Lecture 3: The Bellman Equation

Melih Kandemir

Semester: Fall 2021

1 Relating the present value to the future

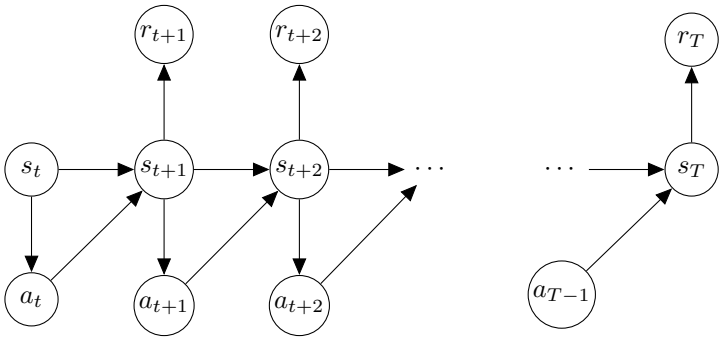

Figure 1: A detailed view of an episode that follows a Markov Decision Process.

An MDP gives a rule to a assign probability to arbitrarily long random variable sequence based on the following assumptions for any time point t:

$$p(s_t, a_t, r_t | s_{t-1}, a_{t-1}) = \underbrace{p(s_t|s_{t-1}, a_{t-1})}_{\substack{\text{Dynamics} \\ \text{model}}} \underbrace{\pi(a_t|s_t)}_{\text{Policy}} \underbrace{p(r_t|s_t)}_{\substack{\text{Reward} \\ \text{model}}}.$$

The joint distribution of an episode of T time steps factorizes as

$$p(s_0, s_1, a_1, r_1, \ldots, s_T, r_T) = p(s_0)\pi(a_0|s_0) \prod_{t=1}^{T} p(s_t|s_{t-1}, a_{t-1})\pi(a_t|s_t)p(r_t|s_t) \tag{1}$$

for some initial state distribution $p(s_0)$ as depicted in the plate diagram in Figure 1.

Define a state-value function as the expected return (cumulative reward) for a state s_t at a time point τ:

Definition 1.1: State Value Function

$$v(s_\tau) = \mathbb{E}_{p(a_\tau, r_{\tau+1}, \ldots, s_T, r_T | s_\tau)} \Big[\underbrace{\sum_{t=\tau+1}^{T} \gamma^{t-\tau-1} r_t}_{G_\tau : \text{Return}} \Big] \tag{2}$$

The following theorem, known as the *Bellman Expectation Equation*, establishes a recursive relationship between state-value functions across consecutive time points. Such a relationship would allow applying divide-and-conquer algorithms for evaluating MDPs for chosen policies. This theorem is a fundamental building block of all the subsequent material. Hence we provide it in great length here.

Theorem 1.1: Bellman Expectation Equation (for state-value)

The following identity holds for any state s_t at any time step τ:

$$v(s_\tau) = \mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)\pi(a_\tau|s_\tau)p(r_{\tau+1}|s_{\tau+1})}[r_{\tau+1}] + \gamma \mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)\pi(a_\tau|s_\tau)}[v(s_{\tau+1})]. \tag{3}$$

Proof. Applying the definition of expectation together with the factorization in Eq. 1, we get

