Exercises on reinforcement learning, IN4080

(Extracted from exams of previous years)

Question 1

You wish to develop a (phone-based) spoken dialogue system that will call random U.S. citizens in order to collect opinion poll data for the next US election, with two candidates on the ballot: Kamala Harris and Ivanka Trump.

This system is framed as a *Markov Decision Process* (MDP) formalised as such:

- We have five possible states:
  - s_1 is the starting state
  - s_2 if the callee indicated their intention to vote for Kamala Harris
  - s_3 if the callee indicated their intention to vote for Ivanka Trump
  - s_4 if the callee expressed something else (that was not understood)
- The set of actions that can be taken by the dialogue system are as follows:
  - a_1 : Say "Hi, I'm a automated bot developed to collect polling data. May I ask you for whom you plan to vote in the next election?"
  - a_2 : Say "Sorry I did not understand. Who do you wish to vote for?"
  - a_3 : Say "Ok, thank you for your help, and have a nice day!"
- The transition model is as follows:
  - In state s_1, only action a_1 is possible, with three possible transitions: $P(s'=s_2|s=s_1, a=a_1) = 0.48, P(s'=s_3|s=s_1, a=a_1) = 0.40, P(s'=s_4|s=s_1, a=a_1) = 0.12$
  - In states s_2 and s_3, only a_3 is possible and terminates the dialogue.
  - In state s_4, only a_2 is possible, with the following transitions: $P(s'=s_2|s=s_4, a=a_2) = 0.36, P(s'=s_3|s=s_4, a=a_2) = 0.32, P(s'=s_4|s=s_4, a=a_2) = 0.32$
- Finally, the reward model is defined as such:
  - $R(s = s_2, a = a_3) = R(s = s_3, a = a_3) = 10$ (if the system manages to register the callee's political preference)
  - $R(s = s_4, a = a_2) = -1$ (to capture the annoyance of asking the callee to repeat)
  - Other actions have a reward of zero.

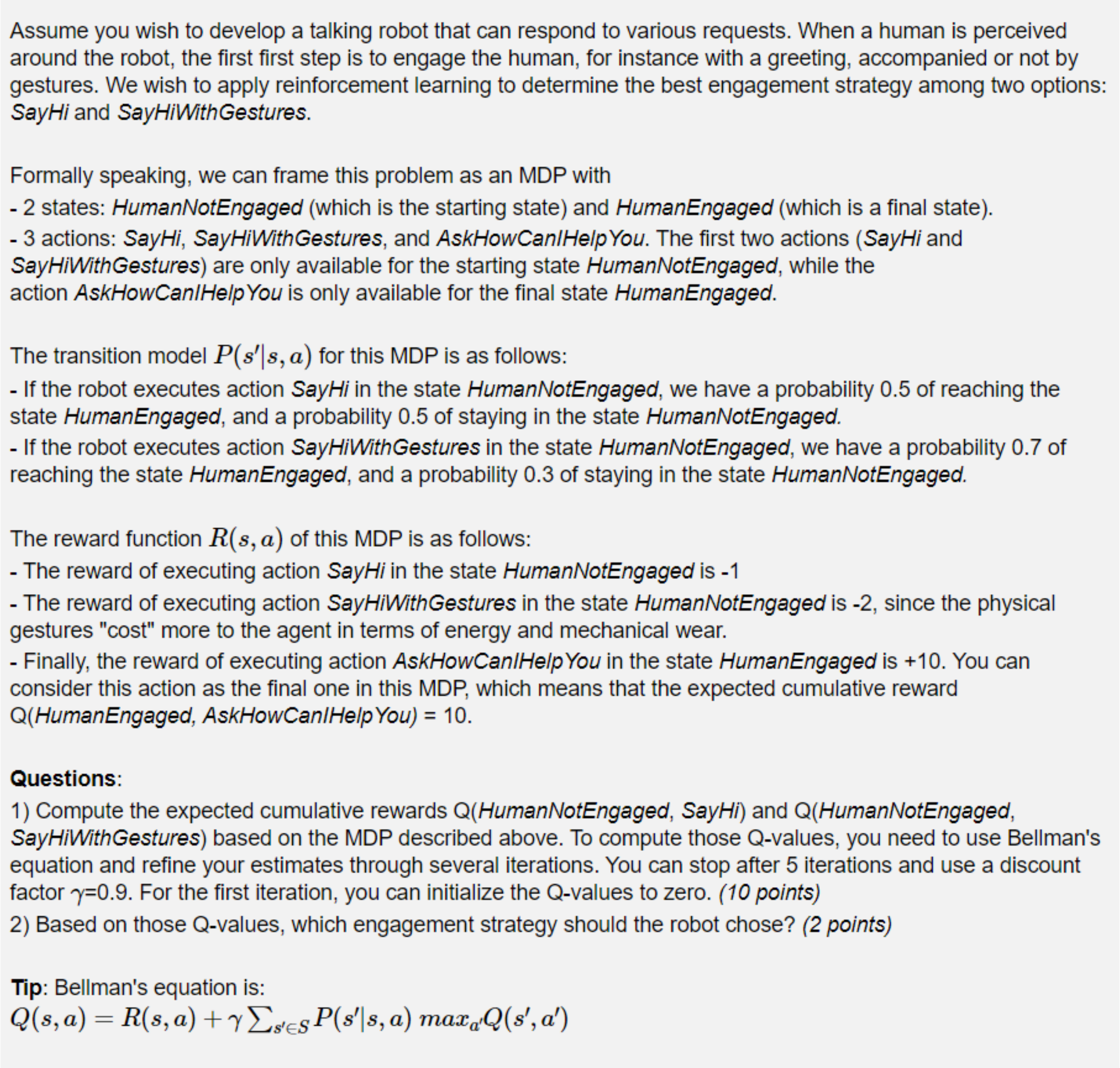
1.1)

