Online and Reinforcement Learning (2025)
Home Assignment 2

Davide Marchi 777881

Contents

1 Short Questions

Determine whether each statement below is True or False and provide a very brief justification.

1. **Statement:** *"In a finite discounted MDP, every possible policy induces a Markov Reward Process."*

   Answer: False. This statement assumes that the policy depends only on the current state. If we allow policies to depend on the *entire* past history (*history-dependent* policies), then the resulting transitions in the state space may no longer satisfy the Markov property, since the chosen action at each step might be a function of all previous states and actions. Hence not *every* (fully history-dependent) policy necessarily induces a Markov Reward Process in the *original* state space.

2. **Statement:** *"Consider a finite discounted MDP, and assume that π is an optimal policy. Then, the action(s) output by π does not depend on history other than the current state (i.e., π is necessarily stationary)."*

   Answer: False. While it is true that there *exists* an optimal policy which is stationary deterministic, it does not follow that *all* optimal policies must be so. In fact, multiple distinct policies (some stationary, others possibly history-dependent or randomized) can achieve exactly the same optimal value. Hence it is incorrect to say that *any* optimal policy π must be purely state-dependent (stationary).

3. **Statement:** *"n a finite discounted MDP, a greedy policy with respect to optimal action-value function, Q^*, corresponds to an optimal policy."*

   Answer: True. From the Bellman optimality equations for Q^*, a policy that selects

   $$\arg\max_a \; Q^*(s, a)$$

   at each state s is indeed an optimal policy. This policy attains the same value as Q^* itself, thus achieving the optimal value.

4. **Statement:** *"Under the coverage assumption, the Weighted Importance Sampling Estimator $\widehat{V}_{\text{wIS}}$ converges to V^π with probability 1."*

   Answer: True. The coverage assumption ensures that the target policy's state-action probabilities are absolutely continuous w.r.t. the behavior policy. Under this assumption, Weighted Importance Sampling (though slightly biased) is a *consistent* estimator of V^π, meaning it converges almost surely to V^π as the sample size grows unbounded.

2 MDPs with Similar Parameters Have Similar Values

Setup: We have two discounted MDPs

$$M_1 \;=\; (S, A, P_1, R_1, \gamma) \quad \text{and} \quad M_2 \;=\; (S, A, P_2, R_2, \gamma),$$

sharing the same discount factor $\gamma \in (0,1)$, the same finite state–action space, and rewards bounded in $[0, R_{\max}]$. For all state–action pairs (s,a):

$$\bigl|\, R_1(s,a) \;-\; R_2(s,a) \bigr| \;\le\; \alpha, \quad \bigl\|\, P_1(\cdot \mid s,a) \;-\; P_2(\cdot \mid s,a) \bigr\|_1 \;\le\; \beta.$$

Consider a fixed stationary policy π, and let V_1^π and V_2^π be its value functions in M_1 and M_2, respectively. The goal is to show that

$$\bigl|\, V_1^\pi(s) \;-\; V_2^\pi(s) \bigr| \;\le\; \frac{\alpha \;+\; \gamma\, R_{\max}\, \beta}{(1-\gamma)^2} \quad \text{for every state } s \;\in\; S.$$

Step 1: Write down the Bellman equations for each MDP.
By definition of π, the Bellman fixed-point form is:

$$V_1^\pi \;=\; r_1^\pi + \gamma\, P_1^\pi\, V_1^\pi, \quad V_2^\pi \;=\; r_2^\pi + \gamma\, P_2^\pi\, V_2^\pi,$$

where

$$r_m^\pi(s) \;=\; R_m\bigl(s, \pi(s)\bigr), \quad (P_m^\pi f)(s) \;=\; \sum_{s'} P_m\bigl(s' \mid s, \pi(s)\bigr)\, f(s'), \quad m = 1, 2.$$

Define $\delta \;=\; V_1^\pi - V_2^\pi$. Then

$$\delta \;=\; \bigl(r_1^\pi - r_2^\pi\bigr) \;+\; \gamma\, \bigl(P_1^\pi V_1^\pi - P_2^\pi V_2^\pi\bigr).$$

