Scalable and Efficient Probabilistic Inference for Bayesian Deep Learning and Generative Modeling

Ruqi Zhang
Purdue University
610 Purdue Mall
West Lafayette, IN 47907 USA
ruqiz@purdue.edu

Abstract

Probabilistic inference is a fundamental challenge in machine learning, spanning tasks from approximate Bayesian inference to generative AI. In this talk, I will present theoretically-guaranteed scalable and efficient probabilistic inference with applications in Bayesian deep learning and generative modeling. First, I will introduce a new compute paradigm for probabilistic inference that leverages modern accelerators, specifically low-precision and sparsity, to significantly speed up inference while preserving accuracy. Next, I will present a new framework for efficient inference in discrete domains, utilizing gradient information—a largely overlooked feature of discrete distributions—to enable more informed and directional exploration. Finally, I will showcase experimental results demonstrating the effectiveness of these methods across various ML tasks, including Bayesian neural networks, energy-based models, and large language models.

Overview

My research centers on probabilistic machine learning, with a particular focus on probabilistic inference for uncertainty quantification and generative modeling. My approaches have been applied to different models such as graphical models, Bayesian neural networks, energy-based models, diffusion models, and large language models, addressing critical applications in trustworthy AI, such as uncertainty estimate, privacy, and alignment.

Fast Probabilistic Inference by Low-precision Compute and Sparsity We develop fast and scalable probabilistic inference by leveraging advancements in accelerators, specifically low-precision compute and sparsity. Zhang, Wilson, and De Sa (2022) introduces the first framework for low-precision probabilistic inference algorithms, demonstrating that inference costs can be significantly reduced without sacrificing accuracy. Besides, our work (Li et al. 2024b) introduces a fully sparse probabilistic inference method that maintains a consistently high sparsity ($>$90%) through the entire inference process.

Gradient-based Discrete Probabilistic Inference Our recent work introduces a new framework for discrete inference that utilizes gradient information, achieving significant improvements over traditional discrete inference methods (Zhang, Liu, and Liu 2022; Pynadath et al. 2024). We introduced Discrete Langevin Proposal (DLP), the first discrete analog of the Langevin algorithm, which efficiently explores high-probability regions and updates all coordinates in parallel. Notably, our framework sets a new standard for efficient discrete inference techniques, showing superior results across various applications, including Ising models, restricted Boltzmann machines, graphical models, binary neural networks, and discrete generative models.

Applications in Trustworthy AI: Uncertainty Quantification and Alignment We improved Bayesian neural network (BNN) generalization through a geometric approach, creating an auxiliary variable inference algorithm (Li and Zhang 2024). For language model alignment, we introduced a decoding-time alignment method based on probabilistic inference that ensures high-reward, ethical responses aligned with human preferences (Li et al. 2024a; Ding, Li, and Zhang 2025; Pynadath and Zhang 2025).

References

Ding, Y.; Li, B.; and Zhang, R. 2025. ETA: Evaluating Then Aligning Safety of Vision Language Models at Inference Time. *ICLR*.

Li, B.; Wang, Y.; Grama, A.; and Zhang, R. 2024a. Cascade reward sampling for efficient decoding-time alignment. *arXiv*.

Li, B.; and Zhang, R. 2024. Entropy-MCMC: Sampling from Flat Basins with Ease. *ICLR*.

Li, J.; Miao, Z.; Qiu, Q.; and Zhang, R. 2024b. Training Bayesian Neural Networks with Sparse Subspace Variational Inference. *ICLR*.

Pynadath, P.; Bhattacharya, R.; Hariharan, A.; and Zhang, R. 2024. Gradient-based Discrete Sampling with Automatic Cyclical Scheduling. *NeurIPS*.

Pynadath, P.; and Zhang, R. 2025. Controlled LLM Decoding via Discrete Auto-regressive Biasing. *ICLR*.

Zhang, R.; Liu, X.; and Liu, Q. 2022. A Langevin-like Sampler for Discrete Distributions. In *ICML*.

Zhang, R.; Wilson, A. G.; and De Sa, C. 2022. Low-Precision Stochastic Gradient Langevin Dynamics. In *ICML*.