RL-Course 2025/26: Final Project Report

abcdef: Julian Jurcevic

February 14, 2026

1 Introduction

2 Method

2.1 Twin Delayed Deep Deterministic Policy Gradients

Twin Delayed Deep Deterministic Policy Gradients (TD3) [1] 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_{ϕ_1} and Q_{ϕ_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] 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.