2022 6th IEEE Electron Devices Technology & Manufacturing Conference (EDTM) | 978-1-6654-2178-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/EDTM53872.2022.9798256

Heteroscedastic Gaussian Process Regression for ReRAM Device Modeling

Imtiaz Hossen[1], Yi Zang[2], Mark A. Anders[3], Lin Wang[2] and Gina C. Adam[1*]

[1]Department of Electrical and Computer Engineering, George Washington University, Washington, DC 20052, USA

[2]Department of Statistics, the George Washington University, Washington, Dc 20052, USA

[3]Alternative Computing Group, National Institution of Standards and Technology, Gaithersburg, MD 20899, USA

*Corresponding author Email: ginaadam@gwu.edu

Abstract

Jump tables are useful modeling approaches for emerging memory devices, e.g. ReRAM, and their use in neural network simulations because they rely only on experimental data. While binning is traditionally used for such modeling, this work proposes the use of Heteroscedastic Gaussian Process Regression (hetGP) to estimate signal mean and standard deviation simultaneously and develop a robust jump table model. The binning and hetGP approaches are verified using Kolmogorov-Smirnov (K-S) and maximum mean discrepancy (MMD) test.

Keywords: ReRAM, Device modeling, Jump table

Introduction

New types of ReRAM (or memristive) devices are continually being developed for their promising application in efficient unconventional computing [1]. As new devices are developed, it takes extensive research to formulate their physical models [2-5]. Models based only on experimental data are beneficial since they provide a fast way to capture the switching behavior at the device population level and get reliable neural network simulations [6].

Jump table modeling is an example of such modeling based entirely on experimental data [7]. The conductance response of ReRAM devices can exhibit stochastic behavior even when the programming pulses are identical. Moreover, there can be asymmetry between SET and RESET programming, even after the pulse amplitudes, shapes have been optimized. Jump tables are cumulative distribution functions used to model this stochastic nature of conductance change (*ΔG*) vs. initial conductance (*G*) in device programming. All these behaviors can be compactly captured in the form of a pair of jump tables, one for SET and one for RESET. While jump table methodology is not applicable for circuit modeling, it is useful in neural network simulations for the weight updating.

Traditionally, these tables are derived using a binning and linear interpolation approach that can introduce unwanted artifacts or lose important device behavior (Fig 1). Prior work showed that Gaussian Process Regression can predict the mean signal well [8], but robust standard deviation modeling is still missing. This work proposes an alternative approach for ReRAM modeling called heteroscedastic Gaussian Process Regression (hetGP) [9-10] which can be used to model both the mean signal and the switching noise standard deviation simultaneously. The hetGP has been extensively used in research areas such as machine learning, statistics, and engineering to solve non-linear regression problem.

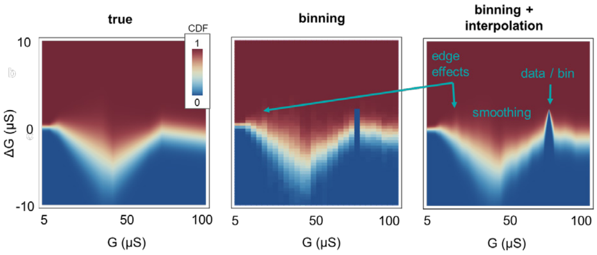


Fig. 1: Challenges with the traditional binning and interpolation approach shown on a synthetic distribution with known mean and standard deviation.

Methods

A. Binning and interpolation

Data binning is a process of grouping data G and ΔG into equally spaced bins using:

