Question 1 // → Comment

Initialize

$\pi(s) \in A(s)$ (arbitrarily), for all $s \in S$

$Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in S$, $a \in A(s)$

~~Returns(s,a) ← empty list~~ for all ~~$s \in S$, $a \in A(s)$~~

① Count(s,a) ∈ W, initially 0 for all $s \in S$, $a \in A(s)$

//maintain count of occurrence of each state-action pair

Loop forever (for each episode):

Choose $S_0 \in S$, $A_0 \in A(S_0)$ randomly such that all pairs have $p > 0$

Generate episode from S_0, A_0 following $\pi : S_0, A_0, R_1, \cdots S_{T-1}, A_{T-1}, R_T$

$G \leftarrow 0$

Loop for each step of episode, $t = T-1, T-2, \cdots, 0$:

$G \leftarrow \gamma G + R_{t+1}$

Unless pair $S_t A_t$ appears in $S_0, A_0, S_1, A_1, \cdots, S_{t-1}, A_{t-1}$:

③ Count(S_t, A_t) += 1 //increase count

② $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{C(S_t, A_t)} [G - Q(S_t, A_t)]$

$C(S_t, A_t) \rightarrow$ Count(S_t, A_t)

~~$C(S_t, A_t) \leftarrow C(S_t, A_t) + 1$~~ //Increase count for this action

$\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

Explanation

Changes are lines marked ①, ② and ③

① was originally a list Returns(S_t, A_t) for all state action pairs

② originally calculated the new average of all returns for S_t, A_t

New Avg = (Old avg × (count − 1) + new value) / (count + 1)

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) \times \frac{1}{\text{Count}(S_t, A_t)} [G - Q(S_t, A_t)]$$