DS-GA 3001.005
NYU Center for Data Science

Reinforcement Learning

Homework 01

Exercice 1 (40 points)

Introducing Reinforcement Learning:

- **1.1 (10 points)**: List three key differences between a Supervised Learning problem and a Reinforcement Learning problem
- **1.2 (5 points)**: What is the name of the function that maps specific states to specific actions?
- **1.3 (5 points)**: What is a possible difference between what an RL agent observes and what defines its state at a given time step?
- **1.4 (5 points)**: What is the difference between a deterministic policy and a stochastic policy?
- **1.5 (5 points)**: For each of the following scenarios, determine whether the given policy is deterministic or stochastic. Briefly justify your answer.

  (a) Scenario: A robot moves in a grid world.
  Policy: The robot always moves right if possible; otherwise, it moves up.

  (b) Scenario: A self-driving car is approaching an intersection.
  Policy: The car chooses to turn left with 70% probability and right with 30% probability.

  (c) Scenario: A chess engine selects a move given a specific board state.
  Policy: The engine always picks the move with the highest evaluation score.

  (d) Scenario: A robot is navigating through rough terrain.
  Policy: The robot chooses an action based on a probability distribution over safe movements.

  (e) Scenario: A thermostat controls room temperature.
  Policy: It turns the heater on when the temperature drops below 18°C and off when it rises above 22°C.

- **1.6 (5 points)**: Define in plain english what is $v(s)$, and what is $q(s, a)$?
- **1.7 (5 points)**: What is the Bellman equation for $v(s)$ given a discount factor γ?

Exercice 2 (20 points)

The Multi-Armed Bandit problem:

- **2.1 (10 points)**: For an ϵ-greedy action selection method with only two actions and $\epsilon = 0.5$, what is the probability that the greedy action is selected?
- **2.2 (5 points)**: What is the key difference between a Bandit *vs.* a more general RL problem?
- **2.3 (5 points)**: What is the key difference between a *Contextual* Bandit and a more general RL problem?

Exercice 3 (20 points)

In the *10-armed Testbed* Bandit problem presented in the lecture (also detailed in Chapter 2, Section 2.3 of the Sutton & Barto book):

- **3.1 (15 points)**: Which action-selection method, amongst the three options listed below, will perform best in the long run in term of cumulative reward and probability of selecting the best action? These methods are visualized in Fig 2.3 in the book and slide 14 in the lecture.

  1. ϵ-greedy with $\epsilon = 0.1$
  2. ϵ-greedy with $\epsilon = 0.01$
  3. ϵ-greedy with $\epsilon = 0$ (greedy policy) and initial q-values set to -10

  Explain your answer.

- **3.2 (5 points)**: Suppose we have a 10-armed bandit problem similar to above but where the reward for each of the 10 arms is deterministic (= always same reward for a given arm) and in the range (-5, +5). Same question as 3.1 but for the following action-selection methods:

  1. ϵ-greedy with $\epsilon = 0.1$
  2. ϵ-greedy with $\epsilon = 0.01$
  3. ϵ-greedy with $\epsilon = 0$ (greedy policy) and initial q-values set to -10
  4. ϵ-greedy with $\epsilon = 0$ (greedy policy) and initial q-values set to $+10$

  Explain your answer.

Exercice 4 (20 points) [Programming]

Produce the exact code (not pseudo code) to do the following:

- Import the necessary libraries, including `gymnasium`.
- Register the Atari environments using `ale_py`.
- Create an instance of the Gym environment called *Seaquest*.
- Run the environment for 20 episodes. For each episode:
  - Initialize the environment and assign the starting state to a variable.
  - Loop for 1000 steps or until the episode terminates:
    - Sample a random action from the space of possible actions.
    - Step forward in the environment and assign the next state, observed reward, termination boolean variable, truncation boolean variable, and auxiliary information to respective variables.
    - If the episode terminates early (either by termination or truncation), stop and proceed to the next episode.
- Close the environment.