Every Failure Is A Lesson: Utilizing All Failure Samples To Deliver Tuning-Free Efficient Yield Estimation

Wei W. Xing[3], Yanfang Liu[2], Weijian Fan[1,4], and Lei He[1,5*]
[1] Eastern Institute of Technology, Ningbo, China [3] SoMas, The University of Sheffield, U.K. [5] University of California, U.S.
[2] School of Integrated Circuit Science and Engineering, Beihang University, Beijing, China
[4] College of Mechatronics and Control Engineering, Shenzhen University, Shenzhen, China
wxingphd@gmail.com,liuyanfang@buaa.edu.cn,fanweijian2021@email.szu.edu.cn,lhe@eitech.edu.cn

ABSTRACT

Yield estimation and optimization have become increasingly important for circuit design as technology nodes scale down. Simple yet well-established minimal norm importance sampling (MNIS) still serves as an industrial standard due to its robustness and reliability. In this study, we generalize the classic MNIS and propose Every Failure Is A Lesson (EFIAL) to utilize every failure sample (instead of one in MNIS) to construct the proposal distribution. EFIAL is completely tuning-free and the update computation complexity is only $O(M)$ (M is the number of failure samples) by utilizing the blessing of dimensionality. The idea of EFIAL is then extended to the state-of-the-art (SOTA) pre-sampling method, onion sampling, to significantly boost efficiency, by up to 9.08x (4.68x on average). Extensive evaluations against SOTA yield estimation methods reveal that EFIAL achieves a speedup of up to 13.54x (5.16x on average) and an accuracy improvement of up to 24.91%.

KEYWORDS

Yield Estimation, Importance Sampling, Norm Minimization

1 INTRODUCTION

As integrated circuit technology advances, microelectronic devices are shrinking to submicrometer scales. Consequently, factors like intra-die mismatches, doping fluctuations, and threshold voltage variations, which result from random process variations, have become increasingly critical considerations in circuit design. This challenge becomes even more pronounced in contemporary circuit designs, particularly those where certain components are replicated millions of times within a single circuit, such as in the case of SRAM cell arrays. To address the rising concerns about yield, the development of efficient yield estimation methods has become pivotal. These methods aim to provide rapid and precise assessments of failure probabilities for specific circuit designs under particular process variations. The cornerstone solution in this context remains the Monte Carlo (MC) simulation, widely adopted in both industry and academia. In essence, MC involves running SPICE simulations (Simulation Program with Integrated Circuit Emphasis) for instance configurations drawn from the distribution of process variations. The yield is then estimated by counting the number of failures. It is thus evident that MC is computationally expensive and can easily become infeasible for low-yield problems, which are increasingly common in modern circuit designs.

In recent years, artificial intelligence (AI) has made significant strides, opening the door to new possibilities and solutions to resolve almost all challenges in the EDA pipeline. For yield estimation, most AI approaches rely on a data-driven surrogate model to approximate the unknown performance function. For instance, LRTA [1], a low-rank approximated polynomial chaos expansion (PCE), is proposed to approximate the performance function. Similarly, ASDK [2] and AYEBO [3] employ a Gaussian process (GP) with different strategies to deliver sequential updates of yield estimation. Recently, deep learning has also been introduced to yield estimation, e.g., [4] uses normalizing flow to fit the distribution of failure areas for yield estimation. Despite their success, this type of method requires careful model training and hyperparameter tuning, making them less attractive to the industry and pragmatists.

To deliver stable and efficient yield estimation, importance sampling (IS)-based methods are normally employed in industry. They draw samples from a proposal distribution. The main advantages of IS-based methods are that they will not introduce bias and will always converge, i.e.,, even when the designed proposal distribution is far from the ideal one, the estimation will still converge to the truth value provided enough samples. The key to improving IS-based methods is to design a proposal distribution that approximates the failure distribution. To this end, the foundational work, minimization of norm IS (MNIS), chooses a normal distribution with the mean being the closest failure sample to the origin as the proposal distribution [5]. To approximate the ideal optimal mean shift vector (OMSV) instead of choosing the closest failure sample, Gradient Importance Sampling (GIS [6]) uses gradient descent to search for the OMSV. Based on this idea, Fast Sensitivity Importance Sampling (FSIS [7]) uses sensitivity analysis to replace the gradient descent in GIS to further improve effectiveness. To further improve efficiency, adaptive importance sampling (AIS [8]) proposes to update OMSV as more data is collected. To deal with the common multi-region failure challenge, HSCS [9] proposes a cluster-based model at the pre-sampling stage. This idea is further improved by ACS [10], which introduces a multi-cone clustering for failure regions and iteratively updates its proposal distribution.

Due to their robustness and simplicity, OMSV-based methods are widely adopted in industry. However, almost all OMSV-based methods [6, 8–10] are based on a single failure sample that is closest to the origin among all failure samples, which overlooks the rich information carried by other failure samples and leads to sub-optimal performance. To resolve this issue, we propose Every Failure Is A

[*] Corresponding author.

