Monte Carlo Control with Exploring Starts

Input:

- Environment with states S, actions A, transition probabilities P, rewards R, and discount factor γ.
- Number of episodes N.

Initialization:

- Initialize action-value function $Q(s, a)$ arbitrarily for all s and a.
- Initialize state-action visit count $N(s, a) = 0$ for all s and a.
- Initialize policy π with random actions for each state-action pair.

Algorithm:

1. **For** each episode $i = 1, 2, \ldots, N$ **do**:
   - Choose a state s_0 and action a_0 arbitrarily, using exploring starts.
   - Generate an episode following policy π: $(s_0, a_0, r_1, s_1, a_1, r_2, \ldots, s_T)$.
   - $G \leftarrow 0$
   - **For** each step $t = T - 1, T - 2, \ldots, 0$ **do**:
     - $G \leftarrow \gamma G + r_{t+1}$
     - **If** (s_t, a_t) is not in the episode history from time step 0 to $t - 1$ **then**:
       - Increment $N(s_t, a_t)$ by 1
       - Update action-value function $Q(s_t, a_t)$ based on G and $N(s_t, a_t)$:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \frac{1}{N(s_t, a_t)}(G - Q(s_t, a_t))$$

       - Update policy π to be greedy with respect to Q:

$$\pi(s_t) \leftarrow \text{argmax}_a Q(s_t, a)$$

Output: Optimized policy π based on the estimated action-value function Q.