Reading – Writing: chapter 7

Brian (Boyuan) Lu, blu1, 1371773

n-step temporal difference is kind of the combination of one-step TD and Monte Carlo method. Which means, instead of obtaining returns for each time step like one-step TD, n-steps TD will obtain a return every n-steps. Therefore, when n-steps are large enough like an episode, it will be very similar to a Monte Carlo method. Therefore, the total reward of n-steps is calculated as truncated return for the time t using rewards up until t+1.

n-step Sarsa is used to perform policy control. Like the method in chapter 6, Sarsa method still uses the greedy epsilon method to choose the state-action pairs. Instead of obtaining the reward from previous step, the accumulated reward from previous n-step is used, and it is calculated from accumulated discounted reward of every time step. And this is a method known as on-policy method. There is also n-step off-policy learning. This is done by important sampling, which refers to a method that take into account of the relative probability of the potential of taking an action. The probability is called importance sampling ratio. The actions within the policy is selected by higher probability. Thus, n-step TD can be performed totally by the off-policy method. One of the off-policy method is called per-reward off-policy. The probability is the relative ratio of an action taken in this policy compare to other policies. And it is changing with time t.

Another approach off-policy n-step method without importance sampling is called tree backup algorithm. It is a tree structure state action model, we usually use the 3-step tree backup where the central spine of the tree holds the sample and reward. The two leaves hang along the sides are the actions that are not selected. Because the unselected actions hold no data, therefore we have updated the estimated value of the node at the top of the diagram toward a target and combines the rewards along the way and all the way to the estimated values of the nodes at the bottom. To perform the estimation, each leaf node has contributions to the actual actions corresponds to their probability of occurring under this policy.

This chapter introduced several ways of implementing temporal-distance learning methods lies between one-step TD and Monte Carlo methods. The major difference is that considering reward of n-step rather than one-step, therefore, it is differ from one-step method also at the same time the step chunk is not as large as one episode like Monte Carlo method. One typical approach is called tree backup diagram which is usually 4 to 3 step update diagrams to perform the estimate based on importance sampling concept. Also, Sarsa can be used to perform n-step methods which only perform estimation on accumulated discount reward of n time steps.

Lastly, n-step methods may be complicated to implement. However, it is more efficient on concept construction which are clear and organized. Importance sampling is the typical prove on this advantage. However, it could cause great variance due to unlikely used actions.