Reinforcement Learning: Tutorial 13

Planning and learning

Week 7
University of Amsterdam

Milena Kapralova
October 2024

Check-in

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

Outline

1. Admin

2. Planning and learning exercises

3. Paper intro + Ask anything about the ERL assignment

Admin

- **We don't have tutorial tomorrow but Thursday, 13-15, @ D1.113**
- Reminder that the reproducible research assignment deadline is Thursday @ 17:00
- Any questions?

Tutorial 13 Overview

1. Planning and learning exercises
2. Paper intro + Ask anything about the ERL assignment

Tutorial 13 Overview

1. Planning and learning exercises
    - Questions 12.1-12.2
2. Paper intro + Ask anything about the ERL assignment

Theory Intermezzo: Types of models

In general, we distinguish between three types of systems we can have that tries to imitate the real environment:

1. A **full** or **distributional** model is a full description of all transition probabilities and rewards.
   - Typically the proxy simulator learned in model-based RL.
2. A **sample** or **generative** model can be viewed as a black-box simulator, where given any state *s* and action *a*, it can sample *a* reward r_t and a next state s'.
3. A **trajectory** or **simulation** model can simulate whole episodes, but is not able to start at any state and action. This is for example the case for a physical model where we cannot start with an arbitrary velocity.
   - Typically assumed in model-free RL.

Q 12.1 *Exam Question: Model based learning (partial)

1. Different types of models are possible:
   - sample/generative model
   - full/distributional model
   - simulation/episodic model

   Which of these model types can be used to do value iteration?

Q 12.1 *Exam Question: Model based learning (partial)

1. Different types of models are possible:
    - sample/generative model
    - full/distributional model
    - simulation/episodic model

   Which of these model types can be used to do value iteration?

   Full/distributional model only, as value iteration is a dynamic programming method where we need to know the full model of the transition probabilities.

Q 12.1 *Exam Question: Model based learning (partial)

2. Which of the models can be used to generate samples from the on-policy distribution?
   - Both simulation/episodic model and full/distributional model can be used
   - Full/distributional model only
   - Sample/generative model only
   - Both simulation/episodic model and sample/generative model can be used
   - All model types can be used

Q 12.1 *Exam Question: Model based learning (partial)

2. Which of the models can be used to generate samples from the on-policy distribution?
    - Both simulation/episodic model and full/distributional model can be used
    - Full/distributional model only
    - Sample/generative model only
    - Both simulation/episodic model and sample/generative model can be used
    - All model types can be used

All model types can be used.

Q 12.1 *Exam Question: Model based learning (partial)

3. Consider the following statements:
   a) Learning with a model becomes increasingly attractive when computational resources get more expensive.
   b) Learning with a model becomes increasingly attractive when gathering data from the real system gets cheaper/faster.

   - a) is false, but b) is true
   - a) and b) are both false
   - a) and b) are both true
   - a) is true, but b) is false

Q 12.1 *Exam Question: Model based learning (partial)

3. Consider the following statements:
   a) Learning with a model becomes increasingly attractive when computational resources get more expensive.
   b) Learning with a model becomes increasingly attractive when gathering data from the real system gets cheaper/faster.
   - a) is false, but b) is true
   - a) and b) are both false
   - a) and b) are both true
   - a) is true, but b) is false

   a) and b) are both false.

Q 12.2 *Exam Question: MCTS and AlphaGo

1. How does MCTS manage the exploration/exploitation trade off during the search?

Q 12.2 *Exam Question: MCTS and AlphaGo

1. How does MCTS manage the exploration/exploitation trade off during the search?

   By using the Upper Confidence Bound (UCB) criterium which functions as an exploration bonus that gets lower with each visit. Note that actions are selected according to:

$$\pi(s) = arg\ max_a \left[Q(s,a) + c\sqrt{\frac{\ln(N(s))}{N(s,a)}} \right]$$

   where $N(s)$ is the number of visits to state s, $N(s,a)$ denotes the number of times that action a has been selected in state s, and the number $c > 0$ controls the degree of exploration.

Q 12.2 *Exam Question: MCTS and AlphaGo

2. AlphaGo can be seen as an adaption of Monte Carlo Tree Search which uses Neural Networks. Briefly explain the two functions of the Neural Network(s), compared to 'standard' MCTS.

Q 12.2 *Exam Question: MCTS and AlphaGo

2. AlphaGo can be seen as an adaption of Monte Carlo Tree Search which uses Neural Networks. Briefly explain the two functions of the Neural Network(s), compared to 'standard' MCTS.

(1) Evaluate / estimate the value / winning probability of a state / position, and
(2) Define a prior / initial probability distribution over the actions / moves (to guide the search).

Q 12.2 *Exam Question: MCTS and AlphaGo

3. Why do we expect the tree search policy to be better than the 'raw' Neural Network policy?

Q 12.2 *Exam Question: MCTS and AlphaGo

3. Why do we expect the tree search policy to be better than the 'raw' Neural Network policy?

   The raw neural network policy is limited by the capacity of the neural network / due to function approximation error / task is hard for raw policy. The tree search policy can improve it performing lookahead search (which can focus on the current state).

WELL DONE!

Tutorial 13 Overview

1. Planning and learning exercises
2. Paper intro + Ask anything about the ERL assignment

Paper intro: Deterministic PG's (Silver et al., 2014)

- What do authors bring? For example, theorems?
- What do the authors solve with their idea? In terms of efficiency or performance?
- Do authors integrate their idea with the already existing frameworks well?
- Environments, comparison methods, hyperparams & fairness, quantities measured, procedure, reporting spread, clarity and interpretability
- The report should not be a research proposal but a critical review of another research paper

Ask anything about the reproducible research assignment

- Is everything clear?

That's it!

Good luck with the HW and see you on Thursday