Homework 1: Bandits

T-747 Reinforcement Learning

Fall, 2023

1. *(Total points: 60)* Design and conduct an experiment to demonstrate the difficulties that sample-average methods have for non-stationary problems. Use a modified version of the 10-armed testbed in which all the $q_*(a)$ start out equal and then take independent random walks by adding a normally distributed increment with mean 0 and standard deviation 0.01 to all the q*(a) on each step.

   Prepare plots like Figure 2.2 (page 29 in the book) for an action-value method using sample averages, incrementally computed, and another action-value method using a constant step-size parameter, $\alpha = 0.1$. Use $\epsilon = 0.1$ and longer runs, say of 10,000 steps.

   (a) (30 points) The provided code implements the stationary k-armed bandit environment as well as an epsilon-greedy based agent which learns with the sample-average method. You should modify the code in the following way:

   - Add an environment that implements a 10-armed testbed where the true values of the actions (i.e., the mean of the reward distribution) changes slowly over time.
   - Add a different agent (or modify the existing one) that learns with a constant step-size instead of using sample averages.
   - Setup an experiment similar to the code in figure_2_2().

   (b) (30 points) Run the experiment and interpret the results. What do you see in the results? What does that tell you about the performance of these different methods for this particular environment?

   Hint: It might also help to look at the difference between the estimated values $Q_t(a)$ and true values $q_*(a)$ over time to understand what is happening.

   (c) **(20 bonus points)** Setup an experiment looking into the relationship of the step-size parameter and how quickly the environment is changing. For example, you could look at how accurate the estimated values $Q_t(a)$ are compared to the true values $q_*(a)$ over time with either a small or a large step size (e.g. 0.01 vs. 0.3) in an environment that is slowly changing vs. one that is quickly changing (standard deviation for the change in true value 0.01 vs. 0.1).

2. *(Total points: 40)* Suppose you face a 2-armed bandit task whose true action values change randomly from time step to time step. Specifically, suppose that, for any time step, the true values of actions 1 and 2 are respectively 10 and 20 with probability 0.5 (case A), and 90 and 80 with probability 0.5 (case B).

   (a) (20 points) If you are not able to tell which case you face at any step, what is the best expected reward you can achieve and how should you behave to achieve it?

   (b) (20 points) Now suppose that on each step you are told whether you are facing case A or case B (although you still don't know the true action values). This is an associative search task. What is the best expected reward you can achieve in this task, and how should you behave to achieve it?