Efficient Yield Analysis for SRAM and Analog Circuits using Meta-Model based Importance Sampling Method

ABSTRACT

Performance failure has become the major threat to the robustness and reliability of various memory and analog circuits. It is challenging to accurately estimate the extremely small failure probability when failed samples are distributed in multiple disjoint failure regions. In this paper, we develop a novel meta-model based importance sampling (MIS) method. MIS utilizes Gaussian Process meta-model to construct quasi-optimal importance sampling distribution, and performs Markov Chain Monte Carlo (MCMC) simulation to generate new samples from the proposed distribution. By updating our global Importance Sampling estimator in an iterated framework, MIS leads to better efficiency and accuracy. For SRAM bit cell with single failure region, MIS uses 4-6X fewer samples and reaches better accuracy when compared to several recent methods. For a two-stage amplifier circuit with multiple failure schemes, MIS is 213X faster than MC without compromising accuracy, while other methods fail to cover all failure regions in our experiment.

CCS CONCEPTS

• **Hardware → Failure prediction**;

KEYWORDS

Process Variation, Failure Probability, Meta-Model, Adaptive Importance Sampling

1 INTRODUCTION

As microelectronic devices shrink to nano-meter scale, variation has become a area of growing concern due to the uncertainty during integrated circuit (IC) manufacturing. For highly replicated standard cells, or critical circuit modules, such as multi-stage amplifiers, an extremely rare failure event could cause a catastrophe for the entire chip.

In general, conventional methods such as worst-case analysis cannot work because deterministic analysis is infeasible. Modern stochastic circuit simulation approaches consider process variation, and estimate the underlying statistical information to evaluate the probability that a circuit does not meet the target performance metric. Among these approaches, standard Monte Carlo (MC) method is regarded as the golden standard, which repeatedly collects samples and perform transistor-level simulations. However, MC is extremely time-consuming under the "rare-event" scenario because we need to perform millions of simulations to capture one single failure event.

Prior Work. In order to avoid expensive MC simulation runs, more efficient approaches have been proposed to collect samples from the likely-to-fail regions, which can be categorized into two major groups:

(1) Classification: In order to speed up the failure probability estimation, classification based methods seek to construct classifier and filter out more likely-to-fail MC samples. For example, Statistical Blockade (SB) [1] applies a classifier to only simulate the rare samples in the tail distribution, and proposes a safety margin to decrease classification error. More recently, recursive SB [2] and REscope [3] improve the classifier to conditional classifier and SVM non-linear classifier, respectively. However, training such classifiers is expensive for high-dimensional circuit cases and the effectiveness deteriorates very quickly as failure rate decreases.

(2) Importance Sampling: As a classic modification of MC method, Importance Sampling (IS) method samples from a "distorted" distribution that assembles the failure region. For example, Mixture Importance Sampling (MixIS) [4] mixes a uniform distribution, original distribution and a shifted one as the optimal sampling density. Norm Minimization (MNIS) [5] and Spherical Sampling (SS) [6] methods spherically search the parametric space, and then shift the sampling distribution toward the minimum L_2-norm point. In order to deal with multi-failure-region circuit problems, methods in [7, 8] attempt to construct multiple shift vectors and perform mixture importance sampling. However, these approaches are highly dependent on the choice of sampling distribution. This makes them vulnerable to poor initialization conditions.

Among others, various strategies have been proposed to modify the conventional static IS sampling distribution. For instance, Particle Filter [**?**] performs a resampling process to accelerate failure region exploration and failure rate evaluation. Adaptive Impotance Sampling (AIS) [9] method takes one step forward. It improves the resampling iterations with an unbiased estimator which can eliminate the time-consuming static IS simulations. However, these modified IS methods suffer from sample diversity degeneracy. The resampling scheme tend to sample from the regions with higher

importance, at the expense of ignoring less important regions. This property leads to biased sampling density, which makes estimated failure probability smaller.

Paper Contributions. In this paper, we propose a novel and efficient meta-model based importance sampling method to tackle the challenging high-sigma yield analysis problem. The specific contributions include:

- We propose an effective Gaussian Process (GP) meta-model to reduce expensive SPICE simulations. By iterative calibration on the indicator function, our model is extremely accurate on the failure region boundaries. For circuits with multiple failure schemes, our GP meta-model can be parallelized to construct a set of failure boundaries independently, which guarantees failure region coverage.
- We present a novel approach to build up quasi-optimal importance sampling density. We prove that sampling from this density is efficient and it can provide unbiased estimation.
- We develop an adaptive sampling scheme that can explore the entire parameter space. The sampling procedure consists a series of augmented sampling iterations and successively update failure probability estimation. Moreover, this sampling scheme is robust to different initialization states.

The remainder of this paper is organized as follows. In Section 2, the rare event analysis problem and static IS approach are briefly reviewed. In section 3, we present the proposed MIS algorithm in detail, including Gaussian Process assisted density approximation, adaptive sampling strategy, and the analysis of MIS estimator. Experimental results are showed in Section 4 to validate the accuracy and efficiency of proposed method. Finally, we conclude this paper in section 5.

2 BACKGROUND

2.1 Rare Event Analysis

Define $\boldsymbol{x}$ as the d-dimensional random process variation, representing design parameters, such as transistor channel length/width offset, resistance/capacitance values, bias current/voltage, etc. Generally, this stochastic behavior of variation $\boldsymbol{x}$ is provided by manufacturer. Without loss of generality, in this paper, the multivariate probability density function (PDF) of $\boldsymbol{x}$ is expressed as mutually independent Gaussian distribution:

$$f(\boldsymbol{x}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}x_i^2\right) \tag{1}$$

We emphasize that this PDF assumption is not a necessary condition. Since our proposed method is developed based on importance sampling framework, it also works well with correlated process variations or other distributions.

