Sumeet Shanbhag

Student Id : 641020714

Reinforcement Learning Assignment #1

1. What is the benefit of incremental update for action-value function, versus non-incremental?

   a) **Streamlined Operations:** Incremental updates operate in a streamlined manner, processing new information as it arrives. This continuous adjustment ensures that the system remains updated without waiting for a batch of data, which is crucial for real-time applications.
   b) **Better Resource Management:** Instead of storing every reward to calculate the mean, incremental methods use the existing estimate to incorporate the new reward. This is akin to a person updating their mental note based on new experiences, without revisiting every past experience.
   c) **Relevance to Changing Environments:** Imagine a scenario where older information loses its relevance over time. Incremental updates naturally weigh recent experiences more, making them inherently adaptive to environments that evolve.
   d) **Ease of Integration:** Incremental update mechanisms can be effortlessly integrated into various reinforcement learning architectures. Their modular nature allows for customization, whether it's adjusting the learning rate or incorporating other optimization techniques.
   e) **Continuous Learning Paradigm:** Incremental updates embody the spirit of continuous learning. Just as humans constantly adapt and learn from new experiences without always reflecting on the entirety of past experiences, incremental methods tweak their knowledge with each new piece of information.
   f) **Swift Adaptation:** In dynamic environments where the underlying mechanisms might shift suddenly (think of a game where rules can change), incremental methods are swift to adapt, catching up with the new regime faster than batch methods that wait for a substantial amount of data before recalculating.

2. Why balancing exploration/exploitation is important? Name a few ways to balance exploration/exploitation in multi-armed bandits.

   Importance of Balancing Exploration and Exploitation:

   a) **Achieving Comprehensive Learning**: Without a balance, an agent might overcommit to a known strategy (exploitation) and miss out on potentially better options, or it could keep wandering in search of better options (exploration) and fail to capitalize on good strategies it has already discovered.

b) **Navigating Dynamic Landscapes**: Some environments change over time. If an agent solely exploits, it could be left behind when changes occur. Conversely, continuous exploration without utilizing learned knowledge can lead to wasted efforts.

c) **Ensuring Robust Performance**: A purely exploitative approach might work in one setting but fail in slightly altered conditions. Exploration introduces variability, making the agent more adaptable to a range of situations.

d) **Maximizing Long-Term Rewards**: The short-term benefits of exploiting known strategies might be eclipsed by the long-term gains from occasionally exploring new strategies and integrating them into the agent's approach.

Methods to Harmonize Exploration/Exploitation in Multi-Armed Bandits:

a) **ε-Greedy Variation**: The agent usually opts for the best-known strategy but occasionally takes a leap into the unknown, thus weaving together both exploration and exploitation.

b) **Temperature-Adjusted Strategies**: By adjusting the "temperature" in a Softmax action selection mechanism, actions are chosen based on a calculated probability, offering a structured way to explore while leaning on known rewards.

c) **UCB Approach**: This method weighs the potential upside of an action. It factors in not just the estimated rewards, but also the uncertainty around those estimates, naturally balancing the known and the unknown.

d) **Thompson Sampling**: By assigning probability distributions to rewards, this Bayesian approach lets agents pick actions that align with current knowledge while remaining open to new discoveries.

e) **Adaptive ε Techniques**: Rather than a static exploration rate, the ε value is adjusted over time, usually diminishing as the agent becomes more knowledgeable, thus blending early exploration with later-stage exploitation.

f) **Starting with Rose-Colored Glasses**: Setting initial optimistic values for action rewards nudges the agent to try out all actions before settling, promoting early-stage exploration.

3. In the equation
$$NewEstimate = OldEstimate + StepSize * [\, Target - OldEstimate \,]$$
what is the target ?

The term "Target" represents the most recent piece of information or feedback an agent receives about the true value it's trying to estimate. Think of it as the most recent data point indicating where the estimate should be directed.

In the context of reinforcement learning:

a) **Value Prediction**: Here, the "Target" is the reward you just received plus the expected future rewards, which can be written as: reward + gamma * V(next state).

b) **Q-learning**: This is an action-value method where the "Target" is the reward from a selected action plus the maximum expected future rewards of the next state, written as: reward + gamma * max Q(next state, all possible actions).

c) **Temporal Difference (TD) Learning**: In this approach, the "Target" is the immediate reward plus the expected value of the next state.

