Intelligent Multi Agent Systems

TECHNISCHE UNIVERSITÄT DARMSTADT

Summer Semester 2016, Homework 4 (46 points + 38 bonus)

G. Neumann, G. Gebhardt, O. Arenz

Due date: 1469577300 Unix time, i.e., Tuesday 26th July, 2016 23:55 CEST

How to hand in: Besides the Matlab code, the solutions have to be handed in on paper (either in human readable calligraphy or LaTeX) by putting them into the mailbox in front of S2|02 E315. The well commented Matlab code has to be handed in via moodle. Working in groups of up to three people is allowed, however every student needs to submit on its own (both, paper work and Matlab code) and name his collaborators. The department rules regarding plagiarism apply.

Problem 4.1 Theoretical Questions

Keep your answers short! As a rule of thumb: not more than two sentences per point! Explain your answers.

a) **POMDP** **[4 Points]**

- What is a POMDP?
- How is it defined? (formal)
- How is it different from a common MDP?
- Given an example of a POMDP.

b) **State Definitions** **[3 Points]**

- Name the different state definitions, introduced in the lecture.
- How do they differ from one another?
- What are their advantages/disadvantages?

c) **Alpha Vectors** **[3 Points]**

- What are alpha vectors?
- How do they relate to POMDPs?
- Which property allows us to use them?

d) **Belief-State Value Iteration** **[4 Points]**

- Describe Belief-State Value Iteration
- How are alpha vectors used?
- What are the main update steps?

e) **Point-Based Value Iteration** **[4 Points]**

- Describe Point-Based Value Iteration
- How are alpha vectors used?
- What are the main update steps?

f) **Bayes Adaptive MDPs** **[4 Points]**

- What is the main concept of BAMDPs? When are they useful?
- Why is a Dirichlet distribution a good prior for discrete systems?
- When does a BAMDP become a BAPOMDP?
- Can we use Belief-State Value Iteration on BAPOMDPs? Is it a good idea?

g) **Decentralized POMDPs** **[4 Points]**

- How do they differ from common to POMDPs??
- How is a Dec-POMDP defined? (formal)
- Can we use the same algorithms as for common POMDPs? Why?
- Are there other ways to solve Dec-POMDPs? Explain one conceptually.

Problem 4.2 Pen & Paper Exercises: Learning Control Policies for Unknown Worlds

a) **Back to the west ...** **[20 Points]**

It's a hot Tuesday afternoon in the great Valley of San Pedros. It used to be the place of the intelligent, the autonomous, the ones who knew their way around the system. Coyote and Roadrunner used to have a Casita here for their romantic getaways during the season breaks. But this is all in the past, lets just say... where the rich go, the wicked follow!

It is this dusty dystopia where the three outlaws, Big kahuna fiery Filipe, Roberto the Kid and Jolly Jumper meet once again. Obviously they didn't learn anything from their last encounter, so they face each other in an good old fashioned IAS stand off. They are too busy throwing threats around to realize, that this time, there is a new man in the picture, Geraldo Newman.

While Filipe, Roberto and Jolly point there weapons at each other, Geraldo throws a smoke grenade in front of them and knocks them all out. When they wake up they are chained one behind the other to Geraldos Horse. First Jolly, then Filipe and finally Roberto. Geraldo looks at them and says: "I AM NOT CRAZY. I will proof it too you!", even though none of the three even said anything at all.

He pulls five lucky rabbit feet out of his hat and shows them to the guys. Three black ones and two white ones. He attaches one rabbit foot to each of the outlaws, such that the respective outlaw himself can not see which color his rabbit food has. However, each rabbit foot can be clearly seen from behind, i.e., none of them knows the color of their own rabbit foot, Roberto and Filipe know the color of Jollys rabbit foot and Roberto additionally knows the color of Filipes rabbit foot. The two remaining rabbit feet go back into Geraldos hat, without disclosing their colors to the outlaws.

Geraldo explains: "Lets play a little game! Each of you has a rabbit foot attached to him. But neither of you knows its color. Each of you has three options:

a_1 Be silent. Nothing will happen.

a_2 Tell me the wrong color and I shoot you all right here.

a_3 Tell me the correct color and I'll let all three of you go! And only you get to keep the rabbit foot as a lucky charm!

One little rule though... if you try to talk to each other I'll shoot all three of you! Is that something a crazy person would do?!". Still no one even suggested he would be crazy.

All three of them really want to live and ideally keep the lucky charm. Therefore, each of them would immediately respond if he would be a 100% sure about the color of his rabbit foot!

Needless to say that the three outlaws weaseled their way out of this dilemma. They quickly made their decisions based on their *Visual* and *Auditory* observations. All three of them were quite for a little while. Finally one of them successfully announced the color of his rabbit foot.

- Which of the outlaws saved the day?
- What color has his rabbit foot?

