Overview

Submit your writeup including all code and plots as a PDF via Gradescope. We recommend reading through the entire homework beforehand and carefully using functions for testing procedures, plotting, and running experiments. Taking the time to test, maintain, and 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 hand-written answers are legible, as we may deduct points otherwise.

1 Simulation Study of Bandit Algorithms

In this problem, we evaluate the performance of three 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 $K = 2$ 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) (2 points) Consider the multi-armed bandit where the arm $A_t \in \{1, 2\}$ is chosen according to the ε-greedy algorithm (below) with $c = 4$ and $\varepsilon = 1/2$. 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 $B = 200$ times and save the values $G_n^{(b)}$, $b = 1, \ldots, B$ in a list. Report the expected value $\frac{1}{B}\sum_{b=1}^{B} G_n^{(b)}$ of the total reward and plot a normalized histogram of the rewards.

Algorithm 1 ε-Greedy Algorithm

input: Number of initial pulls c per arm; additional exploratory fraction $\varepsilon \in (0, 1)$
for $t = 1, \ldots, cK$: **do**
  Choose arm $A_t = (t \bmod K) + 1$
end
for $t = cK + 1, cK + 2, \ldots, n$: **do**
  Toss coin with success probability ε:
  if *success* **then**
    Choose arm A_t uniformly at random
  else
    Choose arm A_t with the highest average reward so far.
  end
end

(b) (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 $c = 4$. Repeat the simulation in Part (a) using the explore-then-commit algorithm, again reporting the expected total reward and the histogram of $G_n^{(b)}$ after $n = 50$ rounds. How does the reward G_n compare to your results from part (a)? Compare both the average and the overall distribution.

Algorithm 2 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

(c) (4 points) Consider the multi-armed bandit where the arm $A_t \in \{1, 2\}$ is chosen according to the UCB algorithm (below) with $c = 4$. Repeat the simulation in Part (a) using the UCB algorithm, again reporting the expected total reward and the histogram of $G_n^{(b)}$ after $n = 50$ rounds. How does the reward G_n compare to your results from part (b)? Compare both the average and the overall distribution. *Note:* if $T_A(t)$ denote the number of times arm A has been chosen (up to and including iteration t) and $\hat{\mu}_{A,t}$ is the average reward from choosing arm A (i.e. the average of the first t rewards from arm A), 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 as we are using an initial exploration phase.

Algorithm 3 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

2 Regret of Explore-then-Commit

In this problem, we analyze the regret of the Explore-then-Commit algorithm for the multi-armed-bandit (MAB) problem.

We consider a stochastic MAB problem with a set of $K = 2$ arms $\mathcal{A} = \{1, 2\}$. Recall from Lab 8 that each arm $A \in \mathcal{A}$ is associated with a reward distribution $X_A \sim \mathbb{P}_A$, with corresponding mean $\mu_A = \mathbb{E}[X_A]$. **We will assume throughout this problem that the first arm has higher average reward, i.e. $\mu_1 > \mu_2$.** At each round $t = 1, \ldots, n$ our algorithm chooses an arm $A_t \in \mathcal{A}$ and receives a corresponding reward $X_{A_t}^{(t)} \sim \mathbb{P}_{A_t}$, independent of all previous rewards.

If we knew arm 1 has higher average reward we would choose $A_t = 1$ each round in order to maximize the expected total reward. In practice, however, we do not know which arm is better since the means $\{\mu_1, \mu_2\}$ are unknown. The expected reward of our algorithm will always be less than $n\mu_1$, and we quantify the price we pay for not knowing the better arm via the *regret*

$$R_n := n\mu_1 - \mathbb{E}\left[\sum_{t=1}^{n} X_{A_t}^{(t)}\right].$$

In this problem we analyze the regret of the explore-then-commit (ETC) algorithm from problem 1(b). Recall from Algorithm 2 that ETC proceeds in two phases. In the exploration phase, each arm $A \in \mathcal{A}$ is pulled c times in order to produce an estimate $\hat{\mu}_A = \frac{1}{c}\sum_{t \le cK : A_t = A} X_A^{(t)}$ of the mean reward for that arm. In the commit phase, i.e. for every $t > cK$, we choose $A_t = \hat{A}$, where $\hat{A} := \arg\max_{a \in \mathcal{A}} \hat{\mu}_A$ is the apparent best arm at the end of the exploration phase.

In the first part of our analysis, we evaluate the probability that we incorrectly identify arm 2 as the best arm, i.e. $\mathbb{P}(\hat{A} = 2)$.

(a) (2 points) Assume each reward is in the unit interval $[0, 1]$, i.e. $0 \le X_A \le 1$ for $A \in \{1, 2\}$. Show that

$$\mathbb{P}\left(\hat{A} = 2\right) \le \exp\left(-\frac{c\Delta^2}{2}\right),$$

where $\Delta = \mu_1 - \mu_2$. *Hint:* apply Hoeffding's inequality from Lecture 18.

In parts (b) and (c), we write the regret in terms of the probability $\mathbb{P}(\hat{A} = 2)$.

(b) (3 points) Let m denote the number of times arm 2 has been pulled, up to and including time n. Show

$$R_n = \Delta\mathbb{E}[m].$$

Hint: Start from the following:

$$\begin{aligned} R_n := n\mu_1 - \mathbb{E}\left[\sum_{t=1}^{n} X_{A_t}^{(t)}\right] &= \mathbb{E}\left[\sum_{t=1}^{n}\left(\mu_1 - X_{A_t}^{(t)}\right)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{I}\{A_t = 1\}\left(\mu_1 - X_1^{(t)}\right)\right] + \mathbb{E}\left[\sum_{t=1}^{n}\mathbb{I}\{A_t = 2\}\left(\mu_1 - X_2^{(t)}\right)\right]. \end{aligned}$$

Note also that for all t, A_t is independent of $X_A^{(t)}$ for $A \in \{1, 2\}$.

(c) (1 point) Show that if $n > 2c$, then

$$\mathbb{E}[m] = c + (n - 2c)\mathbb{P}\left(\hat{A} = 2\right)$$

Hint: If $n > 2c$, both arms are pulled deterministically c times during the first $2c$ rounds. Afterward, an arm is only pulled if it is the one we have committed to.

In parts (d) and (e), we finalize our bound on the regret R_n.

(d) (1 point) Show that

$$R_n \leq \Delta \left(c + (n - 2c) \exp\left(-\frac{c\Delta^2}{2} \right) \right).$$

Hint: combine parts (a)-(c).

(e) (3 points) Suppose you knew the sub-optimality gap Δ. Solve for (and report) a value of c which guarantees that:

$$\exp\left(-\frac{c\Delta^2}{2} \right) \leq \frac{1}{n}.$$

For this number of exploratory pulls c, what is the upper bound on the regret from part (d)? Your answer should be in terms of n and Δ. Does this bound grow linearly in n, or does it do better (i.e. is it sublinear)?