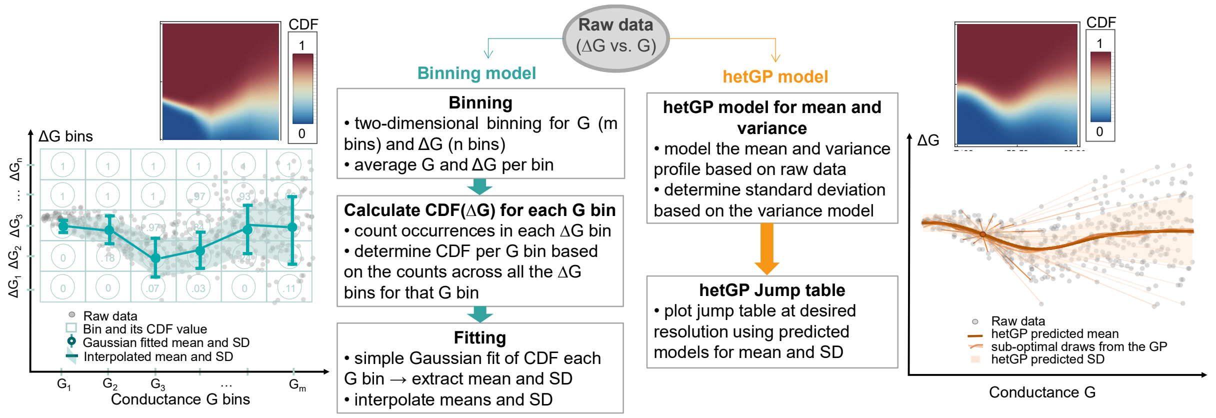
$$\mathrm{G}_{\mathrm{m}} = \frac{1}{N_{\mathrm{m}}} \sum_{\mathrm{m}}^{N_{\mathrm{m}}} \mathrm{G}_{\mathrm{m},i} \quad \text{and} \quad \Delta\mathrm{G}_{\mathrm{n}} - \frac{1}{N_{\mathrm{n}}} \sum_{\mathrm{n}}^{N_{\mathrm{n}}} \Delta\mathrm{G}_{\mathrm{n},j} \qquad [1]$$

where G_m is the average of the m^th^ G bin, $G_{\mathrm{m},i}$ = G of datasets *i* in bin m, $\Delta\mathrm{G}_\mathrm{n}$ = Average ΔG in bin n, $G_{\mathrm{n},j}$ = ΔG of datasets *j* in bin n, N_m = Number of datasets in bin m, N_n = Number of datasets in bin n.

The mean and standard deviation (SD) information can be extracted from binned data through a simple Gaussian fit. The mean and SD values are then linearly interpolated across the G bins to form a jump table. Choosing the correct number of bins to avoid artifacts is challenging. In this work, the number of G bins is 10 and of ΔG bins is 44.

A. hetGP modeling

HetGP is a method of interpolation where the mean and variance (standard deviation) in data can be modeled simultaneously. It assumes a Gaussian



Binning
- two-dimensional binning for G (m bins) and ΔG (n bins)
- average G and ΔG per bin

Calculate CDF(ΔG) for each G bin
- count occurrences in each ΔG bin
- determine CDF per G bin based on the counts across all the ΔG bins for that G bin

Fitting
- simple Gaussian fit of CDF each G bin → extract mean and SD
- interpolate means and SD

hetGP model for mean and variance
- model the mean and variance profile based on raw data
- determine standard deviation based on the variance model

hetGP Jump table
- plot jump table at desired resolution using predicted models for mean and SD

Fig. 2: Binning modeling approach vs. proposed hetGP modeling approach.

process for *ΔG* as stochastic function of conductance *G*. Given the experimental data $D_N = (G_N, \Delta G_N)$, the response to any conductance g value for desired prediction fits a Gaussian process $\boldsymbol{N}$:

$$\Delta G(g)|D_N \sim \mathcal{N}(\mu(g), \sigma^2(g))$$

where the mean is defined as:

$$\mu(g) = \mathbf{k}_N(g)^T \mathbf{K}_N^{-1} \Delta \mathbf{G}_N \qquad [3]$$

and the variance as:

$$\sigma^2(g) = \nu K_\theta(g, g) - \nu \mathbf{k}_N(g)^T \mathbf{K}_N^{-1} \mathbf{k}_N(g) \qquad [4]$$

Here $K_\theta(g;g)$ is the correlation function with θ specifying the decay in correlation; *v* is the magnitude of variances attributed to the conductance; k_N is the vector of correlations between modeled g and the experimental ΔG_N; and K_N is the sample covariance matrix. Once the type of correlation function $K_\theta(g;g)$ is chosen, the mean and variance can be determined through maximum likelihood estimation. In comparison with other Gaussian progress regression methods, hetGP supports joint likehood estimation for both the mean and variance models, as needed by our ReRAM jump table modeling. We used the Matern type covariance function and the hetGP package from R for modeling [11]. A comparative sketch of the binning and hetGP approaches is shown in Fig. 2.

