. Input normal distribution (mean=5, sigma=1) for N=10000 samples

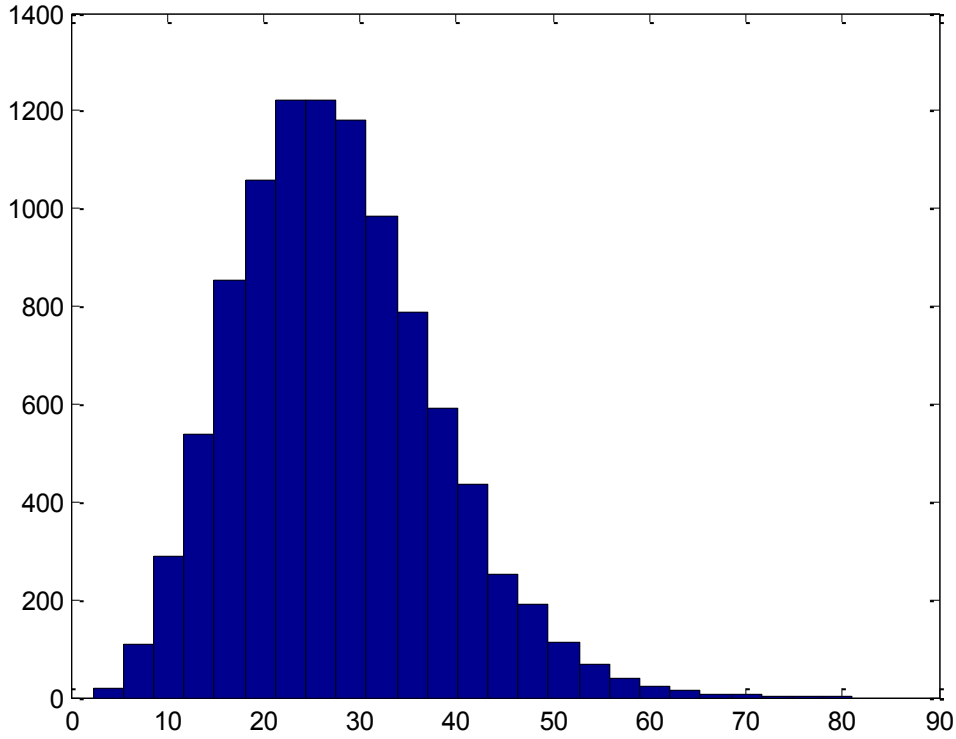

Figure 4.6. Output distribution $f_2(x) = x^2 - e^{-x} + \log(x)$ for given input distribution

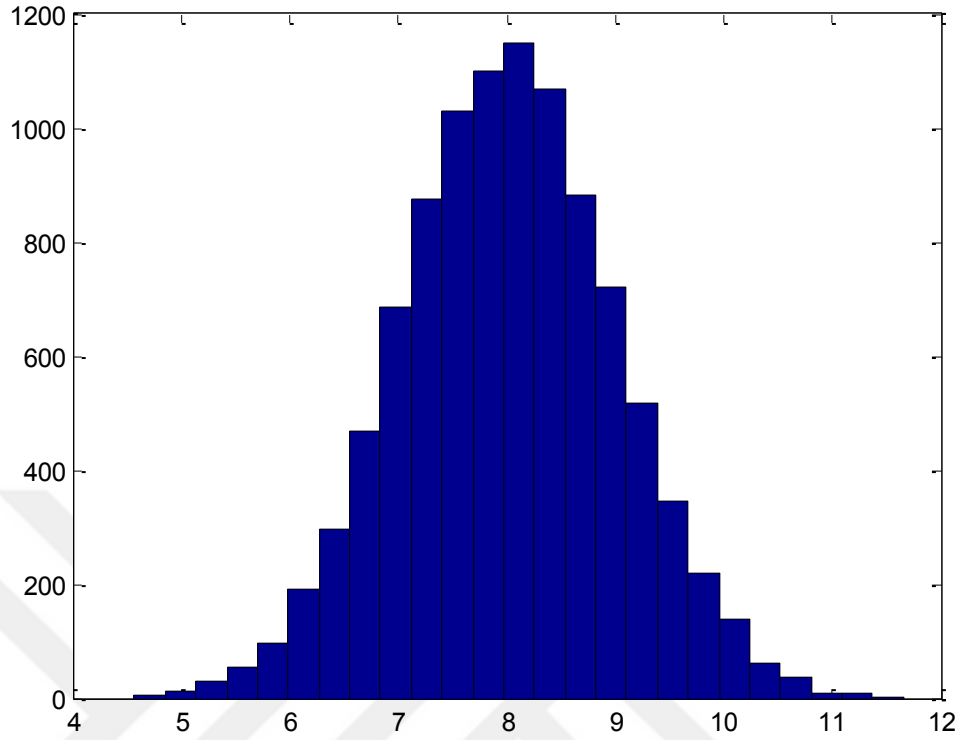

Figure 4.7. Shifted input normal distribution (mean=8, sigma=1) for N=10000 samples

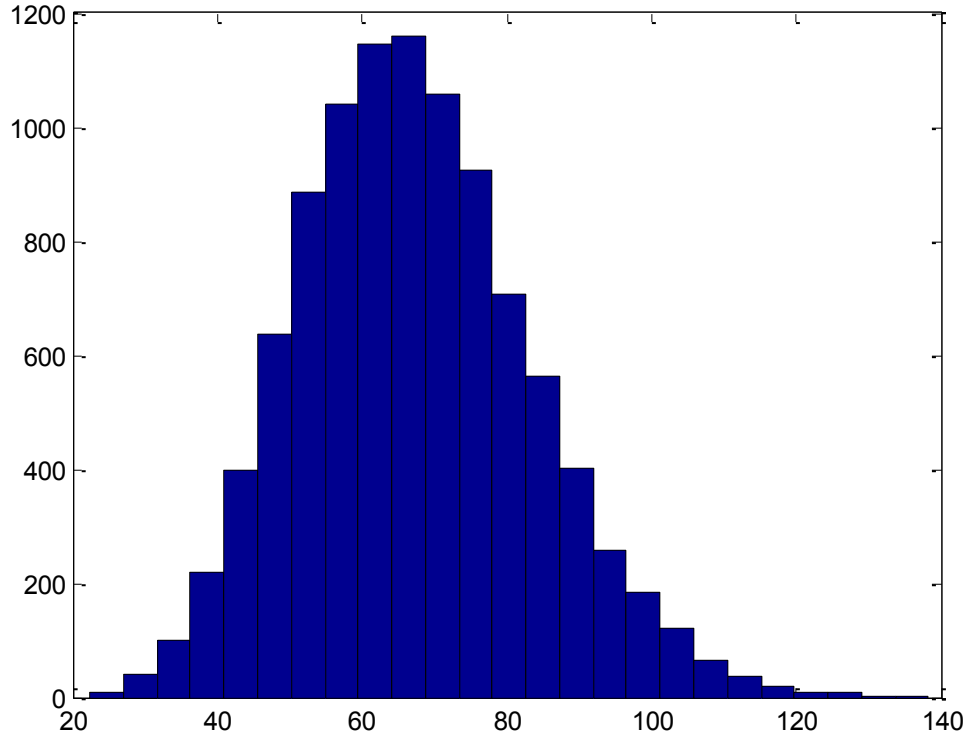

Figure 4.8. Shifted output distribution $f_2(x) = x^2 - e^{-x} + \log(x)$ for shifted input distribution

