ε-greedy method on the 10-armed bandit problem

Author: Rodriguez Noh Santiago Miguel
Professor: Ph.D. Anabel Martin Gonzalez
Link to code: https://github.com/Santiagomrn/e-greedy.git

I. Introduction

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

$$f(x) = \begin{cases} a^* & \text{with a probability } 1-\varepsilon \\ & \text{random action with probability } \varepsilon \end{cases} \quad (1)$$

where

$$a^* = argmax Q_t(a) \quad (2)$$

and

$$Q_t(a) = \frac{r_1 + r_2 + ... + r_{ka}}{k_a} \quad (3)$$

II. Implement the ε -greedy algorithm

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.

A. Selected actions

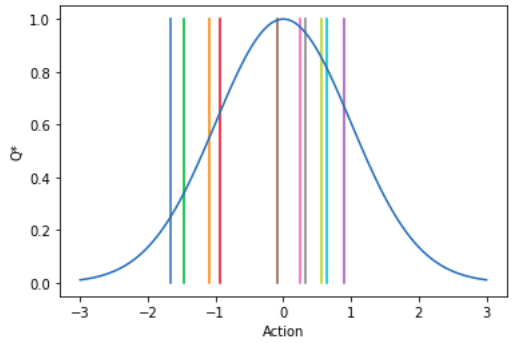

Fig. 1. Selected actions.

action	Q*(action)
-1.6632868303027664	0.25075935621140255
-1.1015853193623366	0.5451222988588382
-1.4720205885886015	0.33843531206700206
-0.9445856178933321	0.6401063178621635
0.890390307202549	0.6727392628577681
-0.09121436155683778	0.9958486110608391
0.24063434730783517	0.9714626617100937
0.3316409482801192	0.9464919065554036
0.571229286032379	0.8494625287524952
0.6427606132380028	0.8133688331783826

Fig. 2. Selected actions.

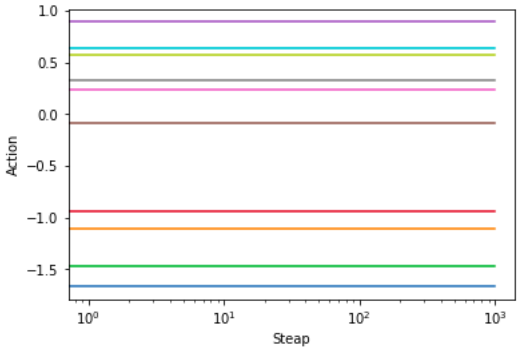

Fig. 3. Selected actions.

B. Results after 2000 iterations with $\varepsilon = 0.1$

The best actions were chosen 333,447 times on average.

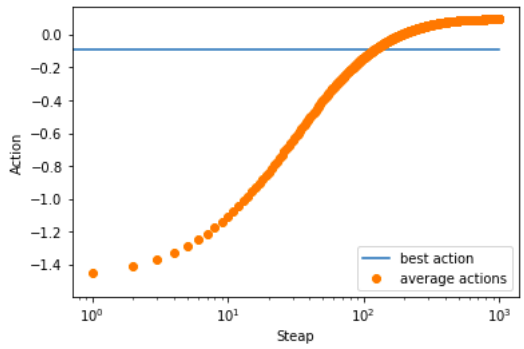

Fig. 4. Average actions.

C. *Results after 2000 iterations with* $\varepsilon = 0.01$

The best actions were chosen 192.1585 times on average.

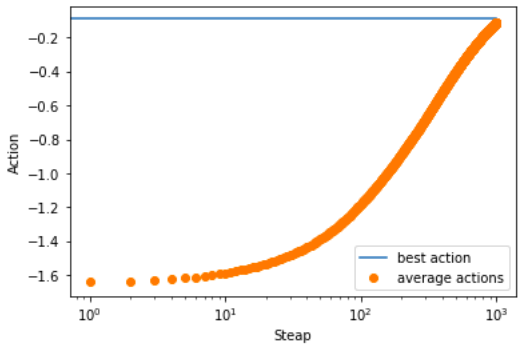

Fig. 5. Average actions.

D. *Results after 2000 iterations with* $\varepsilon = 0.00$

The best actions were chosen 0.0 times on average.

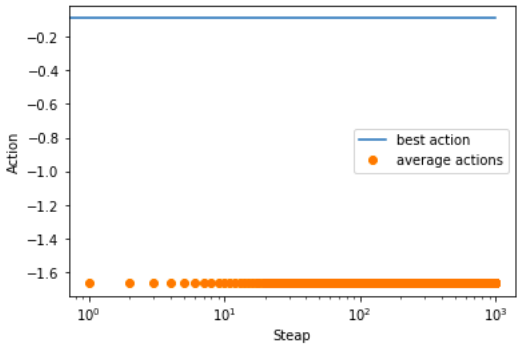

Fig. 6. Average actions.