Lecturer: Rafael F. Cunha, University of Groningen
Course: Reinforcement Learning Practical WBAI015-05 (2024/2025)

Assignment 1 November 18, 2024

Exploration strategies in the Multi-Arm bandit testbed.

This is the course's first assignment. We will use the exploration strategies together with the 10-arm bandit testbed to help you get familiar with the main tools used in a reinforcement learning project.

1. Implement a 10-armed bandit in Python, each arm following a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with $\sigma = 1$ and $\mu \sim \mathcal{N}(0, 1)$.
2. Conduct 500 different learning runs, each initializing a distinct 10-armed bandit setup.
3. Produce two plots and corresponding tables:
   (a) Plot 1. Plot the average reward learning curve per step for the ϵ-greedy algorithm with $\epsilon = 0$, and $\epsilon = 0.1$.
   - Utilize the sample-averaged method for the action-value updates.
   - Initialize the starting estimation of action values at 0 for each action a in the action set $\mathcal{A}$ ($Q_0(a) = 0, \forall a \in \mathcal{A}$).
   - Run the simulation for 1000 steps.

   The plot should resembles Figure 1. Note that your version should have legend, smoothing, mean, and standard deviation, as in Figure 2.

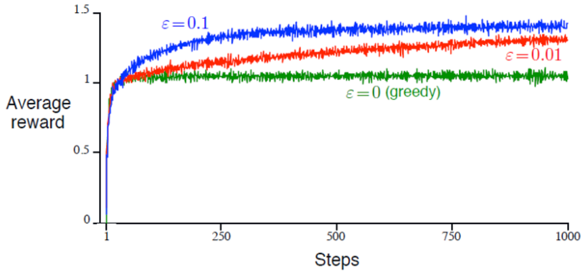

Figure 1: Example of a learning curve.

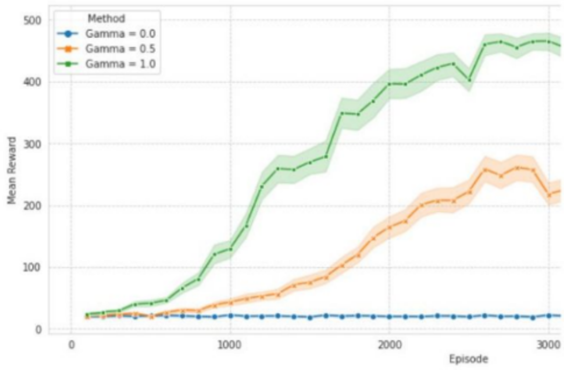

Figure 2: Example of smoothing, legend, mean, and standard deviation in a learning curve.

   (b) Plot 2. Compare the average rewards over the first 1000 steps across different exploration strategies with varying hyperparameters.
   - Analyze the following algorithms:
     – Greedy with optimistic initialization and weighted-average method. (Q_0, $\alpha = 0.1$)
     – ϵ-greedy with sample-averaged method. (ϵ)
     – Upper-Confidence Bound with sample-averaged method.(UCB) (c)
     – SoftMax with sample-averaged method.(τ)
     – Gradient Bandit (Action preferences with baseline) (α)
   - Include a plot with the average and standard deviation of rewards.

   The plot should resemble Figure 3. Note that your version should have legend, smoothing, mean, and standard deviation, as in Figure 2, and that the example does not provide a curve for the SoftMax policy.
4. This assignment is designed to familiarize you with various tools utilized in a reinforcement learning project:
   (a) Conduct experiments on Hábrók.

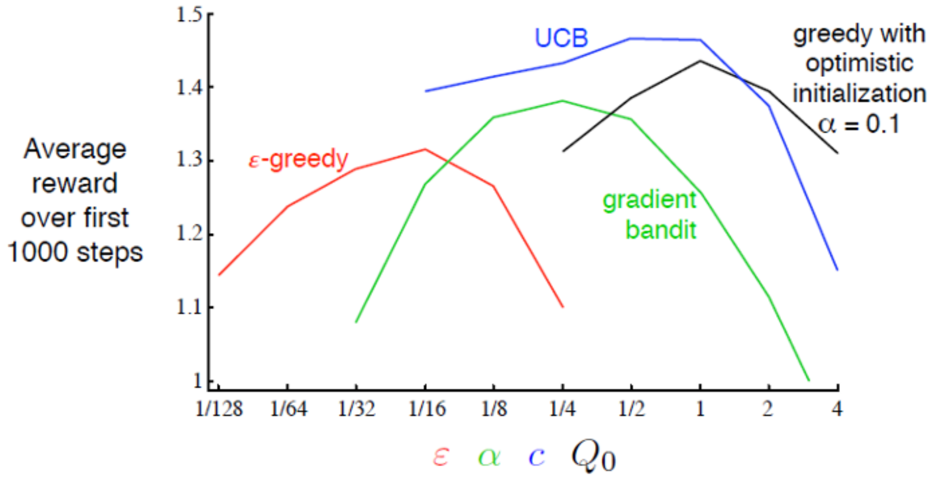

Figure 3: Example of curve for the second plot of the assignment.

(b) Manage file transfers between Hábrók and GitHub using git commands.

(c) Utilize TensorBoard for data visualization during training.

(d) Develop a method to convert data into a Pandas dataframe and plot it using Seaborn or Matplotlib, considering data smoothing, mean, and standard deviation.

(e) Each plot should include a caption inspired by “Experimentation in RL” from Tutorial 3 of the theoretical RL course.

(f) Ensure reproducibility by selecting a consistent starting seed.

(g) Follow the “bachelor’s project” report template.

5. The report should include:

   (a) A section describing the algorithms.

   (b) A section detailing the plots, tables, and discussions of results.

   (c) A conclusion section.

   (d) The report should not exceed five pages.

For additional support, refer to the materials provided in the theoretical reinforcement learning course:

- Theory: Lecture 1 or Chapter 2 of ”Reinforcement Learning: An Introduction”.
- Coding: Tutorial 2, Part 2.
- Hábrók usage: Tutorial 7, Part 2.
- Presenting statistical results: Tutorial 3, Part 2.
- Captions’ plot: Tutorial 3, Part 2.