TÉCNICO LISBOA

Planning, Learning and Decision Making

MSc in Computer Science and Engineering

Second examination — July 10, 2023

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.
- In the multiple choice questions, *you do not get negative points* if you get the answer wrong.
- Please provide your answer in the space below each question, and make sure to include all relevant computations. 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.**)

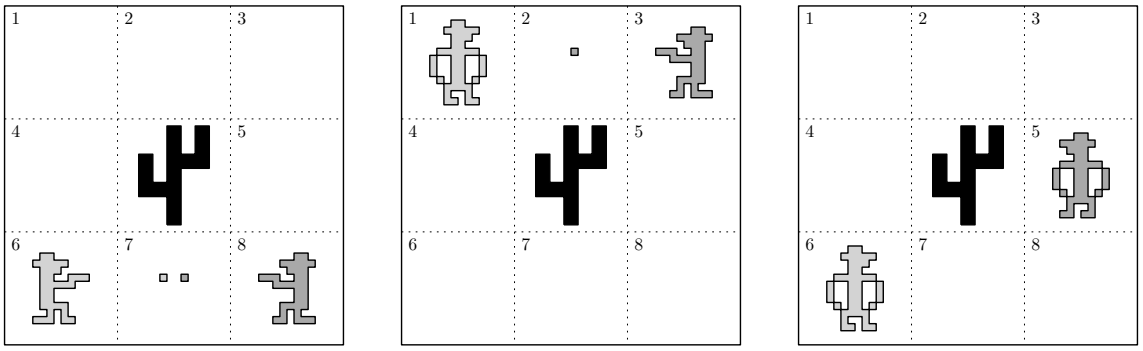


Figure 1: Three example states of a simplified version of the Outlaw game.

Consider the diagrams in Fig. 1, representing three example states from a very simplified version of the game *Outlaw*, from the Atari 2600 console. In it, two cowboys try to shoot one another. The cowboy on the left can only stand in one of the cells 1, 4, and 6. Similarly, the cowboy on the right can only stand in one of the cells 3, 5, and 8. In this game, the player controls the cowboy on the left, while the cowboy on the right is controlled by the game engine. The goal of the game is to shoot the cowboy on the right and avoid being shot.

The player has 4 actions available: "Move up", "Move down", "Stay", and "Shoot". The first two move the cowboy in the corresponding direction, if a cell exists in that direction. The action "Stay" has no effect, while the action "Shoot" shoots a bullet in the direction of the other cowboy if the player is in cells 1 or 6, and has no effect otherwise. The cowboy on the right has the similar actions available, with similar behavior, and at each time step selects its actions uniformly at random.

When a cowboy selects the action "Shoot" at time step t in a position where such action is valid, a bullet will be in the intermediate cell (2 or 7) in time step $t + 1$, and will either disappear or kill the other cowboy at time step $t + 2$ if there is a cowboy in the cell "in front of the bullet". For example, in Fig. 1 on the left panel, both cowboys chose the action "Shoot" in the previous time step and will be shot in the next time step if neither moves from their positions. In the center panel, only the left cowboy will be shot in the next time step if it does not move from its position.[1]

The player wins the game (cost = 0) if it shoots the cowboy on the right without being shot. If the player is shot, it loses the game (cost = 1). If both cowboys are shot, the game is considered a tie (cost = 0.5). All remaining states should have a residual cost (e.g., cost = 0.1).

Describe the decision problem faced by the player 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 spaces (if relevant);
- The transition probabilities **only for the state depicted on the left panel of Fig. 1**, when the agent selects the action "Move up".

[1] Note that, if there is a bullet in a central square at time step t, the cowboy that shot it must be in the cell where it shot it, since it could not have moved in the meantime.

Solution 1.

In order to select its actions, the agent must know the position of the two cowboys and whether there are bullets flying between the two. The problem can be modeled as a *Markov decision problem* $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$, where:

