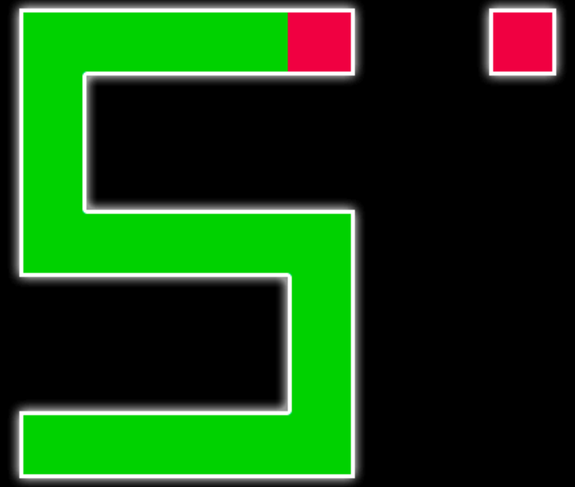
MetaBot Learning to Play

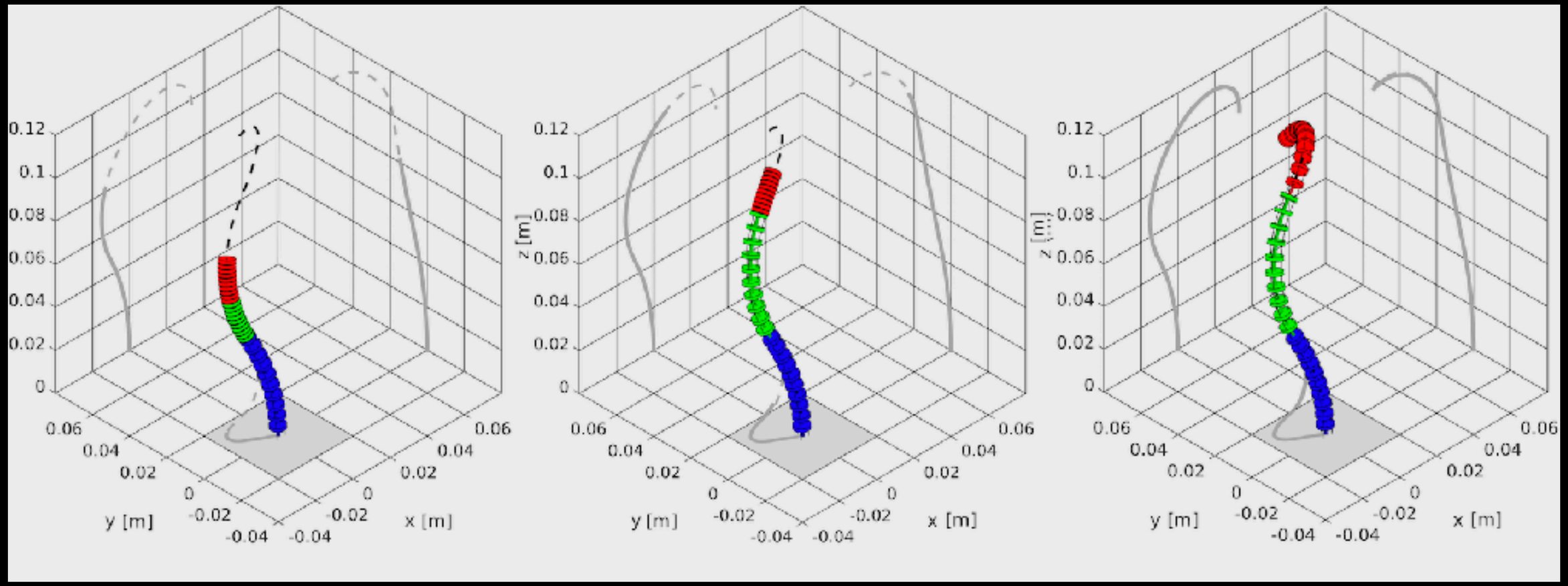
Reinforcement Learning Approach



Abdulwasay Mehar
Supervisors: Jessica Burgner-Kahrs & Reinhard Grassmann

UNIVERSITY OF TORONTO

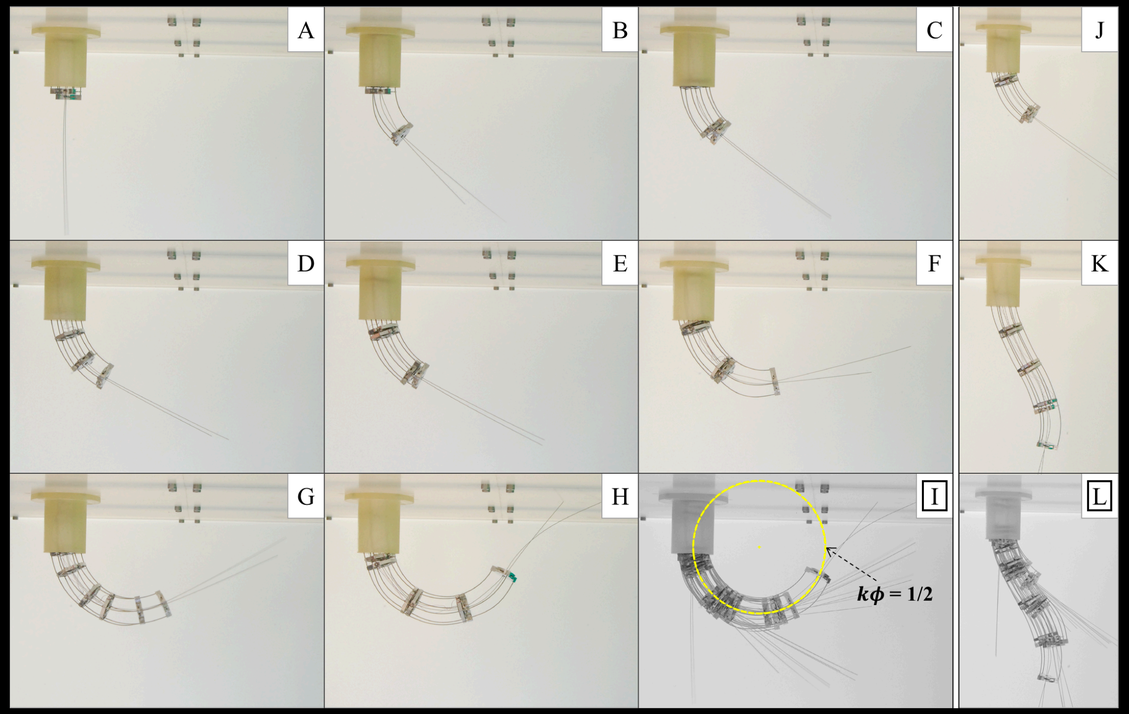
Motivation

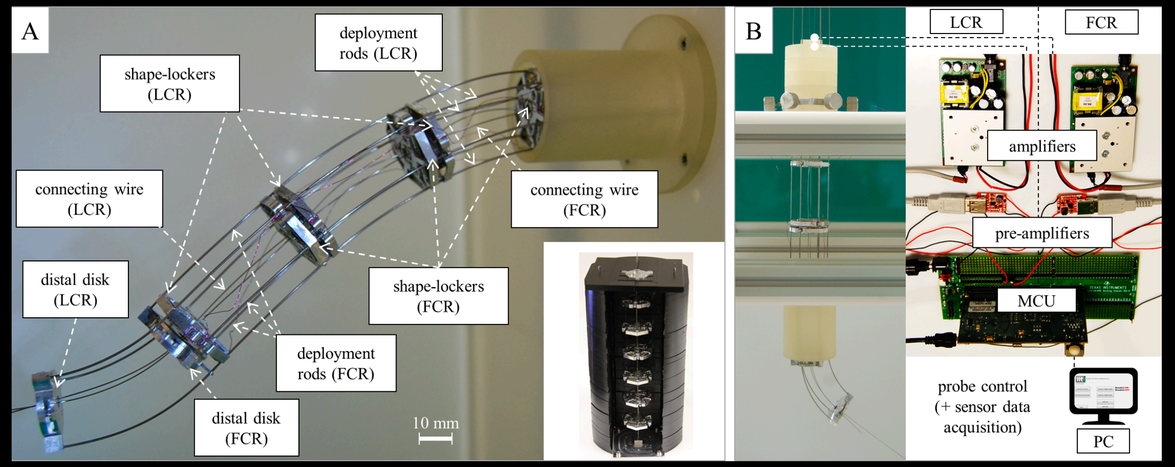


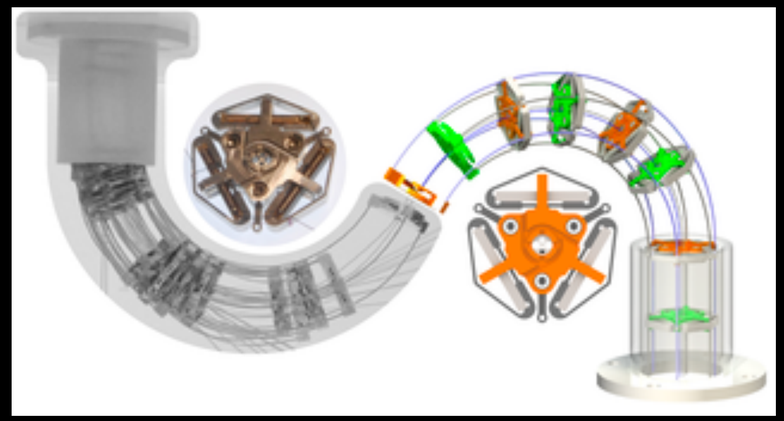
Considerations for Follow-the-Leader Motion of Extensible Tendon-driven Continuum Robots
Maria Neumann and Jessica Burgner-Kahrs, Member, IEEE, May 2016

UNIVERSITY OF TORONTO

Motivation



