Exercise 5.4
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. It would be more efficient to use techniques similar to those explained in Section 2.4 to maintain just the mean and a count (for each state–action pair) and update them incrementally. Describe how the pseudocode would be altered to achieve this.

Answer
We can get an average from all returns just by keeping the number of returns and last average.
Qty(s,a) += 1
Q(s,a) = (Q(s,a)+Rt) * (Qty(s,a) - 1) / Qty(s,a)

Exercise 5.5
Consider an MDP with a single nonterminal state and a single action that transitions back to the nonterminal state with probability p and transitions to the terminal state with probability 1-p. Let the reward be +1 on all transitions, and let γ = 1. Suppose you observe one episode that lasts 10 steps, with a return of 10. What are the first-visit and every-visit estimators of the value of the nonterminal state?

Answer

step	0	1	2	3	4	5	6	7	8	9	10
state	s	s	s	s	s	s	s	s	s	s	T
reward	1	1	1	1	1	1	1	1	1	1	-
action	p	p	p	p	p	p	p	p	p	1-p	-

First-visit
v(s) = 10 / 1 = 10

Every-visit
v(s) = (sum(range(11)) / 10 = 5.5

Exercise 5.6
What is the equation analogous to (5.6) for action values q(s, a) instead of state values v(s), again given returns generated using b?

$$V(s) \doteq \frac{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s)} \rho_{t:T(t)-1}},$$

Answer

$$q(s,a) = \frac{\sum_{t \in \mathcal{T}(s,a)} \rho'_{t:T(t)-1} G_t}{\sum_{t \in \mathcal{T}(s,a)} \rho'_{t:T(t)-1}}$$

$$\rho'_{t:T(t)-1} = \frac{p\left(S_{t+1}|s,a\right) \prod_{k=t+1}^{T-1} \pi\left(A_k|S_k\right) p\left(S_{k+1}|S_k,A_k\right)}{p\left(S_{t+1}|s,a\right) \prod_{k=t+1}^{T-1} b\left(A_k|S_k\right) p\left(S_{k+1}|S_k,A_k\right)}$$

$$\rho'_{t:T(t)-1} = \prod_{k=t+1}^{T-1} \frac{\pi\left(A_k|S_k\right)}{b\left(A_k|S_k\right)}$$