Evaluating Model-Free Algorithms for Learning World Value Functions

Abigail Naicker, Supervised by Dr Steven James and Dr Benjamin Rosman

University of the Witwatersrand, Johannesburg

Introduction

- What is the performance of the different model-free algorithms for learning world value functions in a simulated environment?
- Reinforcement Learning deals with the development of algorithms that allow an agent to learn how to decide on decisions in a setting where some idea of reward is maximized.
- Recent research proposed a goal-oriented function, world-value functions
- Choosing the right model-free algorithm for a specific RL task can be challenging.
- This WVF is expressed like this:

$$Q(s,g,a) = \mathbb{E}_s\left[R(s,g,a,s_1) + \sum_{t=1}^{\infty} \gamma R(s_t,g,a_t,s_{t+1})\right]$$

Model-Free Algorithms

- They refrain from using the transition probability distribution and reward function related with the MDP.
- They estimate the value function related to experience of the agent in the environment.
- State-Action-Reward-State-Action (SARSA)
- Deep-Q Network (DQN)
- Q-Learning
- Soft Actor Critic (SAC)

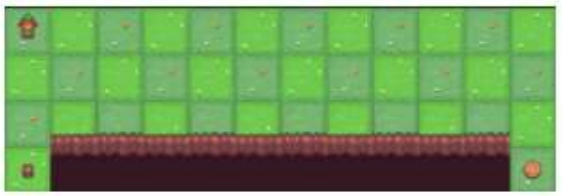

Figure 1. Cliff-walking domain used for SARSA & Q-learning

World-Value Functions

- These extend the concept of value functions in reinforcement learning by incorporating a broader notion of value, including not just the state-action pairs but the entire environment.
- They provably encode how to reach all achievable goals
- Defined by the agents pseudo-reward function:

$$R(s,g,a,s') = \begin{cases} R_{min}, & \text{if } g \neq s \text{ and s' is terminal} \\ R(s,a,s'), & \text{otherwise} \end{cases}$$

Comparison of Algorithms

- SARSA generally performs better as seen in figure 2
- SARSA agent will be cautious in its actions as it wont want to get near the cliff and fall off. Whereas the q-learning agent is an off-policy method and this will want to take more risks and explore more
- Agent showed improvement in learning the WVF's. The agent can balance the exploitation-exploration trade-off of which this can show that this agent can be a choice that is used in different environments.
- In DQN, the score, which is the number of time steps the agent managed to keep the pole balanced, fluctuated up and down, from episode 6 it increased
- Increasing the learning rate affects the score negatively as the score is lower in every episode

Episode	Score	Exploration Rate
1	40	0.96
2	15	0.89
3	17	0.82
4	8	0.79
5	33	0.67
6	13	0.62
7	21	0.56
8	25	0.50
9	34	0.42
10	52	0.32

Figure 1. DQN Results

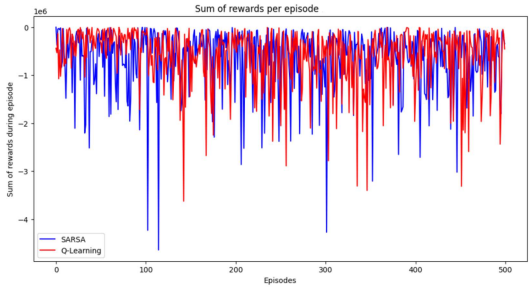

Figure 2. SARSA & Q-learning Results

Conclusion

- SARSA has generally performed better than Q-learning
- SARSA may be better than Q-learning in one environment, further research can be done to investigate if all this is true in all environments
- World-Value Functions contributed to enhancing the agents' learning capabilities and the agent to make informed decisions in complex environments.
- DQN's performance showed fluctuating results and this can further be evaluated by exploitation-exploration balance and parameter optimization analysis.

References

[1] Andrew G Barto, Philip S Thomas, and Richard S Sutton.
Some recent applications of reinforcement learning.
In *Proceedings of the Eighteenth Yale Workshop on Adaptive and Learning Systems*, 2017.

[2] J. Clifton and E. Laber.
Q-learning: Theory and applications.
Annual Review of Statistics and Its Application, 7(1):279–301, 2020.

[3] S. Jordan, Y. Chanda, D. Cohen, M. Zhang, and P. Thomas.
Evaluating the Performance of Reinforcement Learning Algorithms.
2:4962–4973, Nov. 2020.

[4] R. Sutton and A. Barto.
Reinforcement learning: An introduction.
In *2018 MIT Press*, volume 2, Nov. 2018.

[5] G. Tasse, S. James, and B. Rosman.
World Value Functions: Knowledge Representation for Multitask Reinforcement Learning.
May. 2022.

Figure 4. Research Paper