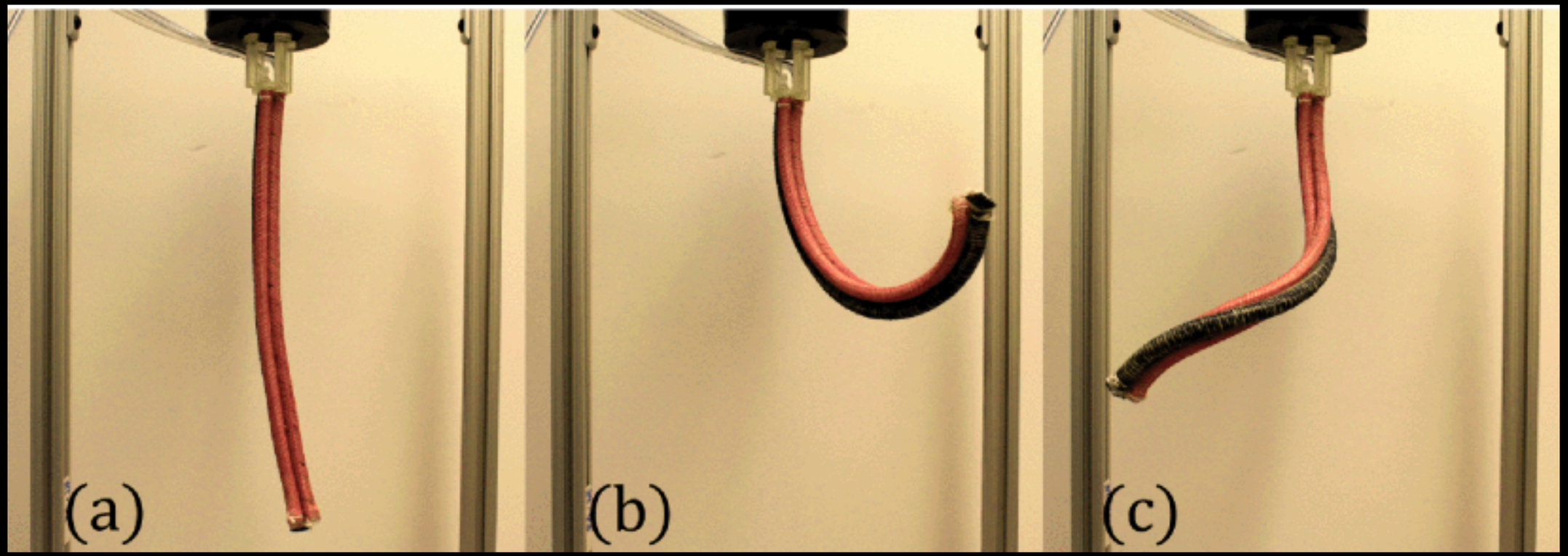




The First Interlaced Continuum Robot, Devised to Intrinsically Follow the Leader
Byungjeon Kang, Risto Kojcev, Edoardo Sinibaldi | February 2016 (*Link*)

UNIVERSITY OF TORONTO

Motivation



Open Loop Position Control of soft continuum arms using Deep Reinforcement learning | ICRA 2019

Sreeshankar Satheeshbabu , Naveen Kumar Uppalapati , Girish Chowdhary and Girish Krishnan

UNIVERSITY OF TORONTO

Problem

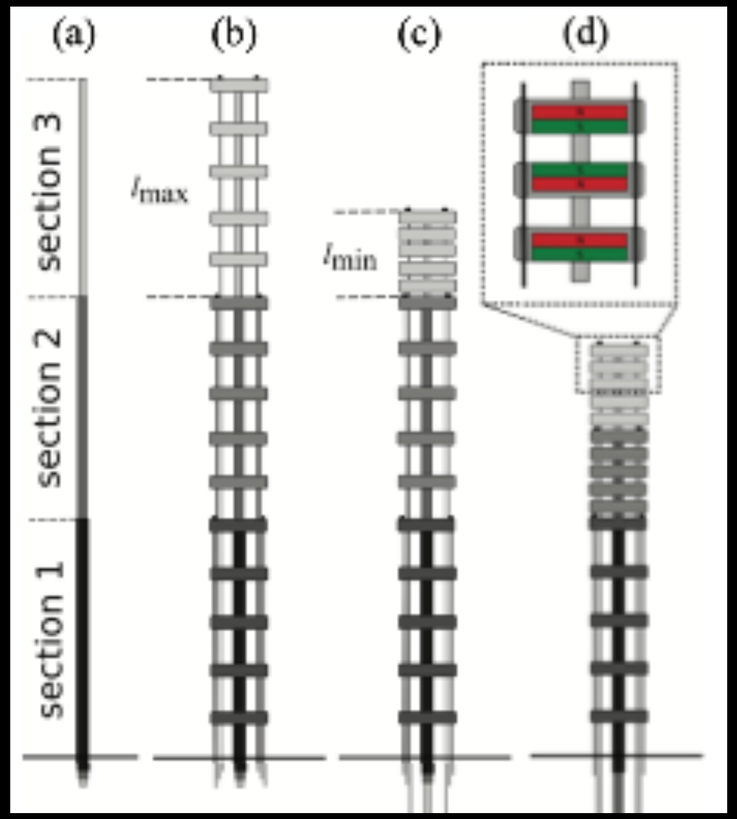
Designing an algorithm to achieve Follow-The-Leader behaviour is complicated