4.3. Cross Entropy Minimization

4.3.1. Background

Cross entropy minimization is a well known method in information theory. It is generally used for approximating the optimal solution of NP hard combinatorial optimization problems and estimation of probability of rare event [47]. Our main goal here is minimizing cross entropy. If the original distribution is fixed, then cross entropy between two distributions becomes identical to Kullback-Leibler divergence between two distributions. Kullback-Leibler divergence of two distributions such as *g* and *h* can be given as [47]

$$D(g, h) = E_g \log \frac{g(X)}{h(X)} = \int g(x)\log g(x)dx - \int g(x) \log h(x)\, dx \qquad (4.7)$$

where f(x, u) represents the original probability distribution function, g(x) represents the optimal importance sampling density, and h(x) = f(x, p). Then, optimal importance sampling density can be found as

$$g^*(x) = \frac{I\{S(x) \geq \gamma\}\, f(x,u)}{l} \qquad (4.8)$$

Minimizing KL divergence between g^* and f(x, p) is the same as choosing p such that expression (4.9) is minimized. Therefore, it is not hard to see that expression (4.9) is equivalent to expression (4.10).

$$- \int g^*(x) \log f(x, p)\, dx \qquad (4.9)$$

$$\max_p \int g^*(x)\log f(x, p)dx \tag{4.10}$$

If we substitute g^* which is given in expression (4.8) into expression (4.10), we obtain following maximization problem

$$\max_p \int \frac{I\{S(x)\geq\gamma\}\, f(x,u)}{l} \log f(x, p)dx \tag{4.11}$$

which is equivalent to following optimization problem.

$$\max_p D(p) = \max_p E_u\, I\{S(x)\geq\gamma\}\log f(X,p) \tag{4.12}$$

4.3.2. Algorithm Details

In cross entropy minimization method, our main goal is to solve the optimization problem presented in (4.12). The main objective is finding a sequence of tuples $\{ \gamma_t, p_t \}$ which converges to a small neighborhood of the optimal tuple (γ^*, p^*). The parameters γ_t and p_t are adaptively updated in each step in order to converge to the optimal tuple. The measure of this convergence is defined by a rarity parameter which is denoted as ρ. During initialization stage, $p_0 = u$ is set as initial starting point, the rarity parameter ρ is chosen such that it is not very small. Then, we iteratively update γ_t and p_t parameters.

Let γ_t be the $(1-\rho)$ quantile of S(X) under p_{t-1}. A simple estimator of γ_t can be obtained by drawing a random sample $X_1,\ldots,X_N$ from $f(x, p_{t-1})$. Then, the associated

function values $S(X_1),\ldots,S(X_N)$ and their order statistics $S_{(1)},\ldots,S_{(N)}$ can be calculated. Assigning γ_t to be order order statistics' $(1-\rho)$ quantile gives us following update formula.

$$\gamma_t = S_{(\lfloor(1-\rho)\, N\rfloor+1)} \tag{4.13}$$

Now, we have the adaptive update formula of γ_t.We can derive p_t for fixed γ_t and p_{t-1} from the solution of the following problem.

$$\max_{p_t} D(p_t) = \max_{p_t} E_{p_{t-1}} I\{S(x)\geq\gamma_t\} \log f(bx, p_t) \tag{4.14}$$

If we solve the problem (4.14) for a discrete n-dimensional probability distribution function with independent components, we can obtain following analytical expression for updating $p_{t,ij}$ where i=1,…,n and j=1,…,m

$$p_{t,ij} = \frac{E_{p_{t-1}} I\{x_i=j\}\; I\{S(x)\geq\gamma_t\}}{E_{p_{t-1}} I\{S(x)\geq\gamma_t\}} \tag{4.15}$$

But we need stochastic counterparts of equations (4.14) and (4.15), in order to simplify the calculation [47] resulting following equations (4.16) and (4.17). Keep in mind that equation (4.16) is the stochastic counterpart of the equation (4.14) whereas the equation (4.17) is the stochastic counterpart of the equation (4.15).

$$\max_{\tilde{p}_t} D(p_t) = \max_{\tilde{p}_t} \frac{1}{N} \sum_{i=1}^{N} I\{S(X_i)\geq\gamma_t\} \log f(X_i\, \tilde{p}_t) \tag{4.16}$$

$$\tilde{p}_{t,ij} = \frac{\sum_{i=1}^{N} I\{x_{ki}=j\}\ I\{S(x_k)\geq\gamma_t\}}{\sum_{i=1}^{N} I\{S(x_k)\geq\gamma_t\}} \tag{4.17}$$

Instead of using $\tilde{p}_t$ obtained from (4.17), the algorithm uses smoothed version $\bar{p}_t$ presented in (4.18). The reason for using the smoothed vector is to reduce the probability of some of components of $\tilde{p}_t$ being zero or unities at the first few iterations. If the smoothing is not applied, the algorithm may converge fast to a local optimum which will result in wrong solution. Therefore, $\tilde{p}_t$ vector is smoothed with a smoothing parameter which is denoted as α. The smoothing parameter should take a value between zero and one. For α=1, it is obvious that smoothed vector $\bar{p}_t$ will be same with the original vector $\tilde{p}_t$.

$$\bar{p}_t = \alpha\tilde{p}_t + (1-\alpha)\tilde{p}_{t-1} \tag{4.18}$$

By using adaptive update equations, we have obtained a sequence of tuples $\{ \gamma_t, p_t \}$ which converges to a small neighborhood of the optimal tuple (γ^*, p^*). Overall cross entropy minimization algorithm is presented in Figure 4.9.

Choose some initial p_0

Set iteration counter t = 1

while solution not converged

Generate a sample $X_1,\ldots,X_N$ from density $f(x, p_{t-1})$

Compute γ_t from Eq. 4.13;

Use the sample $X_1,\ldots,X_N$ and find $\tilde{p}_t$ from Eq. 4.17;

Smooth out the vector $\tilde{p}_t$ to obtain $\bar{p}_t$ using Eq. 4.18;

if $\gamma_t = \gamma_{t-1} = \gamma_{t-2} = \ldots = \gamma_{t-d}$ for some $d \leq t$ **then**

Solution found, stop iteration, break while loop

else

Increment iteration counter, t = t + 1

Continue iteration

end if

end while

Figure 4.9. Main Cross Entropy Minimization Algorithm

5. MULTI LEVEL CROSS ENTROPY MINIMIZATION BASED IMPORTANCE SAMPLING METHOD

In the previous chapter, IS is shown to be a trivial method without determining optimum shift which will result in the optimal practical distribution for IS. Optimal practical distribution for IS can be found with various numerical optimization techniques. For example, norm minimization can be used for one dimensional problems [48]. For high dimensional problems, a variant of norm minimization called spherical sampling can be utilized [37]. However, it is shown that these methods may have suboptimal performance due to either suboptimal shift or performance degradation for high dimensional problems [41, 49]. Minimum cross entropy method [47] finds the optimal practical distribution for IS which is closest in distance to ideal distribution for IS. This distance can be defined with cross entropy which is used excessively in information theory.

Let's define p to be original probability distribution and q to be ideal probability distribution. Then, we can define cross entropy between these two probability distributions as following:

$$H(p, q) = E_p[-\log q] = H(p) + D_{KL}(p \parallel q) \tag{5.1}$$

where $H(p, q)$ is the cross entropy between p and q probability distributions, $H(p)$ is the entropy of p, $D_{KL}(p||q)$ is the Kullback-Leibler divergence of q from p or also called relative entropy of p with respect to q [46], and E[.] as expected value.

If p probability distribution is fixed, then cross entropy between p and q is identical to Kullback-Leibler divergence with an additive constant. In this case, minimizing cross entropy will be the same as minimizing KL divergence. In the literature, the principle of minimizing KL divergence is known as Principle of Minimum Cross Entropy (MCE) or Minxent [47]. And if p and q probability distribution overlay each other meaning $p = q$, then both cross entropy and KL divergence will be zero. But this is not practically possible for our case.

