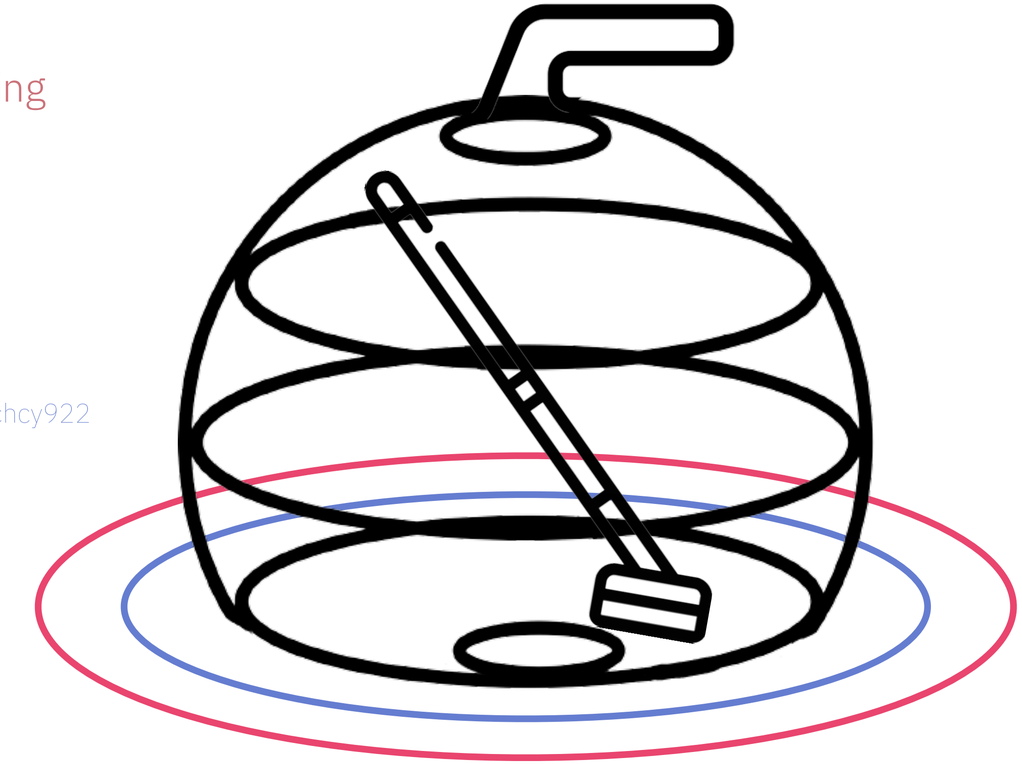
Q-urling

Curling Strategy Prediction using Quantum Approach in Reinforcement Learning

Jaeun Kim @jaeunkim
Chiyoon Kim @chcy922
Kyoung Keun Park @johnparkn
Kyungmin Lee@anfry15rudals
Jeonghoon Lee @WestGround

