Prescriptive Dirichlet Task Allocation Policy with Deep Reinforcement Learning

Ali Fuat Sahin, Mustafa Emre Gürsoy, Matthias Schuller

EPFL

Abstract

We propose a Dirichlet policy for continuous task allocation in UAV fleets. This method is bias-free and shows faster convergence and better performance compared to traditional Gaussian-softmax policies. A UAV simulation for search and rescue missions demonstrates the efficacy of our approach.

Introduction

Background

Efficient task allocation is critical in multi-agent systems. Traditional methods using Gaussian policies with softmax functions often lead to suboptimal performance due to bias and variance issues.

Objective

Improve UAV task allocation in search and rescue missions by integrating a Soft Actor-Critic (SAC) algorithm with a Dirichlet policy to enhance reliability and performance.

Methodology

Previous work [5] introduced the Dirichlet policy to address continuous action-space allocation tasks. It avoids bias and reduces variance, outperforming the Gaussian-softmax policy. This policy is integrated with the SAC framework for improved task allocation.

Implications of the Gaussian Policy

Classic reinforcement learning setting for continuous action space:

- Policy π parameterized by a **conditioned Gaussian distribution**
- **Action sampled** from the policy: $\mathbf{a} \sim \pi(\cdot|\mathbf{a})$
- Constraint $\sum_{i=0}^{N} \mathbf{a}_i = 1$ not satisfied
  $\longrightarrow$ **not directly applicable** to allocation tasks
- **Solution**: pass $\mathbf{a}$ through a softmax function to obtain a point on a simplex
- **Problem**: Side effects
  - Biased estimation
  - Larger variance

  $\longrightarrow$ Negative impact on learning [5].

Target Entropy in SAC

In the SAC algorithm, the objective is to solve the following constrained optimization problem [4], [3].

$$\max_{\pi_{0:T}} \mathbb{E}_{\rho_\pi}[\sum_{t=0}^{T} r(s_t, a_t)], \quad s.t. \quad \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}[-\log(\pi(a_t|s_t)] \geq H$$

In general, with Gaussian policy target entropy, denoted as H, is defined as $\log(|A|)$. However, Dirichlet distribution entropy is bounded, unlike Gaussian distribution. Therefore, in our approach, we derived the target entropy for the Dirichlet policy based on the maximum achievable entropy. The final equation is:

$$H_{\text{target}} = -C \cdot \log \Gamma(|A|)$$

where C is an environment-specific constant.

Case Studies

The Dirichlet policy was evaluated in a UAV fleet simulation for search and rescue missions. Results indicate superior performance in extending search duration and reliability over traditional task allocation strategies.

