HW1

Theo McArn

January 2026

Question 1: 10-Armed Bandit Testbed

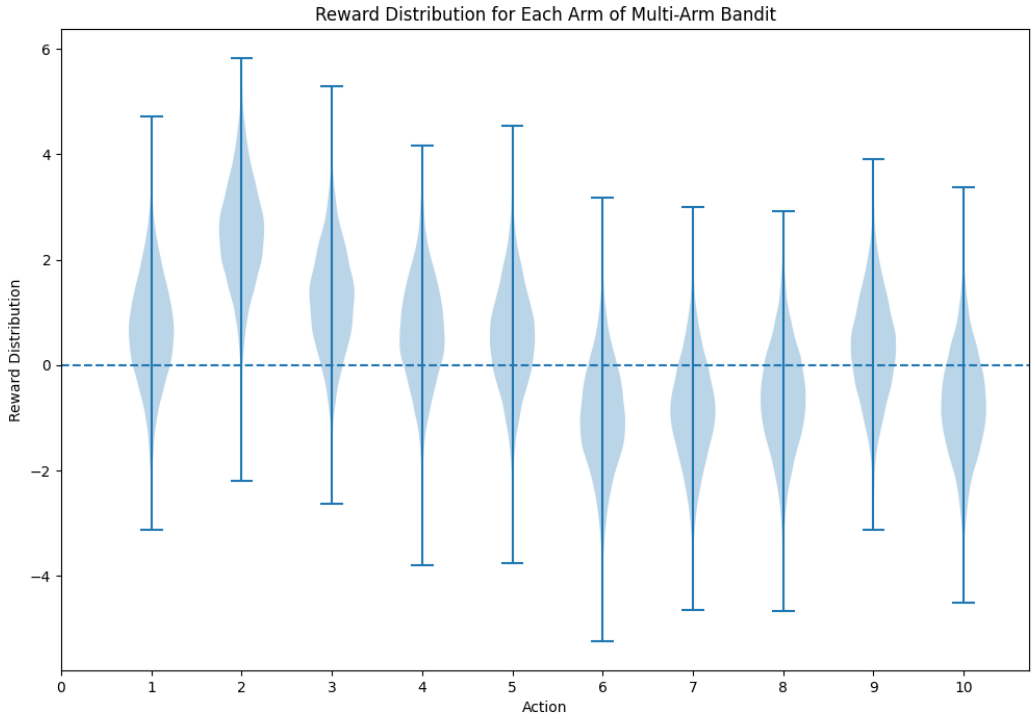


Figure 1: Reward Distribution for each Bandit

Question 2: ϵ-Greedy

When $\epsilon = 0$, there is no exploration. As a result the agent quickly converges on a bandit that is suboptimal. In this case, this results in an average reward of about 1.

When $\epsilon = 0.01$, there is a slight amount of exploration. 1% of actions are random. As a result, the average reward seems to steadily increase as more exploration happens and more optimal bandits are discovered. However, because the exploration is so slow, the average rewards per step does not seem to converge by the end of 1000 time steps. However, at the end of 1000 steps, the average reward is about 1.3.

When $\epsilon = 0.1$, there is a lot of exploration. 10% of actions are random. As a result, the average reward seems to quickly increase as more exploration happens and more optimal bandits are discovered. And it seems to converge with an average reward of about 1.4 per step once it has found the best bandit.

