Decision making, Markov decision processes

Solved tasks

Collected by: Jiří Kléma, klema@fel.cvut.cz

Spring 2020

The main goal:

The text presents solved tasks to support labs in the B4B36ZUI course.

1 Simple decisions, Bayesian decision making

Example 1. *(AIMA, 16.10): 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 a_1 and that there is time to carry out at most one test t_1 which costs 1,000 Kč and which can help to figure out the quality of the car. A car can be in good shape (s_+) or in bad shape (s_-), and the test might help to indicate what shape the car is in. There are only two outcomes for the test: pass (t_{1+}) or fail (t_{1-}). Car a_1 costs 30,000 Kč, and its market value is 40,000 Kč if it is in good shape; if not, 14000 Kč in repairs will be needed to make it in good shape. The buyers estimate is that a_1 has 70% chance of being in good shape. The test is uncertain: $Pr(t_{1+}(a_1)|a_{1+}) = 0.8$ a $Pr(t_{1+}(a_1)|a_{1-}) = 0.35$.*

Calculate the expected net gain from buying car a_1, given no test.

$EU(buy + |\{\}) = \sum_{s\in\{+,-\}} U(s)Pr(s|buy+) = 40,000 - (0.7 \times 30,000 + 0.3 \times 44,000) = 40,000 - 34,200 = 5,800$ Kč

An analogy in classic decision making:
$d^*(t) = \underset{buy+,buy-}{\text{argmin}} \sum_{s\in\{+,-\}} l(d,s)Pr(s|t) = \underset{buy+,buy-}{\text{argmin}} \sum_{s\in\{+,-\}} l(d,s)Pr(s) =$
$= \underset{buy+,buy-}{\text{argmax}} (10000 \times 0.7 - 4000 \times 0.3, 0) = \text{argmax}(5800, 0) = buy+$

Conclusion 1: It pays-off to buy the car without a test.

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.

$Pr(a_{1+}|t_{1+}(a1)) = \frac{Pr(t_{1+}(a1)|a1+)\times Pr(a_{1+})}{Pr(t_{1+}(a1))} = \frac{0.8\times 0.7}{0.8\times 0.7+0.35\times 0.3} = \frac{0.56}{0.665} = 0.842$
$Pr(a_{1+}|t_{1-}(a1)) = \frac{Pr(t_{1-}(a1)|a_{1+})\times Pr(a_{1+})}{Pr(t_{1-}(a1))} = \frac{0.2\times 0.7}{0.2\times 0.7+0.65\times 0.3} = \frac{0.14}{0.335} = 0.418$

Calculate the optimal decisions given either a pass or a fail, and their expected utilities.

$EU(\alpha_{t_1}|t_{1+}(a1)) = 40,000-(0.842\times 30,000+0.158\times 44,000) = 40,000-32,240 = 7,788$ Kč
$EU(\alpha_{t_1}|t_{1-}(a1)) = 40,000-(0.418\times 30,000+0.582\times 44,000) = 40,000-38,120 = 1,852$ Kč

An analogy in classic decision making:
$d^*(t_{1+}(a1)) = \underset{buy+,buy-}{\text{argmin}} \sum l(d,s)Pr(s|t) = \underset{buy+,buy-}{\text{argmin}} (10,000\times 0.842 - 4,000\times 0.158, 0) = \underset{buy+,buy-}{\text{argmin}} (7788, 0) = buy+$
$d^*(t_{1-}(a1)) = \underset{buy+,buy-}{\text{argmin}} \sum l(d,s)Pr(s|t) = \underset{buy+,buy-}{\text{argmin}} (10,000\times 0.418 - 4,000\times 0.582, 0) = \underset{buy+,buy-}{\text{argmin}} (1852, 0) = buy+$

Conclusion 2: It pays-off to buy the car for both the test outcomes. This immediately suggests zero VPI of the test – the test has no potential to change buyer's decision.

Calculate the value of (perfect) information of the test. Should the buyer pay for t_1?

$EU(\alpha|\{\}) = \max(5800, 0) = 5800$ Kč
$EU(\alpha_{t_1}|t_{1+}(a1)) = \max(7788, 0) = 7788$ Kč
$EU(\alpha_{t_1}|t_{1-}(a1)) = \max(1852, 0) = 1852$ Kč