Simplex Regression Task

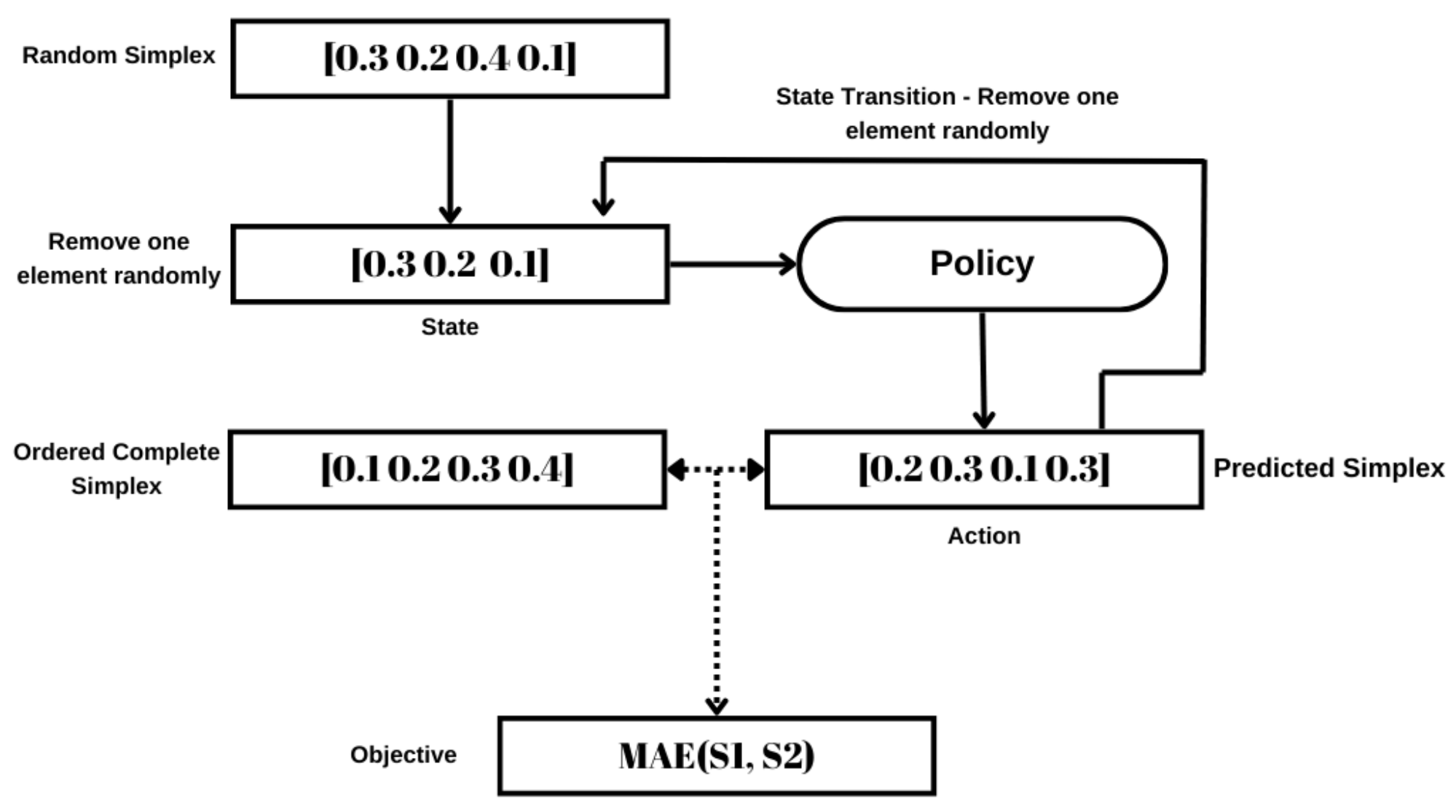

Figure 1. Simplex regression task RL problem.

- **Performance Comparison**: We compare the performance of Dirichlet and Gaussian-softmax policies in task allocation.
- **Experiment**: Given a random simplex and removing an element, we expect the policy to return an ordered complete simplex.

