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, 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 1.0 where is a list of returns for state , and is a counter for the number of times the state has been visited.

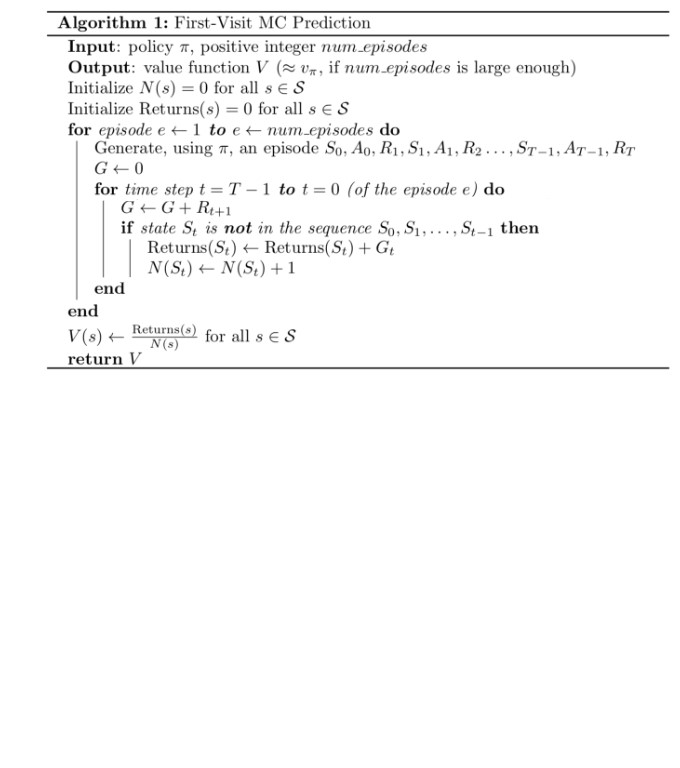


Figure 1.0: Section of algorithm that shows, returns for a particular state is calculated only if it has not been visited before in an episode

Figure 1.1 shows the pseudocode for first-visit MC prediction. Firstly, the number of episodes , policy , state-value function and a list of what will be returns are initialised. A loop is generated for episodes to generate each episode according to a policy. Then a nested loop for each step within the episode is run to sum the estimated return , and while this happens, the algorithm checks if the state it is summing the return from has been visited before. If it hasn’t been visited before, then the calculated return is appended to the list of returns. Finally, the mean of the returns is used to calculate an estimate of .

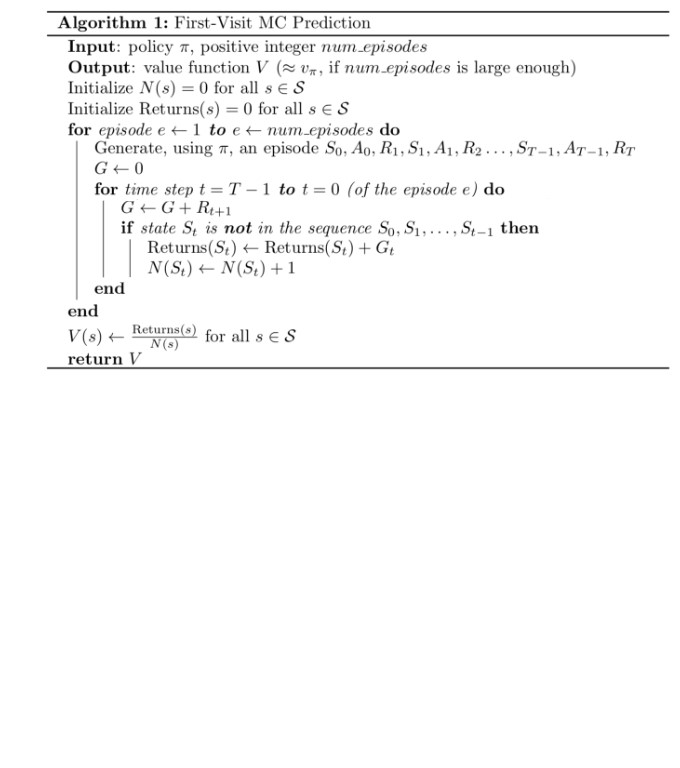


Figure 1.1: First-visit MC algorithm

With every-visit MC algorithms, 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 shows the pseudocode for this. The logic of the code is essentially the same as first-visit however, the condition that checks if the state it is currently on does not exist and instead, every state visited will contribute to the returns and thus the estimate of the value function .

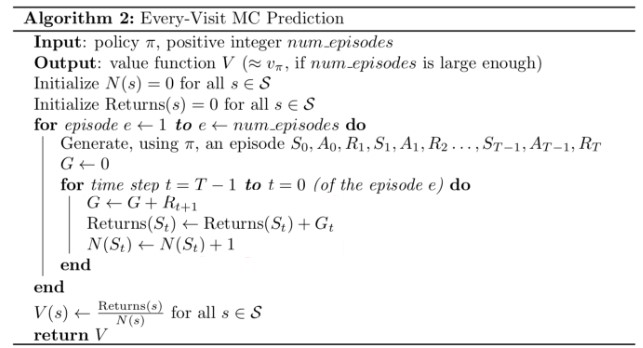


Figure 1.2: Every-visit MC algorithm

Both variants will converge to as the number of visits to tend to infinity. Depending on how the different states were visited during an episode, if the same state never occurs more than once, both variants will produce exactly the same results. For first-visit, each return is an independent and identically distributed (i.i.d) estimate of . Furthermore every average is also an unbiased estimate with standard deviation of error fall as where is the number of returns averaged. First visit over time has been more widely used whereas every-visit MC has been used in function approximation [1].

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