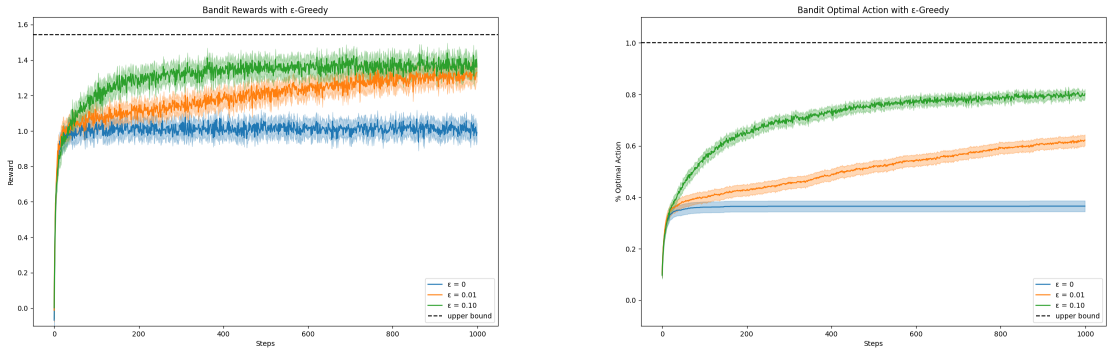


Figure 2: Average Reward and Percentage of Optimal Action for ϵ-Greedy Agents

Question 3: ϵ-Greedy vs. UCB

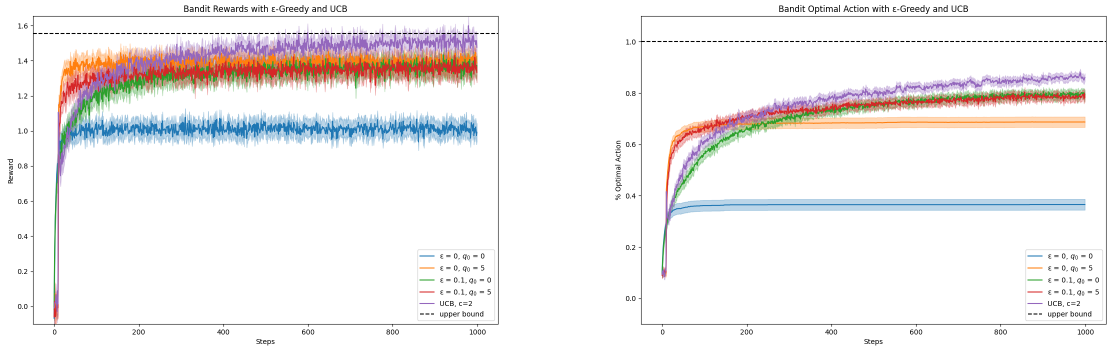


Figure 3: Average Reward and Percentage of Optimal Action for ϵ-Greedy Agents and UCB Agents

Why spikes appear in UCB and Optimistically Initialized Agents:

For both UCB agents and optimistically initialized agents, every bandit is tried once before returning to any duplicate actions. As a result, for the first 10 timesteps, every bandit is tested, regardless of what the previous rewards were. After 10 steps, the "best" action is then chosen based on the bandit that had the highest reward after the first 10 actions. This chosen action represents the "up" part of the spike. However, after this selection, the UCB score for this action is reduced because it has been visited twice and the scores for every other action is increased uniformly since they have only been visited once. As a result, this shift in UCB scores outweighs the Q-Values and as a result, the agent selects a new action that has a lower expected reward than the initial choice at step 10. Taking this action with a lower expected reward results in the "down" component of the spike.

This hypothesis is further supported by figure 3 because you can see from the rewards plot that this spike (in purple) happens roughly at timestep 10.

