Problem Set 1

INF368A - Reinforcement Learning
Spring 2023

Part I.
Problems

In the first part (*Problems*), you are required to provide answers for all the *obligatory* questions, that is, questions marked by a star (*). The remaining questions are optional; however, you are invited to solve them because: (i) they are stepping-stones towards the obligatory questions; (ii) they are similar to the questions you will be asked at the oral exam.

1. Foundations

Exercise1 Characterizing distributions

Consider interacting with a slot-machine which returns the following amount of money R: with probabiliy 0.35 it returns nothing; with probability 0.3 it returns 5$; with probability 0.25 it returns 10$; with probability 0.1 it returns 20$.

1. Plot the distribution of $P(R)$?
2. Compute $E[R]$ and $Var[R]$?
3. What is the probability $P(R \geq 10\$)$?
4. What is the probability $P(R > 10\$)$?
5. If I play the slot-machine twice, what is the probability that the total reward is greater than 15$?

Consider now interacting also with a second slot-machine which returns the following amount of money S: with probabiliy 0.3 it returns nothing; with probability 0.35 it returns 5$; with probability 0.3 it returns 10$; with probability 0.05 it returns 20$.

6. Plot the distribution of $P(S)$?

7. If you were to choose which slot-machine to play, which one would you choose and why?
8. What is the distribution of $P(R, S)$?
9. What is the distribution of $P(S, R)$? Does it differ from $P(R, S)$?
10. (*) If I play both slot machine, what is the distribution of the sum of the rewards? How do you express it formally, and how does it differ from $P(R, S)$?
11. If I play the first slot machine with probability 0.4 and the second slot machine with probability 0.6, what is the distribution of my reward? How do you express it formally, and how does it differ from $P(R, S)$?
12. If I play both slot machines, and I receive 10$ in outcome from the first one, what is the probability that after playing the second slot machine I win less than 10$? How do you express it formally?
13. If I play both slot machines, and I receive 10$ in outcome from the first one, what is the probability that after playing the second slot machine I win at least 10$? How do you express it formally?
14. (*) If I play both slot machines, and I receive 10$ in outcome from the first one, what is the probability that after playing the second slot machine I win more than 20$? How do you express it formally?

Exercise2 Bayes' formula

Bayes' formula is an important identity in probability and statistics, stating:

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

.

1. (*) Using the two probability laws presented in the class, prove that Bayes' formula reduces to an identity.

Exercise3 Inferring distributions

Assume someone played one of the two slot machines in the first exercise above, and reported the following list of ten rewards $D_i = [10, 0, 5, 10, 5, 0, 0, 10, 0, 0]$.

1. Estimate $\hat{E}[D]$?
2. Which of the two slot machines do you expect the person was playing?

The same person plays more the same machine, and provide you with additional ten samples $D_j = \{10, 20, 0, 5, 10, 5, 10, 0, 20, 10\}$.

3. Update your estimate $\hat{E}[D]$ using the twenty sample.
4. (*) Does your opinion on which of the two slot machines the person was playing change?
5. (*) If the person plays once more (21st sample), do you have particular expectations on the reward that will be produced?

2. Basics of Reinforcement Learning

Exercise4 Modelling a RL Problem

Consider the following scenario. A company wants to develop an automatic floor-sweeping robot to be able to effectively clean house rooms. The idea is to train such a robot with reinforcement learning within the premises of the company, and you have been put in charge of the project.

1. Do you consider the project sensible?
2. (*) How would you define the main elements of your RL project (states, actions, rewards)?
3. Is this a planning, prediction, or control problem?

After satisfactorily training your robot, the management gets its hand on the robot of another firm and it asks you to compare your prototype against the competitor.

4. Do you have now a planning, prediction, or control problem?
5. What metric would you use to compare the robots?
6. (*) The competing robot seems to perform better, although it often hits tables and bookshelves, causing papers and mugs to crash on the floor. Do you think this might have something to do with the training of the robot? How would you explain its behaviour?

