Definition of Value, Q function and the Bellman equations

Alex Binder @SUTD

August 16, 2016

1 the definition of value

Ever concerned how the definition of a policy value $V^\pi(s)$ (as an expectation of discounted future rewards) is related to $V(s) = R(s) + \gamma \sum'_s P(s'|s)V(s')$?? If not then stop reading here :D

Wall lizards feel lonely, so I attempt a nightly answer.

Definition 1.1. *Value of a policy*
the value of a policy π in state s is the expected discounted reward when starting in state s

$$V^\pi(s) = E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t = \pi(s_t)) \mid s_0 = s]$$

The idea is: at each time step we are in some state s, chose an action a, then get some reward $R(s, a)$ and land in some new state s' with a probability that depends on the old state s and the chosen action. We start at timestep $t = 0$ with state s_0 being s.
This reward is collected over an infinite number of time steps and is thus the so-called infinite horizon value. The discounting by $\gamma^t, 0 < \gamma < 1$ ensured that this sum is finite.

the finite horizon value (for end at K) would be the sum of un-discounted rewards $V^{\pi,u}(s) = E[\sum_{t=0}^{K} R(s_t, a_t = \pi(s_t)) \mid s_0 = s]$.

The takeaway is: the value $V^{\pi}(s)$ in state s is actually the expected reward when starting in state s. That can be a reward for an infinite time horizon (then it is usally discounted with a factor $\gamma \in (0, 1)$), or with a finite time horizon.

2 some more definitions

What is $V(s)$? $V(s)$ is often a notation for the value of a state of the optimal policy, that is a policy π^* which maximizes our reward $E[\sum_t \gamma^t R(s_t, a_t = \pi^*(s_t))]$.

What is the difference between $Q(s, a)$ and $V(s)$?

$Q(s, a)$ is the value (that is expected reward) when we start in state s and perform action a.
$Q(s, a)$ means to take action a in state s, and after that always continue with the given policy (usually the optimal policy). The difference to $V(s)$ is: we do not execute the given/optimal policy $\pi(s)$ in state s, but chose to perform action a instead.
One can think of a modified policy where one is executing action a in the first step and then is always following the policy. In the last section we will deduce an equation for that.

3 How does one arrive from the definition of the value V to the value iteration equations $(V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s, \pi(s)) V^{\pi}(s'))$?

If math causes migraine in you, pls skip this section or read it only while wearing sunglasses at night in an outdoor location.

$$
\begin{aligned}
V^{\pi}(s) &= E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t = \pi(s_t)) \mid s_0 = s] \\
&= E[R(s_0, a_0 = \pi(s_0)) + \sum_{t=1}^{\infty} \gamma^t R(s_t, a_t = \pi(s_t)) \mid s_0 = s] \\
&= E[R(s, a = \pi(s_0)) + \gamma \sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_0 = s] \\
&= R(s, a = \pi(s)) + \gamma E[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_0 = s]
\end{aligned}
$$

the last equation holds because we assume that rewards are deterministic (but this is not crucial, one could use expected reward in state $s_0 = s$) and because the expectation is a linear operation, so $E[c_1 + c_2 X] = c_1 + c_2 E[X]$.

Now we will use a conditional probability equation. Lets start with the definition of the expectation for a random variable Z that lives in a discrete space $Z = Z(w)$, $\{w\}$ is discrete, that is it is a countable finite or countably infinite set. Then its expectation is defined as

$$E[Z] = \sum_{w} P(w) Z(w)$$

Now lets go to conditional probabilities. For a discrete set of states we have

$$E[X] = \sum_{u} P(u) E[X|u]$$

To see this consider a random variable $X = X(u, v)$ depending on a two-dimensional probability space made of pairs (u, v). Then:

$$\begin{aligned}
E[X] &= \sum_{u,v} P(u,v)X(u,v) \text{ by definition of the expectation for discrete states!} \\
&= \sum_{u,v} P(u)P(v|u)X(u,v) \\
&= \sum_{u} P(u) \sum_{v} P(v|u)X(u,v) \\
&= \sum_{u} P(u)E[X|u]
\end{aligned}$$

This holds no matter what dimensionality u and v have. For completeness note that $P(u, v)$ is no probability over the v's, but $P(v|u)$ is!

