# RL-Course 2025/26: Final Project Report

abcdef: Julian Jurcevic

February 15, 2026

## 1 Introduction

Reinforcement Learning (RL) enables agents to learn decision-making strategies through interaction with an environment. At each time step, the agent observes the current state, selects an action, and receives a scalar reward. The objective is to maximize the expected cumulative reward over time. Continuous control tasks are particularly challenging due to high-dimensional action spaces and instability in value estimation.

In this project, we consider a competitive air hockey environment implemented in Gymnasium [2] and based on a physics simulation using Box2D [1]. Two agents compete in a zero-sum setting. Each player controls translation, rotation, and shooting through continuous actions. The observation space consists of positional and velocity information for both players and the puck. Episodes terminate when a goal is scored or when a time limit is reached.

The agent is trained by maximizing the reward signal provided by the environment. The reward includes goal outcomes and additional shaping components. Performance is evaluated separately using the win rate against predefined weak and strong opponents.

The goals of this work are:

1. Implement the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm for continuous control in the hockey environment.
2. Analyze the effect of different replay strategies and exploration noise models.
3. Investigate the impact of self-play and pretrained initialization on learning stability and final performance.
4. Evaluate the learned policy using reward progression and win rate against fixed opponents.

## 2 Method

### 2.1 Twin Delayed Deep Deterministic Policy Gradients

Twin Delayed Deep Deterministic Policy Gradients (TD3) [3] is an off-policy actor-critic algorithm for continuous action spaces. It improves DDPG by reducing overestimation bias and stabilizing training.

TD3 uses two independent critic networks, $Q_{\phi_1}$ and $Q_{\phi_2}$. The Bellman target is computed using the smaller of the two target Q-values:

$$y = r + \gamma(1-d)\min_{i=1,2} Q_{\phi'_i}(s', a'(s')).$$

This technique is known as clipped double Q-learning. It reduces systematic overestimation.

TD3 further applies delayed policy updates. The actor is updated less frequently than the critics. This prevents unstable feedback between policy and value estimates.
In addition, TD3 uses target policy smoothing. Noise is added to the target action:

$$a'(s') = \text{clip}\left(\pi_{\theta'}(s') + \text{clip}(\epsilon, -c, c)\right), \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

This smooths the Q-function with respect to actions and reduces exploitation of sharp value errors.
The actor is trained to maximize the critic estimate:

$$\max_{\theta} \mathbb{E}_{s\sim\mathcal{D}} \left[ Q_{\phi_1}(s, \pi_\theta(s)) \right].$$

Together, these three modifications make TD3 significantly more stable than standard DDPG in continuous control problems.

## References

[1] E. Catto. Box2d physics engine documentation. `https://box2d.org/documentation/md_simulation.html`, 2023. Accessed: 2026-02-15.

[2] Farama Foundation. Gymnasium: A standard api for reinforcement learning environments. `https://github.com/Farama-Foundation/Gymnasium`, 2023. Accessed: 2026-02-15.

[3] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1587–1596, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.