Problem

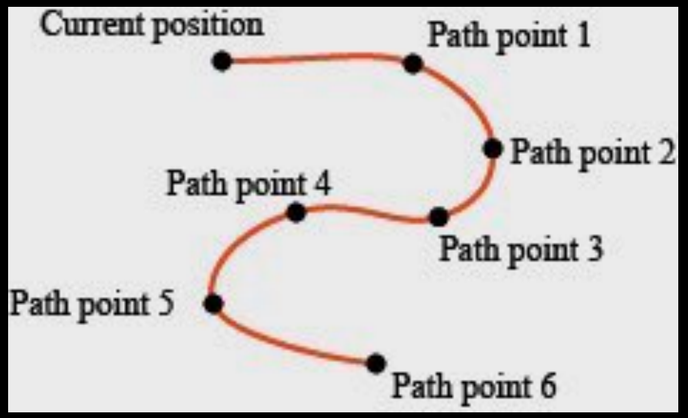
Designing an algorithm to achieve Follow-The-Leader behaviour is complicated

So we will try learning the Follow-The-leader behaviour with Machine Learning

Solution



+ **ML** =



Existing Continuum Robot Models

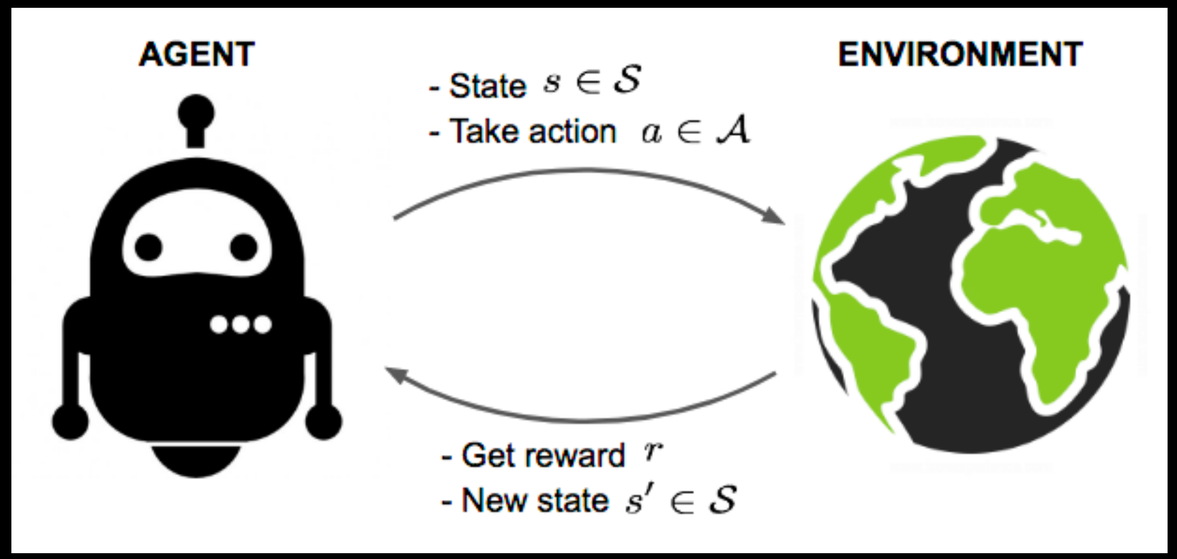
Machine Learning

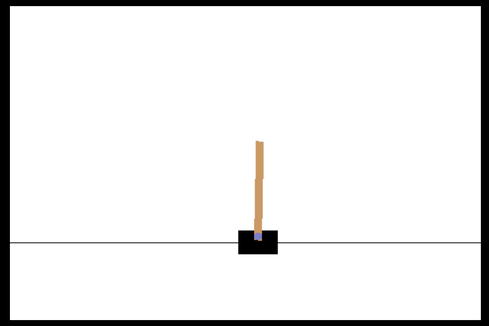
Follow the leader Model

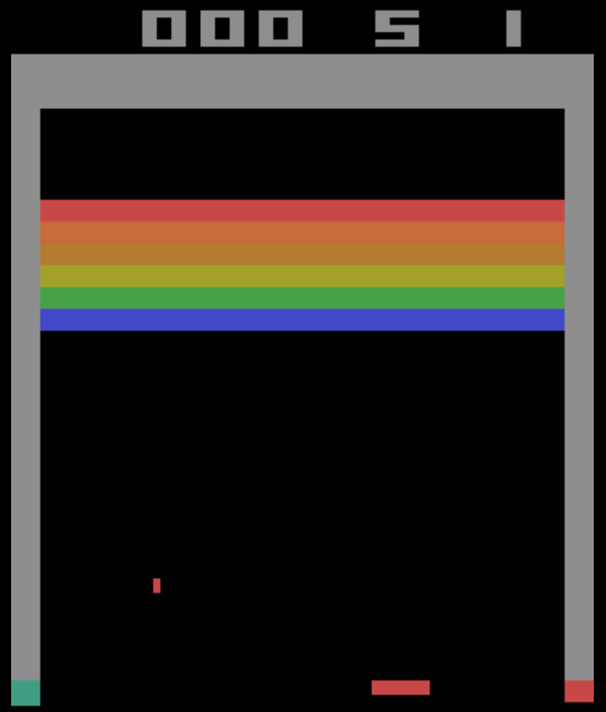
UNIVERSITY OF TORONTO

Solution

The Power of Reinforcement Learning (RL)



