Monte Carlo Control Algorithm

Input:

- Policy π
- Number of episodes N

Initialization:

- Initialize empty arrays $Returns(s, a)$ and $Visits(s, a)$ for each state s and action a
- Initialize action-value function $Q(s, a)$ for each state s and action a with arbitrary values

Algorithm:

1. **For** each episode $i = 1, 2, \ldots, N$ **do**:
   - 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**:
     - $G \leftarrow \gamma G + r_{t+1}$ (where γ is the discount factor)
     - **If** (s_t, a_t) is not in the episode history from time step 0 to $t - 1$ **then**:
       * Append G to $Returns(s_t, a_t)$
       * Increment $Visits(s_t, a_t)$ by 1
       * Update action-value function $Q(s_t, a_t)$ based on $Returns(s_t, a_t)$ and $Visits(s_t, a_t)$:

$$Q(s_t, a_t) = \frac{1}{Visits(s_t, a_t)} \sum_{i=1}^{Visits(s_t, a_t)} Returns(s_t, a_t)[i]$$

Output: Estimated action-value function $Q(s, a)$ and optimal policy π^*