✂️ Where to cut: Efficient ADC quantization for analog in-memory computing with discrete values

ISCAS 2025
IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS
LONDON, UK | May 25-28, 2025

Poster No. 2386

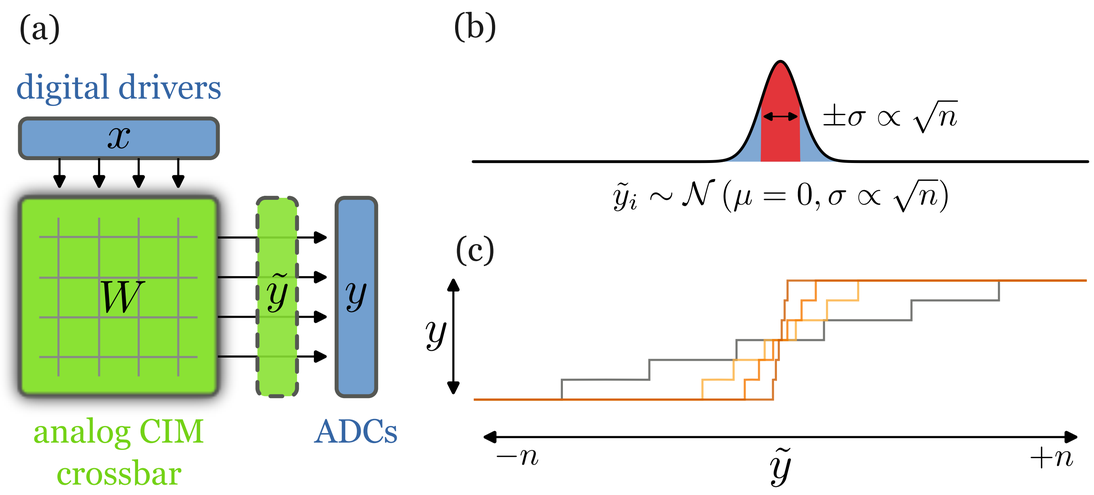
Johannes Leugering†, Shashank Bansal†, Zhaoyi Liu†, Jiahao Song†, Bouchaib Cherif‡, Gert Cauwenberghs†

† University of California, San Diego, CA, USA
{jleugering, shbansal, zhl101, jis090, gcauwenberghs}@ucsd.edu

‡ Northrop Grumman, Falls Church, VA, USA
bouchaib.cherif@ngc.com

INTRODUCTION

Many proposed in-memory-computing systems use **memristive crossbars** to compute **matrix-vector products** over **discrete domains**, e.g. binary or ternary inputs and weights. This yields **analog** output values, e.g. voltages, that are distributed around **discrete** values that can cover a wide range. But as implied by the central limit theorem, typical results are **highly concentrated** only in a small **central region**. Lossless quantization of the entire range requires costly high-precision analog-to-digital converters (ADCs), but this is not necessary for many applications like **neural network accelerators**. Instead, an ADC with **lower resolution** that is linear only in this central region and saturates outside it can achieve **almost full accuracy** at only a fraction of the cost. In this work, we derive optimal parameters for such an ADC.