DAC '24, June 23–27, 2024, San Francisco, CA, USA

ACM ISBN 979-8-4007-0601-1/24/06…$15.00
https://doi.org/10.1145/3649329.3657381

Lesson (EFIAL), which utilizes all failure samples to construct the proposal distribution. EFIAL inherits the advantages of MNIS as a tuning-free method and lightning-fast computation. Inspired by EFIAL, the state-of-the-art (SOTA) pre-sampling method, onion sampling, is also improved with the weighting trick to boost efficiency. The novelty of this work includes:

1. EFIAL: A generalization of MNIS utilizing every failure sample (instead of using one); it is completely tuning-free; the computation complexity is $O(M)$ (M is the number of failure samples).
2. EFIAL-onion sampling: a novel pre-sampling method by implementing EFIAL to the SOTA pre-sampling method, onion sampling.
3. Comprehensive experiments in three circuits showcase a 2.21%-24.98% enhancement in accuracy and a 2.65x-12.99x speedup in efficiency over six SOTA baseline methods.
4. EFIAL-onion sampling consistently improves all IS-based methods by 8.04%-23.28% in accuracy and 2.12x-9.08x in efficiency.

2 BACKGROUND

2.1 Problem Definition

Let $\mathbf{x} = [x^{(1)}, x^{(2)}, \cdots, x^{(D)}]^T \in \mathcal{X}$ denote variation process parameters. Each element within $\mathbf{x}$ represents the variation parameters associated with a circuit during the manufacturing process. These parameters could include quantities such as the length or width of transistors. Generally, the components of $\mathbf{x}$ are regarded as being mutually independent and Gaussian distributed, $p(\mathbf{x}) = (2\pi)^{\frac{D}{2}} \exp\left(-\frac{1}{2}||\mathbf{x}||^2\right)$. Upon having a specific configuration of $\mathbf{x}$, one can evaluate the performance of the circuit $y^{(k)}$ (e.g., memory read/write time and amplifier gain), using a SPICE simulation, $\mathbf{y} = \mathbf{f}(\mathbf{x})$. If all performance metrics satisfy certain pre-defined thresholds $t^{(k)}$, i.e., $y^{(k)} \leq t^{(k)}$ for $k = 1, \cdots, K$, the circuit is considered a success; otherwise, it is a failure.

To succinctly represent the failure status of a circuit, an indicator function $I(\mathbf{x})$ is normally introduced. Specifically, $I(\mathbf{x})$ equals 1 if the corresponding $\mathbf{x}$ leads to a design failure, and 0 otherwise. The ground-truth failure rate $\hat{P}_f$ is determined as the integral of $I(\mathbf{x})$ over the variation process parameter space $\mathcal{X}$, weighted by the probability density function $p(\mathbf{x})$: $\hat{P}_f = \int_{\mathcal{X}} I(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$. Due to the unknown nature of $I(\mathbf{x})$, the direct calculation of the yield is intractable.

2.2 Monte Carlo Yield Estimation

A prevalent strategy to approximate the failure rate is the Monte Carlo (MC) method, which draws samples $\mathbf{x}_i$ from $p(\mathbf{x})$, and subsequently approximates the failure rate by establishing the ratio of failure samples to the total number of samples. In precise terms, the estimated failure rate $\hat{P}_f$ is approximated as $P_f = \frac{1}{N}\sum_{i=1}^{N} I(\mathbf{x}_i)$, where $\mathbf{x}_i$ indicates the i-th sample obtained from $p(\mathbf{x})$, and N represents the number of samples. As N tends towards infinity, P_f asymptotically approaches $\hat{P}_f$. To attain an estimation with an accuracy of $1-\varepsilon$ and a confidence level of $1-\delta$, the required sample number N can be approximated as $N \approx \frac{\log(1/\delta)}{\varepsilon^2 \hat{P}_f}$. In practice, when aiming for a moderate accuracy of 90% ($\varepsilon = 0.1$) with a confidence level of 90% ($\delta = 0.1$), the required sample number is $N \approx 100/\hat{P}_f$. Clearly, this approach becomes impractical when dealing with scenarios characterized by small values of $\hat{P}_f$, such as $\hat{P}_f = 10^{-5}$. This impracticality can be intuitively understood by recognizing that, on average, at least $1/\hat{P}_f$ samples are needed before we can observe a failure event.

