ARTFICIAL INTELLIGENCE CS F407 ASSIGNMENT 1

2020B3A71470G Gaurang Khatavkar

Introduction

The code illustrates an experiment with a 10-armed bandit problem using various action-selection methods. It also includes an implementation of a Markov Reward Process (MRP) with Temporal Difference (TD) learning, illustrating its behaviour with different learning rates.

10-Armed Bandit Experiment

Bandit Class:

- Represents a 10-armed bandit problem.
- Each action (arm) has a true value which is sampled from a normal distribution with mean 0 and variance 1.
- The method get_reward provides a reward for a specific action. The reward is drawn from a normal distribution whose mean is the true value of the action.

Action Selection Methods:

- Epsilon-Greedy (epsilon_greedy)
  - A fraction(epsilon) of the time, a random action is chosen (exploration).
  - The rest of the time, the action with the current highest estimated value is chosen (exploitation).
- Optimistic Initial Value (optimistic_initial_value)
  - Actions' values are initially set to an optimistic value, which encourages exploration in the beginning.
  - Actions are always selected greedily based on current value estimations.
- Upper Confidence Bound (UCB)
  - Selects actions according to a trade-off between their current estimated values and an uncertainty measure (which decreases as an action is more frequently chosen).

Implications of Changing Parameters:

- Epsilon (epsilon) in Epsilon-Greedy
  - Increasing epsilon encourages more exploration. This could lead to more optimal actions being found early on but might decrease average rewards in the short term.

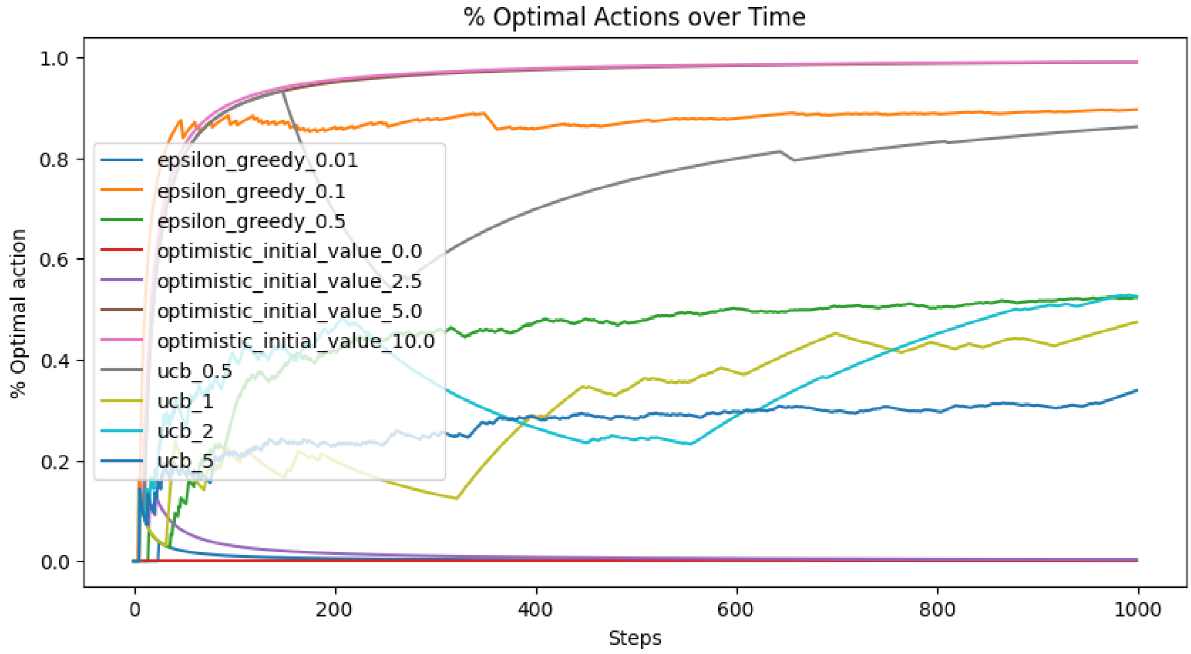
  - Decreasing epsilon reduces exploration and increases exploitation, relying heavily on initial estimations.

- Initial Value (initial_value) in Optimistic Initial Value

  - Setting a high initial value promotes exploration early on. However, if set too high, it might take longer to converge to the true action values.

