Monte Carlo

Aayush Adlakha

September 2, 2023

1 Overview

Monte Carlo methods, unlike DP, require only experience—sample sequences of states, actions, and rewards from actual or simulated interaction with an environment. Monte Carlo methods solve the reinforcement learning problem based on averaging sample returns.

2 First Visit

- Each occurrence of state s in an episode is called a visit to s.
- The first-visit MC method estimates $v_*(s)$ as the average of the returns following first visits to s.
- The multiple occurrences of a particular state means that they are no longer independent, and therefore, First Vist MC is unbiased.

3 Every Visit

- The Every-visit MC method estimates $v_*(s)$ as the average of the returns following all visits to s.
- The multiple occurrences of a particular state mean that they are no longer independent, and therefore, Every-Visit MC is biased.

Both First Visit and Every Visit are based on the **Law of Large Numbers**.

4 On Policy Monte Carlo

- We initialize a random ϵ- soft policy.
- Generate an Episode following the ϵ-soft policy, storing the state, action, and rewards in a list.

- Itearate in reverse order on the maintained lists, add the cumulative reward into the Q-function after averaging appropriately.
- Now assign the best action a probability of $1 - \epsilon + \frac{\epsilon}{A}$, assign $\frac{\epsilon}{A}$ to the rest.
- This way the chance of taking any path is non-zero, but the best path has the highest probability

This process is repeated until we find a stable policy.

5 Off Policy Monte Carlo

- We maintain two policies, one for behavior and one for target.
- We learn from behavior policy and update the policy through **Importance Sampling**.
- One key point is to have a non-zero probability for all actions in the behavior policy to ensure convergence.

Value iteration effectively combines, in each of its sweeps, one sweep of policy evaluation and one sweep of policy improvement.

6 Monte Carlo vs. DP

Monte Carlo learns from a sample experience, whereas DP methods use their previous values to generate new values(**Bootstrapping**). DP methods require complete knowledge of the environment, whereas Monte Carlo methods can work if one is able to simulate the environment without proper knowledge of its functioning.