COMP532-202324 Assignment 1

You need to solve each of the following problems. You must also include a brief report describing and discussing your solutions to the problems. Students can do the assignment in groups or individuals.

- This assignment is worth 15% of the total mark for COMP532
- 80% of the assignment marks will be awarded for correctness of results
- 20% of the assignment marks will be awarded for the quality of the accompanying report
- Students will do the assignment in groups
- The assignment marks will be awarded for correctness of results
- We expect 2-8 students in one group (it would be fine to have groups of 1 as well, but it is suggested to have groups of 5), please find your team members on your own.
- Only one single submission is needed for each group
- The same marks will be granted to all the members in the same group
- Please list all your group members (names, emails, student ids) and individual contributions in your submitted report

Submission Instructions

- **Deadline: 18 MAR 2024 17:00 (UK Time)**
- Send all solutions as a single PDF document containing your answers, results, and discussion of the results. Attach the source code for the programming problems as separate files.
- Submit your solution via Canvas.
- Penalties for late submission apply in accordance with departmental policy as set out in the student handbook, which can be found at https://intranet.csc.liv.ac.uk/student/msc-handbook.pdf and the University Code of Practice on Assessment, found at https://www.liverpool.ac.uk/media/livacuk/tqsd/code-of-practice-on-assessment/code_of_practice_on_assessment.pdf

Problem 1 (40 marks)

Implement a multi-armed bandit algorithm.
Similar to Figure 2.1 (two figures) in the Sutton & Barto book (Reinforcement Learning: An Introduction, see https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf), provide a solution for a n-armed bandit, where n can be 5 or 10 or 20. The total number of plays should be 2000. The solution should consider exploration and exploitation.

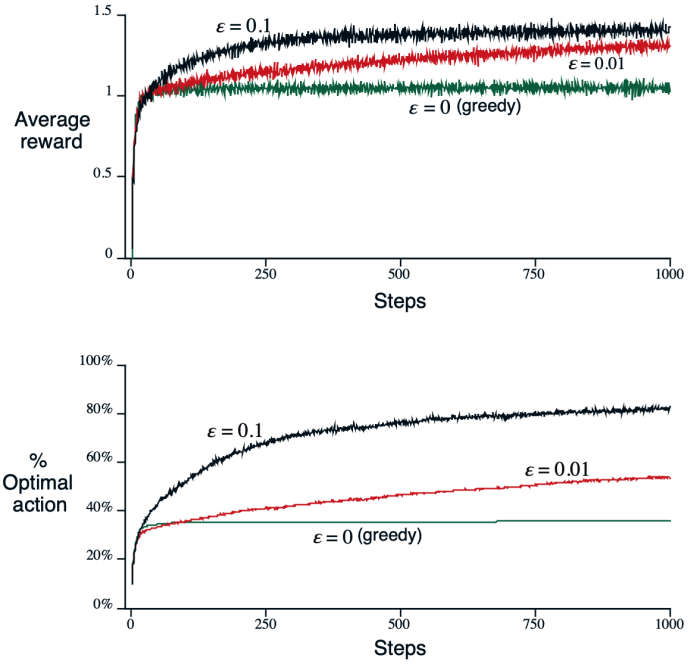

Figure 2.1: Average performance of ε-greedy action-value methods on the 10-armed testbed. These data are averages over 2000 tasks. All methods used sample averages as their action-value estimates. The detailed structure at the beginning of these curves depends on how actions are selected when multiple actions have the same maximal action value. Here such ties were broken randomly. An alternative that has a similar effect is to add a very small amount of randomness to each of the initial action values, so that ties effectively never happen.

Prepare a report explaining your solution and containing your results, and discussion of the results.

Attach the source code as separate files. For example, .ipynb - an ipython notebook file.

Problem 2 (30 marks)

Explain exploration and exploitation for multi-armed bandits.

Problem 3 (30 marks)

Explain action-value methods.