Quantum Enhanced Sampling for Reinforcement Learning with energy based models*

Sujay Srivastava*[3], Arun Vellat Sadashivan[1], Rajarsi Pal[3], and Shailesh Kumar[2]

[1] Jio Platforms Ltd, Bengaluru, India
[2] Jio Platforms Ltd, Hyderabad, India
[3] Indian Institute of Technology Madras, Chennai, India

Keywords: Quantum sampling· Sampling from Boltzmann distribution· Reinforcement Learning · Energy-based models

1 Introduction

Reinforcement Learning (RL) is a paradigm where agents learn optimal behaviors through interactions with an environment, receiving rewards or penalties to maximize reward over time [1]. Deep Reinforcement Learning (Deep RL) combines deep learning techniques with RL, utilizing neural networks to approximate complex value functions and policies, enabling agents to handle high-dimensional state and action spaces. Deep energy models for reinforcement learning have shown advantages over standard deep RL machinery in learning performance, but they face certain computational bottlenecks [3].

We employ a novel quantum approach to tackle the sample, employing the quantum-enhanced Markov Chain Monte Carlo (MCMC) method to improve the efficiency and performance of energy-based models. Our contributions include:

- Incorporating quantum subroutines in deep energy-based Reinforcement Learning algorithm and highlighting the potential advantages for action spaces.
- Implementing a classical version corresponding to the quantum-enhanced sampling procedure and experimentally demonstrating a comparison of both classical and quantum methods in a 4x4 Grid-World environment.

Next, we briefly explain the hybrid training algorithm and discuss our experimental setup and results. Finally, we conclude with insights and future directions.

2 Methodology

We use function approximations like neural networks for training RL model because they can generalize Q-value over unobserved states and actions and hence suitable for larger state and action space. For real-life problems where the model's optimal policy can be very complex, we need neural networks with high expressive power that can learn a complex policy. Here, we consider deep energy-based networks(DEBN), which are neural network model the empirical probability distribution of observed data vectors v, where v is string of binary variables. We use this model to propose the next action to be taken probabilistically, and update the RL policy(train the model) according to the rewards. However, there are some bottlenecks in the training process using Deep Energy Models.
We need to sample the action according to the probability:

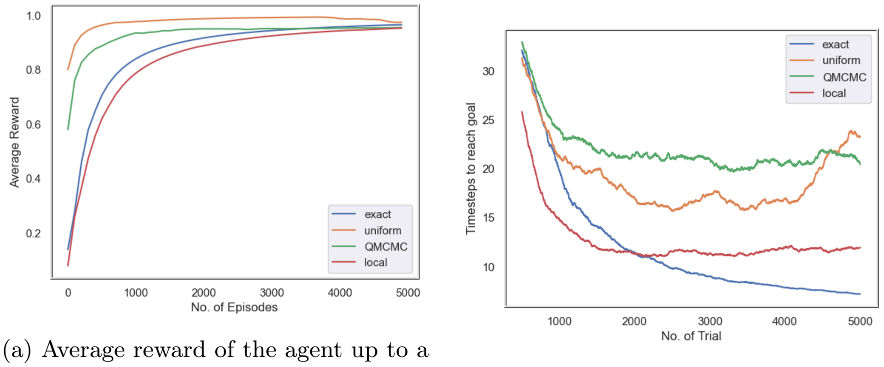
$$\pi_\theta(\mathbf{a}|\mathbf{s}) = \mathbb{P}(\mathbf{a}|\mathbf{s}) = \frac{e^{-\beta F_\theta(s,a)}}{\sum_{a'} e^{-\beta F_\theta(s,a')}} \tag{1}$$

* *Sujay Srivastava's contribution to this work was while he was employed as an intern at Jio Platforms Ltd., Bengaluru, India

where, $F_\theta(s,a)$ is the DEBN output over given state s and action a as input. Finding the probability distribution involves finding $\sum_{a'} e^{-\beta F_\theta(s,a')}$ which is at least $O(|A|)$, making it infeasible for larger action space.

