TÉCNICO LISBOA

Learning and Decision Making 2015-2016

MSc in Computer Science and Engineering

First test – April 5, 2016

Instructions

- You have 90 minutes to complete the test.
- 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 all others if you want to be safe).
- The test has a total of 6 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 test.* 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 test 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.

1 Utility theory

Question 1. (**3 pts.**)
Adam, Brandon and Charles are University students that share an apartment. The three students take all house-related decisions democratically, i.e., when deciding between any two alternatives, they follow the majority vote.

a) **(1 pt.)** Knowing that, when deciding between any two brands of breakfast cereals,

- Adam prefers brand A to B and B to C;
- Brandon prefers brand B to C and C to A;
- Charles prefers C to A and A to B,

does their choice correspond to a valid preference relation? If not, briefly explain why.

b) **(1 pt.)** Knowing that, when having lunch together,

- Adam prefers to order sushi (S) than hamburgers (H) and the latter to eating microwave mac & cheese (M);
- Brandon prefers H to S and any of the two to M;
- Charles also prefers S to the other alternatives, but since he does not like meat, he prefers M to H,

does their choice correspond to a valid preference relation? If not, briefly explain why.

c) **(1 pt.)** When buying the laundry detergent, the three students typically go to a big department store, where they can select among all existing brands, and usually take brand X. When they are pressed with work, however, they buy their laundry detergent from the local supermarket that sells only brands X and Y, and usually take brand Y.

Do the detergent choices correspond to a valid preference relation? If not, briefly explain why, indicating the preferene axiom that is violated.

Solution 1.

a) Consider all possible pairings between the three alternatives. When comparing A and B, Adam and Charles vote for A, which implies that, as a group, $A \succ B$. When comparing B and C, Adam and Brandon vote for B, which means that, as a group, $B \succ C$. Finally, comparing A and C, Brandon and Charles vote for C, meaning that $C \succ A$. This violates the second axiom of preferences, which renders the relation an invalid preference.

b) Once again, comparing S to H Adam and Charles vote for S, implying $S \succ H$. Comparing H with M, Adam and Brandon vote H, implying that $H \succ M$. Finally, comparing S and M, all three vote S, meaning that $S \succ M$. This is a valid relation, where $S \succ H \succ M$.

c) The detergent choice violates the first axiom. When selecting among all brands, the students select brand X which means, in particular, that $X \succ Y$. On the other hand, when choosing between X and Y alone, they choose Y, which means that $Y \succ X$.

Question 2. (**2 pts.**)

Consider the situation faced by a student who, last night, had to decide whether to study for the ADI test or go out with her friends. If she does not study, there is a 0.25 probability that she will pass the test, and a 0.75 probability that she will fail. On the other hand, if she studies there is a 0.8 probability that she will pass the test and a 0.2 probability that she will fail.

When going out, she usually has fun, this happening with a probability of 0.8. On the other hand, when studying, she sometimes has fun learning about a topic that is really exciting, this happening with a probability of 0.4. Knowing that her preference over possible outcomes can be expressed by the utility function:

$$u(\text{Pass; have fun}) = 2.5; \qquad u(\text{Fail; have fun}) = 0.25;$$
$$u(\text{Pass; not have fun}) = 1.0; \qquad u(\text{Fail; not have fun}) = 0.0,$$

indicate what should be the student's choice according to the expected utility theory.

Solution 2.

