TÉCNICO LISBOA

Planning, Learning and Decision Making

MSc in Computer Science and Engineering

Final examination — May 3, 2022

Instructions

- You have 120 minutes to complete the examination.
- Make sure that your test has a total of 10 pages and is not missing any sheets, then write your full name and student n. on this page (and your number in all others).
- The test has a total of 5 questions, with a maximum score of 20 points. The questions have different levels of difficulty. The point value of each question is provided next to the question number.
- *If you get stuck in a question, move on.* You should start with the easier questions to secure those points, before moving on to the harder questions.
- *No interaction with the faculty is allowed during the exam.* If you are unclear about a question, clearly indicate it and answer to the best of your ability.
- Please provide your answer in the space below each question. If you make a mess, clearly indicate your answer.
- The exam is open book and open notes. You may use a calculator, but any other type of electronic or communication equipment is not allowed.
- Good luck.

Question 1. (**2.5 pts.**)

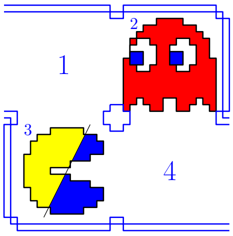


Figure 1: Pacman and ghost in a 2×2 grid.

Consider the environment in Fig. 1, representing the Pacman scenario that you already encountered in several homework assignments. Unlike the homework, you will now model *the ghost* as a decision maker.

The ghost has 4 actions available: "Up", "Down", "Left", and "Right", each of which moves the ghost character one step in the corresponding direction, if an adjacent cell exists in that direction. Otherwise, the ghost remains in the same place. The ghost can be in any of the 4 numbered cells; the cell in the top left corner (cell 1) is adjacent, to the left, to the cell in the lower right corner (cell 4). In other words, if the ghost "moves left" at cell 1, it will end up in cell 4, and vice-versa.

Pacman does not move: it is always in cell 3. However, it can be in one of two states: "normal" or "super-powered". The ghost does not observe the state of Pacman, except when standing in the same cell as Pacman.

If the ghost lies in the same cell as a normal-state Pacman (in cell 3) the game should transition to a "Victory" state, irrespectively of the ghost's action. However, if the ghost lies in the same cell as a super-powered Pacman, it is "eaten" by Pacman. In that case, the game should transition to a "Defeat" state, irrespectively of the ghost's action. The ghost's goal is to reach the "Victory" state.

Describe the decision problem faced by the agent using the adequate type of model. In particular, you should indicate:

- The type of model needed to describe the decision problem of the agent;
- The state, action, and observation space (if relevant);
- The transition probabilities corresponding to the action "Move Left";
- The observation probabilities for the action "Move Left" (if relevant);
- The immediate cost function.

Make sure that

- Both victory and defeat are *absorbing states.*
- The cost function is as simple as possible and verifies $c(x, a) \in [0, 1]$ for all states $x \in \mathcal{X}$ and actions $a \in \mathcal{A}$.
- The cost depends only on the state of the environment.
- Both absorbing states (victory and defeat) should have a cost of 0.

Solution 1.

Since the ghost is unable to determine the state of Pacman, and this information is relevant for its decision-making process, the decision problem has partial observability and, as such, is modeled using a POMDP. We have:

- The states include information about the position of the ghost and the state of Pacman. This yields $\mathcal{X} = \{(1, \emptyset), (2, \emptyset), (3, \emptyset), (4, \emptyset), (1, P), (2, P), (3, P), (4, P), V, D\}$, where states V and D correspond to "Victory" and "Defeat", and states (g, p) include the position of the ghost, g, and the state of Pacman, p, with $g \in \{1, 2, 3, 4\}$ and $p \in \{\emptyset, P\}$.
- The actions correspond to the movements of the ghost, "Up", "Down", "Left", and "Right", yielding $\mathcal{A} = \{U, D, L, R\}$ (each action is represented by its first letter).
- The observations comprise the position of the ghost and, when in cell 3, the state of Pacman, yielding $\mathcal{Z} = \{1, 2, (3, \emptyset), (3, P), 4, V, D\}$.
- The action "Move left" moves Pacman to the adjacent cell to the left, yielding the transition and observation probabilities

$$\mathbf{P}_L = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}; \qquad \mathbf{O}_L = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

- Finally, the cost function comes

$$\boldsymbol{C} = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 1.0 & 1.0 & 1.0 & 1.0 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}.$$