To facilitate the separation of terms, we introduce and subtract $\gamma\, P_1^\pi V_2^\pi$, which allows us to rewrite the second term as:

$$P_1^\pi V_1^\pi - P_2^\pi V_2^\pi \;=\; (P_1^\pi V_1^\pi - P_1^\pi V_2^\pi) \;+\; (P_1^\pi V_2^\pi - P_2^\pi V_2^\pi).$$

Substituting this back, we obtain:

$$\delta \;=\; \bigl(r_1^\pi - r_2^\pi\bigr) \;+\; \gamma\, P_1^\pi\bigl(V_1^\pi - V_2^\pi\bigr) \;+\; \gamma\, \bigl(P_1^\pi - P_2^\pi\bigr)\, V_2^\pi.$$

Step 3: Take norms and use triangle/inequality bounds.
Taking the supremum norm ($\|\cdot\|_\infty$) on both sides,

$$\|\delta\|_\infty \;\le\; \|\, r_1^\pi - r_2^\pi \|_\infty \;+\; \gamma\, \|\, P_1^\pi\, \delta\, \|_\infty \;+\; \gamma\, \|\, \bigl(P_1^\pi - P_2^\pi\bigr)\, V_2^\pi \|_\infty.$$

- *Reward difference:*
  Since $\big|R_1(s,a) - R_2(s,a)\big| \le \alpha$, it follows that $\|r_1^\pi - r_2^\pi\|_\infty \le \alpha$.

- *Term with $P_1^\pi\,\delta$:*
  We have
  $$\big\|\, P_1^\pi\,\delta \,\big\|_\infty \;\le\; \|\delta\|_\infty,$$
  since P_1^π is a probability kernel and thus a contraction in sup norm.

- *Term with $(P_1^\pi - P_2^\pi)\,V_2^\pi$:*
  For each s,
  $$\big|\,(P_1^\pi - P_2^\pi)\,V_2^\pi(s)\big| \;\le\; \sum_{s'}\big|\,P_1(s' \mid s,\pi(s)) - P_2(s' \mid s,\pi(s))\big|\,\big|\,V_2^\pi(s')\big|.$$
  By assumption, $\|P_1(\cdot \mid s,a) - P_2(\cdot \mid s,a)\|_1 \le \beta$, and $\|V_2^\pi\|_\infty \le \frac{R_{\max}}{1-\gamma}$. Hence,
  $$\big\|\,(P_1^\pi - P_2^\pi)\,V_2^\pi\big\|_\infty \;\le\; \beta\,\frac{R_{\max}}{1-\gamma}.$$

Putting these bounds together,

$$\|\delta\|_\infty \;\le\; \alpha \;+\; \gamma\,\|\delta\|_\infty \;+\; \gamma\,\beta\,\frac{R_{\max}}{1-\gamma}.$$

Step 4: Solve for $\|\delta\|_\infty$.
We isolate $\|\delta\|_\infty$ on one side:

$$(1-\gamma)\,\|\delta\|_\infty \;\le\; \alpha \;+\; \gamma\,\beta\,\frac{R_{\max}}{1-\gamma}.$$

Thus

$$\|\delta\|_\infty \;\le\; \frac{\alpha}{1-\gamma} \;+\; \frac{\gamma\,\beta\,R_{\max}}{(1-\gamma)^2}.$$

Since $\alpha/(1-\gamma) \;\le\; \alpha/(1-\gamma)^2$ whenever $0 < \gamma < 1$, we can write

$$\|\delta\|_\infty \;\le\; \frac{\alpha \;+\; \gamma\,\beta\,R_{\max}}{(1-\gamma)^2}.$$

Hence, for every state $s \in S$,

$$\big|\,V_1^\pi(s) - V_2^\pi(s)\big| \;\;\le\;\; \|\delta\|_\infty \;\le\; \frac{\alpha \;+\; \gamma\,R_{\max}\,\beta}{(1-\gamma)^2}.$$

This is the desired result.

3 Policy Evaluation in RiverSwim

4 Solving a Discounted Grid-World

5 Off-Policy Evaluation in Episode-Based River-Swim