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 $\widehat{V}_{\mathrm{wIS}}$ 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

We have two finite discounted MDPs

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

with the same discount factor $\gamma \in (0, 1)$ and state–action space $S \times A$. Suppose the reward functions satisfy $R_m(s, a) \in [0, R_{\max}]$, and for all (s, a):

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

Fix a stationary (deterministic) policy π. Define the "reward" and "transition" components in each MDP as follows:

$$r_1^\pi(s) \;=\; R_1\big(s, \pi(s)\big), \quad (P_1^\pi f)(s) \;=\; \sum_{s'} P_1\big(s' \mid s, \pi(s)\big)\, f(s'),$$

$$r_2^\pi(s) \;=\; R_2\big(s, \pi(s)\big), \quad (P_2^\pi f)(s) \;=\; \sum_{s'} P_2\big(s' \mid s, \pi(s)\big)\, f(s').$$

Then the respective value functions satisfy the Bellman equations:

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

Let $\delta \;=\; V_1^\pi \;-\; V_2^\pi$. Subtracting the two equations gives

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

We add and subtract the term $\gamma P_1^\pi V_2^\pi$:

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

or equivalently

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

Taking the supremum norm (i.e. $\|\cdot\|_\infty$) on both sides yields

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

Step 1: Bounding the Rewards. By assumption,

$$\big|r_1^\pi(s) - r_2^\pi(s)\big| \;=\; \big|R_1(s, \pi(s)) - R_2(s, \pi(s))\big| \;\le\; \alpha,$$

so $\|\, r_1^\pi - r_2^\pi\|_\infty \le \alpha$.

Step 2: Bounding $\|P_1^\pi \delta\|_\infty$. Since P_1^π is a probability kernel,

$$\big|(P_1^\pi \delta)(s)\big| \;=\; \Big|\sum_{s'} P_1\big(s' \mid s, \pi(s)\big)\, \delta(s')\Big| \;\leq\; \sum_{s'} P_1(s' \mid s, \pi(s))\, |\delta(s')| \;\leq\; \|\delta\|_\infty.$$

Hence $\big\|P_1^\pi \delta\big\|_\infty \leq \|\delta\|_\infty$.

Step 3: Bounding $\|\,(P_1^\pi - P_2^\pi)\,V_2^\pi\|\_\infty$. We have, for each s,

$$\big|\big((P_1^\pi - P_2^\pi)\,V_2^\pi\big)(s)\big| \;=\; \big|\sum_{s'}\big[P_1(s' \mid s, \pi(s)) - P_2(s' \mid s, \pi(s))\big]\,V_2^\pi(s')\big|.$$

By the triangle inequality and definition of the L_1 norm,

$$\leq\; \sum_{s'}\big|P_1(s' \mid s, \pi(s)) - P_2(s' \mid s, \pi(s))\big|\,\big|V_2^\pi(s')\big| \;\;\leq\;\; \Big\|\big(P_1 - P_2\big)(s, \cdot)\Big\|_1 \;\cdot\; \|V_2^\pi\|_\infty.$$

Since $\|(P_1 - P_2)(s, \cdot)\|_1 \leq \beta$, and V_2^π is bounded by $\|V_2^\pi\|_\infty \leq \frac{R_{\max}}{1-\gamma}$, we get

$$\big\|\,(P_1^\pi - P_2^\pi)\,V_2^\pi\big\|_\infty \;\;\leq\;\; \beta\,\frac{R_{\max}}{1-\gamma}.$$

Combining the Bounds. Putting all parts together in

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

we isolate $\|\delta\|_\infty$:

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

As a final simplification, we note that $\alpha/(1-\gamma) \;\leq\; \alpha/(1-\gamma)^2$ since $0 < \gamma < 1$. Hence

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

In other words, for every $s \in S$,

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

This completes the proof.

3 Policy Evaluation in RiverSwim

4 Solving a Discounted Grid-World

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