Reinforcement Learning

Model-Free Reinforcement Learning

aka "Playing without the rules"

Temporal Difference Methods

Or how to combine DP and MC

Akka Zemmari

Introduction to Temporal Difference (TD) Methods

Temporal Difference (TD) methods are a class of model-free reinforcement learning algorithms.

TD methods combine ideas from **Monte Carlo methods** and **Dynamic Programming** (DP).

Key characteristics:

- They estimate the value function directly from experience.
- They update estimates based partially on the learned estimate (bootstrapping).

TD methods are central in reinforcement learning due to their efficiency in online learning, making them suitable for situations where the environment's model is unknown or not easily computable.

TD Update Rule

TD learning updates the value of a state based on the difference between the current estimate and a more accurate estimate.

The general update rule for TD is:

$$V(s) \leftarrow V(s) + \alpha \left[r + \gamma V(s') - V(s) \right]$$

where:

- $V(s)$ is the value of state s,
- α is the learning rate,
- r is the immediate reward,
- γ is the discount factor,
- s' is the next state.

The term $r + \gamma V(s') - V(s)$ is called the **TD error**, and represents the difference between the predicted value of s and the target value based on the next state s'.

SARSA

SARSA stands for **State-Action-Reward-State-Action**. It is an on-policy TD method.

SARSA updates the Q-value as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma Q(s', a') - Q(s, a) \right]$$

where:

- $Q(s, a)$ is the action-value function,
- s' is the next state and a' is the next action taken under the current policy.

Key characteristics:

- SARSA is **on-policy**, meaning it updates the policy that is being followed.
- It takes into account the action actually taken by the policy.

Q-Learning

Q-Learning is an off-policy TD method.

It updates the Q-value using the following rule:

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

where $\max_{a'} Q(s',a')$ is the maximum Q-value for the next state over all possible actions.

Key characteristics:

- Q-Learning is **off-policy**, meaning it learns the optimal policy independent of the policy followed by the agent.
- It directly approximates the optimal action-value function $Q^*(s,a)$.

Double Q-Learning

Double Q-Learning is a variant of Q-Learning designed to reduce overestimation bias.

In Double Q-Learning, two Q-value functions are maintained:

$$Q_1(s, a), \ Q_2(s, a)$$

The update rule alternates between updating Q_1 and Q_2:

$$Q_1(s, a) \leftarrow Q_1(s, a) + \alpha \left[r + \gamma Q_2(s', \arg\max_{a'} Q_1(s', a')) - Q_1(s, a) \right]$$

Key characteristics:

- Double Q-Learning reduces the overestimation of action-values seen in traditional Q-Learning.
- It alternates between the two Q-functions to reduce bias.

Conclusion

Temporal Difference (TD) Methods are a powerful set of algorithms in reinforcement learning, combining ideas from Monte Carlo methods and dynamic programming.

Key TD algorithms include:

- **SARSA** (on-policy)
- **Q-Learning** (off-policy)
- **Double Q-Learning** (reduction of overestimation bias)

Each of these methods has its own strengths and trade-offs, making them suitable for different types of problems and environments.