Overview

Submit your writeup, including any code, as a PDF via gradescope.[1] We recommend reading through the entire homework beforehand and carefully using functions for testing procedures, plotting, and running experiments. Taking the time to reuse code will help in the long run!

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 r_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 $r_t \sim \mathcal{N}(A_t - 1, 1)$, 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 $G_n = \sum_{t=1}^{n} r_t$ 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) average pseudoregret $\frac{1}{B} \sum_{b=1}^{B} \left(50\mu^* - G_n^{(b)}\right)$ (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) average pseudoregret and the histogram of $G_n^{(b)}$ for $b = 1 \ldots B$ for $B = 2000$. How does the pseudoregret compare to your results from part (a)?

[1] In Jupyter, you can download as PDF or print to save as PDF

Note: If $T_A(t)$ denote the number of times arm A has been chosen (before time t) and $\hat{\mu}_{A,t}$ is the average reward from choosing arm A (up to time 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 lab and lecture as we are using an initial exploration phase.

Algorithm 2 UCB Algorithm

```
input: Number of initial pulls $c$ per arm
for $t = 1, \dots, cK$ : do
    Choose arm $A_t = (t \bmod K) + 1$
end
for $t = cK + 1, cK + 2 \dots$ : 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.