
Proceedings Paper:
Liu, Y. orcid.org/0009-0001-6499-3705, He, L. orcid.org/0000-0002-5266-3805 and Xing, W.W. orcid.org/0000-0002-3177-8478 (2024) Beyond the yield barrier: Variational importance sampling yield analysis. In: Xiong, J. and Wille, R., (eds.) Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design. ICCAD '24: 43rd IEEE/ACM International Conference on Computer-Aided Design ACM , p. 36. ISBN 9798400710773

https://doi.org/10.1145/3676536.3676672

Beyond the Yield Barrier: Variational Importance Sampling Yield Analysis†

Yanfang Liu[1], Lei He[2], Wei W. Xing[3*]
[1] School of Integrated Circuit Science and Engineering, Beihang University, Beijing, China
[2] Eastern Institute of Technology, Ningbo, China, [3] SoMas, The University of Sheffield, U.K.
liuyanfang@buaa.edu.cn,lhe@eitech.edu.cn,w.xing@sheffield.ac.uk

ABSTRACT

Optimal mean shift vector (OMSV)-based importance sampling methods have long been prevalent in yield estimation and optimization as an industry standard. However, most OMSV-based methods are designed heuristically without a rigorous understanding of their limitations. To this end, we propose VIS, the first variational analysis framework for yield problems, enabling a systematic refinement for OMSV. For instance, VIS reveals that the classic OMSV is suboptimal, and the optimal/true OMSV should always stay beyond the failure boundary, which enables a free improvement for all OMSV-based methods immediately. Using VIS, we show a progressive refinement for the classic OMSV including incorporation of full covariance in closed form, adjusting for asymmetric failure distributions, and capturing multiple failure regions, each of which contributes to a progressive improvement of more than 2×. Inheriting the simplicity of OMSV, the proposed method retains simplicity and robustness yet achieves up to 29.03× speedup over the state-of-the-art (SOTA) methods. We also demonstrate how the SOTA yield optimization, ASAIS, can immediately benefit from our True OMSV, delivering a 1.20× and 1.27× improvement in performance and efficiency, respectively, without additional computational overhead.

KEYWORDS

Yield Estimation, Importance Sampling, Variational Analysis

1 INTRODUCTION

With the continual advancement of integrated circuit technology, microelectronic devices are shrinking to submicrometer scales. This trend has elevated the significance of random process variations, including intra-die mismatches, doping fluctuations, and threshold voltage variations, as critical factors in circuit design. In modern circuit designs, particularly in scenarios like SRAM cell arrays where certain cells can be replicated millions of times, addressing yield concerns has become paramount. Efficient yield estimation methods are crucial for providing accurate and rapid failure rate assessments in the presence of specific process variations.

Monte Carlo (MC) simulation, the industry-standard baseline, is commonly employed for yield estimation. MC involves running SPICE (Simulation Program with Integrated Circuit Emphasis) simulations with parameters drawn from the process variation distribution millions of times, counting failures to obtain precise estimates. However, MC is computationally intensive and becomes impractical for practical problems where the yield can be as low as 10^{-5}, a common setup in a 65nm SRAM cell array.

One pivotal avenue toward efficient yield estimation involves harnessing the potential of machine learning (ML) to construct data-driven surrogate models, approximating the unknown indicator function. Active learning techniques are then employed to iteratively refine the surrogate. Notably, [1] leverages a Gaussian process (GP) for modeling the underlying performance functions and employs an entropy reduction strategy in active learning. Absolute shrinkage deep kernel learning (ASDK) replaces the GP with a nonlinear-correlated deep kernel method and feature selection to identify crucial features for focused analysis [2]. [3] adopts a low-rank tensor approximation (LRTA) to approximate the performance function. Recently, Optimal Manifold Importance Sampling (OPTIMIS) proposes to use normalizing flow model to capture the optimal failure manifold [4]. Despite their success, surrogate-based methods are less favored due to their susceptibility to instability and the demand for data for surrogate model training. Surrogate-based methods are vulnerable to the highly nonlinear optimization problems inherent in model training, which, if not addressed correctly, can yield erroneous surrogate models and consequently inaccurate yield estimation scenarios the industry cannot afford.

Currently, the most extensively applied methods in the industrial landscape for Electronic Design Automation (EDA) tools are the Scaled-sigma Sampling (SSS) and Optimal Mean Shift Vector (OMSV)-based techniques due to their simplicity and robustness [5]. SSS generates random samples from a distorted distribution for which the standard deviation is scaled up to reduce the samples of simulations and enhance estimation efficiency [6]. OMSV-based methods employ the importance sampling (IS) technique, constructing a Gaussian distribution with OMSV as the mean and using it as the proposal distribution from which samples are drawn to accelerate yield estimation. For the OMSV-based methods, finding OMSV is most critical. Minimum Norm Importance Sampling (MNIS) identifies the Minimum Norm (MN) failure sample, a.k.a. the most probable failure point (MPFP), as the OMSV [7]. Owing to its success, many subsequent studies redirect their focus towards

*Corresponding author.

†The title of this paper is a homage to the seminal 2008 work of Lara Dolecek et al., "Breaking the simulation barrier: Sram evaluation through norm minimization" which laid the groundwork for contemporary yield analysis. The key word Beyond is twofold: 1) the optimal shift vector literately should lie beyond the failure boundary and 2) the performance of the proposed method can go beyond the classic one.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

finding MN-OMSV. Gradient Importance Sampling (GIS) enhances efficiency in finding MN-OMSV by employing gradient descent [8]. Fast Sensitivity Importance Sampling (FSIS) uses transient sensitivity analysis instead of gradient-descent to find MN-OMSV [9]. To resolve the challenge of multiple failure regions, Hyperspherical Clustering and Sampling (HSCS) employs clustering to identify them and finds the MN-OMSV for each failure region [10]. To keep MN-OMSV updated with more simulations, Adaptive Importance Sampling (AIS) introduces a dynamically updated sampling distribution, enhancing the accuracy of yield estimation [11]. By integrating the key ideas of HSCS and AIS, Adaptive Clustering and Sampling (ACS) further enhances the efficiency of yield estimation with multiple failure regions [12]. The importance of OMSV-based yield estimation is self-evident and it is *de facto* an industry standard due to its simplicity and robustness, which also lays the foundation for the state-of-the-art (SOTA) yield optimization, e.g., All Sensitivity Adversarial Importance Sampling (ASAIS) [9].

Despite their success, most progresses are heuristic in nature, it is unclear when or where their assumptions are valid. To this end, we introduce the first variational analysis framework for yield analysis: Variational Importance Sampling (VIS), which serves as an unifying framework for various SOTA methods.