In the remainder of the test, consider the POMDP $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{Z}, \{\mathbf{P}_a\}, \{\mathbf{O}_a\}, c, \gamma)$ where

- $\mathcal{X} = \{A, B, C\}$;
- $\mathcal{A} = \{a, b, c\}$;
- $\mathcal{Z} = \{u, v\}$;
- The transition probabilities are

$$\mathbf{P}_a = \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \end{bmatrix}; \qquad \mathbf{P}_b = \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 1.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}; \qquad \mathbf{P}_c = \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}.$$

- The observation probabilities are

$$\mathbf{O}_a = \mathbf{O}_b = \mathbf{O}_c = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \\ 0.0 & 1.0 \end{bmatrix}.$$

- The cost function c is given by

$$\mathbf{C} = \begin{bmatrix} 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 1.0 \\ 0.2 & 0.2 & 0.0 \end{bmatrix}.$$

- Finally, the discount is given by $\gamma = 0.9$.

You may also find useful the fact that, given a 3 × 3 matrix

$$\mathbf{A} = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix},$$

it holds that

$$\mathbf{A}^{-1} = \begin{bmatrix} \frac{1}{a} & -\frac{b}{ad} & \frac{be-cd}{adf} \\ 0 & \frac{1}{d} & -\frac{e}{df} \\ 0 & 0 & \frac{1}{f} \end{bmatrix}.$$

Question 2. (8 pts.)

For each of the following questions, indicate the *single most correct answer.*

(a) **(0.8 pts.)** Let $\succ$ denote a strict preference relation on some set $\mathcal{X}$. Then, it holds that...

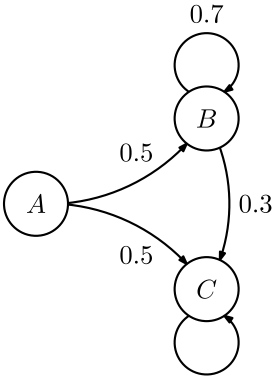
□ ... given two outcomes x, y, it is possible that both $x \succ y$ and $y \succ x$.

□ ... given two outcomes x, y, either $x \succ y$, or $y \succ x$.

☒ **... given three outcomes x, y, z, if $x \not\succ y$ and $y \not\succ z$, then $x \not\succ z$.**

□ None of the above.

(b) **(0.8 pts.)** Consider the Markov chain defined by the transition diagram



The transition probability matrix for the chain is:

□ $\mathbf{P} = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 0.7 & 0.3 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$.

☒ $\mathbf{P} = \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.0 & 0.7 & 0.3 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$.

□ $\mathbf{P} = \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.5 & 0.7 & 0.0 \\ 0.5 & 0.3 & 1.0 \end{bmatrix}$.

□ None of the above.

(c) **(0.8 pts.)** Consider once again the Markov chain in question (b), and suppose that the initial state distribution is $\boldsymbol{\mu}_0 = \begin{bmatrix} 0.0 & 0.5 & 0.5 \end{bmatrix}$. The state distribution at time step $t = 1$ is

□ $\boldsymbol{\mu}_1 = \begin{bmatrix} 0.0 & 0.7 & 0.3 \end{bmatrix}$.

☒ $\boldsymbol{\mu}_1 = \begin{bmatrix} 0.0 & 0.35 & 0.65 \end{bmatrix}$.

□ $\boldsymbol{\mu}_1 = \begin{bmatrix} 0.0 & 0.0 & 1.0 \end{bmatrix}$.

□ None of the above.

(d) **(0.8 pts.)** Consider the HMM obtained from $\mathcal{M}$ when the agent follows a fixed policy that always selects action c. Suppose that the initial state distribution is $\boldsymbol{\mu}_0 = \begin{bmatrix} 0.0 & 0.5 & 0.5 \end{bmatrix}$. Further suppose that, at time step $t = 1$, the agent observes $z_1 = u$. It holds that

- □ $\boldsymbol{\alpha}_1 = \begin{bmatrix} 0.0 & 0.5 & 0.5 \end{bmatrix}$.
- ☒ $\boldsymbol{\alpha}_0 = \begin{bmatrix} 0.0 & 0.5 & 0.5 \end{bmatrix}$**.**
- □ $\boldsymbol{\mu}_1 = \begin{bmatrix} 0.0 & 0.5 & 0.5 \end{bmatrix}$.
- □ None of the above.