- $\mathcal{X}$ is the state space, and is given by

$$\begin{aligned}\mathcal{X} = \{&(1,3,\emptyset,\emptyset),(1,3,\rightarrow,\emptyset),(1,3,\emptyset,\leftarrow),(1,3,\rightarrow,\leftarrow),(1,5,\emptyset,\emptyset),(1,5,\rightarrow,\emptyset),\\ &(1,8,\emptyset,\emptyset),(1,8,\rightarrow,\emptyset),(1,8,\emptyset,\leftarrow),(1,8,\rightarrow,\leftarrow),(4,3,\emptyset,\emptyset),(4,3,\emptyset,\leftarrow),\\ &(4,5,\emptyset,\emptyset),(4,8,\emptyset,\emptyset),(4,8,\emptyset,\leftarrow),(6,3,\emptyset,\emptyset),(6,3,\rightarrow,\emptyset),(6,3,\emptyset,\leftarrow),\\ &(6,3,\rightarrow,\leftarrow),(6,5,\emptyset,\emptyset),(6,5,\rightarrow,\emptyset),(6,8,\emptyset,\emptyset),(6,8,\rightarrow,\emptyset),(6,8,\emptyset,\leftarrow),\\ &(6,8,\rightarrow,\leftarrow)\}\end{aligned}$$

  The first element corresponds to the position of the left cowboy; the element position to the position of the right cowboy; the third element indicates whether there is a bullet shot by the left cowboy; the last element indicates whether there is a bullet shot by the right cowboy. For example, the three panels in Fig. 1 correspond, respectively, to states $(6,8,\rightarrow,\leftarrow)$, $(1,3,\emptyset,\leftarrow)$, and $(6,5,\emptyset,\emptyset)$.

- $\mathcal{A}$ is the action space, and is given by $\mathcal{A} = \{U, D, St, Sh\}$.

- If the agent is in state $(6,8,\rightarrow,\leftarrow)$ and chooses the action "Move up", the cowboy on the left will move to cell 4. The cowboy on the right will "Move up" with a probability 0.25 (ending up in cell 5) and, with probability 0.75, will choose an action that does not change its position. In that case, it will be shot and the player will win the game. The transition probabilities are, thus,

$$\mathbf{P}_U(y \mid (6,8,\rightarrow,\leftarrow)) = \begin{cases} 0.25 & \text{if } y = (4,8,\emptyset,\emptyset);\\ 0.5 & \text{if } y = (4,5,\emptyset,\emptyset);\\ 0.25 & \text{if } y = (4,8,\emptyset,\leftarrow);\\ 0.0 & \text{otherwise.}\end{cases}$$

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

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

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

- The observation probabilities are

$$\mathbf{O}_a = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \\ 1.0 & 0.0 & 0.0 \end{bmatrix}; \qquad \mathbf{O}_b = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.5 & 0.5 \end{bmatrix}; \qquad \mathbf{O}_c = \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.8 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}.$$

- The cost function c is given by

$$\mathbf{C} = \begin{bmatrix} 0.3 & 0.2 & 0.1 \\ 0.0 & 0.2 & 0.1 \\ 1.0 & 0.2 & 1.0 \end{bmatrix}.$$

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

Question 2. (**8 pts.**)
For each of the following questions, indicate the *single most correct answer.*

(a) **(0.8 pts.)** Consider the Markov chain $(\mathcal{X}, \mathbf{P}_a)$, where $\mathcal{X}$ and $\mathbf{P}_a$ are as defined in $\mathcal{M}$. The stationary distribution for the chain is

- ☐ $\mu = \begin{bmatrix} 0.816 & 0.408 & 0.408 \end{bmatrix}$.
- ☐ $\mu = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$.
- ☒ $\mu = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$.
- ☐ $\mu = \begin{bmatrix} 0.577 & 0.577 & 0.577 \end{bmatrix}$.

(b) **(0.8 pts.)** Consider again the Markov chain $(\mathcal{X}, \mathbf{P}_a)$, and suppose that, at time step $t = 0$, the state of the chain is selected uniformly at random. What is the distribution over states at time step $t = 1$?

- ☐ $\mu_1 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$.
- ☒ $\mu_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{3} \end{bmatrix}$.
- ☐ $\mu_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$.
- ☐ None of the above.

