Reading-Writing Assignment: chapter 5

Brian (Boyuan) Lu, 1371773

Unlike previous learning methods, Monte Carlo Methods doesn't assume a complete knowledge of the environment. It's actions, rewards and states are estimated from the actual or simulated interactions with an environment. Monte Carlo method is episodic method since it assumes the experience is divided in to episodes and each episode is value evaluated and has policy changed. MDP is nonstationary because its states are changed for every episode. Therefore, general policy iteration is introduced to compute the optimal policies upon states changes.

In MC prediction, we study the number of visit of certain policies. Two kinds of visit are studied: first-visit and every-visit. First-visit defines the first time a state 's' is visited in an episode. As the number of visits go to infinite, both first-visit MC and every-visit MC converges to an expected v(s). When an model is not in present, action value estimation is a better approach than state value estimation. Without a model, value estimation becomes more important because we need the value from previous action to suggest a new policy. Therefore, primary goal of Monte Carlo method is to evaluate action-value. The every-visit MC which estimate the average of every episode's non-first-visit to under a state 's' with action 'a' and first-visit MC defines the average of every episode's first-visit of a state 's' performing action 'a'. Both MC estimates are expected to converge to each other with infinite number of visits or episodes.

Monte Carlo estimation can be used to perform optimal policies approximation by performing policy evaluation and policy improvement alternatively. Policy evaluation seeks the optimal action taken under the policy. Policy improvement estimates a better policy for certain action to be taken. Therefore, it is a DP problem encounter this problem. MC estimation also use a bias to give a probability that a random action could be selected. In this way, exploring starts. Therefore, epsilon soft policies are used to find the closest greedy for epsilon.

Learning control methods need to learn action value to obtain optimal behavior, however they also need non-optimal condition to explore other possible better actions. On-policy is a proper approach to balance two sides. In this case two policies which one policy is used to learn optimal, another one is used to explore. Learning policy is called target policy, another one is called behavior policy. Off-policy methods require more conceptions and notations which will be slower to converge. However, off-policy is more general and powerful since it contains on-policy to be a special case. Off-policy methods can be used to for prediction problem in case that both target and behavior polices are fixed. Importance sampling is the technique used to evaluate the expected values under a distribution if off-policy methods. It learns by weighting the returns corresponds to the relative probability of their trajectories under the target and behavior policy. The important-sampling ratio is the name of this weighting method.

Off-policy Monte Carlo control is reached by following behavior policy while learning and improving target policy. Which means behavior policy has a small probability to select all actions that might be selected by target policy. At the same time the behavior policy need to be epsilon soft so that the behavior policy can explore all possibilities. Also, target policy need to select greedy actions to accelerate the learning.