RL-Course 2025/26: Final Project Report

Ansel Cheung, Jannik Mänzer, Niklas Abraham

January 26, 2026

Abstract

This report presents our application and comparative study of modern Reinforcement Learning (RL) algorithms—specifically, Twin Delayed Deep Deterministic Policy Gradient (TD3), Soft Actor-Critic (SAC), and the model-based TD-MPC2—within the challenging Laser Hockey environment. We discuss the environment's design, its state and action spaces, and the unique characteristics making it a compelling testbed for RL research. Our work details the methodological approaches taken with these algorithms, summarizes key experimental findings, and reflects on lessons learned in this multi-agent, continuous control setting.

1 Introduction

1.1 Environment Overview

The Laser Hockey environment [**?**] is a custom reinforcement learning benchmark built on the Gymnasium Python API and powered by the Box2D physics engine. In this multi-agent setting, two agents each control a hockey stick and compete to score goals by striking a puck into the opponent's net. The environment is designed such that both agents receive identical but mirrored observations at each timestep, which eliminates the need for side-specific learning strategies.

$$s_t = (\underbrace{x_1, y_1, \theta_1, v_{x,1}, v_{y,1}, \omega_1}_{\text{Player 1}}, \underbrace{x_2, y_2, \theta_2, v_{x,2}, v_{y,2}, \omega_2}_{\text{Player 2}}, \underbrace{x_p, y_p, v_{x,p}, v_{y,p}}_{\text{Puck}}, \underbrace{t_{\text{puck},1}, t_{\text{puck},2}}_{\text{Puck possession time}})$$

(1)

At each timestep, the agent receives an 18-dimensional continuous state vector that captures the complete game state. This observation includes the position and orientation of both players relative to the center of the field, along with their linear and angular velocities. The state also contains the puck's position and velocity, as well as time remaining for each player's puck possession in keep mode, which ranges from 0 to 15 seconds. Player 1 refers to the agent currently being controlled, while Player 2 represents the opponent. The observation structure ensures that when the viewpoint switches during training, the indices are automatically mirrored so that each agent always perceives itself as Player 1, maintaining a consistent learning perspective.

The action space consists of a 4-dimensional continuous vector that controls the agent's stick. The first two components specify forces applied for stick movement in the x and y directions, while the third component controls the torque applied to adjust the stick's angle. The fourth component is a thresholded scalar that determines whether to release the puck when the agent is in possession.

Reward feedback in this environment is sparse and goal-oriented. An agent receives a reward of +10 for scoring a goal, -10 for conceding one, and 0 for a draw. Additionally, there is a small reward signal

for maintaining proximity to the puck, which helps guide exploration during early learning stages. Each episode begins with all entities positioned at the center of the field and terminates when a goal is scored or a timeout occurs. The environment supports both scripted opponents, such as the BasicOpponent, and learning agents as adversaries, making it a versatile and challenging benchmark for continuous control and multi-agent reinforcement learning research.

2 Related Work

3 Method

3.1 TD-MPC 2 - Niklas Abraham

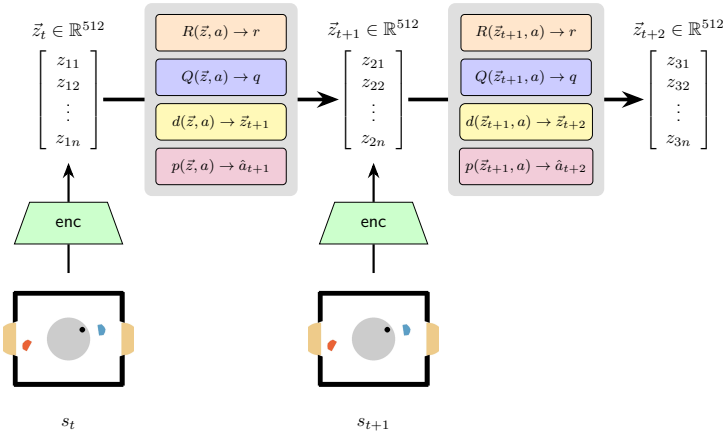

Figure 1: TD-MPC2 agent architecture, the reward, Q-value, dynamics, and action heads are grouped together in the world model, and the latent space flow is shown in the background.