Based on VIS, we discover a surprising fact: the true/optimal OMSV is not the widely used MN-OMSV first proposed in MNIS in 2008 [7], but rather, it always stays beyond the failure boundary (vs. on the failure boundary as MN-OMSV). This insight instantly grants us free improvement without extra costs for SOTA methods built on the OMSV assumption, e.g., ASAIS. VIS further reveals that MNIS and SSS are special cases of the same assumed proposal distribution, and there exists a closed-form solution where these two methods can be unified. To showcase the power of VIS, we further introduce extra refinements, including a skew normal distribution to further boost efficiency and a mixture of distributions to handle multiple failure regions. In summary, the novelty of this work includes:

(1) VIS, the first variational analysis framework for IS-based yield analysis, paving the way for the rigorous design and analysis for computational yield problems, with the following novelty as demonstration.
(2) True OMSV, the calibrated version of the canonical MN-OMSV, generating a free-lunch speedup up to 10×.
(3) Full SSS, a complete version of SSS, which admits a closed-form solution for the covariance matrix, offering another up to 2× speedup at no extra computational cost.
(4) Skew Normal (SN) OMSV, introducing asymmetric distribution to offer an extra 1.4× speedup.
(5) Mixture of Skew Normals (MSN) OMSV, which is used to handle multiple failure region challenges.
(6) The combination of (2)-(5) as a novel yield estimation method, which we call BEYOND (to suggest the importance of True OMSV and superior performance).
(7) Variational-ASAIS, demonstrating how VIS can immediately improve SOTA yield optimization, ASAIS, by 1.20× in performance and 1.27× in efficiency.
(8) The superiority of BEYOND is validated on multiple SRAM and analog circuits with thoughtful experiments, ablation study and robustness study, which demonstrate a 2.50×-29.03× speedup (9.78× on average) and a 0.11%-24.49% improvement (7.33% on average) in yield estimation accuracy.

2 BACKGROUND

2.1 Problem Definition

Let $\mathbf{x} = [x^{(1)}, x^{(2)}, \cdots, x^{(D)}]^T \in \mathcal{X}$ denote the variation variables, with $\mathcal{X}$ representing the parameter space for such variations. Typically, $\mathcal{X}$ is a high-dimensional space, denoted as D, where each element within the vector $\mathbf{x}$ signifies a specific manufacturing-related parameter affecting a circuit, such as the dimensions (length or width) of PMOS and NMOS transistors. In the context of our analysis, we make a general assumption that the elements of $\mathbf{x}$ are statistically independent and follow a Gaussian distribution:

$$p(\mathbf{x}) = (2\pi)^{\frac{D}{2}} \exp\left(-\frac{1}{2}||\mathbf{x}||^2\right). \tag{1}$$

Given $\mathbf{x}$, we can measure the performance of the circuit, denoted as $\mathbf{y}$ (e.g., metrics like memory read/write time and amplifier gain), by using SPICE simulation. Denote this as $\mathbf{y} = \mathbf{f}(\mathbf{x})$, where $\mathbf{f}(\cdot)$ represents the SPICE simulator. If $\mathbf{y}$ satisfies all pre-defined conditions $\boldsymbol{t}$, e.g., $y^{(k)} \leq t^{(k)}$ for $k = 1, \cdots, K$, then the design is considered a success; otherwise, it is a failure. Introducing an indication function $I(\mathbf{x})$ to denote the failure case, the ground-truth failure rate $\hat{P}_f$ is defined as:

$$\hat{P}_f = \int_{\mathcal{X}} I(\mathbf{x})\, p(\mathbf{x}) d\mathbf{x}. \tag{2}$$

2.2 Monte Carlo Yield Estimation

The direct calculation of the yield is intractable due to the unknown $I(\mathbf{x})$. A straightforward approach to estimate the failure rate is MC, which involves sampling $\mathbf{x}_i$ from $p(\mathbf{x})$ and evaluating the failure rate by the ratio of failure:

$$P_f = \frac{1}{N} \sum_{i=1}^{N} I(\mathbf{x}_i), \tag{3}$$

where $\mathbf{x}_i$ is the i-th sample from $p(\mathbf{x})$, and N is the number of samples. To obtain an estimate of $1-\varepsilon$ accuracy with $1-\delta$ confidence, $N \approx \log(1/\delta)/\varepsilon^2 \hat{P}_f$ is required. For a modest 90% accuracy ($\varepsilon = 0.1$) with 90% confidence ($\delta = 0.1$), we need $N \approx 100/\hat{P}_f$ samples, which is infeasible in practice for small $\hat{P}_f$, say, 10^{-5}. We can also see this intuitively from the fact that it requires on average $1/\hat{P}_f$ samples just to observe a failure event.

2.3 Importance Sampling Yield Estimation

In contrast to sampling directly from the distribution $p(\mathbf{x})$, IS-based methods utilize a proposal distribution $q(\mathbf{x})$ to draw samples and estimate the failure rate as follows:

$$P_f = \int_{\mathcal{X}} \frac{I(\mathbf{x})p(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x}) d\mathbf{x} \approx \frac{1}{N} \sum_{i=1}^{N} \frac{I(\mathbf{x}_i)p(\mathbf{x}_i)}{q(\mathbf{x}_i)} = \sum_{i=1}^{N} I(\mathbf{x}_i) w(\mathbf{x}_i), \tag{4}$$

where $\mathbf{x}_i$ are samples drawn from $q(\mathbf{x})$ and are used to approximate the integral as in MC. For convenience, we define the importance weight $w(\mathbf{x}) = p(\mathbf{x})/q(\mathbf{x})$. Eq. (4) is proved to be more efficient

than traditional MC, provided that the proposal distribution $q(\mathbf{x})$ is thoughtfully selected.

2.4 MN-OMSV

A canonical approach to design a proposal distribution $q(\mathbf{x})$ involves employing a normal distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$, effectively shifting the original Gaussian distribution centered at the origin to $\boldsymbol{\mu}$. The optimal shift vector $\boldsymbol{\mu}^*$, referred to as OMSV (a.k.a. MPFP), can be computed by solving the following optimization problem as delineated in MNIS [7]:

$$\boldsymbol{\mu}^* = \arg\min ||\mathbf{x}||^2 \quad \text{s.t.} \quad I(\mathbf{x}) = 1, \tag{5}$$

where $||\mathbf{x}||^2 = \sum_{d=1}^{D} (x^{(d)})^2$ represents the Euclidean norm. As illustrated in Fig. 1a, MNIS essentially uses the existing failure samples with the minimal norm as the OMSV to propose new samples.

3 PROPOSED APPROACH

Despite that MN-OMSV is intuitive, it was never rigorously justified in the literature. We first introduce a variational analysis framework for yield analysis, VIS, which will provide an analysis of the MN-OMSV. We then use VIS to progressively improve OMSV with rigorous mathematical analysis, which is also illustrated in Fig. 1.