Varying Step-Size Weights

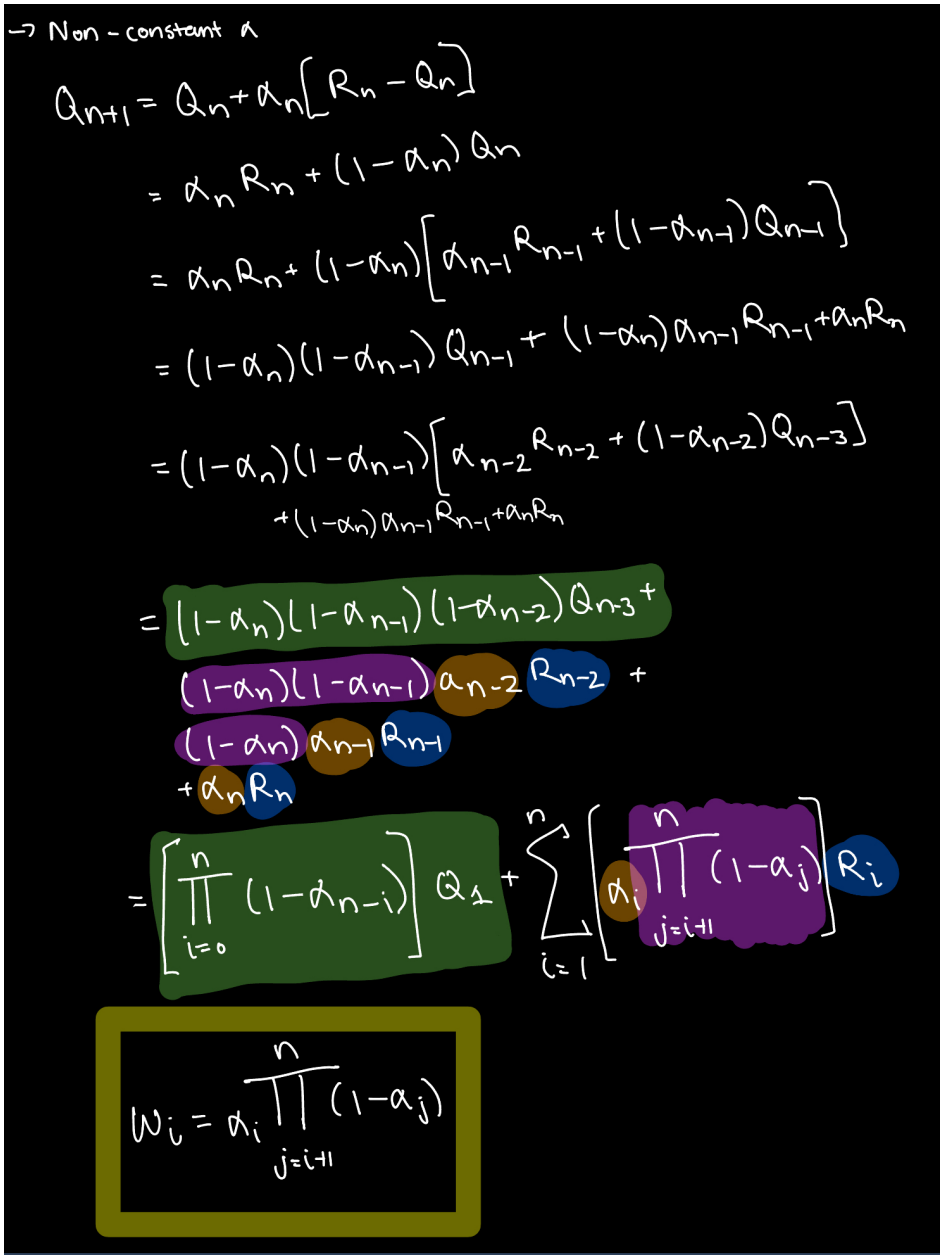
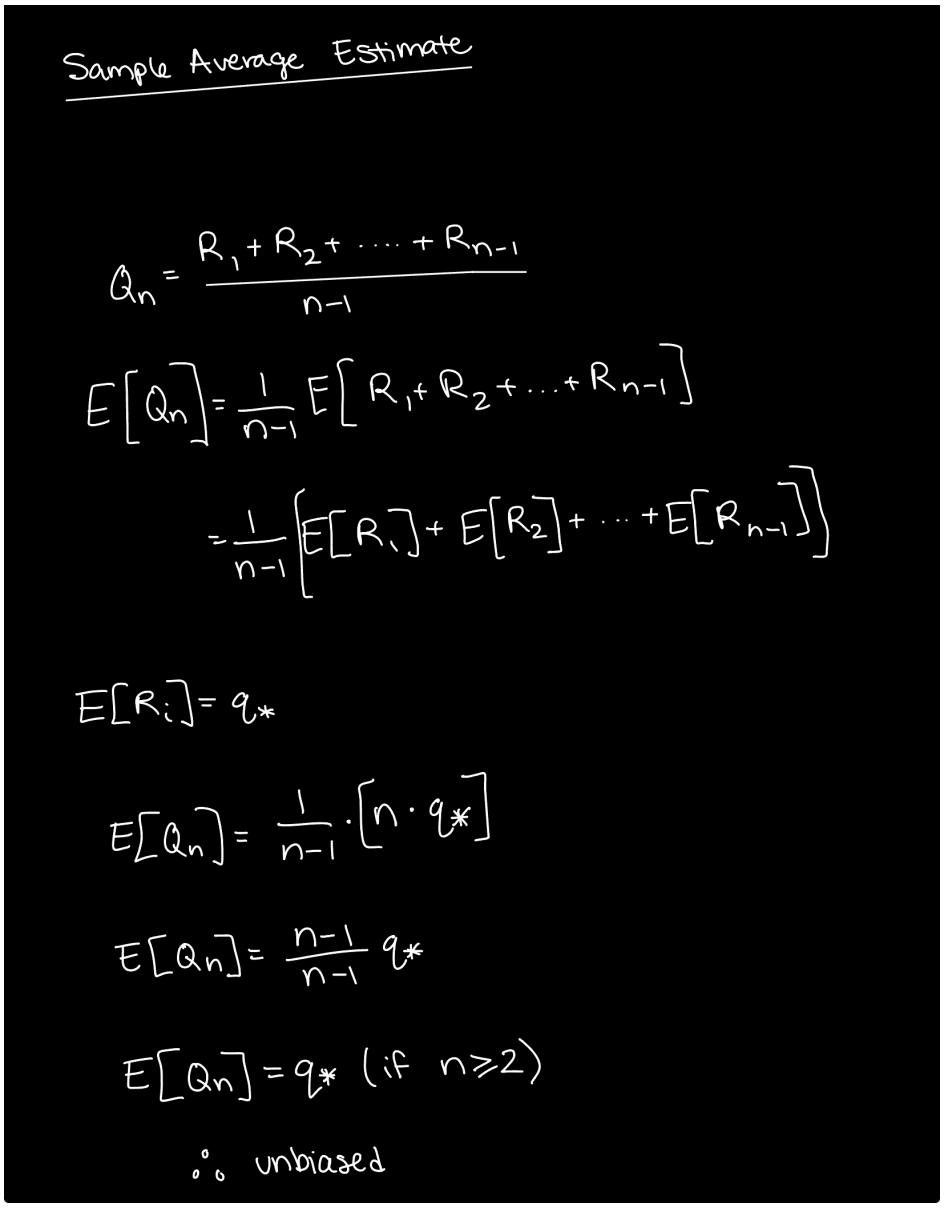