2.3 Importance Sampling Yield Estimation

Instead of drawing samples directly from $p(\mathbf{x})$, the IS-based methods adopt an alternate strategy by drawing samples from a distinct proposal distribution denoted as $q(\mathbf{x})$ to estimate the failure rate P_f, which can be formulated as follows:

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

where $\mathbf{x}_i$ denotes samples drawn from $q(\mathbf{x})$, which are subsequently employed to approximate the integral in a manner akin to MC. For ease of notation, the importance weight is defined as $w(\mathbf{x}) = \frac{p(\mathbf{x})}{q(\mathbf{x})}$. When the proposal distribution $q(\mathbf{x})$ is designed properly, Eq. (1) is more efficient than MC. Utilizing Lagrange multiplier rule for calculus of variations, we can show that the optimal proposal distribution is given by

$$q^*(\mathbf{x}) = p(\mathbf{x})I(\mathbf{x})/\hat{P}_f. \tag{2}$$

3 PROPOSED APPROACH

3.1 Minimization Norm IS Yield Estimation

A canonical method that sets the stage for IS-based yield estimation is MNIS [5], which samples from a normal distribution centered at $\boldsymbol{\mu}^*$, which is called the optimal mean shift vector (OMSV) and is determined by solving the following optimization problem:

$$\boldsymbol{\mu}^* = \operatorname{argmin} ||\mathbf{x}||^2 \quad \text{s.t.} \quad I(\mathbf{x}) = 1, \tag{3}$$

where $||\mathbf{x}||^2 = \mathbf{x}^T\mathbf{x}$ represents the Euclidean norm. Certainly Eq. (3) is intractable due to the unknown $I(\mathbf{x})$. MNIS derives an approximation by drawing samples from a uniform distribution over the parameter space $\mathcal{X}$ and subsequently forms an initial collection of failure samples $\mathcal{D}$. The OMSV is then obtained by identifying the sample closest to the origin, i.e.,

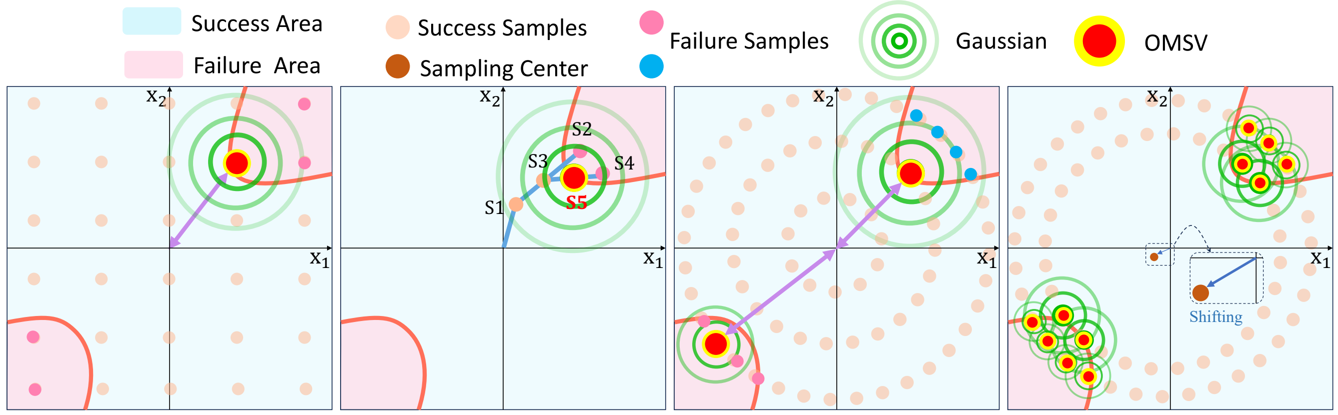
$$\boldsymbol{\mu}^* \approx \mathbf{x} = \underset{i \in \mathcal{D}}{\operatorname{argmin}} ||\mathbf{x_i}||^2. \tag{4}$$

An illustrative example is also given in Fig. 1a where the initial failure samples are indicated by pink dots.

3.2 Searching OMSV With Gradient

Despite its intuitive appeal and success, MNIS is suboptimal. This is obvious from Fig. 1a and Eq. (4) where the OMSV is selected from a finite set of failure samples based on randomly generated samples. To solve Eq. (3) more accurately, we should utilize the gradient information, as suggested by Gradient Importance Sampling (GIS [6]). GIS uses gradients computed from finite difference methods, and searches for the OMSV with a sequential update scheme (illustrated from S1 to S5 in Fig. 1b). To prevent a large step size that leads to search outside the failure region, GIS uses a dichotomy strategy to halve the step size when crossing the failure boundary (illustrated from S2 to S3 and S4 to S5 in Fig. 1b). The search for the OMSV is terminated when the step size reaches a predefined threshold.

While being effective, GIS loses an important advantage of MNIS, i.e., the simplicity of the algorithm without hyperparameter tuning, which is highly desirable in industry standard design flow. For GIS,



(a) MNIS **(b) GIS** **(c) HSCS** **(d) EFIAL**

Figure 1: Illustration of classic OMSV-based IS yield estimation (MNIS, GIS, and HSCS) and the proposed method (EFIAL).

Table 1: Classic OMSV-based Method Comparison

Method (chronologic)	MNIS	HSCS	GIS	ACS	FSIS	EFIAL
# hyperparameters	0	3	2	2	3	0
# Parameters	0	0	1	0	1	0
Multi-regions?	No	Yes	No	Yes	No	Yes
Adaptive?	No	No	No	Yes	No	Yes

we need to fine-tune the hyperparameter to achieve good performance. Inspired by GIS, FSIS [7] proposes a more precise gradient calculation based on sensitivity analysis and incorporates an adjustable learning rate to prevent searching outside the failure area, at the cost of introducing another hyperparameter.