4. What is the purpose of using Upper Confidence Bound (UCB)?

In the exploration-exploitation dilemma, the Upper Confidence Bound (UCB) emerges as a guiding light in this challenging terrain. The UCB is used for the following purposes:

a) **Optimism in the Face of Uncertainty**: UCB operates on a profoundly philosophical yet practical premise: treat uncertainty as an opportunity. When an agent has limited knowledge about an action, UCB gives it an optimistic nudge. This ensures that no stone is left unturned, no path unexplored.
b) **Balanced Decision-making**: While most strategies might blindly gravitate towards known rewards or gamble on uncertain actions, UCB provides a mathematical compass. It considers not just the rewards an action might give but also the uncertainty surrounding those rewards, balancing the scales of knowns and unknowns.
c) **Dynamic Adjustments**: The magic of UCB lies in its dynamic nature. As an agent interacts with its environment and gathers more data, the UCB's estimates evolve. Actions that were once deemed promising might be deprioritized, and underexplored actions are given their fair chance under the spotlight.
d) **Reducing Regret**: In reinforcement learning, we're often concerned about the cumulative regret, which measures how much reward an agent has missed out on by not always choosing the best action. By systematically ensuring exploration, UCB helps minimize this regret over the long run.
e) **Universal Applicability**: Though UCB originated in the multi-armed bandit problem, its principles extend beautifully to more complex scenarios in reinforcement learning. Whether navigating a maze, playing a game, or any RL environment, UCB stands as a beacon, guiding agents to balance curiosity with caution.

5. Why do you think in Gradient Bandit Algorithm, we defined a soft-max distribution to choose the actions from, rather than just choosing action?

Using a softmax distribution in the Gradient Bandit Algorithm, as opposed to directly choosing actions based on their estimated value, is influenced by several key considerations:

a) **Stochastic Nature of Decision Making**: A softmax distribution introduces a probabilistic framework for action selection. This ensures that while the agent is biased towards actions with higher expected rewards, there's still a non-zero probability of selecting other actions. This inherently promotes a balance between exploration and exploitation.
b) **Smooth Transition between Actions**: As the agent's understanding of the environment evolves, the softmax probabilities shift smoothly. Instead of drastically switching between actions, there's a gradual transition which stabilizes the learning process. It avoids potentially abrupt changes in action preferences based on small fluctuations in value estimates.
c) **Graceful Handling of Ties and Near Ties**: In scenarios where multiple actions have nearly identical value estimates, deterministic policies may oscillate between these actions or might need complex tie-breaking mechanisms. Softmax provides a natural way to handle these situations by assigning similar probabilities to both actions.
d) **Temperature Parameter Flexibility**: The softmax function in the Gradient Bandit Algorithm typically includes a temperature parameter. This parameter provides a control knob for the level of exploration. High temperatures make action selection more uniform (increased exploration), while low temperatures make it more deterministic based on the current value estimates (increased exploitation). This flexibility is valuable in different stages of learning or different environments.
e) **Consistency with Policy Gradient Methods**: The Gradient Bandit Algorithm can be viewed as a precursor to more advanced policy gradient methods in reinforcement learning. By using a softmax distribution, it naturally aligns with these methods, where actions are chosen based on a policy parameterized by weights, and the weights are adjusted in the direction of the gradient to optimize some objective.

6. Read the article below and summarize what you learned in a paragraph: https://www.spotx.tv/resources/blog/developer-blog/introduction-to-multi-armed-bandits-with-applications-in-digital-advertising/

The article provides a deep dive into decision-making algorithms, especially on the Multi-Armed Bandits (MABs) approach. At its core, the article explores the epsilon-greedy strategy, a fascinating method that juggles between exploring new avenues (like testing different slot machines) and exploiting known winners. This strategy is brought to life with Python code examples, simulating how it would operate in real-world scenarios, such as choosing the most effective ads based on Click-Through Rates (CTRs). The concept of Bayesian Updating is also introduced, illustrating how algorithms can refine their beliefs and strategies based on accumulating evidence. However, the article doesn't shy away from pointing out potential

pitfalls; notably, how the epsilon-greedy method can sometimes become overly attached to decent, but not optimal, choices. This exploration is wrapped in a broader narrative of how to strike a balance between curiosity and confidence, ensuring that decisions lead to the best possible outcomes while minimizing the sting of missed opportunities.