3.1 Variational Analysis of IS Yield Estimation

From Eq. (4), we can see that the optimal proposal distribution $q^*(\mathbf{x})$ is the one that minimizes the approximate variance, i.e.,

$$q^*(\mathbf{x}) = \underset{q}{\operatorname{argmin}}\, \mathbb{E}_q \left[w^2(\mathbf{x}) \left(I(\mathbf{x}) - \hat{P}_f \right)^2 \right]. \tag{6}$$

Utilizing the Lagrange multiplier rule for the calculus of variations, we can show that the optimal proposal distribution is given by

$$q^*(\mathbf{x}) = p(\mathbf{x}) I(\mathbf{x}) / \hat{P}_f. \tag{7}$$

Thus, the optimal IS yield estimation is equivalent to minimizing the Kullback-Leibler (KL) divergence between the true optimal proposal distribution, $q^*(\mathbf{x})$, and its approximate counterpart, $q(\mathbf{x})$

$$\mathrm{KL}(q^*(\mathbf{x}) || q(\mathbf{x})) = \mathbb{E}_{q^*(\mathbf{x})} \left[\log q^*(\mathbf{x}) \right] - \mathbb{E}_{q^*(\mathbf{x})} \left[\log q(\mathbf{x}) \right], \tag{8}$$

with a limited number of samples. Although alternative divergence metrics (e.g., $\mathrm{KL}(q(\mathbf{x})||q^*(\mathbf{x}))$ and Wasserstein distance) exist, this work specifically employs $\mathrm{KL}(q^*(\mathbf{x})||q(\mathbf{x}))$. As we will see later, this choice yields closed-form solutions to avoid extensive hyperparameter tuning and computational overhead, a common issue with modern ML-based approaches. Notably, $\mathbb{E}_{q^*(\mathbf{x})} [\log q^*(\mathbf{x})]$ is an unknown constant denoting the entropy of the optimal proposal distribution, leaving the optimization focus on the second term. Thus, minimization of the KL divergence is equivalent to maximizing $\mathbb{E}_{q^*(\mathbf{x})} [\log q(\mathbf{x})]$, which admits an approximated solution by only keeping the failure samples

$$\int p(\mathbf{x}) I(\mathbf{x}) / \hat{P}_f \log (q(\mathbf{x}))\, \mathrm{d}\mathbf{x} \approx \frac{1}{\hat{P}_f} \sum_{i=1}^{N'} g(\mathbf{x}_i) p(\mathbf{x}_i) \log (q(\mathbf{x}_i)) \tag{9}$$

where $\mathbf{x}_i$ are failure samples, i.e., $I(\mathbf{x}_i) = 1$, N' is the number of failure samples, and $g(\mathbf{x}_i)$ is distribution that generates the samples. To better explore the failure regions, which are crucial for approximating the integral with a small number of samples, this work uses a uniform distribution, i.e., $g(\mathbf{x}_i) = 1$.

Eq. (9) is the key insight of this work—to provide a variational framework, VIS, using a numerical approximation to the ideal KL divergence. No assumption of the unknown function $I(\mathbf{x})$ is made and the approximation is exact when the number of samples N' approaches infinity.

3.2 True OMSV

Based on VIS, let's now revisit the OMSV [7] and assume that the proposal distribution is a Gaussian with a mean shift $\boldsymbol{\mu}$, i.e., $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \mathbf{I})$. Substituting this $q(\mathbf{x})$ into Eq. (9) and taking the derivative w.r.t $\boldsymbol{\mu}$, we achieve the optimal $\boldsymbol{\mu}$

$$\boldsymbol{\mu} = \frac{\sum_{i=1}^{N'} p(\mathbf{x}_i) \mathbf{x}_i}{\sum_{i=1}^{N'} p(\mathbf{x}_i)}. \tag{10}$$

This elegant closed-form solution reveals that the "True OMSV" that maximizes the objective function is the weighted average of the failure samples and **it always resides beyond the failure boundary NOT on the failure boundary**. We can also see that the importance of each failure sample decreases as it moves away from the origin, explaining why the classic MN-OMSV can still work well as a special case of using just one failure sample with the maximum weight.

3.3 Full SSS

With VIS, it is now possible to transcend the limitations of a fixed variance Gaussian distribution for the proposal distribution. The concept of employing varying variances for the proposal is not novel itself, as it has been previously explored in the pioneer work [6]. However, this method only considers a single variance scaling factor for all dimensions, thereby ignoring the correlations between dimensions and leading to suboptimal performance. Moreover, the selection of variance relies on a heuristic approach and expert knowledge, which is not practical for deployment in real-world applications. Here, we take a more ambitious step by assuming a full covariance matrix for the proposal distribution, i.e., $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$, which may seem overkill and can lead to overfitting. As we will see soon, the covariance matrix Σ will admit a closed-form solution under VIS. Substituting the proposal into Eq. (9), taking the derivative w.r.t $\boldsymbol{\mu}$ and Σ and setting them to zero, we can derive the optimal $\boldsymbol{\mu}$ and Σ. Not surprisingly, the optimal $\boldsymbol{\mu}$ is exactly Eq. (10), and the optimal Σ is

$$\Sigma = \frac{\sum_{i=1}^{N'} p(\mathbf{x}_i)(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T}{\sum_{i=1}^{N'} p(\mathbf{x}_i)}. \tag{11}$$

As a special case of the full covariance matrix, we can also derive the optimal variance for SSS by forcing $\Sigma = \sigma^2 \mathbf{I}$ and get the optimal

$$\sigma^2 = \frac{\sum_{i=1}^{N'} p(\mathbf{x}_i)(\mathbf{x}_i - \boldsymbol{\mu})^T (\mathbf{x}_i - \boldsymbol{\mu})}{\sum_{i=1}^{N'} p(\mathbf{x}_i)}. \tag{12}$$

A diagonal form of Σ can also be assumed. Since the full covariance matrix Σ is available in closed form, there is no need to use a diagonal form unless overfitting becomes an issue, which was not encountered in our experiments.

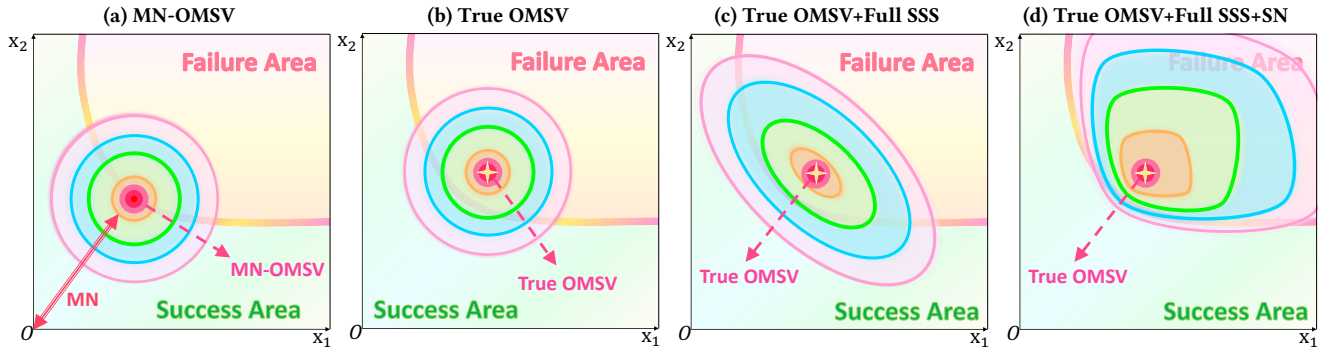

Figure 1: Illustration of progressive refinement of the classic MN-OMSV using VIS.