The management is concerned with your prototype being competitive enough, and so it considers hitting the less crowded market of factories instead of homes. Your trained robot will be deployed on factory floors instead of house floors.

7. (*) Do you think this choice sensible from a technical point of view? Is there any issue you think worth discussing with the management?

Exercise5 RL Assumptions

Consider the following scenario. A company is developing a software for streamlining the hiring process, and it decided to use RL to develop an artificial interviewing agent to optimize the whole process.

1. (*) Do you consider the choice of RL sensible? Discuss.

3. Multi-armed Bandits

Exercise6 MAB Assumptions

Consider the following scenario. Facing a new epidemics, the government is considering the deployment of three different vaccines. However, the success rate of these vaccines is still uncertain. Officials of the government heard that multi-armed bandits are an

effective tool to design the deployment of new medicines, and so they consider adopting the same model.

1. (*) Do you consider the choice of MABs sensible? Discuss.

Exercise7 Hoeffding's Inequality

You are tossing a biased coin C, and collected 10 samples such that $\hat{E}[C] = 0.5$.

1. (*) Use Hoeffding's Inequality to evaluate the probability you are underestimating the true expected value $E[C]$ by $0.01, 0.1$ and 0.5.

Part II.
Report

In the second part (*Report*), you are required to provide a report of maximum 2 pages describing the experiments you have run and analyzed to answer the *obligatory* questions, that is, questions marked by a star (*). The remaining questions are optional; however, you are invited to solve them because: (i) they are stepping-stones towards the obligatory questions; (ii) they are similar to the questions you will be asked at the oral exam. You will be evaluated only on your report. Review the documents on `coding tips` and `writing tips` for advices on how to tackle this part.

Exercise8 MAB Algorithms

Three new medicines A, B, C have been developed to fight headache, and a multi-armed bandit deployment has been prepared in order to learn which one is the most effective medicine in the most efficient way. Instatiate the environment `Bandits_one()` from `bandit.py` in order to solve this problem.

1. Implement the ϵ-greedy, decaying ϵ-greedy and UCB algorithm and use them to solve the MAB in the environment `Bandits_one()`.
2. Run your algorithms for 1000 episodes. What do you observe? What is the optimal medicine?
3. Repeat the above experiments 20 times and average your results. What do you observe? What is the optimal medicine?
4. (*) Plot the regret of the three algorithms averaged across the 20 repetitions. Does it match the asymptotic behaviour presented in class?
5. Observe how the regret of the three algorithms may change when tuning their hyperparameter ϵ, α, c.

A fourth medicine D is added to the available choices. Instatiate the environment `Bandits_two()` from `bandit.py` to interact with all four drugs A, B, C, D.

6. (*) How would you solve this new MAB problem? Would you restart from scratch?

A new set of three medicines X, Y, Z is now proposed in the environment `Bandits_three()` from `bandit.py`.

7. Use the algorithms that you have already implemented to solve this MAB problem.
8. Run your algorithms for 1000 episodes. Repeat it for 20 times. What do you observe? What is the optimal medicine?
9. (*) Consider the regret. How does the MAB of `Bandits_one()` compare with the MAB of `Bandits_three()`? Which MAB was easier to solve? Why so?

A study revealed that the three medicines X, Y, Z are very responsive to the gene G which can take two values 0 or 1. Before each episode, you are now provided with an observation of the gene value for the current patient. This is implemented in `Bandits_four()` from `bandits.py`.

10. Train two bandits, one for patients with gene $G = 0$ and one with gene $G = 1$.
11. Formalize and compute the expected rewards for each medicine X, Y, Z and for each value of the gene $0, 1$.
12. (*) Do you learn the same optimal action? How and why does the optimal action differ from the MAB of `Bandits_three()`?

Further study shows that the gene G can actually have a continuous activation value between 0 and 1. Also, clinical studies confirm that the response of the three drugs X, Y, Z to the gene G is linear.

13. (*) How would you estimate the expected reward of providing medicine X if the gene expression is 0.5?