3.3 Handling Multiple Failure Regions

As clearly demonstrated in Fig. 1, traditional methods such as MNIS, GIS, and FSIS cannot handle multiple failure regions because they define only one OMSV, which seems to be an easy fix just by introducing multiple OMSVs. Hyperspherical Clustering and Sampling (HSCS [9]) first introduces a hyperspherical k-means clustering to identify multiple failure regions, based on which multiple OMSVs are determined using MNIS. An illustrative example is given in Fig. 1c where the two failure regions are indicated by pink and blue dots, respectively. Identifying multiple failure regions becomes the vital task for HSCS, which is then resolved by employing a spherical pre-sampling technique to enhance computational efficiency. Nonetheless, the success comes with a extra cost of hyperparameter tuning.

All the aforementioned OMSV-based methods determine the OMSV(s) based on pre-sampling failure samples, which seems to be a waste of resource as more failure samples are observed and collected during the yield estimation process. To resolve and push the frontier of OMSV-based IS yield estimation, Adaptive Clustering and Sampling (ACS [10]) improves HSCS by introducing an adaptive update framework to update the OMSVs during the estimation process.

3.4 Remark on OMSV-based Methods

The evolution of OMSV-based methods is summarized in Table 1, where we can see that the methods have evolved to address increasingly complex problems at the cost of additional complexity and the need for meticulous tuning. This shift shows a departure from their original intent, which is to provide methods that are both reliable and efficient—qualities highly prized in industrial applications. Given the complexities associated with tuning hyperparameters in HSCS and ACS, users might consider transitioning to more advanced machine learning-based approaches, such as Low Rank Tensor Approximation (LRTA [1]) or Normalizing Flow [4].

3.5 Every Failure Is A Lesson

We seek to make a renaissance of MNIS, i.e., proposing a novel tuning-free method that can handle increasingly complex modern yield challenges with minimal computational cost and robustness. We notice that once the OMSV is determined in MNIS, the other failure samples are discarded, despite the fact that they also contain useful information e.g., where the failure region may reside and the volume of the failure region. As we will see soon, we should learn from every failure sample to construct a better proposal distribution.

Revisiting the optimal proposal distribution in Eq. (2), the optimal proposal distribution is proportional to $p(\mathbf{x})I(\mathbf{x})$, which is approximated by using a simple normal distribution $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I})$, where $\boldsymbol{\mu}_i$ is the failure sample with minimal norm. Inspired by this, let us utilize more than one failure sample to construct a mixture Gaussian distribution, serving as the proposal distribution $q(\mathbf{x})$, i.e.,

$$q(\mathbf{x}) = \sum_{i=1}^{M} \beta_i \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{I}), \tag{5}$$

where β_i is the weight, $\boldsymbol{\mu}_i = \mathbf{x}_i$ is the failure sample, and M is the number of failure samples in our current collection. This idea is illustrated in Fig. 1d where the failure samples are indicated by pink dots. To optimize the mixture weight β_i, we can substitute Eq. (5) into Eq. (2), yielding,

$$\sum_{i=1}^{M} \beta_i \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \mathbf{I}) = \frac{1}{\hat{P}_f} p(\mathbf{x})I(\mathbf{x}). \tag{6}$$

Because all successful samples substituted into Eq. (6) will make the right hand side zero, they are not useful for determining β_i. Substituting all available failure samples into Eq. (6), we will have a linear system of M equations with M unknowns, i.e., β_i.

$$\begin{pmatrix} 1 & e^{-\frac{||\mathbf{x}_1-\mathbf{x}_2||^2}{2}} & \cdots & e^{-\frac{||\mathbf{x}_1-\mathbf{x}_M||^2}{2}} \\ e^{-\frac{||\mathbf{x}_2-\mathbf{x}_1||^2}{2}} & 1 & \cdots & e^{-\frac{||\mathbf{x}_2-\mathbf{x}_M||^2}{2}} \\ \vdots & \vdots & \ddots & \vdots \\ e^{-\frac{||\mathbf{x}_M-\mathbf{x}_1||^2}{2}} & e^{-\frac{||\mathbf{x}_M-\mathbf{x}_2||^2}{2}} & \cdots & 1 \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} e^{-\frac{||\mathbf{x}_1||^2}{2}} \\ e^{-\frac{||\mathbf{x}_2||^2}{2}} \\ \vdots \\ e^{-\frac{||\mathbf{x}_M||^2}{2}} \end{pmatrix} \frac{1}{\hat{P}_f} \tag{7}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)^T$ are the target weights. Note that $\hat{P}_f$ does not matter as we will normalize $\boldsymbol{\beta}$ to make $\sum_{i=1}^{M} \beta_i = 1$, which will not affect our proposal Eq. (5). We can use any linear equation solver to obtain the optimal weights $\boldsymbol{\beta}$ from Eq. (7), which requires a computational cost of $O(M^3)$