We can see how powerful VIS is, as it allows us to derive closed-form solutions for the optimal mean and covariance of the proposal distribution. In other words, we can derive better solutions than the conventional methods, while preserving tractability to minimize the computational costs and model complexity, which are important key merits the industry is looking for.

3.4 Skew Normal Distribution

A ubiquitous assumption made in the OMSV-based literature is that the proposal distribution is a symmetric Gaussian distribution. While this assumption is convenient for the analysis and simple enough to prevent overfitting, it can also lead to a significant reduction in efficiency by proposing many samples in the failure region (about 50%, see Fig. 1c for an example). This is intuitive to see because at least half of the samples will be generated inside the failure region, for the simplest cases with one failure region. In practice, it can get worse when the failure regions have a narrow shape. This issue is not resolved in the literature due to the lack of proper analysis tools and the difficulty in deriving a feasible solution. With VIS, we can now take a further step by amending the proposal distribution to be the multivariate skew normal distribution, which is a generalization of the normal distribution and can better fit the optimal proposal distribution (see Fig. 1d). The Probability Density Function (PDF) of the multivariate skew normal is

$$\mathcal{SN}(\mathbf{x}|\boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha}) = 2\phi(\mathbf{x}; \boldsymbol{\mu}, \Sigma)\Phi(\boldsymbol{\alpha}^T\mathbf{x}). \tag{13}$$

Here: $\phi(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ is the PDF of the normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ; $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution; $\boldsymbol{\alpha}$ is a D-dimensional vector of shape parameters. The vector $\boldsymbol{\alpha}$ determines the skewness in each dimension. When $\boldsymbol{\alpha} = \mathbf{0}$, the multivariate skew normal distribution reduces to the standard multivariate normal distribution.

To get the parameters of the multivariate skew normal distribution, we can substitute Eq. (13) into Eq. (9), take the derivative w.r.t $\boldsymbol{\mu}$, Σ and $\boldsymbol{\alpha}$ and set them to zero.

$$\underset{\boldsymbol{\mu},\Sigma,\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{N'} p(\mathbf{x}_i) \log\left(\mathcal{SN}(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha})\right). \tag{14}$$

Unfortunately, this does not lead to a closed-form solution for the parameters, which is not surprising as the estimation of the parameters in the multivariate skew normal distribution itself is a known challenge. To deliver a practical solution, we use the mean and covariance estimated from the previous sections and only optimize Eq. (14) w.r.t the shape parameter $\boldsymbol{\alpha}$ using gradient descent. This turns out to be an excellent workaround as it fits well with our motivation to have an asymmetric proposal distribution, instead of deriving a well-fitting distribution from scratch.

3.5 Mixture of Skew Normal Distribution

Finally, all simple OMSV-based methods can only deal with a single failure region, which poses a significant limitation for real-world applications. This problem can be simply resolved by introducing a mixture of skew-normal distribution, i.e.,

$$q(\mathbf{x}) = \sum_{m=1}^{M} w_m \mathcal{SN}(\mathbf{x}|\boldsymbol{\mu}_m, \Sigma_m, \boldsymbol{\alpha}_m), \tag{15}$$

where w_i is the weight and M is the number of mixture components. Substituting Eq. (15) into Eq. (9) and doing the optimization, we can derive the optimal mixture of skew normal distribution. However, this optimization is extremely difficult. As a workaround, we first cluster the failure samples using silhouette coefficient [13, 14], which automatically determines the number of clusters and the cluster label for each failure sample. The weight w_m is approximated by the number of samples in each cluster divided by the total number of failure samples. Finally, the parameters $\{\boldsymbol{\mu}_m, \Sigma_m, \boldsymbol{\alpha}_m\}$ for each cluster are optimized by Eq. (14), Eq. (11) and Eq. (10).

3.6 Complexity and Implementation

Given N as the number of failure samples, the computation of the silhouette coefficient is $O(ND)$. The computation of True OMSV and Full SSS is $O(N')$ and $O(N'D^2)$, respectively, where N' is the number of failure samples in a cluster. Updating the skew normal shape parameters is $O(N')$ each iteration. The overall algorithm is summarized in Algorithm 1. Note that the algorithm is flexible by using only the True OMSV, Full SSS.

3.7 Calibration of SOTA Yield Optimization

Many SOTA yield optimization methods rely on yield estimation by MN-OMSV [9, 15, 16], which has been revealed to be sub-optimal in this work. Nonetheless, just by using the True OMSV instead of the MN-OMSV and keeping other parts of the method unchanged, we can achieve better performance for no extra cost. We choose the latest advanced OMSV-based yield optimization method, ASAIS [9], as the baseline method. It optimizes the design parameter $\mathbf{z}$ by

Algorithm 1 BEYOND Algorithm

Require: SPICE-based indication function $I(\mathbf{x})$

1: Use Onion Sampling [4] to form initial failure samples set $\mathcal{D}$
2: **repeat**
3: Update iteration $t = t + 1$
4: Use silhouette coefficient to get M clusters and weight w_m
5: Fit each cluster with a skew normal distribution using Eq. (14), Eq. (11) and Eq. (10) and form $q(\mathbf{x})$ with Eq. (15)
6: Draw K samples from $q^t(\mathbf{x})$ and calculate importance weights: $w_k^t = I(\mathbf{x}_k)p(\mathbf{x}_k)/q^t(\mathbf{x}_k)$ for $k = 1, 2, \ldots, K$.
7: Estimate failure rate $\hat{P}_f = \frac{1}{tK} \sum_{j=1}^{t} \sum_{k=1}^{K} w_k^t$.
8: Update failure sample collection $\mathcal{D}$
9: **until** Figure of Merit (FOM), $std(\hat{P}_f)/\hat{P}_f < 0.1$
10: **return** Failure rate estimation $\hat{P}_f$

maximizing the following objective function:

$$\underset{\mathbf{z}}{\text{argmax}} \, ||\boldsymbol{\mu}(\mathbf{z})||^2, \tag{16}$$

where $\boldsymbol{\mu}(\mathbf{z})$ is the OMSV computed by Eq. (5) for design parameter $\mathbf{z}$. This optimization is solved by gradient descent with gradient $2\boldsymbol{\mu}\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{z}}$ given by adjoint method implemented in the SPICE solver. According to the True OMSV, we simply modify the computation of $\boldsymbol{\mu}(\mathbf{z})$ using Eq. (10). The gradient is then given by the weighted sum of the gradient of all failure samples. We call the modified method Variational-ASAIS.