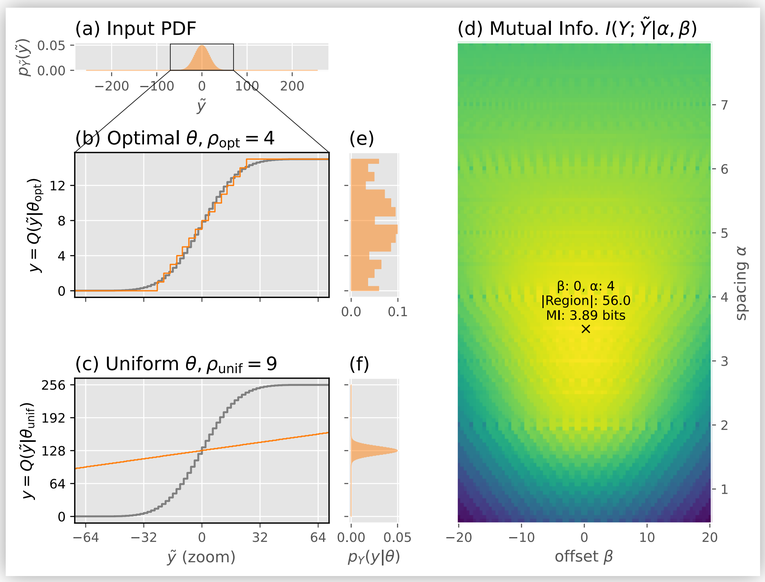
Example: (a) analog compute-in-memory system with $y = Q(\tilde{y} + \varepsilon) = Q(Wx + \varepsilon)$, where $x \in \{-1,1\}^n, W \in \{-1,1\}^{m \times n}, \tilde{y} \in \mathbb{R}^m, y \in Y^m = \{-n,n\}^m, \varepsilon \in \mathbb{R}^m, Q\colon \mathbb{R}^m \to Y^m$.
(b) The central limit theorem states that approximately $\tilde{y} \sim \mathcal{N}(0, \sigma)$, where $\sigma \propto \sqrt{n}$.
(c) Saturating quantization functions Q with different linear "regions of interest".

OBJECTIVES

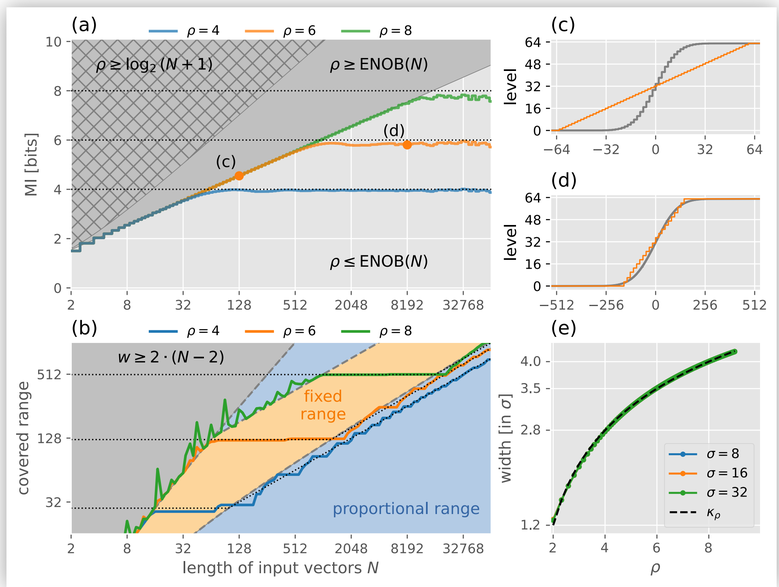
Using tools from information theory and numerical simulations, we

- systematically evaluate the optimal choice of **ADC resolution,**
- define a most informative **region of interest**,
- and explore the influence of **noise**.

