ROB537: HW2

Ammar Kothari

1 Introduction

2 Bandit Problem

A 10 step and 100 step version of the Bandit Problem were tested. For both tests, XXXX total steps were run. The solution quality for the 100 step agent is superior to the 10 step problem.

2.1 Problem Description

The goal fo 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 **??** shows the expected values as estimated by the learning for both situations compared to the actual distribution of the learners.

– Actino Value learning tables.

– Show something about progress? like reward as a function of number of steps learned over.

3 Grid World

3.1 Problem Description

4 Conclusion

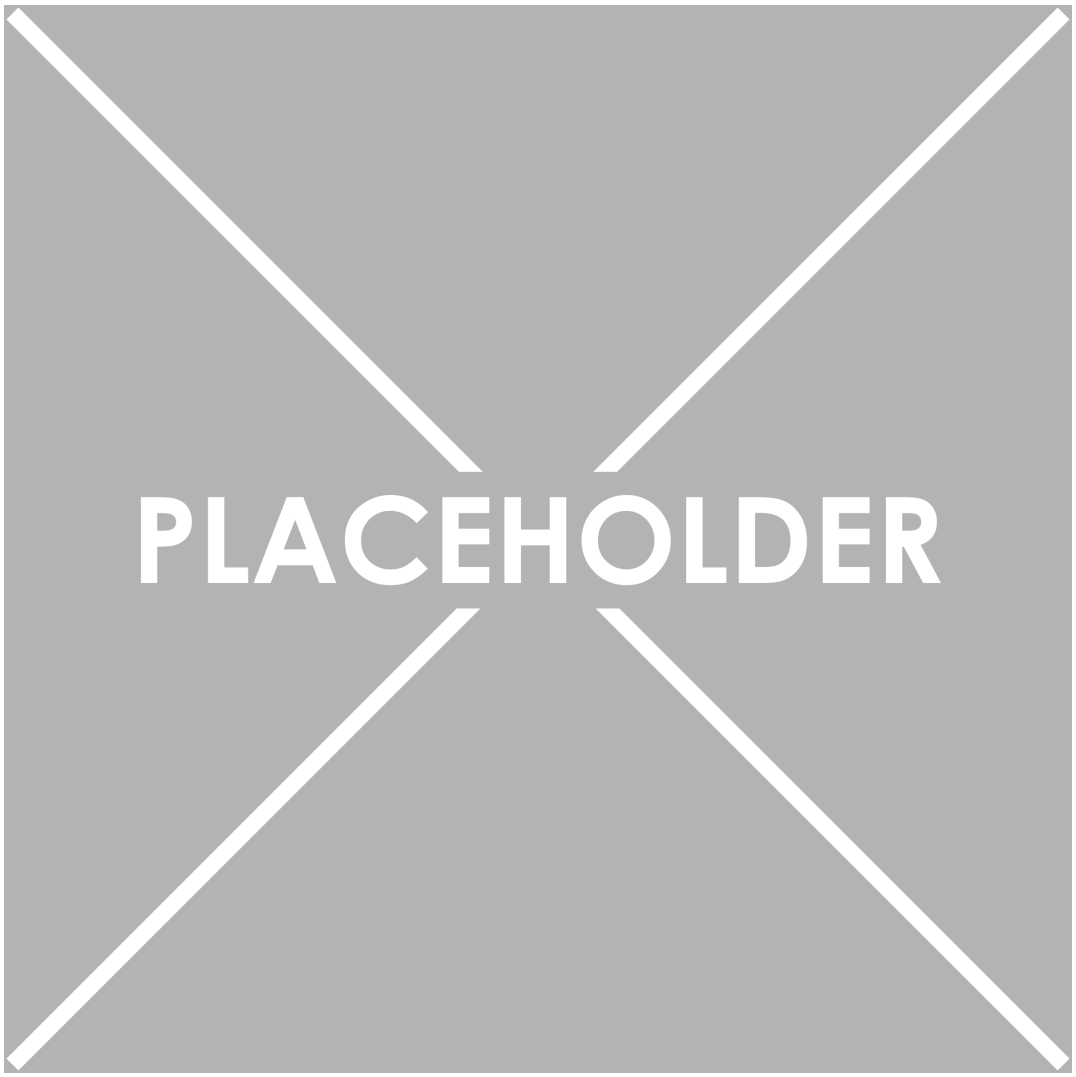

Figure 1: Comparison of performance between greedy and epsilon greedy selection over 10 steps for an action-value learner on a multiarmed bandit problem.