Simulation Environment

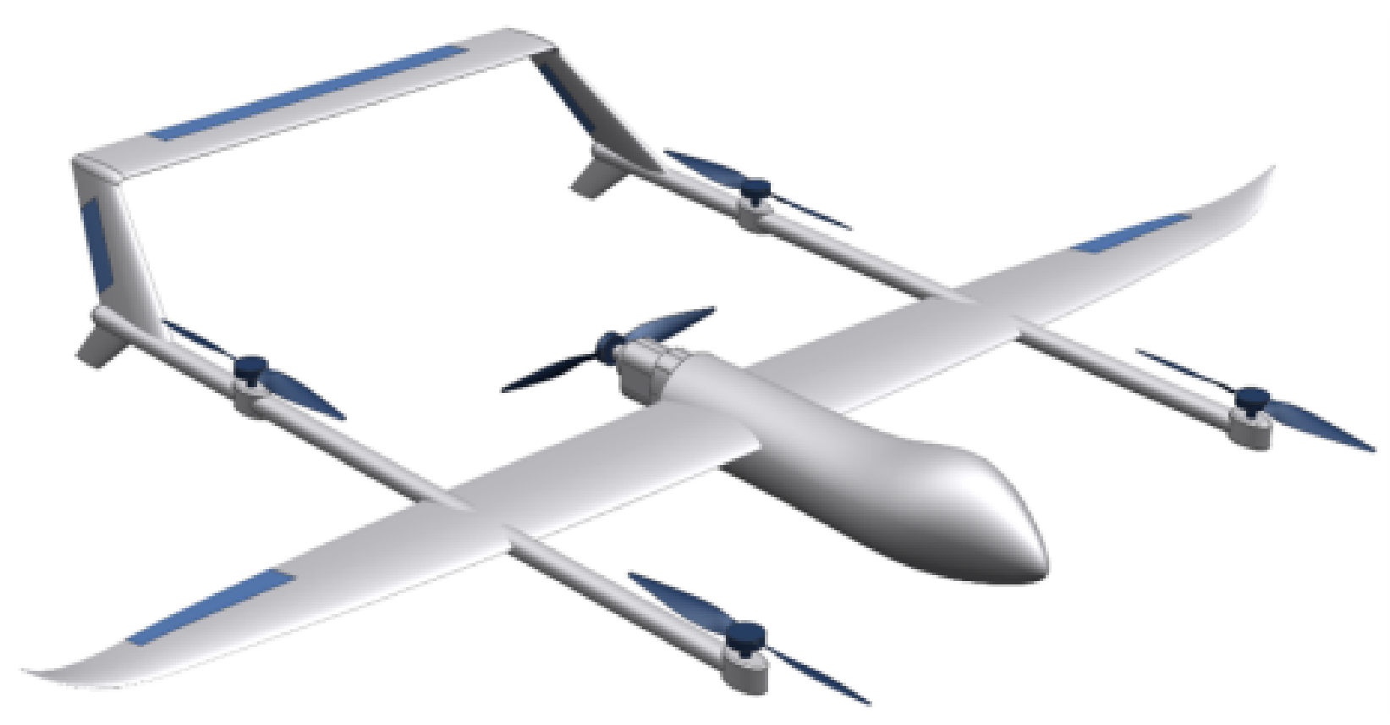

Figure 2. Hybrid VTOL UAV—Skyhunter from the ACoRUs Project Considered In the Simulation.

Simulated UAV components:

- Hover Motors
  - 4 bearings
  - 4 coils
- Pusher Motor
  - 1 bearing
  - 1 coil

Randomized Degradation Dynamics: UAVs experience degradation based on hover and pusher bearings and coils, with randomized dynamics to simulate real-world variability [1].

