SRAM Memory Margin Probability Failure Estimation using Gaussian Process Regression

Manish Rana, Ramon Canal
Department of Computer Architecture
Universitat Politecnica de Catalunya
Barcelona, Catalunya, Spain
Email: mrana@ac.upc.edu, rcanal@ac.upc.edu

Jie Han, Bruce Cockburn
Department of Electrical and Computer Engineering
University of Alberta
Edmonton, Alberta, Canada
Email: jhan8@ualberta.ca, cockburn@ualberta.ca

***Abstract*—Estimating the failure probabilities of SRAM memory cells using Monte Carlo or Importance Sampling techniques is expensive in the number of SPICE simulations needed. This paper presents a methodology for estimating the dynamic margin failure probabilities by building a surrogate model of the dynamic margin using Gaussian Process regression. Additive kernel functions that can extrapolate the margin values from the simulated samples are presented. These proposed kernel functions decrease the out-of-sample error of the surrogate model for a 6T cell by 32% compared with a six-dimensional universal kernel such as a Radial-Basis-Function kernel (RBF). Finally, the failure probability values predicted by a surrogate model built using 1250 SPICE simulations are reported and compared with Monte Carlo analysis with 10^6 samples. The results show a relative error of 30% at 0.4V (predicted value of $4\text{x}10^{-6}$ for the Monte Carlo estimate of $3\text{x}10^{-6}$) and a relative error of 172% at 0.3V (predicted value of $3\text{x}10^{-5}$ for the Monte Carlo estimate of $1.1\text{x}10^{-5}$) for the dynamic read margin. These accuracy numbers are similar to those reported in previous proposals while the reduction in SPICE simulations is between 4x and 23x relative to these proposals and 800x compared to Monte Carlo method.**

I. Introduction

The need for faster SPICE analysis of SRAM memory circuits in the presence of process variations has led to the adoption of advanced statistical sampling methods such as Importance Sampling based methods (Mixture Importance Sampling [1], Minimum-Norm Importance Sampling [2]) and extreme value statistics [3] to estimate the failure probabilities of the memory. In order to achieve robust memory operation, extremely small memory failure probabilities (such as $< 10^{-6}$) are required. As such, the aforementioned statistical sampling methods still need tens of thousands of SPICE simulations to estimate the memory failure probabilities. An alternative approach is to first build a surrogate model for the memory margins using SPICE simulations, and then estimate the memory failure probability using only the predictions of the surrogate model. For instance, the effectiveness of using "Kriging" surrogate models to reduce the SPICE simulations was shown in [4] using the spherical covariance function as the kernel. Importance Sampling from the surrogate models was used for faster high-sigma yield analysis of the SRAM cells in [5] where the Radial Basis Function (RBF) kernel network was used to first build the surrogate model. Furthermore, the surrogate models (built using Gaussian process regression) were also shown in [6] to be effective in reducing the corner simulations by up to 95% for high-confidence design verification compared to the full-factorial analysis of the circuit performance corners.

These kernels (such as RBF kernels) provide highly flexible models and as such are called universal kernels. The difficulty with the use of these kernels is faced when the number of variability sources increase. These universal kernels (e.g. RBF kernels) suffer from the curse of dimensionality [7], that is, the required number of training samples needed to accurately model the memory margin increases exponentially with the increase in variation sources. Consequently the regression becomes slower and SPICE simulations increase.

This paper investigates the use of Gaussian Process regression for modeling the dynamic noise margins of the 6T SRAM bit-cell at sub-threshold voltages under the presence of threshold voltage variations. At ultra-low voltages, the SRAM bit-cell read current has exponential dependence on the threshold voltage [8]. Thus, the presence of threshold voltage variations results in a non-linear response of the dynamic margins to these variations. Regression using Gaussian Processes [9] can be used to build flexible non-linear models of memory margins. In this paper, we present a methodology to build surrogate models of the non-linear behaviour of SRAM dynamic noise margins at sub-threshold voltages using additive kernel based Gaussian Process regression [10].