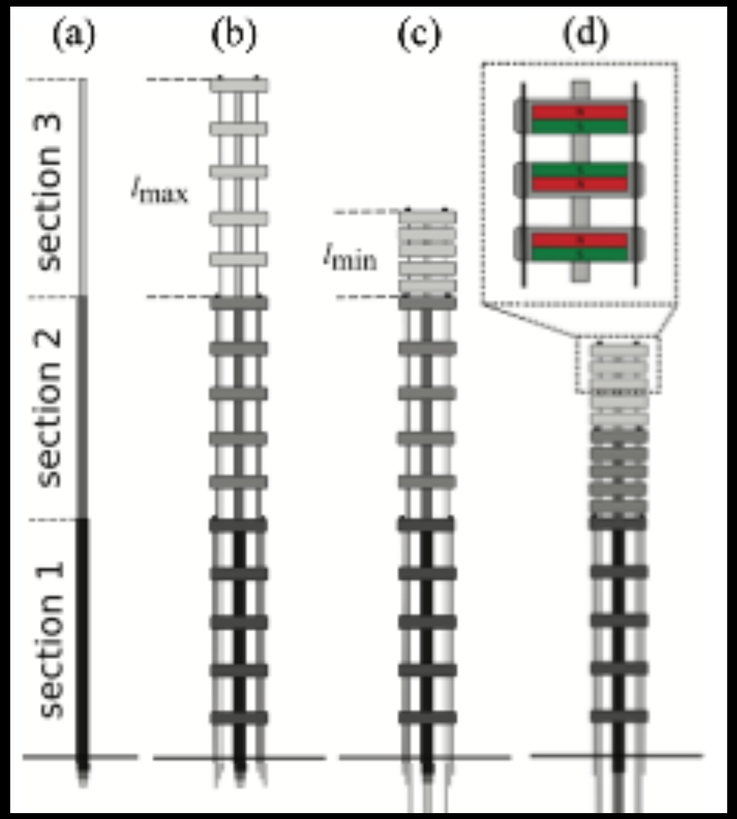




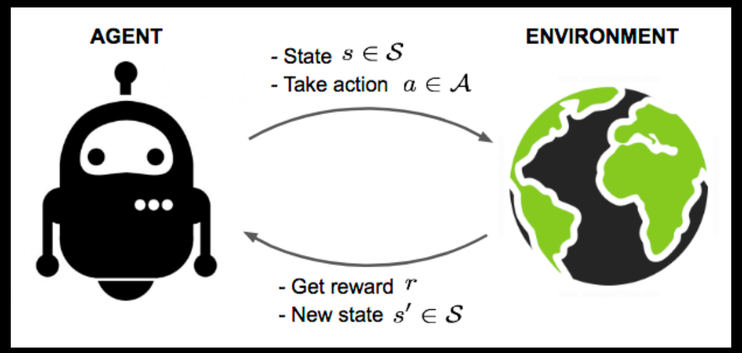
"What we want is a machine to learn from Experience" - Allan Turing