$VPI(t_1(a_1)) = (Pr(t_{1+}(a1))\times 7788 + Pr(t_{1-}(a1))\times 1852) - 5800 = (0.665\times 7788 + 0.335\times 1852) - 5800 = 5800 - 5800 = 0$ Kč
It is "hard" zero, can be confirmed as follows:
$Pr(t_{1+}(a1))\times(10000\times Pr(a_{1+}|t_{1+}(a1)) - 4000\times Pr(a_{1-}|t_{1+}(a1))) + Pr(t_{1-}(a1))\times(10000\times Pr(a_{1+}|t_{1-}(a1)) - 4000\times Pr(a_{1-}|t_{1-}(a1))) = 10000\times(Pr(a_{1+}, t_{1+}(a1)) + Pr(a_{1+}, t_{1-}(a1))) - 4000\times(Pr(a_{1-}, t_{1+}(a1)) + Pr(a_{1-}, t_{1-}(a1))) = 10000\times Pr(a_{1+}) - 4000\times Pr(a_{1-}) = 5800$ Kč
$VPI(t1(a1)) - Cost(t1(a1)) = -1000 < 0$

Conclusion 3: A logical resolution. The test cannot change decision, it has zero value, when considering its cost it brings negative outcome. The best strategy is to buy the

car without the test. The test would need better sensitivity to pay-off. Accuracy of the test (note that the trivial "good state" classifier shows accuracy 0.7):
$Pr(t_{1+}(a1), a_{1+}) + Pr(t_{1-}(a1), a_{1-}) = 0.8 \times 0.7 + 0.65 \times 0.3 = 0.755$

Example 2. *You are going on a trip from San Francisco to Oakland. You have two options to get to Oakland, you want to get there as soon as possible. You can drive your car across the Bay Bridge or go by train through the tunnel under the bay. Bay Bridge is often jammed (on the given part of the day it is in about 40 % of cases). During normal operation, it takes 30 minutes drive. If there is traffic congestion, it takes 1 hour. The train journey always takes 40 minutes.*

When having no traffic information, does it pay off to drive or take a train?

$EU(train|\{\}) = 40$ min
$EU(car|\{\}) = \sum_{z\in\{+,-\}} U(z)Pr(z|car) = 0.4 \times 60 + 0.6 \times 30 = 42$ min

Conclusion 1: The train journey is faster.

Let us assume, that the traffic information for Bay Bridge is available on web, you can get it in 5 minutes. You know, that for congested bridge, the web page says the same with 90% probability. For normal traffic, the page indicates a traffic jam in 20% cases.

What is the congestion probability when having the traffic information?

We employ Bayes theorem (z . . . congestion, real . . . real situation, pred . . . traffic information prediction):

$P(z_{real}|z_{pred}) = \frac{P(z_{pred}|z_{real})P(z_{real})}{P(z_{pred})} = \frac{0.9\times0.4}{P(z_{pred})} = \frac{0.36}{P(z_{pred})}$
$P(\neg z_{real}|z_{pred}) = \frac{P(z_{pred}|\neg z_{real})P(\neg z_{real})}{P(z_{pred})} = \frac{0.2\times0.6}{P(z_{pred})} = \frac{0.12}{P(z_{pred})}$
$P(z_{real}|z_{pred}) = \frac{0.36}{0.36+0.12} = 0.75$
$P(\neg z_{real}|z_{pred}) = \frac{0.12}{0.36+0.12} = 0.25$

Note: $P(z_{pred}) = 0.48$, and $P(\neg z_{pred}) = 0.52$.

$P(z_{real}|\neg z_{pred}) = \frac{P(\neg z_{pred}|z_{real})P(z_{real})}{P(\neg z_{pred})} = \frac{0.1\times0.4}{0.52} = 0.07$
$P(\neg z_{real}|\neg z_{pred}) = \frac{P(\neg z_{pred}|\neg z_{real})P(\neg z_{real})}{P(\neg z_{pred})} = \frac{0.8\times0.6}{0.52} = 0.93$

What should we do if the traffic information predicts normal operation / congestion?

$EU(car|z_{pred}) = P(z_{real}|z_{pred}) \times 60 + P(\neg z_{real}|z_{pred}) \times 30 = 0.75 \times 60 + 0.25 \times 30 = 52.5$ min
$EU(car|\neg z_{pred}) = P(z_{real}|\neg z_{pred}) \times 60 + P(\neg z_{real}|\neg z_{pred}) \times 30 = 0.07 \times 60 + 0.93 \times 30 = 32$ min

Conclusion 2: If congestion is predicted, we will take train, otherwise we will go by car.

Is it efficient to spend 5 minutes by finding out the traffic information or is it better to simply set out?