Let's make the following assumptions to ease the formulation of the problem:

- The outlaws choose their action sequentially in the order: 'Roberto, Filipe, Jolly, Roberto, Filipe, Jolly, ..., Roberto, Filipe, Jolly'.
- The state is a triple describing the color of the rabbit foot for each outlaw.
- Each action taken, no matter by which outlaw increments the time step.
- Observations, beliefs and policies only exist for an agent if it is his turn to take an action.
- Observations can only occur if they would not have lead to a final state in the previous time step.

Let's now formulate the problem:

- Write down the state space (per agent if necessary) $s \in \mathcal{S}$.
- Write down the action space (per agent if necessary) $a \in \mathcal{A}$.
- Write down the observation space (per agent if necessary) $o \in \mathcal{O}$.
- Write down the transition dynamics (per agent if necessary) $\mathcal{P}(s'|s,a) = \mathcal{P}^a_{s's}$.
- Write down the observation probabilities (per agent if necessary) $\mathcal{O}(o|s) = \mathcal{O}_{os}$.
- Is this a Dec-POMDP? Explain your answer.
- Write down the history for each time step, until the three are free. Assume the history is shared by all three of them.

b) **Back to the west continued** **[30 Bonus Points]**

Given the scenario and the assumptions of the previous question, we now want to compute the outlaws' reasoning for the first four time steps.

Let's start with the assumption that each observation space is the Cartesian product of two disjoint sets. $\mathcal{O}_i^x = \mathcal{O}_i^{x|V} \times \mathcal{O}_i^{x|A}, x \in \{R, F, J\}, i :=$ time step.

- What does $\mathcal{O}_i^{x|V}$ represent? What parts of an observations does it contain?[1]
- What does $\mathcal{O}_i^{x|A}$ represent? What parts of an observations does it contain?[1]
- Which of the two, $\mathcal{O}_i^{x|V}, \mathcal{O}_i^{x|A}$, changes because of x and which one because of i, i.e., is unique in x or in i.
- What is the meaning of the probability $p(o_x|s), o_x \in \mathcal{O}_i^x, s \in \mathcal{S}$?
- Why are these terms equivalent $p(o_x|s) = p\left(o_{x|V}|s\right) p\left(o_{x|A}|s\right), o_x \in \mathcal{O}_i^x, o_{x|V} \in \mathcal{O}_i^{x|V}, o_{x|A} \in \mathcal{O}_i^{x|A}, s \in \mathcal{S}$?

All three outlaws, make their decisions based on their current belief sate $b_i^x(s|o), s \in \mathcal{S}, o \in \mathcal{O}_i^x$. As we know form the lecture the policy for each outlaw is therefore defined as $a \sim \pi_i^x\left(a|b_{i|o}^x\right), a \in \mathcal{A}^x$, where $b_{i|o}^x$ is a vector of probabilities describing the belief state of x at time step i for the observation o.

- Why can we describe the belief state $b_i^x(s|o)$ as a vector $b_{i|o}^x$?

We begin with the first time step. It's Robertos turn.

- Derive Robertos belief state, $b_1^R(s|o), s \in \mathcal{S}, o \in \mathcal{O}_1^R$, for the first time step.
- Considering the given scenario, write down a table showing Robertos belief state for the first time step. The table dimensions should be [# possible states x # possible observations]

The text mentions that "each of them would immediately respond if he would be a 100% sure about the color of his rabbit foot."

- What kind of policy is described?
- Write down Robertos policy $\pi_1^R\left(a|b_1^R\right), a \in \mathcal{A}^R$ for the first time step, given his belief state $b_1^R(s|o)$. It is sufficient to write down the action for each possible observation.
- Considering the scenario, which action did Roberto choose? What observations could he possibly have made?

We are also interested in Robertos policy given the actual state $\pi_1^R(a|s), a \in \mathcal{A}^R, s \in \mathcal{S}$

- How is $\pi_1^R(a|s)$ related to $\mathcal{O}_1^{R|A}$?
- How can we compute $\pi_1^R(a|s)$?
- Derive $\pi_1^R(a|s)$. Consider the assumptions we made about the observation space.

[1] OK, this might not be obvious, so I put some *hints* in the text. Check the story again and keep your *eyes* and *ears* open...

Lets continue with the next time step. It is Filipes turn now.

- Derive Filipes belief state, $b_2^F(s|o), s \in \mathcal{S}, o \in \mathcal{O}_2^F$, for the second time step. Why did we compute $\pi_1^R(a|s)$ before?
- Considering the given scenario. Write down a table showing Filipes belief state for the second time step. The table dimensions should be [# possible states x # possible observations]
- Write down Filipes policy $\pi_2^F\left(a|b_2^F\right), a \in \mathcal{A}^F$ for the second time step, given his belief state $b_2^F(s|o)$. It is sufficient to write down the action for each possible observation.
- Considering the scenario, which action did Filipe choose? What observations could he possibly have made?
- What influence did Robertos previous action have on Filipes decision?
- Derive $\pi_2^F(a|s)$. Don't forget the assumptions we made about the observation space.

This concludes Filipes turn. Jolly is next.

