Adaptive Entropy Regularization for Stable Multi-Agent PPO in PettingZoo MPE

Abstract

This paper investigates stability and exploration in Multi-Agent Reinforcement Learning (MARL) using the Multi-Agent Proximal Policy Optimization (MAPPO) framework on the PettingZoo Multi-Agent Particle Environment (MPE), simple_spread_v3. Although MAPPO achieves strong performance with centralized training and decentralized execution, its behavior is sensitive to the entropy coefficient that regulates policy exploration. Fixed entropy weights often cause either premature convergence (entropy collapse) or excessive stochasticity. To address this, we introduce Adaptive Entropy Regularization, where the entropy coefficient β is automatically tuned through a target-entropy mechanism during training. We evaluate four variants baseline MAPPO, entropy-scheduled MAPPO, auto-tuned entropy MAPPO, and recurrent MAPPO with auto-tuning across twelve training runs. Results demonstrate that the adaptive variant achieves the highest average step reward (≈ −2.3 vs. −2.8 and −5.0), maintains consistently high policy entropy (≈ 1.5–1.6), and stabilizes value loss. These findings provide strong evidence that adaptive entropy control enhances both exploration and training stability, offering a simple and effective improvement for cooperative MARL.

Introduction

Multi-Agent Reinforcement Learning (MARL) enables multiple autonomous agents to learn coordinated behaviors through interaction within shared environments. It underpins a range of applications such as robotic swarms, traffic management, and cooperative control. However, MARL remains challenging due to non-stationarity, partial observability, and the exploration–exploitation trade-off across agents. As each agent's policy evolves, the environment dynamics perceived by others change, leading to unstable training and poor convergence if exploration is not adequately regulated.

The Multi-Agent Proximal Policy Optimization (MAPPO) algorithm has emerged as a strong baseline for cooperative MARL. It extends PPO by incorporating a centralized critic that conditions on the global state and decentralized actors that operate on local observations, thereby realizing the Centralized Training and Decentralized Execution (CTDE) paradigm. Despite its effectiveness, MAPPO's performance is highly sensitive to the entropy coefficient (β) used to encourage stochastic exploration. A fixed β can cause entropy collapse, reducing exploration too early, or sustain excessive randomness late in training, hindering convergence.

To overcome these limitations, this study investigates entropy-regularization strategies for MAPPO on the PettingZoo MPE benchmark (simple_spread_v3). We compare three extensions: (i) Entropy-Scheduled MAPPO, where β decays linearly during training; (ii) Auto-Tuned Entropy MAPPO, which adaptively adjusts β through a target-entropy mechanism; and (iii) Recurrent MAPPO + Auto-Tuning, which incorporates temporal memory via GRU units. By systematically evaluating these variants, we aim to determine whether adaptive entropy control can enhance training stability, cooperative behavior, and long-term performance in multi-agent settings.

Methodology

This section describes the experimental framework, algorithmic modifications, and training configuration adopted to evaluate entropy regularization strategies within the Multi-Agent Proximal Policy Optimization (MAPPO) paradigm. We first outline the baseline MAPPO formulation that serves as the foundation for all experiments. Next, we introduce two enhanced variants — one employing a linear entropy schedule and another featuring an adaptive target-entropy mechanism that dynamically tunes the entropy coefficient during training. We further extend the adaptive approach by incorporating recurrent neural architectures to capture temporal dependencies in partially observable environments. Finally, we detail the implementation setup, including environment configuration, hyperparameters, and optimization settings used in the PettingZoo MPE benchmark. Together, these components define a systematic methodology to assess how different entropy-control strategies influence exploration stability, convergence behavior, and cooperative performance in multi-agent reinforcement learning.

Baseline: Multi-Agent PPO (MAPPO)

We adopt Multi-Agent Proximal Policy Optimization (MAPPO) as the baseline algorithm. MAPPO extends the standard PPO algorithm to the multi-agent setting using the centralized training with decentralized execution (CTDE) paradigm. Each agent maintains a decentralized actor policy $\pi_\theta(a_t \mid o_t)$, conditioned only on its own local observation o_t. A shared centralized critic $V\phi(s_t)$, where s_t denotes the global state or concatenation of agent observations, is used during training to stabilize value estimation.

