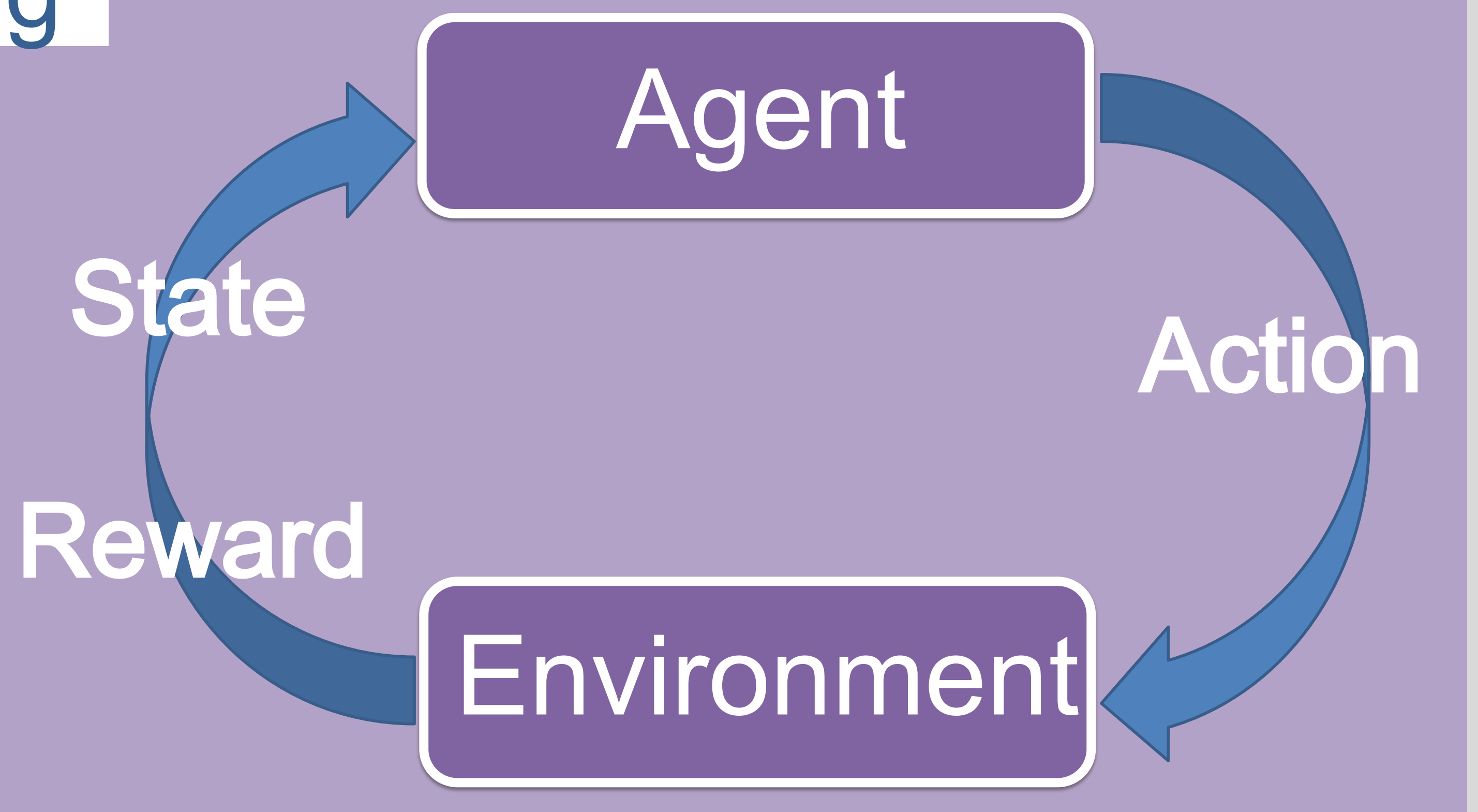
# Game Theory Based Distributed Coordination with Multi-Agent Reinforcement Learning

Charlotte M. Morrison

Faculty Mentor: Dr. Ayan Dutta

## Reinforcement Learning

In reinforcement learning systems, the agent learns through interactions with the environment. The agent takes an action, and this changes the environment. The agent receives a reward or penalty based on how the action affects the environment. This process is repeated allowing the agent to develop a policy that can be used to determine what is a good action given a new environmental observation.



## Problem

Reinforcement Learning (RL) is a powerful tool for training multi-agent robotic systems. This project is focused on using RL in a multi-agent robotic system consisting of two agents to manipulate objects.

Current system considerations

- continuous action space
- massive state space
- cooperative learning

## Multi-Agent System

### Baxter

Each of Baxter's arms is treated as an independent agent that learns separately from the other. Baxter uses RL to learn a policy that utilizes all 7 joints on each arm to best move in cooperation with the other arm to complete an object manipulation task.