3.6 Blessing of Dimensionality

In practice, the variational space has high dimensionality, leading to another challenge known as the curse of dimensionality. Gorban et al. [11] discover that the curse of dimensionality can be transformed into a blessing of dimensionality in some applications of machine learning. Particularly, the classical concentration of measure theorems in high-dimensional space indicates that as the dimensionality increases, the volume of the space grows rapidly, leading to an increase in the average distance between samples to approximate a large constant [11]. This means that the non-diagonal elements in Eq. (7) will become extremely small, and the large $M \times M$ matrix in Eq. (7) will become an identity matrix. Thus, the solution to Eq. (7) is simply $\beta_i = e^{-\frac{||\mathbf{x}_i||^2}{2}} / \hat{P}_f$. Further considering the normalization (i.e., $\sum \beta_i = 1$) and substituting $\boldsymbol{\beta}$ into Eq. (5), we obtain the optimal proposal distribution based on M failure samples,

$$q(\mathbf{x}) = \frac{1}{\sum_{m=1}^{M} p(\mathbf{x}_m)} \sum_{i=1}^{M} p(\mathbf{x}_i) \mathcal{N}(\mathbf{x} | \mathbf{x}_i, \mathbf{I}). \tag{8}$$

This results in a closed-form solution for the optimal proposal distribution, introducing zero parameter/hyperparameter and requiring a computational cost of $O(M)$. Empirically, we find that even for a dimension of 18 (a standard 6T-SRAM bit cell), the values of the non-diagonal elements are on the order of 10^{-8}, which validates our hypothesis of the dimensional blessing. The main steps of EFIAL yield estimation are summarized in Algorithm 1.

3.7 EFIAL For Pre-sampling

Pre-sampling is a critical step in IS-based yield estimation, especially for non-updating methods such as MNIS and GIS. As we will show in Fig. 7, the accuracy of yield estimation is significantly influenced by both the size and methodology of pre-sampling. Even for updating methods (e.g., ACS), pre-sampling is the key to high efficiency and accuracy, as it provides an initial understanding of the geometry of the parameter space. Naive pre-sampling methods include uniform sampling or quasi-random uniform sampling (e.g., Latin Hypercube Sampling). These methods are promising if given enough budget but may lack efficiency, as many failure samples will lie far from the origin, accounting for a small weight is in Eq. (1).

Onion Sampling (OS [4]) is an SOTA tuning-free quasi-random pre-sampling. Starting from the origin, the sampling space expands like a growing onion, where the growing volume is controlled to be the same according to the cumulative probability distribution of a Gaussian function. The growing volume, termed a layer, is selected during sampling with a uniform probability distribution. A sample inside the layer is then generated using another uniform distribution. The key ingredient of OS is that after a certain number of samples are generated, the inner layers that contain no failure samples are removed, and the pre-sampling will focus on the remaining layers. This way, the pre-sampling is more effective by avoiding the successful area normally residing near the origin.

Algorithm 1 EFIAL Yield Estimation Algorithm

```
Require: $M$ failure samples $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_M\}$; SPICE-based $I(\mathbf{x})$
repeat
    Update iteration $t = t + 1$
    Build proposal $q^t(\mathbf{x})$ based on Eq. (8) with $\mathcal{D}$
    Sample $K$ points from $q(\mathbf{x})$ and calculate importance weight:
    Calculate the failure rate $\widehat{P}_f = \frac{1}{tK} \sum_{j=1}^{t} \sum_{k=1}^{K} w_k^j$
    Collect all $K'$ failure samples $\mathcal{D}'$ and update failure sample collection $\mathcal{D} \longleftarrow \mathcal{D} \cup \mathcal{D}'$
until Figure of Merit (FOM): $std(\widehat{P}_f)/\widehat{P}_f < 0.1$
```

The solution in Eq. (8) suggests that the centroid of the proposal distribution should be weighted by the failure samples and their failure rate. Thus, a simple improvement to OS is that, when discarding the inner layers, we also shift the centroid of the OS from the origin the weighted centroid of the failure samples, $\hat{\boldsymbol{\mu}} = \sum_{i=1}^{N} p(\mathbf{x}_i)\mathbf{x}_i$.

