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 recall the setting: 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 $0 < \gamma < 1$ and finite state–action space. For each (s, a) we have:

$$\big|R_1(s,a) - R_2(s,a)\big| \;\le\; \alpha, \quad \big\|P_1(\cdot \mid s,a) - P_2(\cdot \mid s,a)\big\|_1 \;\le\; \beta, \quad R_1(s,a),\, R_2(s,a) \;\in\; [0, R_{\max}].$$

Let π be any fixed stationary policy (deterministic or randomized), and let $V_1^\pi,\ V_2^\pi$ denote its value functions in M_1 and M_2, respectively. We wish to show that, for every $s \in S$,

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

Bellman Operators. Define the Bellman operator T_m^π for each M_m $(m = 1, 2)$ by

$$(T_m^\pi V)(s) \;=\; \sum_{a \in A} \pi(a \mid s)\Big[\, R_m(s,a) \;+\; \gamma \sum_{s'} P_m(s' \mid s,a)\, V(s')\Big].$$

Then V_m^π is the unique fixed point: $V_m^\pi = T_m^\pi\, V_m^\pi$, i.e.

$$V_m^\pi(s) \;=\; (T_m^\pi V_m^\pi)(s) \;=\; \sum_{a} \pi(a \mid s)\Big[R_m(s,a) \;+\; \gamma \sum_{s'} P_m(s' \mid s,a)\, V_m^\pi(s')\Big].$$

Step 1: Decompose the difference. For each $s \in S$, we have

$$V_1^\pi(s) \;-\; V_2^\pi(s) \;=\; \big(T_1^\pi V_1^\pi\big)(s) \;-\; \big(T_2^\pi V_2^\pi\big)(s).$$

Add and subtract $\big(T_1^\pi V_2^\pi\big)(s)$ inside the absolute value:

$$\begin{aligned} \big|V_1^\pi(s) - V_2^\pi(s)\big| &= \big|\, T_1^\pi V_1^\pi(s) \;-\; T_2^\pi V_2^\pi(s)\big| \\ &\le \underbrace{\big|\, T_1^\pi V_1^\pi(s) \;-\; T_1^\pi V_2^\pi(s)\big|}_{\text{(1) same operator, diff in values}} \;+\; \underbrace{\big|\, T_1^\pi V_2^\pi(s) \;-\; T_2^\pi V_2^\pi(s)\big|}_{\text{(2) same value, diff in operators}}\,. \end{aligned}$$

(1) Same operator, difference in the value functions.
Since T_1^π has discount factor γ,

$$\big|\, T_1^\pi V_1^\pi(s) - T_1^\pi V_2^\pi(s)\big| \;=\; \gamma\, \sum_{a} \pi(a \mid s)\, \Big|\sum_{s'} P_1(s' \mid s,a)\, V_1^\pi(s') - \sum_{s'} P_1(s' \mid s,a)\, V_2^\pi(s')\Big| \;\le\; \gamma\, \max_{x} \big|\, V_1^\pi($$

Define

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

Thus the first portion is at most $\gamma\,\Delta$.

(2) Same value function, difference in operators.
Next,

$$\bigl|\,T_1^{\pi}V_2^{\pi}(s) - T_2^{\pi}V_2^{\pi}(s)\bigr| \;=\; \Bigl|\,\sum_a \pi(a\mid s)\,\bigl[R_1(s,a)+\gamma\sum_{s'} P_1(s'\mid s,a)\,V_2^{\pi}(s')\bigr] - \sum_a \pi(a\mid s)\,\bigl[R_2(s,a)+\gamma\sum$$

We can group the terms:

$$\le\; \sum_a \pi(a\mid s)\Bigl|\underbrace{R_1(s,a)-R_2(s,a)}_{\le\alpha} + \gamma\sum_{s'}\bigl[P_1(s'\mid s,a)-P_2(s'\mid s,a)\bigr]\,V_2^{\pi}(s')\Bigr|.$$

Hence

$$\bigl|\,T_1^{\pi}V_2^{\pi}(s) - T_2^{\pi}V_2^{\pi}(s)\bigr| \;\le\; \sum_a \pi(a\mid s)\Bigl[\alpha \;+\; \gamma\,\Bigl|\sum_{s'}\bigl[P_1(s'\mid s,a)-P_2(s'\mid s,a)\bigr]\,V_2^{\pi}(s')\Bigr|\Bigr].$$

Since

$$\sum_{s'}\bigl|P_1(s'\mid s,a)-P_2(s'\mid s,a)\bigr| \;\le\; \beta, \quad \text{and} \quad \bigl|V_2^{\pi}(s')\bigr| \;\le\; \frac{R_{\max}}{1-\gamma},$$

it follows that

$$\Bigl|\sum_{s'}\bigl[P_1(s'\mid s,a)-P_2(s'\mid s,a)\bigr]\,V_2^{\pi}(s')\Bigr| \;\le\; \beta\,\frac{R_{\max}}{1-\gamma}.$$

Hence

$$\bigl|\,T_1^{\pi}V_2^{\pi}(s) - T_2^{\pi}V_2^{\pi}(s)\bigr| \;\le\; \alpha \;+\; \gamma\,\beta\,\frac{R_{\max}}{1-\gamma}.$$

Combine (1) & (2).
Putting the two pieces together:

$$\bigl|V_1^{\pi}(s) - V_2^{\pi}(s)\bigr| \;\le\; \underbrace{\gamma\,\Delta}_{\text{from (1)}} \;+\; \underbrace{\alpha + \gamma\,\beta\,\frac{R_{\max}}{1-\gamma}}_{\text{from (2)}}.$$

Therefore, taking supremum over $s \in S$:

$$\Delta \;=\; \sup_s\bigl|\,V_1^{\pi}(s) - V_2^{\pi}(s)\bigr| \;\le\; \gamma\,\Delta \;+\; \alpha \;+\; \gamma\,\beta\,\frac{R_{\max}}{1-\gamma}.$$

Rearranging gives

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

As in the usual derivations, since $\alpha/(1-\gamma) \;\le\; \alpha/(1-\gamma)^2$ whenever $0<\gamma<1$, one can in fact write

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

Thus, for every $s \in S$,

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

This completes the alternate proof via Bellman equations.

3 Policy Evaluation in RiverSwim

4 Solving a Discounted Grid-World

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