TecNM Campus Ciudad Guzmán
Maestría en Ciencias Computacionales
Materia: Aprendizaje por Refuerzo
Docente: Dr. Julio Alberto García Rodríguez

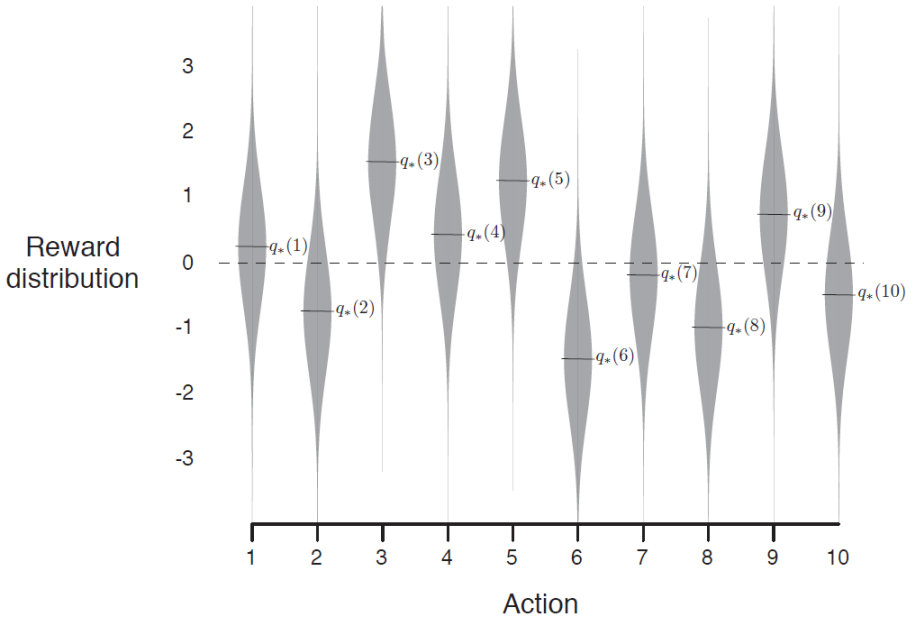
Class Project U2. Part 01.

The k-armed bandit problem

The k-armed bandit problem is a classic reinforcement learning challenge that explores the fundamental trade-off between exploration and exploitation.

1. In this project, you must propose a scenario to solve an optimization problem involving 10 actions, using the theoretical approach of the k-armed bandit problem. Remember that each action gives a reward. The scenario you propose could be, for example: a medical experiment, the actions of a slot machine, a video game, a movie application, an advertising website, etc. You will describe your scenario in one paragraph.

2. Explain in one paragraph and plot the stationary probability distribution of the reward for each possible action. Remember that this distribution does not change and is unknown to the agent.



3. Create your first agent that estimates action-values using the sample average method with a random selection policy (random exploration). You must present the following plots for 500, 1000, 5000 and 10000 simulation steps.
    a. Numbers of times each action was tried.
    b. Estimation of action values over time simulation steps.
    c. Accumulated reward over time simulation steps.
    d. Percentage of optimal action for 2000 runs.

4. Now create your second agent, called greedy, which estimates action values using the argmax() function to exploit the action of greatest value (exploitation-only). You must present the following plots for 500, 1000, 5000 and 10000 simulation steps.
    a. Numbers of times each action was tried.
    b. Estimation of action values over time simulation steps.
    c. Accumulated reward over time simulation steps.
    d. Percentage of optimal action for 2000 runs.

Note: If multiple actions have the same maximum value, choose one of them at random.

5. Make a comparative framework of the performance of your greedy agent when the action values are initialized in different ways: at zero, randomly, and optimistically. Choose a fixed simulation step.
    a. Numbers of times each action was tried.
    b. Estimation of action values over time simulation steps.
    c. Accumulated reward over time simulation steps.
    d. Percentage of optimal action for 2000 runs.

Note: For an optimistic initialization, it is suggested that the initial value should be greater than any possible reward in the distribution. Example: rewards between 0 and 1, initialize at $q = 5$. The aim of this is to ensure that the agent's values remain optimistic as it progresses through the simulation steps.

The project will be delivered in .ipynb format and its .pdf version.