ROB537: HW2

Ammar Kothari

1 Introduction

2 Bandit Problem

A 10 step and 100 step version of the Bandit Problem were tested. The same total number of steps was taken, 10,000 steps, but the episode resets occured depnding on the step length of an episode. The solution quality for the 100 step agent is similar to the 10 step problem.

2.1 Problem Description

The goal for the bandit problem is to maximize the expected return given a set of slot machines that give a reward. In this test, the rewards have a gaussian distrubution and is different for each slot machine. The number of steps is the number of steps allowed before the situation is reset. For the Bandit Problem, the initial state is with the accumulated reward as zero.

2.2 Results

An action value learning agent is used in both cases. Figure 1 and Figure2 shows the expected values as estimated by the learning for both episode lenghts with two different exploration amounts along with the actual distribution of the learners. Table 1 shows statistics of the reward obtained by the agents.

2.3 Analysis

All version struggle to approximate the true distribution. This can be caused by an insufficient number of samples. With increased exploration, the true values are generally approximated better. For action 2 and no exploration, appears to be close to the true value while the many of the other estimates are not close to approximating the true distribution.

3 Grid World

For grid world, two different types of learners are tested. The first only estimates the expected reward of actions. The second estimates the rewards based on state action pairs. The performance of the state action pair is better.

3.1 Problem Description

In Grid World, the agent is spawned at a random location in the world. The goal is a location on the map. Every location that is not the goal gives a reward of -1 and the goal gives a reward of 100. The length of an episode is 100 steps. At the end of an episode, the location of the agent is randomly placed on the map.

3.2 Results

The expected value of the action learner is shown in Figure **??**.

Steps	Exploration	Reward	StDev
10	0.2	0.914	1.24
100	0.2	1.49	0.58
10	0.0	1.14	0.94
100	0.0	0.93	1.87

Table 1: Results for different tests for Bandit Problem with an Action Value learner

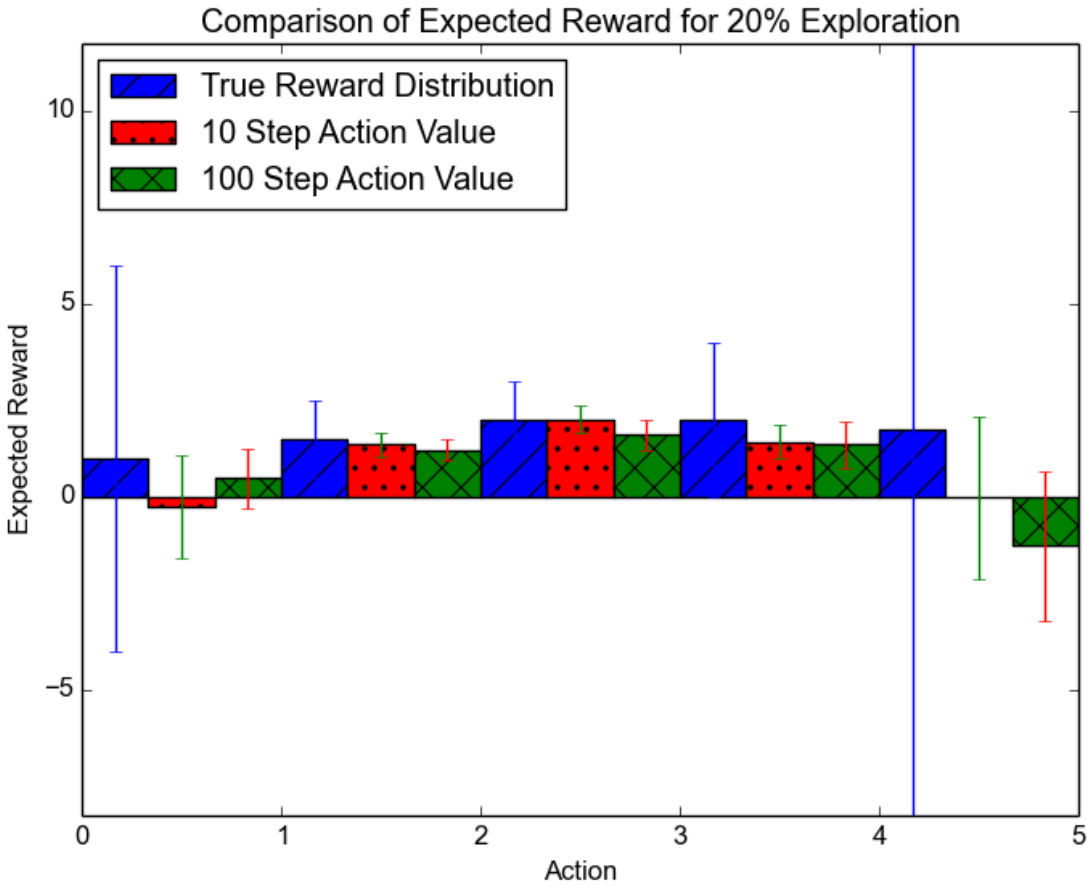

Figure 1: Comparison of Action Value tables between selection for 10 steps and 100 steps for an action-value learner with 20% epsilon greedy exploration on a multiarmed bandit problem.