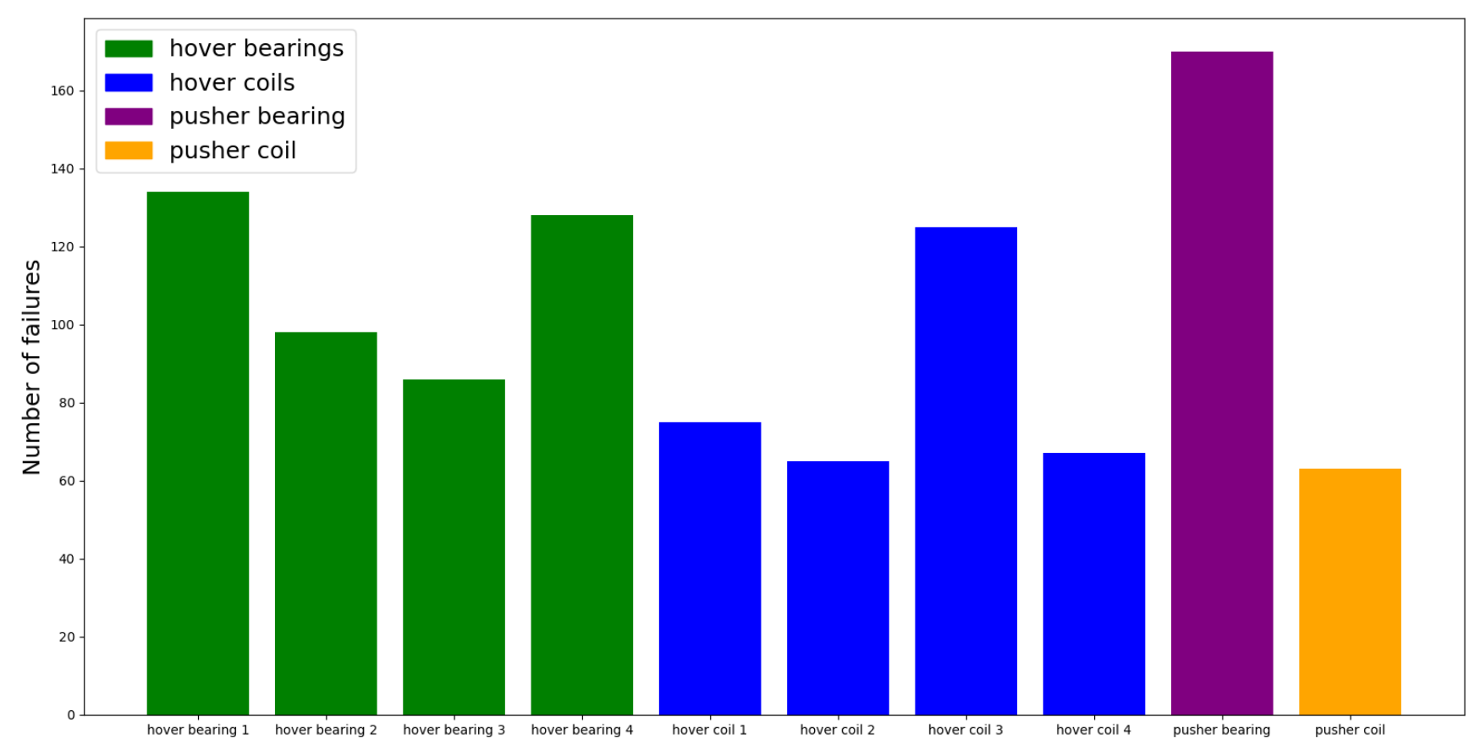

Figure 3. Failure statistics over 1000 runs.

Failure statistics are adjusted based on real-world data. Stator-based faults account for 35-40% of total failures while bearing faults are responsible for 60-65% of the faults if rotor faults are excluded [2].

Randomized Mission Profiles: Search and rescue mission profiles are randomized, and task allocation is performed using the policy.