Our method provides an alternative to the universal kernels such as RBF. The method aims at achieving higher model accuracy with smaller out-of-sample error than the RBF kernel. That is, better extrapolation capability of the memory margin model at variation values largely different from the SPICE simulated values.

This paper makes the following contributions:

1) Gaussian Process models with additive kernels are presented as surrogate models for the SRAM cell's dynamic read margin.
2) Read dynamic failure probabilities using these surrogate models are reported and compared with the traditional Monte Carlo simulation.

The paper is organized as follows. In section II the relevant background material is presented. In section III, the proposed method is discussed for modeling the dynamic read margin of a 6T SRAM cell. Probability failure results are given in section IV. Finally, conclusions are given in section V.

II. Background

A. Gaussian Process Regression

A Gaussian Process (GP) used for non-parametric regression [9] is a distribution over random functions "$f(\cdot)$". Here, a random function "$f(\cdot)$" is defined to be a function chosen randomly

from the set of functions Ω each of which maps an infinite set X to a set Y. As such "$f(\cdot)$" is an infinite collection of random variables. For instance, when the domain $X = N$, i.e. the set of Natural numbers, then the random function "$f(\cdot)$" is the collection of infinite random variables $\{f(1), f(2), f(3), ...\}$ sampled according to some probability distribution. In particular, when its value at a finite set of locations $X_n = \{x_1, x_2, x_3, ..., x_n\}$ i.e, $F_n = \{f(x_1), f(x_1), f(x_2), f(x_3), ..., f(x_n)\}$ has a joint Gaussian distribution, $F_n \sim N(\mu_H(X_n), K_H(X_n, X_n))$, then the random function "$f(\cdot)$" has a Gaussian Process distribution. Here, μ_H is the mean function and K_H is the covariance function (kernel) of the Gaussian Process distribution which are parametrized by the set of hyper-parameters "H".
The parametrized kernel K_H defines the function space "Ω" from which the random function "$f(\cdot)$" is randomly selected according to a Gaussian Process (GP) distribution. When a GP is used for regression, then this function space "Ω" is the hypothesis space for the regression functions. For instance,

1) Hypothesis space of constant functions: Using a constant kernel function $K_H(X, X^*) = \sigma$ where σ is a constant and $H = \{\sigma\}$
2) Hypothesis space of linear functions: Using a linear kernel function $K_H(X, X^*) = \sigma^2\|X - X^*\|$ and $H = \{\sigma\}$
3) Hypothesis space of periodic functions: Using a periodic kernel functions $K_H(X, X^*) = \sigma^2 exp((-2Sin^2(\pi\|X - X^*\|/P))/L)$ and $H = \{\sigma, P, L\}$
4) Hypothesis space of smooth functions (that is, functions having continuous higher-order derivatives): Using a Radial Basis Function kernel (RBF) $K_H(X, X^*) = \sigma^2 * exp((\frac{-\|X-X^*\|^2}{L}))$ and $H = \{\sigma, L\}$

The functions in the hypothesis function space "Ω" are the prior functions for GP regression. As an example, all linear functions are prior functions when the linear kernel is used. Comparison of the marginal likelihood of data from these prior functions provides the set of posterior functions that best explain the observed data. The collection of predictions from these posterior functions at a test point $\{x^*\}$ gives a distribution of predicted values at this test point by the random variable $f(x^*)$. The mean of these predicted values is the mean prediction of the GP regression at test point $\{x^*\}$.

B. *Composite Kernels*

Composite kernels are created by adding kernels or by taking the product of kernels. The product kernels provide more flexible prior functions, while the additive kernels have higher extrapolation capacity [11]. For instance, the extrapolation capability of the product kernels is compared with the additive kernels for data sampled from a quadratic function in Figure 1. When the functional form of a kernel matches the trend in the data, the posterior functions fit the data exactly. This is indeed the case for the "Linear x Linear" kernel. Otherwise, the additive kernels can extrapolate an increasing trend in the observed data farther than the product kernels as is illustrated by the posterior functions of "Linear + RBF" vs. "Linear x RBF". Thus the sum of product kernels provides the extrapolation capability of the additive kernels and also flexible priors from the product kernels.

