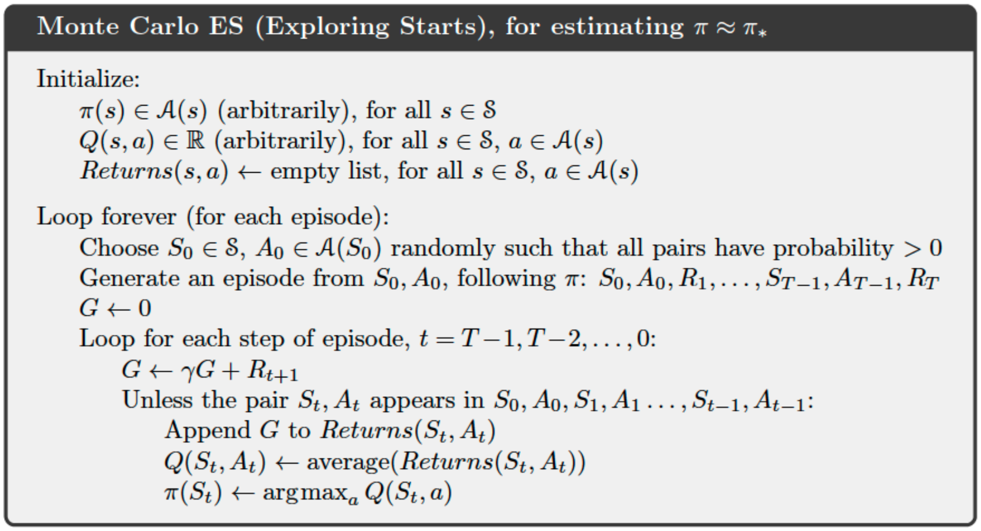
CMPUT 365: Introduction to Reinforcement Learning, Winter 2023
Worksheet #5: Monte-Carlo Methods for Prediction and Control

Manuscript version: *#5875b2* - 2023-03-07 20:36:52-07:00

Question 1. (*Exercise 5.4 S&B*) The pseudocode for *Monte Carlo ES* is inefficient because, for each state-action pair, it maintains a list of all returns and repeatedly calculates their mean. Suggest a modification of the algorithm such that with the same data it computes the same results but it does not use lists for storing returns and such that the update of the action-values takes constant time. The new method should have a memory footprint that doubles the footprint needed just to store the action-values.

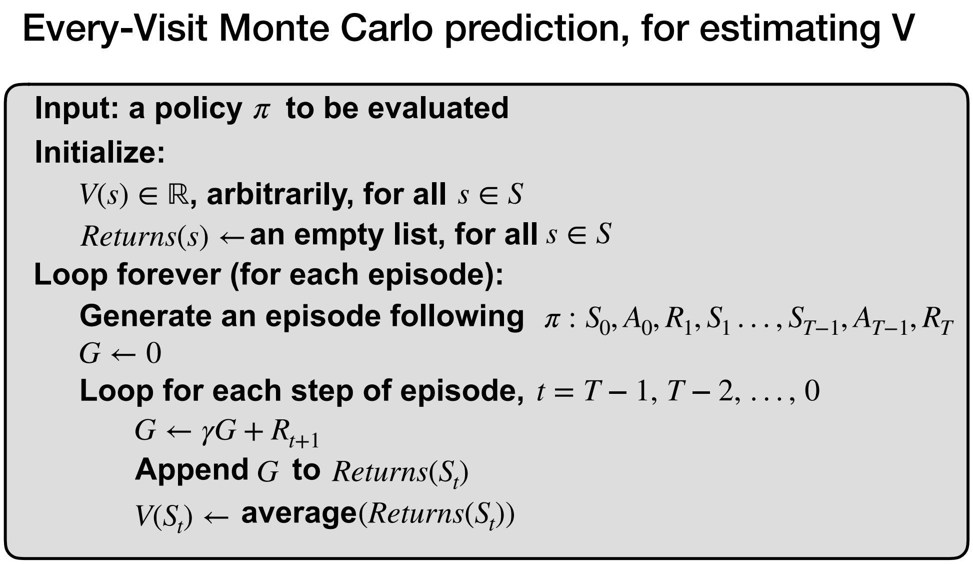


```
Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$

Initialize:
    $\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$
    $Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
    $Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):
    Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability $> 0$
    Generate an episode from $S_0, A_0$, following $\pi$: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
        $G \leftarrow \gamma G + R_{t+1}$
        Unless the pair $S_t, A_t$ appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
            Append $G$ to $Returns(S_t, A_t)$
            $Q(S_t, A_t) \leftarrow$ average$(Returns(S_t, A_t))$
            $\pi(S_t) \leftarrow \text{argmax}_a \, Q(S_t, a)$
```

Question 2. (*Exercise 5.5 S&B*) Consider an MDP with a single nonterminal state s and a single action that transitions back to s with probability p and transitions to the terminal state with probability $1-p$. Let the rewards be $+1$ on all transitions, and let $\gamma = 1$. Suppose you observe one episode that lasts 10 steps, with return of 10. What is the (every-visit) Monte-Carlo estimator of the value of the nonterminal state s?



