Reinforcement Learning Training 2025

Bellman's Equation

- Foundation to solving MDP and RL problems.

Recall (1)

- Markov decision process has transition probabilities

$$\mathbf{P}\Big[S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a\Big]$$

which transitions the agen to state S_{t+1} and a reward of R_{t+1}

- Cumulative reward at time t

$$G_t = R_{t+1} + \gamma R_{t+2} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+1+k}$$

Recall (2)

- A value function is an expected cumulative reward

$$v_\pi(s) = \mathbf{E}_\pi[G_t | S_t = s]$$

- A action-value value function is an expected cumulative reward from taking action a

$$q_\pi(s, a) = E_\pi[G_t | S_t = s, A_t = a]$$

> Note that v and q depend on the policy π.

Bellman Equation

- Allows relationships among v and q.

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma\, v_\pi(s')\right]$$

$$q_\pi(s,a) = \sum_{s',r} p(s', r \mid s, a)\left[r + \gamma \sum_{a'} \pi(a'|s')\, q(s', a')\right]$$

Optimality Condition

- If agent is following the optimal policy π^* (something we want to find), then the value function will also be optimal.

$$v^*(s) = \max_{\pi} v_\pi(s)$$

- It follows that

$$v^*(s) = \max_{a} \left\{ \sum_{s',r} p(s', r|s, a)[r + \gamma v^*(s')] \right\}$$

$$q^*(s, a) = \sum_{s',r} p(s', r|s, a) \cdot \left[r + \gamma \max_{a'} q^*(s', a') \right]$$