## Coordination

### Game Theory Based Cooperation

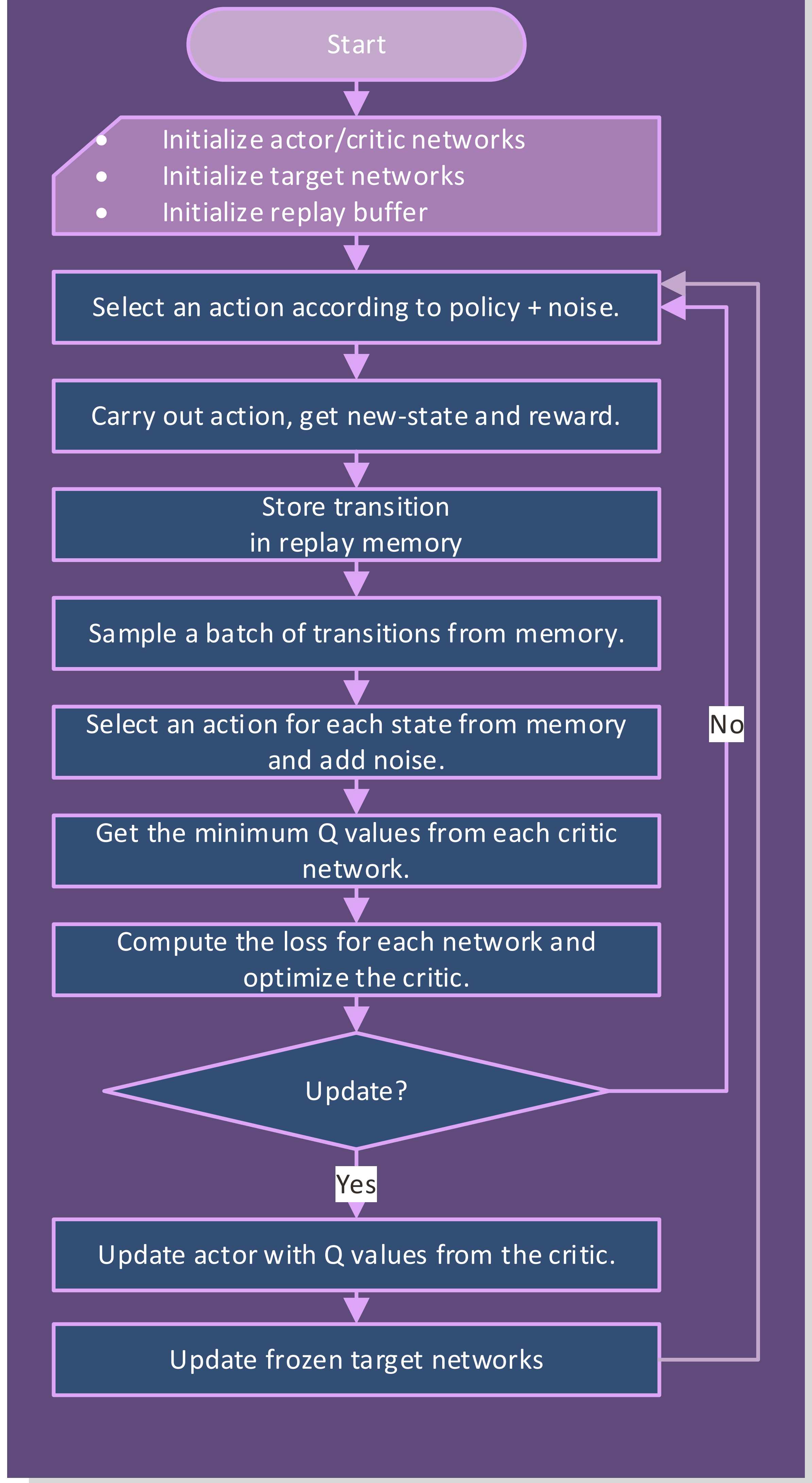
The game participants are the independent agents (each of Baxter's arms) and each can make decisions independently.

The agents receive an individual payoff or reward based on the quality of an action. The individual rewards are combined to produce a shared reward that is incorporated in the learning process. The independent agents learn to cooperate by sharing this reward and using the reward feedback to improve the decision making.

The agents learn to maximize the shared reward and develop a policy for movement through the TD3 algorithm.