Then we define Y as the observed performance metric, such as memory read/write time, amplifier gain/phase, etc. This metric Y usually requires time-consuming transistor-level circuit simulations to evaluate. Figure 1 shows an illustrative mapping relationship between 2-dimensional input variable $\boldsymbol{x}$ and outputs performance metric Y. In stochastic circuit simulation, it is of great interest to estimate the probability that $Y \in S$, where S distinguishes the threshold of required performance specification, depicted as gray cross section in Figure 1. In other words, samples become failure events if corresponding metric Y belongs to S. Thereby, we introduce indicator function $I(\boldsymbol{x})$ to identify pass/fail of samples:

$$I(\boldsymbol{x}) = \begin{cases} 0, & if \quad Y \notin S \\ 1, & if \quad Y \in S \end{cases} \tag{2}$$

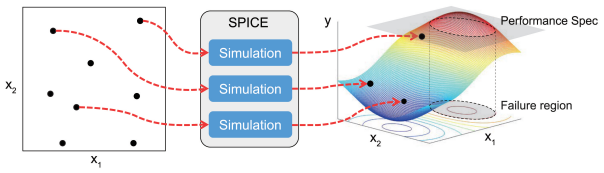

Figure 1: Mapping relationship between process variation and performance metric. The failure region boundary is determined by performance specification.

Using this indicator function $I(\boldsymbol{x})$, the probability P_{fail} can be calculated as

$$P_{fail} = P(Y \in S) = \int I(\boldsymbol{x}) \cdot p(\boldsymbol{x}) d\boldsymbol{x} \tag{3}$$

Note that the integral in Equation (3) is intractable because $I(\boldsymbol{x})$ is unknown in analytical form. Sampling based methods must be applied here to estimate failure probability. For example, MC method is regarded as ground truth, which enumerates a sample set $\{\boldsymbol{x}_i\}_{i=1}^N$ according to $p(\boldsymbol{x})$ and evaluates their indicator values $\{I(\boldsymbol{x}_i)\}_{i=1}^N$ to estimate $\hat{P}_{fail}$:

$$\hat{P}_{fail} = \hat{P}(Y \in S) = \frac{1}{N}\sum_{i=1}^{N} I(\boldsymbol{x}_i) \xrightarrow{N \to +\infty} P(Y \in S) \tag{4}$$

Here $\hat{P}_{fail}$ is an unbiased MC estimator. Its value can converge to $P(Y \in S)$ when sufficient large sample set is given.

2.2 Importance Sampling

When $Y \in S$ is a rare event, standard MC becomes inefficient because it requires millions of simulations to collect one single failure sample. To avoid massive simulations, an intuitive idea is to directly sample from the failure region. The concept of IS is to construct a "distorted" sampling distribution $g(\boldsymbol{x})$ that assembles the failure region. Figure 2 illustrates a one-dimensional example of IS method. By shift the sample mean toward more statistically likely-to-fail (important) regions, IS methods greatly improve the efficiency of collecting failure samples. Failure probability can thus be calculated as:

$$P_{fail} = P(Y \in S) = \int I(\boldsymbol{x}) \cdot \frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \cdot g(\boldsymbol{x}) d\boldsymbol{x} \tag{5}$$

$$= \int I(\boldsymbol{x}) \cdot w(\boldsymbol{x}) \cdot g(\boldsymbol{x}) d\boldsymbol{x} \tag{6}$$

where $w(\boldsymbol{x})$ denotes the importance weight between original PDF $f(\boldsymbol{x})$ and shifted PDF $g(\boldsymbol{x})$. It compensates for the discrepancy between $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$.

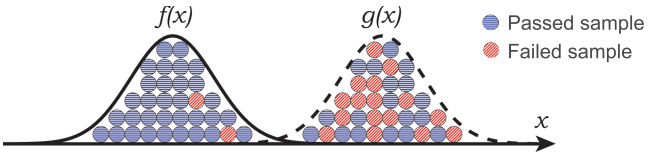

Figure 2: 1D illustrative plot of mean-shift importance sampling. Failed samples are more likely to be captured with appropriate shift vector.

Then, we can construct IS estimator by enumerating samples from shifted $g(\boldsymbol{x})$:

$$\hat{P}_{IS,fail} = \hat{P}_{IS}(Y \in S) = \frac{1}{M}\sum_{j=1}^{M} w(x_j)I(x_j) \xrightarrow{M \to +\infty} P(Y \in S) \tag{7}$$

If shifted PDF $g(\boldsymbol{x})$ is properly chosen, we note that IS sample set $\{x_j\}_{j=1}^{M}$ in Equation (7) is much smaller than MC sample set $\{x_i\}_{i=1}^{N}$ in Equation (4) because failure events in $g(\boldsymbol{x})$ is not rare. Theoretically, the optimal sampling distribution $g^{opt}(x)$ is the ideal failure event distribution, which can be expressed as:

$$g^{opt}(\boldsymbol{x}) = \frac{I(\boldsymbol{x}) \cdot f(\boldsymbol{x})}{P_{fail}} \tag{8}$$

However, $g^{opt}(\boldsymbol{x})$ cannot be evaluated with Equation (8) deterministically because the expression of $I(\boldsymbol{x})$ is unavailable in analytical form. In literature, different strategies have been proposed in order to approximate this $g^{opt}(\boldsymbol{x})$ with mean shift vector. For example, MNIS [5] shifts $f(\boldsymbol{x})$ to the pass/fail boundary, HDIS [10] shifts to the centroid of presampling failure points, and HSCS [7] shifts to multiple centroids by clustering failure samples.