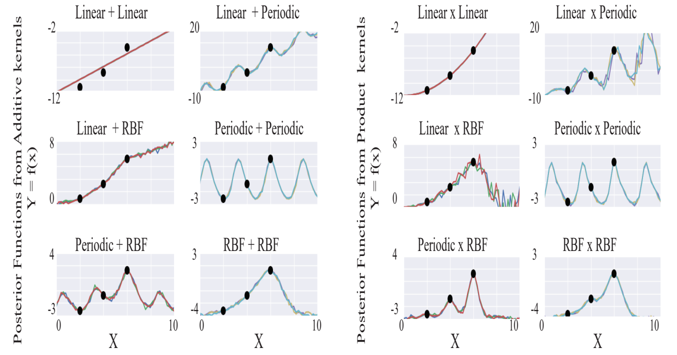


Fig. 1: Three posterior functions for the composite kernels with input data points sampled from a quadratic function. Linear x Linear best captures the quadratic trend.

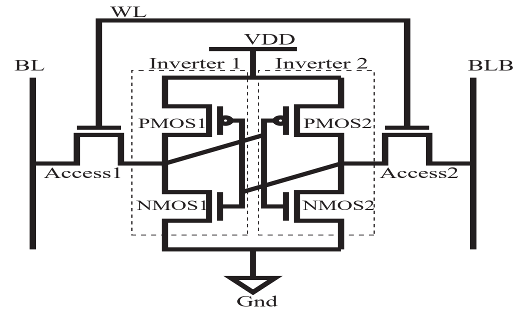


Fig. 2: Schematic of 6T SRAM Cell

III. Modeling 6T SRAM dynamic margins at sub-threshold voltage

A. *Methodology*

The schematic of the 6T SRAM bitcell is shown in Figure 2. The dynamic read margin for this analysis is defined as the voltage difference between the nodes storing logic value "1" and logic value "0" at the end of 20ns read word-line pulse width. HSPICE simulation of the bitcell netlist is done using the Predictive Technology model (PTM) for the 65nm bulk technology node at the sub-threshold supply voltages of 0.3V and 0.4V. The variations in the threshold voltages, V_{th}, of the six transistors due to the random dopant fluctuations (RDF) are considered to be six independent Gaussian random variables, assuming 5% V_{th} variation in the smallest size transistor. Only, the effect of RDF on the threshold voltages is considered because it was shown to be the dominant component of the variations in sub-threshold operation in [12].

The baseline model for the comparison is a six-dimensional RBF kernel which can learn any continuous six dimensional function given enough data.

$$k_{baseline} = RBF([x_1, x_2, x_3, x_4, x_5, x_6], [x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*]) \tag{1}$$

Here $\{x_1, x_2, x_3, x_4, x_5, x_6\}$ is a vector in the six-dimensional space of V_{th} variations in transistors {PMOS1, NMOS1, Access1, PMOS2, NMOS2, Access2}. During the training of the surrogate model, the training set is increased up to 1000 samples. The out-of-sample error of the trained model is then estimated using 10^6 test samples. Mean of the in-sample error and out-sample error are taken from 20 iterations. Lastly, Monte

Carlo simulation is done using 10^6 samples so that the failure probabilities higher than 10^{-6} can be compared.

B. Results

The sensitivity analysis of the dynamic read margin for the 6T SRAM cell showed that the sensitivity of the read dynamic margin is non-linear with respect to {PMOS1, NMOS1, NMOS2, Access2} V_{th} variations. However, it is linear with respect to {Access1, PMOS2} V_{th} variations. In order to minimize the model complexity, only the interaction terms between transistors that are in the same inverter structure are considered. Each constituent kernel in this proposed additive model is a one-dimensional kernel.