To determine the choice according to the expected utility theory, we compute the expected utility associated with each option. Let P = Pass, $\bar{P}$ = Fail, F = Have fun and $\bar{F}$ = Not have fun. Let also S = Study and G = Go out. We have

$$\begin{aligned}
\mathbb{P}\left[PF \mid S\right] &= \mathbb{P}\left[P \mid S\right] \times \mathbb{P}\left[F \mid S\right] = 0.8 \times 0.4 = 0.32 \\
\mathbb{P}\left[\bar{P}F \mid S\right] &= \mathbb{P}\left[\bar{P} \mid S\right] \times \mathbb{P}\left[F \mid S\right] = 0.2 \times 0.4 = 0.08 \\
\mathbb{P}\left[P\bar{F} \mid S\right] &= \mathbb{P}\left[P \mid S\right] \times \mathbb{P}\left[\bar{F} \mid S\right] = 0.8 \times 0.6 = 0.48 \\
\mathbb{P}\left[\bar{P}\bar{F} \mid S\right] &= \mathbb{P}\left[\bar{P} \mid S\right] \times \mathbb{P}\left[\bar{F} \mid S\right] = 0.2 \times 0.6 = 0.12.
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{P}\left[PF \mid G\right] &= \mathbb{P}\left[P \mid G\right] \times \mathbb{P}\left[F \mid G\right] = 0.25 \times 0.8 = 0.2 \\
\mathbb{P}\left[\bar{P}F \mid G\right] &= \mathbb{P}\left[\bar{P} \mid G\right] \times \mathbb{P}\left[F \mid G\right] = 0.75 \times 0.8 = 0.6 \\
\mathbb{P}\left[P\bar{F} \mid G\right] &= \mathbb{P}\left[P \mid G\right] \times \mathbb{P}\left[\bar{F} \mid G\right] = 0.25 \times 0.2 = 0.05 \\
\mathbb{P}\left[\bar{P}\bar{F} \mid G\right] &= \mathbb{P}\left[\bar{P} \mid G\right] \times \mathbb{P}\left[\bar{F} \mid G\right] = 0.75 \times 0.2 = 0.15.
\end{aligned}$$

We thus have:

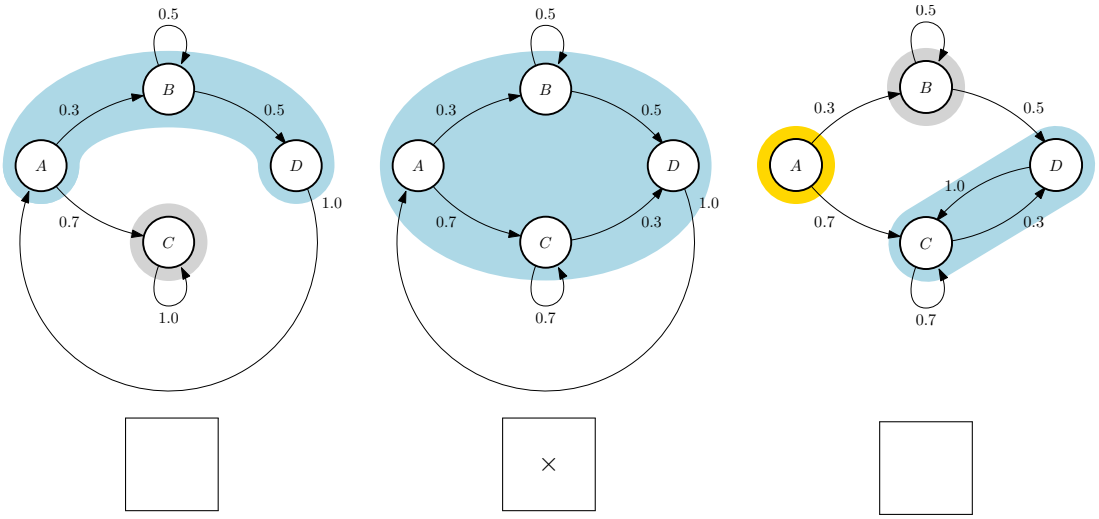
$$\begin{aligned}
Q(S) &= .32 \times 2.5 + .08 \times 0.25 + 0.48 \times 1 + 0.12 \times 0 = 1.3 \\
Q(G) &= .2 \times 2.5 + .6 \times 0.25 + 0.05 \times 1 + 0.15 \times 0 = 0.7.
\end{aligned}$$

The student should choose to stay home and study.