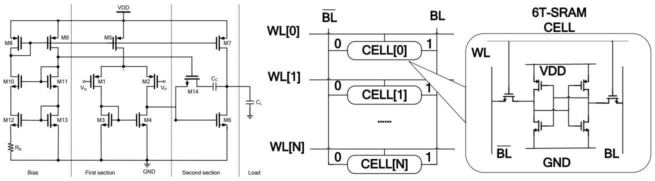


Figure 2: OTA and SRAM array Circuits

4 EXPERIMENTAL RESULTS

In this section, we conduct a thorough assessment of the accuracy and efficiency of our method, referred to as EFIAL, across three circuits: a 6T-SRAM, an operational transconductance amplifier (OTA) and a 6-bit 6T-SRAM array circuit. To establish a robust foundation for comparison, we implement four OMSV-based methods, namely, MNIS [5], HSCS [9], AIS [8], and ACS [10], as well as one surrogate-based method namely, LRTA [1]. MC is used as the gold standard to estimate the true failure rate. The Figure of Merit (FoM), denoted as ρ and calculated as $\rho = \text{std}(P_f)/P_f$ (where $\text{std}(P_f)$ represents the standard deviation of estimated failure rate), serves as the termination criterion for all methods. We set $\rho = 0.1$, a threshold commonly adopted in numerous prior studies, such as [5] and [9]. This threshold corresponds to a minimum 90% accuracy level with a 90% confidence interval. The relative error is defined as the difference between the true failure rate and the estimated value, divided by the true value. The speedup is obtained by dividing simulations required by MC by simulations required by each method.

In the experiments, each method is tested with ten random seeds, and the final experimental results are obtained by averaging the results from these ten runs. Furthermore, we select the best-performing seed for each method among the ten seeds to provide an intuitive visualization of the iterative estimation of failure rate and its FoM. The baseline methods are implemented using their default settings, and in some cases where certain methods fall short, we fine-tune hyperparameters to achieve better performance for different circuits. Note that EFIAL does not require any tuning for any experiments. All experiments are conducted on a Windows system with an AMD 7950x CPU and 32GB RAM.

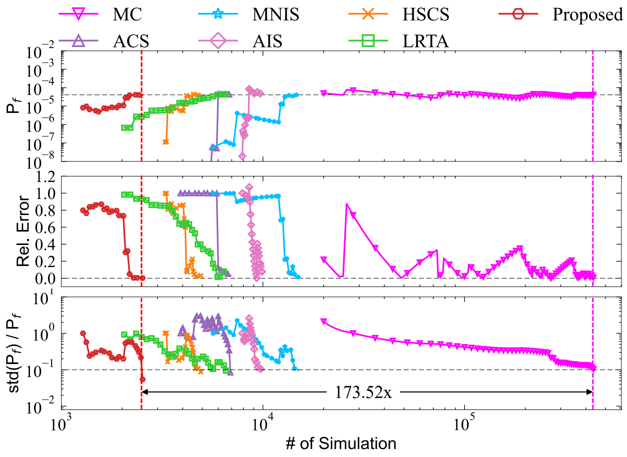


Figure 3: Failure rate estimation with FoM on 6T-SRAM

Table 2: Yield Estimation Results on 6T-SRAM

Method	Fail. Rate	Rel. Error	# Sim	Speedup
MC	4.99e-5	-	406240	1x
MNIS	4.81e-5	3.61%	10030	40.50x
HSCS	4.86e-5	2.61%	4152	97.84x
AIS	4.85e-5	2.81%	9702	41.87x
ACS	4.70e-5	5.81%	9620	42.23x
LRTA	4.86e-5	2.61%	6130	66.27x
Proposed	**4.97e-5**	**0.40%**	**2503**	**162.30x**

4.1 6T-SRAM Circuit

The 6T-SRAM bit cell is shown in Fig. 2 (right). In this design, WL stands for word line, and BL represents bit line. This architecture employs two cross-connected inverters, each using four transistors, as data storage units. Two additional transistors act as control switches for data transfer. Notably, each transistor has three variational parameters, which makes 6T-SRAM bit cell encompass 18 independent random variables. We use the delay time of SRAM read/write operations as the main performance metric.

The yield estimation results are presented in Table 2 and Fig. 3. EFIAL outperforms all baseline methods in terms of both accuracy and efficiency. More specifically, EFIAL achieves an impressively low relative error of 0.40%, which is 2.21% lower than the second-best accuracy achieved by LRTA and HSCS; EFIAL achieves a remarkable 162.30x speedup compared to MC, and a 1.66x to the second best.

4.2 Operational Transconductance Amplifier

The operational transconductance amplifier (OTA) circuit, illustrated in Fig. 2 (left), comprises 14 transistors. Each transistor is characterized by four process variation parameters, including oxide thickness, threshold voltage, and variations in both length and width due to process deviations. In our experimental setup, we evaluate the critical performance metric: quiescent current (I_Q) at a temperature of $27°C$.

The results of our yield estimation experiments, as presented in Table 3 and Fig. 4, highlight the superiority of EFIAL, delivering a 5.29% accuracy improvement and a 3.23x speedup compared with the second best methods in their respective categories. Furthermore, the speedup of EFIAL over MC is 362.92x.