5.1. Background

Let yield be characterized by a random variable *X*. This random variable X is a nonlinear function *F* of circuit variables which are affected by process variations such as transistor width, length, threshold voltage, oxide thickness. Under process variations especially in submicron technologies, these circuit variables deviate from their nominal values significantly. We can define

$$X = F(Y_1, Y_2, \ldots, Y_M) \tag{5.2}$$

where *X* represents random variable, *F* represents nonlinear function of circuit variables, and *Yi* represents the circuit variables. The changes in circuit variables can not be estimated. Therefore, it is difficult to find the distribution of the random variable *X*. But as mentioned in the previous chapter, these process variations make distribution of random variable *X* to have a heavy tail by nature. Therefore, using one of the rare event simulation technique such as IS with combining MCE will be an efficient way to calculate yield.

Deviations in the circuit variables can be modeled as independent Gaussian random variables with mean u_i and variance σ_i^2 and their distribution can be represented as $f(y_i, u_i)$. This type of modelling is fair because process variations are random and their effects on the circuit variables is independent [25]. In IS technique, the original distribution is just shifted towards the rare region, therefore the original distribution mean u_i becomes shifted to a new mean v_i. Cross entropy minimization only improves efficiency of the IS algorithm by making the shift amount optimal. However, the variance σ_i^2 does not change. This can be seen as the disadvantage of proposed algorithm, because it is expected to have a new distribution with optimal mean and variance in the infeasible region for better solutions. However, variance is discarded in our approximation. So, new distribution can be represented as $f(y_i, v_i)$. This new shifted distribution is called optimal practical distribution for IS. In the IS algorithm weight function is used, and it can be calculated as follows

$$w(y, u, v)= \frac{\prod_{i=1}^{M} f(y_i, u_i)}{\prod_{i=1}^{M} f(y_i, v_i)} \tag{5.3}$$

where y_i represents circuit variables, u_i represents original distribution means, v_i represents shifted distribution means, and M the number of circuit variables. For example, BTS OpAmp consists of 12 transistors as can be seen from Figure 3.4. Since we try to analyze the effects of deviations in transistor length, width, oxide thickness and threshold voltage to the yield, we have 4 variables for each transistor. Therefore, M is 48 for BTS OpAmp.

Since we model deviations in the circuit variables as independent random Gaussian variables, weight function $w(y, u, v)$ can be rewritten as follows [25]:

$$w(y, u, v)= \exp\left(- \sum_{i=1}^{M} \frac{2y_i(v_i-u_i)- (v_i^2- u_i^2)}{2\sigma_i^2} \right) \tag{5.4}$$

Now, we have weight function in our hands. The other and the most significant issue is finding the optimal shifted mean vector denoted as v^*. As mentioned before, this optimal shifted mean can be found by minimizing KL divergence between the ideal distribution and optimal practical distribution for IS. Here, ideal distribution for IS will be denoted as $f_{ideal}(y)$, whereas optimal practical distribution for IS will be denoted as $f(y, v^*)$. Therefore, the main optimization algorithm is finding $f(y, v^*)$ which is closest in distance to $f_{ideal}(y)$. Hence our original distribution is fixed; KL divergence will be equal to cross entropy. The cross entropy distance between ideal distribution and optimal practical distribution for IS can be defined as:

$$D = E_{f_{ideal}} \left[\log \frac{f_{ideal}(y)}{f(y, v^*)} \right] \tag{5.5}$$

Minimizing KL divergence by using IS formula for finding ideal distribution, it is not hard to find the following equation [25]:

$$v^{*}=\text{argmax}_{v}\ E_{v}\left[I(y\in A)\log\left(f(y, v)\right)\right] \tag{5.6}$$

Equation (5.6) gives us the optimum shifted mean vector v^{*} of the distribution $f(y, v^{*})$. But finding v^{*} is computationally expensive and we may not get enough samples from rare event infeasible region. Furthermore, we have to use $I(y\in A)$ indicator function which indicates rare event possibility such that it takes value 1 if the chosen sample is indeed in the rare event region. Otherwise, the indicator function takes the value 0. One solution to cope with the possibility of not getting enough samples from the rare event region is using multiple levels of cross entropy method. For this purpose, the original distribution $f(y, u)$ is shifted to some initial approximate distribution $f(y, l)$ by IS technique. Therefore, weight function of applied IS technique will be $w(y, u, l)$. The initial shifted mean vector is denoted as l. And it should be chosen such that current event becomes less rare. Obviously, selection of initial vector l is important. To select proper l, norm minimization approach is utilized. In the next subchapter, initial vector selection will be briefly discussed. With the usage of multiple levels of cross entropy, equation in (5.6) can be modified as

$$v^{*}=\text{argmax}_{v}\ E_{v}\left[I(y\in A)w(y, u, l)\log\left(f(y, v)\right)\right] \tag{5.7}$$

In order to solve above maximum optimization problem, we can rephrase equation (5.7) in terms of KL divergence with respect to v^{*}. Therefore, the problem becomes minimizing KL divergence with respect to v^{*} which is denoted as $D(v^{*})$.

$$D(v^{*})=E_{v}\left[I(y\in A)w(y, u, l)\log\left(f(y, v)\right)\right] \tag{5.8}$$

In order to solve this optimization problem, we have to find a v^{*} such that $D'(v^{*}) = 0$. If we solve equation (5.8) for $D'(v^{*}) = 0$, we get a simple analytical expression [25]:

$$v^{*}= \frac{\sum_{i=1}^{N} I\left(y^{(i)}\in A\right)w(y^{(i)}, u, l)y^{(i)}}{\sum_{i=1}^{N} I(y^{(i)}\in A)w(y^{(i)}, u, l)} \tag{5.9}$$

5.2. Algorithm Implementation

In the first step, we have to find a proper inital vector *l* for the original distribution *f(y, u)*. As mentioned in the previous section, the original distribution *f(y, u)* is shifted to some initial approximate distribution *f(y, l)* by IS technique, so that event becomes less rare. For selection of the initial vector, state of the art norm minimization approach is utilized. The idea behind norm minimization is getting an approximate initial direction for making the event less rare. The aim is not solving the problem completely, it is rather making an approximation to find an initial starting point.

In the first step of norm initialization algorithm, a few random shifts of the original means u_i are generated to get the new shifted means l_i. This step is done for each circuit variable. Now, we have shifted means l_i. In second step, we get shifted distributions $f(y,l_i)$ using shifted means l_i. We generate samples from $f(y, l_i)$ and run simulations on generated samples. After simulation, shifted means l_i which resulted in the infeasible region are filtered out. Then L2 norm is calculated by using these filtered out shifted means. L2 norm can be calculated from formula below:

$$\text{L2 Norm of } l = \sum_{i=1}^{M} \frac{(l_i - u_i)^2}{2\sigma_i^2} \tag{5.10}$$

After calculating the L2 norm, we can choose the inital vector by taking the shifted mean vector l_i which resulted in a minimum L2 norm value.

$$\text{Initial vector } l = \text{argmin} \sum_{i=1}^{M} \frac{(l_i - u_i)^2}{2\sigma_i^2} \tag{5.11}$$

Generate a few uniform random shifts of the original means u_i

Then using random shifts get new shifted means l_i

Use shifted means to obtain $f(y, l_i)$

Generate samples from $f(y, l_i)$ and run simulations

Filter out shifted means which result in rare event region

Use filtered out shifted means to calculate L2-norm using Eq. 5.10;

for i = 1 to M **do**

Choose ith index which makes L2-norm minimum, Eq. 5.11;

end for

Choose l_i as initial shifted mean vector

Figure 5.1. Norm Minimization for Inital Vector Selection Algorithm

As the algorithm in Figure 5.1. suggests, samples which correspond to rare events should be screened out for a proper inital vector selection. Initial vector selection is significant, because it directly affects the efficiency of the algorithm. Therefore, a rare event threshold should be chosen wisely for the algorithm which will be used for screening out the samples that will reside in the infeasible region. The samples that results beyond this chosen rare event threshold will be screened out. In order to choose a proper rare event threshold, we need to have a quick glimpse on the original distribution before running cross entropy minimization based IS algorithm. As stated in chapter 3, Quasi Monte Carlo can be run for general analysis that will give accurate results for the body part of the distribution. We could also use classical Monte Carlo simulation. However, QMC gives similar approximate results compared to classical Monte Carlo with much smaller sample size. Hence, QMC is chosen over classical Monte Carlo for a quick estimation. With QMC, we can get an idea about rare event point that will mark the rare event region threshold. Therefore, QMC algorithm will be run one time before cross

entropy minimization based IS algorithm in order to choose a rare event threshold.

