Theoretical exercise 8 **17. Jan. 2023**

Reinforcement Learning

The solutions will be discussed in the tutorial session

19. Jan 2023, 4-6 p.m. in lecture hall 5901.EG.051

For questions regarding this exercise sheet, please contact: `cosmin.bercea@tum.de`[1]
For general questions, please contact: `course.aim-lab@med.tum.de`

1. (2.5 points) Exercise 1: Multiple choice questions
   (a) (0.5 points) Which of the following is true about reinforcement learning?
      A. Reinforcement learning is a type of unsupervised learning
      B. The agent learns by being rewarded or penalized for its actions
      C. The agent receives explicit instructions on what actions to take
      D. The agent's goal is to minimize the error of its predictions

   (b) (0.5 points) Which of the following are characteristics of an actor-critic algorithm?
      A. It uses two separate networks, the actor and the critic
      B. The actor generates policies while the critic evaluates the policies
      C. It uses the Q-learning update rule
      D. It uses the policy gradient update rule

   (c) (0.5 points) In the Bellman equation for the Q-function, what is the term that represents the maximum expected future reward?
      A. The immediate reward
      B. The discount factor
      C. The Q value of the next state
      D. The action value of the next state

   (d) (0.5 points) Which of the following is a limitation of Q-learning when applied to high dimensional state spaces?
      A. The Q-table becomes too large to store in memory
      B. The algorithm may converge to a suboptimal policy
      C. The algorithm may become too slow to be practical
      D. All of the above

[1] Some of the questions were generated using AI language models, i.e., chatGPT (https://chat.openai.com/chat)

(e) (0.5 points) Which of the following is a limitation of DQNs?

A. They are prone to overfitting
B. They are sensitive to the choice of hyperparameters
C. They have difficulty learning from rare events
D. All of the above

2. (7 points) Exercise 2: Multi-armed bandits in healthcare. Consider a clinical trial where three treatments (X, Y and Z) are being tested on a patient population. Each treatment is represented as an arm of a multi-armed bandit. The goal is to determine which treatment is the most effective. Apply an algorithm using ϵ-greedy action selection, sample average action-value estimates, and initial estimates of $Q_1(X) = 1$, $Q_1(Y) = 0$, and $Q_1(Z) = 0$. Suppose the initial sequence of actions and reward is: $A1 = X, R1 = -2; A2 = Y, R2 = 1; A3 = Z, R3 = 2; A4 = Z, R4 = -1; A5 = X, R5 = 5$.
Hint: Use the incremental update rule:

$$Q_{t+1}^{(a)} = Q_t^{(a)} + \frac{1}{t}(R_t^{(a)} - Q_t^{(a)}) \tag{1}$$

where $Q_t^{(a)}$ is the Q-value of action a at time point t, and $R_t^{(a)}$ is the reward received at time-point t after taking action a.

(a) (2 points) Calculate for each time step $t \in [1, 5]$ the action values $Q_t^{(a)}$ for all possible treatments.

(b) (2 points) On some of these time steps ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possible have occurred?

(c) (1.5 points) Name 3 limitations of ϵ bandits.

(d) (1.5 points) ϵ bandits use a parameter ϵ to control the balance between exploration and exploitation. This can lead to inflexible exploration, where the agent may explore too much or too little at different stages of learning. What is an alternative way to deal with the exploration-exploitation dilemma? How does this method differ from ϵ explorations?

3. (5.5 points) Exercise 3: Deep Reinforcement Learning for Landmark Detection.
Consider the problem of locating anatomical landmarks in a 3D CT image. A kind radiologist annotated the desired landmarks in a subset of the data for you. Now you would like to develop a deep learning method to find those landmarks automatically in the remaining data using reinforcement learning.

(a) (2.5 points) Design a possible approach, especially give information about the set of actions, set of states, reward function, environment, Q function, termination criteria, and neural network. There is not a unique solution to this problem.

(b) (1.5 points) How can you effectively handle the trade-off between including enough context in the state representation (e.g., extracted features from a bounding box) for accurate learning, and keeping the state size manageable for efficient computation when dealing with 3D data in reinforcement learning?

(c) (1.5 points) How would you handle the scenario when the desired anatomical landmark is not present in some of the images due to limited field-of-view, resections, occlusions, or other factors in a 3D CT image localization using reinforcement learning?