Statistical Robustness of Markov Chain Monte Carlo Accelerators Extended Abstract

Xiangyu Zhang, Ramin Bashizade, Yicheng Wang, Sayan Mukherjee, Alvin R. Lebeck

Duke University

1. Motivation

Statistical machine learning often uses probabilistic algorithms, such as Markov Chain Monte Carlo (MCMC) methods, to solve a wide range of problems. As alternatives to Deep Neural Networks, these algorithms provide easier access to interpreting why a given result is obtained through their model transparency and statistical properties. The algorithms, often considered too slow on conventional processors due to inefficient sampling process, can be accelerated with specialized hardware by exploiting parallelism and utilizing various hardware approximations for efficiency, such as reducing bit representation, truncating small values to zero, or simplifying the random number generator. Understanding the influence of these approximations on the correct execution of target algorithms is crucial to meet the quality requirement. A common approach to evaluating correctness is to compare the end-point result quality ("accuracy") against accurately-measured or hand-labeled ground-truth data using community-standard benchmarks and metrics: the hardware execution is considered to be correct if it provides comparable "accuracy" to the software-only implementations that do not have these approximations. Statistical guarantees can be made on the end-point results [6, 9].

However, domain experts in statistics, especially Bayesian Inference are interested in the full distribution of possible results rather than a single-point estimate, including both end-point results and quantified uncertainty—statistical properties of the full distribution, a.k.a., *statistical robustness*. Hardware and/or software solutions should get both aspects correct. Therefore, in the domain of probabilistic computing/algorithms, correctness is defined by more than the end-point result of executing the algorithm, and includes statistical robustness. Statements and guarantees on the application end-point results are necessary but not sufficient to claim correctness. Failure to adequately account for domain-defined correctness can have adverse or catastrophic outcomes, such as a surgeon failing to completely remove a tumor due to incorrect uncertainty in a segmented image [2, 8]. Furthermore, measuring end-point results may not always be possible as ground-truth data is not always accessible. Therefore, *a probabilistic architecture should provide some measure (or guarantee) of statistical robustness.*

This project is supported in part by Intel, the Semiconductor Research Corporation and the National Science Foundation (CNS-1616947).

2. Limitations of the State of the Art

Current methodologies for evaluating probabilistic accelerators are often incomplete or adhoc in evaluating correctness, mostly focusing only on end-point result quality. Previous work addresses some statistical metrics for MCMC accelerators, such as KL-divergence and QQ plots [7], ESS/second [5], and goodness of fit statistical tests [3, 10]. These metrics consider limited aspects of statistical robustness: each metric only addresses one of the three pillars summarized in Sec. 4. *A comprehensive correctness evaluation methodology for a probabilistic architecture is needed and has yet to be proposed in consideration of both end-point result quality and statistical robustness.*

3. Key Insights

This work brings two key insights:

- Only measuring end-point result quality is insufficient to surface the design issues and thus a comprehensive evaluation on statistical robustness is necessary. An MCMC accelerator [12] can achieve good application end-point result quality but compromised statistical robustness. Applications need to run more iterations on the accelerator to achieve satisfactory statistical robustness, reducing the effective speedup.
- Naively applying existing popular metrics from domain experts is problematic in the evaluated applications (stereo vision and motion estimation). The metrics should be modified to account for high dimensionality of the target applications and random variables with zero empirical variance.

4. Main Artifacts

This work takes a first step toward defining metrics and a methodology for quantitatively evaluating correctness of probabilistic accelerators beyond end-point result quality. We propose three pillars of statistical robustness: 1) *sampling quality*, 2) *convergence diagnostic*, and 3) *goodness of fit*. Each pillar has at least one quantitative empirical metric, does not require ground-truth data, and collectively these pillars enable comparison of specialized hardware to a target precision, such as a 64-bit floating-point (FP64) software implementation. We expose several challenges with naively applying existing popular metrics for our purposes, including: high dimensionality of the target applications, and random variables with zero empirical variance. Therefore, we modify the existing methodologies for sampling quality and convergence diagnostic, and propose

a new metric for convergence diagnostic. Below is a summary of each pillar.

Pillar I) Sampling Quality. The intrinsic nature of MCMC methods creates dependency between samples. A sufficient number of independent samples are needed to converge and produce high-quality results. We use *Effective Sample Size* (ESS) [5, 11] to measure the number of independent samples drawn from an MCMC run, and report the arithmetic mean as a scalar metric. The existing method does not consider a practically possible case that a random variable produces empirically zero variance. We modified the method to report "overall" and "active" ESS values separately to account for possible biases. Low ESS indicates that more iterations may be required to generate sufficient independent samples.

Pillar II) Convergence Diagnostic. The total running time of an MCMC run is determined by when it converges. Convergence can be measured by Gelman-Rubin's $\hat{R}$ [1], but this metric is undefined for variables with zero variance. Therefore, we propose a process to determine convergence that accounts for zero variance and a new metric—*convergence percentage*—based on $\hat{R}$, to measure the total percentage of converged results. Low convergence percentage indicates that more iterations are required for the model to converge.

Pillar III) Goodness of Fit. In the absence of ground-truth data (labeled data), it is important to understand the differences between the baseline precision (e.g, FP64) and hardware end-point results to evaluate the overall quality of the hardware. We provide two "goodness of fit" approaches: 1) Root Mean Squared Error (RMSE) on application specific data relative to a baseline reference, and 2) Jensen-Shannon Divergence (JSD) [4] to evaluate all possible data inputs in the binary label case and provide the worst-case distribution divergence.