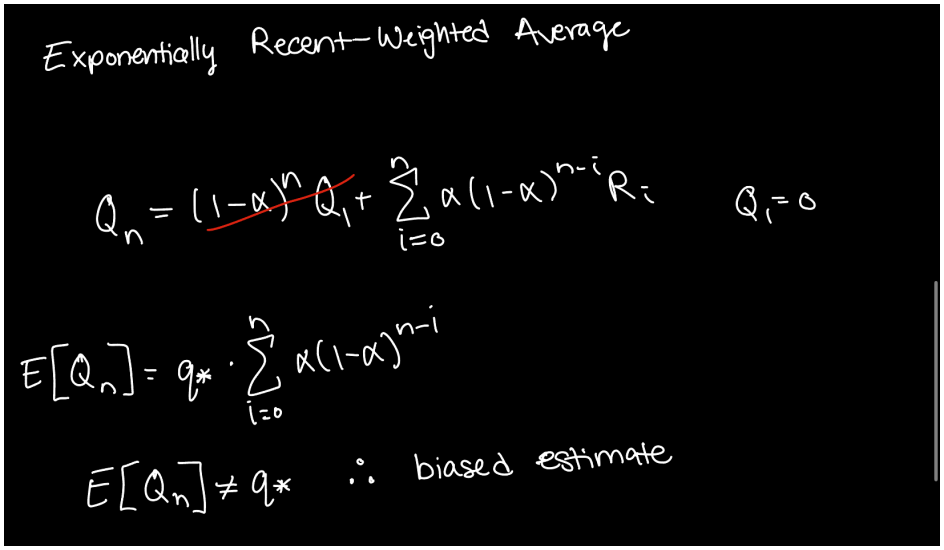
Figure 4: Derivation of Reward Weighting for Non-Constant Step Sizes

Bias in Q-Value Estimates

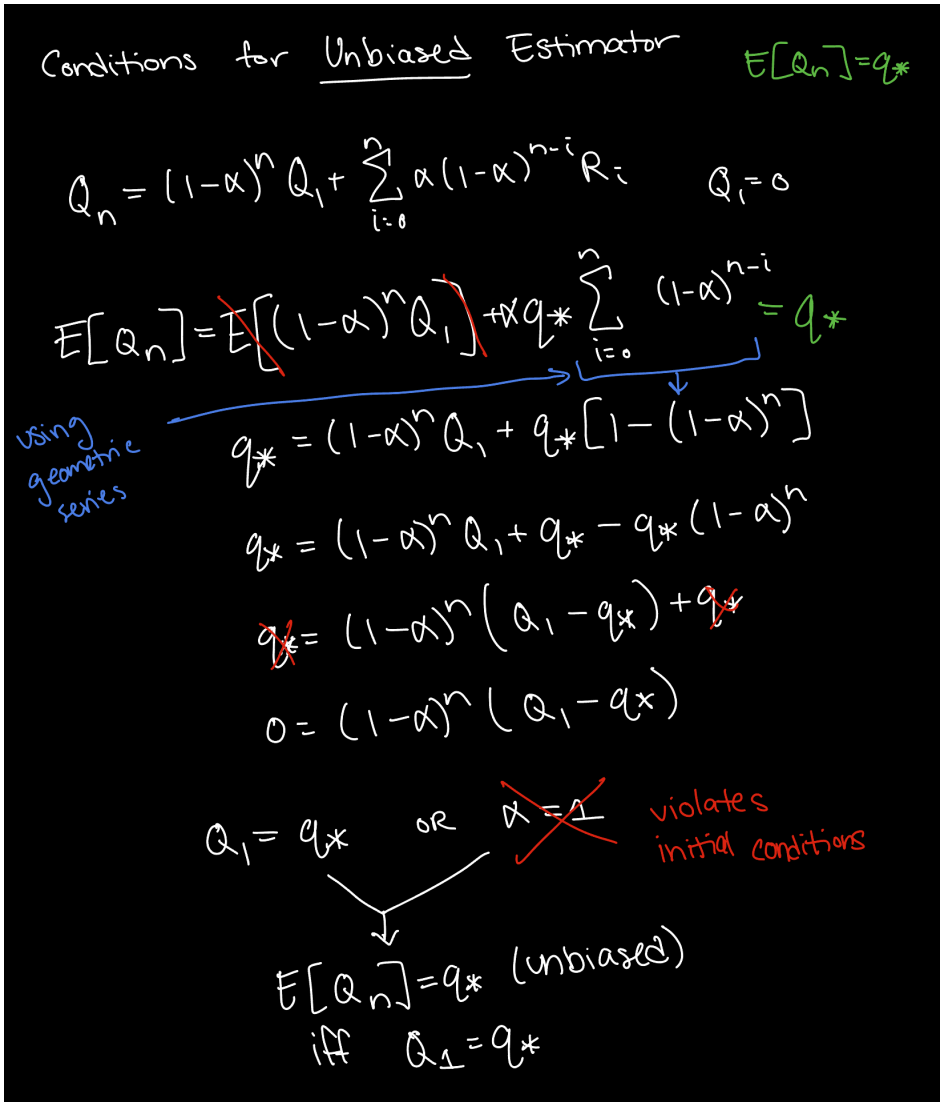
1. The sample-average estimate of Q_n is unbiased.



