Bellman Equation for $v_\pi(s)$

Alexander Schiendorfer Pauline Steffel

July 16, 2025

Stuff we want to write $v_\pi(s)$ in terms of:

$$\pi(a \mid s) \stackrel{\text{def}}{=} \mathrm{P}_\pi(A_t = a \mid S_t = s)$$
$$p(s', r \mid s, a) \stackrel{\text{def}}{=} \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$$

Note that G_t is the return that starts at R_{t+1}, i.e., $R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$. Therefore $G_{t+1} = R_{t+2} + \gamma R_{t+3} + \ldots$.
Some useful sets:

- $a \in A$: a is a specific action (e.g., “left” or “right”); A is the set of all actions
- $r \in R \subseteq \mathbb{R}$: r is a specific reward, R is the set of all possible rewards, e.g. $R = \{-15, -5, 0, 5, 15\}$
- $s' \in S$: s' is a specific next state, S is the set of all states

Expected value of a random variable X:

$$\mathbb{E}[X] = \sum_x x \cdot \mathrm{P}(X = x) \tag{1}$$

Conditional expectation of a random variable X given another random variable Y:

$$\mathbb{E}[X \mid Y = y] = \sum_x x \cdot \mathrm{P}(X = x \mid Y = y) \tag{2}$$

Partition theorem or law of total expectation:

$$\mathbb{E}[X] = \sum_y \mathrm{P}(Y = y)\mathbb{E}[X \mid Y = y] \tag{3}$$

Linearity of expectation:

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] \tag{4}$$

Marginal probabilities of joint probabilities:

$$\mathrm{P}(X = x) = \sum_y \mathrm{P}(X = x, Y = y) \tag{5}$$

Definition of $v_\pi(s)$:

$$v_\pi(s) \stackrel{\text{def}}{=} \mathbb{E}_\pi[G_t \mid S_t = s]$$

Pull one reward out of the return (the return is just a random variable that sums rewards with discounting):

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s]$$

Apply Equation (3) to condition the expectation on actions; “pull out” the action A_t as the random variable Y in Equation (3):

$$v_\pi(s) = \sum_a \underbrace{\mathrm{P}_\pi(A_t = a \mid S_t = s)}_{\pi(a|s)} \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a]$$

Split the expectation of a sum into a sum of expectations by using the linearity of expectation from Equation (4) (note that given an action, the expected immediate reward doesn’t depend on the policy):

$$v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)(\mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] + \mathbb{E}_\pi[\gamma G_{t+1} \mid S_t = s, A_t = a])$$

Pull out γ by linearity of expectation:

$$v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)(\mathbb{E}[R_{t+1} \mid S_t = s, A_t = a] + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_t = s, A_t = a])$$

Write just the expected **immediate** reward R_{t+1} in terms of $p(s', r \mid s, a)$: [1]

$$v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)\left(\sum_{s',r} r \cdot \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_t = s, A_t = a]\right)$$

Now we apply Equation (3) to condition the other expectation on the next state S_{t+1}, basically we "pull out" the next state S_{t+1} as the random variable Y in Equation (3):

$$\begin{aligned} v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)\Bigg(&\sum_{s',r} r \cdot \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \\ &+ \gamma \sum_{s'} \mathrm{P}(S_{t+1} = s' \mid S_t = s, A_t = a)\mathbb{E}_\pi[G_{t+1} \mid S_t = s, A_t = a, S_{t+1} = s']\Bigg) \end{aligned}$$

By the Markov property, knowing S_{t+1} makes the expectation independent of S_t and A_t:

$$\begin{aligned} v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)\Bigg(&\sum_{s',r} r \cdot \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \\ &+ \gamma \sum_{s'} \mathrm{P}(S_{t+1} = s' \mid S_t = s, A_t = a)\mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']\Bigg) \end{aligned}$$

By stationarity of the MDP, we know that starting in some state s' will bring the same expected return, regardless of whether this is at time step t or $t+1$, $\mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s'] = \mathbb{E}_\pi[G_t \mid S_t = s']$ and thus $v_\pi(s')$:
Acknowledging that $v_\pi(s') = \mathbb{E}_\pi[G_{t+1} \mid S_{t+1} = s']$, and combining the summations:

$$\begin{aligned} v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)\Bigg(&\sum_{s',r} r \cdot \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \\ &+ \gamma \sum_{s'} \mathrm{P}(S_{t+1} = s' \mid S_t = s, A_t = a)v_\pi(s')\Bigg) \end{aligned}$$

To get both summations into one, we can just write the joint probability of S_{t+1} and R_{t+1} and "demarginalize" according to eq. (5) applied backwards:

$$\begin{aligned} v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)\Bigg(&\sum_{s',r} r \cdot \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \\ &+ \gamma \sum_{s',r} \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)v_\pi(s')\Bigg) \end{aligned}$$

which we can now combine into one summation:

$$\begin{aligned} v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)\Bigg(&\sum_{s',r} r \cdot \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \\ &+ \gamma \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)v_\pi(s')\Bigg) \end{aligned}$$

now pulling out $\mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$:

[1] Honestly, we wouldn't have to average this over all subsequent states here, we could have marginalized (see eq. (5)) over the next state S_{t+1}, to get just the probabilities over R_{t+1}: $v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)\left(\sum_r r \cdot \mathrm{P}(R_{t+1} = r \mid S_t = s, A_t = a) + \gamma \mathbb{E}_\pi[G_{t+1} \mid S_t = s, A_t = a]\right)$ but we did it to make the next steps easier.

$$v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)\left(\sum_{s',r} \mathrm{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a) \cdot (r + \gamma v_\pi(s'))\right)$$

and that, in fact, can just be written as an expected value using the definition of the expected value (Equation (1)), in particular conditioned on the the current state $S_t = s$ and action $A_t = a$ (Equation (2)):

$$v_\pi(s) = \sum_a \mathrm{P}_\pi(A_t = a \mid S_t = s)\left(\mathbb{E}_{S_{t+1},R_{t+1}}\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s, A_t = a\right]\right)$$

and, finally, using the law of total expectation (Equation (3)) over the action A_t as the random variable Y in Equation (3):

$$v_\pi(s) = \mathbb{E}_\pi(R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s)$$