- Derive Jollys belief state,$b_3^J(s|o), s \in \mathcal{S}, o \in \mathcal{O}_3^J$, for the third time step.
- Write down a table showing Jollys belief state for the third time step.
- Write down Jollys policy $\pi_3^J\left(a|b_3^J\right), a \in \mathcal{A}^J$ for the second time step, given his belief state $b_3^J(s|o)$. The table dimensions should be [# possible states x # possible observations]
- Considering the scenario, which action did Jolly choose? What observations could he possibly have made?
- What influence did Robertos and Filipes previous actions have on Jollys decision?
- Derive $\pi_3^J(a|s)$. Do I really need to mention that you should consider the assumptions we made about the observation space at the beginning of the question...

In time step number four, Roberto takes an action again.

- Wait... is there a fourth step? If yes explain why. If not explain anyways.

Problem 4.3 Programming Exercises (Bonus)

After spending the last night on partying (even though you were supposed to learn for your IMAS course) you find yourself waking up on a dark corridor. Unfortunately, the part of your brain that was supposed to tell you what has happened during the last ten hours was seemingly replaced by a little goblin with a large hammer - causing you some serious headache. In an attempt to deal with this new situation, you look around and quickly identify two doors as well as the only light source for the corridor - a computer screen attached to a dusty pc. Suddenly you hear a voice...

- *"Hi my name is Taka the Tiger, nice to "meat" you! Just kidding, today must be your lucky day! Usually I would just sit behind one of these doors and eat you as soon as you open it. However, after seeing a quite disturbing documentation last week, I decided to become vegan. So let's do it a bit differently today. Once the game is started, the computer will prompt you in fixed intervals which door you would have opened. Enter a '1', if you would have opened the left door, or a '3' if you would have opened the right door. For every time I would have gotten you, you'll have to pay 100€ and for every time you would have escaped, I'll pay you 10€. If you don't enter a number in time, you will have to pay 1€ as punishment. After 100 prompts, the game will end and I will tell you how much money you will get or have to pay. Unfortunately, lately, my stomach is often growling and I'm afraid that you might hear me if you get close to the walls. In order to compensate for that I will change the rooms via my secret tunnel with a probability of roughly 70%. Since you are looking kinda puzzled, I'll give you some time before starting the game."*

What do you want to do?

- Continue reading on page 6: If you want to inspect the pc
- Jump to page 7: If you want to leave the corridor

While inspecting the pc you stumble upon a little script with the cryptic name 'mcts'. After opening it you find several familiar variables and terms, ('POMDP', 'horizon', 'states', ...) and suddenly it strikes you 'It's Monte-Carlo Tree Search for POMDPs, I heard about it in the lecture but didn't quite understand how it works out!'. You figure out, that by getting the script to run, you can kill to birds with one stone: learn for IMAS and make some extra money!
Examining the code, you find out that it iteratively builds a tree by performing four steps in a loop:

a) **select** finds a node of the tree (corresponding to a history) for which not all actions are represented in the tree yet. It also samples a state as well as the reward for getting to the selected node.

b) **expand** adds a new action to the selected node and returns the sampled reward of the corresponding action.

c) **simulate** starts a rollout from the freshly created node and returns encountered reward.

d) **update** is used for improving the statistics of all encountered nodes (selected and expanded).

However, you also notice that some function are not fully implemented yet. Fortunately, the most tricky part of the selection phase is already solved. The fact that observations can not be chosen is dealt with by using separate nodes for representing histories and actions. Selecting the history resulting from a given action is done based on the sampled state and the observation probabilities. Selecting the action for a given history should apparently be done by using UCB. However, this part of the code seems to be missing. Furthermore, there is no rollout-policy implemented yet.

a) **The Vegan-Tiger Problem** **[6 Bonus Points]**

- First you have to enter the correct parameters for the POMDP. You assume that you can't hear the Tiger while standing in front of the computer. If you decide to listen on the wall, you will always hear the tiger, but you only get the door right with a probability of 85%. Also you would not make it back to the pc in time and miss one prompt.
- Now you have to fix HistoryNode.select() in order to select the next action based on its UCB score. Don't worry about the coefficient for the exploration-exploitation trade-off for now.
- Last but not least, a rollout policy has to implemented at HistoryNode.rollout(). Keep it simple and use a uniform policy.

b) **Tuning the parameters** **[2 Bonus Points]**
After fixing the code, you should now be able to use the script for suggesting your first action. However, depending on how you chose the exploration-exploitation coefficient, the results might be quite unsatisfactory. If too much weight is put on exploration, all children of a given HistoryNode will get a roughly equal number of visits. If too much weight is put on exploitation, however, good actions might be wrongly dismissed if their average reward is lower than their expected reward.

- Find a reasonable value for the exploitation-exploration tradeoff by trial-and-error. Use rootNode.getBestPath().printHistory() to find the most-likely start of the game.

c) **The Not-So-Vegan-Anymore-Tiger problem** **[0 Bonus Points]**
You think 'Screw you, Tiger! I need to learn for IMAS and don't have time for any games' and try the left door. Bad choice! After opening the door Taka the tiger immediately jumps at you bareing his teeth.
"*You don't want to play with me??? How dare you!*", were the last words you heard.