4 EXPERIMENTAL RESULTS

In this section, we conduct a comprehensive evaluation of the accuracy and efficiency of our method, namely BEYOND, in yield estimation on three benchmark circuits: a 6T-SRAM, an operational transconductance amplifier (OTA) and a 6-bit 6T-SRAM array circuit. To ensure a SOTA comparison, we implement seven SOTA methods as comparative baselines: MNIS [7], HSCS [10], AIS [11], ACS [12], LRTA [3], ASDK [2], and OPTIMIS [4]. MC serves as the gold standard for estimating the true failure rate. We also utilize the Figure of Merit (FoM), denoted as ρ, calculated as $\rho = \text{std}(P_f)/P_f$, where $\text{std}(P_f)$ is the standard deviation of the estimated failure rate. FoM serves as the termination criterion for all methods and we use $\rho = 0.1$ following [7, 10, 15]. For the assessment, speedup is computed as $\frac{\#Sim_{MC}}{\#Sim}$, and the relative error rate is $(P_f - \hat{P}_{f_{MC}})/\hat{P}_{f_{MC}}$.

In all of our experiments, we conduct ten random seed experiments for each method (ensuring the same seeds for all methods). The final failure rate estimation is obtained by taking the average across these ten experiments. Additionally, we select the best-performing result from the ten random experiments for each method and use it to create a visualization of the iterative estimation of failure rate and its FoM. We implement the baselines with their default configurations, and where necessary, we fine-tune hyperparameters to optimize performance. All experiments are conducted on a Windows system with an AMD 7950x CPU and 32GB RAM.

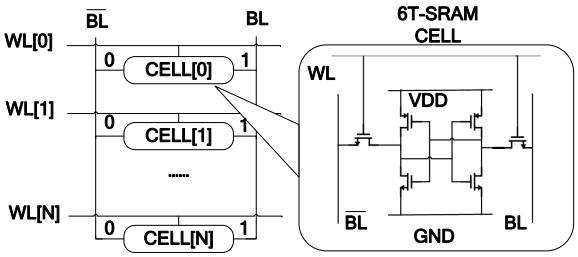

Figure 2: The structure of SRAM column circuit

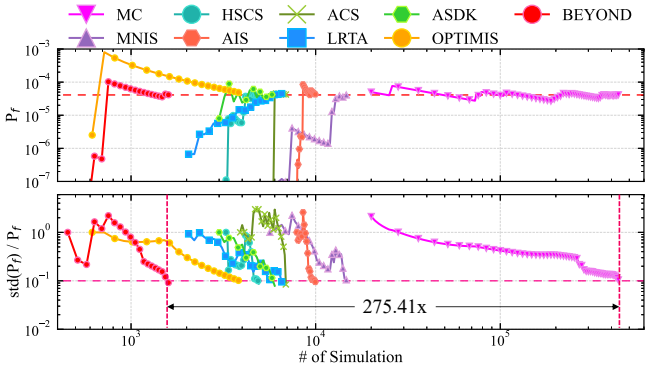

Figure 3: Failure rate estimation with FoM on 6T-SRAM

Table 1: Yield Estimation Results on 6T-SRAM

Model	Fail. Rate	Rel. Err.	# Sim	Speedup
MC	4.99e-5	-	406240	1×
MNIS	4.81e-5	3.61%	10030	40.50×
HSCS	4.86e-5	2.61%	4152	97.84×
AIS	4.85e-5	2.81%	9702	41.87×
ACS	4.70e-5	5.81%	9620	42.23×
LRTA	4.86e-5	2.61%	6130	66.27×
ASDK	4.85e-5	2.81%	6640	61.18×
OPTIMIS	4.93e-5	1.18%	3916	103.74×
BEYOND	**4.98e-5**	**0.16%**	**1564**	**259.74×**

4.1 6T-SRAM Circuit

The 6T-SRAM bit cell, illustrated in Fig. 2, is implemented in a 45nm CMOS process, which includes six transistors. Each transistor has three independent random variables: threshold voltage, mobility, and gate oxide thickness, which are critically impactful on yield among all variation parameters. As a result, the circuit encompasses 18 independent random variables. In our experiments, we focus on the delay time of SRAM read/write as the performance metric of interest.

The yield estimation experimental results are shown in Table 1, and the evolution of failure rate convergence and FoM evaluation is depicted in Fig. 3. As shown in Table 1, it is evident that BEYOND achieves the most accurate estimation with minimal simulations. In terms of accuracy, BEYOND exhibits a relative error rate as low as 0.16%, improving the accuracy by 1.02%-5.65% against other

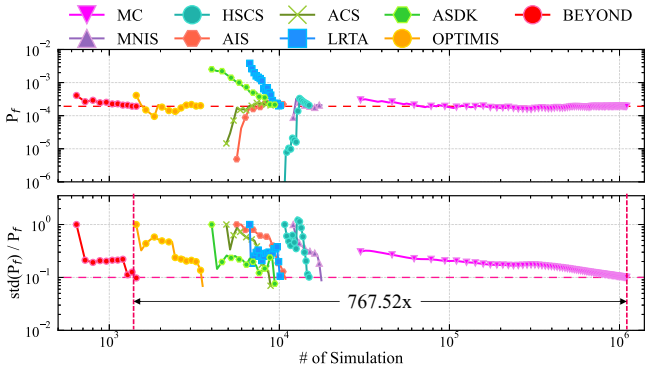

Figure 4: Failure rate estimation with FoM on OTA

Table 2: Yield Estimation Results on OTA

Model	Fail. Rate	Rel. Err.	# Sim	Speedup
MC	1.89e-4	-	1102000	1×
MNIS	1.64e-4	11.94%	21065	52.31×
HSCS	1.70e-4	10.15%	17950	61.39×
AIS	1.74e-4	8.18%	11178	98.59×
ACS	1.78e-4	5.59%	11053	99.70×
LRTA	2.04e-4	7.94%	10100	109.11×
ASDK	2.14e-4	11.68%	9600	114.79×
OPTIMIS	1.92e-4	1.57%	4126	267.09×
BEYOND	**1.90e-4**	**0.41%**	**1441**	**764.75×**

baselines. In terms of efficiency, BEYOND achieves a speedup of up to 259.74× compared to MC, and demonstrates a speedup of 2.50×-6.41× compared to other baselines.