## TD3 Algorithm

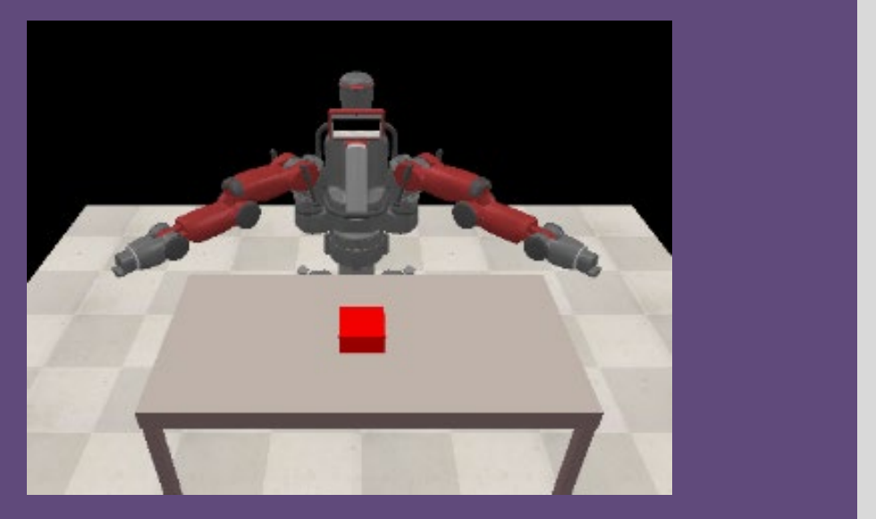
The Twin Delayed Deep Deterministic Policy Gradients (TD3) is an algorithm that can be used for reinforcement learning where the action space is continuous. TD3 is an improvement on Deep Deterministic Policy Gradients (DDPG) because it solves the problems with the accumulation of overestimation errors through the use of a pair of critic networks. TD3 utilizes the minimum value from the pair of critics, delayed updates to the actor, and the addition of regularized noise to improve the stability of the DDPG algorithm.

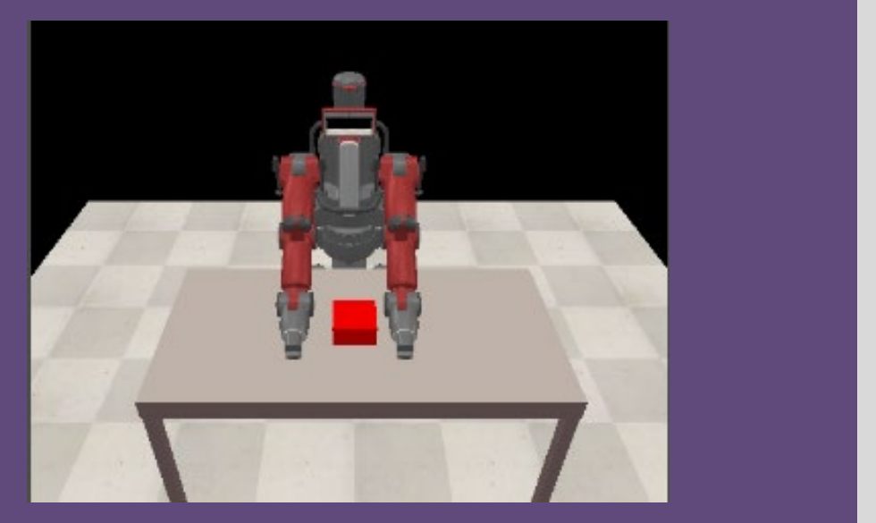

Start
Initialize actor/critic networks
Initialize target networks
Initialize replay buffer
Select an action according to policy + noise.
Carry out action, get new-state and reward.
Store transition in replay memory
Sample a batch of transitions from memory.
Select an action for each state from memory and add noise.
Get the minimum Q values from each critic network.
Compute the loss for each network and optimize the critic.
Update?
No
Yes
Update actor with Q values from the critic.
Update frozen target networks

## Contributions

This research demonstrates that the TD3 algorithm can be applied to a dual-agent system using game theory cooperation.

Baxter is currently learning in a simulated environment to use both arms to approach a fixed target. At each step, his policy provides movements to all the 14 joints. The simulation has learned a policy that directs the arms to successfully approach the target position.





### Current Work

The next phase of research is to train the robot with increasingly complex tasks in the simulation. The next task requires the robot to move an object from a specified location to another specified location. without dropping it. The robot will use both hands to hold the object throughout the movement.

The performance will be tested using fully shared actor and critic networks and independent actors for each agent with shared critic networks.

## References

[1] D. Easley and J. Kleinberg, "Chapter 6," in Networks, Crowds, and Markets: Reasoning about a Highly Connected World, Cambridge University Press, 2010, pp. 155-208.

[2] K. Zhang, Z. Yang and T. Basar, "Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms," arXiv, vol. 1, pp. 1-71, 2019.

[3] S. Fujimoto, H. van Hoof and D. Meger, "Addressing function approximation error in actor-critic methods," arXiv, vol. arVix:1802.09447, 2018.

[4] T. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver and D. Wierstra, "Continuous control with deep reinforcement learning," arXiv, vol. arXiv:1509.02971, 2015.

[5] K. Zhang, Z. Yang, H. Liu, T. Zhang and T. Basar, "Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents," arXiv, 2018.

[6] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, H. S. T. Philip, P. Kohli and S. Whiteson, "Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning," arXiv, vol. arXiv:1702.08887v3, 2018.

[7] J. Ackermann, V. Gabler, T. Osa and M. Sugiyama, "Reducing Overestimation Bias in Multi-Agent Domains Using Double Centralized Critics," arXiv, vol. arXiv:1910.01465v2, 2019.

[8] OpenAI — Spinng Up, 2018: https://spinningup.openai.com/en/latest/algorithms/td3.html#background