This actor–critic design allows for coordinated learning across agents while ensuring that, at execution time, policies remain decentralized and scalable. The PPO clipped surrogate objective serves as the foundation for actor updates, limiting policy changes to improve stability:

$$L_\pi(\theta) = -E_t\big[\min(r_t(\theta)A_t,\ clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t) + \beta H\big(\pi_\theta(\cdot \mid o_t)\big)\big]$$

where $rt(\theta) = \frac{\pi_\theta(a_t|o_t)}{\pi_{\theta_{old}}(a_t|o_t)}$ is the importance sampling ratio, A_t is the advantage function, ϵ is the clipping threshold, and β is the entropy regularization coefficient. The entropy term H encourages exploration by preventing premature policy determinism.

Enhancement 1: Entropy Scheduling

A limitation of standard MAPPO is the use of a fixed entropy coefficient β. In practice, this leads to entropy collapse early in training, as the policy converges too quickly to narrow distributions. To address this, we introduce entropy scheduling, in which the coefficient βt is decayed as training progresses. The modified objective is:

$$L_\pi(\theta) = -E_t\big[\min(r_t(\theta)A_t,\ clip(r_t(\theta), 1 - \epsilon, 1 + \epsilon)At) + \beta H\big(\pi\theta(\cdot \mid ot)\big)\big]$$

where βt is a time-dependent entropy coefficient.

We experiment with two scheduling strategies:

- **Linear decay:**

$$\beta_t = \beta_{start} - \frac{t}{T}(\beta_{start} - \beta_{end})$$

where β_{start} is the initial coefficient, β_{end} is the final coefficient, and T is the total number of updates.

- **Cosine decay:**

$$\beta_t = \beta_{end} + \frac{1}{2}(\beta_{start} - \beta_{end})(1 + \cos(\pi t/T))$$

In both cases, exploration is emphasized early in training and gradually reduced to encourage convergence.

Enhancement 2: Adaptive Target-Entropy Regularization

The Auto-Tuned Entropy MAPPO variant replaces manual scheduling with an adaptive mechanism that learns β_t online to maintain a desired entropy level H_{target}. The update rule for the log-entropy coefficient is:

$$log\beta_{t+1} = log\beta_t + \alpha_\beta\big(H_t - H_{t\,\arg et}\big)$$

where α_β is the learning rate of the entropy coefficient. When the policy's actual entropy H_t falls below H_{target}, β_t increases, encouraging more exploration; conversely, it decreases when entropy is too high.

This adaptive control provides self-regulating exploration, removing the need for manual tuning and allowing the policy to respond dynamically to learning progress.

Enhancement 3: Recurrent MAPPO + Auto-Tuning

The Recurrent MAPPO (R-MAPPO) variant extends MAPPO with a Gated Recurrent Unit (GRU) encoder in each actor network to model temporal dependencies across timesteps. This allows agents to leverage short-

term memory, improving decision-making under partial observability.

Combining GRU with the adaptive entropy mechanism yields R-MAPPO + Auto-Tuning, which simultaneously benefits from stable exploration and temporal credit assignment.

Implementation Details

All experiments are conducted on the PettingZoo MPE environment simple_spread_v3 with 3 cooperative agents. Agents observe local positions, velocities, and relative distances, with discrete actions. Training uses PyTorch with the following configurations:

- **Framework:** PyTorch
- **Algorithm:** MAPPO with PPO-style updates
- **Actors:** Decentralized policies, parameter sharing across agents
- **Critic:** Centralized value function, trained with global state
- **Entropy schedules:**
  - Linear: $=\beta_0$ =0.03, β_{end} =0.008, decay over 400 updates
  - Adaptive tuning: $H_{target}=1.1, \alpha_\beta=10-3, \beta_{min}=10-4, \beta_{max}=0.3$
  - Cosine: same start/end values with cosine profile
- **Hyperparameters:** horizon length = 100, PPO epochs = 6, clip parameter $\epsilon=0.2$, learning rate = 3×10^{-4}, Value loss coefficient: 0.5, batch size adapted to horizon length.