UNIVERSITY OF TORONTO

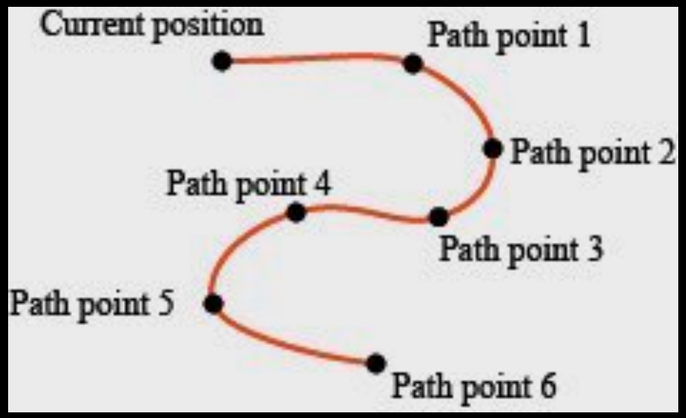
Solution



+



=



Existing Continuum Robot Models

Reinforcement Learning

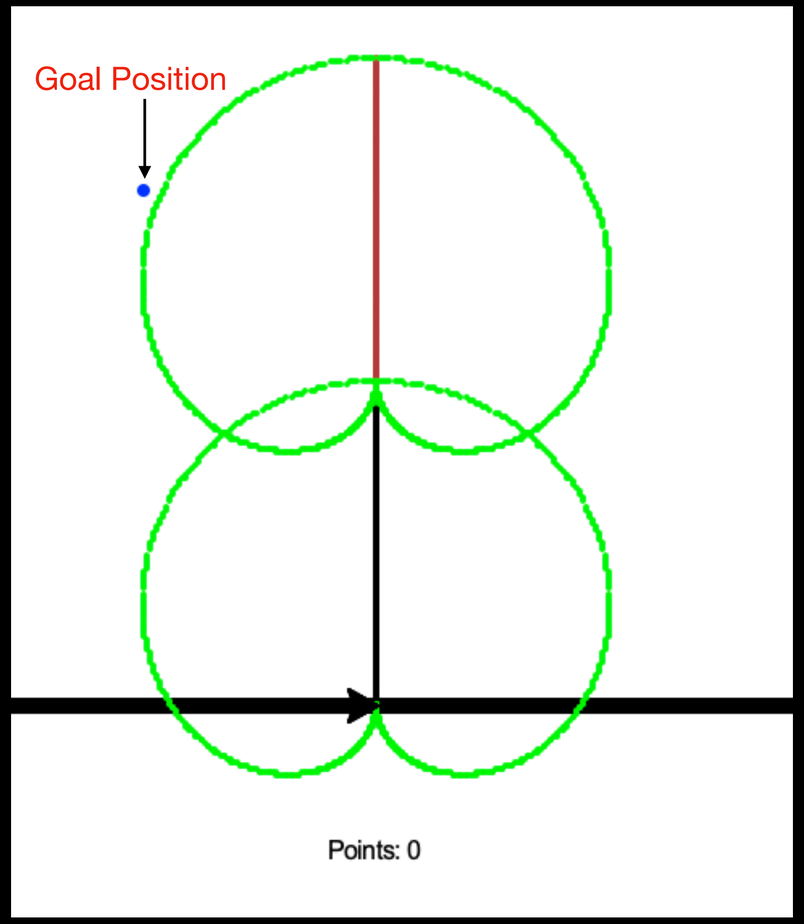
Follow the leader Model

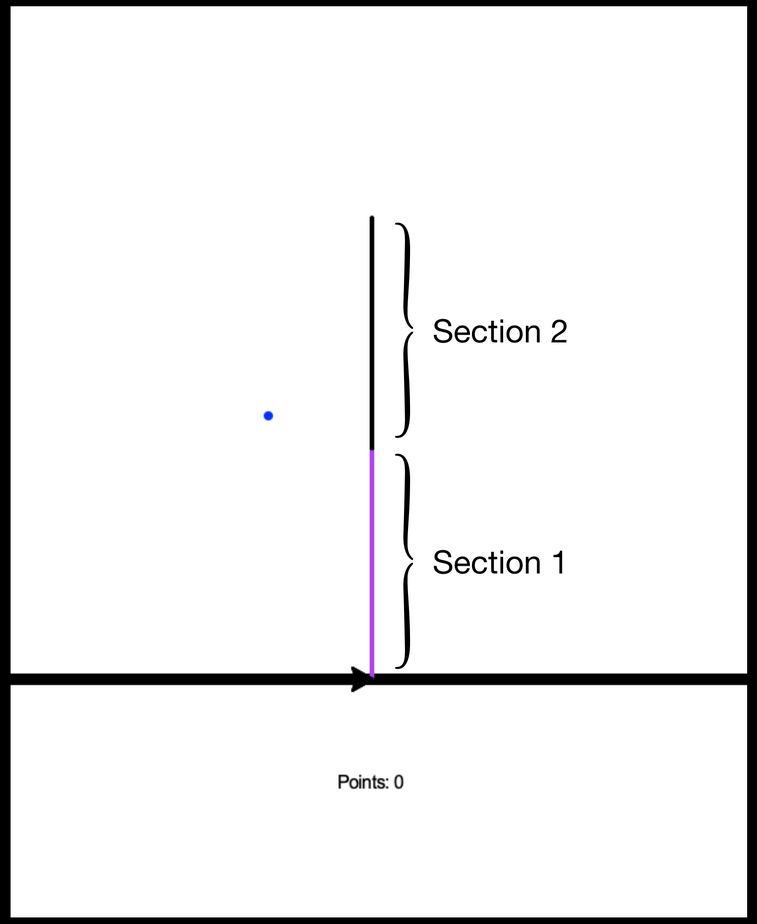
UNIVERSITY OF TORONTO

Project Implementation

Simulation

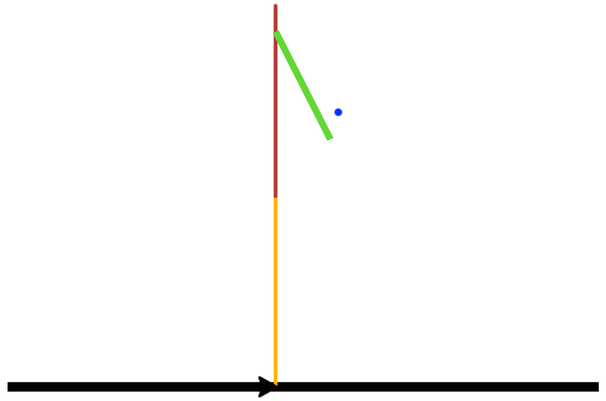
- 2D Dimension
- Dual Sections
- Point Based System
- Constant Curvature
- Actions
  - Left, Right, Extend, Contract





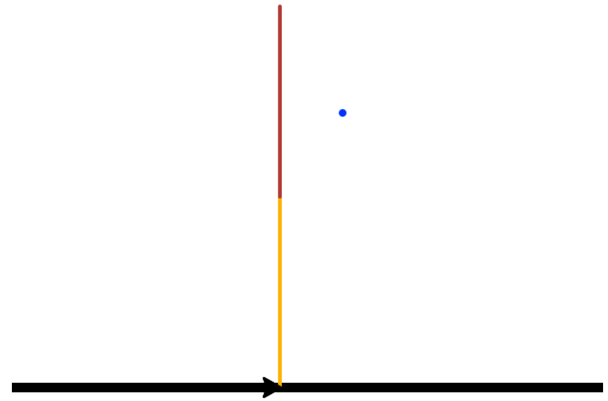
Project Implementation

Reward Policy



Negative distance Reward:

Negative Distance from Tip Position to Goal as the reward for each action



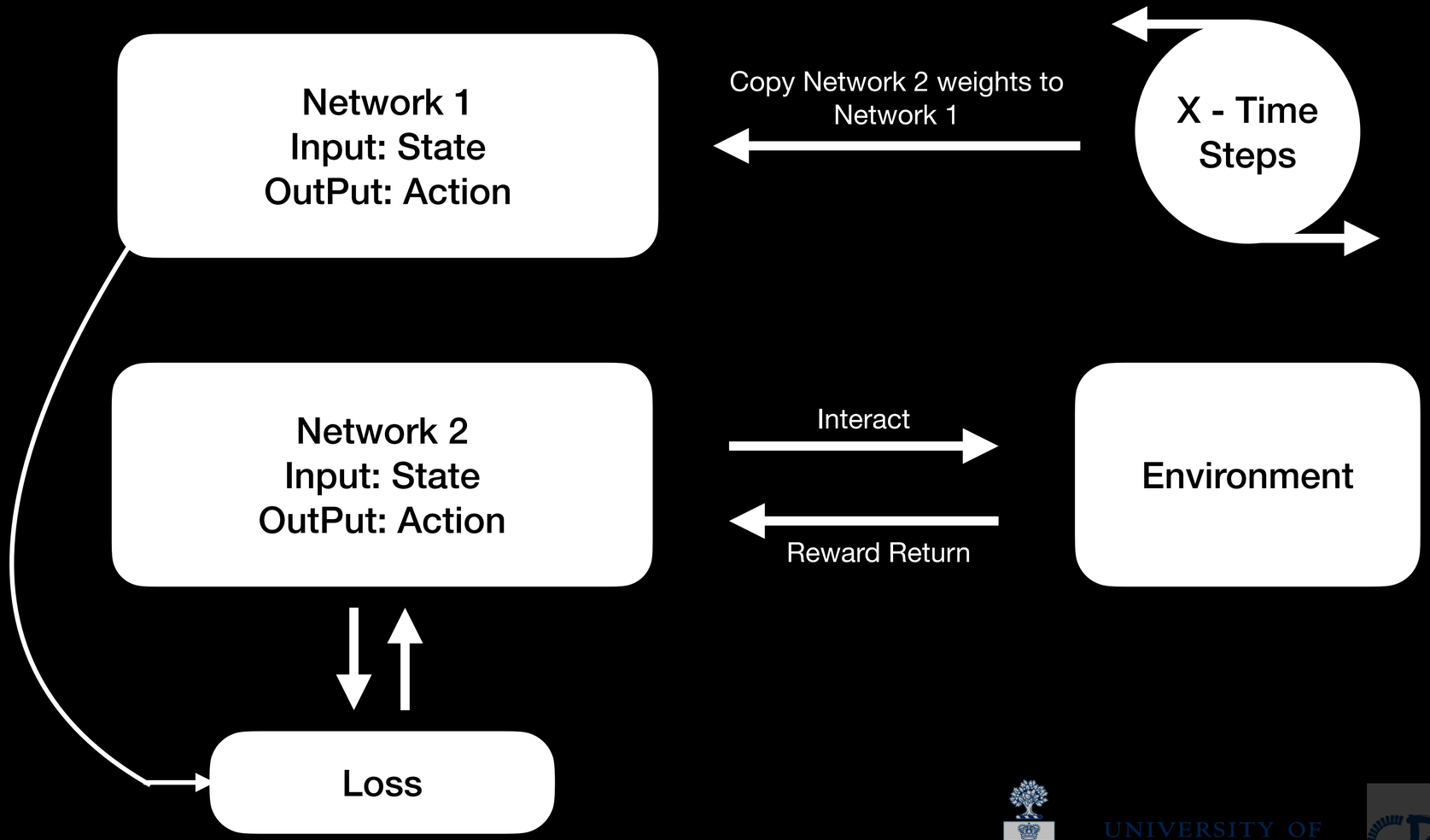
Constant Reward:

Distance Decreased: +1
Distance Increased: -1
Reaching Curvature Limit: -2

Project Implementation

Model 1: Double Deep Q Network (DDQN)

