Closing Report

In order to enable the use of probabilistic models at Amazon at scale, we proposed to exploit synergies between the Clay probabilistic language and the distributed computation effort called Ray that is currently being carried out at UC Berkeley. Clay is a tool for Large Scale Probabilistic Modeling and Inference developed at Amazon by CoreAI. It was targeting single-machine multi-GPU instances, but had a flexible architecture that allows adding new backends with very limited and manageable effort. Ray is a system for distributed computation developed at Berkeley. This project advanced the idea of joining efforts of the two groups, by embedding a backend targeting Ray in order to enable massively distributed probabilistic computing on very-large-scale Datasets at Amazon and beyond. Applications are inventory placement, product pricing, and other business problems that operate on very large datasets.

The first milestone finished as of 8/1/2020 investigated and implemented chain-parallel distribution of Hamiltonian- Monte-Carlo (HMC) inference over autoscaling clusters of GPU EC2 instances managed by Ray. As HMC chains are evaluated independently this method involves very little communication between cluster nodes, and efficient parallelisation. A drawback of this simple scheme is however that nodes that are completing their workload early might run idle waiting for nodes with slower convergence. More efficient schemes are to be evaluated in following milestones, both for sampling (HMC) and approximating inference methods (Laplace Approximation, Variational Inference).

The milestone goal of proving feasibility and evaluation of set-up overhead has been achieved, in tests the system works reliably and meets performance and scaling expectations. The implementation contributed by our BAIR PI has been integrated into the core Clay library, including automated regression testing ensuring long-term stability and compatibility. An introductory Clay+Ray example notebook is shipped as part of the Clay release for AMLC.

The 2nd milestone finished as of 3/1/2021 focuses on accelerating inference for models with close counterparts in production models (particularly CoreAI Berlin's ongoing CPM and PES engagements). The objective is to overcome GPU memory limitations that currently restrict use cases to single GLs or predefined subsets of ASINs, or require to split learning on typically very large datasets into parts, hampering links in model hierarchy, and requiring significant manual intervention for data preparation. We have implemented and integrated distributed logistic regression and lbfgs solver in our library. We implemented a partitioning function that finds the best possible axis to split the clay graph for distributed execution on multiple GPU nodes. The preliminary results are promising and paves a way for implementing ray based distributed computation to other inference methods. With these efforts we pave the way for high-capacity models that enable much higher precision and sensitivity with regard to customer behaviour in pricing and rebate initiatives for individual ASINs. Contributions by BAIR to our business objectives are algorithmically in nature, enabling straightforward and efficient evaluation on synthetic benchmarks and datasets published in the scientific literature. The researched distributed methods evaluate favourably compared to the state of the art, and business decision logic necessarily benefits.