CoMo Computational Modeling Lab

Multi-Agent Learning Systems
Assignment 1

Deadline: 21 october 2010

1 N-Armed Bandit

Implement the 6-Armed bandit problem using each of the following action selection methods:

- Random
- Greedy
- ϵ-Greedy with $\epsilon = 0.2$
- Softmax with $\tau = 3$

All Q_{a_i} are initialized to 0. The rewards are subject to noise according to a normal probability distribution with mean $Q^*_{a_i}$ (see Table 1) and variance 1. (See section 2.2 of "Reinforcement Learning, An Introduction" by Sutton and Barto for a more comprehensive explanation of the problem[1]).

Action	$Q^*_{a_i}$
action #1	2.70
action #2	2.10
action #3	1.50
action #4	0.90
action #5	-1.50
action #6	-2.70

Table 1: $Q^*_{a_i}$ for each action.

1.1 Plotting Results

Run your simulation for 1000 time steps and plot the following graphs:

- For each of the 6 actions produce a plot: Q_{a_i} over time. Each plot should containing curves for all 4 action selection methods and a line showing the expected $Q^*_{a_i}$ for that action from Table 1.
- Plot on a single graph the average reward over time for each of the 4 action selection methods.
- For each of the 4 action selection methods plot a histogram showing the number of times each action is selected.

Discuss your results in a short paragraph.

1.2 Softmax Action Selection

Using the softmax action selection method with $\tau = 3$, re-run your simulation 1000 times, each with 1000 time steps. For each of the 6 actions plot the **average** Q_{a_i} over time and add a line showing the expected $Q^*_{a_i}$ for that action from Table 1. Repeat the same procedure for a softmax selection in which $\tau = 5 * \frac{1000-t}{1000}$, where t is the time step. (NOTE: make sure τ does not become 0!)
Discuss your results in a short paragraph.

[1] http://www.cs.ualberta.ca/%7Esutton/book/ebook/node16.html

1.3 Over/Underestimating Q-values

Re-run your simulation twice for 1000 time steps with **only 3** actions each: action #2, action #4 and action #6. For the first run initialize your Q-values to -10 (i.e., pessimistically) and for the second – to 10 (i.e., optimistically). Use softmax selection with the following 3 different temperature parameters: $\tau = 0.05$, $\tau = 3$, $\tau = 50$ and plot the results. In addition, plot the results for the greedy action selection. Discuss your results in a short paragraph.

NOTE: The choice of the programming language is free. Save yourself work by creating an abstraction for the number of actions so you can re-use your code for Task 1.1 and 1.2.

Good luck !!!

Give observation about the result