However, existing mean-shift IS implementations suffer from a major drawbacks. It requires a presampling stage to determine sampling distribution. This static sampling distribution lacks of flexibility to explore the entire parametric space, especially at high dimensionality. When targeting more complicated circuit cases, the computation complexity for presampling stage increases exponentially because of "curse of dimensionality". We also emphasize that the effectiveness of these approaches are sensitive to initial conditions. Alternatively, our proposed approach applies an adaptive scheme to search for failure regions, which yields more stable performance.

3 META-MODEL BASED IMPORTANCE SAMPLING ALGORITHM

3.1 Motivation

Our motivation to design MIS algorithm is to preserve the strong points of other adaptive importance sampling algorithms (such as AIS[9] or PMC [11]), while exploiting the freedom given by meta-model to develop optimal sampling strategy. Our meta-model is initialized with Latin Hypercube Sampling (LHS) method, and updated iteratively. For each intermediate meta-model, we utilize it to approximate optimal sampling distribution, and enrich the sample set by performing MCMC sampling. We construct intermediate IS estimators for each meta-model and we use Deterministic Mixture (DM) approach to come up with the final global estimator. Compared with static IS, this DM strategy presents advantage in terms of stability and variance. And no additional IS simulations are needed to estimate failure probability.

3.2 Gaussian Process assisted Density Approximation

In order to approximate the optimal IS sampling density $g^{opt}(\boldsymbol{x})$ described in Section 2.2, our MIS method utilizes a series of iterated Gaussian Process meta-models. We choose GP model due to its flexibility for various types of circuit responses and simplicity to implement. Another good property is that it has a built-in error estimator to evaluate the variance of GP predictions, which characterizes the accuracy of model fitting. In literature, GP models have been extensively applied as emulators or surrogates for black-box functions [12].

A typical GP meta-model is defined by a mean function $\mu(\boldsymbol{X})$ and a covariance matrix $\boldsymbol{K}(\boldsymbol{X})$:

$$\mathcal{M}(\boldsymbol{X}) \sim \mathcal{GP}\left(\mu(\boldsymbol{X}), \boldsymbol{K}(\boldsymbol{X})\right) \tag{9}$$

Here $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ stands for a training set of size n, and $\boldsymbol{K}$ is a kernel matrix with expression of:

$$\boldsymbol{K} = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} \tag{10}$$

For a new sample point $\boldsymbol{x}_*$, GP outputs a prediction $[\boldsymbol{y}, y_*]$ in the form of a jointly normal distribution:

$$\begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}(\boldsymbol{X}) \\ \mu(\boldsymbol{x}_*) \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \sigma_\epsilon^2 \boldsymbol{I} & \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x}_*) \\ \boldsymbol{K}(\boldsymbol{x}_*, \boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{bmatrix}\right) \tag{11}$$

The prediction can be further formulated as a Gaussian random variable:

$$P(y_*|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) = \mathcal{N}\left(\mu(\boldsymbol{x}_*), \sigma^2(\boldsymbol{x}_*)\right) \tag{12}$$

The predictive mean and variance are as follows:

$$\mu(\boldsymbol{x}_*) = \mu(\boldsymbol{x}_*) + \boldsymbol{K}(\boldsymbol{x}_*, \boldsymbol{X})(\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \sigma_\epsilon^2 \boldsymbol{I})^{-1}(\boldsymbol{y} - \mu(\boldsymbol{X})) \tag{13}$$

$$\sigma^2(\boldsymbol{x}_*) = \boldsymbol{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{K}(\boldsymbol{x}_*, \boldsymbol{X})(K(\boldsymbol{X},\boldsymbol{X}) + \sigma_\epsilon^2 \boldsymbol{I})^{-1}\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{x}_*) \tag{14}$$

Our MIS algorithm utilizes aforementioned GP meta-model to approximate optimal IS sampling density $g^{opt}(\boldsymbol{x})$. First, we apply GP to construct a probabilistic classification function $\pi(\boldsymbol{x})$, which assembles the indicator function $I(\boldsymbol{x})$. We note that the value of ideal indicator function is either 0 or 1, our $\pi(\boldsymbol{x})$, on the other hand, is a continuous function. The mathematical expression of $\pi(\boldsymbol{x})$ is written as:

$$\pi(\boldsymbol{x}) = \Phi\left(\frac{\mu(\boldsymbol{x}) - T}{\sigma(\boldsymbol{x})}\right) \tag{15}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution.

Next, we introduce quasi-optimal sampling distribution $g(\boldsymbol{x})$ defined as:

$$g(\boldsymbol{x}) = \frac{\pi(\boldsymbol{x})f(\boldsymbol{x})}{\int \pi(\boldsymbol{x})f(\boldsymbol{x})d\boldsymbol{x}} \tag{16}$$

We notice that Equation (16) has the exact same format as $g^{opt}(\boldsymbol{x})$, but in a probabilistic perspective. For a particular GP meta-model, sampling from corresponding quasi-optimal distribution has the

largest possibility to obtain "important" samples, which locate on failure region boundaries.

3.3 Adaptive Sampling Strategy

According to Equation (16), the effectiveness of sampling distribution is strongly dependent on the accuracy of our GP meta-model. However, it is challenging to directly train such GP that can characterize indicator function $I(\boldsymbol{x})$. And the computational cost explodes as circuit complexity increases. To address this issue, we propose to improve our model accuracy through a stepwise adaptive sampling strategy.

Our sampling scheme is initialized with LHS method. It is extensively applied as a space-filling sampling method, which samples from evenly partitioned multidimensional parametric space.