If congestion is predicted, we will take train and vice versa. The train journey takes 40 min, the drive through the free bridge takes 32 minutes. These times must be weighted by their probability given by the probability of both the states of traffic information and add 5 min to both for the time needed to find out the prediction. On average, we would reach Oakland in:

$U(z_{pred}) = 5 + P(z_{pred}) \times 40 + P(\neg z_{pred}) \times 32 = 5 + 0.48 \times 40 + 0.52 \times 32 = 40.8$ min

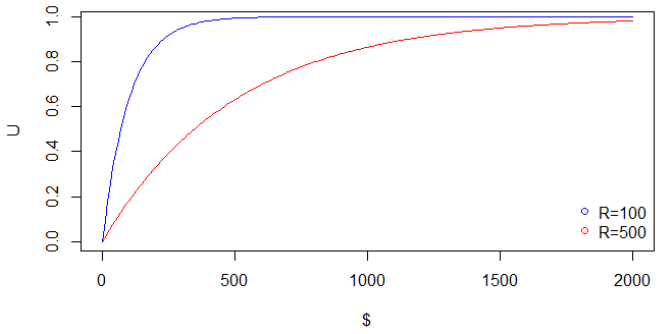
This travel time needs to be compared with the default option without any information. This option is the train journey in 40 minutes.

Conclusion 3: Traffic information helps to increase quality of our decision. However, the 5 minute time for its acquisition is too large. The best option is to simply set out by train.

Example 3. *(AIMA, 16.12): Economists often make use of an exponential utility function for money $U(x) = 1 - e^{-x/R}$, where R is a positive constant representing an individual's risk tolerance. Risk tolerance reflects how likely an individual is to accept a lottery with a particular expected monetary value (EMV) versus some certain payoff. As R (which is measured in the same units as x) becomes larger, the individual becomes less risk-averse.*

Assume Mary has an exponential utility function with $R = \$500$. Mary is given the choice between receiving \$500 with certainty (probability 1) or participating in a lottery that has a 60% probability of winning \$5000 and a 40% probability of winning nothing. Assuming Mary acts rationally, which option would she choose? Show how you derived your answer.

Let us see how the utility function changes with increasing R. The figure below demonstrates that the exponential utility function is concave and leads to risk-avoiding decisions. For larger Rs, the person tends to be more risk neutral for the amounts of money discussed in the excercise.



Mary simply solves the inequality:

$$U(\$500) = 1 - e^{-500/500} = 0.6321 > 0.6 \times U(\$5000) = 0.6(1 - e^{-5000/500}) = 0.6$$

Conclusion 1: Mary would choose the first option, i.e. to receive \$500 with certainty, as it has higher utility.

Consider the choice between receiving \$100 with certainty (probability 1) or participating in a lottery that has a 50% probability of winning \$500 and a 50% probability of winning nothing. Approximate the value of R (to 3 significant digits) in an exponential utility function that would cause an individual to be indifferent to these two alternatives.

We need to solve the equation: $1 - e^{-100/R} = 0.5(1 - e^{-500/R})$. This equation can be converted into: $1 = 2e^{-100/R} - (e^{-100/R})^5$. After substitution $x = e^{-100/R}$ we obtain a quintic equation: $x^5 - 2x + 1 = 0$. This equation has two roots (the first is obvious, the second was reached with numerical methods): $x_1 = 1$, $x_2 = 0.51879$. The first root leads to a trivial solution $R_1 \to \infty$. The second root leads to another solution $R_2 = 152.3796$.

Conclusion 2: The individual becomes indifferent to the two above-mentioned options for $R \approx 152$.

2 Markov decision processes

Example 4. *Concern an episodal process with three states* $(1, 2, 3)$. *The rewards in individual states are* $R(1) = -1$ $R(2) = -2$, *and* $R(3) = 0$, *the process terminates by reaching state* 3. *In the states* 1 *and* 2, *actions a and b can be applied. Action a keeps the current state with 20% probability, with 80% probability it leads to transition from* 1 *to* 2 *resp. from* 2 *to* 1. *Action*

b keeps the current state with 90% probability, with 10% probability it leads to state 3.

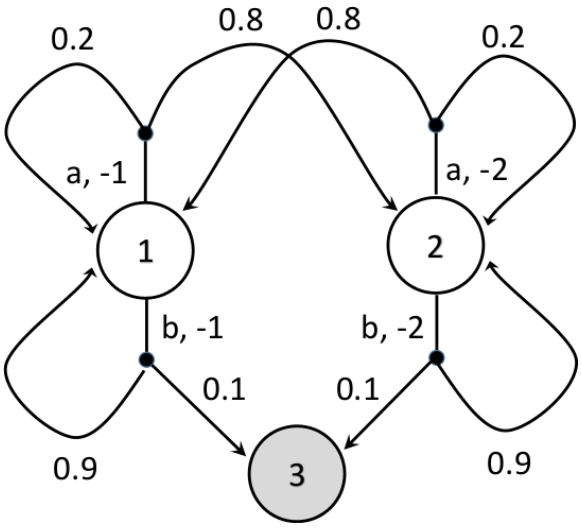
Try to guess the best policy qualitatively for states 1 and 2.