We apply our framework on an representative MCMC accelerator—Stochastic Processing Unit (SPU) [12]—and demonstrate the benefits of using this framework on design space exploration. We implemented our framework, the FP64 software implementation, and a functionally equivalent SPU simulator in MATLAB for statistical robustness analysis. We also implemented SPU in Verilog, Chisel, and HLS all with verified results, for FPGA/ASIC resource usage analysis for design space exploration.

5. Key Results and Contributions

We summarize two most important results:

- A representative accelerator [12], with limited precision and other approximation techniques, achieves the same application end-point result quality as FP64-software, confirming the previous work, but differs from FP64-software with a lower ESS and a lower convergence percentage. Filling the gap requires 2× more iterations on the accelerator, reducing the accelerator's effective speedup.
- A considerable improvement in statistical robustness, comparable to FP64-software, can be achieved by slightly increasing the bit precision from 4 to 6 and removing an approximation technique, with only 1.20× area and 1.10× power overhead, without the commensurate overhead of FP64.

Below summarizes our contributions:

- We proposed a three-pillar framework, to our knowledge, the first attempt to a comprehensive methodology for quantitatively evaluating correctness of probabilistic accelerators in considerations of both end-point result quality and statistical robustness. Previous work [3, 5, 7, 10] belongs to one of three proposed pillars and we argue (main paper Sec. 4 and Sec. 5) all three pillars are needed to fully characterize statistical robustness of an MCMC accelerator.
- We expose challenges with directly applying the existing popular metrics. Thus we modified the existing methodologies in sampling quality and convergence diagnostic and propose a new metric (convergence percentage) in convergence diagnostic.
- We apply our framework to a representative MCMC accelerator and surface design issues that cannot be exposed using only application end-point result quality.
- We demonstrate the benefits of this framework to guide design space exploration in an MCMC accelerator to achieve a design point with satisfactory statistical robustness, avoiding FP hardware overheads.

6. Why ASPLOS

Domain-specific accelerators require performance and correctness evaluation in consideration of the full stack from algorithm to implementation. In the domain of probabilistic computing/algorithms, correctness is defined by more than the end-point result of executing the algorithm, and includes additional statistical properties. A comprehensive methodology to evaluate these properties of a probabilistic architecture is needed and has yet to be defined. Our work takes the first step toward such a methodology. The proposed pillars can inform end-users by characterizing existing hardware and inform hardware designers by using the pillars for design space exploration. We believe our interdisciplinary work in architecture, probabilistic computing/algorithms, and statistics is a good fit to ASPLOS by supporting the community to build efficient and statistically robust hardware/systems.

7. Citation for Most Influential Paper Award

The paper "Statistical Robustness of Markov Chain Monte Carlo Accelerators" argues a comprehensive evaluation on statistical robustness in addition to application end-point result quality is necessary when evaluating probabilistic accelerators. The paper provides a methodology for quantitatively evaluating correctness of MCMC accelerators, including three pillars of statistical robustness: 1) sampling quality, 2) convergence diagnostic, and 3) goodness-of-fit. The paper for the first time brings the notion of "statistical robustness" into the process of designing robust probabilistic accelerators.

References

[1] Stephen P. Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

[2] Wenjun Cheng, Luyao Ma, Tiejun Yang, Jiali Liang, and Yan Zhang. Joint lung ct image segmentation: a hierarchical bayesian approach. *PloS one*, 11(9), 2016.

[3] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[4] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[5] Shuanglong Liu, Grigorios Mingas, and Christos-Savvas Bouganis. An exact mcmc accelerator under custom precision regimes. In *2015 International Conference on Field Programmable Technology (FPT)*, pages 120–127. IEEE, 2015.

[6] Divya Mahajan, Amir Yazdanbakhsh, Jongse Park, Bradley Thwaites, and Hadi Esmaeilzadeh. Towards statistical guarantees in controlling quality tradeoffs for approximate acceleration. *ACM SIGARCH Computer Architecture News*, 44(3):66–77, 2016.

[7] Vikash Mansinghka and Eric Jonas. Building fast bayesian computing machines out of intentionally stochastic, digital parts. *arXiv preprint arXiv:1402.4914*, 2014.

[8] Patrick McClure, Nao Rho, John A. Lee, Jakub R. Kaczmarzyk, Charles Y. Zheng, Satrajit S. Ghosh, Dylan M. Nielson, Adam G. Thomas, Peter Bandettini, and Francisco Pereira. Knowing what you know in brain segmentation using bayesian deep neural networks. *Frontiers in Neuroinformatics*, 13:67, 2019.

[9] Jongse Park, Emmanuel Amaro, Divya Mahajan, Bradley Thwaites, and Hadi Esmaeilzadeh. Axgames: Towards crowdsourcing quality target determination in approximate computing. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, ASPLOS '16, page 623–636, New York, NY, USA, 2016. Association for Computing Machinery.

[10] Feras A. Saad, Cameron E. Freer, Nathanael L. Ackerman, and Vikash K. Mansinghka. A family of exact goodness-of-fit tests for high-dimensional discrete distributions. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1640–1649, 2019.

[11] Madeleine B. Thompson. A comparison of methods for computing autocorrelation time. *arXiv preprint arXiv:1011.0175*, 2010.

[12] Xiangyu Zhang, Ramin Bashizade, Craig LaBoda, Chris Dwyer, and Alvin R. Lebeck. Architecting a stochastic computing unit with molecular optical devices. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, pages 301–314. IEEE, 2018.