2 Sequential models

Question 3. (**5 pts.**)

Consider the Markov chains represented in the transition diagrams below.



a) **(2 pt.)** *For each of the chains*, identify the state space and write down the transition probability matrices.

b) **(1 pt.)** *Indicate in the figure* the communicating classes for each of the chains, and identify with an "×" which one is irreducible. Briefly justify your option.

c) **(2 pt.)** Consider once again the chain identified in b). Which of the alternatives below corresponds to the invariant distribution for that chain? Briefly justify your selection, indicating the *relevant* computations.

a) $\mu = \big[1.0000, \quad 1.0000, \quad 1.0000, \quad 1.0000\big]$;

b) $\mu = \big[0.2027, \quad 0.1216, \quad 0.4730, \quad 0.2027\big]$;

c) $\mu = \big[0.3580, \quad 0.2148, \quad 0.8352, \quad 0.3580\big]$;

d) $\mu = \big[0.2500, \quad 0.2500, \quad 0.2500, \quad 0.2500\big]$;

Solution 3.

a) For all chains, $\mathcal{X} = \{A, B, C, D\}$. As for the transition probability matrices, we have:

$$\boldsymbol{P}_1 = \begin{bmatrix} 0.0 & 0.3 & 0.7 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 \\ 0.0 & 0.0 & 1.0 & 0.0 \\ 1.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \quad \boldsymbol{P}_2 = \begin{bmatrix} 0.0 & 0.3 & 0.7 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 \\ 0.0 & 0.0 & 0.7 & 0.3 \\ 1.0 & 0.0 & 0.0 & 0.0 \end{bmatrix} \quad \boldsymbol{P}_3 = \begin{bmatrix} 0.0 & 0.3 & 0.7 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 \\ 0.0 & 0.0 & 0.7 & 0.3 \\ 0.0 & 0.0 & 1.0 & 0.0 \end{bmatrix}$$

d) The correct option is (b). Options (a) and (c) are not distributions (they do not add to 1). Additionally, for μ to be an invariant distribution, $\mu \boldsymbol{P} = \mu$. For option (d),

$$\mu \boldsymbol{P} = \begin{bmatrix} 0.25 & 0.25 & 0.25 & 0.25 \end{bmatrix} \begin{bmatrix} 0.0 & 0.3 & 0.7 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 \\ 0.0 & 0.0 & 0.7 & 0.3 \\ 1.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.25 & 0.2 & 0.35 & 0.2 \end{bmatrix}.$$

On the other hand, for option (b),

$$\mu \boldsymbol{P} = \begin{bmatrix} 0.2027 & 0.1216 & 0.4730 & 0.2027 \end{bmatrix} \begin{bmatrix} 0.0 & 0.3 & 0.7 & 0.0 \\ 0.0 & 0.5 & 0.0 & 0.5 \\ 0.0 & 0.0 & 0.7 & 0.3 \\ 1.0 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2027 & 0.1216 & 0.4730 & 0.2027 \end{bmatrix}.$$

Question 4. (**3 pts.**)

Suppose that you want to describe the occupation status of a parking space in a shopping mall, where you can observe only the corresponding overhead light signal (red for occupied and green for vacant). You use an HMM $(\mathcal{X}, \mathcal{Z}, \boldsymbol{P}, \boldsymbol{O})$ to model the dynamics of the process, with

- $\mathcal{X} = \{\text{Vacant}, \text{Occupied}\}$.
- $\mathcal{Z} = \{\text{Green}, \text{Red}\}$.
- The transition and observation probabilities are summarized as the matrices

$$\boldsymbol{P} = \begin{bmatrix} 0.2 & 0.8 \\ 0.5 & 0.5 \end{bmatrix} \qquad \boldsymbol{O} = \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix},$$

  where the states and observations are ordered as listed in $\mathcal{X}$ and $\mathcal{Z}$.

