@zblu Reinforcement Learning Theory

INF8250AE Course Notes

Bellman Equations and Operators

Bellman Equations, Bellman Operators and Greedy Policy

Key Topics: Bellman Equations, Bellman Operators, Greedy Policy, Banach Fixed Point Theorem, Contraction & Monotonicity
Reference: `https://amfarahmand.github.io/IntroRL/`

Notes on Reinforcement Learning – Sep 15, 2025

Reinforcement Learning 2: Bellman Equation and operator

Zhuobie

2025 年 10 月 6 日

1. Return,value function and Action-value function(Bellman Equation)

Definition. For an infinite-horizon discounted task with $0 \leq \gamma < 1$, the (policy–dependent) return is

$$G_t^\pi \triangleq \sum_{k \geq t} \gamma^{k-t} R_k.$$

By comparing G_t^π and G_{t+1}^π, we have the recursion

$$G_t^\pi = R_t + \gamma G_{t+1}^\pi. \tag{1}$$

Definition. The value function under policy π is $V^\pi(x) \triangleq \mathbb{E}[G_t^\pi \mid X_t = x]$. By taking the conditional expectation of (1) and expanding it one step, we obtain

$$\begin{aligned} V^\pi(x) &= \mathbb{E}\big[R_t + \gamma G_{t+1}^\pi \mid X_t = x\big] && (2)\\ &= r^\pi(x) + \gamma\, \mathbb{E}[V^\pi(X_{t+1}) \mid X_t = x] && (3)\\ &= r^\pi(x) + \gamma \int P(\mathrm{d}x' \mid x, a)\, \pi(\mathrm{d}a \mid x)\, V^\pi(x'). && (4) \end{aligned}$$

The integral in the second term can be understood as follows: starting from state x, we select an action a according to policy π, then transition to the next state x' according to the environment dynamics P. The value at x' is $V^\pi(x')$, and we take a weighted average over all possible a and x'.

Deterministic policy case. If the policy π is deterministic, i.e., $\pi(x) = a$, then

$$V^\pi(x) = r(x, \pi(x)) + \gamma \int P(\mathrm{d}x' \mid x, \pi(x))\, V^\pi(x'). \tag{5}$$

The example of V^π calculation. Let $\mathcal{X} = x_1, x_2$, with immediate rewards $r^\pi(x_1) = +1$, $r^\pi(x_2) = -1$. The transitions are: $\mathbb{P}(x_1 \mid x_1) = 0.9$, $\mathbb{P}(x_2 \mid x_1) = 0.1$, $\mathbb{P}(x_1 \mid x_2) = 0.5$, $\mathbb{P}(x_2 \mid x_2) = 0.5$. Then,

$$V^\pi(x_1) = 1 + \gamma, [, 0.9, V^\pi(x_1) + 0.1, V^\pi(x_2),],\ V^\pi(x_2) \quad = -1 + \gamma, [, 0.5, V^\pi(x_1) + 0.5, V^\pi(x_2),].$$

Matrix form: $V^\pi = r^\pi + \gamma P^\pi V^\pi \Rightarrow (I - \gamma P^\pi) V^\pi = r^\pi$.

Definition. The action-value function is defined as $Q^\pi(x,a) \triangleq \mathbb{E}[G_t^\pi \mid X_t = x, A_t = a]$. We have

$$Q^\pi(x,a) = r(x,a) + \gamma \int P(\mathrm{d}x' \mid x,a)\, V^\pi(x') \tag{6}$$

$$= r(x,a) + \gamma \int P(\mathrm{d}x' \mid x,a)\, \pi(\mathrm{d}a' \mid x')\, Q^\pi(x',a'), \tag{7}$$

$$V^\pi(x) = \int \pi(\mathrm{d}a \mid x)\, Q^\pi(x,a). \tag{8}$$

Remark

Stationarity In discounted infinite-horizon MDPs, the dynamics P, reward r, and policy π are *stationary*. Hence $V^\pi(x)$ and $Q^\pi(x,a)$ do not depend on the time index t. The only one-step "gap"between viewing from t versus $t{+}1$ is the immediate reward, which is absorbed by the Bellman recursion.

2. Bellman optimality equation and greedy policy

Definition. The optimal value function V^* satisfies the optimal Bellman equation

$$V^*(x) = \max_{a \in \mathcal{A}} \Big\{ r(x,a) + \gamma \int P(\mathrm{d}x' \mid x,a)\, V^*(x') \Big\}, \tag{9}$$

$$Q^*(x,a) = r(x,a) + \gamma \int P(\mathrm{d}x' \mid x,a)\, \max_{a' \in \mathcal{A}} Q^*(x',a'). \tag{10}$$

Lemma 2.1 (Optimal value realized by a stationary policy) *In a discounted infinite-horizon MDP ($0 \le \gamma < 1$), there exists a* stationary *(indeed deterministic) policy π^* such that*

$$V^{\pi^*}(x) = V^*(x), \qquad \forall x \in \mathcal{X}.$$

Equivalently, $V^* = V^{\pi^*}$.

Lemma 2.2 (Stationary suffices vs. non-stationary policies) *For discounted infinite-horizon MDPs,*

$$\sup_{\pi \in \Pi_{\text{all}}} V^\pi(x) \;=\; \sup_{\pi \in \Pi_{\text{stationary}}} V^\pi(x) \;=\; V^*(x), \qquad \forall x.$$

That is, allowing time-dependent (non-stationary) policies does not improve the optimal value.

Remark

The conection between the update in Q-learning and Q^*.