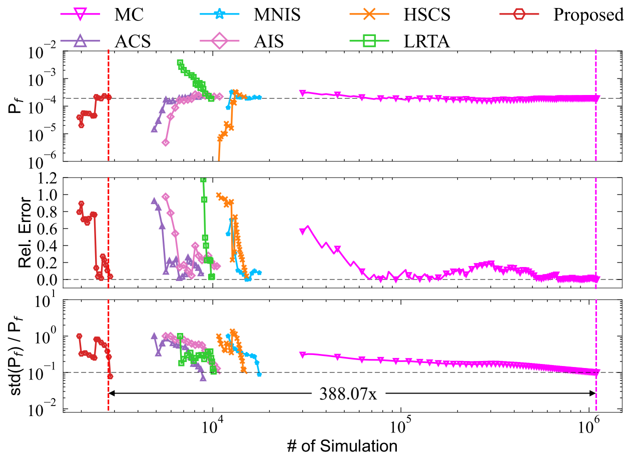


Figure 4: Failure rate estimation with FoM on OTA

Table 3: Yield Estimation Results on OTA

Method	Fail. Rate	Rel. Error	# Sim	Speedup
MC	1.89e-4	-	1135200	1x
MNIS	1.64e-4	13.23%	21065	53.89x
HSCS	1.70e-4	10.05%	17950	63.24x
AIS	1.74e-4	7.94%	11178	101.56x
ACS	1.78e-4	5.82%	11053	102.71x
LRTA	2.04e-4	7.94%	10100	112.40x
Proposed	**1.88e-4**	**0.53%**	**3128**	**362.92x**

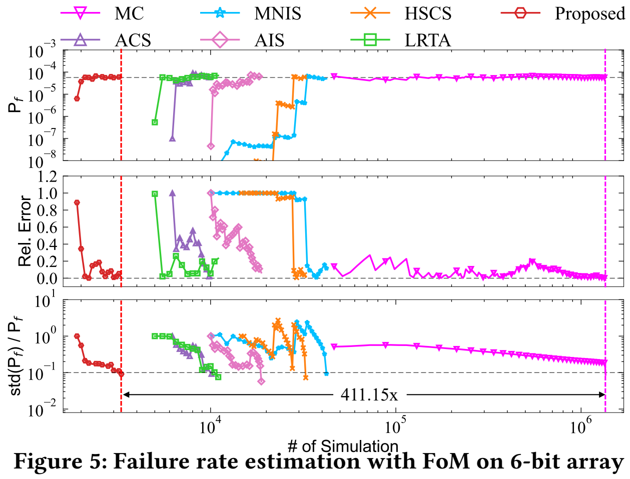


Figure 5: Failure rate estimation with FoM on 6-bit array

Table 4: Yield Estimation Results on 6-bit 6T-SRAM Array

Method	Fail. Rate	Rel. Error	# Sim	Speedup
MC	5.62e-5	-	1417500	1x
MNIS	4.94e-5	12.10%	45174	31.38x
HSCS	4.21e-5	25.09%	47090	30.10x
AIS	4.37e-5	22.24%	15996	88.62x
ACS	4.92e-5	12.46%	14060	100.82x
LRTA	5.96e-5	6.05%	12300	115.24x
Proposed	**5.61e-5**	**0.18%**	**3478**	**407.56x**

Table 5: Comparison for Different Pre-sampling Methods on 6-bit 6T-SRAM Array Circuit

	Original Pre-sampling				Onion Sampling				EFIAL-Onion Sampling			
Method	Fail. Rate	Rel. Error	# Sim	Speedup	Fail. Rate	Rel. Error	# Sim	Speedup	Fail. Rate	Rel. Error	# Sim	Speedup
MNIS	4.94e-5	12.10%	45174	31.38x	5.10e-5	9.25%	21824	64.95x	**5.74e-5**	**2.14%**	**7894**	**179.57x**
HSCS	4.21e-5	25.09%	47090	30.10x	5.02e-5	10.68%	24526	57.80x	**5.52e-5**	**1.78%**	**5185**	**273.38x**
AIS	4.37e-5	22.24%	15996	88.62x	5.17e-5	8.01%	9514	148.99x	**5.75e-5**	**2.31%**	**8870**	**159.81x**
ACS	4.92e-5	12.46%	14060	100.82x	5.23e-5	6.94%	10060	140.90x	**5.37e-5**	**4.45%**	**6645**	**213.32x**

4.3 6-bit 6T-SRAM Array Circuit

According to the previous introduction to SRAM circuit, each transistor contains three variational parameters. So a 6-bit SRAM array, composed of six cells, each with six transistors, results in a total of 108 variational parameters. We continue to choose the delay time of read/write as the interest performance metric $\mathbf{y}$. The experimental results, as shown in Table 4 and Fig. 5, illustrate that as the dimensionality of the process variable parameters increases, both the relative error rates and simulation costs significantly rise for all baseline methods. However, EFIAL remains stable and continues to outperform all baselines in accuracy and efficiency. EFIAL maintains a low relative error at 0.18%, which is significantly lower than other methods. Furthermore, EFIAL achieves a remarkable 407.56x speedup compared to MC, and a 3.54x-13.54x speedup compared to other baseline methods.