Knowing that the parking space was initially vacant, use the forward-backward algorithm to determine the probability that the state at time step $t = 2$ is occupied given the observation sequence $\{\text{Green}, \text{Green}\}$.

Solution 4.

We want to compute $\mathbb{P}\left[X(2) = \text{Occupied} \mid \boldsymbol{Z}_{1:2} = \{\text{Green, Green}\}\right]$. This probability corresponds to the quantity

$$\gamma_2(\text{Occupied}) = \frac{\alpha_2(\text{Occupied})\beta_2(\text{Occupied})}{\boldsymbol{\alpha}_2^\top \boldsymbol{\beta}_2}.$$

However, since we have only 2 observations, $\boldsymbol{\beta}_2 = \mathbf{1}$, and

$$\gamma_2(\text{Occupied}) = \frac{\alpha_2(\text{Occupied})}{\alpha_2(\text{Vacant}) + \alpha_2(\text{Occupied})}.$$

Therefore, the forward-backward algorithm reduces to the forward computation, and we have:

$$\boldsymbol{\alpha}_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top \quad \text{(there is no initial observation)},$$

$$\boldsymbol{\alpha}_1 = \text{diag}(\boldsymbol{O}_{\text{Green}})\boldsymbol{P}^\top \boldsymbol{\alpha}_0^\top = \begin{bmatrix} 0.15 & 0.20 \end{bmatrix}^\top,$$

$$\boldsymbol{\alpha}_2 = \text{diag}(\boldsymbol{O}_{\text{Green}})\boldsymbol{P}^\top \boldsymbol{\alpha}_1^\top = \begin{bmatrix} 0.0975 & 0.055 \end{bmatrix}^\top,$$

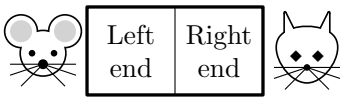
and we finally get

$$\gamma_2(\text{Occupied}) = \frac{0.055}{0.0975 + 0.055} = 0.36.$$

3 Markov decision problems

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

Consider the situation of a mouse trying to escape a cat in a long corridor. Both cat and mouse can be in either of the two ends of the corridor.



At each time step t, the mouse can select between two actions:

- *Move to the other end of the corridor*: If cat and mouse are in the same end of the corridor, such action will succeed with a probability 0.4 and fail with a probability 0.6. If the cat and mouse are in opposite ends of the corridor, the action succeeds with probability 1.
- *Stay still*: This action always succeeds.

The cat always moves in the same way, independently of the action of the mouse: if cat and mouse are in the same end of the corridor, the cat will remain still with probability 1. If not, the cat will move with probability 0.75 and stay still with a probability 0.25.

When cat and mouse are in the same end of the corridor, the mouse suffers some damage (corresponding to a cost of 1). When standing in different ends of the corridor, the mouse suffers no damage.

a) **(1 pt.)** Write down the MDP model for this problem, considering a discount $\gamma = 0.9$. **Suggestion:** It is possible to model this problem with only *two* states. You should try to adopt such representation, as it will facilitate the upcoming computations.

b) **(3 pt.)** Run 2 steps of value iteration to compute V^*, using as initial iterate $\boldsymbol{V}^{(0)} = \mathbf{0}$.

c) **(3 pt.)** Consider the policy in which the mouse always selects the action *Move*. Compute the cost-to-go associated with this policy and indicate whether the policy is optimal, supporting your conclusions with any relevant computations. **Note:** Recall that the inverse of a 2×2 matrix can be computed as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Solution 5.