Now lets apply $E[X] = \sum_u P(u)E[X|u]$ to the case of $E[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_0 = s]$! We will apply it with u being s_1. We will sum over all values s' that s_1 can be.

$$\begin{aligned}
V^{\pi}(s) &= E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t = \pi(s_t)) \mid s_0 = s] \\
&\dots \\
&= R(s, a = \pi(s)) + \gamma E[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_0 = s] \\
&= R(s, a = \pi(s)) + \gamma \sum_{s'} P_{a_0=\pi(s)}(s'|s_0 = s) E[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_0 = s, s_1 = s']
\end{aligned}$$

Note here two things: Firstly, a pure notation thing:

$$P_{a_0=\pi(s)}(s'|s_0 = s) = P_{\pi(s)}(s'|s)$$

- outside of the expectation it is clear that we started in the state $s_0 = s$ with action $a_0 = \pi(s)$.
Secondly, and more importantly

$$E[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_0 = s, s_1 = s']$$

can be computed without knowing s_0 because the sum starts at s_1, and this starting state is given: $s_1 = s'$. All is specified when the starting state is given. looking back further gives no information for computing that sum. Therefore:

$$
\begin{aligned}
V^\pi(s) &= E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t = \pi(s_t)) \mid s_0 = s] \\
&\dots \\
&= R(s, a = \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_0 = s, s_1 = s'] \\
&= R(s, a = \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_1 = s']
\end{aligned}
$$

Allow here pls the notation trolling: $R(s_3) = R(s_3, a_3 = \pi(s_3))$.

$$
\begin{aligned}
V^\pi(s) &= E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t = \pi(s_t)) \mid s_0 = s] \\
&\dots \\
&= R(s, a = \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_1 = s'] \\
&= R(\cdots) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\gamma^0 R(s_1) + \gamma^1 R(s_2) + \gamma^2 R(s_3) + \dots | s_1 = s']
\end{aligned}
$$

Take a closer look!

We arrive at s_2 from s_1 by applying our action $a_1 = \pi(s_1)$ coming from our policy π. The state s_2 might be a random outcome with a certain probability.

BUTTTTTT: The state s_2 does not depend on subscript indices i in s_i. Why is that so? The probability to end up in some new state is in each time step determined by the current state s_t and the chosen action a_t. There is not magic effect from state indices, only their value matters.
That means: if $s_0 = s_1$, then $a_1 = \pi(s_1) = \pi(s_0) = a_0$ and the probability to end up in a state in the next time step are the same for s_0 and s_1.

As a consequence:

$$\begin{aligned} & E[\gamma^0 R(s_1) + \gamma^1 R(s_2) + \gamma^2 R(s_3) + \ldots | s_1 = s'] \\ =& E[\gamma^0 R(s'_0) + \gamma^1 R(s'_1) + \gamma^2 R(s'_2) + \ldots | s'_0 = s'] \end{aligned}$$

That is the same! Therefore replace s_1 by s'_0, s_2 by s'_1, s_3 by s'_2, and so on $\ldots$:

$$\begin{aligned} V^\pi(s) &= E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t = \pi(s_t)) \mid s_0 = s] \\ &\ldots \\ &= R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\sum_{t=1}^{\infty} \gamma^{t-1} R(s_t, a_t = \pi(s_t)) \mid s_{t=1} = s'] \\ &= R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\gamma^0 R(s_1) + \gamma^1 R(s_2) + \gamma^2 R(s_3) + \ldots | s_1 = s'] \\ &= R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\gamma^0 R(s'_0) + \gamma^1 R(s'_1) + \gamma^2 R(s'_2) + \ldots | s'_0 = s'] \\ &= R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\sum_{t=0}^{\infty} \gamma^t R(s'_t) | s'_0 = s'] \\ &= R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\sum_{t=0}^{\infty} \gamma^t R(s'_t, a'_t = \pi(s'_t)) | s'_0 = s'] \end{aligned}$$

