Usually, to obtain an estimation of the policy itself, the state value function would be fed into the policy improvement equation (1) which is then looped for all states and the policy can be extracted.

(1)

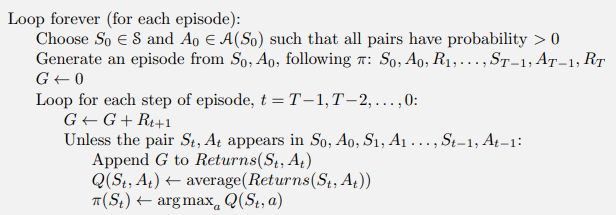
However this equation uses the state transition probability , which is not known in model-free learning since not knowing about the environment means we cannot look ahead for future states.

Therefore, the policy improvement step is instead done by estimating action values, in particular the action value function which returns the expected value of the return (overall reward) for using action in a certain state , and so incorporates future states similar to the transition probability. The policy can then be extracted using equation (2) which shows that for each state the action that maximises the action value function is used as the policy.

(2)

In MC prediction, this step is used to refine the policy as shown in figure 1.0 where at the end of each episode, the new policy is updated with line .

Figure 1.0 Pseudo code that extracts and updates new policy every episode



However, this is not typically used in practice because this is a greedy policy and so can result in an effect called policy freezing where, if all actions are not understood or explored completely, it may continually execute an incorrect action. As a result, the policy will be stuck in a suboptimal result.