INFO8006 Introduction to Artificial Intelligence

Making Decisions

Learning outcomes

At the end of this exercise session, you should be able to:

- Define a Markov Decision Process (MDP).
- Define what is the reward, the Expected Utility, an Optimal Policy.
- Define the Bellman equations of a simple MDP and apply value iteration.
- Define and apply Policy iteration.
- Define a Partially Observable Markov Decision Process.
- Explain the difference between Model-based and Model-free reinforcement algorithms.

Exercise 1: Micro-Blackjack[1]

In Micro-Blackjack, you repeatedly draw a card (with replacement) that is equally likely to be a 2, 3, or 4. You can either Draw or Stop if the total score of the cards you have drawn is less than 6. Otherwise, you must Stop. When you Stop, your utility is equal to your total score (up to 5), or zero if you get a total of 6 or higher. When you Draw, you receive no utility. There is no discount ($\gamma = 1$).

1. Formalise Micro-Blackjack as a MDP.

   - The set of states: $\mathcal{S} = \{0, 2, 3, 4, 5, 6^+\} \times \{\text{Playing}, \text{Done}\}$. It represents the current sum of cards and if the player is still playing.
   - The set of actions: $\mathcal{A} = \{\text{Draw}, \text{Stop}\}$.
   - The transition model: Let $s = (s_1, s_2)$ and $s' = (s'_1, s'_2)$.
     If s_2 is equal to Done then:

     $$P(s'|s, a) = \begin{cases} 1, & \text{if } s'_1 = s_1,\ s'_2 = \text{Done and whatever a.} \\ 0, & \text{Otherwise} \end{cases} \tag{1}$$

     .
     Else if s_2 is equal to Playing then:

     $$P(s'|s, \text{Draw}) = \begin{cases} \text{if } s_1 \in \{4, 5\} \text{ and } s'_1 = 6^+ \text{ and } s'_2 = \text{Playing}: & 1, \\ \text{else if } s_1 = 3 \text{ and } s' = 6^+ \text{ and } s'_2 = \text{Playing}: & \frac{2}{3}, \\ \text{else if } s'_1 - s_1 \in \{2, 3, 4\} \text{ and } s'_2 = \text{Playing}: & \frac{1}{3}, \\ \text{else}: & 0. \end{cases} \tag{2}$$

     $$P(s'|s, \text{Stop}) = \begin{cases} 1, & \text{If } s'_1 = s_1 \text{ and } s'_2 = \text{Done}. \\ 0, & \text{Otherwise}. \end{cases} \tag{3}$$

     .
   - The reward function:

     $$R(s) = \begin{cases} s_1, & \text{if } s_1 \in \{2, 3, 4, 5\} \text{ and } s_2 = \text{Done}. \\ 0, & \text{Otherwise}. \end{cases} \tag{4}$$

2. Derive the optimal policy for this MDP. The Policy iteration algorithm gives the sequence of policies of Table 1 if we initialize the policy by always doing the action Stop. $\pi_{i+1}(s) = \arg\max_{a \in \mathcal{A}} \sum_{s'} P(s'|s, a) V_i(s')$

s_2	Playing						Done					
s_1	0	2	3	4	5	6^+	0	2	3	4	5	6^+
$Action: V_0$	S:0	S:2	S:3	S:4	S:5	S:0	S:0	S:2	S:3	S:4	S:5	S:0
$Action: V_1$	D:3	D:3	S:3	S:4	S:5	S:0	S:0	S:2	S:3	S:4	S:5	S:0
$Action: V_2$	D:$\frac{10}{3}$	D:3	S:3	S:4	S:5	S:0	S:0	S:2	S:3	S:4	S:5	S:0

Table 1: Policy iteration applied to micro blackjack game.

[1] Source: `http://ai.berkeley.edu/sections/section_4_9vzQ4e7e6xqagFuQDinu0lmeSEI29Z.pdf`

Exercise 2: Pursuit Evasion[2]

Pacman is trapped in the following 2×2 maze with a hungry ghost (the horror)! When it is his turn to move, Pacman must move one step horizontally or vertically to a neighboring square. When it is the ghost's turn, he must also move one step horizontally or vertically. The ghost and Pacman alternate moves. After every move (by either the ghost or Pacman) if Pacman and the ghost occupy the same square, Pacman is eaten and receives utility -100. Otherwise, he receives a utility of 1. The ghost attempts to minimize the utility that Pacman receives. Assume the ghost makes the first move.
For example, with a discount factor of $\gamma = 1$, if the ghost moves down, then Pacman moves left, Pacman earns a reward of 1 after the ghost's move and -100 after his move for a total utility of -99.
Note that this game is not guaranteed to terminate.