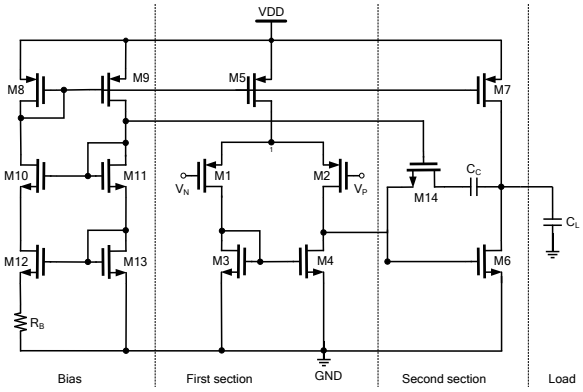

Figure 5: Operational Transconductance Amplifier Circuit

4.2 Operational Transconductance Amplifier

The OTA circuit, depicted in the Fig. 5, contains 14 transistors. Each transistor has four process variation parameters: oxide thickness, threshold voltage, and deviations in length and width due to process variations. Consequently, this circuit comprises 56 independent random variables. In our experiments, the performance of interest

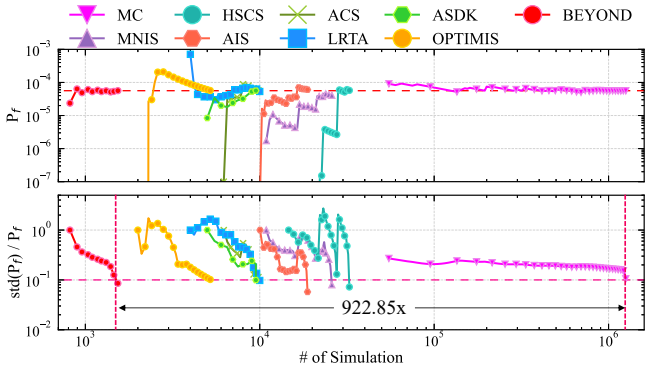

Figure 6: Failure rate estimation with FoM on 6-bit array

Table 3: Yield Estimation Results on 6-bit 6T-SRAM Array

Model	Fail. Rate	Rel. Err.	# Sim	Speedup
MC	5.62e-5	-	1417500	1×
MNIS	4.94e-5	12.03%	45174	31.38×
HSCS	4.21e-5	25.09%	47090	30.10×
AIS	4.37e-5	22.19%	15996	88.62×
ACS	4.92e-5	12.44%	14060	100.82×
LRTA	5.96e-5	6.05%	12300	115.24×
ASDK	5.87e-5	4.44%	12500	113.40×
OPTIMIS	5.66e-5	0.71%	5300	267.45×
BEYOND	**5.59e-5**	**0.60%**	**1622**	**873.92×**

is the quiescent current I_Q at 27°C. The yield estimation results are detailed in Table 2, and the evolution of failure rate convergence and FoM evaluation is illustrated in Fig. 4.

The results indicate that BEYOND consistently delivers highly accurate estimation with the reduced simulations in the analog circuit. In accuracy terms, BEYOND achieves a relative error rate as low as 0.41%, enhancing the accuracy by 1.16%-11.53% over other baselines. In efficiency terms, BEYOND realizes a speedup of up to 764.75× relative to MC and shows a speedup of 2.86×-14.62× compared to other baselines. These results underscore the robustness of BEYOND in varied circuit complexities.

4.3 6-bit 6T-SRAM Array Circuit

Building upon the BEYOND validated in the 6T-SRAM bit cell experiments, we expand to a higher complexity with the 6-bit 6T-SRAM array circuit, which has six such bit cells. This circuit contains a total of 108 variational parameters, offering a comprehensive view that incorporates peripheral circuit influences to enhance failure rate estimation accuracy. Results are captured in Table 3, and the evolution of failure rate convergence and FoM evaluation is shown in Fig. 6.

For the higher-dimensional circuit, BEYOND still maintains its leading edge. Accuracy-wise, BEYOND achieves a relative error rate as low as 0.16% and an accuracy enhancement of 0.11%-24.49% over the baselines. Efficiency-wise, BEYOND exhibits a remarkable speedup, reaching up to 873.92× compared to MC and achieving a

Table 4: Yield Optimization Comparison Results on Adder Circuit

Case	1 (Higher Specification)							2 (Lower Specification)						
Metric	Failure Rate				# Simulation			Failure Rate				# Simulation		
Method	Best	Worst	Mean	Std	Best	Worst	Mean	Best	Worst	Mean	Std	Best	Worst	Mean
WEIBO	7.50e-7	2.08e-5	6.80e-6	8.61e-6	2121	4861	3626	1.50e-5	1.90e-5	1.74e-5	1.36e-6	2681	4391	3536
MESBO	5.00e-8	1.50e-7	6.00e-8	3.00e-8	4070	11200	8640	1.03e-5	2.00e-5	1.70e-5	2.83e-6	4220	7870	5687
KDEBO	5.00e-8	1.30e-6	3.45e-7	4.54e-7	10000	10000	10000	1.10e-5	1.46e-4	5.24e-5	5.40e-5	8000	8000	8000
BYA	4.00e-8	**5.00e-8**	4.50e-8	**5.00e-9**	11000	11000	11000	1.70e-5	1.75e-5	1.72e-5	**2.29e-7**	8000	8000	8000
ASAIS	**2.50e-8**	7.50e-8	4.75e-8	2.08e-8	2417	2427	2422	9.00e-6	2.60e-5	1.83e-5	5.88e-6	2412	2420	2415
V.-ASAIS	**2.50e-8**	**5.00e-8**	**4.00e-8**	1.22e-8	**1875**	**1987**	**1912**	**7.00e-6**	**1.20e-5**	**8.50e-6**	1.50e-6	**1968**	**1992**	**1981**

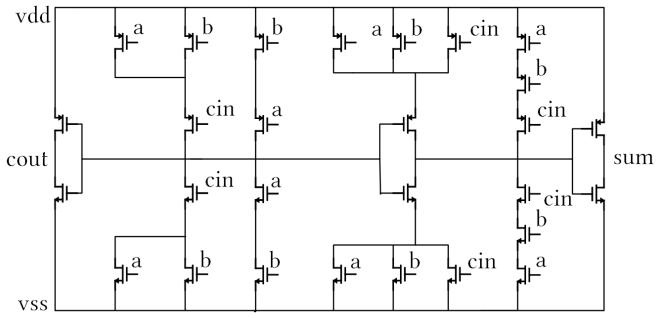

Figure 7: The structure of Adder circuit

speedup of 3.27×-29.03× compared to other baselines. These results not only reinforce the precision of BEYOND but also underscore its efficiency in managing the heightened complexity of advanced SRAM architectures.

4.4 Yield Optimization on Adder Circuit

Building upon our successful yield estimation experiments, we now focus on yield optimization through the ASAIS optimization flow [9], herein referred to as Variational-ASAIS. We benchmark Variational-ASAIS against five SOTA yield optimization methods: Weighted Expected Improvement Bayesian Optimization (WEIBO) [17], Max-value Entropy Search Bayesian Optimization (MESBO) [18], Kernel Density Estimator Bayesian Optimization (KDEBO) [19], Bayesian Yield Analysis (BYA) [1], and ASAIS for a comprehensive comparison.