After choosing initial vector *l*, shifted distribution *f(y, l)* is also obtained. Then, N1 samples are generated from this shifted distribution *f(y, l)*. From equation (5.9), the optimum shifted mean vector v^* can be calculated now. Keep in mind that equation (5.9) needs a weight function to be calculated. So, before calculating v^*, weight function should be calculated from equation (5.4.). After finding the optimum shifted mean vector, we generate a Gaussian distribution $f(y, v^*)$ for the respective shifted mean as if it is a process variation distribution. This distribution is the optimum practical distribution for cross entropy minimization based IS. Since we have 48 process variables for BTS OpAmp, v^* is a vector with length of 48. Hence, after the algorithm is run, we get 48 Gaussian distributions for each shifted mean. As mentioned before, variance does not change because of the algorithm. This is actually the downside of the cross entropy minimization based IS algorithm. After all these generated optimum practical distributions are obtained, SPICE simulation is run. Circuit specification distributions such as bandwidth, phase margin, gain are obtained. Then, the obtained distributions are compared with the ones that resulted after QMC. Normally, it is expected that the distributions that are obtained after cross entropy minimization based IS algorithm should be shifted versions of the ones in QMC.

Choose inital vector l using original distribution f(y, u)

Generate N1 samples from the shifted distribution f(y, l)

Calculate optimum shifted mean vector v^* using Eq. 5.9 and Eq. 5.4;

Obtain optimum practical distribution f(y, v^*)

Generate N2 samples from optimum practical distribution f(y, v^*)

Run circuit simulations

Estimate yield

Figure 5.2. Multi Level Cross Entropy Minimization Based IS Algorithm

5.3. Simulation Results

Cross entropy minimization based IS algorithm is run for bandwidth, phase margin, gain and all three of them simultaneously. As synthesis result, here we present the bandwidth results. Firstly, QMC is run with 500 samples for the sample circuit BTS OpAmp. From QMC analysis, we can obtain the rare event threshold. We may choose 9.1 kHz as our rare event threshold for bandwidth which can be seen from Figure 5.3. In this case, the portion of corresponding rare samples to whole samples can be calculated as 0.004 which corresponds to rare event probability.

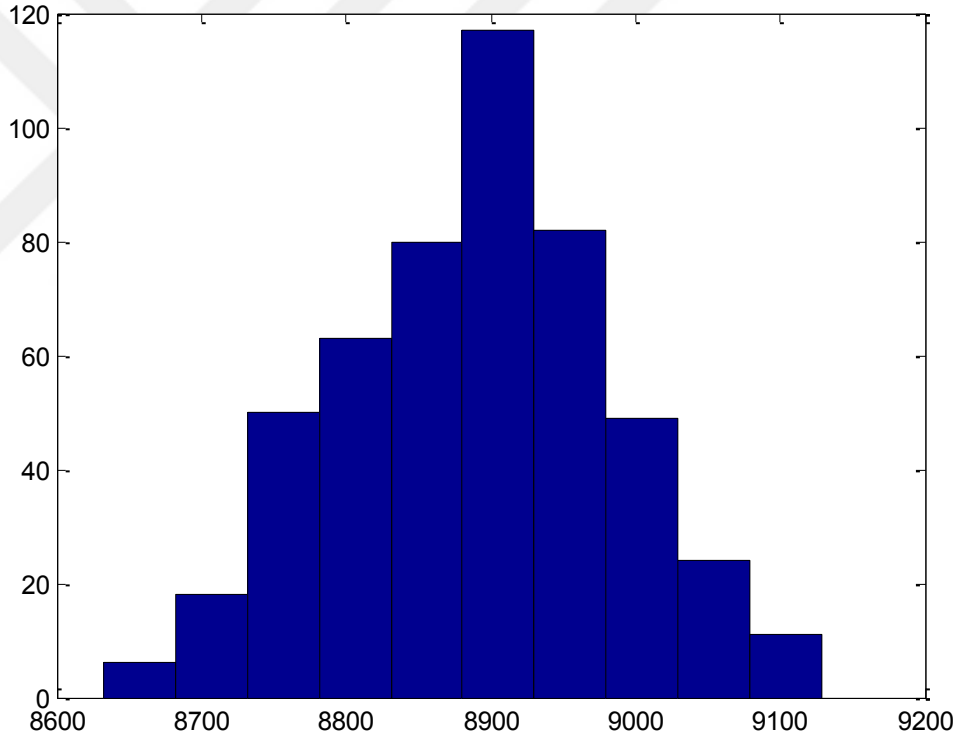

Figure 5.3. Bandwidth distribution after QMC with 500 samples

After applying cross entropy minimization based IS algorithm, corresponding Figure 5.4. is obtained for bandwidth. From Figure 5.4., it can be easily seen that majority of samples are clustered beyond the rare event threshold which is chosen as 9.1 kHz from the previous QMC simulation. In this case, ratio of samples which are beyond the threshold to

the total number of samples can be calculated as 0.9933. The main objective of using rare event sampling techniques was increasing the number of samples in the tail of the original distribution. From the given example, we can say that this goal was achieved, because probability value of 0.004, which is a rare event probability, is increased to 0.993 with the proposed algorithm. Therefore, one can say that entropy minimization based IS algorithm is an efficient way to analyze the distributions that possess long tails.

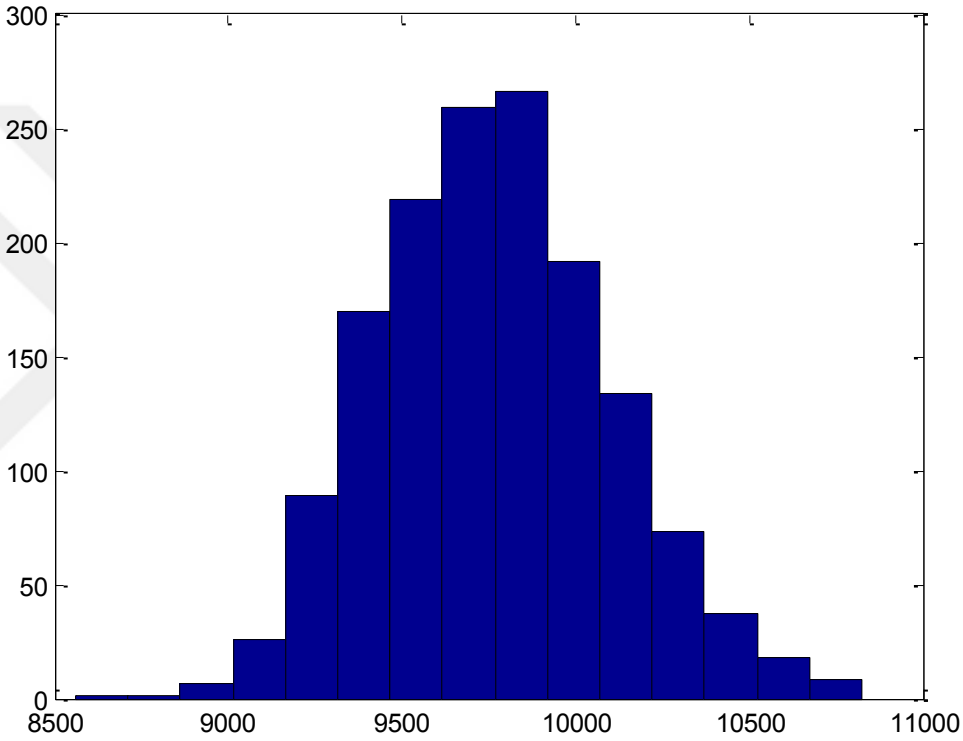

Figure 5.4. Bandwidth distribution after cross entropy based IS with 1500 samples