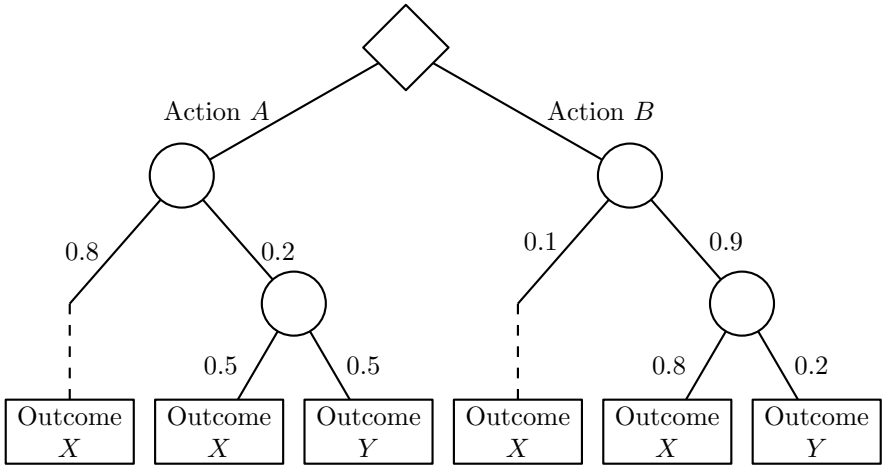
(c) **(0.8 pts.)** Consider now the HMM $(\mathcal{X}, Z, \mathbf{P}_a, \mathbf{O}_a)$, where $\mathcal{X}$, $\mathcal{Z}$, $\mathbf{P}_a$ and $\mathbf{O}_a$ are as defined in $\mathcal{M}$. Further suppose that the initial distribution is given by $\mu_0 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$. Consider the sequence of observations $\boldsymbol{z}_{1:2} = \{u, u\}$. Which of the following is true?

- ☐ $\alpha_0 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$.
- ☐ $\alpha_1 = \begin{bmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{3} \end{bmatrix}$.
- ☐ $\alpha_2 = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$.
- ☒ **All of the above.**

(d) **(0.8 pts.)** Consider once again the HMM $(\mathcal{X}, Z, \mathbf{P}_a, \mathbf{O}_a)$ with initial distribution $\mu_0 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$. Then, given some sequence of observations $\boldsymbol{z}_{1:100}$, ...

- ☒ **... $\mu_{t|0:t} = \mu_0 \mathbf{P}_a^t$ for all $t = 0, \ldots, 100$.**
- ☐ ... $\mu_{t|0:t} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$ for all $t = 0, \ldots, 100$.
- ☐ ... $\mu_{t|0:t} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$ for all $t = 0, \ldots, 100$.
- ☐ ... there is not enough information to compute $\mu_{t|0:t}$ for $t > 0$ ($\boldsymbol{z}_{1:100}$ is necessary).

(e) **(0.8 pts.)** Let X and Y denote the two possible outcomes of a process of decision-making under uncertainty, and suppose that $Y \succ X$ and $u(X) = 0$. Consider also the following decision tree.



Which value for $u(Y)$ ensures that the decision-maker will select action A?

☐ $u(Y) = 1$.

☐ $u(Y) = 0.18$.

☐ Any value of $u(Y) > 0$ will lead the decision maker to choose action A.

☒ **None of the above.**

(f) **(0.8 pts.)** Consider the MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$ obtained from $\mathcal{M}$ by ignoring partial observability, and suppose that after K value iteration steps, we get the Q-function

$$\boldsymbol{Q} = \begin{bmatrix} 1.16 & 1.06 & 0.96 \\ 0.95 & 1.15 & 1.05 \\ 1.95 & 1.15 & 1.95 \end{bmatrix}.$$

The greedy policy associated with $\boldsymbol{Q}$ is...

☐ $\pi = \begin{bmatrix} 0.96 \\ 0.95 \\ 1.15 \end{bmatrix}$.

☐ $\pi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \end{bmatrix}$.

☒ $\pi = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$.

☐ $\pi = \begin{bmatrix} 1.16 \\ 1.15 \\ 1.95 \end{bmatrix}$

(g) **(0.8 pts.)** Consider once again the MDP $(\mathcal{X}, \mathcal{A}, \{\mathbf{P}_a\}, c, \gamma)$, and suppose that the cost-to-go associated with the policy π from (f) is given by

$$\boldsymbol{J}^{\pi} = \begin{bmatrix} 0.96 \\ 0.95 \\ 1.15 \end{bmatrix}.$$

What can you conclude about the policy π?

- ☐ The policy π is a good exploratory policy.
- ☒ **The policy π is optimal.**
- ☐ The policy π is not optimal.
- ☐ None of the above.

(h) **(0.8 pts.)** Suppose you want to solve a given MDP using policy iteration. Then, ...