(e) **(0.8 pts.)** Given a sequence of observations $z_{0:T}$, to compute the most likely sequence of states given $z_{0:T}$, one should use...

- □ ... the forward algorithm.
- ☒ **... the Viterbi algorithm.**
- □ ... the Baum-Welch algorithm.
- □ None of the above.

(f) **(0.8 pts.)** In *inverse reinforcement learning*, the goal is to...

- □ ... given samples of an optimal policy, generalize that policy to unseen states.
- □ ... learn an optimal policy by trial-and-error.
- □ ... learn a policy from examples by a teacher.
- ☒ **... given the transition probabilities of an MDP, identify the cost function that renders a given policy optimal.**

(g) **(0.8 pts.)** Let π^* denote the optimal policy for an arbitrary MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$.

- □ Given an arbitrary policy π, π^* can be obtained from π as $\pi^*(x) = \operatorname{argmin}_{a\in\mathcal{A}} Q^\pi(x, a)$, for all $x \in \mathcal{X}$.
- ☒ **It must hold that** $(\mathbf{P}_{\pi^*} - \mathbf{P}_a)(\boldsymbol{I} - \gamma\boldsymbol{P}_{\pi^*})^{-1}\mathbf{c}_{\pi^*} \leq \mathbf{0}$ **for all** $a \in \mathcal{A}$**.**
- □ Given an arbitrary policy π, J^{π^*} can be obtained as $J^{\pi^*}(x) = \min_{a\in\mathcal{A}} Q^\pi(x, a)$, for all $x \in \mathcal{X}$.
- □ Given an arbitrary policy π, it must hold that $J^{\pi^*}(x) < J^\pi(x)$ for at least one state $x \in \mathcal{X}$.

(h) **(0.8 pts.)** The UCB algorithm...

- □ ... makes use of randomized action selection to avoid being exploited.
- ☒ **... was proposed to address the stochastic multi-armed bandit problem.**
- □ ... makes no assumptions on how costs are generated.
- □ None of the above.

(i) **(0.8 pts.)** Recall the weight update rule for EXP3:

$$w_{t+1}(a_t) = w_t(a_t)e^{-\eta \frac{c_t}{p_t(a_t)}},$$

where a_t and c_t are, respectively, the action selected and the cost incurred by the algorithm at time step t. The term $p_t(a_t)$ in the update above...

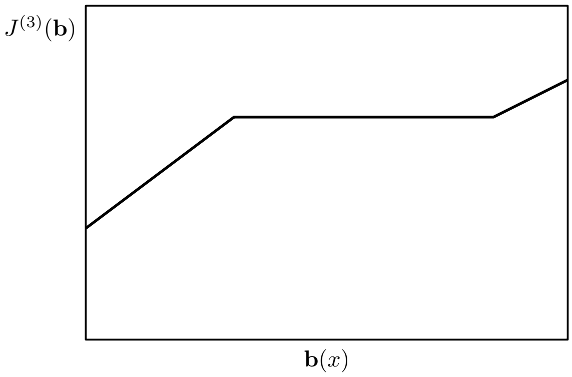
☐ ... is used since the cost of action a_t comes from the distribution $p_t(a_t)$.

☒ **... corresponds to the probability of selecting the action a_t at time step t.**

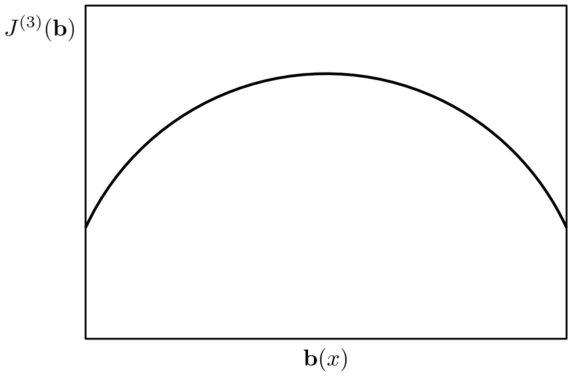
☐ Both of the above.

☐ None of the above.

