# Note on Q functions and V functions in Reinforcement Learning

## Notation and Definitions

**Markov Decision Process** (abbrev *MDP*): a 5-tuple with

* is a set of states (finite or infinite)
* is a set of actions (finite or infinite)
* is the probability to get from state to state with action .
* is the immediate reward after going from state to state with action .
* is the discount factor which determines to what extent the focus is on the most recent rewards. with there is no focus on the most recent rewards only.

**Learning Policy** (or just *Policy*): function which represents mapping from states to probabilities of selecting each possible action.

If the agent is following policy at time , then is the probability that if . Note that is an ordinary function which defines a probability distribution over for each .

We would like to modify the policy with training or experience.

Let us assume that the current state is , and actions are selected according to a stochastic policy then we would like to derive an expression for the expectation of in terms of and .

Recall, the function defines the dynamics of the MDP and is given as:

for all (1)

Then we can write:

(2)

**State-Value Function for Policy** (or simply *Value function*): the value function of a state under a policy , denoted with , is the expected return when starting in s and following thereafter. For MDPs, we can define formally by

for all (3)

where denotes the expected value of a random variable given that the agent follows policy , and is any time step. Note that the value of the terminal state, if any, is always zero.

**Action-Value Function for Policy** (or simply *Value function*):

We define the value of taking action in state under a policy , denoted , as expected return starting from , taking the action , and thereafter following policy :

for all and (4)

Let us express in terms of and . Given a state s, the state value function , given with (3), is equal to the expected cumulative return from that state given a distribution of actions. The action value function is the expectation of the return given state , and taking action a as a starting point, and following policy thereafter. Therefore, given a state the action-value function is the weighted sum of the action-values over all relevant actions weighted by the policy weight:

(5)

Given a state and an action let us express the action-value function in terms of the state value function and the function defining the MDP dynamics . Recall, given a state and an action , the action value function is given by the mathematical expectation of the discounted future rewards i.e. return . The return is the discounted sequence of rewards after the time step and it can be written as:

(6)

It is important to recognize that

. (7)

The first term on the right-hand side of (7) can be expressed as:

. (8)

The expectation in the second term on the right-hand side of (8) can be expressed as:

. (9)

This is the expectation of the return starting at the next time step following the policy given the current state and the action , chosen according to .

Substituting (8) and (9) into (7) gives us:

. (10)

Thus, the action-value function given state s and action following policy is expressed as the sum of the next reward and discounted action-value weighted by probability distribution over the possible next states and next rewards from that action and state .

The value functions and can be estimated from experience. For example, if an agent follows policy and maintains an average, for each state encountered, of the actual returns that have followed that state, then the average will converge to the state value , as the number of times that state is encountered approaches infinity. If separate averages are kept for each action taken in each state, then these averages will similarly converge to the action values, . We call estimation methods of this kind *Monte Carlo methods* because they involve averaging over many random samples of actual returns. Of course, in case there is a large number of states then it would not be feasible to manage separate averages for each state. Instead the agent would have to maintain and as parametrized functions (with fewer parameters than states) and adjust the parameters to better match the observed returns. This approach can produce accurate estimates, although much depends on the nature of the parametrized function approximator.

The value functions satisfy recursive relationships this property of value functions will prove quite useful.

For any policy and for any state s , the following consistency condition holds between the value of s and the value of its successor states. Starting with (6) applied to the definition of :

(11)

Using (5) the last equation becomes:

for all (12)

where it is implicit that the actions, , are taken from the set , that the next states, , are taken from the set , and that the rewards, , are taken from the set . Note that the right-hand side of (12) is interpreted as an expected value obtained as a sum over the values of the triplet , , and . For each triplet the quantity is weighed by its probability, .

Eq. (12) is known as the *Bellman equation* for .