(Presenter) Chiyoon Kim @chcy922



Contents

About Curling

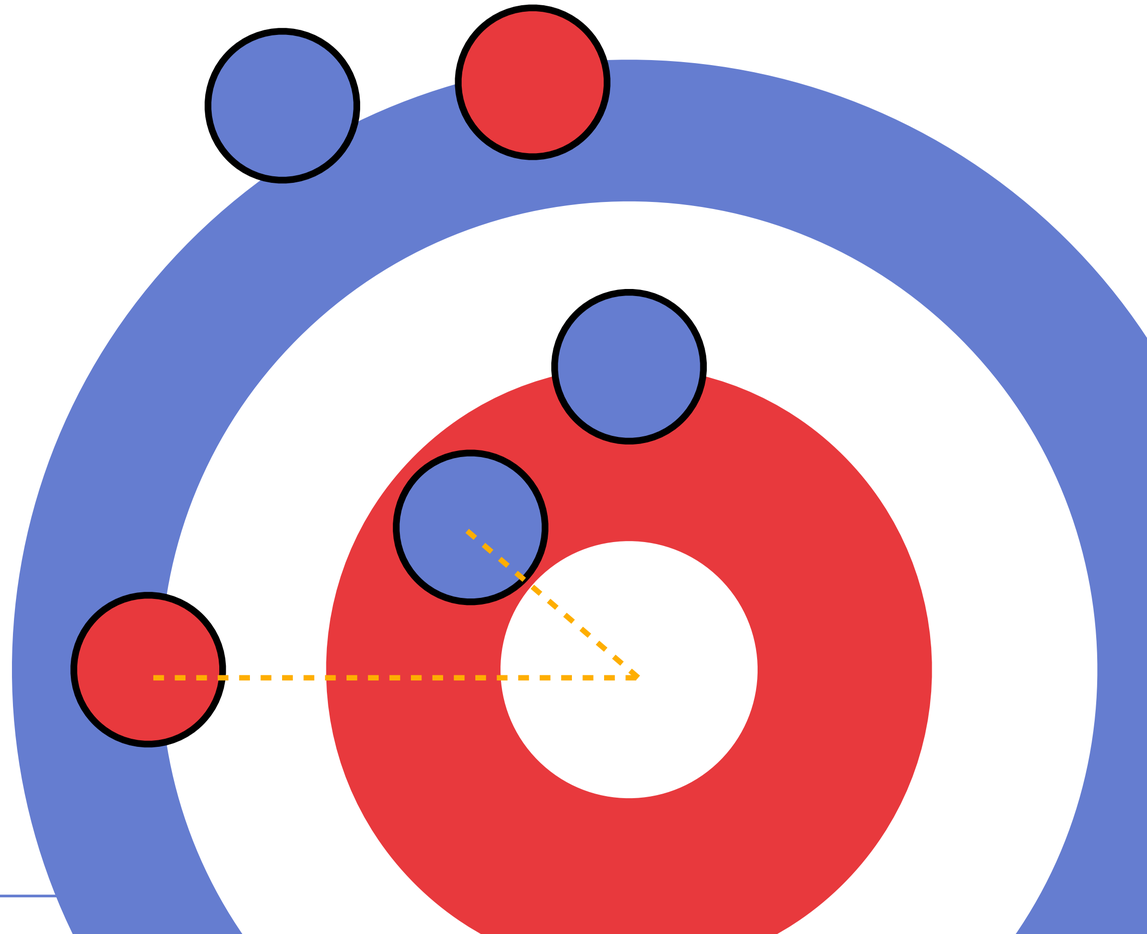
ENDS

Rules of Curling

- Put your stone close to the center of the house
- If your stone is closer to the center than your opponent, you win

Strategies of Curling

- Defensive: try to maintain the score difference
- Offensive: try to gain as many points, even though it takes risks



Predicting the appropriate Strategy in a Curling Game

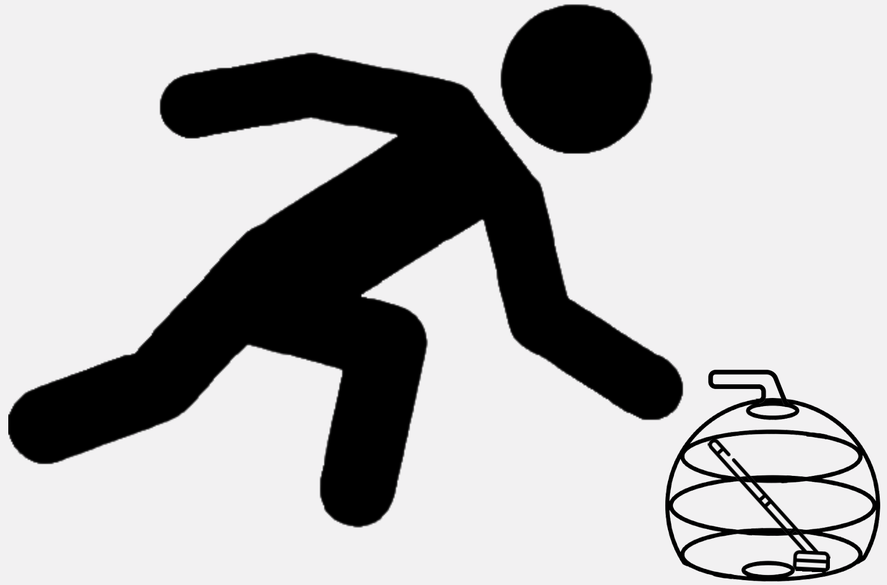
with Qiskit!