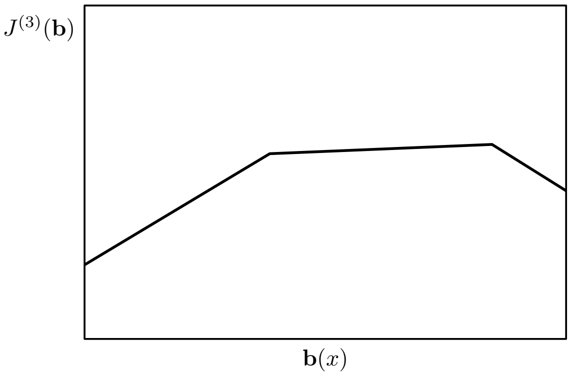
(j) **(0.8 pts.)** Suppose that we run value iteration on a two-state POMDP. Which of the following plots could correspond to the estimate of the cost-to-go function after three iterations of VI?



☐



☐



☒

☐ None of the above.

Question 3. (**4.5 pts.**)

Consider the MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$, obtained from $\mathcal{M}$ by ignoring partial observability. Consider also the matrices

$$c_0 = \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix}, \qquad \mathbf{P}_0 = \begin{bmatrix} 0.0 & 0.5 & 0.5 \\ 0.0 & 0.7 & 0.3 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

(a) **(1.0 pt.)** Indicate a stationary policy π such that $c_\pi = c_0$ and $\mathbf{P}_\pi = \mathbf{P}_0$.

(b) **(1.5 pts.)** For the policy computed in part (a), compute J^π (note that you do not need π to compute J^π).

(c) **(2.0 pts.)** Compute the greedy policy with respect to J^π, $\pi_g^{J^\pi}$. Is the policy π optimal? Explain your reasoning based on the policy $\pi_g^{J^\pi}$ just computed.

Note: If you did not solve (b), you can use

$$\boldsymbol{J}^\pi = \begin{bmatrix} 2 & 2 & 2 \end{bmatrix}^\top.$$

Solution 3.

(a) In state A, all actions yield the same transition probabilities; in state B, action a leads to state B, action b leads to state C, and action c leads to state A. Given the transition probabilities in $\mathbf{P}_0(\cdot \mid B)$, the policy π must select action a with probability 0.7 and action b with probability 0.3. As for state C, the policy π must select action b with probability 1.0. Therefore, the policy π can be, for example

$$\pi = \begin{bmatrix} 0.0 & 0.0 & 1.0 \\ 0.7 & 0.3 & 0.0 \\ 0.0 & 1.0 & 0.0 \end{bmatrix}.$$

(b) We have that

$$\begin{aligned} \boldsymbol{J}^\pi &= (\boldsymbol{I} - \gamma \mathbf{P}_\pi)^{-1} \boldsymbol{c}_\pi \\ &= (\boldsymbol{I} - \gamma \mathbf{P}_0)^{-1} \boldsymbol{c}_0 \\ &= \begin{bmatrix} 1.0 & -0.45 & -0.45 \\ 0.0 & 0.37 & -0.27 \\ 0.0 & 0.0 & 0.1 \end{bmatrix}^{-1} \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix} \\ &= \begin{bmatrix} 1.0 & 1.22 & 7.78 \\ 0.0 & 2.7 & 7.3 \\ 0.0 & 0.0 & 10.0 \end{bmatrix} \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \end{bmatrix} \\ &= \begin{bmatrix} 2.0 \\ 2.0 \\ 2.0 \end{bmatrix}. \end{aligned}$$

(c) To compute the greedy policy, we first compute Q^π. We have that

$$Q^\pi(x, a) = c(x, a) + \gamma \sum_{x' \in \mathcal{X}} \mathbf{P}(x' \mid x, a) J^\pi(x'),$$

which yields

$$\boldsymbol{Q}^\pi = \begin{bmatrix} 0.2 & 0.2 & 0.2 \\ 0.2 & 0.2 & 1.0 \\ 0.2 & 0.2 & 0.0 \end{bmatrix} + \begin{bmatrix} 1.8 & 1.8 & 1.8 \\ 1.8 & 1.8 & 1.8 \\ 1.8 & 1.8 & 1.8 \end{bmatrix} = \begin{bmatrix} 2.0 & 2.0 & 2.0 \\ 2.0 & 2.0 & 2.8 \\ 2.0 & 2.0 & 1.8 \end{bmatrix},$$

, and the greedy policy is given by

$$\pi_g^{J^\pi} = \begin{bmatrix} 0.3 & 0.3 & 0.3 \\ 0.5 & 0.5 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

Since $\pi \neq \pi_g^{J^\pi}$ — namely, in state C, the two policies are clearly different — it follows that π is not optimal.

