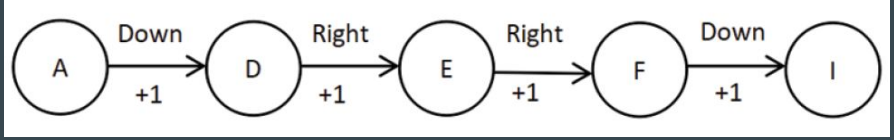
Q Function

A Q function, also called the state-action value function, denotes the value of a state-action pair. The value of a state-action pair is the return the agent would obtain starting from state s and performing action a following policy . The value of a state-action pair or Q function is usually denoted by Q(s,a) and is known as the Q value or state-action value. It is expressed as

$$Q^{\pi}(s,a) = [R(\tau)|s_0 = s, a_0 = a]$$

Note that the only difference between the value function and Q function is that in the value function we compute the value of a state,whereas in the Q function we compute the value of a state-action pair.

Consider the trajectory in generated using policy π



Q function computes the value of a state-action pair. Say we need to compute the Q value of state-action pair A down. That is the Q value of moving down in state A. Then the Q value will be the return of our trajectory starting from state A and performing the action down

$$Q^{\pi}(A,\text{down}) = [R(\tau)|s_0 = A, a_0 = \text{down}]$$

$$Q(A,\text{down}) = 1 + 1 + 1 + 1 = 4$$

Let's suppose we need to compute the Q value of the state-action pair D-right. That is the Q value of moving right in state D. The Q value will be the return of our trajectory starting from state D and performing the action right:

$$Q^{\pi}(\mathrm{D}, \mathrm{right}) = [R(\tau) | s_0 = D, a_0 = \mathrm{right}]$$

$$Q(\mathrm{D}, \mathrm{right}) = 1 + 1 + 1 = 3$$

Similar to what we learned about the value function, instead of taking the return directly as the Q value of a state-action pair, we use the expected return because the return is the random variable and it takes different values with some probability. So, we can redefine our Q function as:

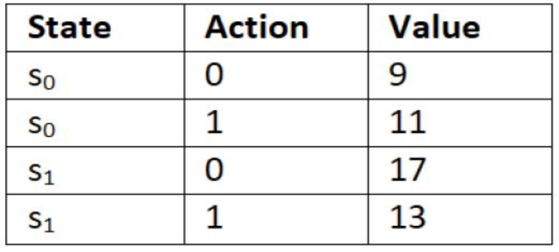
$$Q^{\pi}(s, a) = \underset{\tau \sim \pi}{\mathbb{E}} [R(\tau) | s_0 = s, a_0 = a]$$

Similar to the value function, the Q function depends on the policy, that is, the Q value varies based on the policy we choose. There can be many different Q functions according to different policies. The optimal Q function is the one that has the maximum Q value over other Q functions, and it can be expressed as

$$Q^{*}(s, a) = \max_{\pi} Q^{\pi}(s, a)$$

The optimal policy is the policy π^{*} that gives the maximum Q value.

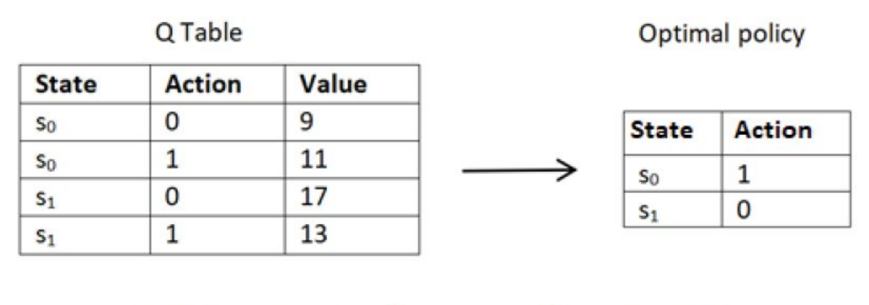
Like the value function, the Q function can be viewed in a table. It is called a Q table. Let's say we have two states s0 and s1, and two actions 0 and 1; then the Q function can be represented as follows