Motivation

- Use hybrid classical-quantum reinforcement learning model
- Curling's strategy can be reduced to relatively simple states (offensive/defensive)

Goal: Q-urling model

- Predict the appropriate strategy for winning a curling game using a quantum approach in Reinforcement Learning



Reinforcement Learning and Policy Gradient

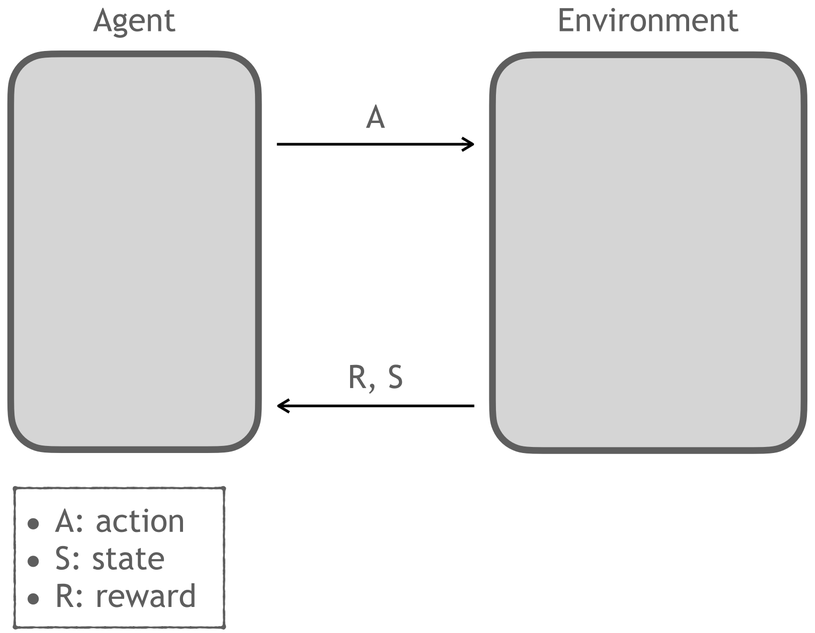
Background

Reinforcement Learning

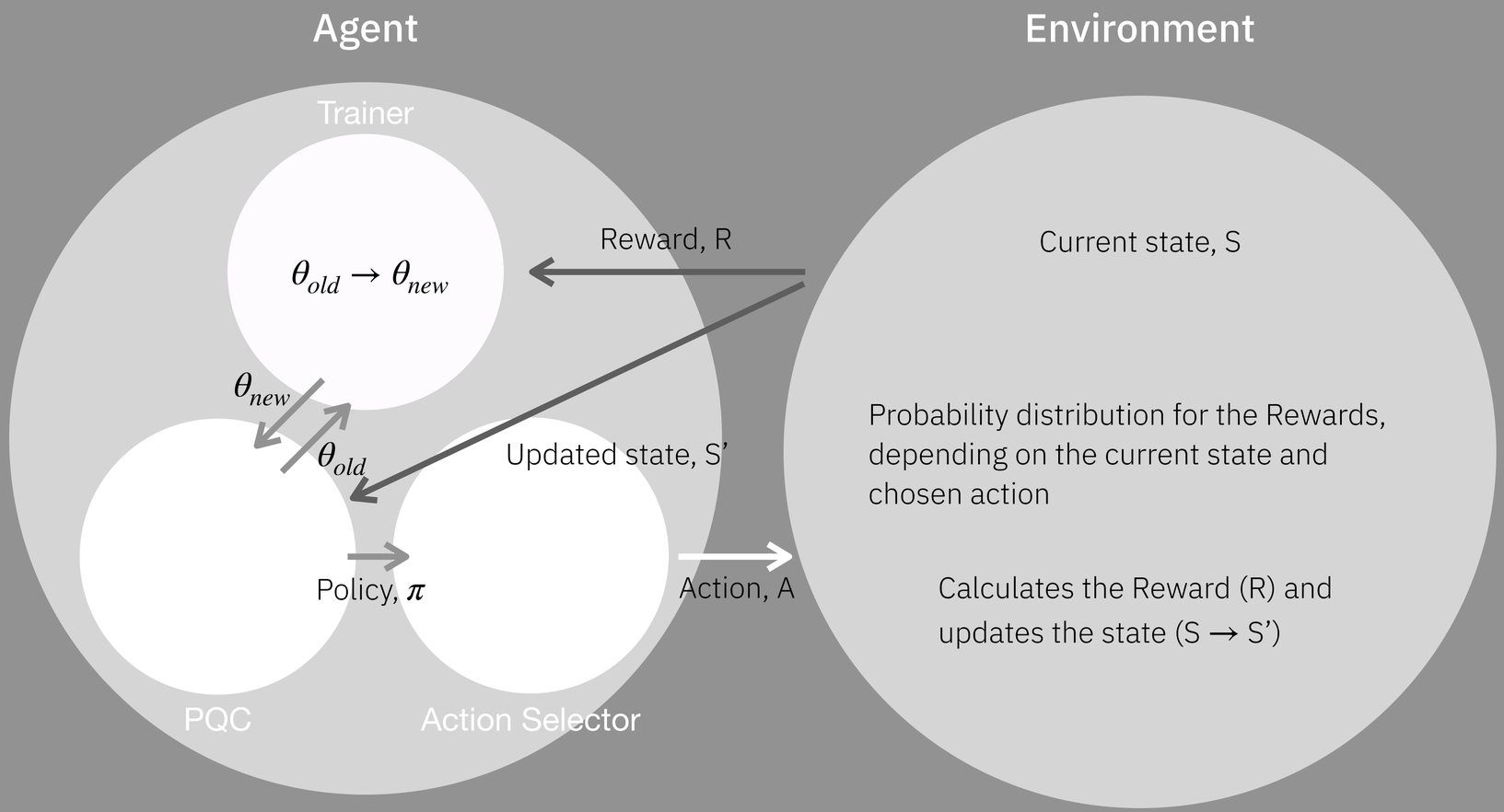
- Agent: Returns “action” to environment and proceeds training to maximize the total rewards from the environment
- Environment: Returns reward and next step to the agent, corresponding to the taken action and the current state.
- Policy: The pattern (probability distribution) of actions by the agent, at a given state

Policy Gradient

- Reward Function, $J(\theta)$: total expected reward, acts as the loss function of the learning process
- Use Gradient Descent on the Reward Function, gradient by parameter θ, $\nabla_\theta J(\theta) \propto E_{\pi_\theta} G_t \nabla_\theta ln[\pi_\theta(a \mid s)]$
- Update the parameter: $\theta \leftarrow \theta + \alpha G_t \nabla_\theta ln[\pi_\theta(a \mid s)]$



Q-urling Model - Overview



Q-urling Model - Agent and Environment

Design of the model for the Curling game

States and Actions of Q-urling

- State vector, S: |S| = 2 → (1) winning/losing/draw, (2) go first / go later
- Action, A: two available actions (strategies) → (1) Offensive, (2) Defensive