a) There are two possible answers for this question. The first considers all possible positions for cat and mouse. Letting L denote the left end of the corridor and R the right end, we can represent the position of the two animals as a pair $(p_{\text{mouse}}, p_{\text{cat}})$. Denoting the two possible mouse actions by M (move) and S (stay still), we get

$$\mathcal{X} = \{(L,L), (L,R), (R,L), (R,R)\};$$
$$\mathcal{A} = \{M, S\};$$
$$\boldsymbol{P}_M = \begin{bmatrix} 0.6 & 0 & 0.4 & 0 \\ 0 & 0 & 0.75 & 0.25 \\ 0.25 & 0.75 & 0 & 0 \\ 0 & 0.4 & 0 & 0.6 \end{bmatrix}, \qquad \boldsymbol{P}_S = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0.75 & 0.25 & 0 & 0 \\ 0 & 0 & 0.25 & 0.75 \\ 0 & 0 & 0 & 1 \end{bmatrix};$$
$$\boldsymbol{C} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

A second solution comes from considering that the only relevant information is whether cat and mouse are co-located (C) or not ($\bar{C}$). This yields the MDP

$$\mathcal{X} = \{C, \bar{C}\};$$
$$\mathcal{A} = \{M, S\};$$
$$\boldsymbol{P}_M = \begin{bmatrix} 0.6 & 0.4 \\ 0.25 & 0.75 \end{bmatrix}, \qquad \boldsymbol{P}_S = \begin{bmatrix} 1 & 0 \\ 0.75 & 0.25 \end{bmatrix};$$
$$\boldsymbol{C} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

b) We use the 2-state version. The VI update is given by

$$\boldsymbol{J}^{(n+1)} = \min_{a \in \mathcal{A}} \left\{ \boldsymbol{C}_{:,a} + \gamma \boldsymbol{P}_a \boldsymbol{J}^{(n)} \right\},$$

where the max is taken component-wise. Since our reward is action independent, the expression above simplifies to:

$$\boldsymbol{J}^{(n+1)} = \boldsymbol{c} + \gamma \min_{a \in \mathcal{A}} \left\{ \boldsymbol{P}_a \boldsymbol{J}^{(n)} \right\},$$

We have

$$\begin{aligned}
\boldsymbol{J}^{(1)} &= \boldsymbol{c} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
\boldsymbol{J}^{(2)} &= \boldsymbol{c} + 0.9 \times \min \left\{ \begin{bmatrix} 0.6 & 0.4 \\ 0.25 & 0.75 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0.75 & 0.25 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\} \\
&= \boldsymbol{c} + 0.9 \times \min \left\{ \begin{bmatrix} 0.6 \\ 0.25 \end{bmatrix}, \begin{bmatrix} 1 \\ 0.75 \end{bmatrix} \right\} \\
&= \begin{bmatrix} 1.6 \\ 0.25 \end{bmatrix}.
\end{aligned}$$

c) Associated with the provided policy we have

$$\boldsymbol{P}_\pi = \begin{bmatrix} 0.6 & 0.4 \\ 0.25 & 0.75 \end{bmatrix},$$

and we can compute

$$\begin{aligned}
\boldsymbol{J}^\pi &= (\mathbf{I} - \gamma \boldsymbol{P}_\pi)^{-1} \boldsymbol{c} \\
&= \begin{bmatrix} 0.46 & -0.36 \\ -0.225 & 0.325 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
&= 14.6 \times \begin{bmatrix} 0.325 & 0.36 \\ 0.225 & 0.46 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} 4.745 \\ 3.285 \end{bmatrix}
\end{aligned}$$

Let us compute a VI update from J^π. We have:

$$\begin{aligned}
\boldsymbol{J}^{\text{new}} &= \boldsymbol{c} + 0.9 \times \min \left\{ \begin{bmatrix} 0.6 & 0.4 \\ 0.25 & 0.75 \end{bmatrix} \begin{bmatrix} 4.745 \\ 3.285 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0.75 & 0.25 \end{bmatrix} \begin{bmatrix} 4.745 \\ 3.285 \end{bmatrix} \right\} \\
&= \boldsymbol{c} + 0.9 \times \min \left\{ \begin{bmatrix} 4.16 \\ 3.65 \end{bmatrix}, \begin{bmatrix} 4.745 \\ 4.38 \end{bmatrix} \right\} \\
&= \begin{bmatrix} 4.745 \\ 3.285 \end{bmatrix}
\end{aligned}$$

and we can conclude that the policy is optimal.

