Reinforcement Learning: Tutorial 14

Partial observability and Exam recap

Week 7
University of Amsterdam

Milena Kapralova
October 2024

Check-in

- How is it going?
- How is the reproducible research assignment going?
- How is exam revision?
- If you have any feedback so far, please mail me at *m.kapralova@uva.nl*

Outline

1. Admin
2. Partial observability exercise
3. Previous years' exams & Q&A Session

Admin

- The reproducible research assignment deadline is today @ 17:00
- Reminder: exam is next Tuesday, 9-12, @ Piet Heinkade 27 (**not SP**)
- Any questions?

Tutorial 14 Overview

1. Partial observability exercise
2. Q&A Session

Tutorial 14 Overview

1. Partial observability exercise
    - Question 13.1
2. Q&A Session

Q 13.1 *Exam Question: Partially Observable MDPs

1. In POMDPs, we usually maintain an internal state s based on the interaction history. Two ways to do that include maintaining a belief state or using frame stacking. Give an advantage and a disadvantage of belief states over frame stacking. (In case it is relevant, assume the number of possible latent system states x is much larger than the dimensionality of observations times the number of stacked frames.)

Q 13.1 *Exam Question: Partially Observable MDPs

1. In POMDPs, we usually maintain an internal state s based on the interaction history. Two ways to do that include maintaining a belief state or using frame stacking. Give an advantage and a disadvantage of belief states over frame stacking. (In case it is relevant, assume the number of possible latent system states x is much larger than the dimensionality of observations times the number of stacked frames.)

   1. **Advantages:**
      - Belief states are Markov → use them to compute optimal policies,
      - belief states are interpretable,
      - belief states are reasonably compact (s has as many dimensions as x has states),
      - belief state can be computed recursively without memorizing the history.

Q 13.1 *Exam Question: Partially Observable MDPs

1. In POMDPs, we usually maintain an internal state s based on the interaction history. Two ways to do that include maintaining a belief state or using frame stacking. Give an advantage and a disadvantage of belief states over frame stacking. (In case it is relevant, assume the number of possible latent system states x is much larger than the dimensionality of observations times the number of stacked frames.)

2. **Disadvantages:**
   - Belief updates are harder to implement than frame stacking,
   - only for discrete states,
   - underlying models are needed,
   - underlying models can be difficult to learn,
   - belief updates are computationally expensive.

Q 13.1 *Exam Question: Partially Observable MDPs

2. While in POMDPs, we have separate latent system states x, observations o and internal agent states s, in MDPs we have $s = x = o$. In general, frame stacking does not result in an internal state that is a Markov function of a history. Is this also the case when the environment is an MDP? Explain why, using the criterion for internal Markov states in your answer.

Q 13.1 *Exam Question: Partially Observable MDPs

2. While in POMDPs, we have separate latent system states x, observations o and internal agent states s, in MDPs we have $s = x = o$. In general, frame stacking does not result in an internal state that is a Markov function of a history. Is this also the case when the environment is an MDP? Explain why, using the criterion for internal Markov states in your answer.

In MDPs, the environment state (x) is directly observed ($x = o$). So only the last frame is already a Markov state. If we stack a couple of frames, that only adds information but s is still Markov. So: it is not the case that frame stacking result in a non-Markov internal state.

Q 13.1 *Exam Question: Partially Observable MDPs

2. While in POMDPs, we have separate latent system states x, observations o and internal agent states s, in MDPs we have $s = x = o$. In general, frame stacking does not result in an internal state that is a Markov function of a history. Is this also the case when the environment is an MDP? Explain why, using the criterion for internal Markov states in your answer.

From the definition of Markov state we can see that any time the internal state is encountered, the probability of future state is the same. Since the stack includes x, and we already know that x contains enough all information about future system behavior (x is a Markov state), the stack is so as well.

WELL DONE!

Tutorial 14 Overview

1. Policy gradient methods: REINFORCE exercises
2. Previous years' exams & Q&A Session

Previous years' exams

- Ask away!

Q&A Session

- Ask anything about the ERL assignment or the course

That's it!

Thank you for joining the tutorial sessions and good luck on the exam