1. Assume a discount factor $\gamma = 0.5$, where the discount factor is applied once every time either Pacman or the ghost moves. What is the minimax value of the truncated game after 2 ghost moves and 2 Pacman moves? (Hint: you should not need to build the minimax tree) If Pacman plays optimally he will avoid the ghost and so he will exactly do the opposite move of this latter. Then the score of pacman will be equal to: $\sum_{i=0}^{3} 1\gamma^i = 1 + 0.5 + 0.5^2 + 0.5^3 = 1.875$.

2. Assume a discount factor $\gamma = 0.5$. What is the minimax value of the complete (infinite) game? (Hint: you should not need to build the minimax tree) $\sum_{i=0}^{\infty} 1\gamma^i = \frac{1}{1-\gamma} = 2$

3. Why is value iteration superior to minimax for solving this game? Because value iteration converges to a value whereas the size of the minimax tree is infinite. Value iteration gives a policy whereas minimax only provides the next action.

4. This game is similar to a MDP because rewards are earned at every timestep. However, it is also an adversarial game involving decisions by two agents. Let s be the state (e.g. the position of Pacman and the ghost and who is playing), and let $A_P(s)$ be the space of actions available to Pacman in state s (and similarly let $A_G(s)$ be the space of actions available to the ghost). Let $N(s, a) = s'$ denote the successor function (given a starting state s, this function returns the state s' which results after taking action a). Finally, let $R(s)$ denote the utility received by pacman after moving (ghost or pacman) to state s. Write down an expression for $V^\star(s)$, the value of the game to Pacman as a function of the current state s (analogous to the Bellman equations). Hint: your answer should include $V^\star(s)$ on the right hand side. $V^\star(s) = \max_{a \in A_p(s)} [R(N(s,a)) + \gamma \min_{a' \in A_G(N(s,a))} R(N(N(s,a),a')) + \gamma V^\star(N(N(s,a),a'))]$

Exercise 3: Crying Baby Problem

To earn some money you have decided to do babysitting. Tonight you will watch a 9 months old child. His parents asked you to take care of feeding him when he is hungry. Because you are a super-babysitter you know from your experience that the probability of the baby crying when he is hungry is equal to 0.8 and to cry when he is not hungry is equal to 0.1. The cost of feeding the baby is equal to 5 euros whereas if you don't feed him while he is hungry it costs you 10 euros (because the parent will be upset and will reduce your pay). You will have many things to do tonight and so you decide to optimize your time by just checking every 10 minutes whether the baby is crying. Between two intervals of time you assume that the baby has a probability of 0.2 to get hungry. You can also assume that the baby has a uniform probability of being hungry when you start your babysitting.

1. Draw the dynamic Bayesian network (as well as the conditional probability tables) associated with your babysitting of tonight. Let $F_t \in \{\text{True, False}\}$ denote the fact that you feed the baby or not at time t, $H_t \in \{\text{True, False}\}$ the unknown fact that the baby is hungry or not at time t, $C_t \in \{\text{True, False}\}$ the fact that the baby is crying or not at time t and $R_t \in \{0, -5, -10\}$ your cost. The dynamic Bayesian network is presented by Figure 1 and the conditional probabilities are given by Tables 2, 3 and 4. For completeness we also have to define $P(H_0 = \text{True}) = 0.5$.

F_t	True		False	
H_t	True	False	True	False
$P(H_{t+1} = \text{True} \mid H_t, F_t)$	0	0	1	0.2

Table 2: CPT of H_{t+1}

H_t	True	False
$P(C_t = \text{True} \mid H_t)$	0.8	0.1

Table 3: CPT of C_t

[2] Source: `http://ai.berkeley.edu/sections/section_4_9vzQ4e7e6xqagFuQDinu0lmeSEI29Z.pdf`