$$\begin{aligned}
v(s_\tau) &= \sum_{a_\tau}\sum_{r_{\tau+1}}\sum_{s_{\tau+1}}\cdots\sum_{r_T} p(a_\tau, r_{\tau+1}, \ldots, s_T, r_T|s_\tau)[r_{\tau+1} + \gamma r_{\tau+2} + \cdots + \gamma^{T-\tau} r_T] \\
&= \sum_{a_{\tau+1}, r_{\tau+2}, \cdots, r_T} \prod_{t=\tau+1}^{T} p(s_t|s_{t-1}, a_{t-1})\pi(a_t|s_t)p(r_t|s_t)[r_{\tau+1} + \gamma r_{\tau+2} + \cdots + \gamma^{T-\tau} r_T]. \\
&= \sum_{a_{\tau+1}, r_{\tau+2}, \cdots, r_T} \prod_{t=\tau+1}^{T} p(s_t|s_{t-1}, a_{t-1})\pi(a_t|s_t)p(r_t|s_t)\Big[\sum_{t=\tau+1}^{T} \gamma^{t-\tau-1} r_t\Big] \\
&= \sum_{a_{\tau+1}, r_{\tau+2}, \cdots, r_T} \prod_{t=\tau+1}^{T} p(s_t|s_{t-1}, a_{t-1})\pi(a_t|s_t)p(r_t|s_t)\Big[r_{\tau+1} + \sum_{t=\tau+2}^{T} \gamma^{t-\tau-1} r_t\Big] \\
&= \sum_{a_{\tau+1}, r_{\tau+2}, \cdots, r_T} \Big[\prod_{t=\tau+1}^{T} p(s_t|s_{t-1}, a_{t-1})\pi(a_t|s_t)p(r_t|s_t) r_{\tau+1} \\
&\qquad + \prod_{t=\tau}^{T} p(s_t|s_{t-1}, a_{t-1})\pi(a_t|s_t)p(r_t|s_t) \sum_{t=\tau+2}^{T} \gamma^{t-\tau-1} r_t\Big] \\
&= \sum_{a_\tau}\sum_{r_{\tau+1}}\sum_{s_{\tau+1}} p(s_{\tau+1}|s_\tau, a_\tau)\pi(a_\tau|s_\tau)p(r_{\tau+1}|s_{\tau+1}) r_{\tau+1} \cancelto{1}{\sum_{a_{\tau+1}, r_{\tau+2}, \cdots, r_T} \prod_{t=\tau+2}^{T} p(s_t|s_{t-1}, a_{t-1})\pi(a_t|s_t)p(r_t|s_t)} \\
&+ \cancelto{1}{\sum_{r_{\tau+1}} p(r_{\tau+1}|s_{\tau+1})} \sum_{a_\tau}\sum_{s_{\tau+1}} p(s_{\tau+1}|s_\tau, a_\tau)\pi(a_\tau|s_\tau) \sum_{a_{\tau+1}, r_{\tau+2}, \cdots, r_T} \prod_{t=\tau+2}^{T} p(s_t|s_{t-1}, a_{t-1})\pi(a_t|s_t)p(r_t|s_t)\gamma^{t-\tau-1} r_t \\
&= \mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)\pi(a_\tau|s_\tau)p(r_{\tau+1}|s_{\tau+1})}[r_{\tau+1}] \\
&\qquad + \sum_{a_\tau}\sum_{s_{\tau+1}} p(s_{\tau+1}|s_\tau, a_\tau)\pi(a_\tau|s_\tau)\mathbb{E}_{\prod_{t=\tau+2}^{T} p(s_t|s_{t-1},a_{t-1})\pi(a_t|s_t)p(r_t|s_t)}\Big[\sum_{t=\tau+2}^{T} \gamma^{t-\tau-1} r_t\Big] \\
&= \mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)\pi(a_\tau|s_\tau)p(r_{\tau+1}|s_{\tau+1})}[r_{\tau+1}] \\
&\qquad + \gamma \sum_{a_\tau}\sum_{s_{\tau+1}} p(s_{\tau+1}|s_\tau, a_\tau)\pi(a_\tau|s_\tau)\mathbb{E}_{\prod_{t=\tau+2}^{T} p(s_t|s_{t-1},a_{t-1})\pi(a_t|s_t)p(r_t|s_t)}\Big[\underbrace{\sum_{t=\tau+2}^{T} \gamma^{t-\tau-2} r_t}_{G_{\tau+1}}\Big] \\
&= \mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)\pi(a_\tau|s_\tau)p(r_{\tau+1}|s_{\tau+1})}[r_{\tau+1}] + \gamma \sum_{a_\tau}\sum_{s_{\tau+1}} p(s_{\tau+1}|s_\tau, a_\tau)\pi(a_\tau|s_\tau)v(s_{\tau+1}) \qquad \blacksquare
\end{aligned}$$

Next, define the action-value function as the expected return (cumulative reward) for being in a state s_t and taking action a_t at a time step τ.

Definition 1.2: Action Value Function

$$q(s_\tau, a_\tau) = \mathbb{E}_{p(r_{\tau+1},\dots,s_T,r_T|s_\tau,a_\tau)}\Big[\sum_{t=\tau+1}^{T} \gamma^{t-\tau-1} r_t\Big]. \tag{4}$$

Comparing the definitions in Eqs. 2 and 4, we see that the state-value function and the action-value function are related as below

$$v(s_\tau) = \sum_{a_\tau} \pi(a_\tau|s_\tau) q(s_\tau, a_\tau). \tag{5}$$

Plugging this identity into the derivation of the Bellman Expectation Equation above, we attain its version for the action-value functions.

