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

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

We let π be any fixed *stationary deterministic* (or stationary randomized) policy, and write V_m^π to denote its value function in M_m. We want to prove:

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

Proof of (ii): Fix $s \in S$. By definition of the value function under policy π, we have

$$V_m^\pi(s) \;=\; \sum_{a \in A} \pi(a \mid s) \Bigl[R_m(s,a) \;+\; \gamma \sum_{x \in S} P_m\bigl(x \mid s,a\bigr)\, V_m^\pi(x)\Bigr] \quad \text{for } m = 1, 2.$$

Taking their difference:

$$\bigl|V_1^\pi(s) - V_2^\pi(s)\bigr| \;=\; \Bigl|\sum_a \pi(a \mid s)\Bigl[R_1(s,a) + \gamma \sum_x P_1(x \mid s,a)\, V_1^\pi(x) - \bigl(R_2(s,a) + \gamma \sum_x P_2(x \mid s,a)\, V_2^\pi(x$$

Use the triangle inequality, plus linearity of the sum:

$$\le \sum_a \pi(a \mid s) \Bigl| \underbrace{R_1(s,a) - R_2(s,a)}_{\le \alpha} + \gamma \sum_x P_1(x \mid s,a)\, V_1^\pi(x) \;-\; \gamma \sum_x P_2(x \mid s,a)\, V_2^\pi(x) \Bigr|.$$

Hence

$$\bigl|V_1^\pi(s) \;-\; V_2^\pi(s)\bigr| \;\le\; \sum_a \pi(a \mid s) \Bigl[\alpha \;+\; \gamma \Bigl|\sum_x P_1(x \mid s,a)\, V_1^\pi(x) \;-\; \sum_x P_2(x \mid s,a)\, V_2^\pi(x)\Bigr|\Bigr].$$

We now split that big absolute difference into two parts:

$$\begin{aligned}
&\Bigl|\sum_x P_1(x \mid s,a)\, V_1^\pi(x) \;-\; \sum_x P_2(x \mid s,a)\, V_2^\pi(x)\Bigr| \\
&\le \Bigl|\sum_x P_1(x \mid s,a)\, \bigl(V_1^\pi(x) - V_2^\pi(x)\bigr)\Bigr| \;+\; \Bigl|\sum_x \bigl(P_1(x \mid s,a) - P_2(x \mid s,a)\bigr)\, V_2^\pi(x)\Bigr| \\
&\le \sum_x P_1(x \mid s,a)\, \bigl|V_1^\pi(x) - V_2^\pi(x)\bigr| \;+\; \sum_x \bigl|P_1(x \mid s,a) - P_2(x \mid s,a)\bigr|\, \bigl|V_2^\pi(x)\bigr|.
\end{aligned}$$

Since $\left|V_2^\pi(x)\right| \le \frac{R_{\max}}{1-\gamma}$ for discounted MDPs, and $\sum_x \left|P_1(x \mid s,a) - P_2(x \mid s,a)\right| \le \beta$, it follows that:

$$\Big|\sum_x P_1(x \mid s,a)\, V_1^\pi(x) \;-\; \sum_x P_2(x \mid s,a)\, V_2^\pi(x)\Big| \;\le\; \sup_x \left|V_1^\pi(x) - V_2^\pi(x)\right| \;+\; \beta\, \frac{R_{\max}}{1-\gamma}.$$

Let

$$\Delta \;=\; \sup_{s \in S} \left|\, V_1^\pi(s) \;-\; V_2^\pi(s)\right|.$$

Then combining everything above,

$$\left|V_1^\pi(s) \;-\; V_2^\pi(s)\right| \;\le\; \alpha \;+\; \gamma \Big(\Delta \;+\; \beta\, \frac{R_{\max}}{1-\gamma}\Big).$$

Taking the supremum in s on the left side gives

$$\Delta \;\le\; \alpha \;+\; \gamma\,\Delta \;+\; \gamma\,\beta\, \frac{R_{\max}}{1-\gamma}.$$

Hence,

$$(\,1-\gamma)\,\Delta \;\le\; \alpha \;+\; \gamma\,\beta\, \frac{R_{\max}}{1-\gamma} \quad\Longrightarrow\quad \Delta \;\le\; \frac{\alpha}{1-\gamma} \;+\; \frac{\gamma\,\beta\, R_{\max}}{(1-\gamma)^2}\;.$$

Finally, we can note that $\alpha \le \alpha/(1-\gamma)$, or equivalently multiply out and observe

$$\frac{\alpha}{1-\gamma} \;=\; \frac{\alpha\,\big(1-\gamma\big)}{(1-\gamma)^2} \;\le\; \frac{\alpha}{(1-\gamma)^2}\;.$$

So we get

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

Thus, for every state s,

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

This completes the proof of part (ii).

3 Policy Evaluation in RiverSwim

4 Solving a Discounted Grid-World

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