Background Study - Reinforcement Learning

1 Introduction

Reinforcement Learning is an interdisciplinary area of machine learning and optimal control where an agent is trained to learn an optimal policy to maximize the return by interacting with the environment where the only feedback the agent gets to evaluate it's actions is reward [**?**].

2 Basic Terms

2.1 Agent

An agent is an entity that makes decisions by taking actions in an environment to achieve a goal. The agent learns from the feedback received from the environment.

2.2 Environment

The environment is everything that the agent interacts with and makes decisions about. It provides the agent with feedback in the form of rewards or punishments based on the actions taken.

2.3 State

A state is a representation of the current situation or configuration of the environment. The agent observes the state to make decisions.

2.4 Action

An action is a decision or move made by the agent that affects the environment. The set of all possible actions is called the action space.

2.5 Policy

A policy is a strategy used by the agent to decide which actions to take given a state. It can be deterministic (always choosing the same action for a state) or stochastic (choosing actions based on probabilities).

2.6 Reward

A reward is feedback from the environment that indicates the value of the agent's action in a given state. The goal of the agent is to maximize the cumulative reward.

2.7 Return

The return is the total accumulated reward that the agent seeks to maximize. It is often calculated as the sum of discounted future rewards.

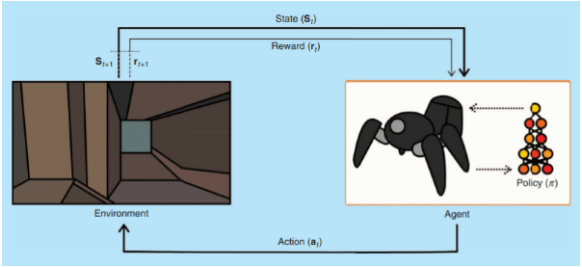

Fig. 1: Agent - Environment Loop

2.8 Episode

An episode is a sequence of states, actions, and rewards that ends in a terminal state. The agent's objective is to maximize the return over each episode.

2.9 Q-Learning

Q-Learning is a model-free reinforcement learning algorithm that aims to learn the optimal action-selection policy by updating Q-values based on the rewards received from actions taken. It does not require a model of the environment.

2.10 Model-Free

Model-free algorithms do not require a model of the environment. They learn the optimal policy directly from interactions with the environment, as opposed to model-based algorithms that use a model to predict future states and rewards.

2.11 Q-Value

A Q-value (or action-value) is a function that estimates the expected return for a specific action in a given state following a particular policy. It helps the agent to choose the best action.

2.12 Deep Q Network (DQN)

A Deep Q Network is a type of neural network used to approximate the Q-values in reinforcement learning. It allows the agent to handle large state spaces by using deep learning techniques.

2.13 Epsilon

Epsilon is a parameter used in the epsilon-greedy exploration strategy. It represents the probability of choosing a random action instead of the action suggested by the policy. This helps the agent to explore the environment and avoid getting stuck in local optima.

2.14 Learning Rate

The learning rate is a parameter that determines the step size in updating the Q-values or the parameters of the policy. It controls how quickly the agent learns from new experiences.

2.15 Discount Factor

The discount factor, denoted by γ, is a parameter that determines the importance of future rewards. It is used to calculate the return and ensures that the sum of future rewards converges.

2.16 Bellman Equation

The Bellman equation provides a recursive decomposition for the value function. It expresses the value of a state as the immediate reward plus the discounted value of the subsequent state, assuming the agent follows the optimal policy.

$$Q(s,a) = r + \gamma \max_{a'} Q(s',a')$$

3 Additional Terms

3.1 Exploration vs Exploitation

Exploration involves trying new actions to discover their effects, while exploitation involves choosing the best-known action to maximize reward. Balancing these two is crucial for effective learning.

3.2 Experience Replay

Experience replay is a technique used in deep reinforcement learning where the agent stores past experiences and replays them to learn from them. This helps to break the correlation between consecutive experiences and improves learning stability.

3.3 Target Network

In DQN, a target network is a copy of the Q-network that is used to calculate the target Q-values. It is updated less frequently to stabilize training.

4 Conclusion

Understanding these basic terms and concepts is essential for studying and applying reinforcement learning techniques. For more detailed information, please refer to standard textbooks and research papers on reinforcement learning.