Results

Simplex Regression Task

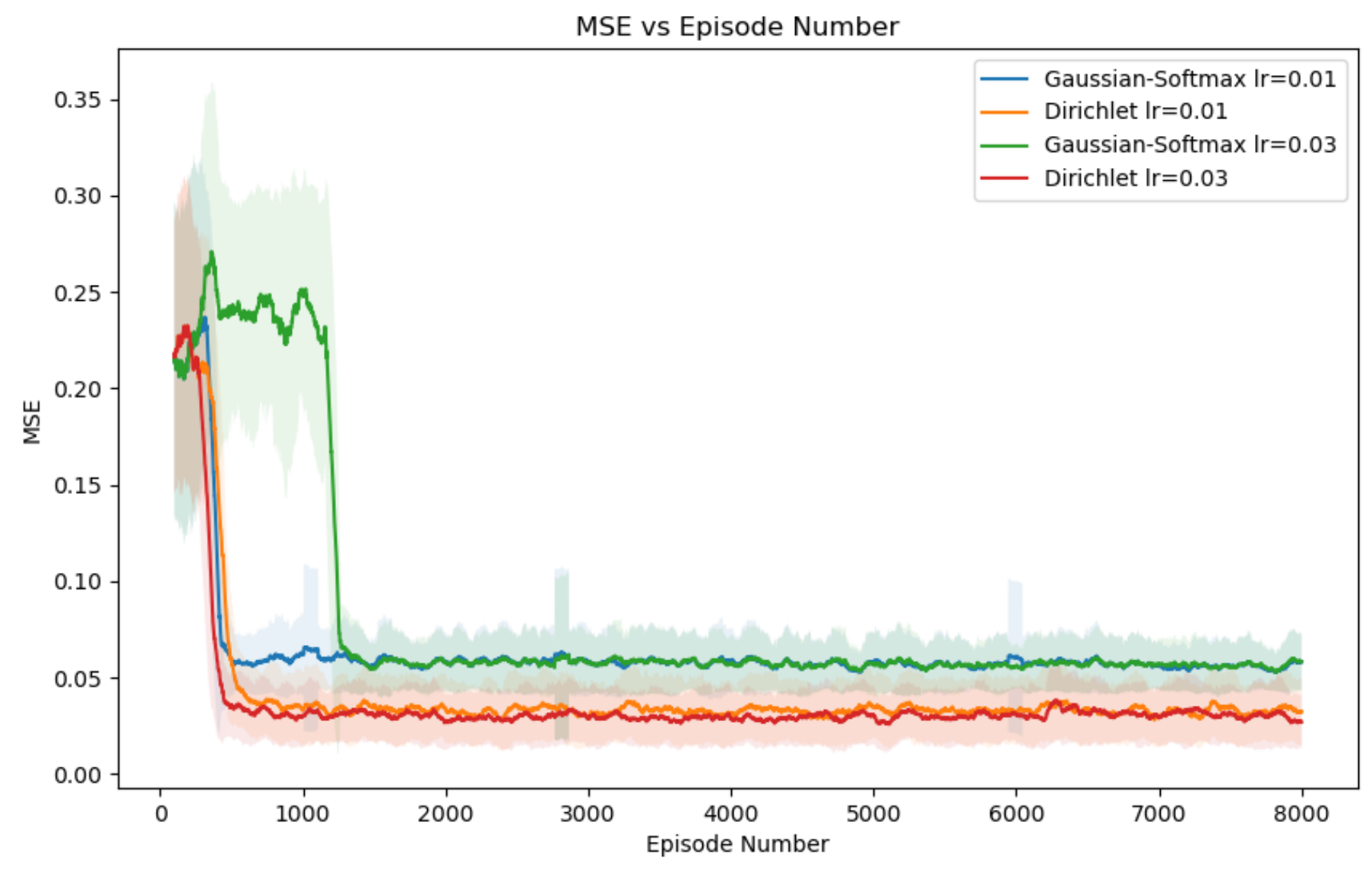

Figure 4. Dirichlet vs Gaussian-Softmax Learning Curves.

- The result shows that the Dirichlet distribution performs better and is more robust to different learning rates.
- For the given learning rates Dirichlet policy performs two times better than the Gaussian policy for the simplex regression task.

UAV Fleet Task Allocation

In our UAV simulation, search distances are allocated based on the health status of UAVs, particularly their bearings and coils. The Dirichlet policy dynamically assigns search distances, adapting to the health status of each UAV.