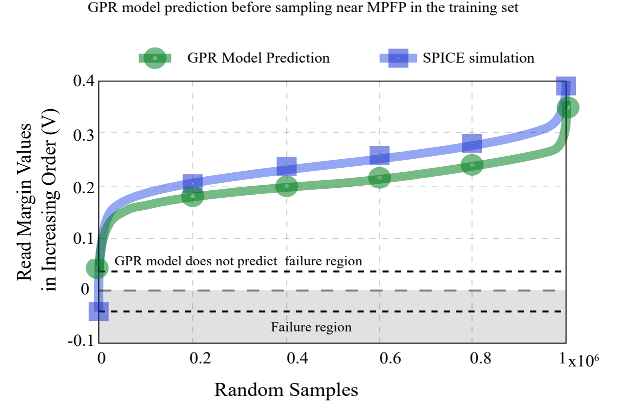
$$
\begin{aligned}
k_{proposed} = & \text{ //Sum of kernels for main effect//} \\
& RBF(x_1, x_1^*) + RBF(x_2, x_2^*) + Linear(x_3, x_3^*) \\
& + Linear(x_4, x_4^*) + RBF(x_5, x_5^*) + RBF(x_6, x_6^*) \\
& \text{//Sum of product kernels for interactions in Inverter-1//} \\
& + RBF(x_1, x_1^*) * RBF(x_2, x_2^*) \\
& + RBF(x_2, x_2^*) * Linear(x_3, x_3^*) \\
& + Linear(x_3, x_3^*) * RBF(x_1, x_1^*) \\
& \text{//Sum of product kernels for interactions in Inverter-2//} \\
& + Linear(x_4, x_4^*) * RBF(x_5, x_5^*) \\
& + RBF(x_5, x_5^*) * RBF(x_6, x_6^*) \\
& + RBF(x_6, x_6^*) * Linear(x_4, x_4^*)
\end{aligned} \tag{2}
$$

The proposed additive model achieves a lower out-sample error compared to the baseline RBF kernel and the other additive kernels as seen in Figure 3(1). After training the models on 400 samples, the out-sample error of the proposed additive model reaches 2.3x10^{-2} while for the baseline RBF model the Out-sample error at 400 simulations is 3.6x10^{-2}.

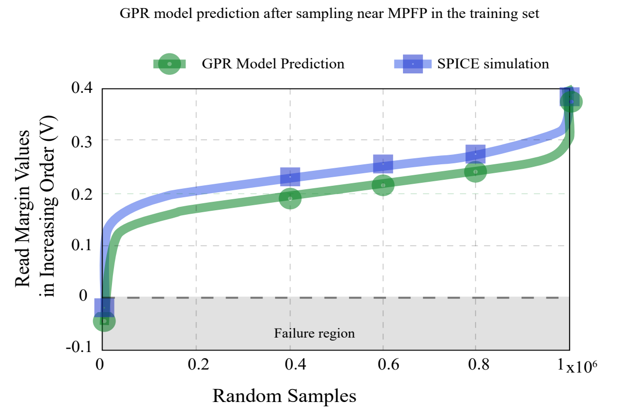
Fig. 3: Mean out-sample prediction error and mean in-sample error for dynamic read margin modeling of 6T SRAM at 0.3V for 20 iterations of GP regression. The proposed model has the minimum out-sample error. (R = RBF kernel and L = Linear kernel)

IV. Dynamic Margin Failure Probability

The additive kernel described in the previous section was used to model the 6T SRAM cell's dynamic read margin. Since the model's learning rate (decrease in its out-sample error) for the proposed additive model does not increase significantly after around 1000 training samples, as seen in Figure 3, the initial sampling stage randomly samples 1000 training samples using the Latin Hyper-cube Sampling (LHS) method to ensure that each sampling space dimension is uniformly sampled. Figure 4a compares the predicted margin values by the additive model



(a) Before sampling near the minimum norm failure point in the training set. Note that no failure points (points with margin value less than zero) are predicted by the model.