We maximize our reward, the rewards are not positive, the process should be terminated as soon as possible, i.e., state 3 should be reached. At the same time, the transition to 3 by b is relatively improbable. The expected number of b trials to terminate the process is 10 (for $p = 0.1$: $E = p + 2p(1-p) + 3p(1-p)^2 + \cdots = 1/p = 10$). State 1 has twice smaller penalty, we will directly try to terminate the process through b when being in it. On the contrary, in state 2 it pays off to switch to state 1 first through a and only then to employ b to terminate the process by transition to state 3.

Conclusion 1: The best policy seems to be $\pi^* = (b, a, NULL)$.

Formalize as MDP. Apply policy iteration. Start with the policy $\pi_0 = (b, b, NULL)$ and illustrate its convergence to the optimal policy in detail.

MDP is a quadruple {S,A,P,R}, where S={1,2,3}, A={a,b} and P and R correspond to the description above. MDP can also be summarized as a state-transition diagram:



Policy iteration proceeds as follows:
Init: $\pi_0 = (b, b, NULL)$, $U_0 = (0, 0, 0)$.
Iteration 1:
Evaluation: $U_1(1) = -1 + 0.9 \times U_1(1)$, $U_1(2) = -2 + 0.9 \times U_1(2)$, $U_1(3) = 0$,
$U_1(1) = -10$, $U_1(2) = -20$,

(can be solved by DP, an analytical solution is available too).
Improvement: $Q_1(a,1) = -1 + 0.2 \times U_1(1) + 0.8 \times U_1(2) = -19$,
$Q_1(b,1) = -1 + 0.9 \times U_1(1) + 0.1 \times U_1(3) = -10$,
$Q_1(a,1) < Q_1(b,1) \rightarrow$ for state 1 we pick b,
$Q_1(a,2) = -2 + 0.8 \times U_1(1) + 0.2 \times U_1(2) = -14$,
$Q_1(b,2) = -2 + 0.9 \times U_1(2) + 0.1 \times U_1(3) = -20$,
$Q_1(a,2) > Q_1(b,2) \rightarrow$ for state 2 we pick a,
$\pi_1 = (b, a, NULL)$.
Iteration 2:
Evaluation: $U_2(1) = -1 + 0.9 \times U_2(1)$, $U_2(2) = -2 + 0.8 \times U_2(1) + 0.2 \times U_2(2)$,
$U_2(1) = -10$, $U_2(2) = -12.5$,
(analytical solution again).
Improvement: $Q_2(a,1) = -1 + 0.2 \times U_2(1) + 0.8 \times U_2(2) = -13$,
$Q_2(b,1) = -1 + 0.9 \times U_2(1) + 0.1 \times U_2(3) = -10$,
$Q_2(a,1) < Q_2(b,1) \rightarrow$ for state 1 we pick b,
$Q_2(a,2) = -2 + 0.8 \times U_2(1) + 0.2 \times U_2(2) = -12.5$
$Q_2(b,2) = -2 + 0.9 \times U_2(2) + 0.1 \times U_2(3) = -13.25$
$Q_2(a,2) > Q_2(b,2) \rightarrow$ for state 2 we pick a,
$\pi_2 = (b, a, NULL)$, no policy change, STOP.

Conclusion 2: We confirmed our qualitative guess, the optimal policy is $\pi^* = (b, a, NULL)$.

Now apply value iteration. Briefly compare with policy iteration.

The iteration starts with the zero value vector $U_0 = (0, 0, 0)$. Then, value iteration works with the above-written action value formulas ($Q(a,1)$, $Q(b,1)$,$Q(a,2)$, $Q(b,2)$). It always takes the better action out of the two actions available (in each state and iteration), formally $U_i(s) = max_\alpha Q_i(\alpha, s)$:

iteration	$Q(a,1)$	$Q(b,1)$	$Q(a,2)$	$Q(b,2)$	action1	action2
1	0	0	0	0	any	any
2	-1	-1	-2	-2	any	any
3	-2.8	**-1.9**	**-3.2**	-3.8	b	a
4	-3.94	**-2.71**	**-4.16**	-4.88	b	a
5	-4.87	**-3.44**	**-5**	-5.74	b	a
...						
69	-12.99	**-9.99**	**-12.49**	-13.24	b	a

Conclusion 3: The optimal policy is again $\pi^* = (b, a, NULL)$. The policy does not change since the third iteration step. If we choose $\epsilon = 0.001$ (the threshold for the maximum change in state values between two consecutive steps), the iteration stops

in the 69th step, the state values are nearly identical with the previously found values U(1)=-10, U(2)=-12.5.