Question 6. (**3 pts.**)

a) **(2 pt.)** Let $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \boldsymbol{P}, r, \gamma)$ denote a finite MDP. A *potential function* is any real-valued function $\phi : \mathcal{X} \to \mathbb{R}$. The optimal value function for $\mathcal{M}$ is an example of a potential function. Given a potential function ϕ, the function $f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ defined as

$$f(x, a) = \phi(x) - \gamma \sum_{y \in \mathcal{X}} \boldsymbol{P}(y \mid x, a)\phi(y)$$

is called a *potential-based shaping function*. Let ϕ be an arbitrary potential function and f its associated potential-based shaping function. Letting $c'(x, a) = c(x, a) + f(x, a)$, show that the optimal policies for $\mathcal{M}$ and for the modified MDP $\mathcal{M}' = (\mathcal{X}, \mathcal{A}, \boldsymbol{P}, c', \gamma)$ are the same.

b) **(1 pt.)** Briefly explain the statement (no math needed): "*In a POMDP the belief state is a sufficient statistic for the history.*"

Solution 6.

a) Let Q and Q' denote the optimal Q-functions associated with the original cost $c(x, a)$ and the modified cost $c'(x, a) = r(x, a) + f(x, a)$, respectively. Then

$$\begin{aligned}
Q'(x, a) &= c'(x, a) + \gamma \sum_{y \in \mathcal{X}} \boldsymbol{P}(y \mid x, a) \min_{b \in \mathcal{A}} Q'(y, b) \\
&= c(x, a) + \phi(x) - \gamma \sum_{y \in \mathcal{X}} \boldsymbol{P}(y \mid x, a)\phi(y) + \gamma \sum_{y \in \mathcal{X}} \boldsymbol{P}(y \mid x, a) \min_{b \in \mathcal{A}} Q'(y, b) \\
&= c(x, a) + \phi(x) + \gamma \sum_{y \in \mathcal{X}} \boldsymbol{P}(y \mid x, a) \left[\min_{b \in \mathcal{A}} Q'(y, b) - \phi(x) \right].
\end{aligned}$$

Since ϕ does not depend on the action, the above yields,

$$Q'(x, a) = c(x, a) + \phi(x) + \gamma \sum_{y \in \mathcal{X}} \boldsymbol{P}(y \mid x, a) \min_{b \in \mathcal{A}} \left[Q'(y, b) - \phi(x) \right]$$

and, finally,

$$Q'(x, a) - \phi(x) = c(x, a) + \gamma \sum_{y \in \mathcal{X}} \boldsymbol{P}(y \mid x, a) \min_{b \in \mathcal{A}} \left[Q'(y, b) - \phi(x) \right],$$

which implies that $Q(x,a) + \phi(x) = Q'(x,a)$. But then,

$$\operatorname*{argmin}_{a \in \mathcal{A}} Q'(x,a) = \operatorname*{argmin}_{a \in \mathcal{A}} \{Q(x,a) + \phi(x)\} = \operatorname*{argmin}_{a \in \mathcal{A}} Q(x,a).$$

b) It was shown that the distribution over states in a POMDP at any time-step t,

- can be computed from the history up to time step $t-1$ and the most recent events—namely the action a_{t-1} and the observation z_t.
- can equivalently be computed from the belief $\mathbf{b}_{t-1}$, the action a_{t-1} and the observation z_t.

This means that the belief $\mathbf{b}_{t-1}$ contains all the information contained in the history h_{t-1}, i.e., it is a sufficient statistic for the history.