Environment Design

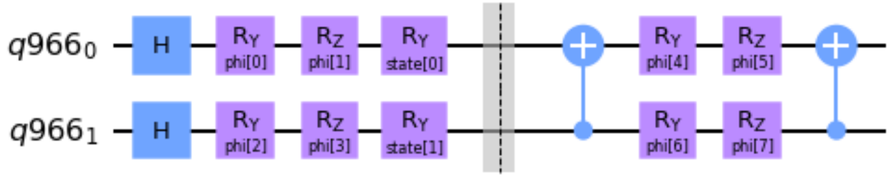
- Available score difference from one trial (end) → -3, -2, -1, 0, 1, 2, 3
- Probability for available score difference, depending on (1) state: go first / go later, (2) actions: Offensive / Defensive
- The result of each game (end) is determined stochastically, following given probability distribution
- The reward is defined by the result of the end (if win, return 1 / if lose, return 0)

Process of returning Reward

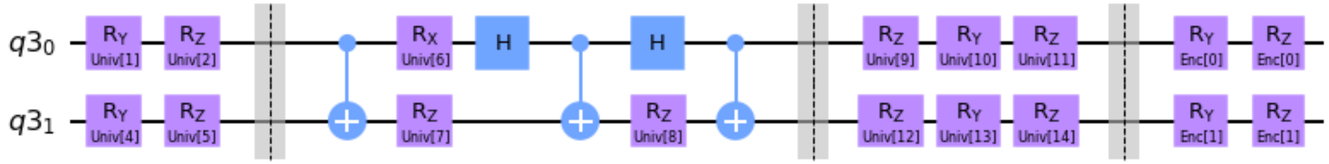
1. Get the current status, S from the environment (go first / go later) and the proceeding action from the agent
2. Do the stochastic calculation using the probability distribution of the score difference
3. Obtain the result (win/lose) as output
4. Update state and calculate reward
5. Return updated state to PQC / Return reward to Trainer

Q-urling Model - Policy

Policy of the Agent mapped on Quantum Circuit, the parameterized quantum circuit (PQC)

