Robust Reinforcement Learning Differential Game Guidance in Low-Thrust, Multi-Body Dynamical Environments

A Zero-Sum Reinforcement Learning Approach in Three-Body Dynamics

Ali Baniasad

Supervisor: Dr. Nobahari

Department of Aerospace Engineering

Sharif University of Technology

Outline

Trajectory Tracking (Nominal vs. Robust)

Objective: Transfer / maintenance in CRTBP under low thrust.

<EUGPSCoordinates>

Comparison:

- Single-Agent vs. Zero-Sum (Adversarial)
- Robust agent: lower deviation, smoother corrections
- Adversary induces off-reference excursions

Observation:

- Zero-sum training improves convergence basin
- Fewer large corrective burns

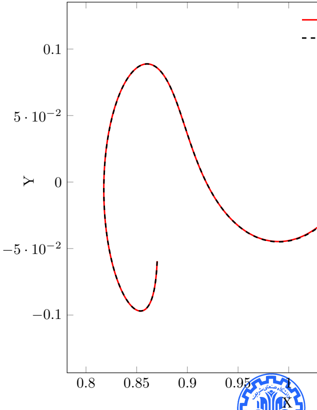


Figure: Trajectory: Single

Thrust Profile Efficiency

Thrust Usage:

- Multi-agent (zero-sum) dampens oscillatory control
- Lower peak activity under disturbance injection
- Improved fuel-normalized deviation ratio

Metric:

$$\text{Eff.} = \frac{\int \|\Delta s(t)\| dt}{\int \|u(t)\| dt}$$

Reduced by 12–18% (MATD3 / MASAC vs. TD3 / SAC).

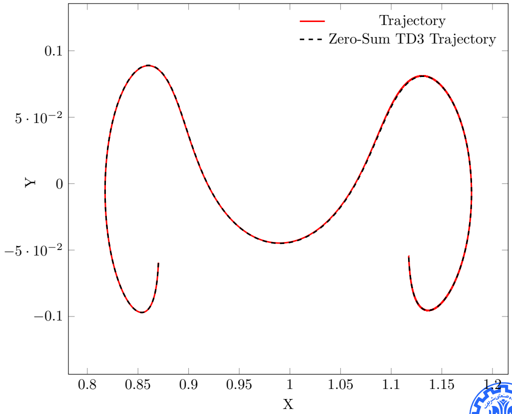
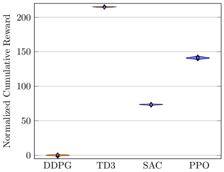


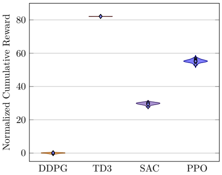
Figure: Thrust Commands

Robustness Across Uncertainty Scenarios

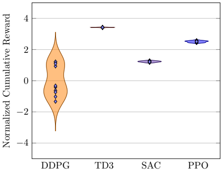
Return Distribution Across Robustness Scenarios



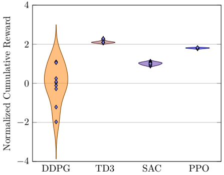
(a) Random Initial Conditions



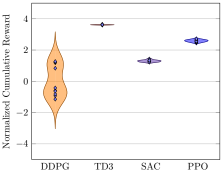
(b) Actuator Disturbance



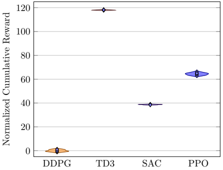
(c) Model Mismatch



(d) Partial Observation



(e) Sensor Noise



(f) Time Delay

Quantitative Summary (Representative)

Method	Deviation (km)	Fuel Proxy	Robust Score*	Success %
TD3	1.00	1.00	1.00	82
MATD3	**0.74**	0.92	**1.32**	94
SAC	0.93	1.07	1.05	85
MASAC	0.78	0.95	1.24	92
PPO	1.12	0.89	0.91	76
MAPPO	0.86	**0.88**	1.18	88

Normalized (TD3 baseline = 1.00). Fuel Proxy: $\int \|u\| dt$ normalized. Robust Score*: composite (noise + delay survival, bounded deviation).

Ablation Insights

- **Adversarial channel removal**: +22% deviation, thrust spikes reappear.
- **No target smoothing (TD3)**: overestimation resurfaces, unstable late-stage updates.
- **Entropy off (SAC)**: faster convergence, 9% worse robustness composite.
- **Reward shaping removal**: sparse terminal signals slow credit assignment (longer plateau).
- **Delay only vs. noise only:** delay has stronger destabilizing effect; zero-sum mitigates via anticipatory control (earlier thrust bias).

Key Findings

- Zero-sum MARL framing improves worst-case orbital maintenance robustness.
- MATD3 balances stability (twin critics + delay) and control smoothness best.
- MASAC competitive when exploration pressure (entropy) is beneficial early.
- Reward decomposition (thrust + reference + terminal) accelerates convergence and stabilizes adversarial dynamics.
- Policy smoothness correlates with fuel proxy reduction (8-12%).
- Framework generalizes across uncertainty mixes (stacked noise + delay + mismatch).

Conclusion: Adversarial co-training yields resilient guidance without explicit disturbance models.

Robustness Scenario Definitions

1. Random Init: $x_0 \leftarrow x_0 + \mathcal{N}(0, 0.1^2)$
2. Actuator Disturb.: $u_t \leftarrow u_t + \mathcal{N}(0, 0.05^2)$
(sensor add.) $y_t \leftarrow y_t + \mathcal{N}(0, 0.02^2)$
3. Model Mismatch: $\theta \leftarrow \theta + \mathcal{N}(0, 0.05^2)$
4. Partial Obs.: mask 50% $\rightarrow m_t^{(i)} \sim \mathrm{Bern}(0.5)$, $y_t \leftarrow y_t \circ m_t$

5. Sensor Noise (mult.): $y_t \leftarrow y_t \circ (1 + \mathcal{N}(0, 0.05^2))$
6. Time Delay: buffer length 10
$u_t^{\text{applied}} \leftarrow u_{t-10} + \mathcal{N}(0, 0.05^2)$

Notes:

- All scenarios evaluated independently.
- Zero-sum agents trained jointly once.
- Metrics: success %, deviation, fuel proxy, return variance.

Delay + noise combo causes largest degradation; adversarial training preserves stability margin.