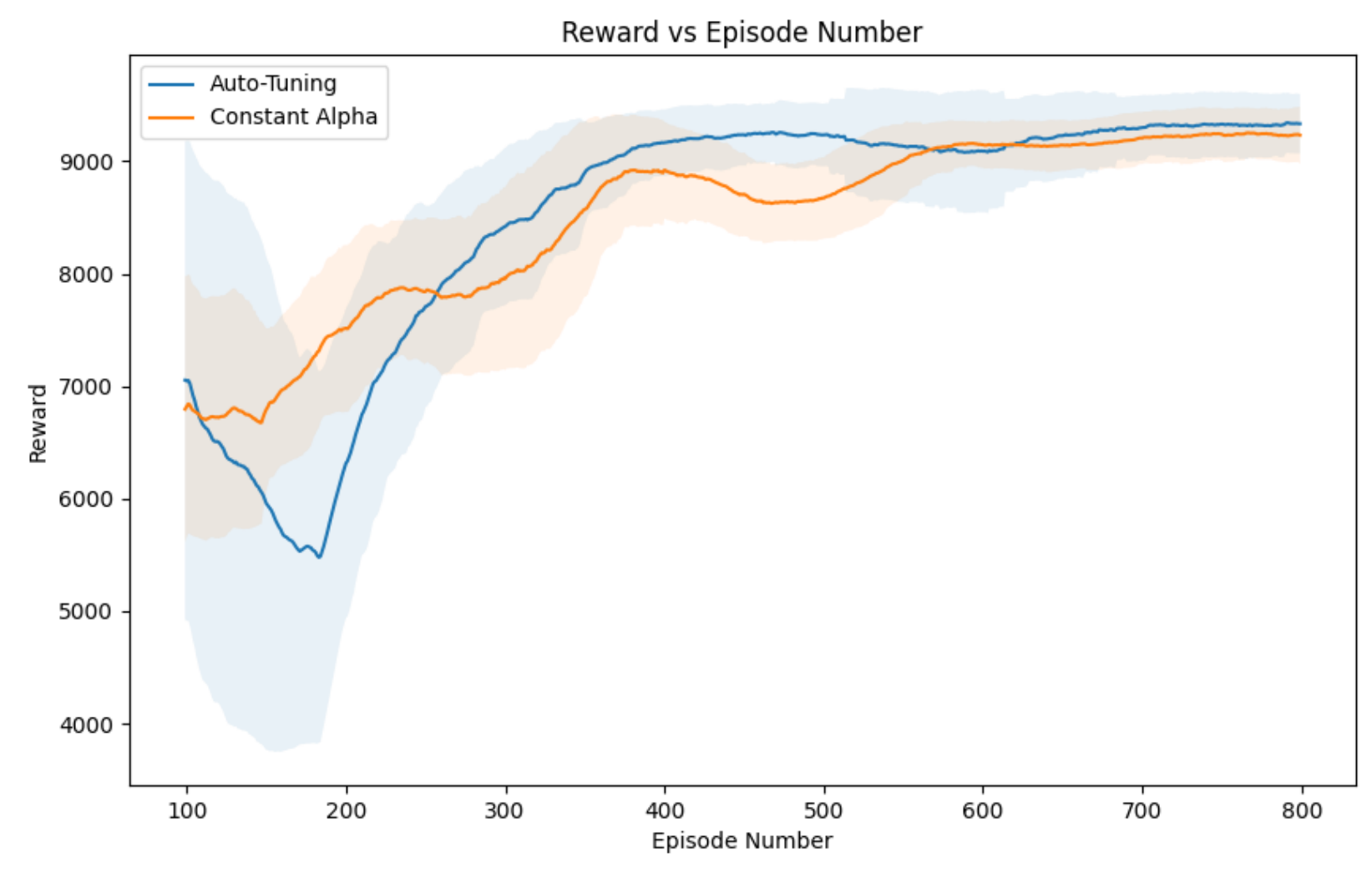

Figure 5. Score vs. Episode graph for 4 UAV simulation.

Figure 5 illustrates the distinction between our target entropy function and the use of a constant $\alpha = 0.6$ without automatic entropy tuning. In our approach, we have incorporated the α value that yielded the best performance. However, our target entropy function is more resilient to variations in UAV numbers and environmental changes.