Based on this MDP model, calculate the expected cumulative reward of asking the callee to repeat when their answer was not properly understood, that is: $Q(s=s_4, a=a_2)$. You can assume a discount factor of 0.9.

1.2)

One limitation of this MDP model is that is assumes that the dialogue system will always be 100 % certain it has correctly understood the political preference expressed by the callee. In practice, this will not always be the case, because of e.g. speech recognition or NLU errors, or because the callee may intentionally provide unclear or misleading information. How could this model be modified to capture those uncertainties?

Question 2



Assume you wish to develop a talking robot that can respond to various requests. When a human is perceived around the robot, the first first step is to engage the human, for instance with a greeting, accompanied or not by gestures. We wish to apply reinforcement learning to determine the best engagement strategy among two options: *SayHi* and *SayHiWithGestures*.

Formally speaking, we can frame this problem as an MDP with

- 2 states: *HumanNotEngaged* (which is the starting state) and *HumanEngaged* (which is a final state).
- 3 actions: *SayHi*, *SayHiWithGestures*, and *AskHowCanIHelpYou*. The first two actions (*SayHi* and *SayHiWithGestures*) are only available for the starting state *HumanNotEngaged*, while the action *AskHowCanIHelpYou* is only available for the final state *HumanEngaged*.

The transition model $P(s'|s,a)$ for this MDP is as follows:

- If the robot executes action *SayHi* in the state *HumanNotEngaged*, we have a probability 0.5 of reaching the state *HumanEngaged*, and a probability 0.5 of staying in the state *HumanNotEngaged.*
- If the robot executes action *SayHiWithGestures* in the state *HumanNotEngaged*, we have a probability 0.7 of reaching the state *HumanEngaged*, and a probability 0.3 of staying in the state *HumanNotEngaged.*

The reward function $R(s,a)$ of this MDP is as follows:

- The reward of executing action *SayHi* in the state *HumanNotEngaged* is -1
- The reward of executing action *SayHiWithGestures* in the state *HumanNotEngaged* is -2, since the physical gestures "cost" more to the agent in terms of energy and mechanical wear.
- Finally, the reward of executing action *AskHowCanIHelpYou* in the state *HumanEngaged* is +10. You can consider this action as the final one in this MDP, which means that the expected cumulative reward Q(*HumanEngaged, AskHowCanIHelpYou)* = 10.

Questions:

1) Compute the expected cumulative rewards Q(*HumanNotEngaged*, *SayHi*) and Q(*HumanNotEngaged*, *SayHiWithGestures*) based on the MDP described above. To compute those Q-values, you need to use Bellman's equation and refine your estimates through several iterations. You can stop after 5 iterations and use a discount factor γ=0.9. For the first iteration, you can initialize the Q-values to zero. *(10 points)*
2) Based on those Q-values, which engagement strategy should the robot chose? *(2 points)*

Tip: Bellman's equation is:
$Q(s,a) = R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a)\ max_{a'} Q(s',a')$