Q-Learning in Reinforcement Learning

September 25, 2023

1 Introduction

Q-Learning is a model-free reinforcement learning algorithm used to find the optimal policy for a Markov Decision Process (MDP) [Watkins and Dayan(1992)]. It provides agents with a way to learn the optimal action-selection policy using a Q-function, even when the model of the environment is unknown.

2 Q-Learning Algorithm

The Q-Learning algorithm iteratively updates the Q-values (action values) for each state-action pair in order to converge to the optimal policy. The update rule for Q-Learning is defined as:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left(r + \gamma \max_{a'} Q(s',a')\right)$$

where:

- $Q(s,a)$ is the current estimate of the Q-value for state s and action a.
- α is the learning rate, controlling how much of the new Q-value estimate we adopt.
- r is the immediate reward received after taking action a in state s.
- γ is the discount factor, determining the agent's consideration for future rewards.
- $\max_{a'} Q(s',a')$ is the maximum estimated future reward when at the next state s'.

3 Exploration-Exploitation Dilemma

A significant aspect of Q-Learning is the balance between exploration and exploitation. The agent must explore by taking random actions to discover the reward dynamics of the environment and exploit by choosing actions that maximize the expected reward based on current knowledge. Common strategies to address this include ϵ-greedy, softmax, and upper confidence bounds (UCB) [Sutton and Barto(2018)].

4 Convergence

Under certain conditions, such as if every state-action pair is visited infinitely often and the learning rate α is appropriately reduced over time, Q-Learning is guaranteed to converge to the optimal policy [Jaakkola et al.(1994)Jaakkola, Jordan, and Singh].

5 Application to the Frozen-Lake environment

For this project, we will use the Gym environment *"FrozenLake-v1"*. Everything will be clearly explained in the Python code. The ϵ-greedy policy is a common policy used in Reinforcement Learning for balancing exploration and exploitation and is the one we will use in this project. It is defined as follows:

$$\pi(a|s) = \begin{cases} 1-\epsilon & \text{if } a = \arg\max_{a' \in A(s)} Q(s,a') \\ \epsilon & \text{otherwise} \end{cases},$$

meaning that with probability $1-\epsilon$, the agent selects the action that has the maximum current estimated value, and with probability ϵ, it selects an action at random.

6 Conclusion

Q-Learning is a foundational algorithm in reinforcement learning, providing a method for agents to learn optimal policies in unknown environments. Balancing exploration and exploitation is crucial for its success, and under suitable conditions, the algorithm converges to the optimal solution.

References

[Jaakkola et al.(1994)Jaakkola, Jordan, and Singh] Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural computation*, 6(6):1185–1201, 1994.

[Sutton and Barto(2018)] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press Cambridge, 2018.

[Watkins and Dayan(1992)] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.