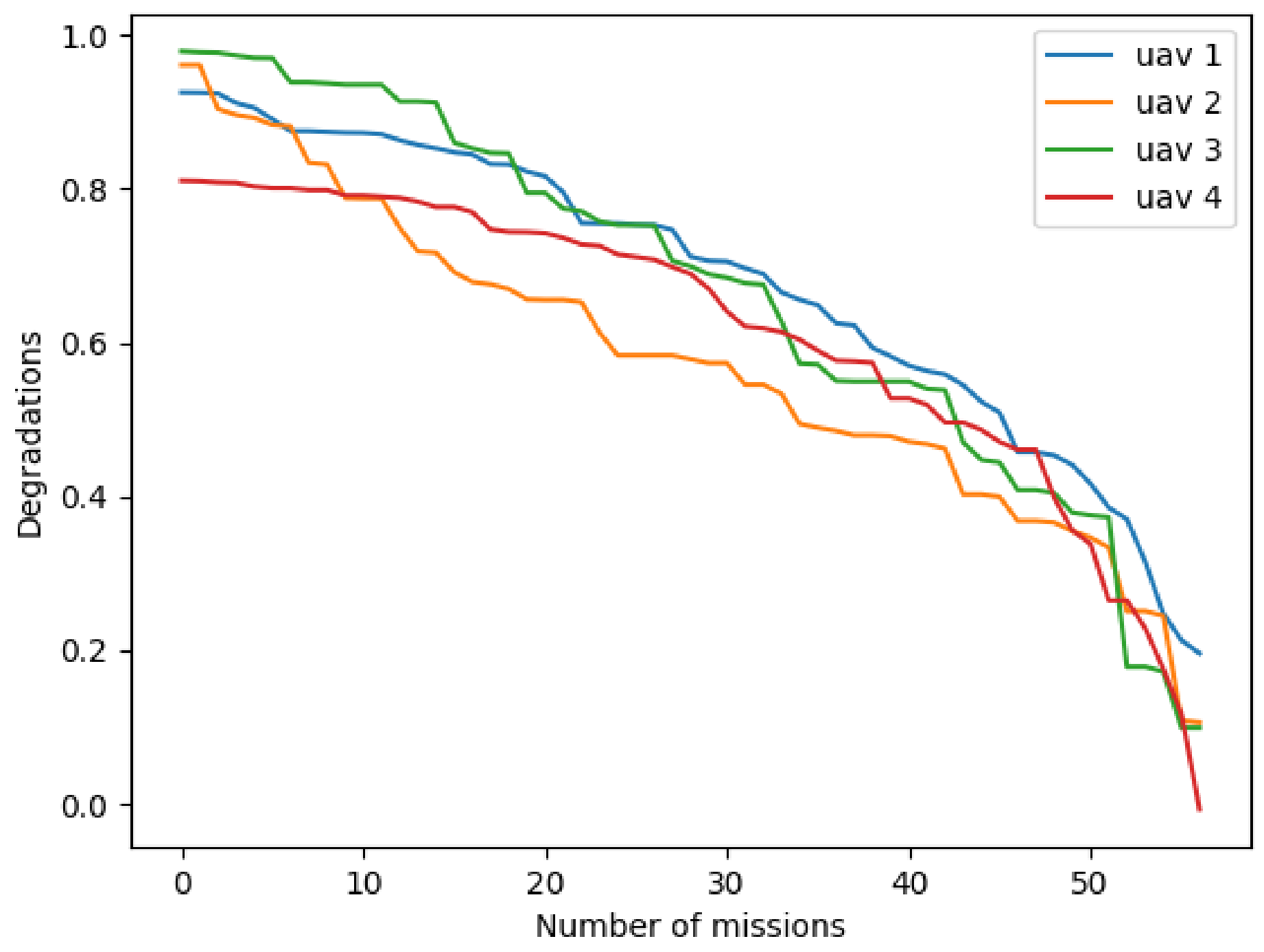

Figure 6. Lowest component degradation values for each UAV.

Figure 6, demonstrates that the policy can allocate tasks based on UAV health states and none of the UAVs are exploited.