- Future Action Estimation
- Avoids Overestimation
- Replay Buffer

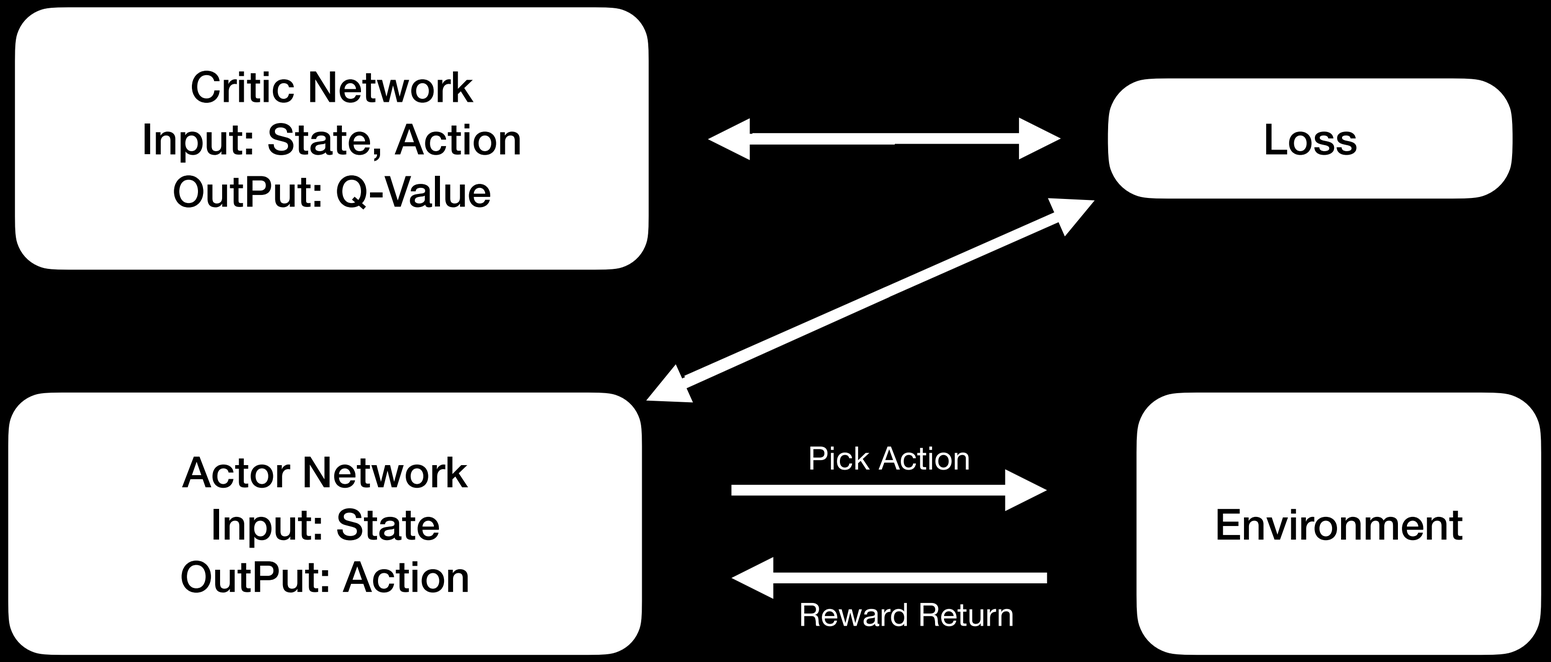


UNIVERSITY OF TORONTO

Project Implementation

Model 2: Actor Critic (A2C)

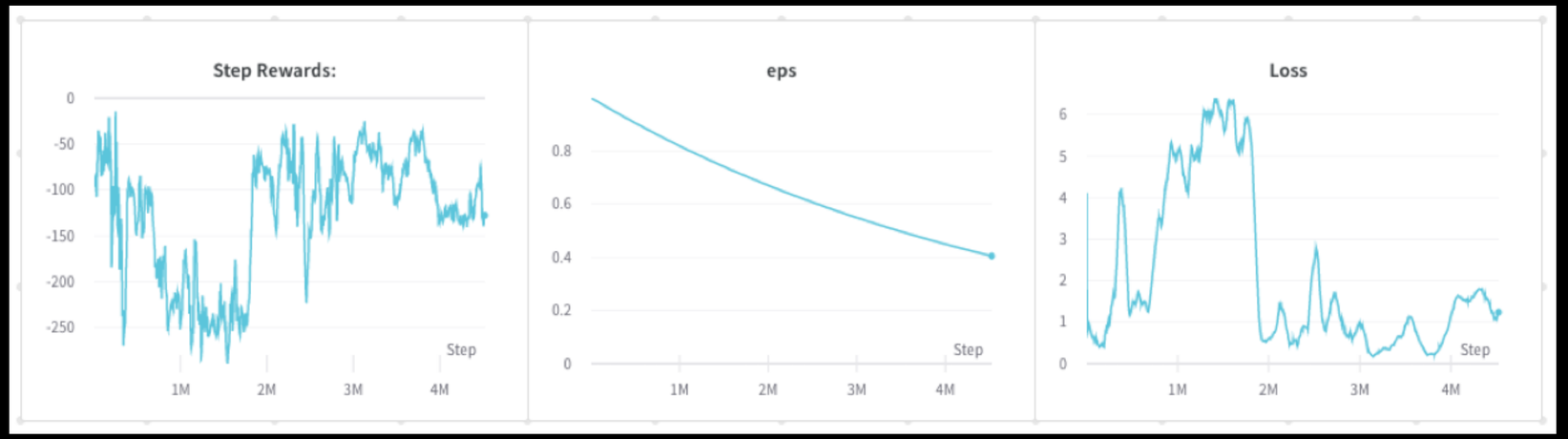
- Continuous Task Efficient
- Action Criticism

