10/10

Reinforcement Learning
Exercise 1 - Solution

Jonathan Schnitzler - st166934
ErickVillanuevaVillasenor - st190300
Erik Choquet - st160996

April 21, 2024

1 Multi-Armed Bandits

2p

Task 1a)

Either the greedy action is selected with a probability of $p = 0.5$ or with an additional probability of $p = 0.5 * 0.5$ the greedy action is picked randomly. Therefore the the probability of the greedy action is

$$p_{\text{greedy}} = 0.75$$

.

Task 1b)

A k-armed bandit problem with $k = 4$ with following rules

1. ϵ-greedy action selection, i.e. either choose the greedy action

$$A_t = \text{argmax}_a Q_t(a) \tag{1}$$

with probability $1 - \epsilon$ or choose a random action with probability ϵ

2. sample-average action-value estimates, i.e.

$$Q_t(a) = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}} = \frac{\text{sum of rewards of } a}{\text{number of } a} \tag{2}$$

3. initial estimates of $Q_1(a) = 0$ for all a

Following sequence of actions and rewards is observed:

Step	Action A_i	Reward R_{i+1}
1	1	1
2	2	1
3	2	2
4	2	2
5	3	0

Where the ϵ-case definetly occured

- Step 2: since $Q_2(1) = 1$ and $Q_2(2) = 0$, the greedy action is 1, but the action 2 was chosen.
- Step 5: since

$$Q_5(1) = 1 \tag{3}$$

$$Q_5(2) = \frac{5}{3} \tag{4}$$

$$Q_5(3) = 0, \tag{5}$$

2p

  the greedy action is $a = 2$, but the action 3 was chosen.

Where the ϵ-case possibly occured In all other steps it could be possible that the ϵ-case occured, since when the action is random, still the epsilon-greedy action could be chosen and for step 3 it is even ambiguous, which action is the greedy one.

2 Action selection strategies

Task 2c)

As can be observed in Figure 1 the best strategy over many timesteps is the ϵ-greedy strategy.

6/6

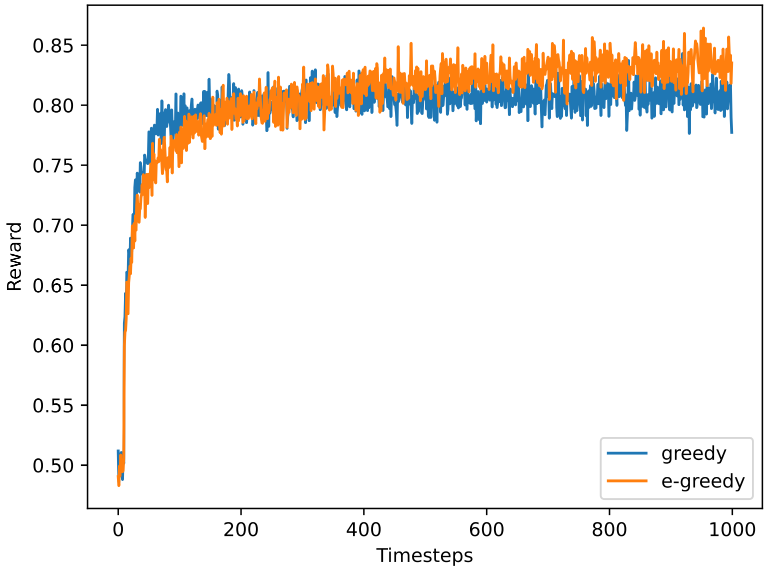

Figure 1:

Task 2d)

- Make the actions not random but according to the Q distribution. Therefore actions which are more likely to actually be the greedy action are chosen more often.
- Decrease the epsilon over time, so that the agent is more and more exploiting the environment.
- Make exploration dependent on timesteps, i.e. exploration is more encouraged for larger timesteps. For only ten timesteps maybe just exploit.