Exercise 1 Bandits

[Avik Banerjee (3374885), Soumyadeep Bhattacharjee (3375428)]

Thursday, 23. April 2020 21:32

(1) Multi-armed bandits

(a) Two actions, $k=2$, $\epsilon = 0.5$

Probability that the greedy action is selected

$= (1-\epsilon) + \epsilon \times 0.5 = 0.5 + 0.25 = 0.75$

(b) In step 0:

$Q_1(1) = Q_1(2) = Q_1(3) = Q_1(4) = 0$

In step 1, $A_1 = 1$, $R_1 = 1$ ← possibly ϵ

$Q_1(1) = 1$, $Q_1(2) = Q_1(3) = Q_1(4) = 0$

In step 2, $A_2 = 2$, $R_2 = 1$ ← definitely ϵ

$Q_2(1) = 1$, $Q_2(2) = 1$, $Q_2(3) = 0$, $Q_2(4) = 0$

In step 3, $A_3 = 2$, $R_3 = 2$ ← possibly ϵ

$Q_3(1) = 1$, $Q_3(2) = 1.5$, $Q_3(3) = 0$, $Q_3(4) = 0$

In step 4, $A_4 = 2$, $R_4 = 2$ ← possibly ϵ

$Q_4(1) = 1$, $Q_4(2) = 1.7$, $Q_4(3) = 0$, $Q_4(4) = 0$

In step 5, $A_5 = 3$, $R_5 = 0$ ← definitely ϵ

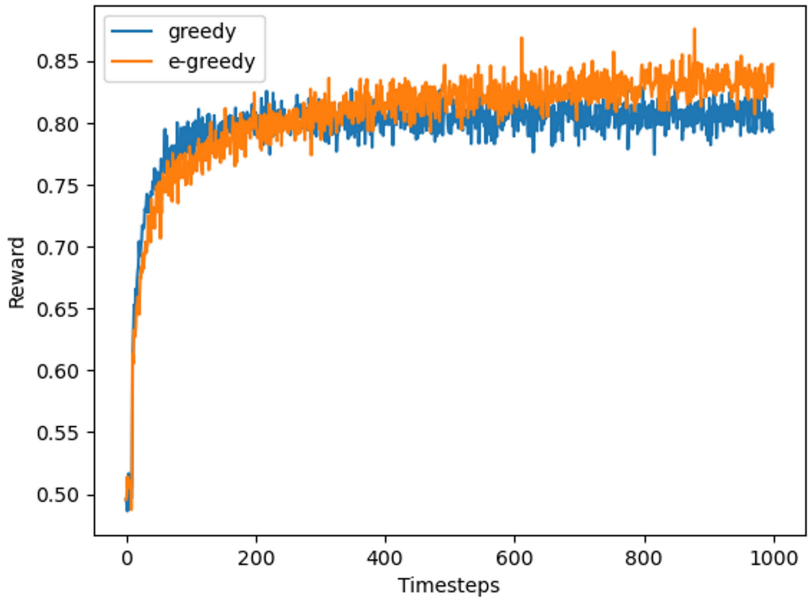
$Q_5(1) = 1$, $Q_5(2) = 1.7$, $Q_5(3) = 0$, $Q_5(4) = 0$

(1) Definitely ϵ on timesteps 2 and 5

(2) Possibly ϵ on steps 1, 3, 4.

(2) Action Selection Strategies :

(c)



The ϵ-greedy strategy performs better than the greedy strategy as it produces a higher average return. This occurs because the ϵ-greedy strategy enforces continued exploration of the action space. Thus, as the number of steps increases, every action will be sampled an infinite number of times and $Q_t(a)$ (estimate of the action-value at time t) will eventually converge to $q^*(a)$ (expected reward)

(d) For improved rewards :

- A decreasing epsilon value could be used : similar to the ϵ-greedy strategy except that the value of ϵ decreases with each epoch, resulting in highly explorative behaviour at the beginning and highly exploitative behaviour at the end.

- An adaptive ϵ-value can be used, where high

- An adaptive ϵ-value can be used, where high fluctuations in the value estimates lead to a high ϵ-value (high exploration, low exploitation) and low fluctuations lead to a low ϵ-value (low exploration, high exploitation).

Further improvements can be achieved by a softmax-weighted action selection in case of exploratory actions.