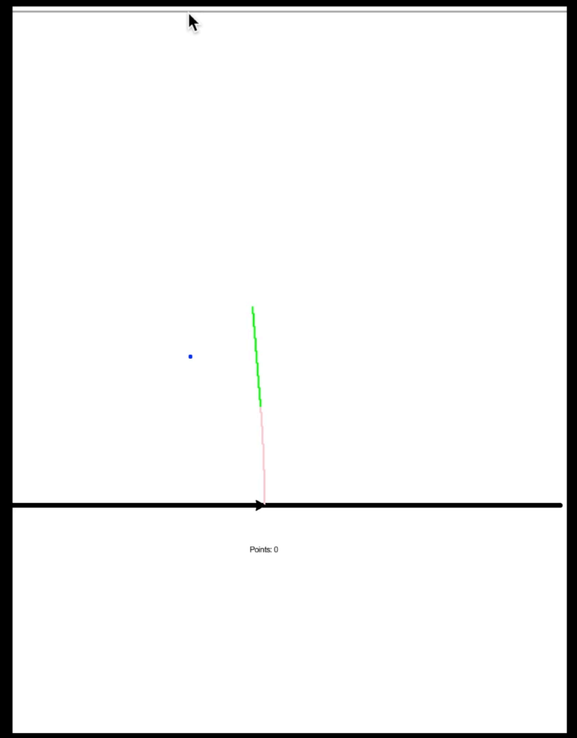


Results

Model 1: DDQN

- Negative Distance as Reward
- 6000 Epoch/Model Updates



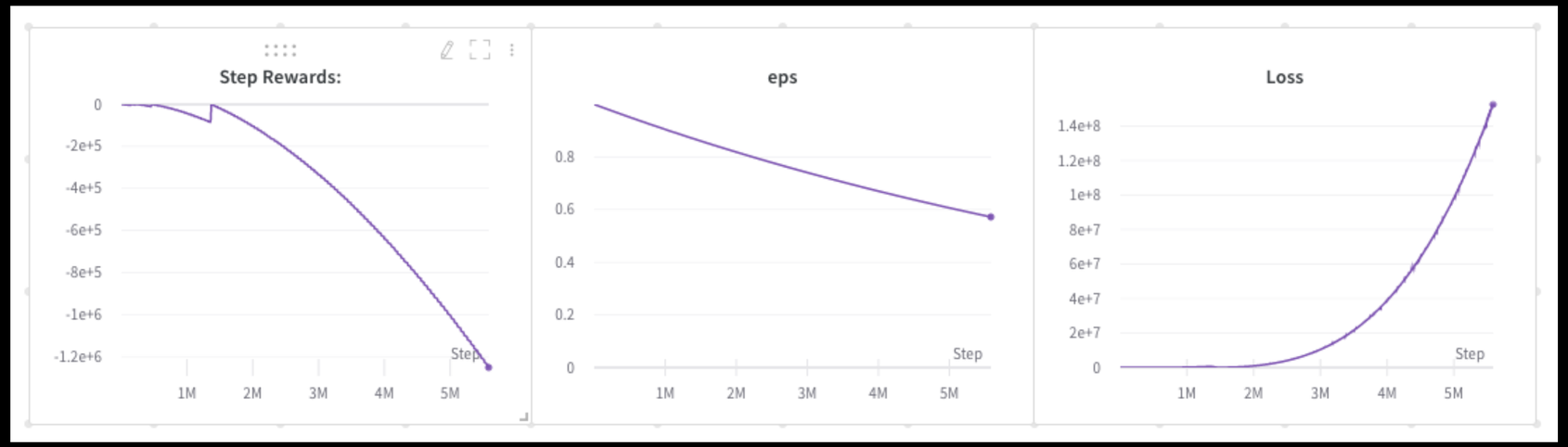


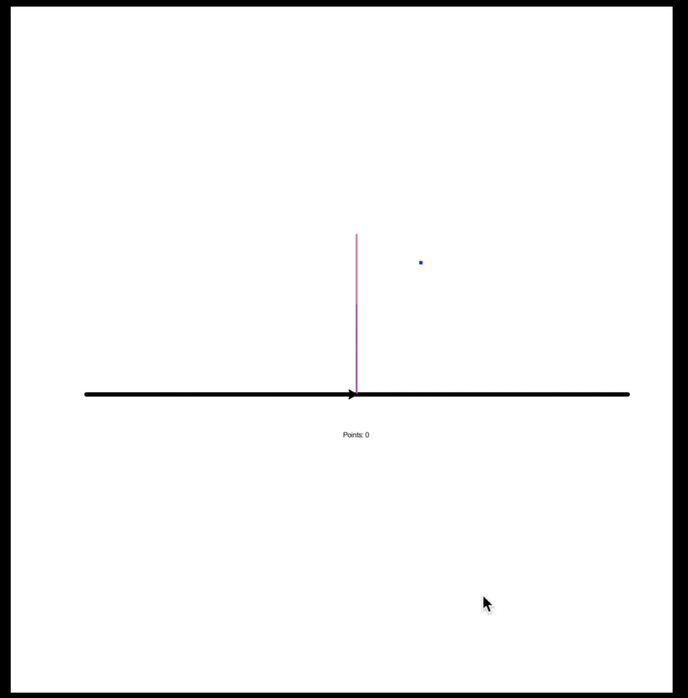
Issue: Taking too long to converge

Results

Model 1: DDQN

- Constant Reward
- 7000 Epoch/Model Updates





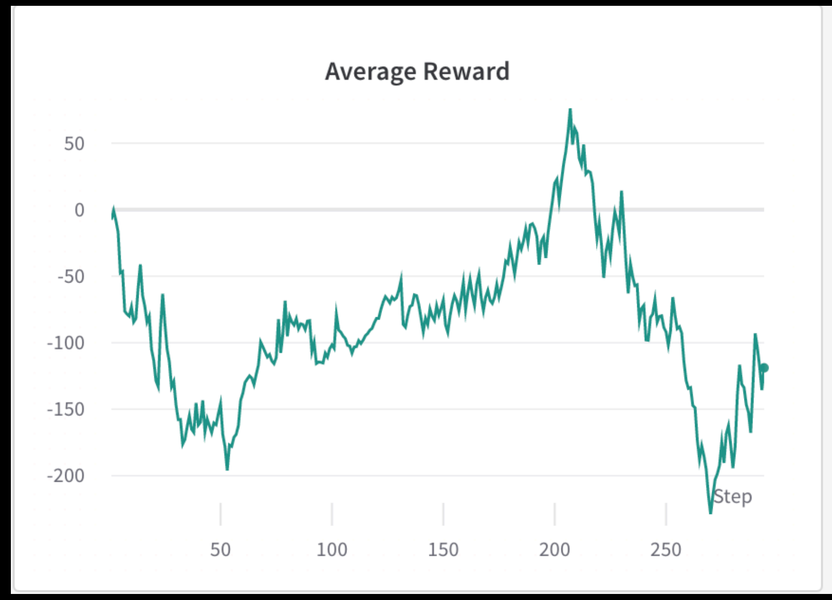
Possible Issue: Calculating expected reward too far into the future