Every-Visit Monte Carlo prediction, for estimating V

```
Input: a policy $\pi$ to be evaluated
Initialize:
    $V(s) \in \mathbb{R}$, arbitrarily, for all $s \in S$
    $Returns(s) \leftarrow$ an empty list, for all $s \in S$
Loop forever (for each episode):
    Generate an episode following $\pi : S_0, A_0, R_1, S_1 \ldots, S_{T-1}, A_{T-1}, R_T$
    $G \leftarrow 0$
    Loop for each step of episode, $t = T-1, T-2, \ldots, 0$
        $G \leftarrow \gamma G + R_{t+1}$
        Append $G$ to $Returns(S_t)$
        $V(S_t) \leftarrow$ average$(Returns(S_t))$
```

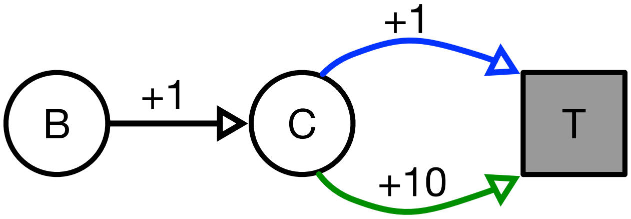
Question 3. Off-policy Monte Carlo prediction allows us to use sample trajectories to estimate the value function for a policy that may be different than the one used to generate the data. Consider the following MDP, with two states B and C, with 1 action in state B and two actions in state C, with $\gamma = 1.0$. In state C both actions transition to the terminating state with $A = 1$ following the blue path to receive a reward $R = 1$ and $A = 2$ following the green path to receive a reward $R = 10$. Assume the target policy π has $\pi(1|C) = 0.9$ and $\pi(2|C) = 0.1$, and that the behaviour policy b has $b(1|C) = 0.25$ and $b(2|C) = 0.75$.

1. What are the true values v_π?

2. Imagine you got to execute π in the environment for one episode, and observed the episode trajectory $S_0 = B, A_0 = 1, R_1 = 1, S_1 = C, A_1 = 1, R_2 = 1$. What is the return for B for this episode? Additionally, what are the value estimates V_π, using this one episode with Monte Carlo updates?

3. You do not actually get to execute π; the agent follows the behaviour policy b. You get one episode when following b and the observed the episode trajectory is $S_0 = B, A_0 = 1, R_1 = 1, S_1 = C, A_1 = 2, R_2 = 10$. What is the return for B for this episode? Notice that this is a return for the behaviour policy, and using it with Monte Carlo updates (without importance sampling ratios) would give you value estimates for b.

4. What is the return for B using this episode if we use importance sampling ratios where the goal is to get value estimates for policy π? Additionally, what is the resulting value estimate for v_π using this return?



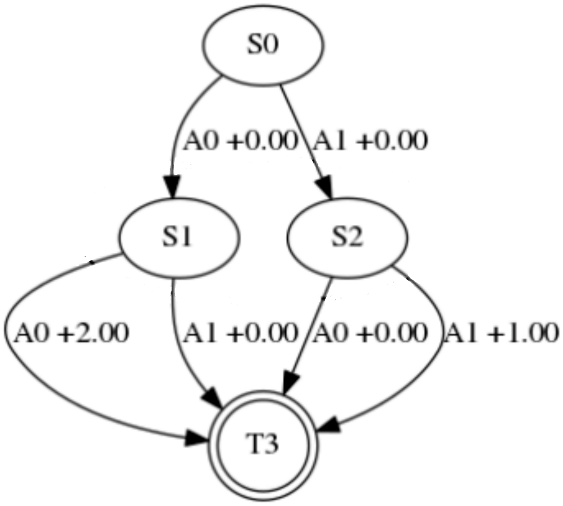
Question 4. Let $\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$. Let $s \in \mathcal{S}$ be such that $\mathbb{P}_b(S_t = s) > 0$, where $\mathbb{P}_b$ is the probability distribution that arises from following b from some arbitrary fixed initial distribution. Let $\mathbb{E}_b$ be the expectation that corresponds to $\mathbb{P}_b$. **Assume that for any $a \in \mathcal{A}$ such that $\pi(a|s) > 0$, it holds that $b(a|s) > 0$.**

1. (*) Verify that $\mathbb{E}_b[\rho_t|S_t = s] = 1$
2. (*) Verify that $\mathbb{E}_b[\rho_t R_{t+1}|S_t = s] = \mathbb{E}_\pi[R_{t+1}|S_t = s]$.
3. (***) What is the variance of the importance corrected one-step reward, $\mathbb{V}_b(\rho_t R_{t+1}|S_t = s)$? When would this variance be large?