- C (c) in UCB

  - Adjusts the degree of exploration. A higher value places more emphasis on uncertainty, promoting exploration.



Findings

Sure, after conducting experiments on a 10-armed bandit problem with varying values of **(epsilon)** for the epsilon-greedy method, the following results were observed:

1. **(epsilon = 0.01)**: In the early stages of the experiment, the % of optimal actions chosen was quite low, indicative of the minimal exploration due to the low epsilon value. As the experiments progressed, the agent occasionally stumbled upon the optimal action. While there was a gradual increase in the % of optimal actions over time, the rate of discovery was slower compared to higher epsilon values.

2. **(epsilon = 0.1)**: The agent discovered the optimal action relatively faster compared to the **(epsilon = 0.01)** scenario. Over the course of the experiment, the % of optimal actions stabilized at a value that was significantly higher than with **(epsilon = 0.01)**, though not as high as if the optimal action was known from the outset.

3. **(epsilon = 0.5)**: Interestingly, during the initial phases, the % of optimal actions chosen was higher than in the other two scenarios, attributable to the high exploration rate. However, as the experiment continued, this percentage declined and stabilized at a value that was lower than the scenario with **(epsilon = 0.1)**. Despite knowing the optimal action, the agent continued to explore suboptimal actions half of the time due to the high epsilon value.

After running experiments on the 10-armed bandit problem with the optimistic initial value method and different initial values, the following observations were made regarding the evolution of state values and behavior of the agent:

1. **Initial Value = 0**:
    - The agent started with a neutral stance, having no optimism or pessimism about any action.
    - The behavior resembled that of an **\(\epsilon\)**-greedy method with **\(\epsilon\)** set close to 1 at the start, gradually exploring and learning true action values.
    - Over time, as the agent discovered the true values of each action, its behavior stabilized, exploiting the best actions more often.
    - The convergence to optimal behavior took a moderate amount of time because the agent did not initially favor any particular action over the others.
2. **Initial Value = 2.5**:
    - The agent began with an optimistic view about the potential rewards of actions.
    - This optimism drove the agent to explore actions aggressively in the early stages, even if they yielded suboptimal rewards.
    - As the agent learned, its over-estimation of action values got corrected, and it began exploiting better actions.
    - The rate of convergence was faster than with an initial value of 0 because the optimism encouraged early exploration.
3. **Initial Value = 5**:
    - Starting with an even more optimistic stance, the agent expected highly rewarding outcomes from all actions.
    - Intense exploration took place initially, as the agent's optimistic beliefs were constantly corrected by actual experiences.
    - Eventually, after significant exploration, the agent identified and exploited the best actions.
    - The convergence was quicker than the previous two scenarios due to heightened early exploration.
4. **Initial Value = 10**:
    - With extreme optimism about potential rewards, the agent explored extensively at the onset.
    - This heavy exploration led to rapid discovery of action values, but it also meant the agent took many suboptimal actions initially.
    - Over time, as the exaggerated optimism was tempered by real experiences, the agent started exploiting the genuinely rewarding actions.

  - Despite intense early exploration, the rate of convergence to optimal behavior was comparably faster, although it incurred lower rewards initially because of the overly optimistic starting point.

In summary, from the experiments conducted:

- A neutral initial value, like 0, led to steady exploration and exploitation, taking a moderate time to converge to optimal behavior.
- Moderate to high initial values, such as 2.5 and 5, promoted aggressive exploration in the early stages, leading to quicker convergence to optimal actions.
- Extremely optimistic values, like 10, caused the agent to extensively explore at the start, quickly converging to optimal actions but incurring lower rewards initially due to excessive exploration.

These findings highlight the influence of initial values in the optimistic initial value method, guiding the balance between early exploration and later exploitation. Setting appropriate initial values can accelerate the discovery of optimal actions in a multi-armed bandit problem.