(b) After sampling additional points near the minimum norm failure point in the training set. The model is able to predict failure points.

Fig. 4: Read dynamic margin values at 0.3V predicted by the additive kernel model compared with the SPICE simulated margin values for 10^6 samples. Both the predicted and SPICE simulated margin values are sorted in increasing order.

with the SPICE results for 10^6 test samples. The predicted margin values and the SPICE simulated margin values are sorted in increasing order to make the comparison easier to visualize. The proposed additive model fails to predict margin values below 0 (i.e., no Failure points). The reason is the lack of enough failure points in the training set of 1000 samples which results in the set of posterior functions containing mostly functions with no failure regions. Thus, in order to improve the failure region prediction of the model, the training set is increased with an 250 more points (1000/4). The ratio (1/4) was empirically found for this specific case-study to be the ratio that least over-estimated the predicted dynamic read margin failure probability. This ratio will change for different supply voltages and range of variation in threshold voltages. These points are sampled from a normal distribution centered at the minimum norm failure point (MPFP), which is the "failure point" in the training set with the highest probability to be sampled under the distribution of threshold voltage variations. Addition of these samples to the training set increases the prediction accuracy near

TABLE I: Predicted dynamic read margin and write margin failure probabilities

Method	Dynamic Margin	#SPICE Simulations	Estimated Failure Probability
Monte Carlo	Read Margin @ 0.3V	10^6	$1.1\text{x}10^{-5}$
Proposal	Read Margin @ 0.3V	1250	$3\text{x}10^{-5}$
Monte Carlo	Read Margin @ 0.4V	10^6	$3\text{x}10^{-6}$
Proposal	Read Margin @ 0.4V	1250	$4\text{x}10^{-6}$

the failure boundary. Note that in our case-study of 6T bitcell at ultra-low voltages of 0.3 and 0.4V, the initial set of 1000 training samples gave failure samples among which the MPFP could be selected. However if there are no failure samples in the initial training set, then additional LHS sampling will be required. Figure 4b compares the predicted dynamic read margin values after sampling near MPFP. This additive model of memory margin is used as a surrogate model for Monte Carlo analysis to estimate margin failure probability. This step can also be performed using Importance Sampling on the surrogate model. Since sampling from surrogate model is not computationally as expensive as SPICE simulation, the focus of this paper is on reducing SPICE simulations to generate a surrogate model and then traditional Monte Carlo sampling from the surrogate model is used to estimate the failure probability. Table I shows that the relative error of the predicted dynamic read margin at 0.4V supply voltage is 30% compared to its Monte Carlo estimate, while the relative error at 0.3V is 172%. The relative error is larger for higher failure probability values because the same number of 250 points are sampled near MPFP in step 4 in both the cases. The fraction of these samples that are failure points is higher in the case $1.1\text{x}10^{-5}$ (Monte Carlo) failure probability for 0.3V. As such the method overestimates the failure probability to $3\text{x}10^{-5}$. This approach can be improved by using the generalized Pareto distribution (GPD) to accurately fit the tail of the dynamic margin distribution, as proposed in [13]. The failure regions can be classified using proposed additive kernels for Gaussian Processes instead of using the Gaussian Radial Basis kernel (GRBF) based state vector machine (SVM) [13].

The comparison of accuracy given in [14] for predicting the Monte Carlo estimate of $2.3\text{x}10^{-4}$ with REscope [13] and recursive statistical blockade [3] shows a relative error between 20% and 64% for estimating failure (Figure 5). Thus the proposed method provides similar accuracy numbers (minimum relative error of 30%) with speed-up in computation between 4x and 23x compared to these previous methods.

