Data 102, Spring 2024
Homework 5

Due: **5:00 PM** Friday, April 5, 2024

Submission Instructions

Homework assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

Written Portion:

- Every answer should contain a calculation or reasoning.
- You may write the written portions on paper or in LaTeX.
- If you type your written responses, please make sure to put it in a markdown cell instead of writing it as a comment in a code cell.
- Please start each question on a new page.
- It is your responsibility to check that work on all the scanned pages is legible.

Code Portion:

- You should append any code you wrote in the PDF you submit. You can either do so by copy and paste the code into a text file or convert your Jupyter Notebook to PDF.
- Run your notebook and make sure you print out your outputs from running the code.
- It is your responsibility to check that your code and answers show up in the PDF file.

Submitting:

You will submit a PDF file to Gradescope containing all the work you want graded (including your math and code).

- When downloading your Jupyter Notebook, make sure you go to File → Save and Export Notebook As → PDF; do not just print page from your web browser because your code and written responses will be cut off.
- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so. As a Berkeley student, you get free access to Adobe Acrobat, which you can use to merge as many PDFs as you want.
- Please see this guide for how to submit your PDF on Gradescope. In particular, for each question on the assignment, please make sure you understand how to select the corresponding page(s) that contain your solution (see item 2 on the last page).

Late assignments will count towards your slip days; it is your responsibility to ensure you have enough time to submit your work.

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, please include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

Simulation Study of Bandit Algorithms

In this problem, we evaluate the performance of two algorithms for the multi-armed bandit problem. The general protocol for the multi-armed bandit problem with K arms and n rounds is as follows: in each round $t = 1, \ldots, n$ the algorithm chooses an arm $A_t \in \{1, \ldots, K\}$ and then observes reward X_t for the chosen arm. The bandit algorithm specifies how to choose the arm A_t based on what rewards have been observed so far. In this problem, we consider a multi-armed bandit for $K = 2$ arms, $n = 50$ rounds, and where the reward at time t is $X_t \sim \mathcal{N}(A_t - 1, 1)$, i.e. $\mathcal{N}(0, 1)$ for arm 1 and $\mathcal{N}(1, 1)$ for arm 2.

(a) (4 points) Consider the multi-armed bandit where the arm $A_t \in \{1, 2\}$ is chosen according to the explore-then-commit algorithm (below) with $c = 4$. Let $G_n = \sum_{t=1}^{n} X_t$ denote the total reward after $n = 50$ iterations. Simulate the random variable G_n a total of $B = 2000$ times and save the values $G_n^{(b)}$, $b = 1, \ldots, B$ in a list. Report the empirical averaged regret $\frac{1}{B}\sum_{b=1}^{B}\left(50\mu^* - G_n^{(b)}\right)$ (where μ^* is the mean of the best arm) and plot a normalized histogram of the rewards.

Algorithm 1 Explore-then-Commit Algorithm

input: Number of initial pulls c per arm
for $t = 1, \ldots, cK$: **do**
  Choose arm $A_t = (t \bmod K) + 1$
end
Let $\hat{A} \in \{1, \ldots, K\}$ denote the arm with the highest average reward so far.
for $t = cK + 1, cK + 2, \ldots, n$: **do**
  Choose arm $A_t = \hat{A}$
end

(b) (4 points) Consider the multi-armed bandit where the arm $A_t \in \{1, 2\}$ is chosen according to the UCB algorithm (below) with $c = 4$, $n = 50$ rounds. Repeat the simulation in Part (a) using the UCB algorithm, again reporting the empirical averaged regret and the histogram of $G_n^{(b)}$ for $b = 1 \ldots B$ for $B = 2000$. How does the empirical averaged regret compare to your results from part (a)?
Note: If $T_A(t)$ denote the number of times arm A has been chosen (up to and including time t) and $\hat{\mu}_{A,t}$ is the average reward from choosing arm A (up to and including t), then use the upper confidence bound $\hat{\mu}_{A,T_A(t-1)} + \sqrt{\frac{2\log(20)}{T_A(t-1)}}$. Note also that this algorithm is slightly different than the one used in the lab and lecture as we are using an initial exploration phase.

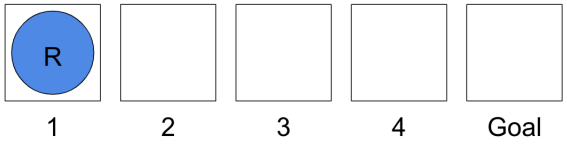
Algorithm 2 UCB Algorithm

input: Number of initial pulls c per arm
for $t = 1, \ldots, cK$: **do**
| Choose arm $A_t = (t \bmod K) + 1$
end
for $t = cK + 1, cK + 2 \ldots$: **do**
| Choose arm A_t with the highest upper confidence bound so far.
end

(c) (1 point) Compare the distributions of the rewards by also plotting them on the same plot and briefly justify the salient differences.

Markov Decision Process for Robot Soccer

A soccer robot R is on a fast break toward the goal, starting in position 1. From positions 1 through 3, it can either shoot (S) or dribble the ball forward (D). From 4 it can only shoot. If it shoots, it either scores a goal (state G) or misses (state M). If it dribbles, it either advances a square or loses the ball, ending up in state M.



In this Markov Decision Process (MDP), the states are 1, 2, 3, 4, G, and M, where G and M are terminal states. The transition model depends on the parameter y, which is the probability of dribbling successfully (*i.e.*, advancing a square). Assume a discount of $\gamma = 1$. For $k \in \{1, 2, 3, 4\}$, we have

$$\begin{aligned} \Pr(G \mid k, S) &= \frac{k}{6} \\ \Pr(M \mid k, S) &= 1 - \frac{k}{6} \\ \Pr(k+1 \mid k, D) &= y \\ \Pr(M \mid k, D) &= 1 - y, \\ R(k, S, G) &= 1 \end{aligned}$$

and rewards are 0 for all other transitions.

(a) (3 points) Denote by V^π the value function for the specific policy π. What is $V^\pi(1)$ for the policy π that always shoots?

(b) (4 points) Denote by $Q^*(s, a)$ the value of a q-state (s, a), which is the expected utility when starting with action a at state s, and thereafter acting optimally. What is $Q^*(3, D)$ in terms of y?

(c) (3 points) For what range of values of y is $Q^*(3, S) \geq Q^*(3, D)$? Interpret your answer in plain English.