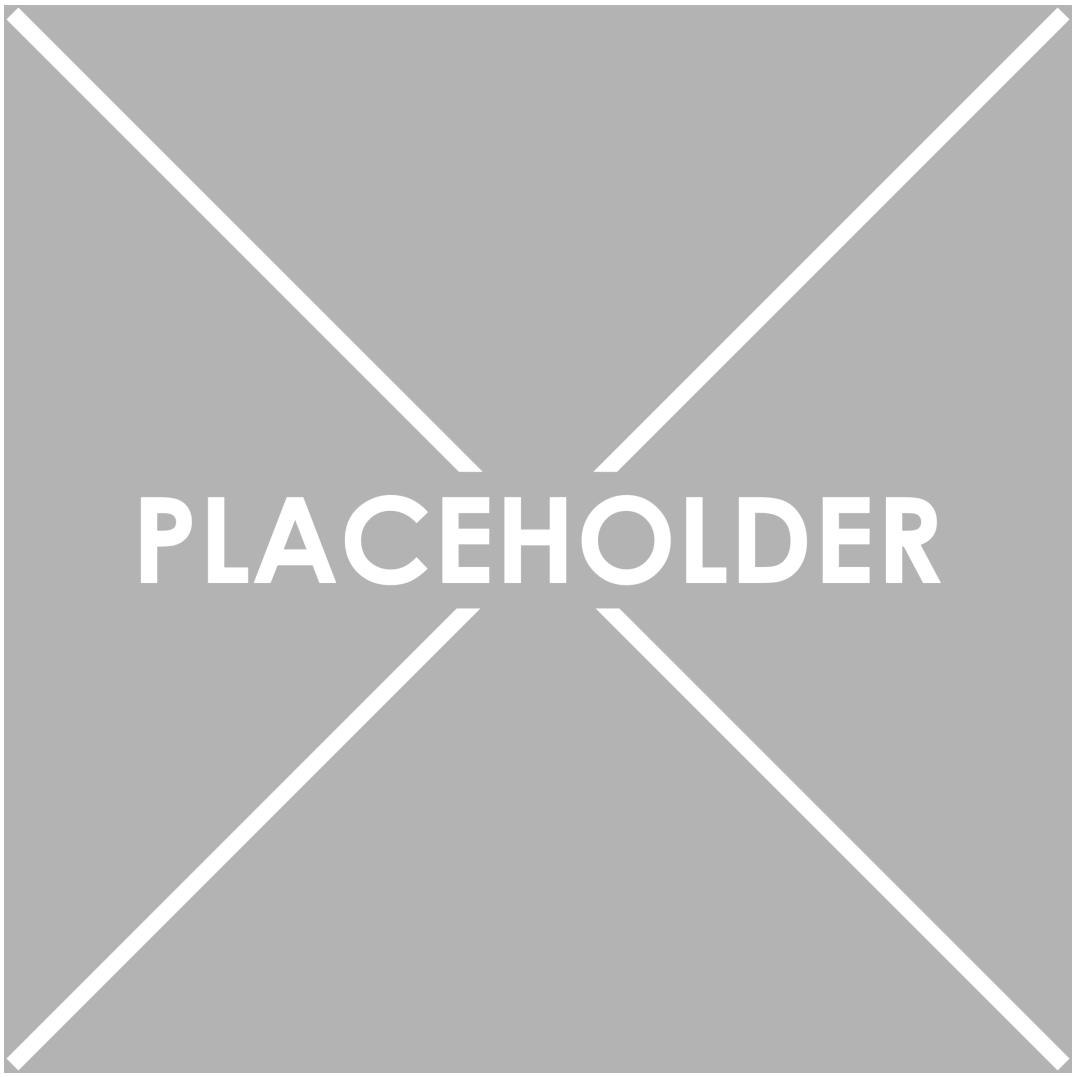

Figure 2: Comparison of performance between greedy and epsilon greedy selection over 100 steps for an action-value learner on a multiarmed bandit problem.

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.

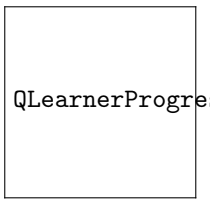

Figure 5: Q learner learning progression.

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