Problem 2: We would like to evaluate the result of the First-Visit Monte Carlo algorithm used for policy evaluation in the episodic, discounted setting. Recall that this algorithm estimates the value function under policy π in any state s, i.e.,

$$V^\pi(s) = \mathbb{E}_{\substack{A_t \sim \pi(.|S_t) \\ S_{t+1} \sim P(.|S_t, A_t)}} \left[\sum_{t=0}^{\infty} \gamma^t R(S_t, A_t) | S_0 = s \right]$$

by the empirical mean of a set of sample returns. Let $\mathcal{T}(s) = \{\tau_1^s, \tau_2^s, \dots, \tau_{N^s}^s\}$ be the sample trajectories starting from s and ending in a terminal state obtained from different episodes of the algorithm. Let $\mathcal{G}(s) = \{G_1^s, G_2^s, \dots, G_{N^s}^s\}$ be the sample returns corresponding to $\mathcal{T}(s)$, i.e.,

$$G_i^s = \sum_{t=0}^{|\tau_i^s|} \gamma^t R(S_{t,i}^s, A_{t,i}^s),$$

where $S_{t,i}^s$ and $A_{t,i}^s$ are the state and action observed at time t in trajectory τ_i^s, respectively. The Monte Carlo algorithm estimates $V^\pi(s)$ as follows:

$$\hat{V}^\pi(s) = \frac{1}{N^s} \sum_{i=1}^{N^s} G_i^s = \frac{1}{N^s} \sum_{\tau_i^s \in \mathcal{T}(s)} \sum_{t=0}^{|\tau_i^s|} \gamma^t R(S_{t,i}^s, A_{t,i}^s).$$

1. Assume that the reward function is bounded, $|R(s,a)| \le R_{max}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Find an upper bound and lower bound on G_i^s, i.e., find α and β such that

$$\alpha \le G_i^s \le \beta.$$

Answer:

Since $|R(s,a)| \le R_{max}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then

$$G_i^s = \sum_{t=0}^{|\tau_i^s|} \gamma^t R(S_{t,i}^s, A_{t,i}^s) \le \sum_{t=0}^{|\tau_i^s|} \gamma^t R_{max} = R_{max} \sum_{t=0}^{|\tau_i^s|} \gamma^t \le R_{max} \sum_{t=0}^{\infty} \gamma^t = \frac{R_{max}}{1-\gamma}$$

Similarly, since $R(s,a) \ge -R_{max}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$, then

$$G_i^s = \sum_{t=0}^{|\tau_i^s|} \gamma^t R(S_{t,i}^s, A_{t,i}^s) \ge \sum_{t=0}^{|\tau_i^s|} \gamma^t (-R_{max}) = -R_{max} \sum_{t=0}^{|\tau_i^s|} \gamma^t \ge -R_{max} \sum_{t=0}^{\infty} \gamma^t = \frac{-R_{max}}{1-\gamma}$$
$$\implies \frac{-R_{max}}{1-\gamma} \le G_i^s \le \frac{R_{max}}{1-\gamma}$$

2. Let $E(s) = \sum_{i=1}^{N^s} G_i^s$ be the sum of all sample returns for state s. Recall that the expected value of each sample return is the true value function, i.e., $\mathbb{E}_{\substack{A_t \sim \pi(.|S_t) \\ S_{t+1} \sim P(.|S_t, A_t)}} [G_i^s] = V^\pi(s)$. Derive and express $\mathbb{E}_{\substack{A_t \sim \pi(.|S_t) \\ S_{t+1} \sim P(.|S_t, A_t)}} [E(s)]$ in terms of $V^\pi(s)$.

Answer:

$$
\begin{aligned}
\mathbb{E}_{\substack{A_t\sim\pi(.|S_t)\\ S_{t+1}\sim P(.|S_t,A_t)}}[E(s)] &= \mathbb{E}_{\substack{A_t\sim\pi(.|S_t)\\ S_{t+1}\sim P(.|S_t,A_t)}}[\sum_{i=1}^{N^s} G_i^s] \\
&= \sum_{i=1}^{N^s} \mathbb{E}_{\substack{A_t\sim\pi(.|S_t)\\ S_{t+1}\sim P(.|S_t,A_t)}}[G_i^s] && \text{due to linearity of expectation} \\
&= \sum_{i=1}^{N^s} V^\pi(s) && \text{by definition} \\
&= V^\pi(s)\sum_{i=1}^{N^s} 1 \\
&= N^s V^\pi(s)
\end{aligned}
$$

3. Apply Hoeffding's inequality (or other concentration inequalities) to bound the probability that $E(s)$ deviates from its expected value obtained in the previous part, i.e., bound $\mathbb{P}\left(|E(s)-\mathbb{E}[E(s)]|\geq\epsilon\right)$ (the subscript for the expectation operator is omitted for simplifying the notation) for any $\epsilon>0$. Notice that the samples G_i^s are independent random variables.

   Answer: For $j\neq i$, G_i^s and G_j^s are independent since they are coming from different episodes. As $E(s)$ is a summation of independent variables G_i^s and $\alpha\leq G_i^s\leq\beta$, then we can directly apply Hoeffding's inequality.

$$
\begin{aligned}
\mathbb{P}\left(|E(s)-\mathbb{E}[E(s)]|\geq\epsilon\right) &\leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{N^s}(\beta-\alpha)^2}\right) \\
&= 2\exp\left(-\frac{2\epsilon^2}{N^s(\frac{2R_{max}}{1-\gamma})^2}\right) \\
&= 2\exp\left(-\frac{(1-\gamma)^2\epsilon^2}{2N^sR_{max}^2}\right)
\end{aligned}
$$

4. Now, considering $\hat{V}^\pi(s)=\frac{1}{N^s}E(s)$, bound the probability that $\hat{V}^\pi(s)$ deviates from $V^\pi(s)$, i.e., bound $\mathbb{P}\left(|\hat{V}^\pi(s)-V^\pi(s)|\geq\epsilon'\right)$ for any $\epsilon'>0$.

   Answer:

$$
\begin{aligned}
\mathbb{P}\left(|\hat{V}^\pi(s)-V^\pi(s)|\geq\epsilon'\right) &= \mathbb{P}\left(\left|\frac{1}{N^s}E(s)-\frac{1}{N^s}\mathbb{E}[E(s)]\right|\geq\epsilon'\right) && \text{by definition of } \hat{V}^\pi(s) \text{ and by part 2} \\
&= \mathbb{P}\left(\frac{1}{N^s}\left|E(s)-\mathbb{E}[E(s)]\right|\geq\epsilon'\right) \\
&= \mathbb{P}\left(|E(s)-\mathbb{E}[E(s)]|\geq N^s\epsilon'\right)
\end{aligned}
$$

   Setting $\epsilon=\epsilon' N^s$ in part 3

$$
\mathbb{P}\left(|\hat{V}^\pi(s)-V^\pi(s)|\geq\epsilon'\right)\leq 2\exp\left(-\frac{(1-\gamma)^2\epsilon'^2N^s}{2R_{max}^2}\right)
$$