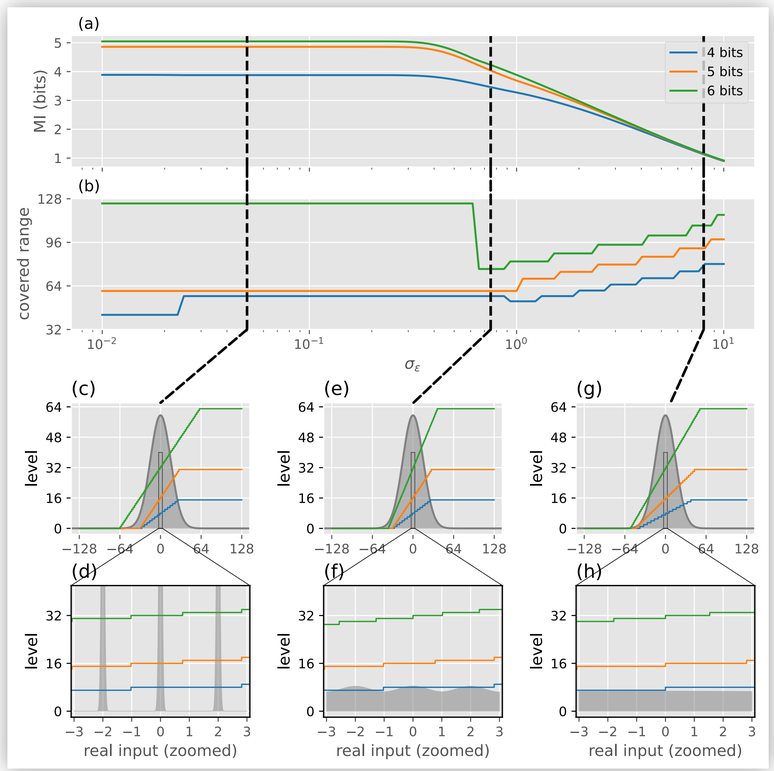
RESULTS



Quantization functions $y = Q(\tilde{y}|\theta)$ for linearly spaced thresholds at $\theta_i = (i - {}^{K-1}/_2)\alpha+\beta$ with scale parameter α, offset parameter β and $N = 256$. **(a)** Distribution of the input $\tilde{y}$. **(b)** Q for ρ_{opt}= 4bit, α=4 (optimal range, lossy, saturation). **(c)** Q for ρ_{unif} = 9bit, α=1 (full range, lossless). **(d)** Mutual information as a function of α, β. **(e,f)** The corresponding distributions of y for both ADCs.



Optimal ADC range as a function of vector length N. **(a)** MI for three ADCs with resolutions $\rho \in \{4, 6, 8\}$ (solid lines). **(b)** Size of optimal ADC range for each ρ as a function of N. **(c, d)** Optimal quantization functions for $\rho = 6$ and $N \in \{128, 8192\}$. **(e)** Numerically estimated optimal proportionality constant κ_ρ as a function of ρ.



The optimal quantization versus noise amplitude σ_ε for N = 256.
(a) MI for $\rho \in \{4, 5, 6\}$. **(b)** The optimal ADC range for each ρ.
(c, e, g) Distribution of unquantized results (gray fill) and quantization function for each ρ (solid lines) at $\sigma_\varepsilon \in \{0.05, 0.75, 8\}$.
(d,f, h) Corresponding zoomed-in views inside the center region.

- To maximize mutual information **at low resolution** (lossy encoding), the **quantization function** should approximate the CDF → **linear region.**
- **At high resolution**, it should encode each value at least 1:1 (lossless).
- The optimal **region-of-interest** thus depends on ρ and distribution of Y.
- For **high noise**, distribution becomes continuous & ROI scales with $\log \sigma_\varepsilon$

CONCLUSION

Heuristic[1] algorithm for designing ADCs for a compute-in-memory system:

- Pick the ADC resolution ρ (ideal: slightly larger than the **distribution's ENOB).**
- Determine the central ROI of size $\kappa_\rho \sigma_{\tilde{y}}$, where κ_ρ depends on ρ and $\tilde{Y}$.
  - E.g. for 256 bipolar inputs and weights, and $\rho = 8$ bit we get $\kappa_8 \approx 4$.
  - For $\rho = 4$ bit we get $\kappa_4 \approx 2.8$.
- For low resolution ($2^\rho \leq |\mathrm{Y} \cap \mathrm{ROI}|$): scale the ADC's range to **cover (only) the ROI**
- For high resolution ($2^\rho > |\mathrm{Y} \cap \mathrm{ROI}|$): cover **each value** in the ROI with ≥ 1 bin
- For real-valued weights or high noise: scale the ADC's range to **cover the ROI**
  - As noise ε increases beyond an LSB, κ_ρ begins to scale with $\log \sigma_\varepsilon$. κ_ρ

[1]: Future work should verify that these results generalize beyond our simplifying assumptions (i.i.d. bipolar x_i and w_i) to the empirical distributions of real workloads.