	Failure probability	**# Sim. runs**	**Speed -up(x)**	**Error%**
Monte Carlo(MC)	2.300E-04	1M	-	-
Rare Event Microscope (REscope)	3.786E-04	5009	199.6	64.61
Recursive Statistical Blockade (RSB)	2.775E-04	29260	34.2	20.65
Proposed method (IFRD)	2.852E-04	15730	63.6	24.00

Fig. 5: Accuracy numbers for similar approaches reported in [14]

V. Conclusion

In this paper, we show that for modeling SRAM dynamic margins, the extrapolation error (out-sample error) can be decreased with a smaller training set by using additive kernels that encode the structure present in the sensitivity analysis of the dynamic margin functions. We present the case study of modeling the dynamic read margin as an example for the efficacy of additive models made by using one-dimensional kernels and their interactions as sum of product kernels. The response surface generated by Gaussian Process using these proposed models is then used to estimate failure probabilities with 1250 simulations. These predicted failure probability values are then compared with Monte Carlo analysis with 10^6 samples and they show a relative error of 172% for dynamic read margin at 0.3V supply voltage.

Acknowledgment

This work has been supported by the HiPEAC PhD collaboration grant-2015, and the Spanish Ministry of Education and Science under grant TIN2013-44375-R, and Generalitat of Catalunya under grant 2009SGR1250. Manish Rana is supported by FI-DGR-2013-638.

References

[1] R. Kanj, R. Joshi, and S. Nassif, “Mixture importance sampling and its application to the analysis of SRAM designs in the presence of rare failure events,” in *2006 43rd ACM/IEEE Design Automation Conference*, 2006, pp. 69–72.

[2] L. Dolecek, M. Qazi, D. Shah, and A. Chandrakasan, “Breaking the simulation barrier: SRAM evaluation through norm minimization,” in *2008 IEEE/ACM International Conference on Computer-Aided Design*, Nov 2008, pp. 322–329.

[3] A. Singhee and R. 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*. Springer, 2008, pp. 235–251.

[4] O. Okobiah, S. Mohanty, and E. Kougianos, “Fast Design Optimization Through Simple Kriging Metamodeling: A Sense Amplifier Case Study,” *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 22, no. 4, pp. 932–937, April 2014.

[5] J. Yao, Z. Ye, and Y. Wang, “Efficient importance sampling for high-sigma yield analysis with adaptive online surrogate modeling,” in *Design, Automation Test in Europe Conference Exhibition (DATE), 2013*, March 2013, pp. 1291–1296.

[6] M. Shoniker, B. F. Cockburn, J. Han, and W. Pedrycz, “Minimizing the number of process corner simulations during design verification,” in *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*. EDA Consortium, 2015, pp. 289–292.

[7] G. Geenens *et al.*, “Curse of dimensionality and related issues in nonparametric functional regression,” *Statistics Surveys*, vol. 5, pp. 30–43, 2011.

[8] B. H. Calhoun and A. Chandrakasan, “Analyzing static noise margin for sub-threshold SRAM in 65nm CMOS,” in *Proceedings of the 31st European Solid-State Circuits Conference, 2005. ESSCIRC 2005.*, Sept 2005, pp. 363–366.

[9] C. E. Rasmussen, “Gaussian processes for machine learning,” 2006.

[10] D. K. Duvenaud, H. Nickisch, and C. E. Rasmussen, “Additive Gaussian Processes,” in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2011, pp. 226–234.

[11] D. Duvenaud, “Automatic model construction with Gaussian processes,” Ph.D. dissertation, University of Cambridge, 2014.

[12] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, “Analysis and mitigation of variability in subthreshold design,” in *ISLPED ’05. Proceedings of the 2005 International Symposium on Low Power Electronics and Design, 2005.*, Aug 2005, pp. 20–25.

[13] W. Wu, W. Xu, R. Krishnan, Y.-L. Chen, and L. He, “REscope: High-dimensional statistical circuit simulation towards full failure region coverage,” in *2014 51st ACM/EDAC/IEEE Design Automation Conference (DAC)*, June 2014, pp. 1–6.

[14] Y. Zhao, H. Shin, H. Chen, S. X. D. Tan, G. Shi, and X. Li, “Statistical rare event analysis using smart sampling and parameter guidance,” in *2015 28th IEEE International System-on-Chip Conference (SOCC)*, Sept 2015, pp. 53–58.