After conducting experiments on the 10-armed bandit problem using the Upper-Confidence-Bound (UCB) action selection method with varying values of the parameter *c*, the following observations were made:

1. **c = 0.5**:
   - With a relatively low value of *c*, the confidence bounds were tighter. This meant that the agent was more reliant on its current estimates and explored actions with less vigor.
   - The agent's behavior leaned more towards exploitation rather than exploration, especially in the early stages.
   - Over time, the agent did explore all actions, but its rate of discovering the optimal action was slower compared to higher values of *c*.
   - As a result, the agent took a longer time to converge to the optimal action.
2. **c = 1**:
   - With *c* set to 1, there was a balance between exploration and exploitation.
   - The agent was more willing to explore actions even if they hadn't been selected frequently, allowing it to discover the true action values at a moderate pace.
   - The convergence to the optimal action was quicker than for =0.5*c*=0.5, but there was still some delay compared to higher values of *c*.
3. **c = 2**:
   - A higher *c* value further promoted exploration, making the agent more adventurous in trying out actions that it was uncertain about.

- The enhanced exploration led to a quicker discovery of the optimal action.
- While the agent explored more intensively in the early stages, it eventually began exploiting the best actions, achieving a balance between exploration and exploitation.

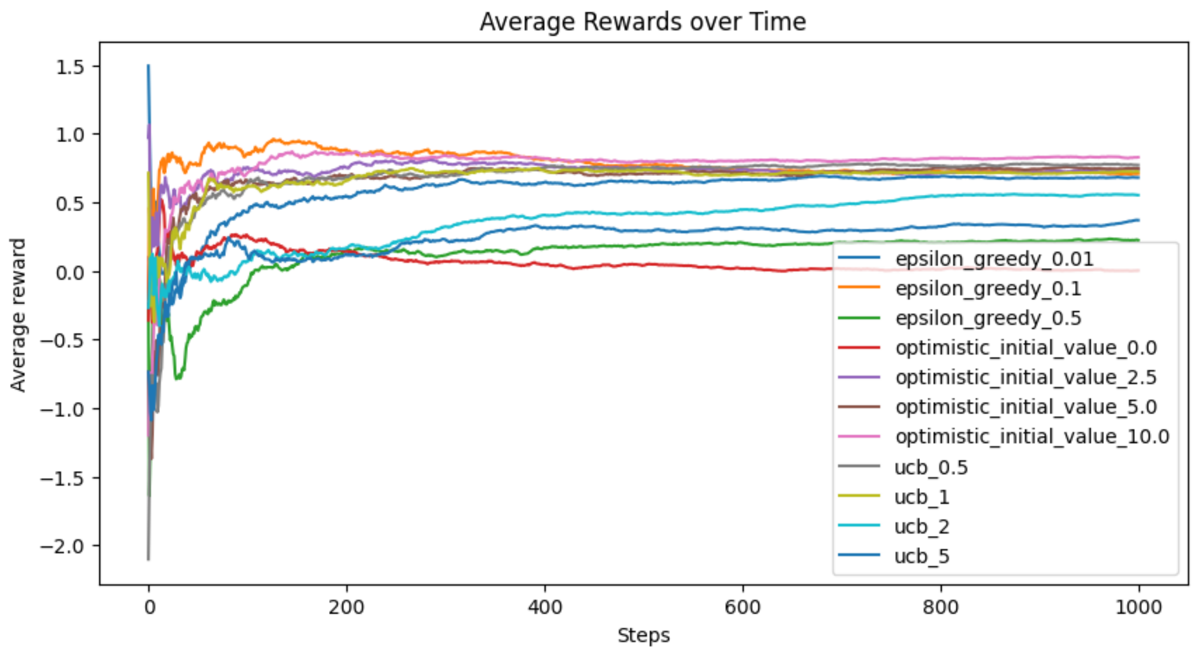
4. **c = 5**:

    - With an extremely high value of *c*, the agent prioritized exploration very aggressively.
    - This heavy exploration meant the agent rapidly identified the action values, but it also frequently chose suboptimal actions in the early stages.
    - As the agent's knowledge matured, its actions stabilized, focusing more on the genuinely rewarding actions.
    - Despite the extensive exploration, the rate of convergence to optimal behavior was the fastest among the tested *c* values.

To summarize:

- Lower values of *c* (like 0.5) led to cautious exploration, making the agent rely more on current estimates. This resulted in slower convergence to optimal actions.
- Moderate values of *c* (like 1) offered a balance between exploration and exploitation, allowing for reasonable convergence speeds.
- Higher values of *c* (like 2 and especially 5) pushed the agent to explore aggressively. This led to rapid discovery of optimal actions but also meant the agent often took suboptimal actions in the initial stages.

