Safe Multi-Agent Reinforcement Learning for Cooperative Drone Transport of Suspended Payloads

Toma Grundler Faculty Advisor: Dr. Siavash Farzan

Department of Electrical Engineering, College of Engineering, California Polytechnic State University, San Luis Obispo

INTRODUCTION

Decentralized multi-agent reinforcement learning (MARL) enables UAV teams to cooperatively transport cable-suspended payloads. However, learned policies do not inherently guarantee collision avoidance, and the coupled UAV–payload dynamics complicate safety enforcement. This motivates a decentralized safety filter that modifies policy commands to satisfy collision-avoidance constraints.

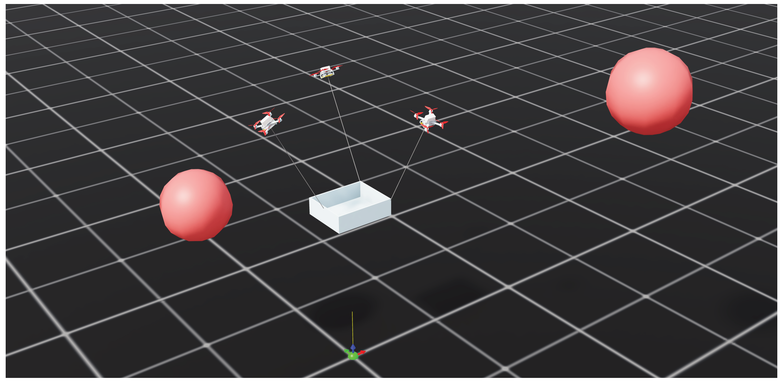


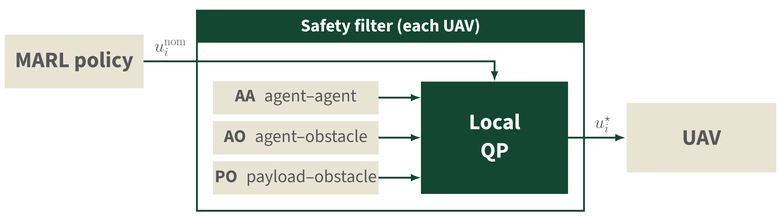
Fig. 1. Three UAVs cooperatively transport a cable-suspended payload in NVIDIA Isaac Sim.

OBJECTIVES

- Integrate a decentralized safety filter based on higher-order control barrier functions (HOCBFs) with a MARL transport policy.
- Enforce agent–agent, agent–obstacle, and payload–obstacle separation constraints using onboard and neighbor-relative measurements.
- Investigate whether conformally calibrated learned residuals reduce conservatism relative to analytic robust bounds.

SAFETY FILTER

At each control step, each UAV solves a local quadratic program (QP) that balances deviation from the MARL acceleration command against control smoothness, subject to collision-avoidance and actuator constraints.



$$u_i^\star = \arg\min_{u_i} \ \tfrac{1}{2}(u_i - u_i^{\mathrm{nom}})^\top W_i (u_i - u_i^{\mathrm{nom}}) + \tfrac{\lambda_{\mathrm{slew}}}{2}\|u_i - u_{i,\mathrm{prev}}^\star\|^2$$
$$\text{s.t.} \quad n_{iq}^\top u_i + c_{iq} \geq 0, \quad q \in \mathcal{A}_i,$$
$$u_{\min} \leq u_i \leq u_{\max}.$$

Each inequality in the active constraint set $\mathcal{A}_i$ corresponds to an agent–agent (AA), agent–obstacle (AO), or payload–obstacle (PO) collision pair.

HOCBF CONSTRAINTS

For collision points A and B, let $r = p_A - p_B$ and $v = v_A - v_B$ denote their relative position and velocity, and let R denote the minimum required separation. The barrier function and first auxiliary function are

$$h_0 = \tfrac{1}{2}\left(\|r\|^2 - R^2\right), \qquad \psi_1 = r^\top v + k_1 h_0.$$

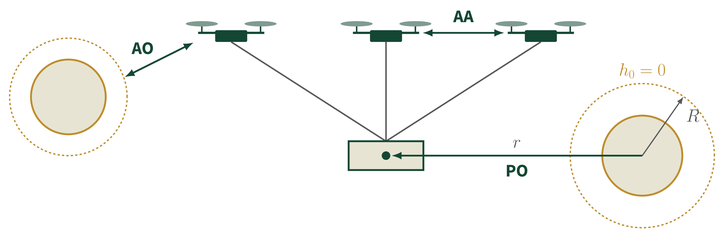


Fig. 2. Collision pairs considered by the safety filter: agent–agent (AA), agent–obstacle (AO), and payload–obstacle (PO). Dashed circles indicate the safety boundary, $h_0 = 0$.