Update formula in Q-learning.

$$Q_{t+1}(s_t,a_t) \leftarrow Q_t(s_t,a_t) + \alpha\big[r_t + \gamma \max_{a' \in \mathcal{A}} Q_t(s_{t+1},a') - Q_t(s_t,a_t)\big]. \tag{11}$$

Intuition. This uses *sampling* to approximate one step of the optimal operator T^*: the target term $r_t + \gamma \max_{a'} Q_t(s_{t+1},a')$ is a sample-based estimate of T_t^Q, and the update moves Q_t closer to the fixed point Q^*.

3. Optimal policy from optimal value function: Greedy policy

If we know V^* or Q^*, we can construct an optimal policy π^* as follows:

$$\pi^*(x) = \arg\max_{a\in\mathcal{A}} Q^*(x,a). \pi^*(x) = \arg\max_{a\in\mathcal{A}} \Big\{ r(x,a) + \gamma \int P(\mathrm{d}x' \mid x,a)\, V^*(x') \Big\}.$$

Definition.

$$\pi_g(x;Q) \in \arg\max_{a\in\mathcal{A}} Q(x,a), \tag{12}$$

$$\tag{13}$$

Greedy depends on the reference function. If $Q_1 \neq Q_2$, in general $\pi_g(\cdot;Q_1) \neq \pi_g(\cdot;Q_2)$; similarly, if $V_1 \neq V_2$, then $\pi_g(\cdot;V_1) \neq \pi_g(\cdot;V_2)$.

Consistency at optimality. Let V^* and Q^* be the optimal value functions. Then the greedy policies coincide and are optimal:

$$\pi_g(\cdot;Q^*) \;=\; \pi_g(\cdot;V^*) \;=\; \pi^*.$$

Remark

Non-uniqueness policy π^*. The optimal policy need not be unique (ties in the argmax may occur). Any selector that is greedy w.r.t. V^* or Q^* is optimal, in the sense that

$$V^{\pi_g(V^*)} \;=\; V^* \;=\; V^{\pi^*} \quad \text{and} \quad V^{\pi_g(Q^*)} \;=\; V^* \;=\; V^{\pi^*}$$

4. Bellman operators and fixed points

Motivation. Why do we introduce the Bellman operators?

Recall that in the previous part, we do not know V^π or Q^π in advance. Treat them as unknowns and compute them by fixed-point iteration with the Bellman operators. Start with any bounded initial guess $V^{(0)}$ or $Q^{(0)}$ under a policy π, then update repeatedly:

- Policy evaluation: $V^{(k+1)} = T^\pi V^{(k)}$ (and similarly $Q^{(k+1)} = T^\pi Q^{(k)}$).
- Value iteration: $V^{(k+1)} = T^* V^{(k)}$ (and similarly $Q^{(k+1)} = T^* Q^{(k)}$).

Thanks to the operators'monotonicity and γ-contraction ($\gamma < 1$), these sequences converge to the unique fixed points—namely V^π or Q^π for T^π, and V^* or Q^* for T^*.

Definition. Given a policy π, define the operators

$$(T^\pi V)(x) \triangleq r^\pi(x) + \gamma \int P(\mathrm{d}x' \mid x, a)\, \pi(\mathrm{d}a \mid x)\, V(x'), \tag{14}$$

$$(T^\pi Q)(x, a) \triangleq r(x, a) + \gamma \int P(\mathrm{d}x' \mid x, a)\, \pi(\mathrm{d}a' \mid x')\, Q(x', a'). \tag{15}$$

Observe that—since V and Q are the unknowns we want to compute—the only difference from the Bellman *equations* is that V^π or Q^π are replaced by the variables V or Q;
all other operations are identical. The same idea applies to the *optimal* case: apart from the unknowns V and Q, the computations match the Bellman optimality equations.

Definition. The optimal operators are

$$(T^* V)(x) \triangleq \max_a \left\{ r(x, a) + \gamma \int P(\mathrm{d}x' \mid x, a)\, V(x') \right\}, \tag{16}$$

$$(T^* Q)(x, a) \triangleq r(x, a) + \gamma \int P(\mathrm{d}x' \mid x, a)\, \max_{a'} Q(x', a'). \tag{17}$$

Definition. Fixed point

$$V^\pi = T^\pi V^\pi, \quad Q^\pi = T^\pi Q^\pi; \qquad V^* = T^* V^*, \quad Q^* = T^* Q^*.$$

Hence, the quantities we seek—V^π, Q^π, V^*, and Q^*—are precisely the *fixed points* of their corresponding Bellman operators.

5. Two Key Properties of the Bellman Operators

Lemma 5.1 (Monotonicity) *If $V_1 \le V_2$ (pointwise), then $T^\pi V_1 \le T^\pi V_2$, and $T^* V_1 \le T^* V_2$.*

Lemma 5.2 (Contraction) *With respect to the supremum norm, for $0 \le \gamma < 1$,*

$$\|T^\pi V_1 - T^\pi V_2\|_\infty \;\le\; \gamma\, \|V_1 - V_2\|_\infty,$$

and the same holds for T^*.

Remark

Key consequence. Monotonicity together with contraction implies *existence and uniqueness* of the solution to the Bellman equation. Moreover, starting from any initial guess, the iterations $V_{k+1} = T^\pi V_k$ or $V_{k+1} = T^* V_k$ converge to the unique fixed point.

Toy example. Let $P^\pi = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$, $r^\pi = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$. Then $T^\pi V = r^\pi + \gamma P^\pi V$, so you can compute by substituting elementwise.