Eligibility Traces

Reinforcement learning (RL) is a powerful technique used to teach agents to make decisions in complex and dynamic environments. The goal of RL is to find the best policy, which maps states to actions, that maximizes the cumulative reward over time. There are several different methods used to achieve this goal, including Monte Carlo methods, Temporal Difference (TD) learning, and Eligibility Traces.

Monte Carlo methods are a popular technique for RL, in which an agent learns by sampling from its experience, updating its policy based on the rewards it receives. The agent starts with an initial policy and then repeatedly interacts with the environment, updating its policy based on the rewards it receives. This process continues until the agent's policy converges to an optimal solution.

While Monte Carlo methods are powerful, they have several limitations. First, they are computationally expensive, as they require a large number of samples to converge to an optimal solution. Second, Monte Carlo methods are not well suited for problems with non-stationary or delayed rewards, as the agent must wait until the end of an episode to receive a reward.

TD learning is an alternative technique that addresses these limitations by updating the agent's policy based on the expected future rewards, rather than the actual rewards received. This allows for faster convergence, as the agent can learn from its experiences more quickly. However, TD methods are prone to overfitting and can struggle with problems with long-term dependencies.

Eligibility Traces

Eligibility traces is a technique used in RL to improve the efficiency and generality of learning algorithms. It is a method that combines elements of both Monte Carlo methods and TD learning to provide a more powerful and versatile approach to RL.

In eligibility traces, the agent maintains a trace of all the states and actions it has encountered, with the trace decayed over time. When the agent receives a reward, it updates its policy based on the expected future rewards and the trace of its previous experiences. This allows the agent to learn from its experiences more efficiently and to take into account the long-term dependencies of the problem.

The eligibility trace for a state-action pair, denoted as e(s,a), is an exponentially decaying trace that starts with 1 at the time step the state-action pair is visited and then decays over time. The agent updates its policy using the following update rule:

$$\theta <- \theta + \alpha * e(s, a) * \nabla\theta\ Q(s, a)$$

Where α is the learning rate, θ is the parameters of the agent's policy, e(s, a) is the eligibility trace for the state-action pair, and Q(s, a) is the value function of the agent. The agent updates its policy based on the gradient of the value function, with the update weighted by the eligibility trace.

Eligibility traces have several advantages over other RL methods. First, they can handle non-stationary and delayed rewards, as the agent can update its policy based on the expected future rewards and the trace of its previous experiences. Second, they can be used to solve problems with continuous states and actions, as the agent can update its policy based on the gradient of the value function. Third, they have been proven to be more sample efficient than TD learning.

One of the main advantages of using eligibility traces is its ability to handle the problem of non-stationary rewards. In traditional RL methods, the agent updates its policy based on the rewards it receives. However, in some problems, the rewards are non-stationary, meaning they change over time. This makes it difficult for the agent to learn the optimal policy, as the rewards it receives at a particular time step may not be relevant at a later time.

Eligibility traces addresses this problem by maintaining a trace of all the states and actions the agent has encountered, with the trace decayed over time. When the agent receives a reward, it updates its policy based on the expected future rewards and the trace of its previous experiences. This allows the agent to take into account the long-term dependencies of the problem, and to learn from its experiences more efficiently.

For example, if the agent receives a high reward for visiting a certain state-action pair, but the reward changes over time, the agent will still have an idea of the previous high reward it received for visiting that state-action pair because of the trace. This allows the agent to update its policy based on the expected future rewards, taking into account the non-stationary nature of the rewards.

In summary, the ability of eligibility traces to handle non-stationary rewards is that it allows the agent to take into account the long-term dependencies of the problem, and to learn from its experiences more efficiently, even if the rewards change over time.

In conclusion, eligibility traces are a powerful and versatile technique for reinforcement learning. By combining elements of both Monte Carlo and TD methods, they provide a more efficient and general solution to RL problems. Their ability to handle non-stationary and delayed rewards, and to work with continuous states and actions, makes them an attractive option for a wide range of RL applications.