6. YIELD CALCULATION

6.1. Rare Event Modelling

Inspired from rare event modelling algorithms, here we propose yield estimation algorithm for both calculating yield and evaluating the efficiency of the cross entropy minimization based IS algorithm. Initially, a conventional MC simulation is run with a large number of samples such as 100,000. The distribution obtained from conventional MC has a long tail where MC is not capable of giving accurate estimation. However, we are certain about the accuracy of the estimation based on samples that reside in the main part of the distribution. Therefore, we determine 3 arbitrary points from MC distribution in which we are certain about their accuracy. Hence, it is best to choose these 3 points from the body part of the distribution and not from the tail, which corresponds to the rare event region. These points are denoted as R1, R2, and R3 respectively. As mentioned before, we need to have a rare event threshold before running cross entropy minimization based IS algorithm. For this purpose, QMC with a small sample size such as 1000 is run. From QMC analysis, rare event threshold is determined. For the sake of reliability, 3 different rare event thresholds are chosen from QMC analysis. These rare event threshold points can be denoted as T1, T2, and T3. For each rare event threshold T, the cross entropy minimization based IS algorithm is run. For different R values, the estimate of yield is calculated by the expression in (6.1).

$$P_T = P_R * (P_T' / P_R') \tag{6.1}$$

where P_T represents rare event probability for selected rare event threshold T, P_R represents reliable probability for a given arbitrary R obtained from QMC simulation, P_T' and P_R' represent probabilities for respective T and R values obtained from cross entropy minimization based IS algorithm. After calculating P_T, we can cross validate the result with conventional MC in order to determine the efficiency of our algorithm. Overall yield estimation algorithm is presented in Figure 6.1.

Run conventional MC with large sample size (100k)
Run QMC (1k) to determine rare event threshold T
For defined T determine 3 arbitrary points R1, R2, R3
Run cross minimization based IS with defined T and R values
Calculate the estimate yield
Cross validate the result with MC result found in first step

Figure 6.1. Yield Estimation Algorithm

6.2. Yield Estimation Results

In the first step, conventional MC analysis with sample size of 100,000 is run. Distribution obtained from MC will be used for cross validation of results obtained from cross entropy minimization based IS algorithm. In Figure 6.2., conventional MC analysis for bandwidth is presented. In order to determine rare event threshold values T1, T2, and T3, QMC simulation is run with the sample size of 1000. Rare event thresholds are chosen as 8150, 8200, and 8300 Hz. We chose rare event thresholds from left tail of the original distribution because yield estimate is determined by the low limit of the bandwidth specification. R values are chosen arbitrarily. The important aspect of choosing R values is choosing points whose probabilities from conventional MC distribution presented in Figure 6.2 are certain. QMC analysis of bandwidth specification for 1000 samples is shown in Figure 6.3. Cross entropy minimization based IS distributions of bandwidth specification for 8.15, 8.2, and 8.3 kHZ respectively are presented in Figure 6.4, Figure 6.5., and Figure 6.6.

After running cross entropy minimization based IS, we have obtained accurate estimate yields. For cross validation, the results are compared with the results obtained from conventional MC. With this validation, we are able to see how approximate our estimation based on cross entropy minimization based IS is.

Simulation results for 3 independent runs for bandwidth specification for BTS OpAmp circuitry is presented on Table 6.1, Table 6.2 and Table 6.3.

Table 6.1. Simulation results for T=8150 Hz for BTS Opamp.

	R(Hz)	IS - Yield(%)	MC – Yield(%)	Relative Error(%)
1	8850	99.52	99.67	0.15
2	8900	99.51	99.67	0.16
3	8950	99.52	99.67	0.15

Table 6.2. Simulation results for T=8200 Hz for BTS Opamp.

	R(Hz)	IS - Yield(%)	MC – Yield(%)	Relative Error(%)
1	8850	99.35	99.46	0.11
2	8900	99.35	99.46	0.11
3	8950	99.35	99.46	0.11

Table 6.3. Simulation results for T=8300 Hz for BTS Opamp.

	R(Hz)	IS - Yield(%)	MC – Yield(%)	Relative Error(%)
1	8850	98.45	98.54	0.09
2	8900	98.48	98.54	0.06
3	8950	98.45	98.54	0.09

