Expected Sarsa

Objectives

- Describe the Expected Sarsa algorithm
- Understand how Expected Sarsa compares to Sarsa control
- Explain how Expected Sarsa generalizes Q-learning

Expected SARSA

- Expected SARSA is a variant of the SARSA algorithm that enhances learning by taking the expected value of action selection instead of selecting a single action deterministically.
- This approach accounts for the stochastic nature of the policy, providing a smoother and more stable update process compared to SARSA.

Expected SARSA

- In Expected SARSA, instead of selecting the action with the maximum Q-value (as in SARSA), the algorithm computes the expected value of all possible actions according to the current policy.
- The update to the state-action value is then based on this expected value.

Expected SARSA

- The Expected SARSA update rule for the state-action value Q(s,a) is given by:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left(r + \gamma \sum_{a'} \pi(a'|s') Q(s',a') - Q(s,a) \right)$$

  - α is the learning rate.
  - r is the immediate reward received after taking action a in state s.
  - γ is the discount factor, representing the importance of future rewards relative to immediate rewards.
  - $\pi(a'|s')$ is the probability of selecting action a′ in the next state s′ under the current policy.
  - Q(s′,a′) is the estimated state-action value for the next state-action pair (s′,a′).

Expected SARSA vs SARSA

- Expected SARSA and SARSA (State-Action-Reward-State-Action) are both reinforcement learning algorithms used for control tasks, aiming to find optimal policies by estimating action-values.
- However, they differ in how they update the action-values and make action selections.

Expected SARSA vs SARSA

- Action Selection:
  - Expected SARSA: action selection is based on the expected value of all possible actions according to the current policy. It computes the expected value of all actions and selects actions probabilistically based on their probabilities under the current policy.
  - SARSA: action selection is typically greedy or ε-greedy, where the action with the maximum Q-value is chosen with probability 1-ε and a random action is chosen with probability ε.

Expected SARSA vs SARSA

- Update Rule:
  - Expected SARSA: It updates the Q-values based on the expected value of all possible actions, considering the probabilities of each action under the current policy.
  - SARSA: It updates the Q-values based on the action actually taken and the subsequent action chosen according to the current policy.

Expected SARSA vs SARSA

- Stability:
  - Expected SARSA: It generally provides more stable updates compared to SARSA(By considering the expected value of all actions)
  - SARSA: It updates can be more volatile, especially in environments with high variability or stochastic policies.

Expected SARSA vs SARSA

- Exploration-Exploitation Trade-off:
  - Expected SARSA: Expected SARSA maintains a balance between exploration and exploitation by considering the expected value of all actions.
  - SARSA: SARSA uses ε-greedy exploration, which can sometimes result in suboptimal actions being chosen due to the random exploration component (SARSA generally prioritizes exploitation over exploration, especially with low ε values)

Expected SARSA vs Q-Learning

- Expected SARSA is a variant of SARSA that combines elements of both SARSA and Q-learning.
- It generalizes Q-learning by considering the expected value of all possible actions at each state, similarly to how Q-learning considers the maximum value of the next state-action pair.

Expected SARSA vs Q-Learning

- Expected Action Selection:
  - In Q-learning, the agent selects the action with the maximum Q-value for the next state (s′) deterministically.
  - In Expected SARSA, the agent computes the expected value of all possible actions in the next state according to the current policy. This is done by averaging the Q-values of all possible actions, weighted by their probabilities under the current policy.

Expected SARSA vs Q-Learning

- Update Rule:
  - In Q-learning, the Q-value for the current state-action pair is updated based on the maximum Q-value of the next state (s').
  - In Expected SARSA, the Q-value is updated based on the expected value of all possible actions in the next state (s'), considering their probabilities under the current policy.

Expected SARSA vs Q-Learning

- Exploration-Exploitation:
  - Q-learning often prioritizes exploitation over exploration, as it updates the Q-value based on the maximum value, which might not necessarily be the action chosen during exploration.
  - Expected SARSA, by considering the expected value of all actions, maintains a balance between exploration and exploitation. It ensures that exploration is guided by the probabilities of actions under the current policy, leading to a smoother exploration process.

Expected SARSA vs Q-Learning

- Stability:
  - Expected SARSA generally provides more stable updates compared to Q-learning. By considering the expected value of all actions, it smoothes the learning process and reduces the variance in updates, resulting in more consistent convergence.

Summary

- Describe the Expected Sarsa algorithm
- Understand how Expected Sarsa compares to Sarsa control
- Explain how Expected Sarsa generalizes Q-learning

Q & A