We conduct the yield optimization experiments on an adder circuit, illustrated in Fig. 7. The adder circuit comprises 28 MOS transistors, each subject to three variational parameters, totaling 84 variables. Our design parameters focus on the width and length of these transistors. We assess the yield by examining the time-to-threshold (TT) performance, which involves simulating the transient response until the sum output attains a specified threshold voltage. To validate the optimization performance of each method, we conduct experiments with ten different random seeds to reduce random fluctuations (ensuring the same seeds for all methods). Furthermore, we employ two distinct circuit specifications - a higher case (Case 1) and a lower case (Case 2) - to assess the robustness of all methods. The optimal design is validated using 4e7 and 1e6 MC simulations for Case 1 and Case 2, respectively. The results are summarized in Table 4.

For Case 1 on the high specification, BYA achieves the lowest standard deviation and its worst-case result equating to that achieved by Variational-ASAIS, which consistently leads in performance. While ASAIS also posts competitive optimization results, Variational-ASAIS outstrips all baselines when considering mean performance, boasting a 1.13×-170× improvement over the other baselines. In efficiency terms, Variational-ASAIS proves to be the most resource-sparing, surpassing the baselines by 1.27× to 5.75×. For Case 2 on the lower specification, BYA continues to show a good result in the lowest standard deviation. But in other aspects, all baselines are inferior to Variational-ASAIS, which achieves a 2×-6.16× optimization performance improvement with a 1.22×-4.03× speedup compared to other baselines based on the mean results. Collectively, these findings highlight Variational-ASAIS's exceptional optimization prowess while conserving simulations.

Table 5: Comparison of Computational Time (CPU Hours)

CPU Hours	MNIS	HSCS	AIS	ACS	LRTA	ASDK	OPT.	BEYOND
6T-SRAM	4.8	2.0	4.6	4.6	3.1	3.2	2.0	**0.8**
OTA	50.5	43.0	26.8	26.5	24.3	25.0	10.0	**3.5**
6-bit Array	270.9	283.2	95.9	84.3	73.9	75.7	32.4	**9.8**

4.5 Computational Time Study

We demonstrate the computational time for the aforementioned yield estimation experiments in this section to highlight the efficiency. Table 5 presents the average computational time for ten random seed runs for each method. Clearly, BEYOND demonstrates a leading edge in computational efficiency, showing a 4.64×, 8.39×, and 13.34× speedup on average for the 6T-SRAM, OTA, and 6-bit 6T-SRAM array circuits, respectively.

Furthermore, we conduct comparative experiments on the training time (find the optimal parameters) between BEYOND and other surrogate-based methods (LRTA, ASDK, OPTIMIS). As depicted in Fig. 8, BEYOND achieves a 41.87×, 39.15×, and 8.27× speedup on average for the 6T-SRAM, OTA, and 6-bit 6T-SRAM array circuits, respectively.

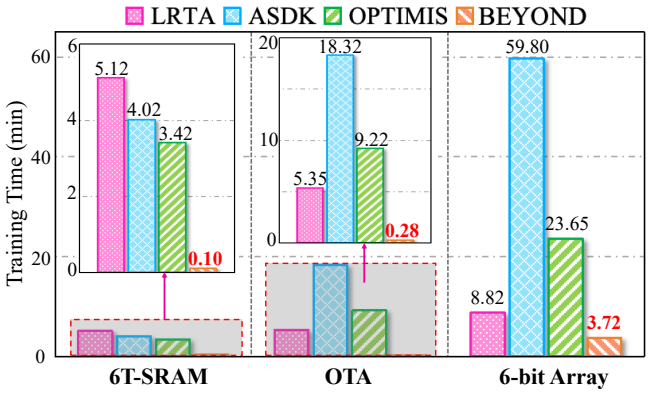

Figure 8: Model training time on three benchmark circuits

4.6 Ablation Study

To assess the contribution of each component, i.e., True OMSV, Full SSS, and the mixture of skew normal distributions, we conduct an ablation study on 6-bit 6T-SRAM array (the most challenging one among our testing examples) by incrementally integrating components into the basic MN-OMSV method. The experimental results are presented in Table 6, which reveal consistent improvement with the incorporation of each component. Fig. 9 more vividly illustrates the trend of accuracy and efficiency.

True OMSV brings the most significant improvement in both accuracy and efficiency, reducing the relative error rate by 9.02% and the number of simulations by a factor of 10.17. Full SSS further improves the accuracy by 1.63% and the efficiency by 2.01×, whereas skew normal distribution brings a marginal improvement of 0.78% in accuracy and 1.25× in efficiency. These results substantiate the superiority of BEYOND (True OMSV+Full SSS+MSN) in yield estimation.

Table 6: Ablation Study of BEYOND on 6-bit Array

Model	Fail. Rate	Rel. Err.	# Sim	Speedup
MC	5.62e-5	-	1417500	1×
MN-OMSV	4.94e-5	12.03%	45174	31.38×
True OMSV	5.45e-5	3.01%	4440	319.26×
True OMSV+Full SSS	5.54e-5	1.38%	2240	632.81×
True OMSV+Full SSS+MSN	**5.59e-5**	**0.60%**	**1622**	**873.92×**

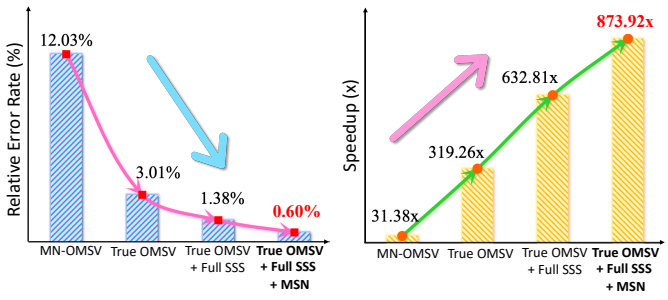

Figure 9: Ablation study of each component

Table 7: The Comparison of Incorrect Estimation Counts

Circuit	MNIS	HSCS	ACS	AIS	LRTA	ASDK	OPT.	BEYOND
6T-SRAM	**1/10**	2/10	2/10	2/10	3/10	5/10	2/10	**1/10**
OTA	3/10	3/10	3/10	3/10	4/10	7/10	4/10	**2/10**
6-bit Array	5/10	6/10	4/10	4/10	4/10	5/10	**2/10**	**2/10**

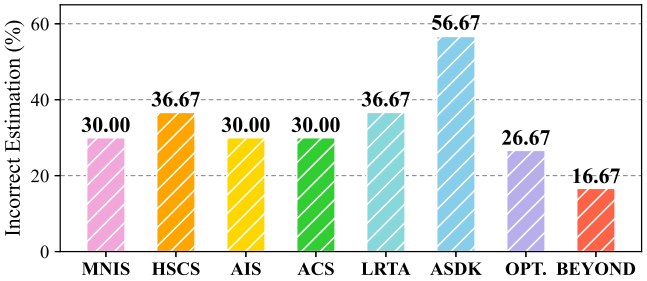

Figure 10: Incorrect estimation ratio in all experiments

4.7 Robustness Study

To highlight the robustness that is highly valued by the industry, we conduct an in-depth robustness study on the three benchmark circuits for all methods. Specifically, each method is executed with the same set of ten consecutive random seeds, and the counts of incorrect estimations—where the relative error rate exceeds 30%—is recorded for each method. The statistical outcomes are presented in Table 7.