Reapply policy iteration. Start with $\pi_0 = (a, a, NULL)$. What happens? What is the solution?

Init: $\pi_0 = (a, a, NULL)$, $U_0 = (0, 0, 0)$.
Iteration 1:
Evaluation: $U_1(1) = -1 + 0.2 \times U_1(1) + 0.8 \times U_1(2)$,
$U_1(2) = -2 + 0.8 \times U_1(1) + 0.2 \times U_1(2)$,
(this system of equations has no solution)
(DP solution diverges, the state values grow towards ∞).

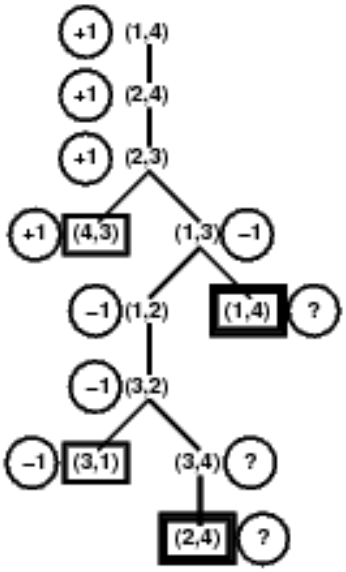
Conclusion 4: The solution is to introduce a discount factor γ. The system of linear equations will not be singular any longer. But, a bit different task gets solved, the best policy can be different, especially for small discount factors. With a small γ, immediate reward gets preferred, one can find b as the best option even in state 2.

Example 5. *Consider a two-player game on a four-field board. Each player has one stone, the goal is to move its stone to the opposite side of the board (A player moves from field 1 to field 4, B player from field 4 to field 1). The player that first reaches its goal field wins. The players may move one filed left or right, they cannot skip their move nor move out of the board. If a neighbor field is occupied by the opponent's stone, the stone can be jumped. (example: if A is in the position 3 and B in the position 2 and A moves left, it ends up in position 1).*

Which player wins? Demonstrate the classic solution based on state space search first.

The state of the game can be represented by the position of both the stones, it can be written down as an ordered pair (s_A, s_B). There are 11 reachable states (state $(4, 1)$ is not reachable). The standard solution is by MiniMax procedure. The game tree is in figure below (the evaluation is for A player, who is a maximization player).
The only bottleneck lies in the fact that the game contains cycles and the standard depth-first MiniMax would fall into an infinite loop. For this reason, we will put the expanded states on a stack. As soon as a cycle is detected, the value of the state is

denoted as "?" and the current branch is terminated. When propagating the evaluation we assume that max(1,?)=1 and min(-1,?)=-1. This improvement is sufficient for the given game which does not distinguish beyond wins and losses.



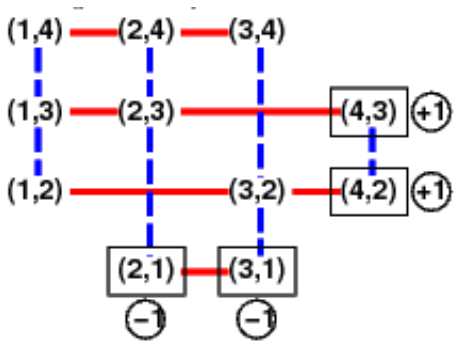
Conclusion 1: Two-player games can be solved by MiniMax. If keeping the optimal game strategy, A player wins (the player who moves first).

Can we formalize this game as MDP? Is it a good choice?

Conclusion 2: Every search problem can be formalized as MDP. The transformation is routine: states and actions do not change, the goal states map to terminal MDP states, the transition matrix is deterministic and the reward is inverted evaluation. However, MDP is not a good choice in the case of deterministic actions. Its formalism is too heavy and time demanding. It is a good choice for stochastic problems.

Formalize this game as MDP. Let $V_A(s)$ be the state s value if A player is on move, $V_B(s)$ be the state s value if B player is on move. Let $R(s)$ be the reward in state s, for the terminal states where wins A it is 1, for the terminal states where wins B it is -1. Draw a state space diagram. Put down Bellman equations for both the players and apply these equations in terms of value iteration. Formulate the iteration termination condition.

The state space diagram is in figure below. The moves of A player are in solid red, the moves of B player in dashed blue.



Bellman equations stem from the MiniMax principal:
$V_A(s) = R(s) + \max_a P^a_{ss'} V_B(s')$
$V_B(s) = R(s) + \min_a P^a_{ss'} V_A(s')$

$R(s)$ will only be used in terminal states, the value of the rest of the states is given solely by its descendants. A player maximizes evaluation, B player does the opposite. As the actions are deterministic, each action has the unit probability for one of the descendant states and zero probability for remaining states.

