Sumeet Shanbhag

Student ID : 641020714

Reinforcement Learning Assignment #3

1. **When is it suited to apply Monte-Carlo to a problem?**

   In RL, Monte-Carlo method is primarily used for policy evaluation and value estimation in model-free learning scenarios. When an agent interacts with an environment, it generates episodes and learns from the experiences, even when the environment's model is unknown. Monte Carlo methods allow the agent to estimate the value of states or state-action pairs by averaging the rewards received over multiple episodes, enabling the agent to learn the value functions and subsequently improve its policy.

   In situations with delayed rewards or long-term dependencies, Monte Carlo methods excel by associating the correct value to actions taken in the past, optimizing cumulative rewards over long horizons. For example, in game-playing applications, an agent can use Monte Carlo simulations to explore various game states and actions, learning optimal strategies without knowing the game's rules. This exploration and evaluation of action sequences through Monte Carlo methods, such as Monte Carlo Tree Search (MCTS), are crucial in navigating large or continuous action spaces efficiently, balancing exploration and exploitation, and handling the complexity inherent in many RL problems.

   In essence, Monte Carlo methods provide a practical and efficient approach to learning and decision-making in reinforcement learning, especially under uncertainty and in the absence of a precise environmental model, enabling the development of robust and adaptive agents capable of performing complex tasks.

2. **When does the Monte-Carlo prediction performs the first update?**

   In Monte Carlo prediction within reinforcement learning, the first update to value estimates is made after the conclusion of the first episode. Monte Carlo methods are episodic and require the completion of entire episodes to perform updates, as they rely on experiencing the full sequence of states, actions, and rewards to calculate returns. Once an episode is complete, meaning the agent has reached a terminal state, the returns for each visited state or state-action pair within the episode are calculated, and the corresponding value estimates are updated based on these returns.

3. **What is off-policy learning and why it is useful?**
   Off-policy learning is a strategy in reinforcement learning where an agent learns the value of the optimal policy independently of the policy it follows, allowing the agent to learn from the exploration of the environment. This means the agent learns about the best

possible actions while it might be taking suboptimal actions to explore the environment, making it a crucial technique for balancing exploration and exploitation.

This approach is particularly useful as it allows for the reuse of experiences, enabling more efficient learning and the development of more sophisticated learning strategies. It provides flexibility, safety, and improved stability in learning, allowing agents to understand and adapt to various environments effectively, even in situations with high risks or uncertainties.

4. **Exercise 5.5 of the textbook (page 105)**

First-Visit Monte Carlo Estimator:
Since the episode has 10 steps with a return of 10, the first-visit estimator for the nonterminal state would be 10, as the state is visited for the first time at the beginning of the episode.

Every-Visit Monte Carlo Estimator:
For the every-visit Monte Carlo estimator, the average returns following every visit to the nonterminal state. Since the nonterminal state is visited at every step of the episode until the terminal state is reached, we average the returns from each visit to the nonterminal state within the episode.

To find the every-visit estimator, we would average these returns from each visit:

Every-Visit Estimator
$= (10 + 9 + 8 + 7 + 6 + 5 + 4 + 3 + 2 + 1)/10$
$= 55/10$
$= 5.5$

Conclusion:
- The first-visit Monte Carlo estimator for the value of the nonterminal state is 10.
- The every-visit Monte Carlo estimator for the value of the nonterminal state is 5.5.

5. **Exercise 5.7 of the textbook (page 108)**

The initial rise and subsequent fall in error observed in learning curves for weighted importance-sampling methods are likely due to the elevated variance encountered early in training. In the beginning phases, the agent's experience is minimal, and the discrepancies between the behaviour and target policies can result in high variance and sometimes extreme importance sampling ratios, leading to inflated error in the estimated values. However, as the agent continues to learn and accrue more experiences, the influence of the extreme ratios is mitigated, the estimates of the state values stabilize, and the error begins to diminish.

6. **Exercise 5.8 of the textbook (page 108)**

For the first-visit, we have

$$\mathbb{E}_b\left[\left(\prod_{t=0}^{T-1}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}G_0\right)^2\right]$$

For the next visits, we will have the following difference

$$\mathbb{E}_b\left[\left(\frac{1}{T-1}\sum_{k=1}^{T-1}\prod_{t=0}^{k}\frac{\pi(A_t|S_t)}{b(A_t|S_t)}G_0\right)^2\right]$$

To get the expected square we need only consider each length of episode, multiplying the probability of the episode's occurrence by the square of its importance-sampling ratio, and add these up:

$$\begin{aligned}
&= \frac{1}{2}\cdot 0.1\left(\frac{1}{0.5}\right)^2 && \text{(the length 1 episode)}\\
&\quad + \frac{1}{2}\cdot 0.9\cdot\frac{1}{2}\cdot 0.1\left(\frac{1}{0.5}\frac{1}{0.5}\right)^2 && \text{(the length 2 episode)}\\
&\quad + \frac{1}{2}\cdot 0.9\cdot\frac{1}{2}\cdot 0.9\cdot\frac{1}{2}\cdot 0.1\left(\frac{1}{0.5}\frac{1}{0.5}\frac{1}{0.5}\right)^2 && \text{(the length 3 episode)}\\
&\quad + \cdots
\end{aligned}$$

Which then yields the following:

$$0.1\sum_{k=0}^{\infty}0.9^k\cdot 2^k\cdot 2 = \ 0.2\sum_{k=0}^{\infty}1.8^k = \ \infty.$$

The variance of the estimator would remain infinite, as the expected return is still 1 for every visit to the state. The only difference between first-visit and every-visit MC in this case is that the number of terms increases proportionally to the number of visits to the state and so would continue to run to infinity.

7. Exercise 5.10 of the textbook (page 109)

The equation 5.7 is as shown below

$$V_n \doteq \frac{\sum_{k=1}^{n-1}W_kG_k}{\sum_{k=1}^{n-1}W_k}, \qquad n\geq 2,$$

The equation 5.8 is as shown below

$$V_{n+1} \doteq V_n + \frac{W_n}{C_n}\Big[G_n - V_n\Big], \qquad n \geq 1,$$

and

$$C_{n+1} \doteq C_n + W_{n+1},$$

Setting n = n+1 in equation 5.7, we will get

$$V_{n+1} \doteq \frac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k}$$

Separating the last term and multiplying numerator and denominator by the term:

$$\sum_{k=1}^{n-1} W_k$$

We get,

$$= \frac{W_n G_n + \sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k} \frac{\sum_{k=1}^{n-1} W_k}{\sum_{k=1}^{n} W_k}$$

The denominator terms are interchanged for working out the problem further.
Now, replacing by weighted terms, we get

$$= \Big[\frac{W_n G_n}{C_{n-1}} + V_n\Big] \frac{C_{n-1}}{C_n}$$

Now further simplifying these terms, we get

$$= V_n + \frac{W_n G_n}{C_n} + \frac{V_n C_{n-1}}{C_n} - V_n$$

$$= V_n + \frac{W_n G_n}{C_n} + \frac{V_n C_{n-1} - V_n C_n}{C_n}$$

$$= V_n + \frac{W_n G_n}{C_n} + \frac{-V_n W_n}{C_n}$$

$$= V_n + \frac{W_n}{C_n}\Big[G_n - V_n\Big]$$

8. Exercise 5.11 of the textbook (page 111)

Since policy π is greedy it means that it is deterministic and hence the probability of taking action At in state St is 1.