In the 6T-SRAM experiments, OMSV-based methods generally demonstrate better stability than surrogate-based methods, with MNIS and BEYOND showing the highest stability. However, as circuit complexity increasing, as observed in the OTA and 6-bit Array experiments, the stability of all methods declines. Despite this, BEYOND continues to exhibit superior stability.

Based on the statistical results of incorrect estimations for the three circuits, the percentage of incorrect estimations for each method is depicted in Fig. 10. From the percentages of incorrect estimations, it is observable that OMSV-based methods generally exhibit higher robustness compared to surrogate-based methods, with BEYOND being the most stable among the all methods.

5 CONCLUSION

We propose VIS, a rigorous analysis framework for yield estimation, which may revolutionize the traditional yield estimation paradigm. Based on VIS, we propose BEYOND, a novel yield estimation method. The capacity of BEYOND is demonstrated by multiple modifications to the classic OMSV method in both yield estimation and optimization. BEYOND's superiority is validated by comprehensive experiments conducted on real-world circuit benchmarks, computational time studies, ablation studies and robustness studies. With the way paved by this work, we expect more innovative methods to be developed in the future. Dealing with high-dimensional circuits remains challenging, a common issue faced by all SOTA methods. We will further investigate the potential of multi-region sampling and dimensionality reduction strategies to address this issue.

REFERENCES

[1] Shuo Yin, Xiang Jin, Linxu Shi, Kang Wang, and Wei W. Xing. Efficient bayesian yield analysis and optimization with active learning. In *Proceedings of the 59th ACM/IEEE Design Automation Conference*, DAC '22, page 1195–1200, New York, NY, USA, 2022. Association for Computing Machinery.

[2] Shuo Yin, Guohao Dai, and Wei W. Xing. High-dimensional yield estimation using shrinkage deep features and maximization of integral entropy reduction. In *Proceedings of the 28th Asia and South Pacific Design Automation Conference*, ASPDAC '23, page 283–289, New York, NY, USA, 2023.

[3] Xiao Shi, Hao Yan, Qiancun Huang, Jiajia Zhang, Longxing Shi, and Lei He. Meta-model based high-dimensional yield analysis using low-rank tensor approximation. In *Proceedings of the 56th Annual Design Automation Conference 2019*, DAC '19, New York, NY, USA, 2019. Association for Computing Machinery.

[4] Yanfang Liu, Guohao Dai, and Wei W. Xing. Seeking the yield barrier: High-dimensional sram evaluation through optimal manifold. In *2023 60th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6, 2023.

[5] António Manuel Lourenc**c**o Canelas, Jorge Manuel Correia Guilherme, and Nuno Cavaco Gomes Horta. *Yield Estimation Techniques Related Work*, pages 65–95. Springer International Publishing, Cham, 2020.

[6] Shupeng Sun, Xin Li, Hongzhou Liu, Kangsheng Luo, and Ben Gu. Fast statistical analysis of rare circuit failure events via scaled-sigma sampling for high-dimensional variation space. In *Proceedings of the International Conference on Computer-Aided Design*, ICCAD '13, page 478–485. IEEE Press, 2013.

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

[8] Thomas Haine, Johan Segers, Denis Flandre, and David Bol. Gradient importance sampling: An efficient statistical extraction methodology of high-sigma sram dynamic characteristics. In *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 195–200, 2018.

[9] Wenfei Hu, Zhikai Wang, Sen Yin, Zuochang Ye, and Yan Wang. Sensitivity importance sampling yield analysis and optimization for high sigma failure rate estimation. In *2021 58th ACM/IEEE Design Automation Conference (DAC)*, pages 895–900, 2021.

[10] Wei Wu, Srinivas Bodapati, and Lei He. Hyperspherical clustering and sampling for rare event analysis with multiple failure region coverage. In *Proceedings of the 2016 on International Symposium on Physical Design*, ISPD '16, page 153–160, New York, NY, USA, 2016. Association for Computing Machinery.

[11] Xiao Shi, Fengyuan Liu, Jun Yang, and Lei He. A fast and robust failure analysis of memory circuits using adaptive importance sampling method. In *Proceedings of the 55th Annual Design Automation Conference*, DAC '18, 2018.

[12] Xiao Shi, Hao Yan, Jinxin Wang, Xiaofen Xu, Fengyuan Liu, Longxing Shi, and Lei He. Adaptive clustering and sampling for high-dimensional and multi-failure-region sram yield analysis. In *Proceedings of the 2019 International Symposium on Physical Design*, ISPD '19, page 139–146, 2019.

[13] Duy-Tai Dinh, Tsutomu Fujinami, and Van-Nam Huynh. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In Jian Chen, Van Nam Huynh, Gia-Nhu Nguyen, and Xijin Tang, editors, *Knowledge and Systems Sciences*, pages 1–17, Singapore, 2019. Springer Singapore.

[14] Ketan Rajshekhar Shahapure and Charles Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748, 2020.

[15] Jian Yao, Zuochang Ye, and Yan Wang. An efficient sram yield analysis and optimization method with adaptive online surrogate modeling. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 23(7):1245–1253, 2015.

[16] Chengzhi Jiang, Xiaoming Fan, Yan Xing, Chao Duan, and Jiaqi Zhang. Min norm failure vector guided yield optimization method for nanometer sram design. In *2019 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–4, 2019.

[17] Mengshuo Wang, Wenlong Lv, Fan Yang, Changhao Yan, Wei Cai, Dian Zhou, and Xuan Zeng. Efficient yield optimization for analog and sram circuits via gaussian process regression and adaptive yield estimation. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 37(10):1929–1942, oct 2018.

[18] Shuhan Zhang, Fan Yang, Dian Zhou, and Xuan Zeng. Bayesian methods for the yield optimization of analog and sram circuits. In *2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC)*, pages 440–445, 2020.

[19] Dennis D. Weller, Michael Hefenbrock, Michael Beigl, and Mehdi B. Tahoori. Fast and efficient high-sigma yield analysis and optimization using kernel density estimation on a bayesian optimized failure rate model. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(3):695–708, 2022.