Next, our MIS adaptive sampling strategy is depicted in Figure 3. It consists of two major steps. In the sample propagation step, at certain iteration t, we first construct quasi-optimal intermediate sampling distribution $g_t(\boldsymbol{x})$ based on our GP model. Next, we implement MCMC simulation to choose the next most informative sample point $\boldsymbol{x}_T$. The Metropolis-Hasting sampler we applied here is detailed in Algorithm 1. We note that generation and acceptance of samples are independently performed, which enables us to obtain vectorized result. In the following model calibration step, we enrich our training sample set with $\boldsymbol{x}_T$, and calibrate our meta-model, as shown in Figure 3(b). Our adaptive sampling procedure iterates between these two steps until the failure probability estimation converge to a stable value with specific confidence interval. In this way, our GP meta-model is refined immediately after sample propagation, providing highest flexibility to explore rare failure regions with limited budget.

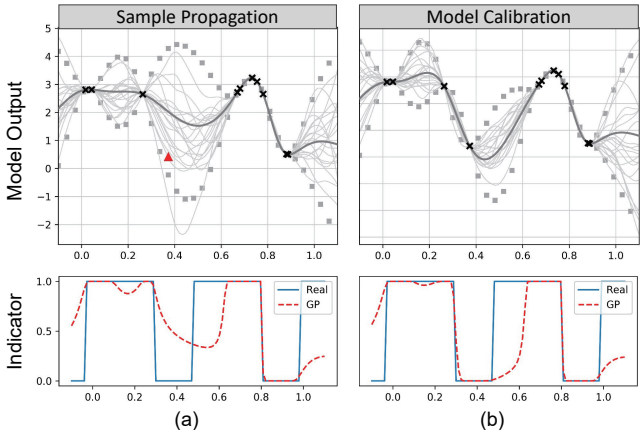

Figure 3: The calibration of GP model through adaptive sampling strategy. The dark crosses are training sample set, and red triangle represents the next sample from MCMC simulation. The dark gray lines show the mean value of GP models, and light gray lines indicate posterior distributions of trained GP model. Dashed lines mark the 95% confidence interval of GP model, which is proportional to prediction variance.

Algorithm 1: Metropolis-Hastings sampler

Input: Initial sample $\boldsymbol{x}^{(0)}$, proposal normal PDF $p(\boldsymbol{x})$, length of Markov chain M, target PDF $g_t(\boldsymbol{x})$
Output: $\boldsymbol{x}_t$
$i = 0$
while $i \leq M$ **do**
  generate a new sample $\boldsymbol{x}^* \sim p(\cdot|\boldsymbol{x}^{(i)})$
  calculate acceptance rate

$$\alpha^{(i+1)} = \min\left(1; \frac{g_t(\boldsymbol{x}^*)p(\boldsymbol{x}^{(i)}|\boldsymbol{x}^*)}{g_t(\boldsymbol{x}^{(i)})p(\boldsymbol{x}^*|\boldsymbol{x}^{(i)})}\right)$$

  generate $u \sim U[0,1]$
  if $u \leq \alpha^{(i+1)}$ **then**
    accept new sample $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^*$
  else
    reject new sample $\boldsymbol{x}^{(i+1)} = \boldsymbol{x}^{(i)}$
  end
end
return $\boldsymbol{x}_t \sim g_t(\boldsymbol{x})$

3.4 MIS Estimator Analysis

The major contribution of our MIS algorithm is that we develop a novel estimator. It is generated by updating a series of intermediate estimators, which can successively provide estimation from a cloud of iterative sampling distributions. Compared with sampling from a single IS distribution, we can eliminate the time-consuming static IS procedure, and it can accurately locate failure region boundaries.

Based on our sampling strategy described in section 3.3, until certain iteration T, sample set $\{\boldsymbol{x}_t\}_{t=1}^T$ have been generated from all the previous sampling distributions $\{g_t(\boldsymbol{x})\}_{t=1}^T$. For current sample point $\boldsymbol{x}_T$, we perform DM strategy to construct its intermediate estimator. First, we construct the mixture density by averaging previous sampling distributions $\frac{1}{T}\sum_{t=1}^T g_t(\boldsymbol{x}_T)$. Thereby, its importance weight $w(\boldsymbol{x}_T)$ is defined as the ratio between original probability density $f(\boldsymbol{x}_T)$ and mixture density:

$$w(\boldsymbol{x}_T) = \frac{f(\boldsymbol{x}_T)}{\frac{1}{T}\sum_{t=1}^T g_t(\boldsymbol{x}_T)} \tag{17}$$

This importance weight can actually be regarded as the extension of conventional static IS importance weight defined in Equation (5). Thus the intermediate estimator $\hat{P}_{fail,T}^{MIS}$ for iteration T is written as:

$$\hat{P}_{fail,T}^{MIS} = \frac{1}{T}\sum_{t=1}^T w(\boldsymbol{x}_t)I(\boldsymbol{x}_t) \tag{18}$$

As iteration continues, this intermediate estimator keeps updating until the termination of sampling procedure, giving final global estimation on failure probability.

3.4.1 ***Adaptation of MIS Estimator.*** The adaptive mechanism of MIS is driven by the uncertainty from the intermediate estimators. To be specific, each sample $\boldsymbol{x}_t$ comes from a corresponding distribution $g_t(\boldsymbol{x})$ that can describe local features of the original distribution $f(\boldsymbol{x})$. After we perform SPICE simulation on typical sample point and use it to calibrate our meta-model, a random walk

will be generated toward regions of higher probabilities. That is, regions with larger mismatches or regions with higher importance. As our meta-model is becoming more and more accurate, our sampling distribution heuristically approaches failure boundary, where the importance weight $w(\boldsymbol{x_T})$ is largest. Thus more observations are added to the sample set, and our final sampling distribution will be very similar to the ideal failure event distribution $g^{opt}(\boldsymbol{x})$.

