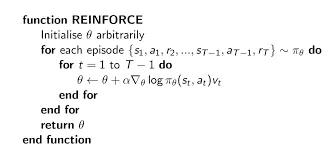
|  |
| --- |
| CS 370 Current/Emerging Trends in CS |
| 6-2 Assignment: Cartpole Revisited |
| Emily Wood |

|  |
| --- |
| emily.wood7@snhu.edu  4-9-2023 |

**Explain how the cartpole problem can be solved using the REINFORCE algorithm.**

The REINFORCE algorithm uses a policy. A policy is a model that takes a state as input and generates the probability of taking an action as output (Noufal, 2020). The objective of the policy is to maximize the expected reward. In the case of the cartpole problem, a random policy is initialized. Then a number of steps are taken are taken and the actions, rewards, and environment are recorded. The discounted reward and expected rewards are calculated. The weights of the policy are adjusted, and the cycle is repeated.

  
image from medium.com

**Explain how the cartpole problem can be solved using the A2C algorithm.**

The Advantage Actor Critic (A2C) algorithm consists of two networks. The advantage function calculates the prediction error. The actor network chooses and action at each step and the critic network evaluates the quality of the input state. As the critic network learns which states are better, the actor uses this to teach the agent to use good states (Wang, 2021).

With the cartpole problem, the advantages cease when the game ends. When the critic sees this, it gives a larger negative value. The actor sees a move as worse than expected and decides to try something else. Many games are played, so the actor and critic work together to balance the pole for longer periods of time.

Chart, diagram, bubble chart

Description automatically generated

image from (Wang, Advantage actor critic tutorial: MINA2C 2021)

**Explain how policy gradient approaches differ from value-based approaches, such as Q-learning.**

Q-Learning uses state-action values (or Q values) for each action. Usually starting from nothing, the agent will interact with the environment, get feedback, and improve the Q-values to derive the optimal policy (Doshi, 2021). The optimal policy is obtained indirectly.

Unlike the Q-Learning approach, which outputs Q-Values for all actions, a policy gradient approach learns a function that outputs the best action that can be taken from a state. Rather than outputting a single action, it will output a probability distributions of the actions. The action is then chosen.

**Explain how actor-critic approaches differ from value- and policy-based approaches.**

A2C and policy-based approaches involve policy evaluations and policy improvement, while a value-based approach finds the optimal value function and finds the policy based on that. With A2C and policy-based approaches the evaluations and improvement are iteratively done to where they meet, while with value-based approaches the finding of the optimal value function and finding the policy are not repeated once completed. A2C and policy-based functions are relatively faster than value-based approaches, but are used in simpler problems (Bhagat et al., 2018).

**References:**

Noufal, S. (2020, November 24). *Reinforce algorithm: Taking baby steps in reinforcement learning*. Analytics Vidhya. Retrieved April 8, 2023, from <https://www.analyticsvidhya.com/blog/2020/11/reinforce-algorithm-taking-baby-steps-in-reinforcement-learning/>

Wang, M. (2021, January 26). *Advantage actor critic tutorial: MINA2C*. Medium. Retrieved April 8, 2023, from <https://towardsdatascience.com/advantage-actor-critic-tutorial-mina2c-7a3249962fc8>

Doshi, K. (2021, April 24). *Reinforcement learning explained visually (part 6): Policy gradients, step-by-step*. Medium. Retrieved April 8, 2023, from <https://towardsdatascience.com/reinforcement-learning-explained-visually-part-6-policy-gradients-step-by-step-f9f448e73754>

Bhagat, S., Banerjee, H., & Ren, H. (2018). Deep reinforcement learning for soft robotic applications: Brief overview with impending challenges. https://doi.org/10.20944/preprints201811.0510.v2