Model the relative acceleration as $\dot{v} = Gu + e$: the commanded acceleration plus an unknown residual e that captures model mismatch. Substituting a lower bound $\underline{\rho} \leq \rho = r^\top e$ on the radial residual into the HOCBF condition $\dot{\psi}_1 + k_2\psi_1 \geq 0$ yields an affine inequality in the control input:

$$r^\top G u + \|v\|^2 + (k_1 + k_2)\, r^\top v + k_1 k_2 h_0 + \underline{\rho} \ \geq \ 0.$$

Residual lower bounds

Nominal	$\underline{\rho} = 0$
Robust	$\underline{\rho}_R = -\|r\|\,\epsilon$
Learned	$\underline{\rho}_{NN} = \|r\|\,(\hat{z} - \beta_z)$
Hybrid	$\underline{\rho} = \max(\underline{\rho}_{NN}, \underline{\rho}_R)$ if V_q, else $\underline{\rho}_R$

EXPERIMENTS

Parameter	Configuration
Simulator	NVIDIA Isaac Sim; three UAVs transporting one suspended payload
Obstacles	Two to five spherical obstacles in easy and hard configurations
Conditions	Unfiltered baseline and seven AA/AO/PO constraint subsets, each evaluated with nominal and robust bounds (15 total)

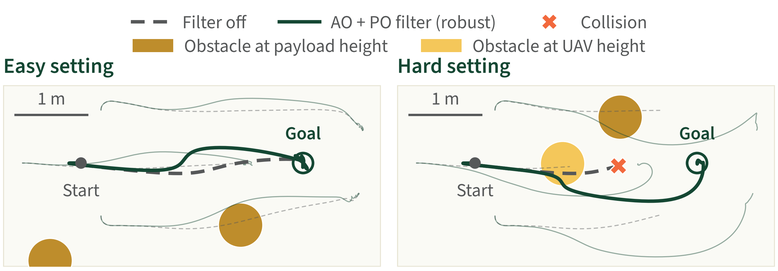


Fig. 3. Example episodes, top view: payload (thick) and UAV (thin) paths from the same start and obstacle layout, with the filter off and with AO + PO constraints.

RESULTS

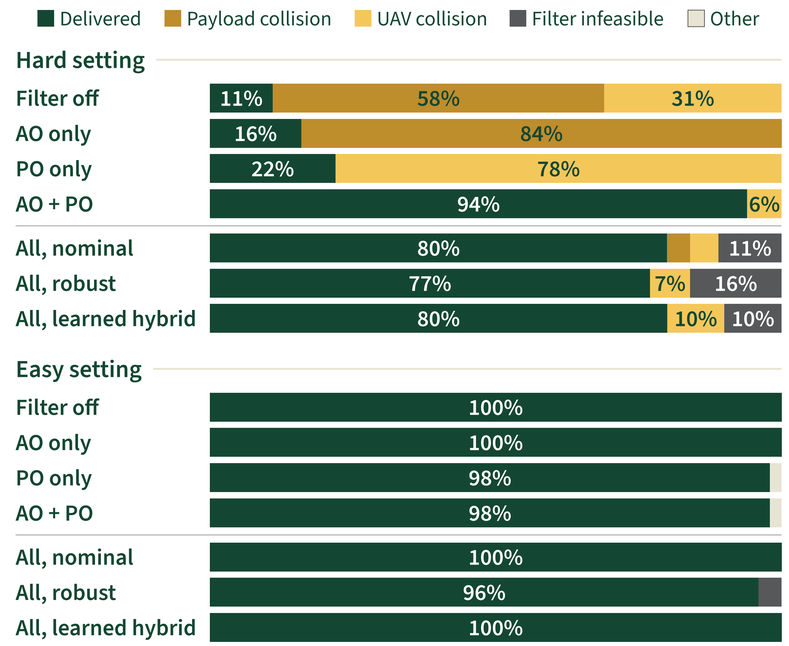


Fig. 4. Outcomes of 100 episodes per condition in the hard and easy obstacle settings (seed 0). The AO-only, PO-only, and AO + PO conditions use robust bounds.

CONCLUSIONS

- Joint enforcement of AO and PO constraints increased successful delivery from 11% without filtering to 94%–95% under nominal and robust bounds in the hard obstacle setting.
- Adding AA constraints reduced delivery to 77%–80%, with 11%–16% of episodes terminating because of QP infeasibility.
- The hybrid bounds showed a 10.7% lower agent–obstacle constraint-binding rate and a 54% smaller normalized infeasibility deficit in QP-failure cases, suggesting potential to alleviate constraint conflicts.
- The effect of calibrated learned residual bounds on delivery performance and QP feasibility remains under evaluation.

FUTURE WORK

With all collision-avoidance constraints active, QP infeasibility caused most of the failed deliveries in the hard setting under both nominal and robust bounds. Future work will incorporate the safety filter into MARL training to reduce visits to states in which the constrained control problem becomes infeasible.

ACKNOWLEDGMENTS

This work was supported by the Lockheed Martin Endowed Professorship and the Cal Poly Graduate Education Office. Simulations ran on the National Research Platform (NRP). This work was supported in part by National Science Foundation (NSF) awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, OAC-2112167, CNS-2100237, CNS-2120019.

Contact: tgrundle@calpoly.edu