Theorem 1.2: Bellman Expectation Equation (for action-value)

The following identity holds for any state s_t at any time step τ:

$$q(s_\tau, a_\tau) = \mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)p(r_{\tau+1}|s_{\tau+1})}[r_{\tau+1}] + \gamma \mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)\pi(a_{\tau+1}|s_{\tau+1})}[q(s_{\tau+1}, a_{\tau+1})]. \tag{6}$$

Proof. Plug the definition of the action value function $v(s) = \sum_a \pi(a|s) q(s,a)$ into the Bellman equation for the state-value functions:

$$v(s_\tau) = \sum_{a_\tau} \pi(a_\tau|s_\tau) \underbrace{\Big[\mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)p(r_{\tau+1}|s_{\tau+1})}[r_{\tau+1}] + \gamma \sum_{s_{\tau+1}} p(s_{\tau+1}|s_\tau, a_\tau) v(s_{\tau+1})\Big]}_{q(s_\tau, a_\tau)}.$$

Plugging the same definition once more to $v(s_{\tau+1})$ gives the desired result

$$\begin{aligned} q(s_\tau, a_\tau) &= \mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)p(r_{\tau+1}|s_{\tau+1})}[r_{\tau+1}] + \gamma \sum_{s_{\tau+1}} p(s_{\tau+1}|s_\tau, a_\tau) \underbrace{v(s_{\tau+1})}_{\sum_{a_{\tau+1}} \pi(a_{\tau+1}|s_{\tau+1}) q(s_{\tau+1}, a_{\tau+1})} \\ &= \mathbb{E}_{p(s_{\tau+1}|s_\tau,a_\tau)p(r_{\tau+1}|s_{\tau+1})}[r_{\tau+1}] + \gamma \sum_{s_{\tau+1}} p(s_{\tau+1}|s_\tau, a_\tau) \sum_{a_{\tau+1}} \pi(a_{\tau+1}|s_{\tau+1}) q(s_{\tau+1}, a_{\tau+1}) \end{aligned}$$ ■

2 Analytical solution of the value function

When both the state space and the action space are discrete, it is possible to repeat the Bellman equation for all possible states and store the outcomes into a vector. Defining $\langle r_\alpha \rangle \triangleq \sum_{r\in\mathcal{R}} p(s', r|s := \alpha)[r]$ and $P_{\alpha,\beta} \triangleq p(s' := \beta | s := \alpha)$ and evaluating these quantities for all possible states, we attain the linear system of equation below:

$$\begin{aligned} \mathbf{v} = \begin{bmatrix} v(s=1) \\ v(s=2) \\ \vdots \\ v(s=n) \end{bmatrix} &= \begin{bmatrix} \langle r_1 \rangle + \gamma \sum_{s'\in\mathcal{S}} P_{1s'} v(s') \\ \langle r_2 \rangle + \gamma \sum_{s'\in\mathcal{S}} P_{2s'} v(s') \\ \vdots \\ \langle r_n \rangle + \gamma \sum_{s'\in\mathcal{S}} P_{ns'} v(s') \end{bmatrix} = \underbrace{\begin{bmatrix} \langle r_1 \rangle \\ \langle r_2 \rangle \\ \vdots \\ \langle r_n \rangle \end{bmatrix}}_{\langle \mathbf{r} \rangle} + \gamma \underbrace{\begin{bmatrix} \sum_{s'\in\mathcal{S}} P_{1s'} v(s') \\ \sum_{s'\in\mathcal{S}} P_{2s'} v(s') \\ \vdots \\ \sum_{s'\in\mathcal{S}} P_{ns'} v(s') \end{bmatrix}}_{\mathbf{P}\mathbf{v}} \\ &= \langle \mathbf{r} \rangle + \gamma \mathbf{P} \mathbf{v}, \end{aligned}$$

where $\mathbf{P}$ is the transition matrix with $P_{ss'} \triangleq p(s'|s)$. We can solve this system simply as

$$\begin{aligned} \mathbf{v} &= \langle \mathbf{r} \rangle + \gamma \mathbf{P} \mathbf{v} \\ \mathbf{v}(\mathbf{I} - \gamma \mathbf{P}) &= \langle \mathbf{r} \rangle \\ \mathbf{v} &= (\mathbf{I} - \gamma \mathbf{P})^{-1} \langle \mathbf{r} \rangle. \end{aligned}$$

Defining $\langle r_s^\pi \rangle \triangleq \sum_{r \in \mathcal{R}} p_\pi(s', r|s)[r]$ and $P_{ss'}^\pi \triangleq p_\pi(s'|s)$, we arrive at the solution for the Bellman equation available for each individual π as

$$\mathbf{v}^\pi = (\mathbf{I} - \gamma \mathbf{P}^\pi)^{-1} \langle \mathbf{r}^\pi \rangle.$$

Although this construction allows us to find an analytical solution to the value function when the environment dynamics $p(s', r|s)$ is given, it has a prohibitive complexity of $O(n^3)$ for n states, which is not feasible for many real-world applications. In the rest of the lectures, we will explore various approximations to make this solution feasible.