State	Action	Value
s_0	0	9
s_0	1	11
s_1	0	17
s_1	1	13

Q table

As we can observe, the Q table represents the Q values of all possible state-action pairs. We learned that the optimal policy is the policy that gets our agent the maximum return (sum of rewards). We can extract the optimal policy from the Q table by just selecting the action that has the maximum Q value in each state. Thus, our optimal policy will select action 1 in state s0 and action 0 in state s1 since they have a high Q value



Q Table

State	Action	Value
s_0	0	9
s_0	1	11
s_1	0	17
s_1	1	13

Optimal policy

State	Action
s_0	1
s_1	0

Table 1.6: Optimal policy extracted from the Q table

Thus, we can extract the optimal policy by computing the Q function.

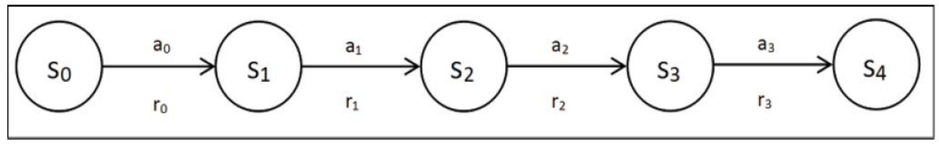
Bellman Concepts

The Bellman equation of the value function:

The Bellman equation states that the value of a state can be obtained as a sum of the immediate reward and the discounted value of the next state (Future Rewards). Say we perform an action a in state s and move to the next state s' and obtain a reward r,then the Bellman equation of the value function can be expressed as,

$$V(s) = R(s, a, s') + \gamma V(s')$$

Let's understand the Bellman equation with an example. Say we generate a trajectory using some policy π.



Let's suppose we need to compute the value of state s2. According to the Bellman equation, the value of state s2 is given as

$$V(s_2) = R(s_2, a_2, s_3) + \gamma V(s_3)$$

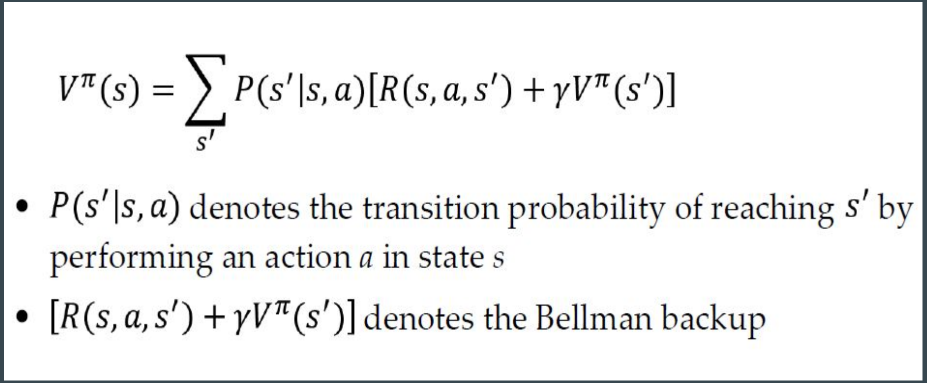
Thus, the Bellman equation of the value function can be expressed as,

$$V^\pi(s) = R(s, a, s') + \gamma V^\pi(s')$$

Where the superscript π implies that we are using policy π .The right-hand side term $R(s, a, s') + \gamma V^\pi(s')$ is often called the **Bellman backup**. The preceding Bellman equation works only when we have a deterministic environment.