Another good property of our MIS estimator is that the adaptation procedure is independent of multiple families of sampling distributions. It means that MIS has the potential to allow multiple meta-models adapting in parallel, in order to explore different regions. Therefore, our MIS method can tackle circuit cases with complex failure regions, or with various failure mechanisms.

3.4.2 ***Unbiasedness and Variance of MIS Estimator.*** In this section, we first prove that our MIS estimator $\hat{P}_{fail}^{MIS}$ is unbiased as iteration proceeds. The unbiasedness can be validated by its expected value, which guarantees $\hat{P}_{fail}^{MIS}$ converge to the ground truth failure probability:

$$\begin{aligned} E\left[\hat{P}_{fail}^{MIS}\right] &= E\left[\frac{1}{T}\sum_{t=1}^{T} w(\boldsymbol{x}_t) I(\boldsymbol{x}_t)\right] \\ &= \frac{1}{T}\sum_{t=1}^{T} E\left[\frac{f(x)I(x)}{\frac{1}{T}\sum_{t=1}^{T} g_t(x)}\right] \\ &= \frac{1}{T}\sum_{t=1}^{T} \int \frac{f(x)I(x)}{\frac{1}{T}\sum_{j=1}^{T} g_j(x)} g_t(x)dx = P_{fail} \end{aligned} \tag{19}$$

Next, we demonstrates that proposed MIS estimator is strictly superior or equal to static IS estimator in terms of lower variance. The worst-case scenario of MIS degenerates into static IS, which occurs at iteration $T = 1$. Mathematically, the variance of two estimators is given by:

$$Var(\hat{P}_{fail}^{MIS}) = \frac{1}{T^2}\sum_{i=1}^{T}\left(\int \frac{f^2(\boldsymbol{x})I^2(\boldsymbol{x})}{\frac{1}{T}\sum_{j=1}^{T} g_j(\boldsymbol{x})} - P_{fail}^2\right) \tag{20}$$

$$Var(\hat{P}_{fail}^{IS}) = \frac{1}{T^2}\sum_{i=1}^{T}\left(\int \frac{f^2(\boldsymbol{x})I^2(\boldsymbol{x})}{g_i(\boldsymbol{x})} - P_{fail}^2\right) \tag{21}$$

By subtracting Equation (20) and Equation (21), we prove that $Var(\hat{P}_{fail}^{MIS})$ is always smaller by deriving the following inequality:

$$\begin{aligned} \frac{1}{\frac{1}{T}\sum_{j=1}^{T} g_j(\boldsymbol{x})} &= \frac{1}{\frac{T-1}{T}\frac{1}{T-1}\sum_{i=1}^{T-1} g_i(\boldsymbol{x}) + \frac{1}{T} g_T(\boldsymbol{x})} \\ &\le \frac{(T-1)/T}{\frac{1}{T-1}\sum_{i=1}^{T-1} g_i(\boldsymbol{x})} + \frac{1/T}{g_T(\boldsymbol{x})} \\ &\le \frac{T-1}{T}\frac{1}{T-1}\sum_{i=1}^{T-1}\frac{1}{g_i(\boldsymbol{x})} + \frac{1}{T}\frac{1}{g_T(\boldsymbol{x})} \\ &= \frac{1}{T}\sum_{i=1}^{T}\frac{1}{g_i(\boldsymbol{x})} \end{aligned} \tag{22}$$

With the unbiased estimation and lower variance, our MIS method exhibits better stability and convergence. It is validated by the accuracy and efficiency comparison in the next section.

4 EXPERIMENT RESULT

In this section, our proposed yield analysis algorithm is first tested on a typical 6T SRAM bit cell with 36 variables. For analog yield analysis, we then validate our MIS on a two-stage amplifier with 84 variables. We implement MC as ground truth for accuracy comparison. To show the efficiency of MIS, we also implement several state-of-the-art approaches including Hyperspherical Clustering and Sampling (HSCS) [7] and Adaptive Importance Sampling (AIS) [9]. The SPICE model is SMIC 40nm transistor model. All the experiments are performed on Linux server with Intel Xeon X5675 CPU @3.07 GHz and 94 GB RAM.

4.1 Experiments on 6T SRAM Bit Cell

The schematic of typical 6T SRAM bit cell is shown in Figure 4. Four transistors MP1, MN2, MP3 and MN4 form two cross-coupled inverters and use two steady states (either '0' or '1') to store data in this cell. The other two transistors MN5 and MN6 work as switches to control access to the storage cell during reading, writing and standby operations. Taking reading failure as an example, it occurs when the voltage difference between BL and BLB is too small to be captured by sense amplifier in a certain period. The performance of the circuit is characterized by the delay of discharging bitline, which should be smaller than a given threshold for reading success. We implement different methods to compare their accuracy and efficiency.

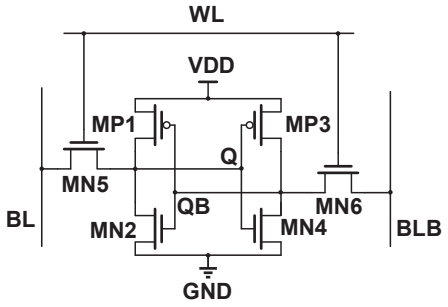

Figure 4: The schematic of typical 6T SRAM Bit Cell

4.1.1 Accuracy Comparison. To evaluate the accuracy of difference methods, we introduce Figure of Merit (FOM), ρ, to characterize the convergence and confident interval of our estimation. Its definition is:

$$\rho = \frac{\sqrt{\sigma^2_{\hat{p}_{fail}}}}{\hat{p}_{fail}} \tag{23}$$

where $\hat{p}_{fail}$ represents the failure probability and $\sigma_{\hat{p}_{fail}}$ denotes its standard deviation. To clarify, if an estimator terminates with $\rho \le \epsilon\sqrt{log(1/\delta)}$, we can claim that $\hat{p}_{fail}$ is $(1-\epsilon)\times 100\%$ accurate with $(1-\delta)\times 100\%$ confidence. To guarantee the accuracy of estimation, we set $\rho = 0.1$ as the convergence criterion to reveal the

estimation reaches 90% accuracy level with 90% confidence interval. It is depicted as dashed line in Figure 5. We observe that the estimation of HSCS, AIS and proposed MIS all succeed to match the ground truth value when sufficient simulations are available.