F_t	True		False	
H_t	True	False	True	False
R_t	-5	-5	-10	0

Table 4: Degenerated (deterministic) CPT of R_t

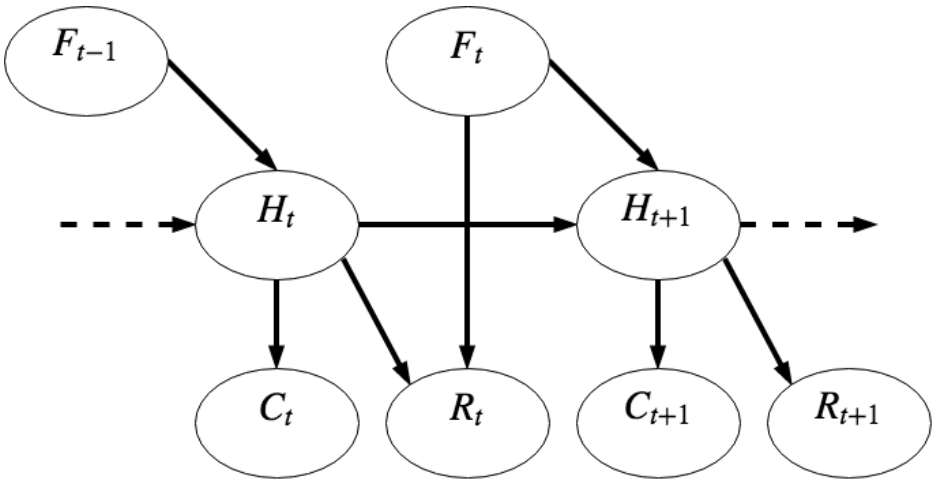


Figure 1: Dynamic Bayesian network of the crying baby problem.

2. Give the sequence of belief states if you assume a uniform prior and that the sequence of (action, observation) is: (No feed, cry), (Feed, no cry), (No feed, No cry), (No feed, No cry).

$$
\begin{aligned}
b_{H_0} &= P(H_0 = \text{True}) && (5)\\
&= 0.5 && (6)\\
b_{H_1} &= P(H_1 | F_0 = \text{False}, C_1 = \text{True}) && (7)\\
&= \sum_{H_0} P(H_0, H_1 | F_0 = \text{False}, C_1 = \text{True}) && (8)\\
&= \alpha \sum_{H_0} P(H_0, H_1, F_0 = \text{False}, C_1 = \text{True}) && (9)\\
&= \alpha \sum_{H_0} P(H_0) P(H_1 | H_0, F_0 = \text{False}) P(C_1 = \text{True} | H_1) && (10)\\
&= \frac{\alpha}{2} P(C_1 = \text{True} | H_1) \sum_{H_0} P(H_1 | H_0, F_0 = \text{False}) && (11)\\
&= \begin{cases} P(H_1 = \text{True} | F_0 = \text{False}, C_1 = \text{True}) & = \frac{\alpha}{2} P(C_1 = \text{True} | H_1 = \text{True}) \\ & \sum_{H_0} P(H_1 = \text{True} | H_0, F_0 = \text{False}) \\ & = \frac{\alpha}{2} 0.8(1 + 0.2) = 0.48\alpha \\ P(H_1 = \text{False} | F_0 = \text{False}, C_1 = \text{True}) & = \frac{\alpha}{2} P(C_1 = \text{True} | H_1 = \text{False}) \\ & \sum_{H_0} P(H_1 = \text{False} | H_0, F_0 = \text{False}) \\ & = \frac{\alpha}{2} 0.1(0 + 0.8) = 0.04 \end{cases} && (12)\\
&= \begin{cases} 0.923 \\ 0.077 \end{cases} && (13)\\
b_{H_2} &= \begin{cases} 0 \\ 1 \end{cases} && (14)\\
b_{H_3} &= \sum_{H_2} P(H_2, H_3 | F_2 = \text{False}, C_3 = \text{False}, b_{H_3}) && (15)\\
&= \alpha \sum_{H_2} P(H_2) P(H_3 | H_2, F_2 = \text{False}) P(C_3 = \text{False} | H_3) && (16)\\
&= \alpha P(C_3 = \text{False} | H_3) \sum_{H_2} P(H_2) P(H_3 | H_2, F_2 = \text{False}) && (17)\\
&= \begin{cases} \alpha \times 0.2(0 \times 1 + 1 \times 0.2) \\ \alpha \times 0.9(0 \times 0 + 1 \times 0.8) \end{cases} && (18)\\
&= \begin{cases} 0.052 \\ 0.948 \end{cases} && (19)\\
b_{H_4} &= \begin{cases} \alpha \times 0.2(0.052 \times 1 + 0.948 \times 0.2) \\ \alpha \times 0.9(0.052 \times 0 + 0.948 \times 0.8) \end{cases} && (20)\\
&= \begin{cases} 0.067 \\ 0.933 \end{cases} && (21)
\end{aligned}
$$