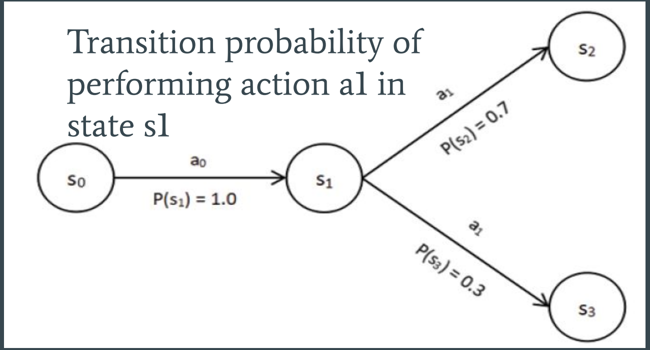
Assume Stochastic Environment:

In this case, we can slightly modify our Bellman equation with the expectations (the weighted average), that is, **a sum of the Bellman backup multiplied by the corresponding transition probability of the next state**.



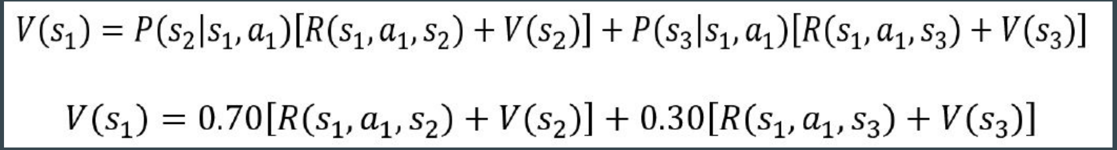
$$V^{\pi}(s) = \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^{\pi}(s')]$$

- $P(s'|s,a)$ denotes the transition probability of reaching s' by performing an action a in state s
- $[R(s,a,s') + \gamma V^{\pi}(s')]$ denotes the Bellman backup



Transition probability of performing action a1 in state s1

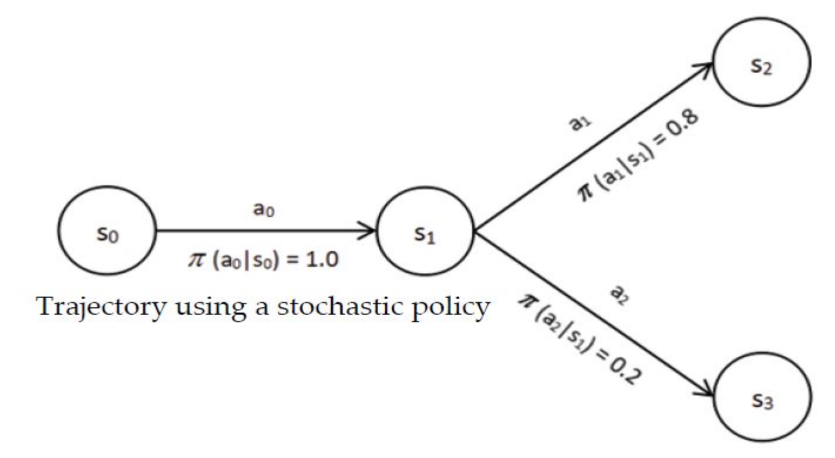
Let's understand this equation better by considering the same trajectory we just used. As we notice, when we perform an action a1 in state s1, we go to s2 with a probability of 0.70 and s3 with a probability of 0.30. Thus, we can write:



$$V(s_1) = P(s_2|s_1,a_1)[R(s_1,a_1,s_2) + V(s_2)] + P(s_3|s_1,a_1)[R(s_1,a_1,s_3) + V(s_3)]$$

$$V(s_1) = 0.70[R(s_1,a_1,s_2) + V(s_2)] + 0.30[R(s_1,a_1,s_3) + V(s_3)]$$

Assume Stochastic Environment with Stochastic Policy:

