Homework #5
The House Always Wins

Problem

Description

In reinforcement learning, the agent is seeking the optimal solution while solving the exploration-exploitation dilemma. Your experiments in this course have given you a chance to appreciate the importance of this dilemma and its effect on the rate of convergence. Given that, it is worth delving into the topic of exploration in more detail to understand its role for an reinforcement learning agent. To that end, we will be exploring the k-armed bandit problem for this assignment.

The k-armed bandit is a problem in which an agent is sitting at a row of 'k' slot machines. The agent has to decide which machines to play, how many times to play a machine, and the order in which the agent should play them. Each time a machine is played, it returns a random reward (based on a distribution). The goal of the agent is to maximize its reward.

Procedure

Implement a k-armed bandit environment along with agents for the following exploration strategies:

- ε -greedy with $\varepsilon = 0.01$
- ε -greedy with $\varepsilon = 0.1$
- Softmax action selection
- Optimistic initialization

Your algorithm using these strategies should make updates to the value function via incremental updates. Otherwise, you are free to choose parameters relevant to each strategy in a manner that produces the best outcome.

Then test your reinforcement learning agent and exploration methods on two different reward functions with a different number of arms. For both testbeds, Q^* is a normal distribution with a mean of 5 and a variance of 1.

100-armed testbed

Number of arms: k = 100

Reward function: normal distribution with mean Q* and variance of 1.

Number of pulls = 50000

Number of bandits = 2000

Softmax Temp = 0.3

Optimistic Init = 15

10-armed testbed

Number of arms: k = 10

Reward function:

"even arms": uniform distribution in the range [0 to 10]

"odd arms": normal distribution with mean Q* and variance of 5

Number of pulls = 20000

Number of bandits = 2000

Softmax Temp = 0.5

Optimistic Init = 15

Resources

The concepts explored in this homework are covered by:

- Lectures
  - Exploring Exploration
- Readings
  - Sutton-Barto Chapter 2 http://incompleteideas.net/sutton/book/ebook/node15.html
  - Kocsis-Szepesvari (2006)

Submission Details

Due Date: **July 9, 2017 (AOE)**

You will submit 8 graphs of Average reward on the 'x' axis versus number of pulls on the 'y' axis (1 graph for each strategy on each testbed). Please put all the graphs with labels in a single PDF and submit it to T-Square.