Optimal Bellman Equations and Solutions

Navneet Kashyap

1 Optimality in MDPs

Optimality refers to finding a policy π^* that maximizes expected cumulative reward:

$$v_{\pi^*}(s) \geq v_\pi(s) \quad \forall s \in \mathcal{S}, \forall \pi$$

Equivalently for action-values:

$$q_{\pi^*}(s,a) \geq q_\pi(s,a) \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}, \forall \pi$$

The **optimal value functions** are:

$$v_*(s) = \max_\pi v_\pi(s)$$
$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

2 Expectation in Bellman Equations

Bellman equations use expectations to average over:

- Stochastic policies ($\pi(a|s)$)
- State transitions ($p(s'|s,a)$)
- Reward distributions ($r(s,a,s')$)

The common functions used with expectations are π:

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

3 Derivation of Optimal Bellman Equations

3.1 Optimal Value Function (v_*)

For optimal policy π^*, the value satisfies:

$$\begin{aligned} v_*(s) &= \max_{a \in \mathcal{A}} q_{\pi^*}(s, a) \\ &= \max_a \mathbb{E}_{\pi^*} \left[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a \right] \\ &= \max_a \sum_{s', r} p(s', r|s, a) \left[r + \gamma v_*(s') \right] \end{aligned}$$

3.2 Optimal Action-Value Function (q_*)

$$\begin{aligned} q_*(s, a) &= \mathbb{E} \left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \mid S_t = s, A_t = a \right] \\ &= \sum_{s', r} p(s', r|s, a) \left[r + \gamma \max_{a'} q_*(s', a') \right] \end{aligned}$$

4 Solutions to Exercises

Exercise 3.25: v_* in terms of q_*

$$\boxed{v_*(s) = \max_a q_*(s, a)}$$

An interesting side note : This equation is exactly what we use in VI method to extract the optimal policy after calculating the optimal value function.

Exercise 3.26: q_* in terms of v_*

$$\boxed{q_*(s, a) = \sum_{s', r} p(s', r|s, a) \left[r + \gamma v_*(s') \right]}$$

To elaborate, here r is the immediate reward we are receiving due to action a (we are iterating over all possible next states and their corresponding rewards) and we add the discounted expected return from the new state.

Exercise 3.27: π_* in terms of q_*

The optimal policy chooses actions that maximize q_*:

$$\boxed{\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a'} q_*(s,a') \\ 0 & \text{otherwise} \end{cases}}$$

non-zero probability only for actions which maximize the q function

Exercise 3.28: π_* in terms of v_*

$$\boxed{\pi_*(a|s) = \begin{cases} 1 & \text{if } a \in \arg\max_{a'} \sum_{s',r} p(s',r|s,a') \left[r + \gamma v_*(s')\right] \\ 0 & \text{otherwise} \end{cases}}$$

Just replace in 3.27 using 3.26

Exercise 3.29: Bellman optimality for q_* with expectations

Define:

$$r(s,a) = \mathbb{E}[R_{t+1} \mid s,a] = \sum_{s',r} r \cdot p(s',r|s,a)$$

$$p(s'|s,a) = \sum_{r} p(s',r|s,a)$$

Then:

$$\boxed{q_*(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) \max_{a'} q_*(s',a')}$$

Interrelationships
$v_*(s) = \max_a q_*(s,a)$
$q_*(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_*(s')$
$v_*(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_*(s')\right]$
$\pi_* = \arg\max_a q_*(s,a) = \arg\max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) v_*(s')\right]$