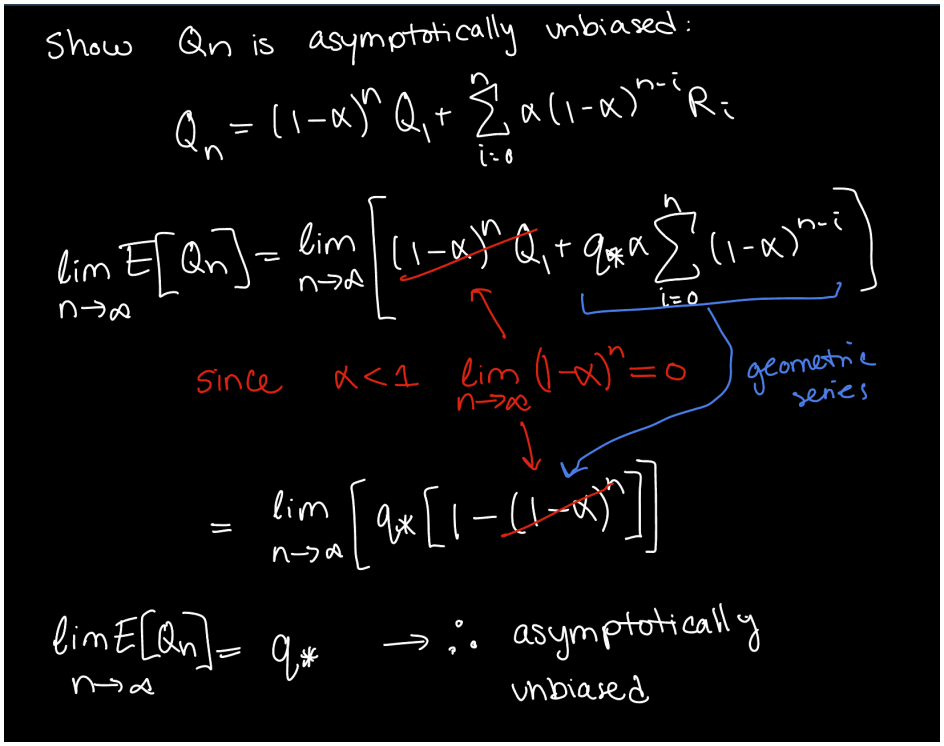
2. For an exponential recency-weighted average estimate, if $Q_1 = 0$, Q_n is biased when when $n > 1$.



3. The exponential recency-weighted average estimate will be unbiased only when $Q_1 = q_*$



4. For an exponential recency-weighted average estimate, Q_n is asymptotically unbiased as $n \to \infty$



5. **Why should we expect that the exponential recency-weighted average will be biased in general?** In general, we can expect that the exponential recency-weighted average will be biased because it is always placing more weight on the current and recent samples and less weight on the older observations. Inherently, this difference in weighting the rewards from the same distribution results in bias.