- ☐ ... you will take less iterations but more time than if you use value iteration.
- ☐ ... you will take more iterations but less time than if you use value iteration.
- ☐ ... the number of iterations will be exponential in the number of states of the MDP.
- ☒ **... depending on the problem, you may take more or less time than if you use value iteration.**

(i) **(0.8 pts.)** Let $\mathcal{A}$ denote the action space in $\mathcal{M}$ and suppose that a decision-maker sequentially chooses actions from $\mathcal{A}$ using the exponentially weighted averager algorithm with $\eta = 1$. At time step $t = 1$ it chooses action $a_1 = a$ and observes the costs $c_1 = \begin{bmatrix} 1 & 0 & 0.1 \end{bmatrix}$. Knowing that all weights are initialized to 1, after performing the corresponding update, the resulting weights are given by:

- ☐ $w = \begin{bmatrix} 2.72 & 1.0 & 1.11 \end{bmatrix}$.
- ☐ $w = \begin{bmatrix} 2.72 & 1.0 & 1.0 \end{bmatrix}$.
- ☐ $w = \begin{bmatrix} 0.56 & 0.21 & 0.23 \end{bmatrix}$.
- ☒ **None of the above.**

(j) **(0.8 pts.)** After running the exponentially weighted averager for a few steps, the resulting weights are $w = \begin{bmatrix} 12.18 & 7.39 & 6.69 \end{bmatrix}$. The action selection probabilities at the next time step are:

- ☐ $p = \begin{bmatrix} 1.0 & 0.0 & 0.0 \end{bmatrix}$.
- ☒ $p = \begin{bmatrix} 0.46 & 0.28 & 0.26 \end{bmatrix}$.
- ☐ $p = \begin{bmatrix} 0.99 & 0.01 & 0.00 \end{bmatrix}$.
- ☐ $p = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$.

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

Consider the POMDP $\mathcal{M}$ provided in the shaded box. Suppose that the belief at time step $t = 0$ is given by $\boldsymbol{b}_0 = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$. The agent takes action $a_0 = b$ and observes $z_1 = v$.

(a) **(1.5 pts.)** Compute the updated belief $\boldsymbol{b}_1$.

(b) **(1.5 pt.)** Suppose now that the agent is following the FIB heuristic, with

$$\boldsymbol{Q}_{\text{FIB}} = \begin{bmatrix} 1.0 & 0.9 & 0.8 \\ 0.8 & 1.0 & 0.9 \\ 1.8 & 1.0 & 1.8 \end{bmatrix}.$$

Compute the action at time step $t = 1$. **Note:** If you did not solve (a), use $\boldsymbol{b}_1 = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$.

Solution 3.

(a) The updated belief is given by

$$\boldsymbol{b}_1 = \rho \boldsymbol{b}_0 \mathbf{P}_b \operatorname{diag}(\mathbf{O}_b(v \mid \cdot)),$$

where ρ is a normalization constant. We thus get:

$$\boldsymbol{b}_1 = \rho \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 & 0.0 \\ 0.5 & 0.0 & 0.5 \\ 0.5 & 0.0 & 0.5 \end{bmatrix} \begin{bmatrix} 0.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.5 \end{bmatrix} = \rho \begin{bmatrix} 0 & \frac{1}{6} & \frac{1}{6} \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

(b) The action selected by the FIB heuristic is given by

$$a_1 = \operatorname*{argmin}_{a \in \mathcal{A}} \boldsymbol{b}_2 \boldsymbol{Q}_{\text{FIB}} = \operatorname*{argmin}_{a \in \mathcal{A}} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 1.0 & 0.9 & 0.8 \\ 0.8 & 1.0 & 0.9 \\ 1.8 & 1.0 & 1.8 \end{bmatrix} = \operatorname*{argmin}_{a \in \mathcal{A}} \begin{bmatrix} 1.3 & 0.95 & 1.3 \end{bmatrix} = b.$$

Question 4. (**5 pts.**)

Consider the MDP obtained from $\mathcal{M}$ by ignoring partial observability, and suppose that we want to use a policy gradient RL algorithm to solve the MDP. Let $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ denote a scalar feature of (x, a) and consider the parameterized family of policies

$$\pi_\theta(a \mid x) = \frac{\exp(-\phi(x, a)\theta)}{\sum_{a' \in \mathcal{A}} \exp(-\phi(x, a')\theta)},$$

where θ is a scalar parameter.