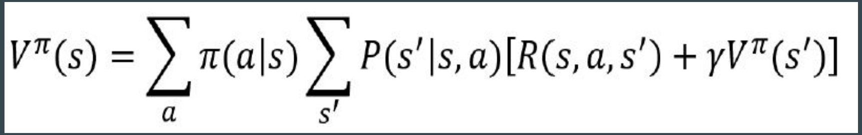


Trajectory using a stochastic policy

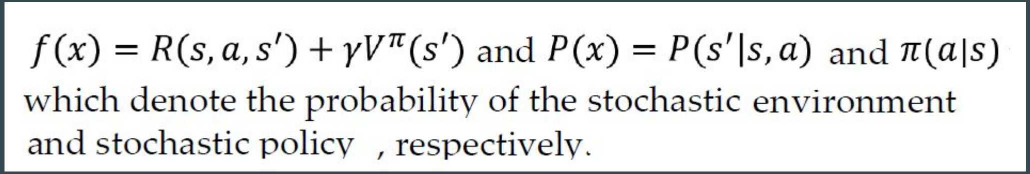
- We learned that to include the stochasticity present in the environment in the Bellman equation, we took the expectation (the weighted average), that is, a sum of the Bellman backup multiplied by the corresponding transition probability of the next state.

- Similarly, to include the stochastic nature of the policy in the Bellman equation, we can use the expectation (the weighted average), that is, a sum of the Bellman backup multiplied by the corresponding probability of action.

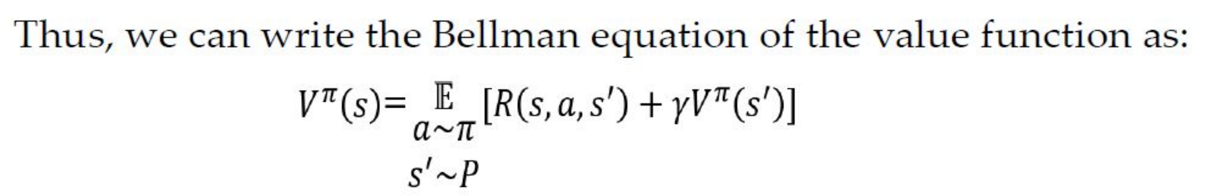
Thus,final Bellman equation of the value function can be written as,



$$V^\pi(s) = \sum_a \pi(a|s) \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^\pi(s')]$$



$f(x) = R(s,a,s') + \gamma V^\pi(s')$ and $P(x) = P(s'|s,a)$ and $\pi(a|s)$ which denote the probability of the stochastic environment and stochastic policy , respectively.



Thus, we can write the Bellman equation of the value function as:

$$V^\pi(s) = \mathop{\mathbb{E}}_{\substack{a\sim\pi \\ s'\sim P}} [R(s,a,s') + \gamma V^\pi(s')]$$

Bellman Equation on the Q Function (State-Value Function):

Similar to the Bellman equation of the value function, the Bellman equation of the Q function states that the Q value of a state-action pair can be obtained as a sum of the immediate reward and the discounted Q value of the next state-action pair.

$$Q^{\pi}(s,a) = R(s,a,s') + \gamma Q^{\pi}(s',a')$$

Where the superscript implies that we are using the policy and the right-hand side term is the Bellman backup.

Suppose we have a stochastic environment, then when we perform an action a in state s. It is not guaranteed that our next state will always be s' ; it could be some other states too with some probability.we can use the expectation (the weighted average), that is, a sum of the Bellman backup multiplied by their corresponding transition probability of the next state, and rewrite our Bellman equation of the Q function as:

$$Q^{\pi}(s,a) = \sum_{s'} P(s'|s,a)\,[R(s,a,s') + \gamma Q^{\pi}(s',a')]$$

Similarly, when we use a stochastic policy, our next state will not always be the same; it will be different states with some probability. So, to include the stochastic nature of the policy, we can rewrite our Bellman equation with the expectation (the weighted average), that is, a sum of Bellman backup multiplied by the corresponding probability of action, just like we did in the Bellman equation of the value function. Thus, the Bellman equation of the Q function is given as:

$$Q^{\pi}(s,a) = \sum_{a} \pi(a|s) \sum_{s'} P(s'|s,a)\,[R(s,a,s') + \gamma Q^{\pi}(s',a')]$$

There is a small change in the previous equation. We don't need to add the term $\sum_{a} \pi(a|s)$ in our equation since we are not selecting any action a based on the policy.we need to select action a' based on the policy π while computing the Q value of the next state-action pair $Q(s',a')$ since a' will not be given.

