Example: Monte Carlo Method for Grid World

Consider a 3x3 grid where the agent can move left, right, up, or down. The grid has a reward of -1 for each step, and the agent receives a reward of $+10$ for reaching the goal state. The discount factor γ is set to 0.9. We'll use the first-visit Monte Carlo method to estimate the value function.

1. **Initialization**: Start with an empty grid world and initialize the value function for each state with arbitrary values. Let's initialize all values to zero.

0	0	0
0	0	0
0	0	0

2. **Episode 1**: The agent starts at an initial state (e.g., $S1$), takes a sequence of actions, and reaches the goal state (G). The sequence of states visited and the rewards received during this episode are recorded:

$$\text{States visited:} \quad S1 \rightarrow S2 \rightarrow G$$

$$\text{Rewards received:} \quad -1 \rightarrow -1 \rightarrow +10$$

3. **Update the Value Function**: Using the first-visit Monte Carlo method, update the value function for each state visited in the episode based on the observed returns. Since this is the first visit to each state in this episode, we can simply average the returns for each state:

$$V(S1) = \frac{-1}{1} = -1$$

$$V(S2) = \frac{-1+10}{1} = 9$$

$$V(G) = \frac{10}{1} = 10$$

4. **Episode 2**: Repeat the process for another episode, starting from a different initial state if desired. Record the states visited and the rewards received during this episode.

5. **Update the Value Function**: Update the value function based on the returns observed in the second episode.

6. **Repeat**: Continue this process for a predefined number of episodes or until convergence.

This example demonstrates how the Monte Carlo method can be used to estimate the value function for a simple grid world environment. The estimated values of the states improve with more episodes, providing a better approximation of the true values.