Exercise Week 8 **7. July 2022**

Reinforcement Learning

Please submit your solutions (via Moodle) until 7. July, 4 p.m.

For questions regarding this exercise sheet, contact Veronika Zimmer:
`veronika.zimmer@tum.de`

Exercise 1

[3 points]

Consider a k-armed bandit problem with $k = 4$ actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is

$A_1 = 1, R_1 = -1, A_2 = 2, R_2 = 1, A_3 = 2, R_3 = -2, A_4 = 2, R_4 = 2, A_5 = 3, R_5 = 0.$

(a) Calculate for each time steps $t \in [1, 5]$ the action values $Q_t(a)$ for all actions a.

(b) On some of these time steps the 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 possibly have occurred?

Exercise 2

[8 points]

Execute for this exercise the notebook `Ex_Reinforcement_Learning.ipynb`, inspect the results and answer the following questions:

(a) Briefly describe the *exploration-exploitation* dilemma. How is this addressed in the class `MultiArmedBandit`?

(b) Run the experiments (with 1000 test runs (episodes) of the Multi-armed bandit with $\epsilon = 0$, $\epsilon = 0.1$, $\epsilon = 0.01$). Describe and discuss the results which are displayed in Figure 1.

(c) Run the experiments (with $k = 4$ and mu ='sequence').

(i) Compare Figure 2 to Figure 1. Which differences do you observe?

(ii) Compare how often the correct action is taken in each of the three models (illustrated in Figure 3). Describe and discuss the results.

(iii) Figure 4 shows the average estimated rewards and the true reward for each arm of the three bandits. Compare and discuss the three subplots.

(d) An improved version of the ϵ-*greedy* action selection method is ϵ-*decay* action selection (implemented in class `MultiArmedBandid_decay`. Here, ϵ is dependent on the time step t, e.g., $\epsilon(t) = \frac{1}{1+t\beta}$ with $\beta < 1$ (use $\beta = \frac{1}{k}$).

(i) In which sense does this improve the ϵ-*greedy* action selection?

(ii) Run the experiments (with $k = 4$ and three bandits). Discuss the results presented in Figure 5 and 6.

Exercise 3

[0 points: optional]

Sample averaging methods are appropriate for stationary bandit problems (the reward probabilities do not change over time). Many reinforcement learning problems are non-stationary, and in such cases it makes sense to give more weight to recent rewards than to long-past rewards, for example by using a constant step-size parameter α, which gives the following update rule:

$$Q_{t+1} = Q_t + \alpha \left(R_t - Q_t \right).$$

Show that this results in Q_{t+1} being a weighted average of past rewards and the initial estimate Q_1.

Exercise 4

[5 points]

Consider the problem of locating anatomical landmarks in a 3D MRI 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. Design a possible approach, especially give information about the set of actions, set of states, reward function, environment, Q function, neural network, etc. There is not a unique solution to this problem.