More detailed comparison of those approaches is illustrated in Table 1. We note that the proposed MIS algorithm provides the most accurate estimation with only 3.2% relative error, while HSCS and AIS gain 4.6% and 6.1% relative error, respectively. It is because MIS iteratively improves the optimal sampling distribution by searching for failure region, and accurately determine the failure boundaries through Gaussian Process model to avoid misclassified samples.

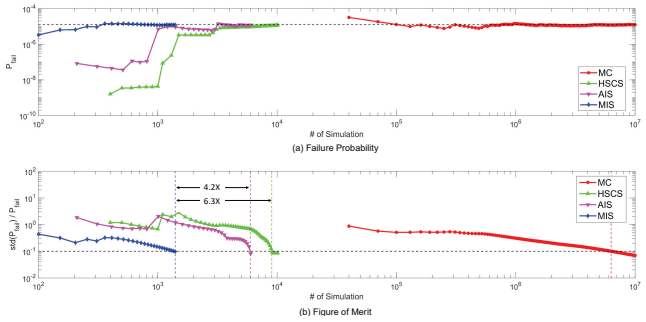

Figure 5: Evolution comparison of failure prob. and FOM on SRAM bit cell

Table 1: Accuracy and efficiency comparison on 36 dimensional SRAM bit cell

	MC	HSCS	AIS	MIS
Failure prob.	1.24e-5	1.18e-5	1.32e-5	1.20e-5
Relative error	golden	4.6%	6.1%	3.2%
# Sim. runs	6.3e6	9019	5962	1413
Speedup	1X	698X	1056X	4458X

4.1.2 Efficiency Comparison. Figure 2 shows the efficiency of MC, HSCS, AIS and MIS. The proposed MIS algorithm can provide fastest convergence. It is attributed to MIS utilizing an adaptive sampling strategy, which is far more efficient than static sampling method in HSCS and resampling scheme in AIS. As detailed in Table 1, HSCS and AIS require 9019 and 5962 samples converge to golden reference, respectively. Contrasting to these methods, the proposed algorithm exhibits very competitive estimation with only 1413 simulations. As a result, we can achieve 4458X, 6.4X and 4.2X speedup over MC, HSCS and AIS method, respectively.

4.2 Experiments on Two-Stage Amplifier

In this section, we verify that the proposed method is capable of handling problems on a multi-performance analog circuit. Figure 6 shows the circuit schematic of two stage operational transimpedance amplifier (OTA) using master-slave structure for low supply voltage application. The slave stage consists of the tail current transistor (*MP5*), the differential pair (*MP1* and *MP2*) and the current-mirror load (*MN1* and *MN2*). The master stage replicates the tail current source and the transconductance transistor of the slave stage circuit (i.e. *MP3*, *MP4*, *MP6* and *MN3* are the copies of *MP1*, *MP2*, *MP5* and *MN1*, respectively). *MP5* and *MP6* operates in linear region to save voltage margin. A robust OTA design should satisfy multiple specification requirements. In our experiments setting, we consider various performance specifications, including voltage gain margin, gain bandwidth, phase margin and 3dB bandwidth. There are in total 84 variation parameters in this case.

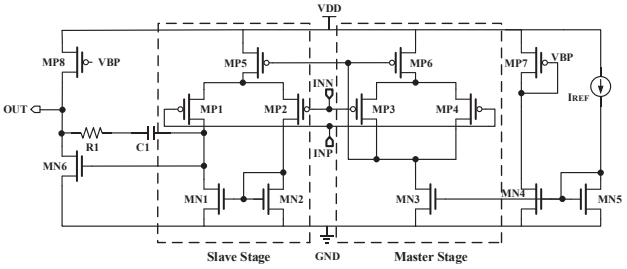

Figure 6: The schematic of two-stage operational transimpedance amplifier

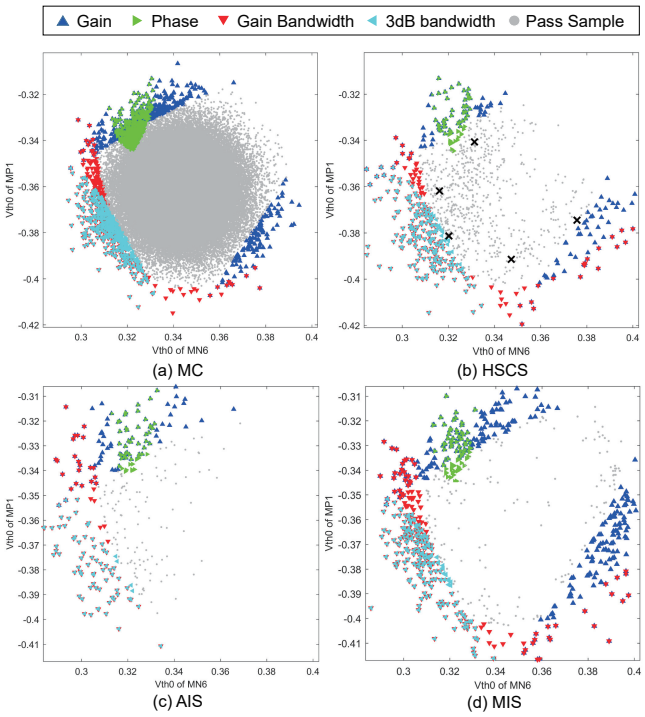

Figure 7: Multiple failure regions coverage comparison. (samples with different failure schemes are colored triangles, dark crosses in HSCS denote the shifted mean)

4.2.1 Comparison of Visualized Failure Regions. In order to compare the capability of locating multiple failure regions, we project the sample points onto two most important dimensions, which are

the threshold voltage of MN6 and MP1. Figure 7 shows the visualized sample points in 2D parametric space for different methods. Here gray dots denote passed samples and multi-colored triangles are samples with different failure schemes. Based on ground truth MC sampling result shown in Figure 7(a), we notice that the failure regions are overlapped and have complex non-convex boundaries.

