Overview

This lecture focuses on establishing theoretical foundations for value-based reinforcement learning by exploring the **Bellman Optimality Equation**, the **existence and uniqueness** of its solutions, and the **convergence guarantees** of classic methods like **value iteration** and **policy iteration**.

We examine key concepts such as fixed points, contraction mappings, and how these relate to the optimal action-value function q^*. The lecture provides both intuition and formal mathematical reasoning for convergence and optimality guarantees.

Bellman Optimality Equation (Stochastic Rewards)

Assume a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, with:

- $\mathcal{S}$: set of states
- $\mathcal{A}$: set of actions
- $P(s'|s,a)$: transition probability
- $R(s,a)$: **stochastic** reward function
- $\gamma \in [0,1)$: discount factor

The **Bellman Optimality Equation** for the action-value function $q^*(s,a)$ is:

$$q^*(s,a) = \mathbb{E}_{s'}\left[R(s,a) + \gamma \max_{a'} q^*(s',a')\right]$$

This equation defines a fixed point relationship over q^*. Our goal is to study the **existence**, **uniqueness**, and **convergence** properties of solutions to this equation.

Questions to Address

1. **Existence**: Does a function q^* exist that satisfies the Bellman equation?
2. **Uniqueness**: Is the solution q^* unique?
3. **Algorithmic Convergence**: Can **value iteration** compute this q^*?

Fixed Point Theory

Definition: Let T be an operator on a function space. A function x is a **fixed point** of T if:

$$T(x) = x$$

In the context of Bellman's equation, q^* is a fixed point of the **Bellman operator** T, defined as:

$$(Tq)(s,a) := \mathbb{E}_{s'}\left[R(s,a) + \gamma \max_{a'} q(s',a')\right]$$

Existence of the Fixed Point

We want to show that T has a fixed point in the space of bounded functions $Q := \{q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$.

Construction of a Sequence

- Start from an arbitrary initial $q_0 \in Q$
- Define the iterative process:

$$q_{n+1} = Tq_n$$

- Inductively, $q_n = T^n q_0$

Intuition: Due to the bounded nature of rewards and the contraction properties of T, this sequence forms a **Cauchy sequence** in a suitable norm, hence it converges.

Normed Space and Convergence

We equip the function space with the supremum norm:

$$\|q\|_\infty := \sup_{s,a} |q(s,a)|$$

Then, we show that the Bellman operator T is a **contraction mapping** under this norm:

$$\|Tq - Tq'\|_\infty \leq \gamma \|q - q'\|_\infty$$

This implies that the sequence $\{q_n\}$ converges to a unique fixed point q^*, by the **Banach fixed-point theorem**.

Uniqueness of the Fixed Point

Assume by contradiction that two fixed points q^* and q'^* exist such that:

$$Tq^* = q^* \quad \text{and} \quad Tq'^* = q'^*$$

Then,

$$\|q^* - q'^*\|_\infty = \|Tq^* - Tq'^*\|_\infty \leq \gamma \|q^* - q'^*\|_\infty$$

This implies:

$$(1-\gamma)\|q^* - q'^*\|_\infty \leq 0 \Rightarrow \|q^* - q'^*\|_\infty = 0 \Rightarrow q^* = q'^*$$

Hence, the fixed point is unique.

Bellman Operator as a Contraction

Let us analyze T explicitly:

$$(Tq)(s,a) = \mathbb{E}_{s'}\left[R(s,a) + \gamma \max_{a'} q(s',a')\right]$$

Then for any two functions q_1, q_2:

$$\begin{aligned}\|Tq_1 - Tq_2\|_\infty &= \sup_{s,a} \left|\mathbb{E}_{s'}\left[\gamma \max_{a'} q_1(s',a') - \gamma \max_{a'} q_2(s',a')\right]\right| \\ &\leq \gamma \sup_{s',a'} |q_1(s',a') - q_2(s',a')| = \gamma \|q_1 - q_2\|_\infty\end{aligned}$$

This shows that T is a γ-contraction.

Value Iteration Converges

Given that:

- T is a γ-contraction
- The sequence $q_n = T^n q_0$ lies in a complete metric space

Then by the **Banach fixed-point theorem**, $q_n \to q^*$ as $n \to \infty$.

Hence, **value iteration** converges to the optimal q^* regardless of the initial guess q_0.

Policy Improvement Intuition

Suppose π is a policy, and we evaluate $q_\pi(s, a)$. If we now define:

$$\pi'(s) = \arg\max_a q_\pi(s, a)$$

then π' is a **greedy policy** w.r.t. q_π.

From policy improvement theorem:

$$v_{\pi'}(s) \geq v_\pi(s) \quad \forall s$$

That is, the greedy policy improves or retains the value of the previous policy. Thus, policy improvement always moves toward the optimal policy.

Policy Iteration Converges

Policy iteration alternates between:

1. **Policy Evaluation**: Compute q_π
2. **Policy Improvement**: Greedify q_π to obtain π'

Because the number of deterministic policies is finite ($|\mathcal{A}|^{|\mathcal{S}|}$), and each improvement strictly increases v_π for at least one state (unless already optimal), policy iteration is **guaranteed to converge** to π^* in finite steps.