Monte Carlo methods can be used to evaluate the performance of a policy in an episodic system by:

1. Generate an episode by running a system with a policy
2. Calculate returns for each state
3. Store calculated returns for each individual state in a table
4. Calculate the average/mean returns estimate the state value function
5. By repeating the following steps above using different policies and generating a significant number of episodes, the actual return for each state will become more accurate and so will the policy itself

There are 2 variants of the Monte Carlo algorithm, first visit and every visit that are both used to estimate a value of for all states .

In first visit MC algorithms (Figure 1.0), for every episode, it stores the return of each state if it the first time that state has been encountered for a particular policy which is shown below in figure 1.1 where is a list of returns for state , and is a counter for the number of times the state has been visited.

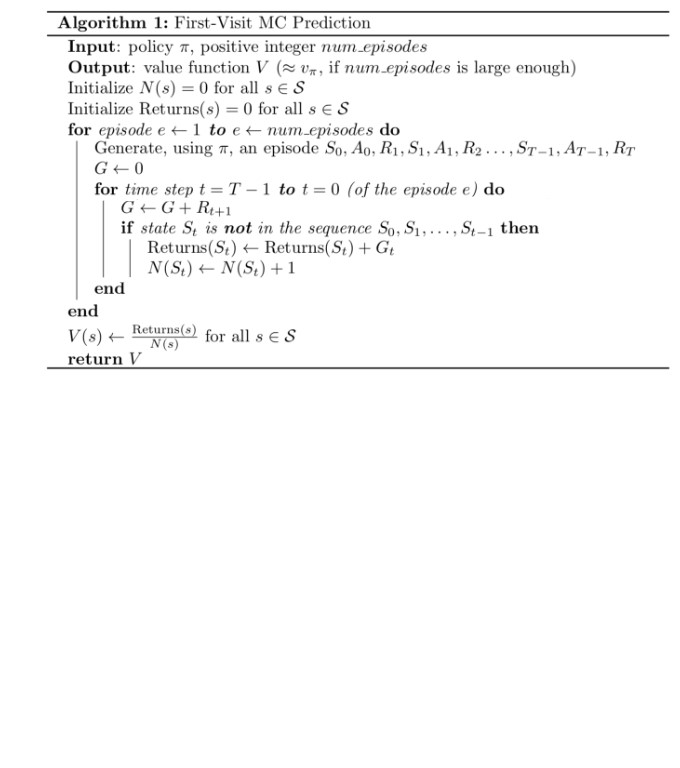


Figure 1.1: Section of algorithm that shows, returns for a particular state is calculated only if it has not been visited before in an episode

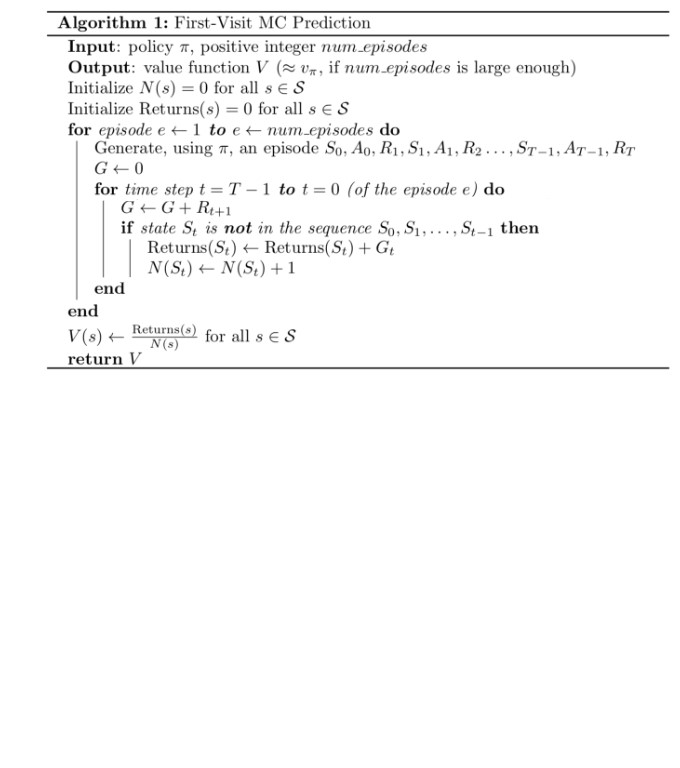


Figure 1.0: First-visit MC algorithm

With every-visit MC algorithms (Figure 1.2), it will update the list of returns for every encounter of within an episode so some states will be updated multiple times.

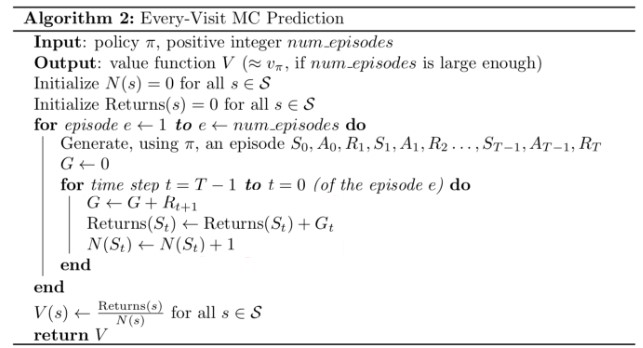


Figure 1.2: Every-visit MC algorithm

Both variants will converge to as the number of visits to tend to infinity. However, for first-visit, each return is an independent and identically distributed estimate of [1].

[1] Richard S. Sutton; Andrew G. Barto, *Reinforcement Learning: An Introduction*. 1998.