The players take moves in turns, we apply the individual Bellman equations in turns too. In the beginning, the terminal states start with their $R(s)$, the rest of the states has zero value. The values gradually propagate, the state space diagram is used, see the table below:

s	(1,4)	(2,4)	(3,4)	(1,3)	(2,3)	(4,3)	(1,2)	(3,2)	(4,2)	(2,1)	(3,1)
V_A	0	0	0	0	0	+1	0	0	+1	-1	-1
V_B	0	0	0	0	-1	+1	0	-1	+1	-1	-1
V_A	0	0	0	-1	+1	+1	-1	+1	+1	-1	-1
V_B	-1	+1	+1	-1	-1	+1	-1	-1	+1	-1	-1
V_A	+1	+1	+1	-1	+1	+1	-1	+1	+1	-1	-1
V_B	-1	+1	+1	-1	-1	+1	-1	-1	+1	-1	-1

The termination condition is no change in value vector for one of the players (i.e., the match between the current $V_A(s)$ vector and the $V_A(s)$ vector generated two moves before, or the same match for $V_B(s)$). In the table above, we observe the match in two last $V_B(s)$ vector instances. Obviously, no change may appear for the next $V_A(s)$ too as it will be derived from the identical $V_B(s)$.

Note that $V_A(s)$ and $V_B(s)$ vectors do not have to match. $V_A(s)$ assumes that A player is on move and vice versa (e.g., (3,2) state switches its value in principle as the player on move always wins whoever it is).

Conclusion 3: MDP solves the problem concurrently for both the players taking the first move. The value of the terminal states is given apriori. In states (2,4) and (3,4), A player wins disregarding turns. In states (1,3) and (1,2), B player wins disregarding turns. In states (1,4), (2,3) and (3,2), the player on move wins.
MiniMax tree shown earlier employs different state values at different tree levels, the tree de facto combines $V_A(s)$ and $V_B(s)$ according to the tree depth.

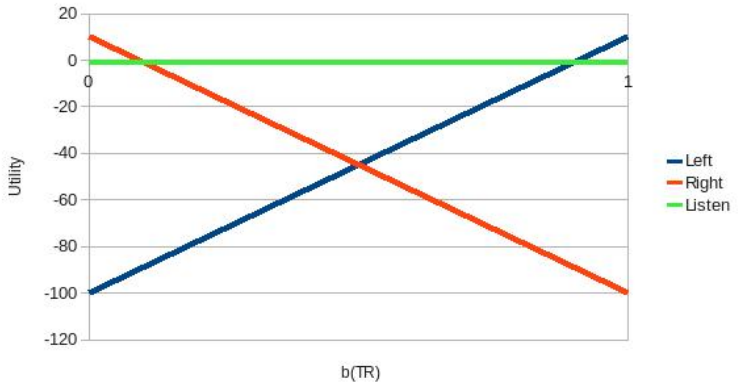
Example 6. The tiger problem (POMDP). *An agent stands in front of two closed doors. Behind one of the doors is a tiger and behind the other is freedom. If the agent opens the door with the tiger, the tiger eats the agent (a large penalty -100 is received). If the agent opens the other door, it obtains a reward, its value is +10. Instead of opening one of the two doors, the agent can listen, in order to gain some information about the location of the tiger. Unfortunately, listening is not free, it costs -1; in addition, it is also not entirely accurate. There is a 15% chance that the agent will hear tiger behind the left-hand door when the tiger is really behind the right-hand door, and vice versa. If the agent listens to the tiger repeatedly, it has to pay its cost repeatedly, however, the mishearings can be considered independent.*

Formalize the tiger problem as a partially observable Markov decision process.

POMDP $= \{S, A, P, R, O, \Omega\}$,
hidden states: S=$\{TL, TR, STOP\}$, TL $\sim$ tiger left, TR $\sim$ tiger right, STOP $\sim$ the end of game (a door was opened)
actions: A=$\{Li, L, R\}$, Li $\sim$ listen=wait for the next roar, L $\sim$ open the left door, R $\sim$ open the right door,
observations: O=$\{LL, LR\}$, LL $\sim$ the tiger heard left, LR $\sim$ the tiger heard right,
transition probs (P): $Pr(TL|TL, Li) = 1$, $Pr(TR|TL, Li) = 0$, $Pr(TL|TL, L) = 0$, $Pr(TR|TL, L) = 0$,$Pr(STOP|TL, L) = 1$ (for TL only, the TR case is symmetric),
sensoric model (Ω): $Pr(LL|TL, Li) = 0.85$, $Pr(LR|TL, Li) = 0.15$,
reward: $R(Li, TL) = -1$, $R(R, TL) = 10$, $R(L, TL) = -100$.