In Figure 7(b), HSCS method performs Kmeans algorithm to group samples into clusters, and generate a set of min-norm points as centroids for each cluster. Then it performs multiple spherical sampling to shift sample mean to these min-norm points. Although it is able to cover the majority of failure regions, the spherical sampling strategy lacks flexibility to characterize the boundaries. Also, we observes that large proportion of samples fall in the passed regions, which makes it inefficient.

Figure 7(c) displays the failure samples collected with AIS method. We notice that AIS cannot detect all the failure regions, which gives relatively smaller failure probability. This property is caused by weight degeneracy during resampling procedure. As iteration proceeds, samples with smaller weights are more likely to be neglected, which lead to missing failure regions on the right hand side.

As shown in Figure 7(d), our MIS method considers each performance metric separately. We first build four initial GP meta-models between each performance metric and input variation vector in a parallel fashion. Next, for each meta-model, we approximate its failure event indicator function $\{I_i^{(0)}(\boldsymbol{x})\}_{i=1}^4$ and generate intermediate sampling distribution $\{g_i^{(0)}(\boldsymbol{x})\}_{i=1}^4$. At each iteration t, four samples are collected from each of intermediate distributions, formulating a mixture sampling distribution $G^{(t)}(\boldsymbol{x})$ by averaging out $\{g_i^{(t)}(\boldsymbol{x})\}_{i=1}^4$. This concept is actually very similar to the methodology in Kernel Density Estimation (KDE). As displayed in Figure 7(d), the whole sample set of MIS is generated by incrementally sampling from all the previous intermediate sampling distribution $g_i^{(t)}(\boldsymbol{x})$. We notice that the MIS sample set is able to capture all the boundaries for different failure schemes, preventing biased estimation on P_{fail}.

Another observation from Figure 7 is that, for this amplifier circuit with multiple failure schemes, our MIS is prone to sample from the distinct boundaries for different performance metrics. That is, our formulated mixture sampling distribution has slight density deviation compared with optimal sampling distribution. The optimal distribution, which is ideal failure event density, is shown as colored triangles in MC. However, instead of training complex global meta-model, our experiment result verifies that training separate meta-model can speed up with order of magnitude. We notice that among all these methods, MIS requires the smallest number of sample points to guarantee all failure-region coverage. Proposed GP meta-model is extremely simple to train with minimum sacrifice in sampling efficiency.

4.2.2 Accuracy and Efficiency Comparison. We evaluate proposed MIS and other methods(MC, HSCS, AIS) and compare their efficiencies and accuracies on two-stage amplifier to validate the improvement of proposed method.

Table 2 shows the analysis performance comparison of Monte Carlo, HSCS, AIS and proposed MIS. According to results of Monte Carlo, the standard failure probability of two-stage amplifier is

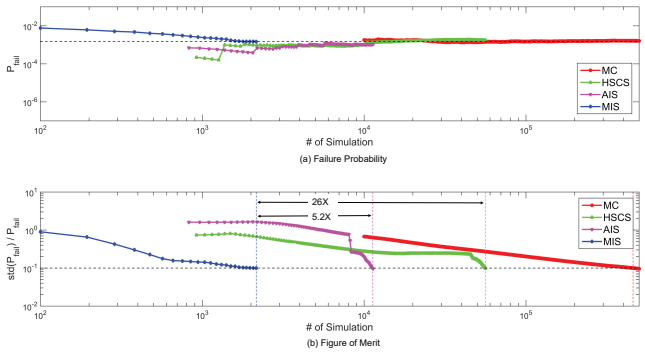

Figure 8: Evolution comparison of failure prob. and FOM on Two-stage OTA

Table 2: Accuracy and efficiency comparison on 84 dimensional Two-Stage OTA

	MC	HSCS	AIS	MIS
Failure prob.	1.5e-3	1.68e-3	1.05e-3	1.60e-3
Relative error	golden	12%	30%	6.7%
# Sim. runs	458700	56290	11316	2156
Speedup	1X	8X	41X	213X

higher than SRAM Bitcell case since its multiple performances requirements and higher dimensions induce more complex failure regions. Under this situation, the yield analysis accuracy and efficiency of both HSCS and AIS are diminished with over 10% relative error and under 100X speedup rate, while proposed method remains relatively stable performance of 8% relative error and 589X speedup. The comparison validates that MIS can be adapted to wide-range failure probability and multiple performances yield analysis. This extendability is contributed by building separate meta-models for each specifications to locate sampling areas instead of brute fail/success classification, which will be further evaluated and analyzed in next Section.

