Progress Report: Reinforcement Learning for 1v1 and 2v2 Aerial Dogfight

Ahmet YILDIZ

1 Introduction and Scope of the Report

This document presents a progress report on an ongoing research effort focused on reinforcement learning for autonomous aerial dogfight scenarios. The primary objective of this work is to develop learning-based agents capable of effective air combat decision-making under continuous control, partial observability, and adversarial conditions.
At this stage, the study covers two scenarios:

- A 1v1 dogfight setting, which serves as the foundational learning environment.
- A 2v2 dogfight setting, representing an extension toward multi-agent interactions.

The purpose of this report is not to provide a final evaluation, but rather to summarize the methodologies explored so far, the design choices made, and the initial experimental outcomes obtained in both scenarios.

2 1v1 Dogfight Scenario

2.1 Environment and Problem Setup

In the 1v1 dogfight scenario, two aircraft with symmetric dynamics engage in an adversarial encounter. The objective is to defeat the opponent by reducing its health points to zero within a fixed time horizon. From a reinforcement learning perspective, this problem involves continuous control actions, delayed rewards, and partial observability.

The agent observes a fixed-dimensional state vector composed of ego-centric kinematic information and relative features describing the opponent. All opponent-related variables are expressed in a relative reference frame to encourage generalization and reduce dependency on absolute positioning.

The action space is continuous and includes maneuvering and engagement commands, namely bank rate, throttle control, and a probabilistic firing trigger. Modeling the firing decision as a probability rather than a binary action allows for smoother exploration and more stable learning behavior.

2.2 Reward Design and Training Strategy

The reward function consists of a sparse terminal component combined with several shaping terms. A large positive reward is assigned upon achieving a kill, while a significant penalty is applied when the agent is defeated. However, relying solely on sparse rewards proved insufficient for efficient learning.

To address this issue, reward shaping terms were introduced to guide the agent toward tactically meaningful behavior. These include continuous tracking rewards based on target alignment, firing rewards conditioned on shot quality, and penalties reflecting ammunition expenditure. Additionally, firing-related rewards are gated by engagement conditions such as weapon engagement zone occupancy, ensuring that firing is encouraged only in realistic combat situations.

Training was conducted using a recurrent variant of Proximal Policy Optimization with an LSTM-based policy architecture. The recurrent structure enables the agent to integrate information over time and compensate for partial observability. To mitigate non-stationarity, training was staged using a heuristic opponent followed by a frozen expert policy.

2.3 1v1 Experimental Results

The trained agent demonstrates consistent improvement over the heuristic baseline and achieves strong performance against the frozen expert opponent. Evaluation was performed over multiple episodes using fixed random seeds to ensure reproducibility.

Performance metrics include win-rate, average time-to-kill, shot accuracy, and ammunition efficiency. Against the frozen expert, the agent achieves a win-rate of approximately 87%, indicating that the learned policy generalizes beyond the training opponent. These results suggest that the reward shaping and training strategy effectively capture key aspects of air combat behavior.

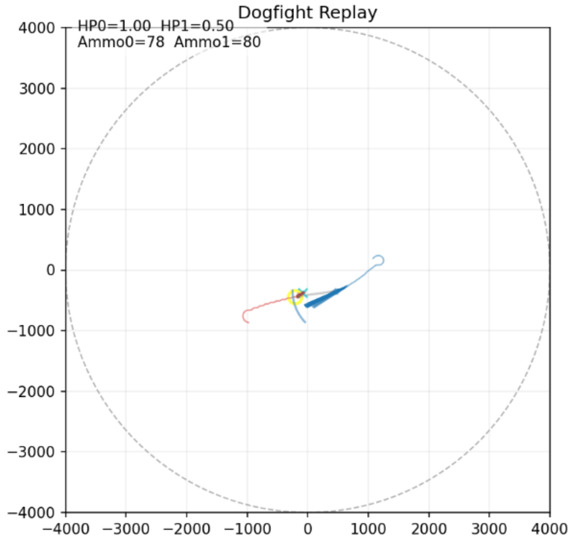

Figure 1: Sample 1v1 Replay - Result: Blue (our agent) wins)

3 Transition to 2v2 Dogfight

3.1 Motivation and Challenges

While the 1v1 scenario provides a controlled environment for learning fundamental combat behaviors, real-world air combat often involves multiple agents. Extending the framework to a 2v2 setting introduces additional challenges, including coordination, teammate awareness, and increased non-stationarity.

In the 2v2 scenario, each agent must reason not only about adversaries but also about allied agents. This significantly expands the state space and complicates credit assignment, making learning more challenging.

3.2 2v2 Environment Setup

The 2v2 environment builds upon the 1v1 framework by introducing two agents per team. Each agent receives observations containing ego-state information as well as relative features for both opponents and the teammate. Action spaces remain continuous and identical to the 1v1 case to maintain architectural consistency.

The initial training setup follows an independent learning paradigm, where each agent optimizes its own policy while treating other agents as part of the environment. This approach allows rapid prototyping and reuse of the 1v1 training infrastructure.

3.3 Initial Training Observations

Initial training runs in the 2v2 environment indicate that learning is feasible but significantly slower compared to the 1v1 case. Early-stage behaviors show limited coordination, with agents primarily focusing on individual opponents rather than team-level strategies.

Despite these limitations, preliminary results demonstrate gradual improvement in survival time and engagement quality. These early observations suggest that the learned representations from the 1v1 scenario provide a useful initialization for multi-agent training.

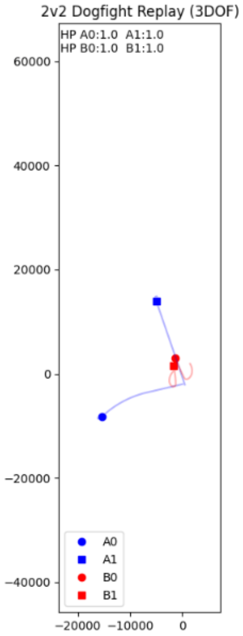

Figure 2: Sample 2v2 Replay - Result: Stalemate (No kill)

4 Discussion and Current Progress

The results obtained so far indicate that the proposed framework successfully scales from a 1v1 to a 2v2 setting at an architectural level. The 1v1 experiments validate the effectiveness of recurrent policies and structured reward shaping, while the 2v2 experiments highlight the emerging challenges of coordination and non-stationarity.

At this stage, the primary contribution lies in demonstrating a stable training pipeline and identifying key difficulties associated with multi-agent extensions. Further work is required to incorporate explicit coordination mechanisms and more advanced opponent modeling techniques.

5 Conclusion and Next Steps

This progress report summarizes the current state of a reinforcement learning framework for aerial dogfight scenarios. The 1v1 setting has been successfully addressed with robust performance against strong opponents, while the 2v2 setting has been established and initial training results obtained.

Future work will focus on improving coordination in the 2v2 scenario, exploring self-play or league-based training methods, and conducting more extensive evaluations. These efforts aim to ultimately scale the framework toward more complex multi-agent air combat environments.