Reinforcement Learning

Worksheet 1

Stoyan Dimitrov
April 24, 2020

Bandits

Exercise 1

(1) At the beginning, every choice is non-epsilon, but $A_1 = 1$ was made.

$Q_2(1) = 1$, $Q_2(2) = Q_2(3) = Q_2(4) = 0$ but chosen action was $A_2 = 2$

$Q_3(1) = Q_3(2) = 1$, $Q_3(3) = Q_3(4) = 0$ both $A_3 = 1$ and $A_3 = 2$ can be chosen

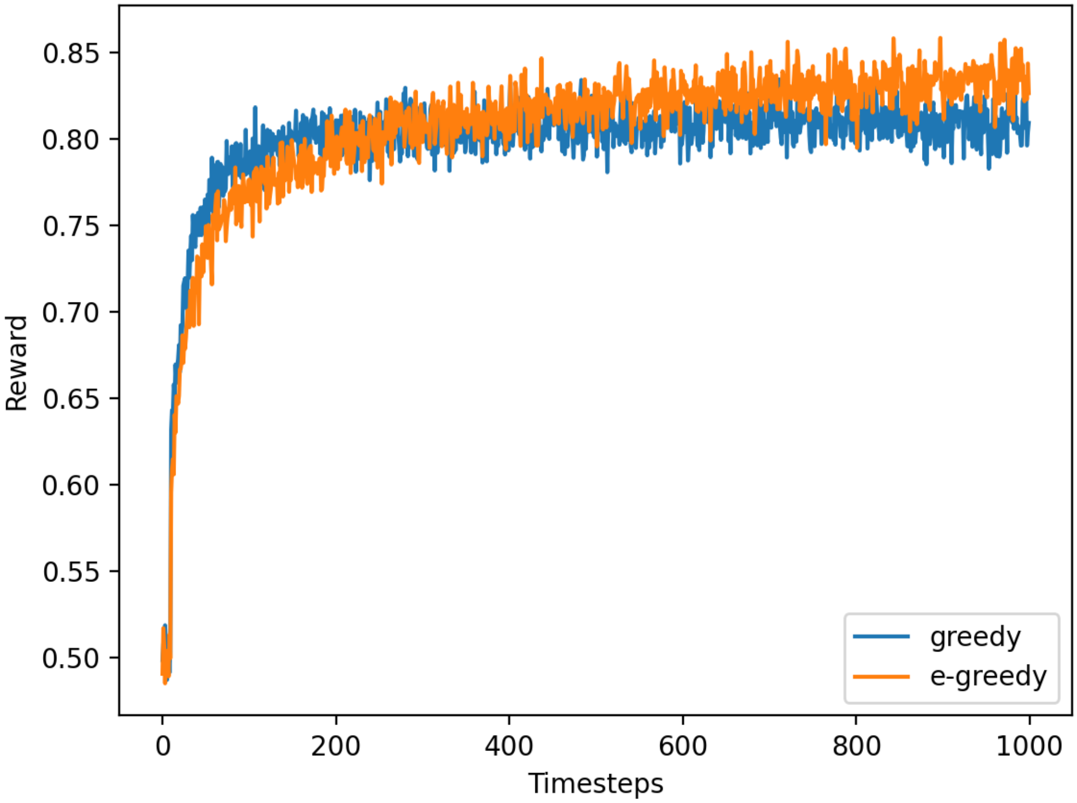
$Q_5(1) = 1, Q_5(2) = 5/3 = 1.7$ and $Q_5(3) = Q_5(4) = 0$ but chosen action was $A_5 = 3$

Definitely, choices 2 and 5 were ϵ choices

(2) Choices 1 and 3 could have been ϵ choices

Exercise 2

(3) The ϵ-greedy method improves slower, but to a better average reward:



(4) We can use initialization with optimistic values to force exploration on the initial stages. Also Upper-Confidence-Bound action selection is shown to perform better than ϵ-greedy search, because it takes into account the uncertainty of the value of the chosen action - more frequently chosen actions are have lower uncertainty.