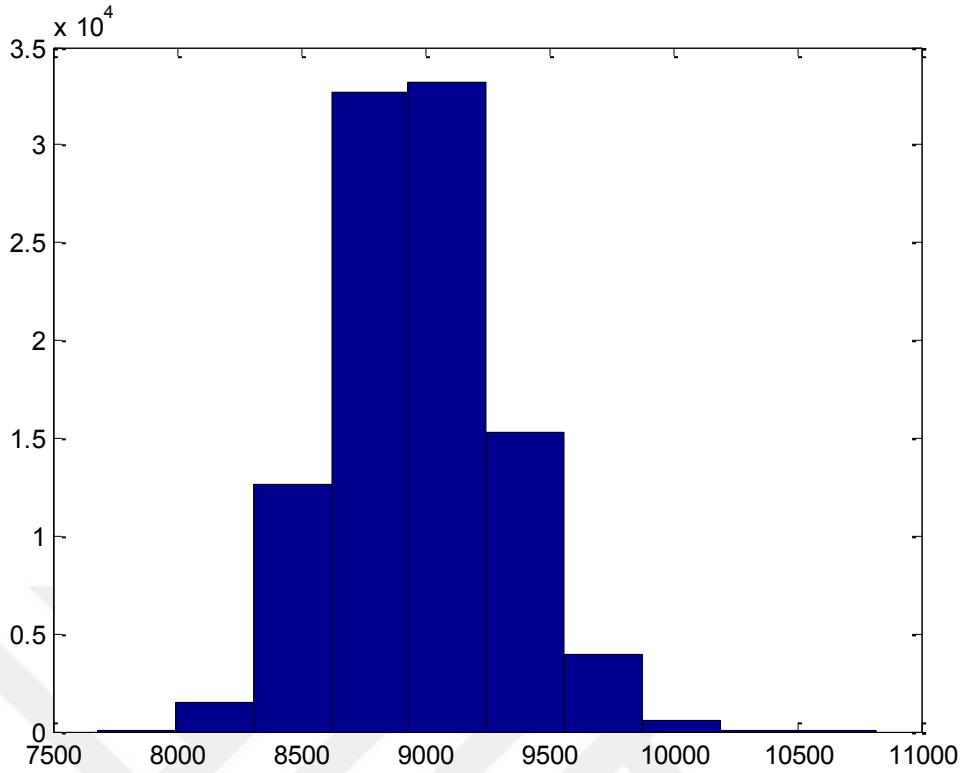

Figure 6.2. Conventional MC with 100k samples for bandwidth in BTS OpAmp

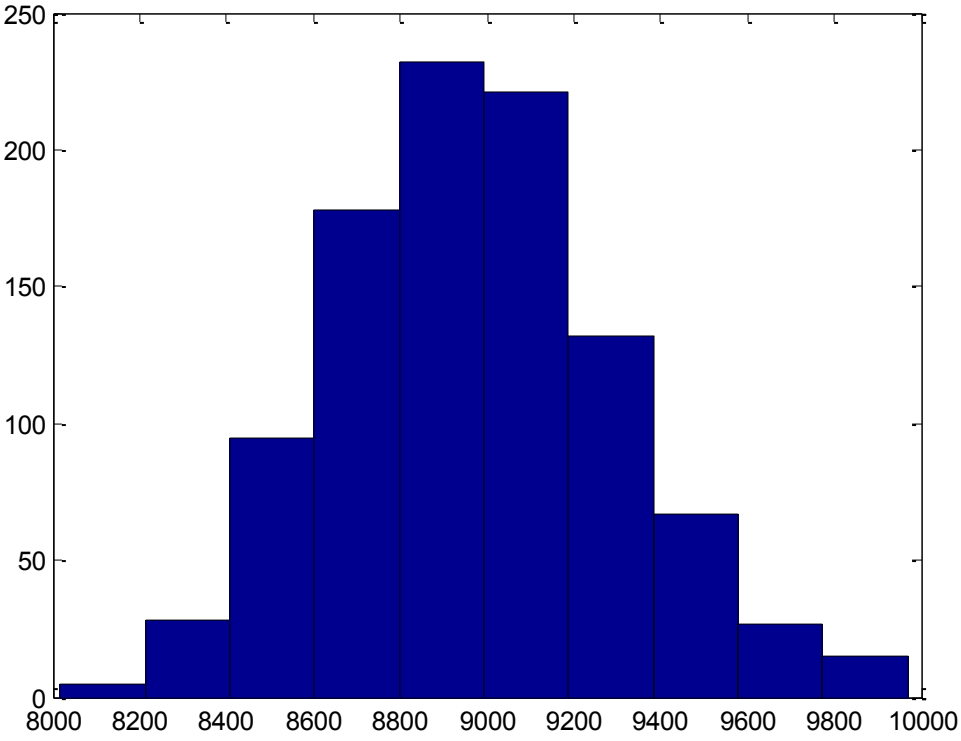

Figure 6.3. QMC with 1k samples for bandwidth in BTS OpAmp

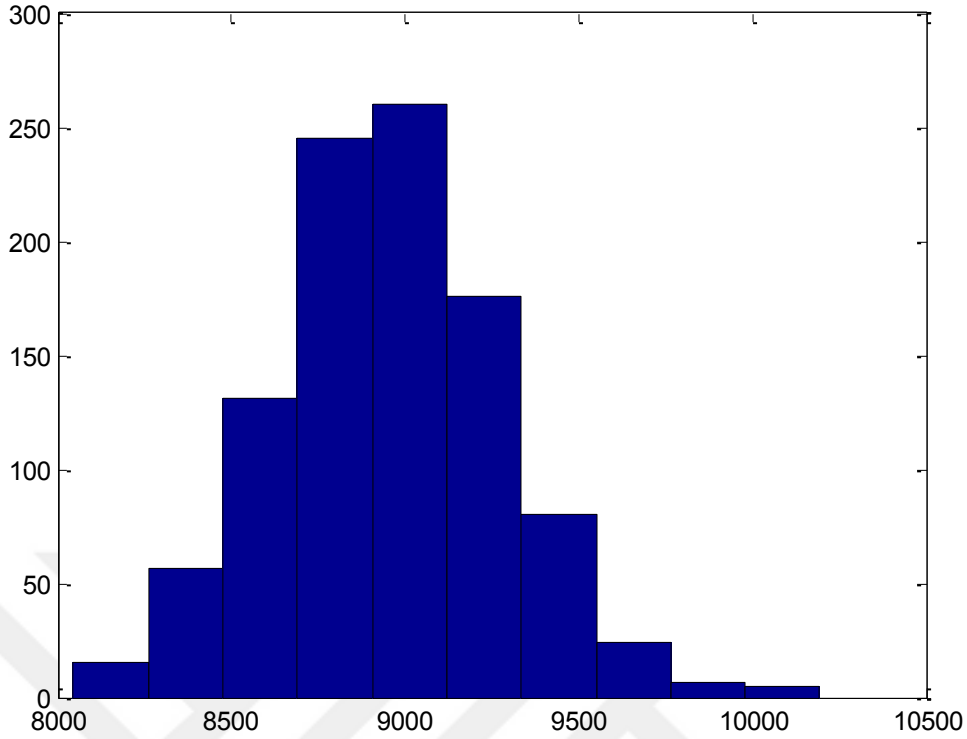

Figure 6.4. MCE based IS with 2k samples for bandwidth in BTS OpAmp (Rare event threshold is chosen as 8.15 kHz)

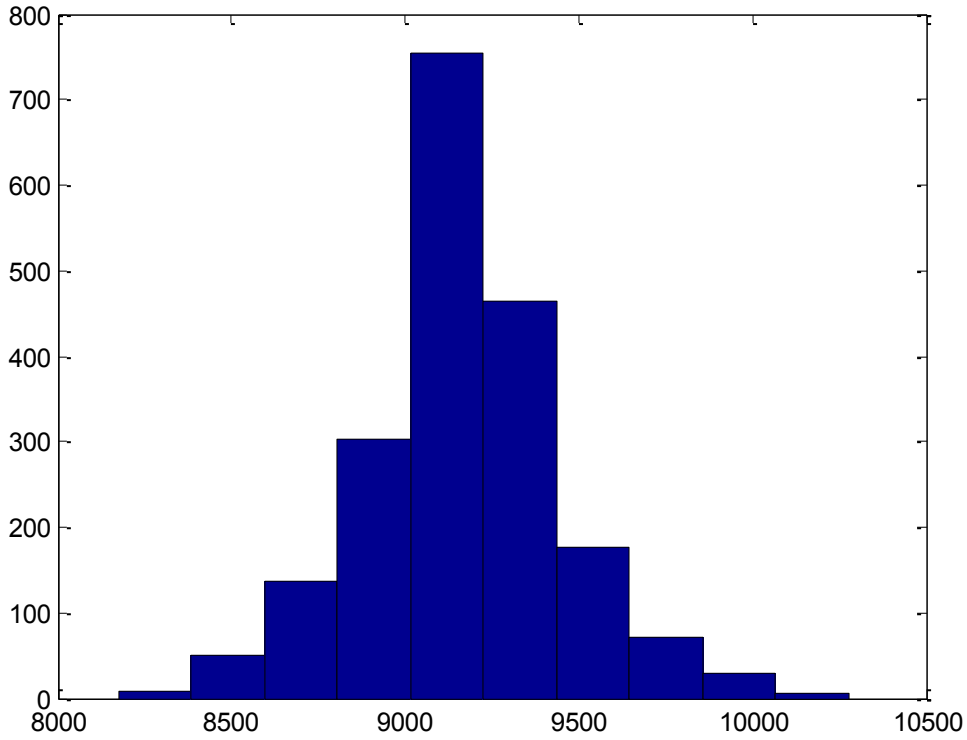

Figure 6.5. MCE based IS with 2k samples for bandwidth in BTS OpAmp (Rare event threshold is chosen as 8.2 kHz)

