

Multi-agent DRL Based Multi-UAV Trajectory Optimization for Energy Efficiency Improvement

Joon-Soo Eom and Oh-Soon Shin

School of Electronic Engineering, Soongsil University, Seoul, Korea

(email: jseom@soongsil.ac.kr; osshin@ssu.ac.kr)

Abstract—This paper addresses a UAV trajectory optimization problem that considers energy efficiency in an environment where multiple UAVs ($N = 2, 4, 6, 8$) move in three-dimensional space to provide uplink service to multiple ground users. User association and bandwidth allocation are determined by heuristic rules based on channel state and remaining data demand, while deep reinforcement learning (DRL) controls the three-dimensional acceleration of each UAV. We compare five DRL algorithms spanning centralized and multi-agent paradigms: DDPG and TD3 as centralized single-agent baselines, MADDPG and MATD3 under the centralized training with decentralized execution (CTDE) framework, and MAPPO as an on-policy multi-agent approach. Experiments across four UAV scales reveal that energy efficiency peaks at $N = 4$ under a fixed-spectrum assumption, and that MADDPG degrades at $N = 8$ due to quadratic critic input growth. MAPPO achieves competitive performance at small N but requires significantly more training episodes to converge at large N , a limitation confirmed by extended 2000-episode experiments.

Keywords—UAV, trajectory optimization, deep reinforcement learning, energy efficiency.

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) can provide flexible coverage in 6G and disaster communication environments owing to their high mobility and rapid deployment capability. However, because their onboard energy is limited, operation must account not only for communication performance but also for the propulsion energy consumed during movement [1]. Accordingly, this paper addresses a trajectory optimization problem that considers energy efficiency in an environment where multiple UAVs provide uplink service to ground users in three-dimensional space.

The resulting continuous control problem is modeled as a Markov Decision Process (MDP). To analyze scalability under varying fleet sizes, we evaluate five algorithms: DDPG and TD3 as centralized baselines, MADDPG [3] and MATD3 under the CTDE framework, and MAPPO [4] as an on-policy cooperative approach. A key contribution is a scalability comparison across $N = 2, 4, 6$, and 8 UAVs and an analysis of MAPPO's sample-efficiency limitation.

II. SYSTEM MODEL AND OPTIMIZATION PROBLEM

Let M denote the number of UAVs and $T = \{1, 2, \dots, S\}$ the set of time slots, with $S = 100$. The position and velocity of UAV $m \in \{1, \dots, M\}$ are denoted $q_m(t)$ and $v_m(t)$, with control input defined as three-dimensional acceleration $a_m(t)$. The UAV kinematics are modeled as $q_m(t+1) = q_m(t) + v_m(t)\Delta t + \frac{1}{2}a_m(t)\Delta t^2$ and $v_m(t+1) = v_m(t) + a_m(t)\Delta t$.

Each UAV must satisfy altitude, velocity, and horizontal flight-area constraints, and to prevent collisions the distance between any two UAVs must remain at least a minimum safety distance d_{min} . The number of ground users is $K = 4M$, and they are distributed uniformly in a 3200×3200 m area. The spectral efficiency between user k and UAV m is computed as $\eta_{k,m}(t) = \log_2(1 + SNR_{k,m}(t))$. Each user is associated with the UAV that maximizes the priority metric $\psi_{k,m}(t) = \eta_{k,m}(t)(0.3 + \rho_k(t))$, and associated users are allocated UAV bandwidth in proportion to their $\psi_{k,m}(t)$ value. Here, $\rho_k(t)$ is the ratio of the remaining information of user k , and the constant offset 0.3 ensures that even a user with a small remaining-data ratio retains a certain level of priority when its channel condition is good. The total system bandwidth $B_{tot} = 4$ MHz is fixed regardless of M . The per-user nominal bandwidth is defined as $B_{nom} = \frac{B_{tot}}{4M^2}$, which reflects the increasing co-channel interference as fleet size grows.