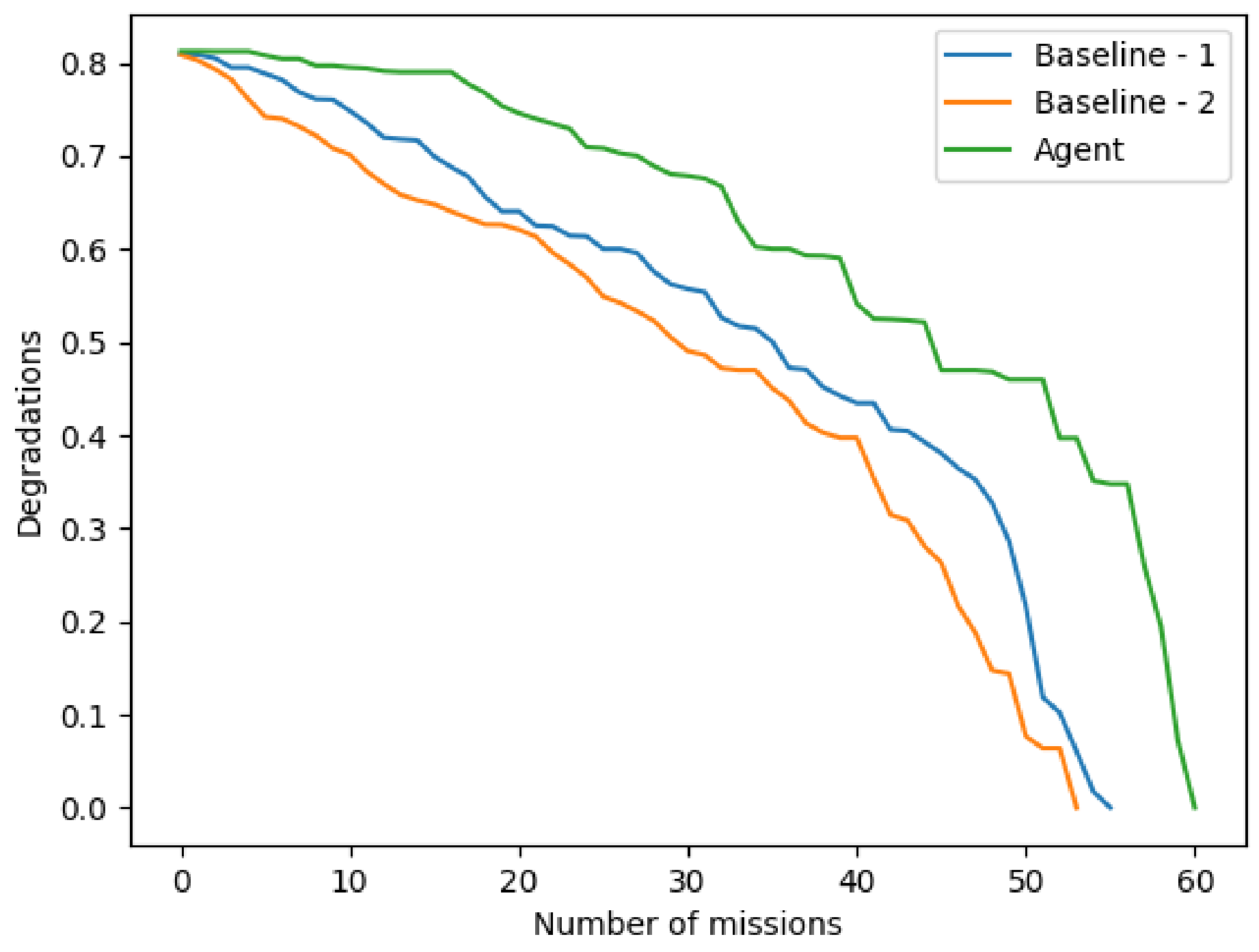

Figure 7. Comparison of performance with baseline strategies. Lowest component health among all UAVs.

- Baseline - 1: Task allocated proportionally to the minimum health state value of the UAV (56 Missions completed).
- Baseline - 2: Task allocated equally to each UAV (54 Missions completed).
- Agent: RL strategy (61 Missions completed).

Conclusion

Our Dirichlet policy framework for continuous allocation tasks mitigates bias and reduces variance, offering significant improvements in efficiency and reliability. This approach applies to various continuous control reinforcement learning algorithms.

In general, our study presents,

- A prescriptive, fully autonomous, scalable, and transferable framework.
- Developed framework improves sustainability and efficiency of multi-agent systems.

Future work includes,

- Applying and deploying the proposed framework to more challenging and extensive real-world problems.
- Evaluating limitations of the proposed framework.
- Extending the algorithm for time-varying action space depending on the UAV availabilities.

References

[1] Lorenz Dingeldein. "Simulation framework for real-time PHM applications in a system-of-systems environment". In: *Aerospace* 10.1 (2023), p. 58.

[2] Tomas Garcia-Calva et al. "Early detection of faults in induction motors—A review". In: *Energies* 15.21 (2022), p. 7855.

[3] Tuomas Haarnoja et al. "Soft actor-critic algorithms and applications". In: *arXiv preprint arXiv:1812.05905* (2018).

[4] Tuomas Haarnoja et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor". In: *International conference on machine learning*. PMLR. 2018, pp. 1861–1870.

[5] Yuan Tian et al. "A prescriptive Dirichlet power allocation policy with deep reinforcement learning". In: *Reliability Engineering & System Safety* 224 (2022), p. 108529.