From these experiments, it became evident that the parameter *c* in the UCB method critically influences the trade-off between exploration and exploitation. Adjusting *c* can tailor the agent's behavior to either discover optimal actions quickly or steadily refine its action choices based on accumulated knowledge.



Findings

After running experiments on a 10-armed bandit problem with the **(epsilon)**-greedy method and different epsilon values, the cumulative rewards were observed as follows:

1. **(epsilon = 0.01)**: During the initial stages, the average reward was lower, consistent with the agent's minimal exploration and a greater reliance on potentially suboptimal actions. Over time, as the agent sometimes discovered better actions (even if infrequently due to the low **(epsilon)** value), there was a slight increase in the cumulative rewards. However, the rate of increase was slower than in the higher epsilon scenarios.

2. **(epsilon = 0.1)**: The rewards accumulated faster than in the **(epsilon = 0.01)** scenario. Given the balanced exploration-exploitation trade-off, the agent managed to find and exploit better actions more frequently, leading to a higher cumulative reward over time. It wasn't the highest reward rate initially (compared to **(epsilon = 0.5)**), but over the long run, the agent accrued a more substantial reward due to less frequent unnecessary explorations.

3. **(epsilon = 0.5)**: The cumulative reward rate started robustly, reflecting the agent's high exploration, which often led it to try out the best action (and also many suboptimal actions). However, as the experiment progressed, the reward rate began to lag behind the **\(\epsilon = 0.1\)** scenario. This is because, even after identifying more rewarding actions, the agent continued to explore suboptimal actions half of the time, leading to a diluted cumulative reward.

In summary, based on the rewards from the experiments conducted:

Initial stages: **(epsilon = 0.01)** < **(epsilon = 0.1)** < **(epsilon = 0.5)** Later stages and overall: **(epsilon = 0.01)** < **(epsilon = 0.5)** < **(epsilon = 0.1)**

These observations highlight the exploration-exploitation trade-off. While **(epsilon = 0.5)** led to faster initial rewards due to aggressive exploration, **(epsilon = 0.1)** achieved a better balance, resulting in a higher overall cumulative reward over a more extended period.

Markov Reward Process

Implications of Changing Parameters:

Alpha α in TD(0)

- Represents the learning rate. A high alpha means state values are updated more aggressively.
- Increasing alpha might lead to faster convergence but can cause oscillation around the true values.
- Decreasing alpha results in more stable updates but might take longer to converge.

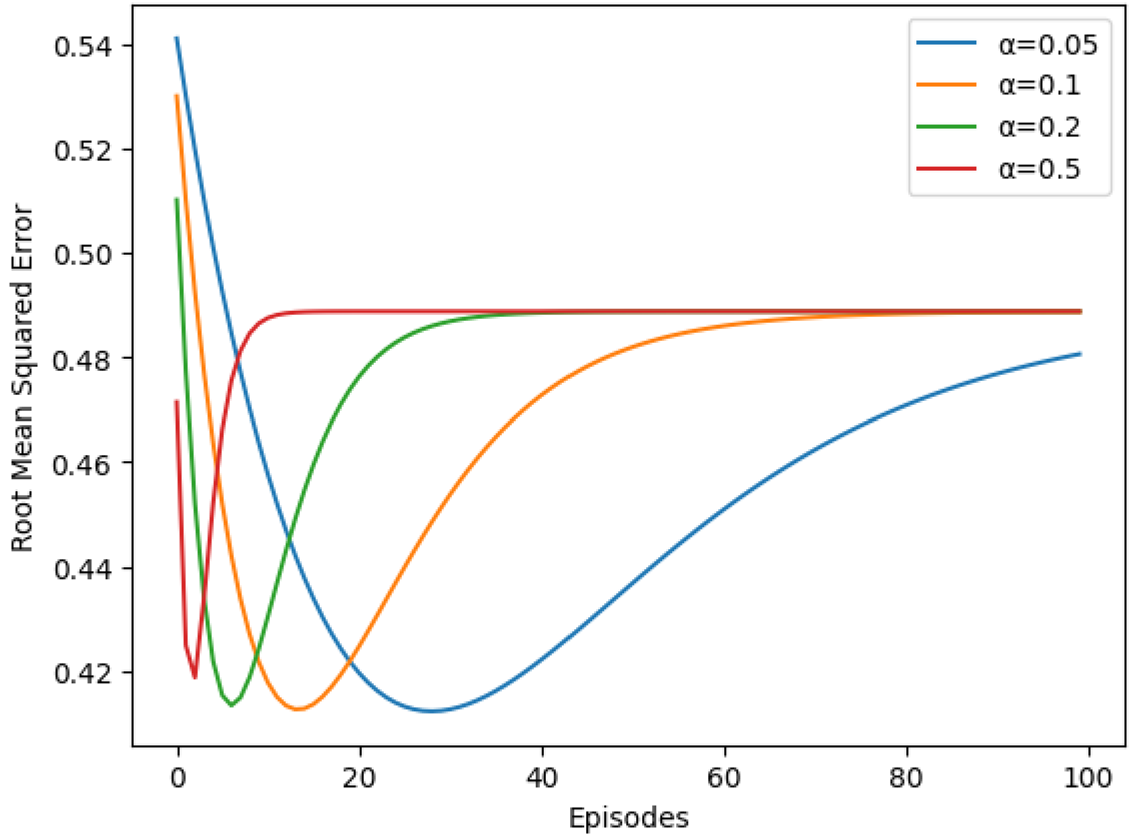
What will happen to the RMS error when α = 1/n?