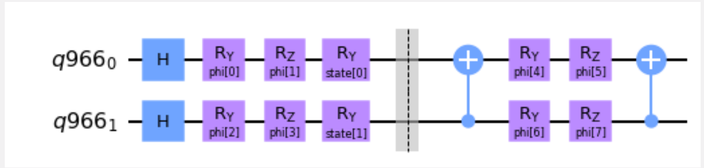


PQC policy composed of tensor product of Pauli Z operators

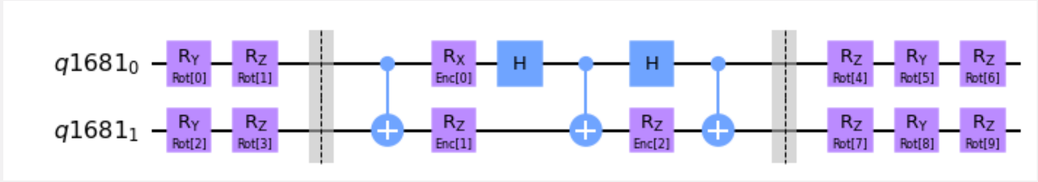


SOFTMAX-PQC policy

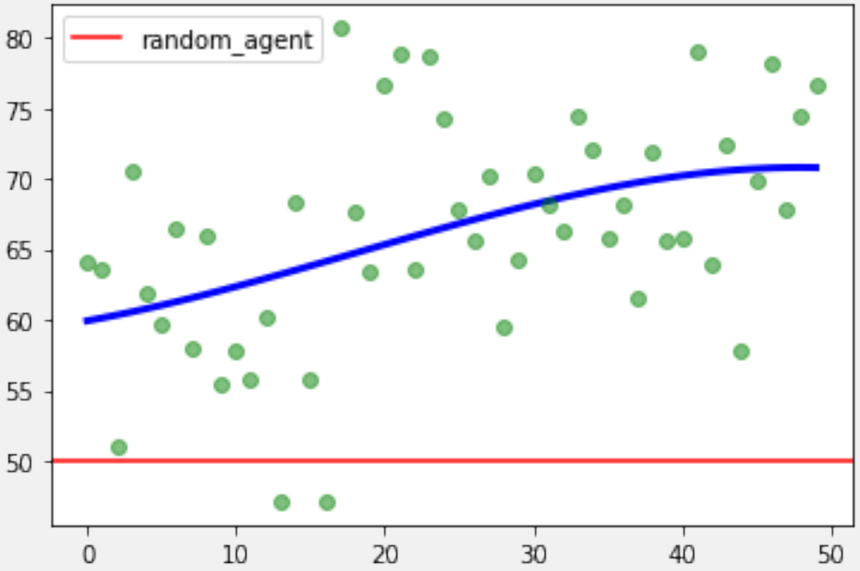
Results

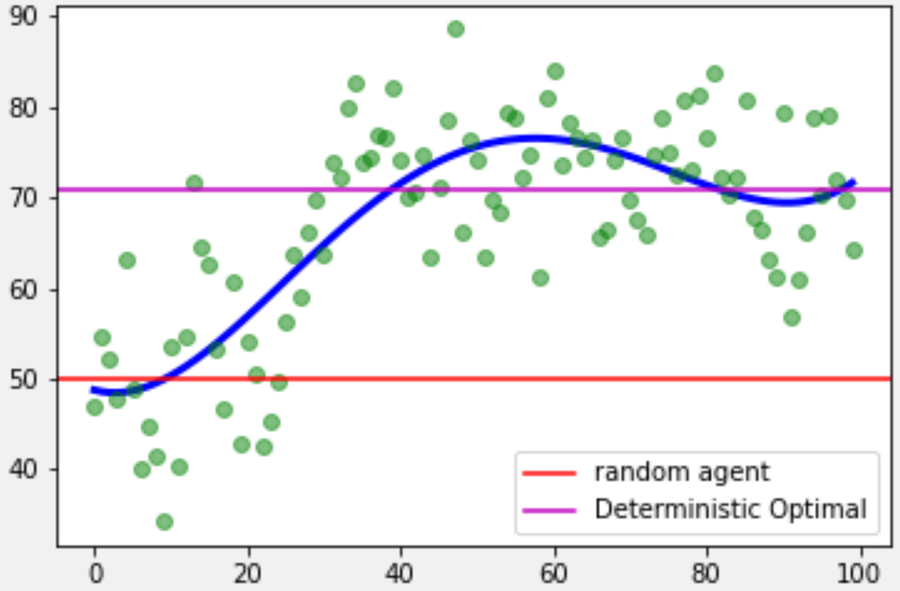


PQC polices composed of tensor product of Pauli Z operators



SOFTMAX-PQC polices





Conclusion

REFERENCES

- Kiwook Bae, K., Park, D., Kim, D., Shin, H., (2016). *Markov Decision Process for Curling Strategies,* Journal of the Korean Institute of Industrial Engineers, Vol. 42, No. 1, pp. 65-72. http://dx.doi.org/10.7232/JKIIE.2016.42.1.065
- Jerbi, S., Gyurik, C., Marshall S. C., Briegel, H. J., Dunjko, V., (2021). *Parameterized Quantum Policies for Reinforcement Learning,* arXiv:2103.05577
- P. B. M. de Sousa, Ramos, R. V., (2006) *Universal quantum circuit for n-qubit quantum gate: A programmable quantum gate*, QuantumInformation & Computation, Vol. 7, Issue 3, pp 228 - 242
- Nielsen, M.A. & Chuang, I.L., 2011. *Quantum Computation and Quantum Information: 10th Anniversary Edition*, Cambridge University Press.

Thank You