Question 4. (**3.5 pts.**)

Consider once again the MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$, obtained from $\mathcal{M}$ by ignoring partial observability. Consider the following trajectory, obtained with some policy π:

$$\tau = \{A, 0.2, B, 0.2, B, 0.2, C\}.$$

Further consider the estimate for J^π

$$\hat{\boldsymbol{J}} = \begin{bmatrix} 0.52 \\ 0.56 \\ 0.40 \end{bmatrix}.$$

(a) **(1.5 pts.)** Perform one update of Monte Carlo policy evaluation to $\hat{J}(A)$. Use a step-size of $\alpha = 0.1$. Indicate the relevant computations.

(b) **(1.0 pt.)** Perform one update of TD(0) to $\hat{J}(A)$. Use a step-size of $\alpha = 0.1$.

(c) **(1.0 pt.)** Do you agree with the sentence: "*Monte Carlo policy evaluation suffers from larger bias but smaller variance than TD(0)*"? Explain your answer.

Solution 4.

(a) To perform one update of Monte Carlo policy evaluation, we start by computing

$$L(\tau) = \sum_{t=0}^{T} \gamma^t c_t = 0.2 + \gamma \times 0.2 + \gamma^2 \times 0.2 = 0.542.$$

The resulting update comes:

$$\begin{aligned} \hat{J}(A) &= \hat{J}(A) + \alpha(L(\tau) - \hat{J}(A)) \\ &= 0.52 + 0.1(0.542 - 0.52) = 0.522. \end{aligned}$$

(b) To perform a TD(0) update, we have that

$$\begin{aligned} \hat{J}(A) &= \hat{J}(A) + \alpha(0.2 + \gamma\hat{J}(B) - \hat{J}(A)) \\ &= 0.52 + 0.1(0.2 + 0.9 \times 0.56 - 0.52) = 0.538. \end{aligned}$$

(c) I do not agree with the sentence. Monte Carlo policy evaluation suffers from no bias, since it does not bootstrap (it does not use its current estimate of J to compute the target). However, since it relies on (typically) long trajectories, it is often plagued with large variance. TD(0), on the other hand, suffers significantly less variance, since it relies on a single transition to perform the update. However, the target of the update uses the current estimate of J, which introduces a significant bias.

Question 5. (1.5 pts.)

As seen in class, the policy gradient theorem states that

$$\nabla_\theta V(\theta) = \sum_{x \in \mathcal{X}} \mu_\theta(x) \sum_{a \in \mathcal{A}} \pi_\theta(a \mid x) \nabla_\theta \log \pi_\theta(a \mid x) Q^{\pi_\theta}(x, a),$$

where V is the discounted expected cost-to-go given the initial state distribution. Show that the gradient remains unchanged if the term Q^{π_θ} is replaced by $Q^{\pi_\theta} - J^{\pi_\theta}$, i.e.,

$$\nabla_\theta V(\theta) = \sum_{x \in \mathcal{X}} \mu_\theta(x) \sum_{a \in \mathcal{A}} \pi_\theta(a \mid x) \nabla_\theta \log \pi_\theta(a \mid x) \big(Q^{\pi_\theta}(x, a) - J^{\pi_\theta}(x)\big).$$

Hint: What can you say about $\sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a \mid x)$?

Solution 5.

We start by noting that

$$\sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a \mid x) = \sum_{a \in \mathcal{A}} \pi(a \mid x) \nabla_\theta \log \pi_\theta(a \mid x)$$

while, on the other hand,

$$\sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a \mid x) = \nabla_\theta \sum_{a \in \mathcal{A}} \pi_\theta(a \mid x) = \nabla_\theta(1) = 0.$$

Therefore,

$$\begin{aligned}\sum_{a \in \mathcal{A}} \pi(a \mid x) \nabla_\theta \log \pi_\theta(a \mid x) J^{\pi_\theta}(x) &= \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a \mid x) J^{\pi_\theta}(x) \\ &= J^{\pi_theta}(x) \underbrace{\sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a \mid x)}_{=0} \\ &= 0.\end{aligned}$$

The conclusion follows.