B. *Experimental vs. reconstructed device data*

Experimental data (Fig. 3) were collected from a 10 nm edge device with 2.5 nm Al_2O_3 / 15 nm TiO_x / 5 nm Ti / 30 nm Pt, via two fast SMUs, a probe station, and no current compliance. The pulse width was 500 ns with a 50 % duty cycle. Four pulse amplitudes (±1.35 V, ±1.5 V, ±1.65 V, ±1.8 V) for set and reset respectively were investigated. From the measured set and reset data, a random subset of 4000 points was chosen for training the models and another non-overlapping 4000 points subset for testing them. This is iterated 20 times for each pulse amplitude for statistical significance. Each reconstructed data from hetGP and binning model were compared against the test subset using the Kolmogorov-Smirnov (KS) [12-13] test to quantify the distance between the distributions and the Maximum mean Discrepancy (MMD) [14] test to determine goodness of fit.

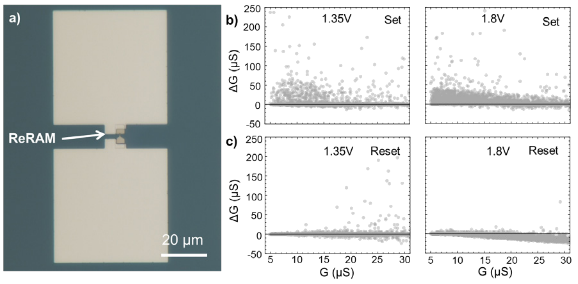


Fig. 3: ReRAM device information a) Representative optical picture b) Experimental data for set and c) reset obtained at 1.35 V and 1.8 V pulse amplitudes.

Results and discussions

In Fig. 4, the averages of the KS and MMD values and their 1-σ deviation of 20 iterations are shown for every pulse amplitude. The K-S and MMD values for the hetGP model are lower than the binning for most of the iterations (16 out of 80 iterations of K-S values) of each pulse amplitude. Lower values of KS and MMD indicate a better model fit to the experimental data, as the distance between the experimental test data and the reconstructured data is lower.

The results in Fig. 5 show that the reconstructed data based on the hetGP model successfully capture the overall distribution even of noisy experimental ReRAM data, with good low values for the K-S distances and MMD particularly for reset. Both models used in this work assume Gaussian data.

Nevertheless, the set data, particularly at low pulse

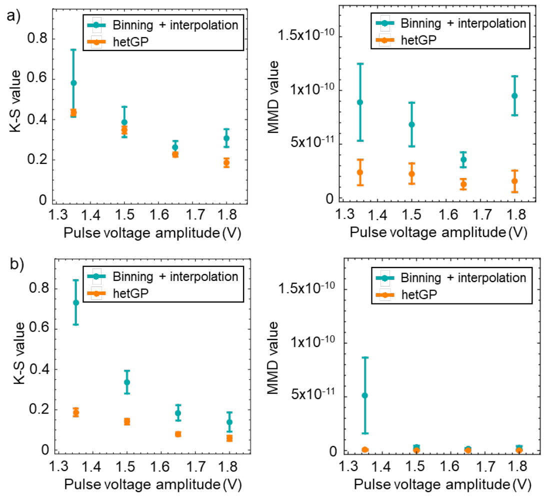


Fig. 4: K-S and MMD metrics for a) set and b) reset.