In the last line I had reverted my notation trolling: $R(s'_t) = R(s'_t, a'_t = \pi(s'_t))$ by definition

$$E[\sum_{t=0}^{\infty} \gamma^t R(s'_t, a'_t = \pi(s'_t)) | s'_0 = s'] = V^\pi(s')$$

plug this now in to get:

$$
\begin{aligned}
V^{\pi}(s) &= E[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t = \pi(s_t)) \mid s_0 = s] \\
&\dots \\
&= R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) E[\sum_{t=0}^{\infty} \gamma^t R(s'_t, a'_t = \pi(s'_t)) | s'_0 = s'] \\
&= R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) V^{\pi}(s')
\end{aligned}
$$

These are the bellman equations. For finite horizon you need to be a bit more careful because $\gamma = 1$ and it would end then at $K - 1$.

4 ways to solve

4.1 Known policy (can be optimal or not optimal at all)

One way to solve it for a given and known policy π is to use iteration:

$$
V^{\pi,k+1}(s) = R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) V^{\pi,k}(s')
$$

Another way - for a given and known policy is linear algebra by writing V, R a vectors and $P_{\pi}(s'|s)$ as a matrix where the s'-th column is given by $P_{\pi}(s'|\cdot)$

$$
\begin{aligned}
V^{\pi}(s) &= R(s, \pi(s)) + \gamma \sum_{s'} P_{\pi(s)}(s'|s) V^{\pi}(s') \\
\Rightarrow V^{\pi} &= R + \gamma \; P_{\pi(s)} V^{\pi} \\
\Leftrightarrow V^{\pi} - \gamma \; P_{\pi(s)} V^{\pi} &= R \\
\Leftrightarrow I V^{\pi} - \gamma \; P_{\pi(s)} V^{\pi} &= R \\
\Leftrightarrow (I - \gamma \; P_{\pi(s)}) V^{\pi} &= R \\
\Leftrightarrow V^{\pi} &= (I - \gamma \; P_{\pi(s)})^{-1} R
\end{aligned}
$$

That goes often cheap and quick in python/numpy, either as inversion or as solving the linear system $(I - \gamma\ P_{\pi(s)})V^{\pi} = R$ for the vector V.
...

4.2 Unknown optimal policy

Warning: in some cases we do not know the optimal policy for a state s and we want to compute the value for the optimal and yet unknown policy. Then we need to take the best action at state s, that is:

$$V^{\pi*,k+1}(s) = \max_{a \in A} \left(R(s,a) + \gamma \sum_{s'} P_a(s'|s) V^{\pi*,k}(s') \right)$$

This cannot be solved anymore by linear algebra if we do not know the optimal policy ahead!!

5 $Q(s,a)$ and $V(s)$

Recall that $Q(s,a)$ means to take action a in state s, and after that always continue with the given policy (usually the optimal policy).
We know that when we take action a in state s, then we will get reward $R(s,a)$ and end up in the next time step in state s' with probability $P_a(s'|s)$ (this probability depends on the action a. different probabilities give rise to different matrices P_a). So ...

$$Q(s,a) = R(s,a) + \gamma \sum_{s'} P_a(s'|s) V(s')$$

The difference to $V(s)$ is: we do not execute the given/optimal policy $\pi(s)$ in state s, but chose to perform action a instead.
Compare to the last equation above. Then you will see:

$$V^{\pi*,k+1}(s) = \max_{a \in A} Q^{\pi*,k}(s,a)$$
$$Q^{\pi*,k}(s,a) = R(s,a) + \gamma \sum_{s'} P_a(s'|s) V^{\pi*,k}(s')$$

6 Final remarks

Tomas and Leslie remarked correctly: in practice research deals with partially observable MDPs (POMDPs) and more involved stuff (more math, ;'-)). The states in which the agent is, are not (fully) directly observable. Instead one has sensor measurements and one infers probabilities of being in a state or another based on the sensor readings.
For example: imagine a robot with a laser scanner. The state - the robot position and orientation - is a probabilistic function of its sensor measurements.