Paper + Code:

REFERENCES

[1] D. Ielmini and H.-S. P. Wong, "In-memory computing with resistive switching devices," Nature Electronics, vol. 1, no. 6, pp. 333–343, Jun. 2018, number: 6 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41928-018-0092-2

[2] Q. Xia and J. J. Yang, "Memristive crossbar arrays for brain-inspired computing," Nature Materials, vol. 18, no. 4, pp. 309–323, Apr. 2019, number: 4 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41563-019-0291-x

[3] A. S. Rekhi, B. Zimmer, N. Nedovic, N. Liu, R. Venkatesan, M. Wang, B. Khailany, W. J. Dally, and C. T. Gray, "Analog/Mixed- Signal Hardware Error Modeling for Deep Learning Inference," in Proceedings of the 56th Annual Design Automation Conference 2019. Las Vegas NV USA: ACM, Jun. 2019, pp. 1–6. [Online]. Available: https://dl.acm.org/doi/10.1145/3316781.3317770

[4] S. Yin, Z. Jiang, J.-S. Seo, and M. Seok, "XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks," IEEE Journal of Solid-State Circuits, pp. 1–11, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8959407/

[5] Z. Jiang, S. Yin, J.-S. Seo, and M. Seok, "C3SRAM: An In- Memory-Computing SRAM Macro Based on Robust Capacitive Coupling Computing Mechanism," IEEE Journal of Solid-State Circuits, vol. 55, no. 7, pp. 1888–1897, Jul. 2020, conference Name: IEEE Journal of Solid-State Circuits. [Online]. Available: https://ieeexplore.ieee.org/document/9094713

[6] H. Sun, Z. Zhu, Y. Cai, X. Chen, Y. Wang, and H. Yang, "An Energy-Efficient Quantized and Regularized Training Framework For Processing-In-Memory Accelerators," in 2020 25th Asia and South Pacific Design Automation Conference (ASP-DAC). Beijing, China: IEEE, Jan. 2020, pp. 325–330. [Online]. Available: https://ieeexplore.ieee.org/document/9045192/

[7] S. K. Gonugondla, C. Sakr, H. Dbouk, and N. R. Shanbhag, "Fundamental Limits on Energy-Delay-Accuracy of In-Memory Architectures in Inference Applications," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 41, no. 10, pp. 3188–3201, Oct. 2022, conference Name: IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems. [Online]. Available: https://ieeexplore.ieee.org/document/9598887/?arnumber=9598887

[8] Y. Kim, H. Kim, and J.-J. Kim, "Extreme Partial-Sum Quantization for Analog Computing-In-Memory Neural Network Accelerators," J. Emerg. Technol. Comput. Syst., vol. 18, no. 4, pp. 75:1–75:19, Oct. 2022. [Online]. Available: https://doi.org/10.1145/3528104

[9] S. Huang, H. Jiang, and S. Yu, "Hardware-aware Quantization/Mapping Strategies for Compute-in-Memory Accelerators," ACM Transactions on Design Automation of Electronic Systems, vol. 28, no. 3, pp. 1–23, May 2023. [Online]. Available: https://dl.acm.org/doi/10.1145/3569940

[10] N. Laubeuf, J. Doevenspeck, I. A. Papistas, M. Caselli, S. Cosemans, P. Vrancx, D. Bhattacharjee, A. Mallik, P. Debacker, D. Verkest, F. Catthoor, and R. Lauwereins, "Dynamic quantization range control for analog-in-memory neural networks acceleration," ACM Trans. Des. Autom. Electron. Syst., vol. 27, no. 5, Jun. 2022. [Online]. Available: https://doi.org/10.1145/3498328

[11] R. Genov and G. Cauwenberghs, "Stochastic mixed-signal vlsi architecture for high-dimensional kernel machines," in Advances in Neural Information Processing Systems, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., vol. 14. MIT Press, 2001. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2001/file/5352696a9ca3397beb79f116f3a33991-Paper.pdf

[12] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," Apr. 2000, arXiv:physics/0004057. [Online]. Available: http://arxiv.org/abs/physics/0004057

ACKNOWLEDGEMENT

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) program Massive Cross-Correlation (MAX) through Air Force Research Laboratory Contract No. FA8650-23-C-7306. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

Institute for Neural Computation
UC San Diego JACOBS SCHOOL OF ENGINEERING
NORTHROP GRUMMAN