CS5180 Reinforcement Learning

Exercise 4: Monte Carlo Methods

Anway Shirgaonkar

Q1 Incremental implementation of Monte-Carlo methods

To implement the first visit MC policy evaluation with an incremental approach, there are a few major changes that need to be made.

$Input$: $a\ policy\ \pi\ to\ be\ evaluated$

$Initialize$:
$V(s) \in \mathbb{R}, arbitratily\ \ \forall s \in S$
$N(s) = 0\ \ \forall\ s \in S$

$Loop\ forever(for\ each\ episode)$:
$Generate\ an\ episode\ following\ \pi: S_0, A_0, R_1, S_1, A_1, R_2, \dots, S_{T-1}, A_{T-1}, R_T$
$G \leftarrow 0$
$Loop\ for\ each\ step\ of\ episode, t = T-1, T-2, \dots, 0$:
$G \leftarrow \gamma G + R_{t+1}$
$Unless\ S_t\ appears\ in\ S_0, S_1, \dots, S_{t-1}$:

$$V(S_t) \leftarrow V(S_t) + \frac{1}{N(S_t)}[G - V(S_t)]$$

The array $N(s)$ keeps a track of the number of times a state has been visited.

Q2 First-visit vs. every-visit

(a) Blackjack would NOT be affected by either using the first-visit or every-visit MC implementation because each action that we take in an episode will always lead to a unique state. Hence the implementation of first visit and every-visit MC would give same results.

(b) Let A be the non-terminal state and B be the terminal state.

For first visit MC evaluation:

For the last action we take, we are in state A and land on B with a reward of +1. Hence as per the MC first visit algorithm,

$G \leftarrow \gamma G + R_{t+1}$

$G = 1$

$V(A) = V(A) + 1\big(G - V(A)\big) = 1$

Hence, $V(A) = 1$ for first visit MC evaluation

For every visit MC evaluation:

The value for state A will be calculated as the average return each time state A is visited. Since there are total of 10 transitions involved here, the value will be calculated as:

$$V(A) = \frac{1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10}{10} = 5.5$$