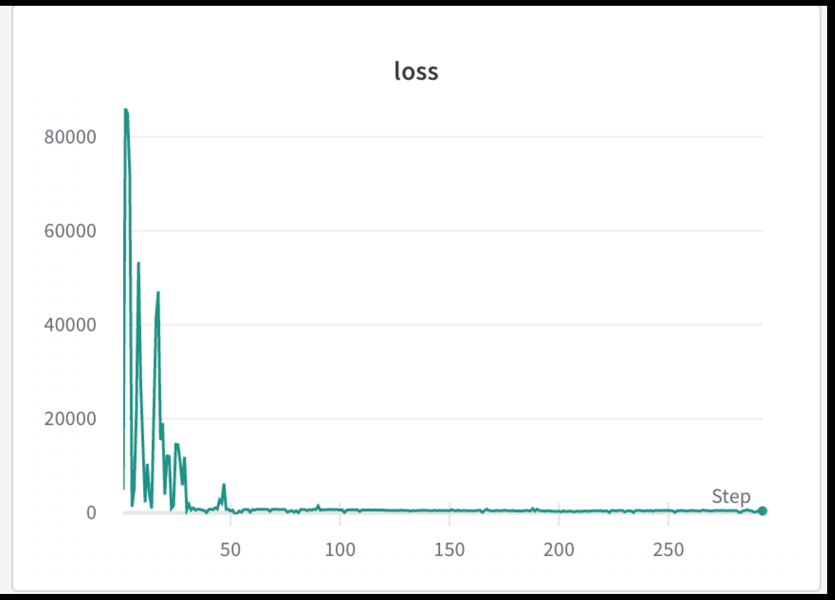
UNIVERSITY OF TORONTO

Results

Model 2: Actor Critic

- Constant Reward
- 250 Episodes

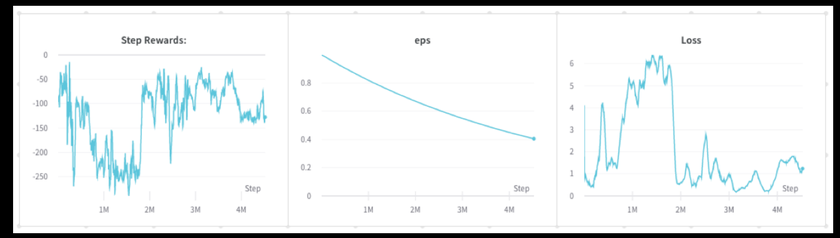


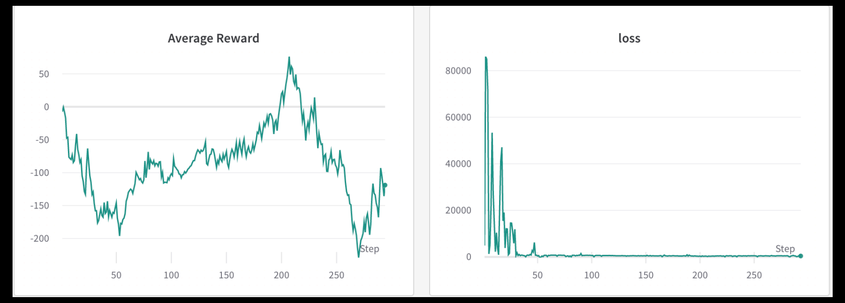


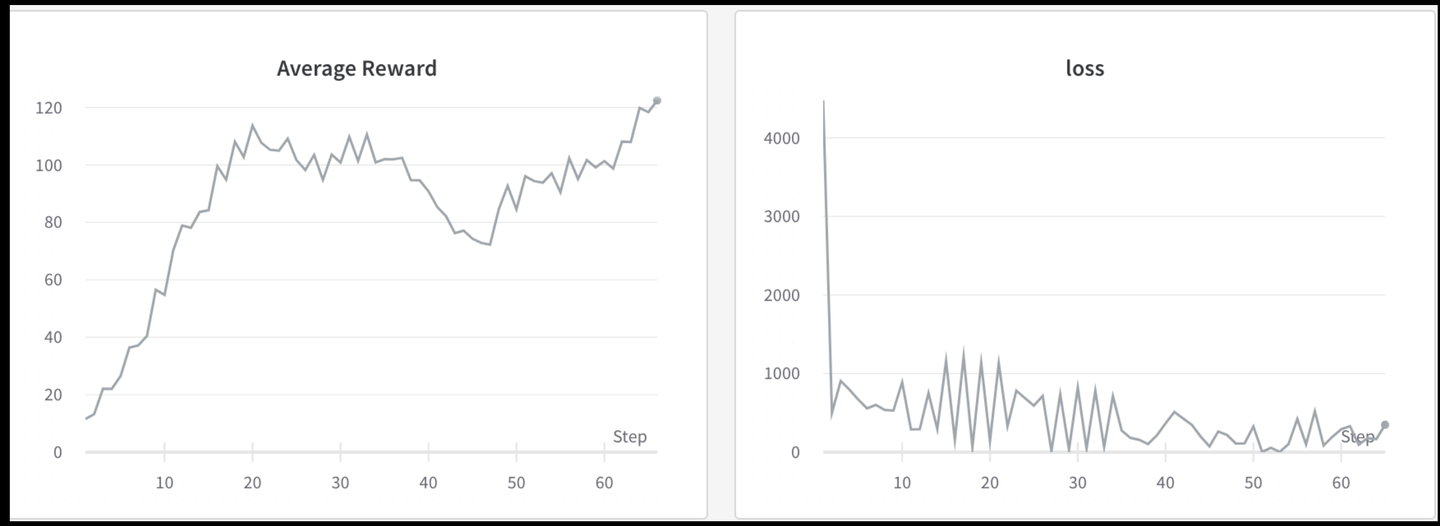
Issue: Numerical Errors

Discussion

- DDQN vs Actor Critic
  - Actor Critic
- Negative Distance vs Constant Reward
  - Constant Reward overall
- NaN Error



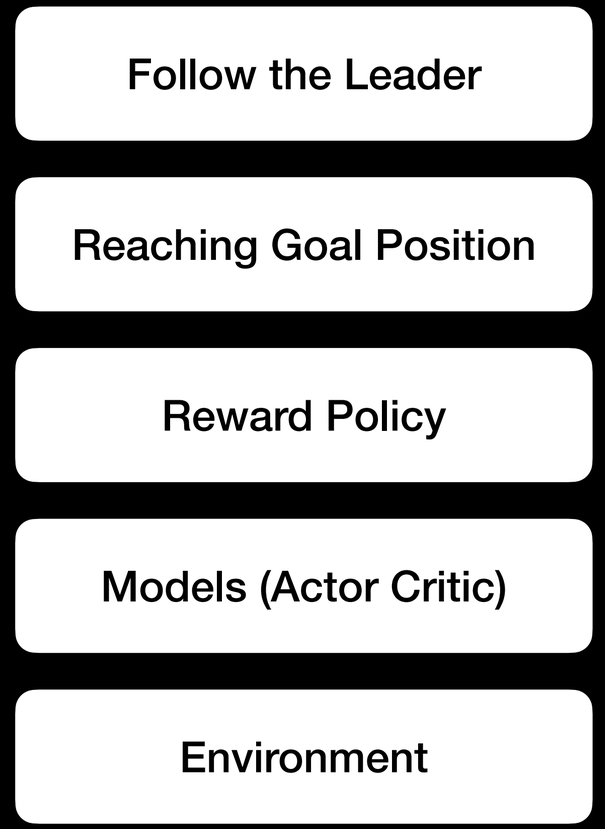




Actor-Critic | Step-Wise Reward System | Normalized Inputs

Conclusion

- Base Foundation is Complete
  - Model Structures and Base Reward Systems
- Sub Goal: Reach a Goal Position from an initial position
- Reward System Modifications

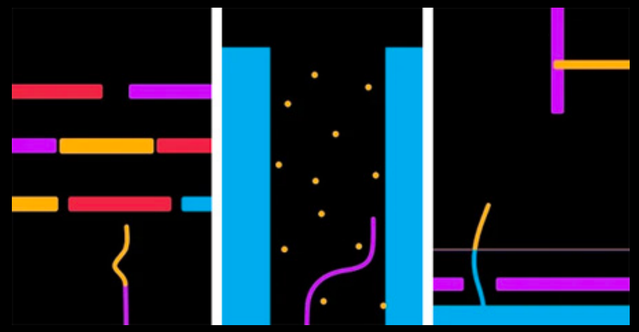


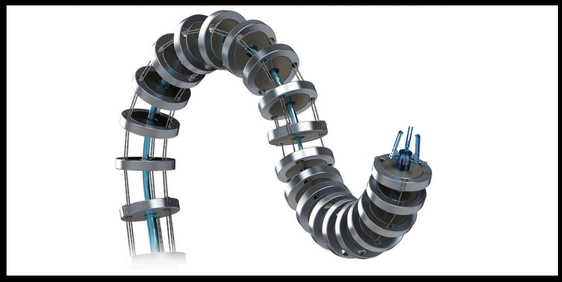
Next Steps

- Resolving NaN issue
- Reward Function Optimization
- Environmental Complexity
- 3D Simulation

DDQN | Negative Distance Reward

DDQN | Constant Reward





UNIVERSITY OF TORONTO

Thank You for Listening

Q/A

UNIVERSITY OF TORONTO