Exercise session 1: Bandits

Guillaume Charpiat Corentin Tallec

January 8, 2017

Exercise
The following exercise is strongly inspired by exercise from the second chapter of [1].

An instance of the 10-*arms-bandit* experiments set is generated by sampling 10 number from a standard gaussian of mean 0, $(q_0^*, \dots, q_{10}^*)$ and take those samples as mean values for each arm of the bandit. When arm i is selected, the agent receives a reward drawn from a gaussian of mean q_i^* and of variance 1.

A sample of 2000 10-*arms-bandit*s allows for evaluation of a bandit algorithm by letting the agent interact with each of the 2000 bandits and reporting performance evaluation. Mean reward and mean optimal action selection on the 2000 bandits are used to measure performance.

You are asked to implement various bandit algorithm on this test-bed and to report performances. A Python file is provided at this address to allow for easy testing. Instructions on how to implement your agent are provided in the file.

Implement an OptimisticEpsilonGreedyAgent, a SoftmaxAgent, a UCBAgent, and report your results on the test-bed. As a reminder, the UCBAgent selects is action as $\underset{k \in \mathcal{A}}{\operatorname{argmax}}\, \mu_{s_k,k} + \sqrt{\frac{2 \log t}{s_k}}$ with t the number of iterations, s_k the number of times arm k was selected and $\mu_{s_k,k}$ the empirical estimator of the value of arm k.

References

[1] Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press, 1998. ISBN: 0262193981.