Find the optimal 1-step plan as a function of belief. In other words, propose the optimal action as a result of your belief in the tiger's current location. Identify the belief space positions where the action changes.

We deal with two states only, the belief can be represented as a real number between 0 and 1. We will denote it as $b(TR)$, $b(TL)$ is the complement to 1. The utility of the individual actions will be functions of one variable, the variable b. Since we know that they are linear, it is sufficient to calculate them in the ultimate points of the belief space, i.e., in the situations where the position of the tiger is clear. See the figure below.



Conclusion 1: The figure suggests that for $b(TR) > 0.9$ it pays off to choose the action L, pro $b(TR) < 0.1$ it pays off to choose the action R. In the central part of the belief space it is advantageous to choose Li. The values 0.1 and 0.9 follow from the equations: $-100b(TR) + 10(1 - b(TR)) = -1$, resp. $-100(1 - b(TR)) + 10 \times b(TR) = -1$.

How many conditional plans of length 2 do we have? Calculate the utility of one of them (it is a function of b again). Will be any of these plans clearly dominated by the other plans?

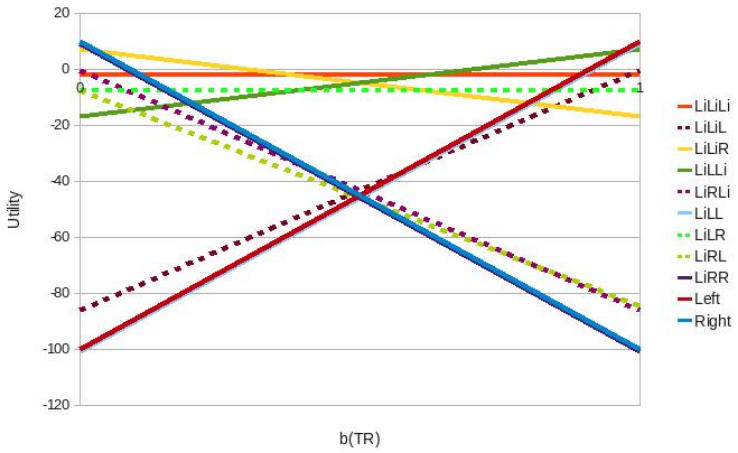
A conditional plan of length 2 is such a plan, which determines the action for the given instant and then the action after the next roar. The roar can be heard from two directions, we need a recommendation for both of them. Consequently, the plan will have three action altogether $[A_1$ if LR then A_2 else $A_3]$, the simplified form is $[A_1A_2A_3]$. In theory, there are 27 plans (sequences of length 3 over the alphabet of three actions, i.e., 3^3 options). However, all the plans that do not start with Li restart the game, i.e., it exists only 9 conditional plans of length 2.

The most trivial case is the plan $[Li, Li, Li]$:
$\alpha_{[LiLiLi]}(b(TR) = 0) = \alpha_{[LiLiLi]}(b(TR) = 1) = -2$,

A more difficult plan is $[LiLLi]$:
$\alpha_{[LiLLi]}(b(TR) = 0) = R(Li, TL) + Pr(TR|TR, Li)(Pr(LL|TR, Li) \times \alpha_{[L]}(0) +$

$Pr(LR|TR,Li)\alpha_{[Li]}(0))+Pr(TL|TR,Li)(Pr(LL|TL,Li)\times\alpha_{[L]}(0)+Pr(LR|TL,Li)\alpha_{[Li]}(0)) = -1+1(0.15\times-100+0.85(-1))+0(\dots) = -16.85,$
$\alpha_{[LiLLi]}(b(TR) = 1) = R(Li,TR) + Pr(TL|TL,Li)(Pr(LL|TL,Li) \times \alpha_{[L]}(1) + Pr(LR|TL,Li)\alpha_{[Li]}(1))+Pr(TR|TL,Li)(Pr(LL|TR,Li)\alpha_{[L]}(1)+Pr(LR|TR,Li)\alpha_{[Li]}(1)) = -1+1\times(0.85\times10+0.15\times-1)+0\times(\dots) = 7.35,$

Out of the 9 plans starting with Li there are only 5 dominating: $[LiRR]$ for beliefs<0.019, $[LiLiR]$ for beliefs 0.019-0.39, $[LiLiLi]$ for beliefs 0.39-0.61, $[LiLLi]$ fr beliefs 0.61-0.981, $[LiLL]$ for beliefs>0.981. In the first case it is so clear that the state is TR that any direction keeps the action R. In the second case, TR has to be confirmed by LR, otherwise the agent keeps listening. The central part of the belief space is insecure, no direction allows to securely open a door. The rest is symmetric ...
It is obvious that the plan $[LiRLi]$ makes no sense, $[LiLiL]$ is its symmetric complement, $[LiRL]$ is clearly counter-intuitive too (open the right door if hearing the tiger from the right and vice versa). $[LiLR]$ is a potentially interesting plan, but it necer breaks through in the current parametrization, the penalty for opening of the wrong door is too large.