⋆ Exercise 4: AIMA 16.15

Consider a student who has the choice to buy or not buy a textbook for a course. We'll model this as a decision problem with one Boolean decision node, B, indicating whether the agent chooses to buy the book, and two Boolean chance nodes, M, indicating whether the student has mastered the material in the book, and P, indicating whether the student passes the course. Of course, there is also a utility node, U. A certain student, Sam, has an additive utility function: 0 for not buying the book and -$100 for buying it; and $2000 for passing the course and 0 for not passing. Sam's conditional probability estimates are as follows:

$$
\begin{array}{ll}
P(p|b,m) = 0.9 & P(m|b) = 0.9 \\
P(p|b,\neg m) = 0.5 & P(m|\neg b) = 0.7 \\
P(p|\neg b, m) = 0.8 & \\
P(p|\neg b, \neg m) = 0.3 &
\end{array}
$$

You might think that P would be independent of B given M, But this course has an open-book final—so having the book helps.

1. Draw the decision network for this problem.
2. Compute the expected utility of buying the book and of not buying it.
3. What should Sam do?

⋆ Exercise 5: AIMA 16.17

A used-car buyer can decide to carry out various tests with various costs (e.g., kick the tires, take the car to a qualified mechanic) and then, depending on the outcome of the tests, decide which car to buy. We will assume that the buyer is deciding whether to buy car c_1, that there is time to carry out at most one test, and that t_1 is the test of c_1 and costs \$50.

A car can be in good shape (quality q^+) or bad shape (quality q^-), and the tests might help indicate what shape the car is in. Car c_1 costs \$1,500, and its market value is \$2,000 if it is in good shape; if not, \$700 in repairs will be needed to make it in good shape. The buyer's estimate is that c_1 has a 70% chance of being in good shape.

1. Draw the decision network that represents this problem.
2. Calculate the expected net gain from buying c_1, given no test.
3. Tests can be described by the probability that the car will pass or fail the test given that the car is in good or bad shape. We have the following information:
   $P(pass(c_1, t_1)|q^+(c_1)) = 0.8$
   $P(pass(c_1, t_1)|q^-(c_1)) = 0.35$
   Use Bayes' theorem to calculate the probability that the car will pass (or fail) its test and hence the probability that it is in good (or bad) shape given each possible test outcome.
4. Calculate the optimal decisions given either a pass or a fail, and their expected utilities.
5. Calculate the value of information of the test, and derive an optimal conditional plan for the buyer.

⋆ Exercise 6: AIMA 17.6

The slides of the course states that the Bellman operator is a contraction.

1. Show that, for any functions f and g,
$$|\max_a f(a) - \max_a g(a)| \leq \max_a |f(a) - g(a)| \ .$$
2. Write out an expression for $|(B\,U_i - B\,U_i')(s)|$ and then apply the result from (a) to complete the proof that the Bellman operator is a contraction.

⋆ Exercise 7: AIMA 17.8

Consider the 3×3 world shown in Figure 2(a). The transition model is the same as in the 4×3 Figure 3: 80% of the time the agent goes in the direction it selects; the rest of the time it moves at right angles to the intended direction.

Implement value iteration for this world for each value of r below. Use discounted rewards with a discount factor of 0.99. Show the policy obtained in each case. Explain intuitively why the value of r leads to each policy.

1. $r = -100$
2. $r = -3$
3. $r = 0$
4. $r = +3$

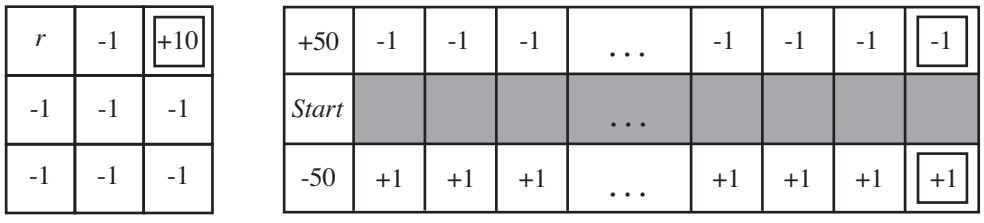
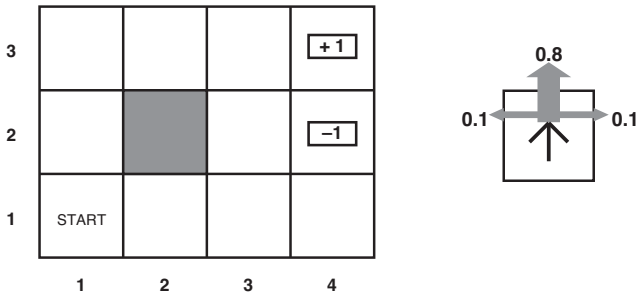


Figure 2: (a) 3×3 world for Exercise 17.8. The reward for each state is indicated. The upper right square is a terminal state. (b) 101×3 world for Exercise 17.9 (omitting 93 identical columns in the middle). The start state has reward 0.