Let the total information delivered during one episode be $\sum_{k=1}^K d_k(s) = B_{del}$ and the total propulsion energy be $\sum_{m=1}^M E_m^{mov} = E_{prop}$; then the energy efficiency per episode is defined as $EE = \frac{B_{del}}{E_{prop}}$. In this work, the step EE, computed

as the ratio of delivered information to propulsion energy at each step, is used as the key metric, and a simple reward function that additionally suppresses only collisions and boundary violations is designed. Accordingly, the optimization problem is formulated as follows.

$$\max_{q_m, v_m, a_m} \sum_{s=1}^S \frac{B_{del}}{E_{prop}} - \lambda_{col} \sum_{s=1}^S P_{col}(s) \quad (1a)$$

$$s.t. \quad 0 \leq \beta_{k,m}(s) \leq \alpha_{k,m}(s), \quad \forall k, m, s \quad (1b)$$

$$v_{min} \leq \|v_m(s)\| \leq v_{max}, \quad \forall m, s, \quad (1c)$$

$$h_{min} \leq z_m(s) \leq h_{max}, \quad \forall m, s, \quad (1d)$$

$$\|q_m(s) - q_n(s)\| \geq d_{min}, \quad \forall m \neq n, \quad \forall s, \quad (1e)$$

$$r_k(s) \geq R_{min}, \quad \forall k, \quad \forall s. \quad (1f)$$

where λ_{col} is a weight that adjusts the trade-off between energy-efficiency performance and safety. Constraint (1b) ensures that resource allocation is consistent with the user association state, while (1c) and (1d) limit the UAV velocity and altitude, respectively. Furthermore, (1e) keeps the distance between any two UAVs above the minimum safety distance to prevent collisions. Finally, (1f) guarantees that the achievable uplink rate $r_k(s)$ of every user remains above the minimum required rate R_{min} .

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00208995, RS-2025-02214082).

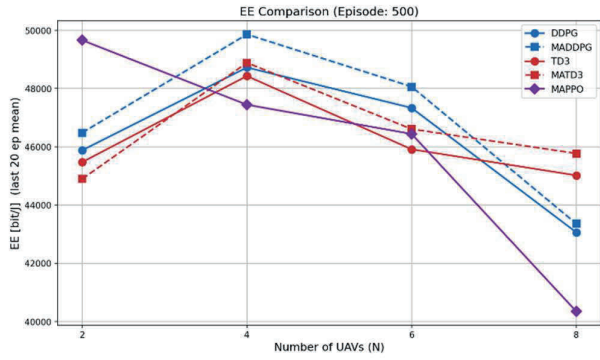


Fig 1. Convergence EE vs number of UAVs(N)

III. DRL ALGORITHMS

We compare five DRL algorithms under a unified training environment.

A. Centralized Approaches (DDPG, TD3)

DDPG and TD3 treat all M UAVs as a single agent with joint state $s \in \mathbb{R}^d$ and joint action $a \in \mathbb{R}^M$. A single actor-critic pair is trained using the full global state. TD3 additionally employs twin critics and delayed policy updates to reduce overestimation bias [5].

B. CTDE Approaches (MADDPG, MATD3)

Under the CTDE framework, each UAV m maintains a local actor $\pi_m(a_m | o_m)$ that observes only its local state o_m during execution, while a centralized critic $Q_m(s, a_1, \dots, a_M)$ accesses the global state and all joint actions during training [3]. MATD3 extends MADDPG with twin critics per agent.

C. On-Policy Approach (MAPPO)

MAPPO employs a shared stochastic actor (parameter sharing across agents) and a centralized value network $V(s)$. Advantages are estimated via Generalized Advantage Estimation (GAE) with $\lambda = 0.95$, and the policy is updated using a PPO clipping objective ($\epsilon = 0.20$) over $K = 4$ epochs per rollout of length $S = 100$ [4]. Unlike off-policy methods, MAPPO discards experience after each update, making it more sample-inefficient but free from replay buffer constraints.