Thus,Bellman equation of the Q function can be written as,

$$Q^{\pi}(s,a) = \sum_{s'} P(s'|s,a)\left[R(s,a,s') + \gamma \sum_{a'} \pi\,(a'|s')Q^{\pi}(s',a')\right]$$

We can also express the in expectation form as,

$$Q^{\pi}(s,a) = \mathbb{E}_{s'\sim P}[R(s,a,s') + \gamma \mathbb{E}_{a'\sim\pi} Q^{\pi}(s',a')]$$

Bellman Optimal Value Function:

Instead of using some policy to select the action, we compute the value of the state using all possible actions, and then we select the maximum value as the value of the state.It's just the same as the Bellman equation, except here we are taking a maximum over all the possible actions instead of the expectation (weighted average) over the policy since we are only interested in the maximum value.

Let's understand this with an example. Say we are in a state s and we have two possible actions in the state. Let the actions be 0 and 1. Optimal Value Function is given as:

$$V^{*}(s) = \max\begin{pmatrix}\mathbb{E}_{s'\sim P}[R(s,0,s') + \gamma V^{*}(s')] \\ \mathbb{E}_{s'\sim P}[R(s,1,s') + \gamma V^{*}(s')]\end{pmatrix}$$

As we can observe from the above equation, we compute the state value using all possible actions (0 and 1) and then select the maximum value as the value of the state.

Bellman optimal Q function

Instead of using the policy to select action a' in the next state s' we choose all possible actions in that state s' and compute the maximum Q value..

$$Q^*(s,a) = \mathbb{E}_{s' \sim P}\left[R(s,a,s') + \gamma \max_{a'} Q^*(s',a')\right]$$

Let's understand this with an example. Say we are in a state s with an action a. We perform action a in state s and reach the next state .We need to compute the Q value for the next state s' . There can be many actions in state . Let's say we have two actions 0 and 1 in state . Then we can write the optimal Bellman Q function as:

$$Q^*(s,a) = \mathbb{E}_{s' \sim P}[R(s,a,s') + \gamma \max \begin{pmatrix} Q^*(s',0) \\ Q^*(s',1) \end{pmatrix}]$$

Thus, to summarize, the Bellman optimality equations of the value function and Q function are:

$$V^*(s) = \max_a \mathbb{E}_{s' \sim P}[R(s,a,s') + \gamma V^*(s')]$$

$$Q^*(s,a) = \mathbb{E}_{s' \sim P}\left[R(s,a,s') + \gamma \max_{a'} Q^*(s',a')\right]$$

We can also expand the expectation and rewrite the preceding Bellman optimality equations as:

$$V^*(s) = \max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^*(s')]$$

$$Q^*(s,a) = \sum_{s'} P\,(s'|s,a)[R(s,a,s') + \gamma \max_{a'} Q^*(s',a')]$$

Can we derive some relation between the optimal value function and optimal Q function?

We know that the optimal value function has the maximum expected return when we start from a state s and the optimal Q function has the maximum expected return when we start from state s performing some action a. So, we can say that the optimal value function is the maximum of optimal Q value over all possible actions, and it can be expressed as follows:

$$V^*(s) = \max_a Q^*(s,a)$$

We learned that the optimal Bellman Q function is expressed as,

$$Q^*(s,a) = \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma \max_{a'} Q^*(s',a')] \longrightarrow Q^*(s,a) = \sum_{s'} P\,(s'|s,a)[R(s,a,s') + \gamma V^*(s')]$$

$$V^*(s) = \max_a Q^*(s,a) \longrightarrow V^*(s) = \max_a \sum_{s'} P(s'|s,a)[R(s,a,s') + \gamma V^*(s')]$$

we just obtained the optimal Bellman value function. Now that we understand the Bellman equation and the relationship between the value and the Q function.