Worksheet 6

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. How can we modify the algorithm to have incremental updates for each state-action pair?

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 π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  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$(Returns(S_t, A_t))$
      $\pi(S_t) \leftarrow \text{argmax}_a Q(S_t, a)$

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 π 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**$(Returns(S_t))$

Worksheet 6

3. In Policy Iteration, we used dynamic programming for the policy evaluation step, to compute v_π. Monte Carlo ES is a Generalized Policy Iteration (GPI) algorithm, that does not do a full policy evaluation step, before greedifying. How might you modify Monte Carlo ES, to do (more) complete policy evaluation steps before greedifying?

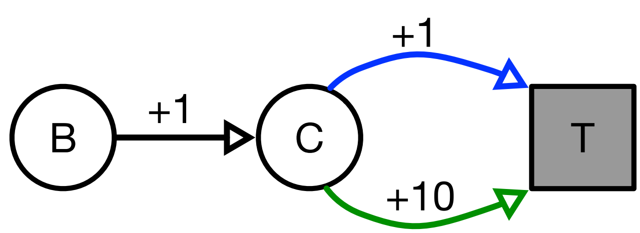
4. Off-policy Monte Carlo prediction allows us to use sample trajectories to estimate the value function for a policy that may be different than the one used to generate the data. Consider the following MDP, with two states B and C, with 1 action in state B and two actions in state C, with $\gamma =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

(a) What are the true values v_π?

(b) 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?

(c) But, you do not actually get to execute π; the agent follows the behaviour policy b. Instead, you get one episode when following b, and observed the episode trajectory $S_0 = B, A_0 = 1, R_1 = 1, S_1 = C, A_1 = 2, R_2 = 10$. 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.

(d) But, we do not actually want to estimate the values for behaviour b, we want to estimates the values for π. So, we need to use importance sampling ratios for this return. What is the return for B using this episode, but now with importance sampling ratios? Additionally, what is the resulting value estimate for V_π using this return?



5. Let $\rho_t = \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$.

(a) Verify that $\mathbb{E}_b[\rho_t|S_t = s] = 1$.

(b) Verify that $\mathbb{E}_b[\rho_t R_{t+1}|S_t = s] = \mathbb{E}_\pi[R_{t+1}|S_t = s]$.

(c) What is the variance of the importance corrected one-step reward, $\mathbb{V}(\rho_t R_{t+1}|S_t = s)$? When would this variance be large?