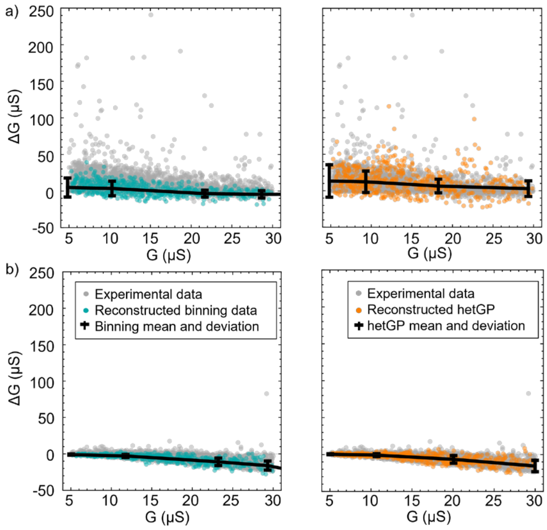


Fig. 5: Experimental and reconstructed data based on binning and hetGP approaches for a) set and b) reset obtained at 1.8V pulse amplitude.

amplitude seems to depart from the Gaussian assumption, so future hetGP expansions to support non-Gaussian distributions are needed.

Conclusions

This work proposes the use of heteroscedastic Gaussian Process regression for ReRAM jump table modeling and tests it vs. the binning approach using K-S and MMD tests. The hetGP modeling can determine the mean and standard deviation trends better than binning for all investigated pulse amplitudes. Future work will expand to non-Gaussian methods that consider physical constraints and test the models in machine learning simulations.

Acknowledgments

This work was supported by DARPA / ONR grant N00014-20-1-2031, the GW University Facilitating fund and the GW Cross-Disciplinary Research fund.

References

[1] J. J. Yang, D. B. Strukov, D. R. Stewart, “Memristive devices for computing.” Nat. Nanotechnol, 8(1), p.13-24 (2013).

[2] E.Abbaspour et al. “Studying the switching variability in redox-based resistive switching devices” J. Comput. Electron., 19, p. 1426-1432 (2020).

[3] X. Guan, S. Yu, H.-S. P. Wong, “On the switching parameter variation of metal-oxide RRAM—part i: Physical modeling and simulation methodology,” IEEE Trans. Electron Devices, 59 (4), 1172–1182 (2012).

[4] D. Ielmini, F. Nardi, C. Cagli, “Physical models of size-dependent nanofilament formation and rupture in NiO resistive switching memories,” *Nanotechnology*, 22 (25), 254022 (2011).

[5] F. M. Bayat, B.D. Hoskins, D.B. Strukov “Phenomenological modeling of memristive devices” Appl. Phys. A, 118(3), p. 779-786 (2015)

[6] G. W. Burr et al. “Experimental demonstration and tolerancing of a large-scale neural network (165 000 synapses) using phase-change memory as the synaptic weight element,” IEEE Trans Electron Devices, 62(11), p. 3498–3507 (2015).

[7] S. Sidler et al. “Large-scale neural networks implemented with non-volatile memory as the synaptic weight element” European Solid- State Device Research Conference, 440–443 (2016).

[8] N. Gong et al., Signal and noise extraction from analog memory elements for neuromorphic computing, Nat. Commun. 9, 2102 (2018).

[9] Goldberg et al., “Regression with input-dependent noise: A gaussian process treatment,” Adv. Neural Inf. Process. Syst, 10, p. 493-499, (1997).

[10] M. Binois, R. B. Gramacy, and M. Ludkovski “Practical heteroscedastic gaussian process modeling for large simulation experiments”, J. Comput Graph. Stat., 27(4), p.808-821 (2018).

[11] cran.r-project.org/web/packages/hetGP

[12] Y. Xiao, “A fast algorithm for two-dimensional Kolmogorov-Smirnov two sample tests,” Comput. Stat. Data Anal., 105, p. 53-58 (2017).

[13] cran.r-project.org/web/packages/Peacock.test/

[14] A. Gretton et al., “A kernel method for the two-sample-problem,” Adv. Neural Inf. Process. Syst, 19, p. 513-520 (2006).