When solving the problem as a whole, these plans have to be compared with the plans of length 1 that finish the game, i.e., L and R. Obviously, they outperform $[LiLL]$, resp. $[LiRR]$ for the ultimate beliefs as there is no sense to listen there with no chance to change the original decision.
Switching between $[LiRR]$ and $[LiLiR]$: $7.35-(16.85+7.35)x = -2, x = 9.35/(16.85+7.35) = 0.39$ Switching between $[LiLiR]$ and $[LiLiLi]$: $7.35-(16.85+7.35)x = 9-(101+9)x, x = (9-7.35)/(110-24.2) = 0.019$

Conclusion 2: There are 9 conditional plans of length 2. 4 of them are dominated by

the other plans. When considering the plans of length 1, there are only 3 conditional plans of length 2 that are not dominated.

How many times does the agent have to hear the tiger from the same direction to open a door? Consider the beginning of the game where $b(TR) = 0.5$. Explain.

In problem b) we learnt that b(TR) must be smaller than 0.1 or greater than 0.9, this can be reached after [LL, LL] or [LR, LR]. Let us test the option [LR, LR] (watch out, $b_2(TR)$ cannot be calculated as $(1 - 0.15^2) = 0.9775$):

$b_0(TR) = 0.5$ (the beginning),
$b_1(TL) = \alpha Pr(LR|TL, Li)(Pr(TL|TL, Li) \times b_0(TL) + Pr(TL|TR, Li) \times b_0(TR)) =$
$\alpha \times 0.15 \times (1 \times 0.5 + 0 \times 0.5) = \alpha \times 0.15 \times 0.5$,
$b_1(TR) = \alpha Pr(LR|TR, Li)(Pr(TR|TR, Li) \times b_0(TR) + Pr(TR|TL, Li) \times b_0(TL)) =$
$\alpha \times 0.85 \times (1 \times 0.5 + 0 \times 0.5) = \alpha \times 0.85 \times 0.5$,
$b_1(TL) + b_1(TR) = 1 \ldots \alpha = 2 \ldots b_1(TL) = 0.15, b_1(TR) = 0.85$,
a single roar is not enough,

$b_2(TL) = \alpha Pr(LR|TL, Li)(Pr(TL|TL, Li) \times b_1(TL) + Pr(TL|TR, Li) \times b_1(TR)) =$
$\alpha \times 0.15 \times (1 \times 0.15 + 0 \times 0.5) = \alpha \times 0.15^2$,
$b_2(TR) = \alpha Pr(LR|TR, Li)(Pr(TR|TR, Li) \times b_1(TR) + Pr(TR|TL, Li) \times b_1(TL)) =$
$\alpha \times 0.85 \times (1 \times 0.5 + 0 \times 0.5) = \alpha \times 0.85^2$,
$b_2(TL) + b_2(TR) = 1 \ldots \alpha = 1.34 \ldots b_1(TL) = 0.03, b_1(TR) = 0.97$,
two roars are sufficient.

Conclusion 3: The tiger must be heard two times from the same direction to open a door. However, the reasoning is indicative only, the waiting penalty considered in problem b) is not final.

Let us consider the following policy π: wait until the number of roars from one direction is higher by 2 than from the other direction then open the door with the smaller number of roars.

Let us consider $b_0(TR) = 0$ (the other extreme is symmetric):
$\alpha_\pi = 0.85^2 \times (10 - 2) + 0.15^2 \times (-100 - 2) + 2 \times 0.85 \times 0.15 \times (\alpha_\pi - 2)$
(the first element corresponds to the correct choice of R, the second one to the incorrect choice of L after two (unlikely) mishearings, the last element stands for the return tio the original state after two opposing hearings and loosing -2 for waiting)
$0.745 \times \alpha_\pi = 3.48 - 0.51 \rightarrow \alpha_\pi = 3.99$

Conclusion 4: This policy reaches the expected utility 4 in all the belief space. It dominates the simple plan $[LiLR]$ and for the initial uniformed state $b_0(TR) = 0.5$ it

outperforms all the other above-mentioned plans. The immediate opening of a door pays off even less than considered before (for stronger beliefs) than it seemed from the plans of length 1.