4 Conclusion

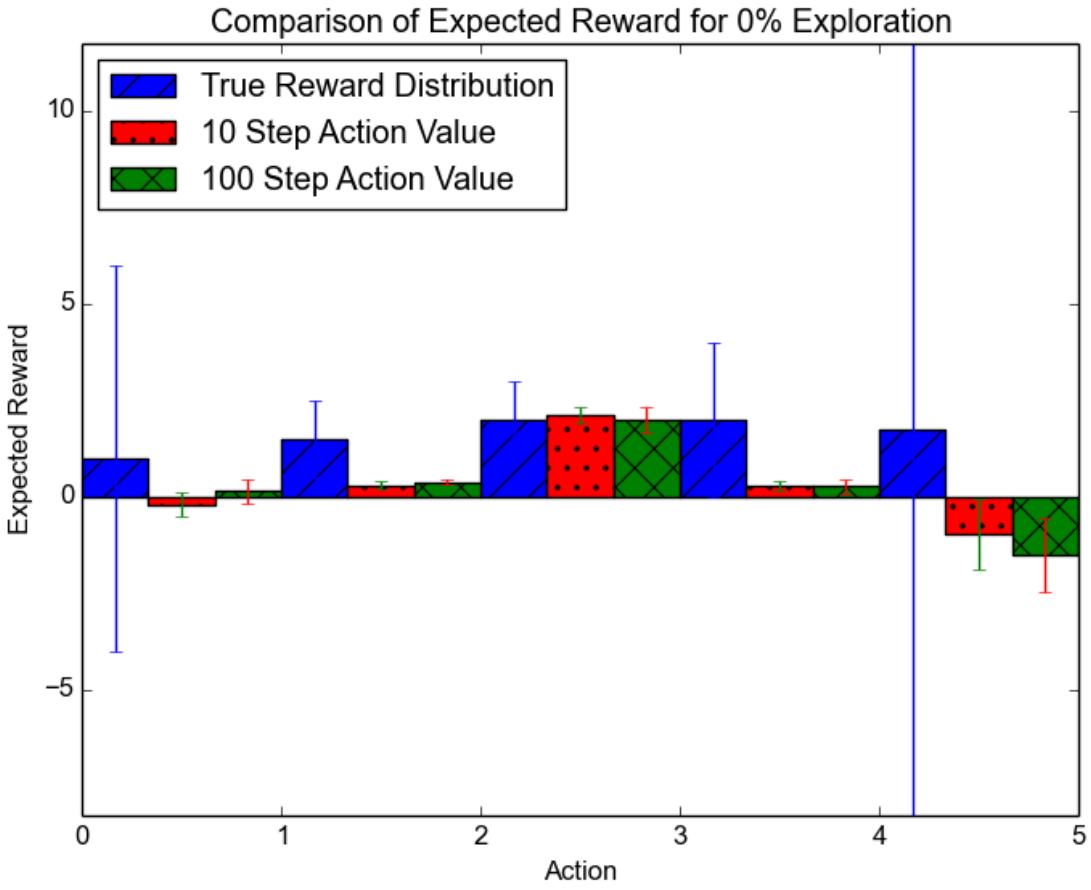

Figure 2: Comparison of Action Value tables between selection for 10 steps and 100 steps for an action-value learner with 0% epsilon greedy exploration on a multiarmed bandit problem.

AVGrid.png

Figure 3: Action value table for epsilon greedy agent for 20 steps.

AVGridSolution.png

Figure 4: Action value table quiver plot for epsilon greedy agent for 20 steps. Arrows are weighted average of the best action at that state.

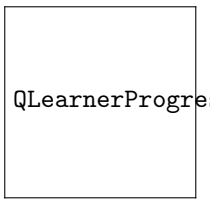

Figure 5: Q learner learning progression.

Figure 6: Q table quiver plot for epsilon greedy agent for 20 steps. Arrows are weighted average of the best action at that state.