4.2.3 MIS Adaptation for Poor Initialization. In this section, we demonstrate that the adaption of MIS enables it to recover from poor initialization states. Here we take the amplifier phase margin as target metric. Figure 8(a) displays the ground truth MC contour plot, where colored contour lines represent different threshold for phase margin. Among these contour lines, the black one with 55.5 marked on it is the failure region. In Figure (b)-(e), we show the evolution of contour lines generated from intermediate GP meta-model. The incremental sample set used to calibrate model is also depicted. To be specific, Figure 8(b) shows a poor initial sample set generated from space-filling LHS method. As iteration proceeds, most samples, through a random walk, have the trend to approach the failure region boundary. Other few samples are generated near the regions where prediction mismatch is largest. After 60 iterations, our meta-model prediction is very similar compared with golden MC, which implies that it can output accurate indicator function $I(\boldsymbol{x})$. Our sampling distribution is thus stabilized around failure region boundaries, providing constant and robust estimation.

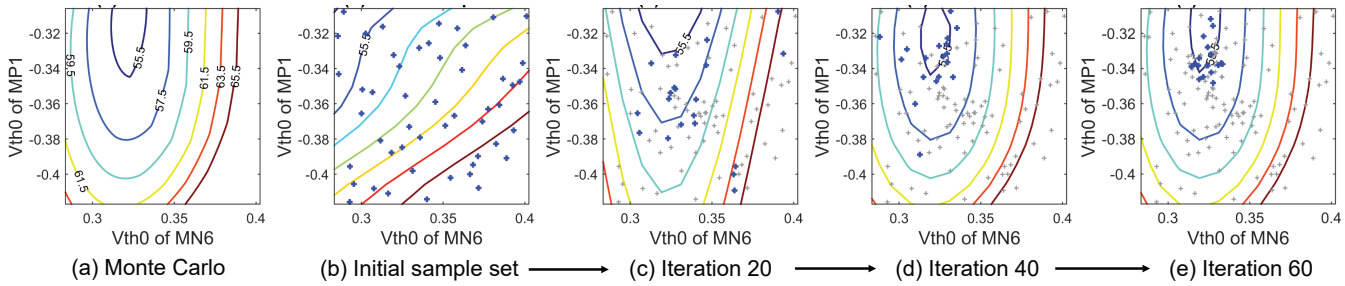

Figure 9: 2D visualized plot of proposed adaptive sampling strategy. Here contour lines denote boundaries of different OTA phase margin threshold. Blue crosses are current sample set to calibrate contour lines, while gray crosses stand for all previous samples.

5 CONCLUSIONS

In this paper, we propose a meta-model based importance sampling (MIS) to tackle the challenging circuit reliability problems with multiple disjoint failure regions. We first apply Gaussian Process meta-model to construct quasi-optimal sampling distribution. Next, we design a adaptive sampling strategy to generate new samples from the proposed distribution, and estimate the failure probability at the end of each iteration. The experimental results demonstrate that the proposed MIS algorithm can provide extremely high accuracy and efficiency. For SRAM bit cell with 36 variables, MIS achieves 4458X speedup over MC and 4-6X over other state-of-the-art methods. For 84-dimensional two-stage amplifier with multiple failure schemes, MIS is 213X faster than MC, while other approaches fail to provide a reasonable accuracy. The experimental results also shows that the potential parallelizability of the proposed MIS, which is an appealing characteristic for multiple failure mechanisms circuit problems.

REFERENCES

[1] Amith Singhee and Rob A Rutenbar. Statistical blockade: a novel method for very fast monte carlo simulation of rare circuit events, and its application. In *Design, Automation, and Test in Europe*, pages 235–251. Springer, 2008.

[2] Amith Singhee, Jiajing Wang, Benton H Calhoun, and Rob A Rutenbar. Recursive statistical blockade: An enhanced technique for rare event simulation with application to sram circuit design. In *VLSI Design, 2008. VLSID 2008. 21st International Conference on*, pages 131–136. IEEE, 2008.

[3] Wei Wu, Wenyao Xu, Rahul Krishnan, Yen-Lung Chen, and Lei He. Rescope: High-dimensional statistical circuit simulation towards full failure region coverage. In *Proceedings of the 51st Annual Design Automation Conference*, pages 1–6. ACM, 2014.

[4] Rouwaida Kanj, Rajiv Joshi, and Sani Nassif. Mixture importance sampling and its application to the analysis of sram designs in the presence of rare failure events. In *Design Automation Conference, 2006 43rd ACM/IEEE*, pages 69–72. IEEE, 2006.

[5] Lara Dolecek, Masood Qazi, Devavrat Shah, and Anantha Chandrakasan. Breaking the simulation barrier: Sram evaluation through norm minimization. In *Proceedings of the 2008 IEEE/ACM International Conference on Computer-Aided Design*, pages 322–329. IEEE Press, 2008.

[6] Masood Qazi, Mehul Tikekar, Lara Dolecek, Devavrat Shah, and Anantha Chandrakasan. Loop flattening & spherical sampling: Highly efficient model reduction techniques for sram yield analysis. In *Proceedings of the Conference on Design, Automation and Test in Europe*, pages 801–806. European Design and Automation Association, 2010.

[7] Wei Wu, Srinivas Bodapati, and Lei He. Hyperspherical clustering and sampling for rare event analysis with multiple failure region coverage. In *on International Symposium on Physical Design*, pages 153–160, 2016.

[8] Mengshuo Wang, Changhao Yan, Xin Li, Dian Zhou, and Xuan Zeng. High-dimensional and multiple-failure-region importance sampling for sram yield analysis. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 25(3):806–819, 2017.

[9] Xiao Shi, Jun Yang, Fengyuan Liu, and Lei He. A fast and robust failure analysis of memory circuits using adaptive importance sampling method. In *2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2018.

[10] Wei Wu, Fang Gong, Gengsheng Chen, and Lei He. A fast and provably bounded failure analysis of memory circuits in high dimensions. In *2014 19th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 424–429. IEEE, 2014.

[11] Olivier Cappé, Arnaud Guillin, Jean-Michel Marin, and Christian P Robert. Population monte carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.

[12] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.