Robot Learning

Assingnment 1

Due Tuesday, April 24th, before class.

UNIVERSITÄT BONN AIS

RHEINISCHE   COMPUTER SCIENCE VI
FRIEDRICH-WILHELMS-   AUTONOMOUS
UNIVERSITÄT BONN   INTELLIGENT SYSTEMS

Prof. Dr. Sven Behnke   Endenicher Allee 19a

1.1)   Given is seven-armed bandit, as introduced in the lecture.
        The first arm shall sample its reward uniformly from the interval [–2, 3).
        The second arm shall sample its reward uniformly from [1, 4).
        The third arm shall sample its reward uniformly from the interval [2, 3).
        The fourth arm shall sample its reward uniformly from [–1, 5).
        The fifth arm shall sample its reward uniformly from [0, 4).
        The sixth arm shall sample its reward uniformly from [1, 4).
        The seventh arm shall sample its reward uniformly from [3, 7).

        What is the expected reward when actions are chosen uniformly?

        4 points

1.2)   Implement the seven-armed bandit from 1.1) and compute the sample average reward for 20
        uniformly chosen actions.
        Compare this to your expectation from 1.1).

        4 points

1.3)   Initialize $Q(a_i)$=0 and chose 2000 actions according to an ε-greedy selection strategy (ε=0.1).
        Update your action values by computing the sample average reward of each action recursively.
        For every 100 actions show the percentage of choosing arm 1, arm 2, arm 3, arm 4, arm 5, arm 6,
        and arm 7 as well as the resulting average reward.

        4 points

1.4)   Redo the experiment, but after 1000 steps sample the rewards of the fourth arm uniformly from
        [6, 8).
        Compare updating action values by computing the sample average reward of each action
        recursively (as done in 1.3) with using a constant learning rate α=0.01.
        For every 100 actions show the percentage of choosing arm 1, arm 2, arm 3, arm 4, arm 5, arm 6,
        and arm 7 as well as the resulting average reward.

        4 points

1.5)   Modify your implementation by using an optimistic initialization $Q(a_i)$=5 and a greedy action
        selection strategy, still using a constant learning rate α=0.01.
        For every 100 actions show the percentage of choosing arm 1, arm 2, arm 3, arm 4, arm 5, arm 6,
        and arm 7 as well as the resulting average reward.
        Compare this to your result from 1.4).

        4 points