IV. SIMULATION RESULTS

Simulations are conducted in a 3200×3200 m horizontal area with altitude range 25–90 m. Each episode consists of $S = 100$ time steps with $\Delta t = 1$ s. The number of ground users is $K = 4M$. All algorithms are trained for 500 episodes with 3 independent random seeds (seeds 21, 42, 63), and results are reported as mean $\pm$ std over seeds. Key parameters: $B_{tot} = 4$ MHz, path-loss exponent $\alpha = 2.2$, noise power $\sigma^2 = -174$ dBm/Hz, and minimum safety distance $d_{min} = 18$ m.

A. Energy efficiency Comparison (500 Episodes)

As shown in Fig. 1, energy efficiency peaks at $N = 4$ for all algorithms. This arises because the per-user bandwidth $B_{nom} = \frac{B_{tot}}{4M^2}$ decreases quadratically with M under the fixed-spectrum assumption, while propulsion energy grows linearly with M . The $N = 4$ operating point achieves the best balance between communication throughput and fleet energy cost. At $N = 8$, MADDPG degrades notably. This is attributed to the

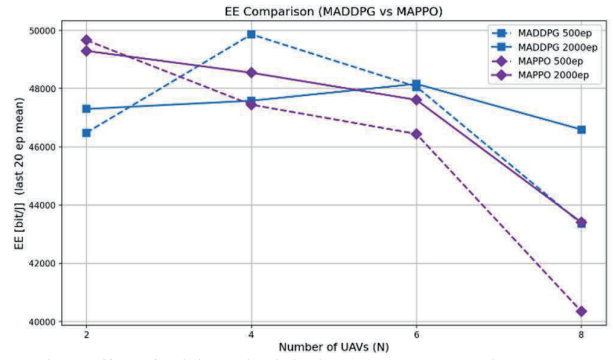


Fig 2. Effect of training episode budget on MADDPG and MAPPO

growth of the centralized critic input dimension, which impairs credit assignment. MATD3 mitigates this via twin critics. MAPPO achieves the highest EE at $N = 2$, confirming that its shared-actor architecture is well-suited to small homogeneous fleets. However, MAPPO shows the steepest degradation with increasing N .

B. MAPPO Extended Training (2000 Episodes)

The degradation of MAPPO at large N is hypothesized to stem from sample inefficiency inherent in on-policy learning rather than from a structural limitation. Unlike off-policy methods that repeatedly reuse replay-buffer samples, MAPPO discards collected trajectories after each update and therefore requires more episodes to converge. To verify this hypothesis, MADDPG and MAPPO are retrained with 2000 episodes. As shown in Fig. 2, MAPPO improves consistently with extended training. Nevertheless, even with 2000 episodes, the results indicate that on-policy methods still face a fundamental sample-efficiency disadvantage relative to off-policy alternatives when training budgets are limited.

V. CONCLUSION

This paper presented a systematic scalability comparison of five DRL algorithms for multi-UAV trajectory optimization across fleet sizes $N \in \{2, 4, 6, 8\}$. Under a fixed-bandwidth assumption, energy efficiency peaks at $N = 4$. Among the compared methods, MATD3 demonstrated the most robust scaling across all N values, while MADDPG suffers from critic dimensionality growth at large N . MAPPO achieves the highest EE at $N = 2$ but degrades most steeply with N ; extended training partially mitigates this limitation. Future work will explore adaptive bandwidth allocation and attention-based critic architectures to further improve large-fleet scalability.

REFERENCES

- [1] J. Mu, R. Zhang, Y. Cui, N. Gao, and X. Jing, "UAV meets integrated sensing and communication: Challenges and future directions," *IEEE Commun. Mag.*, vol. 61, no. 5, pp. 62–67, May 2023.
- [2] D. Silver et al., "Deterministic policy gradient algorithms," in *Proc. ICML*, vol. 32, pp. 387–395, 2014.
- [3] R. Lowe et al., "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. NeurIPS*, 2017.
- [4] C. Yu et al., "The surprising effectiveness of PPO in cooperative multi-agent games," in *Proc. NeurIPS*, 2022.
- [5] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. ICML*, 2018.