Across all runs, Generalized Advantage Estimation (GAE-λ), advantage normalization, value clipping, and KL early-stopping are applied to ensure stable optimization. Each variant is evaluated over three random seeds to estimate performance robustness.

Preliminary Results

Experiments were conducted on the **PettingZoo MPE** environment `simple_spread_v3` using three cooperative agents trained for 500 updates per seed. Each method was run across **three random seeds**, and the reported metrics represent the mean ± standard deviation across these runs. The comparison includes: (1) **Baseline MAPPO**, (2) **Entropy-Scheduled MAPPO** (linear decay), (3) **Auto-Tuned Entropy MAPPO** (target-entropy control), and (4) **Recurrent MAPPO + Auto-Tuning** (GRU-based).

Learning Curves

Baseline MAPPO shows rapid saturation with an average reward around $Ravg\approx-5.0$, indicating premature convergence and poor cooperation. The entropy-scheduled variant significantly improves stability, achieving $Ravg\approx-2.8$. The proposed Auto-Tuned Entropy MAPPO exhibits the best performance with $Ravg\approx-2.3$, maintaining smoother convergence and lower variance across seeds. The recurrent version yields comparable performance ($Ravg\approx-2.6$) but demonstrates enhanced temporal stability in sequential decision-making tasks.

Entropy in the baseline drops from ~1.6 to < 0.5 within the first 100 updates, confirming early exploration collapse.

The scheduled approach sustains entropy between 1.0 and 1.3 for most of training. In contrast, the adaptive entropy variants maintain high, stable entropy around 1.5–1.6 throughout learning, effectively preventing policy determinism.

Adaptive Coefficient Behavior

In both cases, βt automatically increases toward its upper bound (≈ 0.3) when policy entropy begins to decline, thereby restoring exploration. This dynamic adjustment enables a self-regulated balance between exploration and exploitation, without manual hyper-parameter tuning.

Quantitative Comparison

Table 1 summarizes the final 100-update averages (mean ± std) for key metrics. Compared with baseline MAPPO, the adaptive variant achieves a ≈ 55 % relative improvement in average reward and sustains higher entropy with bounded value-loss variance. Formally,

$$R_{\text{avg}}^{\text{adaptive}} > R_{\text{avg}}^{\text{scheduled}} > R_{\text{avg}}^{\text{baseline}}, \quad H_{\text{adaptive}}(t) \approx 1.55 > H_{\text{scheduled}} \approx 1.1 > H_{\text{baseline}} \to 0.3.$$

Critic stability is confirmed by $V_{loss}\in[0.2,0.7]$ and

smooth KL divergence trajectories, showing well-conditioned updates across all seeds.

Method	Avg Step Reward (± std)	Entropy (± std)	π_loss (± std)	V_loss (± std)
Baseline MAPPO	−5.02 ± 0.68	0.37 ± 0.12	−0.021 ± 0.008	0.63 ± 0.14
Entropy-Scheduled MAPPO	−2.83 ± 0.42	1.08 ± 0.18	−0.035 ± 0.009	0.47 ± 0.10
Auto-Tuned Entropy MAPPO	**−2.31 ± 0.33**	**1.56 ± 0.09**	−0.028 ± 0.006	0.40 ± 0.08
Recurrent MAPPO + Auto-Tuning	−2.63 ± 0.39	1.54 ± 0.11	−0.031 ± 0.007	0.43 ± 0.09

Table 1. Comparison of final 100-update metrics (mean ± std over 3 seeds). Auto-Tuned Entropy MAPPO achieves the best reward and entropy balance, while Recurrent MAPPO + Auto-Tuning provides the most stable temporal performance.

Discussion

Overall, the Auto-Tuned Entropy MAPPO delivers the highest cooperative performance, combining strong average rewards, stable entropy, and minimal oscillations in policy loss. The linear-scheduled approach remains an effective baseline improvement over fixed-entropy MAPPO, while the recurrent + adaptive configuration offers superior temporal robustness for partially observable domains. These findings collectively verify that adaptive entropy regulation improves exploration efficiency and training stability in cooperative multi-agent reinforcement learning.

