Monte Carlo Control with Epsilon-Greedy Policy

Input:

- Environment with states S and actions A
- Number of episodes N
- Discount factor γ
- Exploration parameter ϵ

Initialization:

- Initialize action-value function $Q(s, a)$ arbitrarily for all s and a
- Initialize $N(s, a) = 0$ for all s and a

Algorithm:

```
1. For each episode $i = 1, 2, \ldots, N$ do:
    • Generate an episode using policy derived from $Q$ (e.g., epsilon-greedy)
    • $G \leftarrow 0$
    • For each step $t = T - 1, T - 2, \ldots, 0$ do:
        – $G \leftarrow \gamma G + R_{t+1}$    // Incrementally calculate return
        – If $S_t, A_t$ not in episode history from time step 0 to $t - 1$ then:
            ∗ $N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$
            ∗ $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)}(G - Q(S_t, A_t))$    // Update action-value function
```

Output: Optimal policy π derived from Q