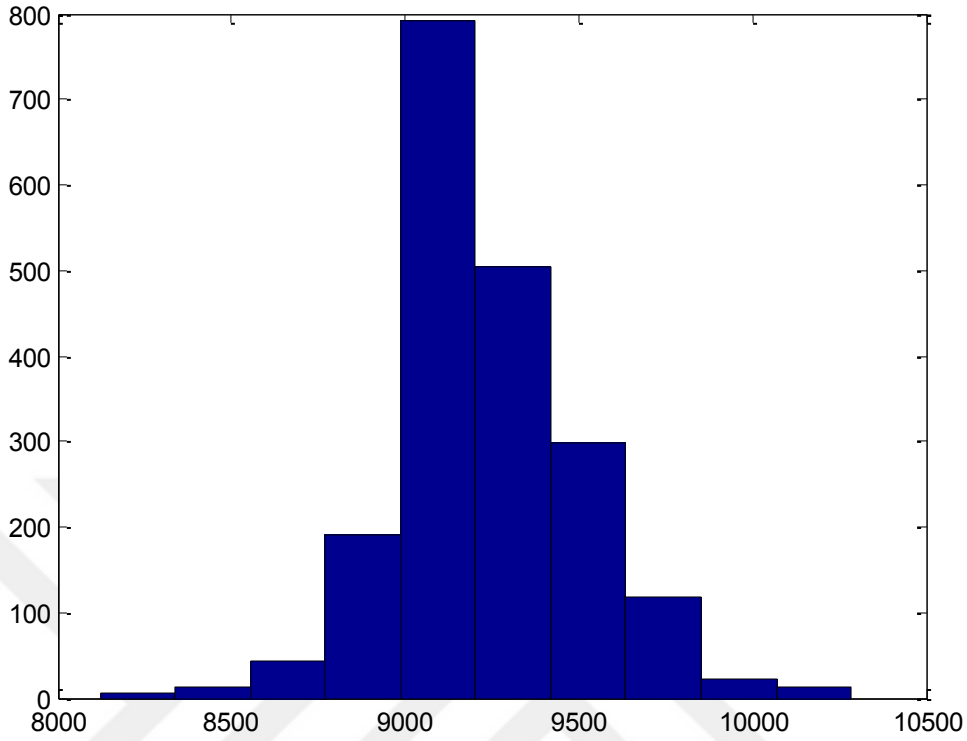

Figure 6.6. MCE based IS with 2k samples for bandwidth in BTS OpAmp (Rare event threshold is chosen as 8.3 kHz)

7. CONCLUSION

With the scaling of feature size and technology, process variation effects have worsened. Especially in submicron technologies, worsening process variation effects result in difference between simulated and measured performances of manufactured ICs. This problem leads to low yields for manufactured ICs for CMOS technology. Furthermore, increased circuit and silicon complexity complicates the analysis of analog circuits. Generally, electronic designers tend to overcome this kind of problems by leaving a margin. However, this results overdesign and causes loss of precious chip area. Therefore, yield-aware optimization has become a must for electronic designers with the worsening process variation effects. Yield aware optimization is a daunting task due to the trade-off between the accuracy and the efficiency of the yield estimation.

For an approximate yield estimation with smaller sample size compared to classical Monte Carlo simulation, QMC based variability analysis is adopted. However, error bound can not be estimated for yield estimation because QMC lacks of natural variance. To tackle this problem, hybrid QMC method which combines both scrambled and conventional QMC methods has been utilized.

Although QMC is a reliable and an efficient method for a quick approximation for yield estimation, it can't be used for certain yield estimations. Because the distribution of design specifications with respect to process variation effects tends to have heavy tail by nature. MC based QMC or conventional MC methods are inefficient for distributions that have heavy tail. Therefore, a rare event sampling method is proposed for increasing number of samples corresponding to tail of the original distribution. For this purpose, cross entropy minimization based IS method is chosen as rare event sampling method for the scope of this thesis due to its efficiency. In this context, cross entropy minimization concept is applied in order to increase the efficiency of IS algorithm. Normally, the major problem of IS is finding optimal shift hence optimal probability distribution. Otherwise, IS becomes a trivial method where one needs to just shift the original distribution. With the usage of cross entropy as a distance of measure between optimal practical distribution and

original distribution. After deploying proposed cross entropy minimization based IS algorithm, the tail of the original distribution becomes oversampled. Accurate yield estimation can be done after rare event region of the original distribution is oversampled.

In chapter 6, rare event statistics and modelling based proposal is presented for yield calculation. Yield estimate that found after algorithm is compared to conventional MC analysis in order to see the efficiency of proposed algorithm.

Our future work will focus on extending the range of the tail estimates to extremely rare events with good confidence.

REFERENCES

1. Gielen, G. G. and R. Rutenbar, "Computer-Aided Design of Analog and Mixed-Signal Integrated Circuits", Proceedings of the IEEE, Vol. 88, No. 12, pp. 1825-1854, 2000.

2. Moore, G. E., “Cramming More Components onto Integrated Circuits", Proceedings of the IEEE , Vol. 86, No. 1, pp. 82-85, 1998.

3. Haron, N. Z., S. Hamdioui, “Why is CMOS scaling coming to an END?”, Design and Test Workshop, IDT 2008, 3rd International, pp. 98-103, 2008.

4. Committee, I. R., “International Technology Roadmap for Semiconductors, 2011 Edition”, Semiconductor Industry Association, http://www.itrs.net/Links/2011ITRS/2011ExecSum.pdf, accessed at May 2016.

5. Afacan, E., G. Berkol, F. Baskaya and G. Dundar, “Sensitivity Based Methodologies for Process Variation-Aware Analog IC Optimization", Microelectronics and Electronics (PRIME), 10th Conference on Ph. D. Research in, pp. 1-4, IEEE, 2014.

6. Conti, M., P. Crippa, S. Orcioni and C. Turchetti, “Parametric Yield Formulation of MOS IC's Affected by Mismatch Effect", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 18, No. 5, pp. 582-596, 1999.

7. Bagad, V. S., *VLSI Design*, 2nd Edition, Technical Publications Pune, India, 2009.

8. Liu, B., Computational Intelligence Techniques for Automated Design of Analog and High-Frequency Circuits, Ph.D. Thesis, Katholieke Universiteit Leuven, 2012.

9. Morio J., Balesdant, M. “A survey of rare event simulation methods for static input-output models”, Simulation Modelling Practice and Theory, Vol. 49, No. 5, pp. 287-304, 2014.

10. Juneja, S., Shahabuddin, P., *Handbooks in Operations Research and Management Science - Simulation*, Elsevier, 2006.

11. Harjani, R., L. R. Carley and R. Rutenbar, “OASYS: A Framework for Analog Circuit Synthesis", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 8, No. 12, pp. 1247-1266, 1989.

12. Degrauwe, M. G., O. Nys, E. Dijkstra, J. Rijmenants, S. Bitz, B. L. Go art, E. Vittoz, S. Cserveny, C. Meixenberger, G. Van Der Stappen and H. J. Oguey, “IDAC: An Interactive Design Tool for Analog CMOS Circuits", Solid-State Circuits, IEEE Journal of , Vol. 22, No. 6, pp. 1106-1116, 1987.

13. El-Turky, F. and E. E. Perry, “BLADES: An Artificial Intelligence Approach to Analog Circuit Design", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 8, No. 6, pp. 680-692, 1989.

14. Hershenson M., S. P. Boyd and T. H. Lee, “Optimal Design of a CMOS Op-amp via Geometric Programming”, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 20, No. 1, pp. 1-21, January 2001.

15. Mandal, P. and V. Visvanathan, “CMOS Op-Amp Sizing Using a Geometric Programming Formulation”, Computer-Aided Design of Integrated Circuits and Sytems, Transactions on, Vol. 20, No. 1, pp. 22-38, January 2001.