Our approach involves a hybrid algorithm for training deep energy-based models (DEBNs) with quantum-enhanced sampling techniques. The energy-based models utilize a tailored version of the quantum-enhanced Markov Chain Monte Carlo(MCMC) to sample from the Boltzmann distribution efficiently.[5] By using the network weights and mapping them to an Ising model for sampling, we can perform quantum sampling in choosing action in an Boltzmann Machine architecture. This hybrid algorithm aims to improve the learning performance of DEBNs in large action space environments. We train a Deep-Energy-based network for reinforcement learning tasks by adding quantum subroutines that can create a quantum advantage to speed up our algorithm. They are:

1. ***Subroutine 1***- To sample the policy $\pi(\mathbf{a_t}|\mathbf{s_t})$ given the state $\mathbf{s_t}$ from the approximated distribution $\frac{e^{\beta M^{\theta}(\mathbf{s_t},\mathbf{a_t})}}{Z_\beta(\mathbf{s_t})}$.
2. ***Subroutine 2***- Explicit numerical estimation of the merit(or the free energy) function $M^{\theta}(\mathbf{s_t},\mathbf{a_t})$ of the model. This can be used for Deep-Boltzmann Machines(DBMs), where we need to sample hidden nodes to approximate merit function.[2]
3. ***Subroutine 3***- The evaluation of the gradient of the merit function $\nabla_\theta M^{\theta}(\mathbf{s_t},\mathbf{a_t})$ which is essential for the parameter updating part of the algorithm.



(a) Average reward of the agent up to a particular episode number. Model receives +1 for completing the task within 50 steps, otherwise -1

(b) Number of timesteps taken by agent to complete the episode(or trial), 50 in case episode not completed

Fig. 1: Plots showing performance using different sampling techniques(local, uniform, quantum, and exact method) in the 4x4 GridWorld Task

3 Experiments and Results

To see the impact of different approximate action sampling techniques in the model performance, we perform RL experiment in a in a 4x4 GridWorld environment and compare the results with quantum sampling on a RBM neural network architecture with classical sampling techniques for choosing action. The actions chosen are performed and recorded, and the weights in the model are trained on these recorded interactions. The model weights are extracted to create the Ising model, which is used for quantum-enhanced action sampling. Except for the sampling-specific parameter, the model parameters for each sampling technique are kept the same. For quantum experiments, we have used the Qulacs[4] simulators.

For our GridWorld simulation, we used a single-layer neural network with state nodes(two-hot encoded for x and y coordinates) and action nodes(binary encoded) as input and 5 units in the layer.
Figure 1a and 1b shows the results of the experiments, comparing average reward and number of timesteps required to reach goal.
The exact method of choosing an action (brute-force) converges to the optimal solution. In contrast, the approximate algorithm finds the solution (shown by the average reward that they are reaching the goal state consistently), but doesn't converge to the most optimal solution, as evident from the average time-steps to reach goal. A larger experiment environment must be tried to show a significant difference between the quantum sampling and other approximate solutions based on classical Monte Carlo sampling. In our experiments, the size of the visible space (in this case, the action space) is rather small, which offsets the advantages we would otherwise expect from a quantum sampling algorithm compared to a classical one. It would be possible to run larger experiments with access to better hardware.

4 Conclusion

In this paper, we talked about how quantum-enhanced sampling can be used for reinforcement learning tasks, specifically QMCMC for sampling actions in Deep-energy based networks (DEBN). Further, we have proposed a method to implement quantum sampling with the Deep Boltzmann Machine framework for evaluation of model output and gradient along with choosing action. Our exploration of quantum enhancements demonstrates avenues for overcoming the classical barriers that deep energy-based models face for solving complex RL tasks, promising to advance the state-of-the-art in reinforcement learning.

References

[1] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* Cambridge, MA, USA: A Bradford Book, 2018. ISBN: 0262039249.

[2] Daniel Crawford et al. *Reinforcement Learning Using Quantum Boltzmann Machines.* 2019. arXiv: 1612.05695 [quant-ph].

[3] Sofiene Jerbi et al. "Quantum Enhancements for Deep Reinforcement Learning in Large Spaces". In: *PRX Quantum* 2 (1 Feb. 2021), p. 010328. DOI: 10.1103/PRXQuantum.2.010328. URL: https://link.aps.org/doi/10.1103/PRXQuantum.2.010328.

[4] Yasunari Suzuki et al. "Qulacs: a fast and versatile quantum circuit simulator for research purpose". In: *Quantum* 5 (Oct. 2021), p. 559. ISSN: 2521-327X. DOI: 10.22331/q-2021-10-06-559. URL: http://dx.doi.org/10.22331/q-2021-10-06-559.

[5] David Layden et al. "Quantum-enhanced Markov chain Monte Carlo". In: (Mar. 2022). arXiv: 2203.12497 [quant-ph].