Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 15

Solution By: Arshia Gharooni

Q1

Regarding the theoretical guarantees for convergence in value-based deep reinforcement learning, which of the following statements are correct? (Select all that apply.)

(A) Because the Bellman operator T is a contraction mapping, value iteration converges to the unique fixed point q^* regardless of the starting function q_0.

(B) Policy iteration converges to the optimal policy π^* in a finite number of iterations since there is a finite number of deterministic policies.

(C) Increasing the discount factor γ (closer to 1) improves the speed of convergence of value iteration because the contraction rate becomes smaller.

(D) The existence of a unique q^* implies that the greedy policy improvement step will always strictly increase the value of the current policy.

Correct Answers: A & B

(A) True. The Bellman operator T satisfies

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

showing it is a contraction mapping. By the Banach fixed-point theorem, value iteration $q_{n+1} = Tq_n$ converges to the unique fixed point q^* from any starting point q_0.

(B) True. Policy iteration alternates between policy evaluation and policy improvement. Since there are finitely many deterministic policies ($|\mathcal{A}|^{|\mathcal{S}|}$), and each improvement step increases the policy value unless it is already optimal, convergence to π^* happens in a finite number of steps.

(C) False. Increasing γ slows down convergence. The contraction factor is γ, and as γ approaches 1, T becomes "less contracting," meaning the distance between successive iterates shrinks more slowly.

(D) False. Policy improvement guarantees that the new policy is no worse than the old one. However, if the current policy is already optimal, the greedy improvement step will not strictly improve the policy — it will remain the same.

Q2

Consider the Bellman operator T defined as

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

for a Markov Decision Process (MDP) with discount factor $\gamma \in [0,1)$ and where q is any bounded function on $\mathcal{S} \times \mathcal{A}$. Which of the following statements are correct? (Select all that apply.)

a) T **is a contraction mapping with a contraction factor equal to** γ**.**
b) T **has at least one fixed point in the space of bounded functions.**
c) For any initial q_0**, repeatedly applying** T **(i.e.,** $q_{n+1} = Tq_n$**) converges to the unique fixed point** q^***.**
d) T **always increases every value of** q **so that** $(Tq)(s,a) \geq q(s,a)$ **for all states** s **and actions** a**.**

Correct Answers: A $\&$ B $\&$ C

- **Option A: True.**
  The Bellman operator satisfies

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

  showing that it is a contraction mapping with contraction factor γ under the supremum norm.
- **Option B: True.**
  By Banach's Fixed Point Theorem, any contraction mapping over a complete metric space (such as bounded functions under the supremum norm) has at least one fixed point. Hence, T has a unique fixed point q^*.
- **Option C: True.**
  Since T is a contraction, the sequence $\{q_n\}$ defined by $q_{n+1} = Tq_n$ converges to the unique fixed point q^* for any initial q_0.
- **Option D: False.**
  The Bellman operator is monotonic (if $q \leq q'$, then $Tq \leq Tq'$), but it does not necessarily satisfy $(Tq)(s,a) \geq q(s,a)$ for all q. It may decrease some values if q is initially overestimated relative to q^*.

Q3

Explain why the Bellman operator is considered a contraction mapping under the supremum norm and how this property guarantees the convergence of value iteration to the optimal action-value function q^*.

Under the supremum norm, the Bellman operator T satisfies:

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

for any two bounded functions q and q', where $\gamma \in [0, 1)$ is the discount factor. This inequality shows that T reduces the "distance" (measured by the supremum norm) between any two functions by at least a factor of γ.
Since $\gamma < 1$, T is a *contraction mapping*.

By the Banach Fixed-Point Theorem, any contraction mapping on a complete metric space (in this case, the space of bounded functions equipped with the supremum norm) has:

- **A unique fixed point, and**
- **Guaranteed convergence of the sequence generated by repeatedly applying T starting from any initial function q_0.**

Thus, starting from any q_0 and defining:

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

the sequence $\{q_n\}$ will converge to the unique fixed point q^*, which is the optimal action-value function satisfying the Bellman optimality equation.