16. Dawson, J. L., S. P. Boyd, M. del Mar Hershenson and T. H. Lee, “Optimal Allocation of Local Feedback in Multistage Amplifiers via Geometric Programming”, Circuits and Sytems I: Fundamentals Theory and Applications, IEEE Transactions, Vol. 48, No. 1, pp. 1-11, 2001.

17. Xu Y., K. L. Hsiung, X. Li, L. T. Pileggi and S. P. Boyd, “Regular Analog/RF Integrated Circuits Design Using Optimization with Resource Including Ellipsoidal Uncertainity”, Computer-Aided Design of Integrated Circuits and Sytems, Transactions on, Vol. 28, No. 5, pp. 623-637, 2009.

18. Koh, H. Y., C. H. Sequin and P. R. Gray, “OPASYN: A Compiler for CMOS Operational Amplifiers", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 9, No. 2, pp. 113-125, 1990.

19. Gielen, G. G., H. C. Walscharts and W. Sansen, “Analog Circuit Design Optimization Based on Symbolic Simulation and Simulated Annealing", Solid-State Circuits, IEEE Journal of , Vol. 25, No. 3, pp. 707-713, 1990.

20. Van der Plas G., G. Debyser, F. Leyn, K. Lampaert, J. Vandenbussche, G. Gielen, W. Sansen, P. Veselinovic and D. Leenarts, “AMGIE-A Synthesis Environment for CMOS Analog Integrated Circuits”, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 20, No. 9, pp. 1037-1058, 2001.

21. Afacan, E., G. Berkol, A. E. Pusane, G. Dundar and F. Baskaya, “Adaptive Sized Quasi-Monte Carlo Based Yield-Aware Analog Circuit Optimization Tool", CMOS Variability (VARI), 2014 5th European Workshop on, pp. 1-6, IEEE, 2014.

22. Phelps, R., M. Krasnicki, R. Rutenbar, L. R. Carley and J. R. Hellums, “Anaconda: Simulation-Based Synthesis of Analog Circuits via Stochastic Pattern Search", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 19, No. 6, pp. 703-717, 2000.

23. Keding, H., M. Willems, M. Coors and H. Meyr, “FRIDGE: A Fixed-Point Design and Simulation Environment", Proceedings of the Conference on Design, Automation and Test in Europe, pp. 429-435, IEEE Computer Society, 1998.

24. Berkol, G., Novel Design Method for Analog Design Automation Tools, MSc. Thesis, Bogazici University, 2015.

25. Shahid, M. A., “Cross Entropy Minimization for Efficient Estimation of SRAM Failure Rate”, Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012 20th Conference, pp. 230-235, IEEE, 2012.

26. Asenov, A., S. Kaya and A. R. Brown, “Intrinsic Parameter Fluctuations in Decananometer MOSFETs Introduced by Gate Line Edge Roughness", Electron Devices, IEEE Transactions on, Vol. 50, No. 5, pp. 1254-1260, 2003.

27. Ye, Y., S. Gummalla, C.-C. Wang, C. Chakrabarti and Y. Cao, “Random Variability Modeling and Its Impact on Scaled CMOS Circuits", Journal of computational electronics, Vol. 9, No. 3-4, pp. 108-113, 2010.

28. Afacan, E., G. Berkol, A. E. Pusane, G. Dündar and F. Başkaya, “A Hybrid Quasi Monte Carlo Method for Yield Aware Analog Circuit Sizing Tool", Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition, pp. 1225-1228, EDA Consortium, 2015.

29. Singhee A., S. Singhal and R. A. Rutenbar, “Practical, Fast Monte Carlo Statistical Static Timing Analysis: Why and How", Proceedings of the 2008 IEEE/ACM International Conference on Computer-Aided Design, pp. 190-195, IEEE Press, 2008.

30. Hlawka, E., “Funktionen Von Beschrankter Variatiou in der Theorie der Gleichverteilung", Annali di Matematica Pura ed Applicata, Vol. 54, No. 1, pp. 325-333, 1961.

31. Asmussen, S. and P. W. Glynn, Stochastic Simulation: Algorithms and Analysis, Springer, 2007.

32. Owen, A. B., Randomly Permuted (t, m, s)-Nets and (t, s)-Sequences, Springer, 1995.

33. Singhee, A. and R. A. Rutenbar, Novel Algorithms for Fast Statistical Analysis of Scaled Circuits, Vol. 46, Springer Science & Business Media, 2009.

34. Singhee, A. and R. A. Rutenbar, “Why Quasi-Monte Carlo is Better Than Monte Carlo or Latin Hypercube Sampling for Statistical Circuit Analysis", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 29, No. 11, pp. 1763-1776, 2010.

35. Hocevar D., M. Lightner, and T. Trick, “A Study of variance reduction techniques for estimating circuit yields”, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. CAD-2, No. 3, pp.180-192, 1983.

36. Singhee A., and R. A. Rutenbar, “Statistical blockade: A novel method for very fast Monte Carlo simulation of rare circuit events, and its application”, Proceedings of the 2007 Design, Automation & Test in Europe Conference & Exhibition, pp. 1-6, 2007.

37. Qazi M., M. Tikekar, L. Dolecek, D. Shah, and A. Chandrakasan, “Loop flattening and spherical sampling: Highly efficient model reduction techniques for SRAM yield analysis”, Proceedings of the 2010 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 801-806, 2010.

38. Kanj R., R. Joshi, and S. Nassif, “Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events”, Proceedings of the 43rd Annual Design Automation Conference, pp. 69-72, 2006.

39. Sun S., X. Li, H. Liu, K. Luo and B. Gu, "Fast statistical analysis of rare circuit failure events via scaled-sigma sampling for high-dimensional variation space", Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 34, No. 7, pp. 1096-1109, 2015.

40. Agarwal K., and S. Nassif, “Statistical analysis of SRAM cell stability”, Proceedings of the 43rd Annual Design Automation Conference, pp. 57-62, 2006.

41. Bayrakci A., A. Demir, and S. Tasiran, “Fast Monte Carlo estimation of timing yield with importance sampling and transistor level circuit similation”, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 29, No. 9, pp. 1328-1341, 2010.

42. Wang J., A. Singhee, R. A. Rutenbar, and B. H. Calhou, “Two fast methods for estimating the minimum standby supply voltage for large SRAMs”, Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on, Vol. 29, pp. 1908-1920, 2010.

43. Rubinstein R. Y., and Kroese D. P., *Simulation and the Monte Carlo method*, John Wiley and Sons, 2011.

44. Owen A., Y. Zhou, “Safe and Effective Importance Sampling”, Journal of the American Statistical Association, Vol. 95, No. 449, 2000.

45. Bishop C.M., *Pattern recognition and machine learning*, Springer, 2006.

46. Kullback S., "Letter to the Editor: The Kullback–Leibler distance". The American Statistician, Vol. 41, pp. 340–341, 1987.

47. Rubinstein R. Y., and Kroese D. P., *The cross-entropy method: a unified approach to combinatorial optimization, monte-carlo simulation and machine learning,* Springer-Verlag, 2004.

48. Dolecek L., M. Qazi, D. Shah, and A. Chandrakasan, “Breaking the simulation barrier: SRAM evaluation through norm minimization”, Proceedings of the 2008 IEEE/ACM International Conference on Computer-Aided Design, pp. 322-329, 2008.

49. Katayama H. T. H. O. K., S. Hagiwara, and T. Sato, “Sequential importance sampling for low-probability and high-dimensional SRAM yield analysis”, Proceedings of the 2010 IEEE/ACM International Conference on Computer-Aided Design, pp. 703-708, 2010.