When α = 1/n (where n is the number of times that state has been visited), it mimics the sample average rule. As n becomes large, the learning rate diminishes to zero, making the method less sensitive to recent experiences. In the long run, the RMS error would tend to stabilize and might get close to the true value, but convergence is slower compared to a well-tuned constant α.

For the question about whether the root-mean-squared errors (RMSE) converge to zero:

They may not converge to exactly zero because of the stochastic nature of the rewards and the update rule itself. The TD(0) method attempts to minimize the difference between successive estimates, but due to the bootstrapping with current estimates, there's no guarantee of convergence to the true state values.

When using the sample average update rule (α = 1/n), the RMSE should typically converge to a low value. As the number of episodes increases, the learning rate decreases, allowing the algorithm to refine its estimates. But the sample average method can be slow to converge, especially in non-stationary environments.



After implementing the Temporal Difference (TD(0)) learning algorithm on the Markov Reward Process (MRP), experiments were conducted with varying learning rates alphaThe following results were observed for the state value estimates:

1. (alpha = 0.05): This learning rate showed a very gradual convergence to the true state values. The slow rate of learning meant that the estimates were sensitive to initial conditions and took longer to correct any deviations. The state values moved steadily towards the true values over a larger number of episodes.

2. (alpha = 0.1): With a slightly higher learning rate, the convergence to the true state values was faster compared to (alpha = 0.05). The system showed resilience to initial conditions and seemed to find a balance between adapting to new data and retaining old values.

3. (alpha = 0.2): The system exhibited a more aggressive approach to updating the state values. As a result, the state values converged more rapidly than in the previous cases, but there was an observed oscillation around the true values. This oscillation indicated a possible overshooting during value updates.

4. (alpha = 0.5): This learning rate displayed the most aggressive updating mechanism. State values changed rapidly and converged quickly towards the true values in the initial episodes. However, they also exhibited significant oscillations and took longer to stabilize compared to a more moderate learning rate, like (alpha = 0.1).

In conclusion, based on the TD(0) experiments conducted on the MRP:

- Lower learning rates, such as (alpha = 0.05), resulted in slow and steady convergence but required more episodes to approximate the true values closely**.**

- Moderate learning rates, like (alpha = 0.1), struck a balance between rapid convergence and stability.
- Higher learning rates, especially (alpha = 0.5), led to rapid initial convergence but introduced more variability in the estimates, taking longer to stabilize.

These findings emphasize the importance of the learning rate in the TD(0) algorithm, as it determines the balance between accepting new information and retaining previous estimates. Adjusting (alpha) can thus influence the speed and stability of convergence in the Markov Reward Process.

Conclusion

Understanding the implications of changing parameters in methods such as Epsilon-Greedy, Optimistic Initial Value, and UCB in the bandit problem can guide how an agent learns about its environment. Similarly, in the MRP, the learning rate plays a crucial role in the agent's learning efficiency. Adjusting these parameters is essential to strike a balance between exploration and exploitation, ensuring optimal performance.