239AS Project Proposal S2021

Samuel Gessow	Sunay Bhat	Yi-Chun Hung	Vahe Gyuloglyan
sgessow@ucla.edu	sunaybhat1@ucla.edu	yichunhung@ucla.edu	vgyulogl@ucla.edu
604781350	905629072	705428593	905528327

Abstract

This proposal outlines a novel take on the standard cart-pole reinforcement learning problem we intend to investigate for our course project. We will explore a multi-agent version in which two carts operate two poles independently that are jointed together and support a pendulum. This extension will allow us to explore the standard cart-pole problem with considerably more complex dynamics. We will further investigate the multi-agent aspect and more complex rewards structures beyond the basic cart-pole problem. Our intention is to expand the same algorithmic techniques for more complex dynamics, and then further the exploration into different rewards structures that can arise with multiple agents.

1 Problem Outline and Background

The problem of a pendulum on a cart or cart pole has been well studied in reinforcement learning (RL) [1] and is one of the widely accepted test environments for a new RL methodology. However this setup has simple dynamics and does not allow for multi agent behaviors. In our project, we propose an extension of the classic problem that will allow the exploration of a more complex physical system as well as multi-agent RL and the collaboration and competition dynamics that can arise. We will develop a new environment consisting of two joint carts as shown in Fig. 1 below. Within this environment, we first want to validate that the pole can be stabilized and that successful "swing-up" (pendulum starting position is at rest pointed down) is possible using the common model-free algorithms used in the traditional cart-pole environment [1]. In order to do this, we will initially use a single-agent setup which has two sets of two actions available for each cart. We would then like to have each cart be an independent agent and see if collaborative behavior can be learned to achieve the same reward which can be far less trivial than we might intuit [3]. Finally, those two agents can be presented with a challenge that forces both collaboration and competition in varying degrees. A simple example might be balancing the pole but having each agent receive a higher or lower reward based on bisecting the space the third joint (under the pendulum) is located. This environment presents a multitude of extensions for our project to expand on the fundamentals and utilize our RL algorithmic knowledge in complex ways.

Our initial literature search focused on extensions of the cart-pole problem, particularly double [2] and triple pendulums [4]. There is plenty of research in more complex cart-pole or pendulum problems, but few we found that explore a more complex variation along with multiple agents. In order to combine the two, it became clear that we needed a system that will not overwhelm any simulation software with overtly complicated dynamics but also offer the needed flexibility to explore two agent interactions. After a review of existing successes and failures in the literature and TA consultation, we settled on this system for the right balance. We will initially reference our goals and performance baselines against an inverted double-pendulum extension [2].

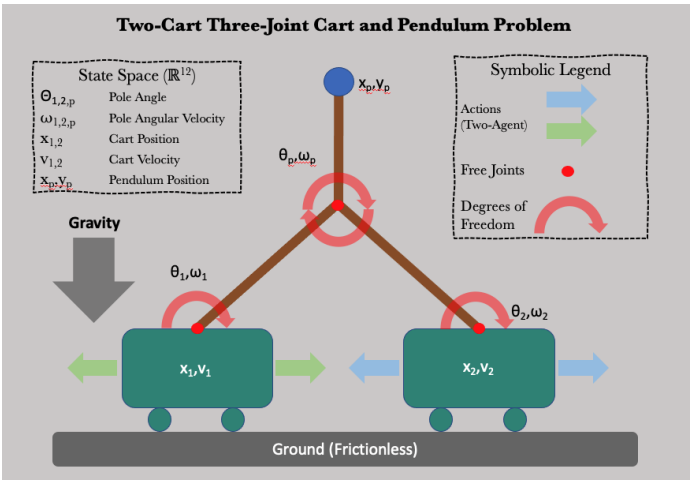

Figure 1: The illustration of our experiment system, including two carts, three joints and pendulums.

2 Proposed Method

In order to implement our environment, we will utilize existing simulation software. We are currently building simulations in multiple python packages, but will likely settle on Pymunk to handle the physics simulation. We will combine this with OpenAi Gym to provide the RL framework for the cleanest implementation. Like most model-free RL based approaches to the cart-pole problem, we will be fully online and rely entirely on data generated in simulation. Based on existing literature and our course work, we will focus on a model-free Q-learning, function approximation based approach [5]. This will allow us to deal with a continuous state space and complicated kinematics in a tractable way. In addition, we will only allow for a discrete action space, likely a set of fixed forces or single force action with only left/right direction vectors. This will further reduce the complexity of the system.

Initially our focus will be on proving the system can be stable. This will involve first demonstrating the pole can be held upright for a target period of time with a single-agent formulation that selects actions for both carts and studying convergence properties for both the upright start and "swing-up". We will then reformulate the problem as a multi-agent system where each agent independently receives a reward. If we are able to stabilize this formulation to explore such dynamics in a meaningful way, we will proceed to more unique reward structures. Although we have many ideas for this phase, we will initially vary the agents rewards such that they will have collaborative and competitive incentives such as varying rewards based on the position of stability. We will let the multi-agent dynamics dictate further exploration as well implement these reward structures.

3 Evaluation Techniques

We will stick to a few common standards in order to ensure our results are valid and meaningful. Our time step will stay within the multi-millisecond range as most analysis do [1], and we will require a 2-5 minute period of the pole being within a strict degree range to declare successfully achieving stability. As our implementation will likely include epsilon-greediness and linear value-function approximation or neural networks, we will report any hyper-parameter and parameter optimization in comparison and convergence plots. The actual simulation will require a visual structure, so we can display the environment states as images for reporting as well as associated videos to show the dynamics. Our primary goal is to report results on multi-agent behavior with the same reward structure as well as unique reward structure. This will entail a multitude of plots and figures to study convergence of state variables, behavior correlation, action variable analysis, and many other dynamics that might be of interest after testing. We will baseline our result reporting against the most popular RL problems explored to ensure consistency with existing literature and confirm the value of our methods [5].

References

[1] Kumar, S. *"Balancing a CartPole System with Reinforcement Learning – A Tutorial"*, arXiv e-prints, 2020.

[2] Gustafsson, F. (2016). Control of Inverted Double Pendulum using Reinforcement Learning. 1–6.

[3] Barton, S. L., Waytowich, N. R., Zaroukian, E., & Asher, D. E. (2019). Measuring Collaborative Emergent Behavior in Multi-agent Reinforcement Learning. Advances in Intelligent Systems and Computing, 876(Ccm), 422–427.

[4] Tobias GlüCk, Andreas Eder, and Andreas Kugi. 2013. Swing-up control of a triple pendulum on a cart with experimental validation. Automatica 49, 3 (March, 2013), 801–808. DOI:https://doi.org/10.1016/j.automatica.2012.12.006

[5] Sutton, R. S., Barto, A. G. (2018). Reinforcement Learning: An Introduction. The MIT Press.