Reinforcement Learning Practical, Assignment 1

Matthia Sabatelli, Nicole Orzan

November 12, 2021

Instructions

In this assignment, you will implement two versions of the multi-armed bandit problem. You can use whichever programming language you feel the most comfortable with. At the end of the assignment, you will have to submit a report explaining your results as well as the code used for your experiments. The multi-armed bandit problems that you have to implement are:

- The Gaussian bandit: a multi-armed bandit in which the reward obtained from each action is sampled from a normal distribution.
- The Bernoulli bandit: a multi-armed bandit in which the reward obtained from each action is sampled from a Bernoulli distribution (each arm has probability p to return 1 and $1-p$ probability to return 0).

As seen throughout the second theoretical lecture, the goal of the agent for each bandit problem is to learn an optimal policy π^*, i.e. the action that brings the maximum reward. This goal is reached through learning. You have to create a set of N randomly generated k-armed bandit problems for both bandit scenarios, where both k and N are parameters of your choice. For each of those problems, you will then train an agent with the different exploration methods that we have seen in lecture 2. These methods are the following:

- Greedy and ϵ-greedy
- Optimistic initial values
- Softmax policy
- Upper-Confidence Bound
- Action Preferences

You will perform one experiment for each exploration method. Each experiment will consist of a number of training steps T. At the end of each training run, we expect your agent to have learned to recognize the action (or actions) that allow it to obtain the maximum possible reward. Furthermore, each experiment will be repeated for a certain number of times (for example $N = 1000$). You will measure the learning performance of the agent by monitoring the average reward it obtains, as well as the percentage of times the agent chooses the best action. Note that you might have to fine-tune the hyperparameters that govern the learning process for every experiment, which could change across the different exploration methods.

Report

In your report you should include:

- A brief description of each algorithm you have used, with their respective equations;
- A description of the experimental setup used, the definition of the reward functions as well as which hyperparameters you chose for each algorithm;

- For both problems, you will plot your results in two clear figures: one containing the trend of the rewards during learning and one containing the percentage of times the best action is selected;
- A section where you compare and discuss the performance of the different exploration algorithms. If there are one or more algorithms with significantly better performance than all the others, explain why.

The **deadline** for submitting this assignment is the 6th of December 2021.