Text

Description automatically generated with medium confidence

A picture containing text

Description automatically generated

A picture containing text

Description automatically generated

It may not be the same as the original scenario since the original one doesn’t have a belief that included all states. TD need more episodes to update every step and bootstrap, but MC can update all states with only one single episode.

A picture containing text

Description automatically generated

In some environments that the agent rewards only when they reach the goal with 1 point and every other step with 0 reward. Take the maze or puzzle as example, MC will perform better than the TD method since MC will update every state along with the trajectory but TD needs to run for a long time and episodes to have a good result.

Text

Description automatically generated

The reason Q-learning is off-policy is because it updates its value by next state and greedy action regardless the behavior policy that used. Thus, we do not need to consider the behavior policy to update action values. So, Q-learning is considered an off-policy control method.

Text

Description automatically generated

If action-selection using the greedy policy, then the SARSA algorithm will be the same as the Q-learning one. Thus, they will make exactly the action selection and weight updates.

Text, letter

Description automatically generated

Text, letter

Description automatically generated

From the graph, we could know that the first episode terminated on state A. All state value is initialized to be 0.5. And the reward is all 0 except for state E for 1.

By taking , we getText, letter

Description automatically generated

So state value of A is changed by -0.05, and all other state remain unchanged.

Text, letter

Description automatically generated

No. From the graph, the results of random walk example are affected by and the results finally converge to an optimal value. TD converges at and MC converges . TD are always had better performance than MC even with different value of , thus, the conclusion about which algorithm is better won't be affected by the range of values of .

Text, letter

Description automatically generated

Different value of like and may cause this. And initialize state value and function of how to approximate may also cause this since TD method always have a bias.

Text, letter

Description automatically generated

In order to compare the effect of different n values on the performance of the TD method. Thus, we need to set the n to a different value for investigating. The original environment use 5 states and it is hard to evaluate the performance.

For a smaller walk problem, the expected value of number of states for travelling is smaller than n. Then the performance is more accurate than a larger n. Therefore, a smaller walk task does affect the results regarding which value of n yields the best performance, the smaller n should be better.

Changing the left-side outcome from 0 to -1 won't make any difference in the best value of n. n only depends on the environment, as long as the state and transition unchanged, the best value of n is unchanged.

Text, letter

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Chart, line chart

Description automatically generated

Chart

Description automatically generated

From the graph, Q-learning > Expected Sarsa > N-step Sarsa > Sarsa > First visit Monte-Carlo.

Text

Description automatically generated

**Chart, line chart

Description automatically generated**

For the eight-action king’s move Q-learning and Sarsa is much better than the original environment and N-step Sarsa remains approximately same level as the original one.

Chart

Description automatically generated with low confidence

The nine’s step will largely reduce the performance of the algorithms. So, it is not recommended to include the stop action.

Text

Description automatically generated

Chart

Description automatically generated

All the algorithms perform worst in the stochastic environment.

Text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

Monte-Carlo won’t affected by the training set since it is not accomplished from state value for bootstrap. But other algorithms will be affected.