Zmh6339
AI & Machine Learning, CS-UH 3260
Professor Keith Ross

Ziad Hassan
January 31, 2024

Assignment 1

Part 1: Continuous Bandit Algorithm

Part 2: Theory

a) *Proof.*

$$\begin{aligned}
P(\text{greedy action}) &= 1 - P(\text{not greedy action}) \\
&= 1 - P(\text{not greedy selection } AND \text{ not greedy action}) \\
&= 1 - P(\text{not greedy selection}) \cdot P(\text{not greedy action} \mid \text{not greedy selection}) \\
&= 1 - \epsilon \cdot \left(\frac{k-1}{k}\right) \qquad {}^{*}\ \textit{Since there is only } \textbf{\textit{one}} \textit{ greedy action} \\
&= 1 - \epsilon \cdot \left(1 - \frac{1}{k}\right) \\
&= 1 - \epsilon + \frac{\epsilon}{k}
\end{aligned}$$

■

b) i) *Proof.*

To determine the probability that the greedy action was chosen for the first time at time T, we need to consider that it was not chosen at any time before T, and that it was chosen at time T.
Thus, the following equation should be quantified:

$$P(\text{greedy at } T) = P(\text{not greedy before } T) \cdot P(\text{greedy at } T)$$

Therefore, the probability that the greedy action was chosen for the first time at time T is:

$$\begin{aligned}
P(\text{greedy at } T) &= P(\text{not greedy before } T) \cdot P(\text{greedy at } T) \\
&= \left(1 - 1 + \epsilon - \frac{\epsilon}{k}\right)^{T-1} \cdot \left(1 - \epsilon + \frac{\epsilon}{k}\right) \\
&= \left(\epsilon - \frac{\epsilon}{k}\right)^{T-1} \cdot \left(1 - \epsilon + \frac{\epsilon}{k}\right)
\end{aligned}$$

■

ii) *Proof.*

To get the expected number of steps, $\mathbb{E}[T]$, until the the greedy action is chosen for the first time is a sum over all possible time steps, each weighted by its probability of being the

first time the greedy action is chosen.
Thus, the following equation should be quantified:

$$\mathbb{E}[T] = \sum_{t=1}^{\infty} t \cdot P(\text{greedy at } t)$$

Following, the equation is

$$\begin{aligned}\mathbb{E}[T] &= \sum_{t=1}^{\infty} t \cdot P(\text{greedy at } t)\\ &= \sum_{t=1}^{\infty} t \cdot \left(\epsilon - \frac{\epsilon}{k}\right)^{t-1} \cdot \left(1 - \epsilon + \frac{\epsilon}{k}\right)\end{aligned}$$

It can be observed that the above equation is a geometric series, which can be simplified to the following:

$$\begin{aligned}\mathbb{E}[T] &= \sum_{t=1}^{\infty} t \cdot \left(\epsilon - \frac{\epsilon}{k}\right)^{t-1} \cdot \left(1 - \epsilon + \frac{\epsilon}{k}\right)\\ &= \frac{1}{\left(\epsilon - \frac{\epsilon}{k}\right)^{t-1} \cdot \left(1 - \epsilon + \frac{\epsilon}{k}\right)}\end{aligned}$$

The above simplifciation is valid due to the definiton of the expected value of a geometric series.

■

c) i) *Proof.*

Firstly, denote $\max(q(1), q(2), \ldots, q(10))$ as q_*.
The algorithm will choose the greedy selection, q_* with probability $1 - \epsilon$, and a non-greedy selection with probability ϵ.
Moreover, the expected value of the an action during non-greedy selection is the average of all the action values, which is $\frac{\sum_{i=1}^{10} q(i)}{10}$.
Therefore, the long-run reward is

$$R = (1 - \epsilon) \cdot q_* + \epsilon \cdot \frac{\sum_{i=1}^{10} q(i)}{10}$$

■

ii) *Proof.*

Since $q(1), q(2), \ldots, q(10)$ are i.i.d. random variables, the expected value of a greedy action becomes $\mathbb{E}[q_*] = b$.
Moreover, since $q(a)$ is $\mathcal{N}(0, 1)$, the expected average of all the action values becomes $\mathbb{E}\left[\frac{\sum_{i=1}^{10} q(i)}{10}\right] = 0$.
Therefore, the long-run reward is

$$\begin{aligned}R &= (1 - \epsilon) \cdot \mathbb{E}[q_*] + \epsilon \cdot \mathbb{E}\left[\frac{\sum_{i=1}^{10} q(i)}{10}\right]\\ &= (1 - \epsilon) \cdot b + \epsilon \cdot 0\\ &= (1 - \epsilon) \cdot b\end{aligned}$$

■