(a) (b)

Figure 3: (a) A simple 4×3 environment that presents the agent with a sequential decision problem. (b) Illustration of the transition model of the environment: the "intended" outcome occurs with probability 0.8, but with probability 0.2 the agent moves at right angles to the intended direction. A collision with a wall results in no movement. The two terminal states have reward $+1$ and -1, respectively, and all other states have a reward of -0.04.

⋆ Exercise 8: AIMA 17.9

Consider the 101×3 world shown in Figure 2(b). In the start state the agent has a choice of two deterministic actions, *Up* or *Down*, but in the other states the agent has one deterministic action, *Right*. Assuming a discounted reward function, for what values of the discount γ should the agent choose *Up* and for which *Down*? Compute the utility of each action as a function of γ. (Note that this simple example actually reflects many real-world situations in which one must weigh the value of an immediate action versus the potential continual long-term consequences, such as choosing to dump pollutants into a lake.)

⋆ Exercise 9: AIMA 17.10

Consider an undiscounted MDP having three states, (1, 2, 3), with rewards -1, -2, 0, respectively. State 3 is a terminal state. In states 1 and 2 there are two possible actions: a and b. The transition model is as follows:

- In state 1, action a moves the agent to state 2 with probability 0.8 and makes the agent stay put with probability 0.2.
- In state 2, action a moves the agent to state 1 with probability 0.8 and makes the agent stay put with probability 0.2.
- In either state 1 or state 2, action b moves the agent to state 3 with probability 0.1 and makes the agent stay put with probability 0.9.

Answer the following questions:

1. What can be determined *qualitatively* about the optimal policy in states 1 and 2?
2. Apply policy iteration, showing each step in full, to determine the optimal policy and the values of states 1 and 2. Assume that the initial policy has action b in both states.
3. What happens to policy iteration if the initial policy has action a in both states? Does discounting help? Does the optimal policy depend on the discount factor?

⋆ Exercise 10: AIMA 17.19

A Dutch auction is similar in an English auction, but rather than starting the bidding at a low price and increasing, in a Dutch auction the seller starts at a high price and gradually lowers the price until some buyer is willing to accept that price. (If multiple bidders accept the price, one is arbitrarily chosen as the winner.) More formally, the seller begins with a price p and gradually lowers p by increments of d until at least one buyer accepts the price. Assuming all bidders act rationally, is it true that for arbitrarily small d, a Dutch auction will always result in the bidder with the highest value for the item obtaining the item? If so, show mathematically why. If not, explain how it may be possible for the bidder with highest value for the item not to obtain it.

⋆ Exercise 11: AIMA 17.21

Teams in the National Hockey League historically received 2 points for winning a game and 0 for losing. If the game is tied, an overtime period is played; if nobody wins in overtime, the game is a tie and each team gets 1 point. But league officials felt that teams were playing too conservatively in overtime (to avoid a loss), and it would be more exciting if overtime produced a winner. So in 1999 the officials experimented in mechanism design: the rules were changed, giving a team that loses in overtime 1 point, not 0. It is still 2 points for a win and 1 for a tie.

1. Was hockey a zero-sum game before the rule change? After?

2. Suppose that at a certain time t in a game, the home team has probability p of winning in regulation time, probability $0.78 - p$ of losing, and probability 0.22 of going into overtime, where they have probability q of winning, $.9 - q$ of losing, and .1 of tying. Give equations for the expected value for the home and visiting teams.

3. Imagine that it were legal and ethical for the two teams to enter into a pact where they agree that they will skate to a tie in regulation time, and then both try in earnest to win in overtime. Under what conditions, in terms of p and q, would it be rational for both teams to agree to this pact?

4. Some experts report that since the rule change, the percentage of games with a winner in overtime went up 18.2%, as desired, but the percentage of overtime games also went up 3.6%. What does that suggest about possible collusion or conservative play after the rule change?