UNIVERSITÄT BONN
RHEINISCHE FRIEDRICH-WILHELMS-UNIVERSITÄT BONN
COMPUTER SCIENCE VI AUTONOMOUS INTELLIGENT SYSTEMS
Prof. Dr. Sven Behnke
Endenicher Allee 19a

Robot Learning

Assignment 2

Due Tuesday, May 5th, before class.

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

2.2) Implement the seven-armed bandit from 2.1)!
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.

6 points

2.3) Consider a student taking an exam, which consists of *k* tasks.
For simplicity, we assume that the tasks *i=1,…,k* can either be solved, which results in the full number r_i of points, or not be solved, resulting in zero points ($r_i=0$).
After working on a task, the student knows whether the task has been solved or not.
The student may attempt to solve each task a second time, but only when it has not been solved before.
For each attempt, the probability p_i of solving the task shall be independent. It depends only on the difficulty of the task and is as follows:

Task i	Points r_i	Solution probability p_i
1	8	0.15
2	6	0.4
3	10	0.25
4	2	0.6
5	7	0.35
6	3	0.5
7	20	0.2

Formulate this problem as a Markov Decision Process! 4 points

2.4) The student considers two policies for choosing the tasks:
π_A: work on the tasks in sequential order, according to index i.
π_B: work on the tasks in the order of increasing difficulty
(decreasing solution probability)
In both cases, the first non-solved task will be attempted again.
Compare the expected return of both policies! 4 points

2.5) Give an example for a process model where the Markov assumption is not justified.
How can the state be augmented to make the assumption valid again?
2 points