(a) **(2.0 pts.)** Show that

$$\nabla \log \pi_\theta(a \mid x) = \sum_{a' \in \mathcal{A}} \pi_\theta(a' \mid x)\phi(x, a') - \phi(x, a).$$

(b) **(2 pts.)** Suppose that $\phi(x, a) = c(x, a)$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$. Further assume that θ is initialized to 0 and that the agent observes a trajectory $\tau = \{x_0, a_0, c_0, x_1, \ldots, a_9, c_9, x_{10}\}$, with $x_0 = A$, $a_0 = a$, and

$$\sum_{t=0}^{9} \gamma^t c_t = 2.3.$$

Compute the value of θ after the first update of REINFORCE. Use $\alpha = 0.1$.

Note: Recall that, given a T-step trajectory $\tau = \{x_0, a_0, c_0, \ldots, x_T\}$, the REINFORCE update rule, in its simplest form, is given by

$$\theta \leftarrow \theta - \alpha \nabla \log \pi_\theta(a_0 \mid x_0) \sum_{t=0}^{T} \gamma^t c_t.$$

(c) **(1.0 pts.)** Identify an advantage of policy gradient methods over value-based methods in large scale problems. Explain your reasoning.

Solution 4.

(a) We have that

$$\log \pi_\theta(a \mid x) = -\phi(x,a)\theta - \log \sum_{a' \in \mathcal{A}} \exp(-\phi(x,a')\theta),$$

yielding

$$\begin{aligned}
\nabla \log \pi_\theta(a \mid x) &= -\phi(x,a) - \frac{\nabla \sum_{a' \in \mathcal{A}} \exp(-\phi(x,a')\theta)}{\sum_{a' \in \mathcal{A}} \exp(-\phi(x,a')\theta)} \\
&= -\phi(x,a) + \frac{\sum_{a' \in \mathcal{A}} \phi(x,a') \exp(-\phi(x,a')\theta)}{\sum_{a' \in \mathcal{A}} \exp(-\phi(x,a')\theta)} \\
&= -\phi(x,a) + \sum_{a' \in \mathcal{A}} \phi(x,a') \pi_\theta(a' \mid x).
\end{aligned}$$

The conclusion follows.

(b) We have that

$$\begin{aligned}
\theta &\leftarrow 0 - 0.1 \times \left(\frac{1}{3}(c(A,a) + c(A,b) + c(A,c)) - c(A,a) \right) \times 2.3 \\
&= 0 - 0.1 \times (0.2 - 0.3) \times 2.3 = 0.023.
\end{aligned}$$

(c) Policy gradient methods can naturally be used in problems with continuous actions (for example). In such problems, value-based methods such as Q-learning require the computation of a maximizing action, which is, in itself, a complex optimization problem. Policy gradient methods do not require the computation of such maximum.

Question 5. (**1.5 pts.**)

In inverse reinforcement learning, an agent must estimate the cost function c given a policy π from an expert (the "teacher"). Show that, in order for a cost function c to render a policy π optimal, it must hold that

$$(\mathbf{P}_a - \mathbf{P}_\pi)(\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1} \boldsymbol{c} \geq 0, \qquad \text{for all } a \in \mathcal{A}.$$

Assume that the cost function depends only on the states.

Hint: Note that, if π is optimal, $Q^\pi(x,a) \geq J^\pi(x)$ for all $a \in \mathcal{A}$ and all $x \in \mathcal{X}$.

Solution 5.

We know that, if π is optimal, then

$$Q^{\pi}(x, a) \geq J^{\pi}(x)$$

for all states $x \in \mathcal{X}$ and all actions $a \in \mathcal{A}$. Equivalently,

$$c(x) + \gamma \sum_{x' \in \mathcal{X}} \mathbf{P}_a(x' \mid x) J^{\pi}(x) \geq c(x) + \gamma \sum_{x' \in \mathcal{X}} \mathbf{P}_{\pi}(x' \mid x) J^{\pi}(x),$$

where we used the fact that the cost only depends on the states. Simplifying the expression, we are left with

$$\sum_{x' \in \mathcal{X}} (\mathbf{P}_a(x' \mid x) - \mathbf{P}_{\pi}(x' \mid x)) J^{\pi}(x) \geq 0,$$

which can be written in vector form as

$$(\mathbf{P}_a - \mathbf{P}_{\pi}) \boldsymbol{J}^{\pi} \geq 0.$$

Since $J^{\pi} = (\mathrm{I} - \gamma \mathbf{P}_{\pi})^{-1} \boldsymbol{c}$, we finally get

$$(\mathbf{P}_a - \mathbf{P}_{\pi})(\mathrm{I} - \gamma \mathbf{P}_{\pi})^{-1} \boldsymbol{c} \geq 0.$$