Question 5. (*Exercise 5.2 S&B*) Suppose every-visit MC was used instead of first-visit MC on the blackjack task. Would you expect the results to be very different? Why or why not?

Question 6. (*Exercise 5.3 S&B*) What is the backup diagram for Monte Carlo estimation of q_π?

Question 7. Consider the three state MDP below with terminal state T_3 and $\gamma = 1$. Suppose you observe three episodes: $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_2, T_3\}$ with a return of 1. What is the (every-visit) Monte-carlo estimate of the value for each of state S_0, S_1, S_2? How would the Monte-Carlo estimates change if $r(S_0, A_1, S_1) = +1.00$?



Note: In the above question, the notation is a bit weird and deviates from what we have been using so far. In the class notes and the RL book, we use upper-case alphabets for random variables, whereas in this question, we referenced actual states and actions with the upper-case alphabets. Ideally, the question should have used lower-case alphabets to reference the actual states and actions, that is, $\mathcal{S} = \{s_0, s_1, s_2, t_3\}$ and $\mathcal{A} = \{a_1, a_2\}$.

Question 8. Consider the three state MDP from the previous question, but let $r(S_1, A_1, T_3) = +1.00$ and $r(S_2, A_0, T_3) = +2.00$. You observe the following episodes: $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_1, T_3\}$ with a return of 2, $\{S_0, S_1, T_3\}$ with a return of 1, $\{S_0, S_2, T_3\}$ with a return of 1. What is the (every-visit) Monte-Carlo estimate of the value for each of state S_0, S_1, S_2?

Question 9. Suppose you would like to estimate the expected outcome for $3X^2$ where X is the outcome of a fair die. You run an experiment and observe the following sequence of rolls

$$1, 1, 6, 3, 4.$$

How would you use the above data to estimate the expected value $\mathbb{E}[3X^2]$ for a fair die if the above rolls were observed from a loaded die with $\mathbb{P}(X = 1) = 1/2$ and $\mathbb{P}(X = i) = 1/10$ for $i = 2, 3, 4, 5, 6$?

Question 10. Off-policy Monte Carlo prediction allows us to use sample trajectories to estimate the value function for a policy that may be different than the one used to generate the data. Assume the behaviour policy is uniform random: it takes each action with equal probability in each state. The environment has two possible states $\{1, 2\}$ and two possible actions in each state $\{1, 2\}$, with $\gamma = 1.0$. Suppose we collect the following episode, by following the behaviour policy

$$S_0 = 1, A_0 = 2, R_1 = 3, S_1 = 2, A_1 = 1, R_2 = -4\,.$$

We want to evaluate target policy π, which selects action 1 with 0.3 probability in each state. Give an estimate of $v_\pi(1)$ and $v_\pi(2)$, using the episode generated under b using ordinary importance sampling.

Question 11. (****, inspired by Section 5.4)

Call a distribution $p \in \mathcal{M}_1(\mathcal{A})$ over the action space $\mathcal{A}$ to be ϵ-soft, if $p(a) \ge \epsilon/A$ holds for any $a \in \mathcal{A}$. Let $\mathcal{M}_{\epsilon+}(\mathcal{A}) = \{p \in \mathcal{M}_1(\mathcal{A}) : \min_{a\in\mathcal{A}} p(a) \ge \epsilon/A\}$ be the set of ϵ-soft distributions over $\mathcal{A}$. Let $\Pi_{\epsilon+}$ be the set of policies that choose ϵ-soft distributions over the actions only:

$$\Pi_{\epsilon+} = \{\pi \in \Pi \,:\, \pi_t(h_t) \in \mathcal{M}_{\epsilon+}(\mathcal{A}) \text{ for any } h_t \in \mathcal{H}_t \text{ and } t \ge 0\}\,.$$

Let $v^*_{\epsilon+} : \mathcal{S} \to \mathbb{R}$ be defined by

$$v^*_{\epsilon+}(s) = \sup_{\pi\in\Pi_{\epsilon+}} v_\pi(s)\,, \qquad s \in \mathcal{S}\,.$$

In words, $v^*_{\epsilon+}(s)$ is the best value achievable by ϵ-soft policies. Show that

$$v^*_{\epsilon+}(s) = \max_{a\in\mathcal{A}} \left[r'(s,a) + \gamma \sum_{s'\in\mathcal{S}} p'(s'|s,a) v^*_{\epsilon+}(s') \right]\,,$$

where

$$\begin{aligned} p'(s'|s,a) &= (1-\epsilon)p(s'|s,a) + \frac{\epsilon}{A}\sum_{a'\in\mathcal{A}} p(s'|s,a')\,, \qquad \text{and} \\ r'(s,a) &= (1-\epsilon)r(s,a) + \frac{\epsilon}{A}\sum_{a'\in\mathcal{A}} r(s,a')\,. \end{aligned}$$

Hint: Repeat the steps of the proof of the fundamental theorem.