4.4 Computational Time

The computational time (by CPU hours) for the aforementioned yield estimation experiments is shown in Fig. 6. EFIAL exhibits a superior computational efficiency, achieving up to 4x (3.17x on average), 6.41x (4.32x on average), and 13.6x (7.72x on average) speedup for the 6T-SRAM, OTA, and 6-bit 6T-SRAM array circuits, respectively.

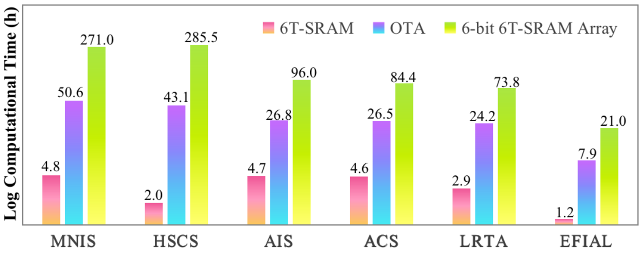


Figure 6: The Comparison of Computational Time

4.5 Pre-sampling Methods Comparison

Finally, we assess how EFIAL can improve the SOTA pre-sampling method, onion sampling [4], which has shown promising results. We first conduct comparison experiments on 6T-SRAM using EFIAL with uniform sampling, onion sampling, and EFIAL-onion sampling. The results are shown in Fig. 7, which show a clear improvement in yield estimation accuracy (1−relative error) with the same budget using EFIAL-onion sampling. When the budget is large, all methods converge to the same accuracy level. In this case, EFIAL reaches an accuracy of about 1 with 1500 samples, whereas onion sampling and uniform sampling require 4000 and 6000 samples, respectively.

To further showcase the effectiveness of EFIAL-onion sampling, we conduct comparison experiments for all OMSV-based baselines, equipped with their original pre-sampling methods, onion sampling, and EFIAL-onion sampling on 6-bit 6T-SRAM array. The results, shown in Table 5, demonstrate that EFIAL-onion sampling consistently provides substantial improvements on top of the gains achieved by onion sampling across the four baseline methods. On average, it enhances the accuracy of these methods by 15.26% and improves efficiency with a 4.68x speedup, indicating the power of EFIAL-onion sampling.

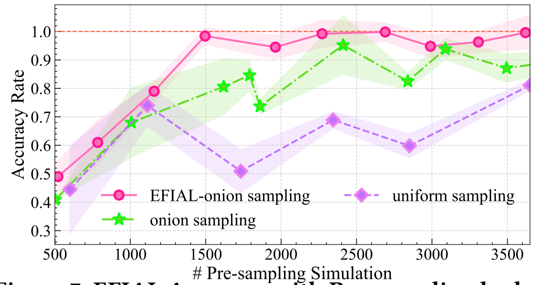


Figure 7: EFIAL Accuracy with Pre-sampling budget

5 CONCLUSION

In this work, we generalize what is arguably the most fundamental yield estimation method, OMSV, by harnessing all failure samples (instead of only the most probable one), without introducing any parameters or hyperparameters. The accuracy and efficiency of the proposed method are well demonstrated and we expect the concept of EFIAL to inspire more novel yield estimation methods.

REFERENCES

[1] 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 *Proc. DAC*, pages 1–6, 2019.

[2] Shuo Yin, Guohao Dai, and Wei W. Xing. High-dimensional yield estimation using shrinkage deep features and maximization of integral entropy reduction. In *Proc. ASP-DAC*. IEEE, 2023.

[3] Shuo Yin, Xiang Jin, Linxu Shi, Kang Wang, and Wei W Xing. Efficient bayesian yield analysis and optimization with active learning. In *Proc. DAC*, pages 1195–1200, 2022.

[4] Yanfang Liu, Guohao Dai, and Wei W Xing. Seeking the yield barrier: High-dimensional sram evaluation through optimal manifold. In *Proc. DAC*, pages 1–6. IEEE, 2023.

[5] Lara Dolecek, Masood Qazi, Devavrat Shah, and Anantha Chandrakasan. Breaking the simulation barrier: Sram evaluation through norm minimization. In *Proc. ICCAD*, pages 322–329. IEEE, 2008.

[6] Thomas Haine, Johan Segers, Denis Flandre, and David Bol. Gradient importance sampling: An efficient statistical extraction methodology of high-sigma sram dynamic characteristics. In *Proc. DATE*, pages 195–200, 2018.

[7] 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 *Proc. DAC*, pages 895–900, 2021.

[8] Xiao Shi, Fengyuan Liu, Jun Yang, and Lei He. A fast and robust failure analysis of memory circuits using adaptive importance sampling method. In *Proc. DAC*, pages 1–6. IEEE, 2018.

[9] Wei Wu, Srinivas Bodapati, and Lei He. Hyperspherical clustering and sampling for rare event analysis with multiple failure region coverage. In *Proc. ISPD*, pages 153–160, 2016.

[10] 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 *Proc. ISPD*, pages 139–146, 2019.

[11] Alexander N Gorban and Ivan Yu Tyukin. Blessing of dimensionality: mathematical foundations of the statistical physics of data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2118):20170237, 2018.