EXERCISE 7

DR. VICTOR UC CETINA

1. ε-GREEDY METHOD ON THE 10-ARMED BANDIT PROBLEM

The ε-greedy is a strategy to balance the tradeoff between exploitation and exploration in reinforcement learning. The ε-greedy policy is the following:

$$a_t = \begin{cases} a^* \text{ with a probability } 1-\varepsilon, \\ \text{random action with probability } \varepsilon. \end{cases}$$

where

$$a^* = \arg\max_a Q_t(a)$$

and

$$Q_t(a) = \frac{r_1 + r_2 + \ldots + r_{k_a}}{k_a}.$$

Implement the ε-greedy algorithm for solving a 10-armed bandit problem with the following setup:

- $n = 10$ possible actions
- Each $Q * (a)$ is chosen randomly from a normal distribution: $\eta(0, 1)$
- Each r_t is also normal: $\eta(Q^*(a_t), 1)$
- 1000 plays
- Repeat the whole thing 2000 times and average the results

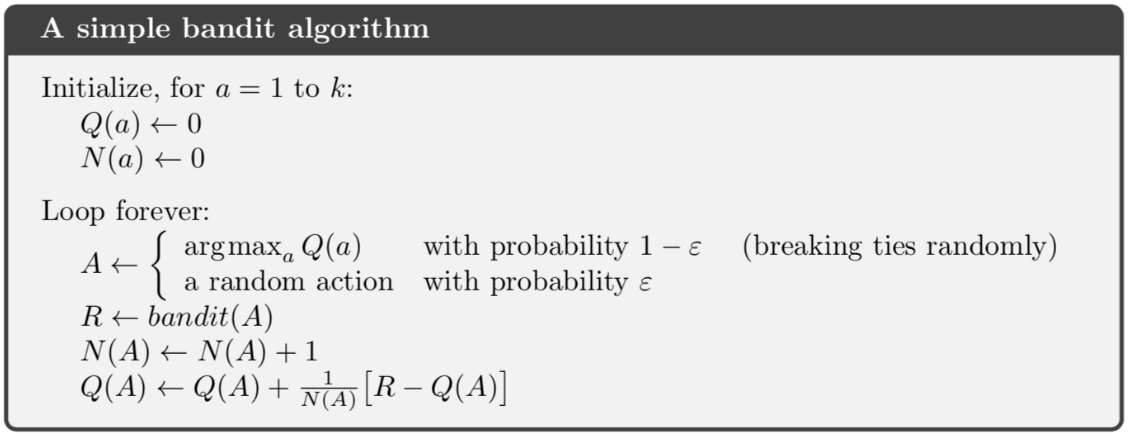
Run experiments with $\varepsilon = 0.1, \varepsilon = 0.01$ and $\varepsilon = 0.0$. Finally, plot the average curves for each value of ε. Plot also the average number of times that the optimal action was selected.

2. Hint

Implement the algorithm version provided in page 32 of the book by Sutton and Barto, available from:

`http://incompleteideas.net/book/RLbook2018.pdf`

The algorithm is provided here for your convenience.



A simple bandit algorithm

```
Initialize, for $a = 1$ to $k$:
    $Q(a) \leftarrow 0$
    $N(a) \leftarrow 0$

Loop forever:
    $A \leftarrow \begin{cases} \arg\max_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$
    $R \leftarrow bandit(A)$
    $N(A) \leftarrow N(A) + 1$
    $Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\big[R - Q(A)\big]$
```

3. Deadline

Please handle your report at most on January 28-30.