Wprowadzenie do RL 7

Zadanie 1

Rozważmy dwa epizody:

1) A, 3, A, 2, B, -4, A, 4, B, -3, T
2) B, -2, A, 3, B, -3, T

Wylicz wartości *v(A)* i *v(B)* stosując metody Monte Carlo: *first-visit* i *each-visit*. Przyjmij $\gamma = 1$.

Zadanie 2

Napisz program implementujący poniższy **algorytm iteracji polityki** do środowiska *Frozen Lake* w celu znalezienia optymalnej polityki π_*. Przyjmij, że polityka π jest **deterministyczna**.

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 $Returns(S_t, A_t)$
$Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$