Technical Validation

The technical validation of this work focuses on confirming that the observed performance improvements of the proposed Auto-Tuned Entropy MAPPO and Recurrent MAPPO + Auto-Tuning are both empirically consistent and theoretically justified. The validation process emphasizes three aspects: (1) consistency across random seeds, (2) stability of optimization dynamics, and (3) alignment of empirical outcomes with theoretical expectations in entropy-regularized policy optimization.

Consistency Across Seeds

All experiments were executed using three independent random seeds to ensure statistical significance and eliminate random initialization bias. For each seed, we monitored the evolution of policy loss, critic loss, entropy, and average step reward. Across runs, the Auto-Tuned Entropy MAPPO exhibited low variance in both reward and entropy trajectories (standard deviation < 0.4 in average reward), confirming that the adaptive mechanism consistently guides the policy toward stable exploration–exploitation behavior.

In contrast, the Baseline MAPPO runs showed higher variance and frequent oscillations, particularly during the early training phase, reflecting sensitivity to initial conditions. This consistency across seeds provides strong evidence that the performance gains of the adaptive variant are not coincidental but emerge as a repeatable effect of the algorithmic design.

Optimization Stability and Bounded Loss Behavior

To assess learning stability, we examined three key indicators: critic value loss, policy loss, and approximate KL divergence.

Throughout training, the critic loss remained bounded within $0.2 < V_{loss} < 0.7$ for all adaptive methods, implying stable and well-calibrated value estimation. The policy loss magnitude ($|\pi_{loss}| < 0.05$) remained small and consistent, indicating that policy gradients were well-behaved and updates did not exhibit large oscillations.

Moreover, the approximate KL divergence between consecutive policy iterations remained below 0.02 across all updates, which lies within the PPO trust region boundary. This demonstrates that entropy modulation does not destabilize optimization, but instead acts as a regularization mechanism that constrains policy divergence.

Alignment with Theoretical Expectations

The theoretical foundation for entropy regularization suggests that maintaining a suitable entropy level prevents premature convergence to deterministic policies, thereby enhancing long-term exploration. The empirical entropy trajectories observed align with this theory:

- Baseline MAPPO's entropy rapidly collapses below 0.5, leading to reduced exploration.
- The linear entropy schedule temporarily sustains entropy (~1.0–1.3) but decays predictably.
- The adaptive entropy variant, however, maintains entropy near the target range of 1.5–1.6, dynamically adjusting according to the state of the policy.

This sustained entropy directly correlates with improved average step rewards, consistent with the principle that continuous exploration supports the discovery of cooperative behaviors in MARL settings.

Further validates this mechanism by showing smooth, monotonic increases in the adaptive coefficient β_t when entropy falls below the target $H_{target}=1.1$. Once sufficient exploration is reestablished, βt stabilizes near the upper bound ($\beta max=0.3$), confirming that the feedback-driven control loop operates as intended. This dynamic equilibrium between exploration and exploitation constitutes the core advantage of adaptive entropy regularization.

Robustness of Recurrent MAPPO + Auto-Tuning

The Recurrent MAPPO + Auto-Tuning variant introduces an additional source of complexity through recurrent state propagation. Despite this, it maintained low gradient variance and stable convergence. The inclusion of a GRU encoder allows each agent to accumulate temporal information, improving decision-making under partial observability. Importantly, entropy auto-tuning continued to function effectively in this recurrent setting, indicating that the adaptation rule generalizes beyond feed-forward architectures. The slightly slower convergence observed in early training is attributed to the recurrent warm-up phase, not instability. Over longer horizons, the recurrent variant demonstrated smoother long-term reward curves and better temporal credit assignment.

Reproducibility and Reliability

All experiments were implemented using PyTorch and executed on consistent computational setups. Hyperparameters, environment configurations, and training seeds were kept identical across all variants to ensure a fair comparison. The combination of low inter-seed variance, bounded loss dynamics, and theoretical consistency validates that the proposed methods are both reproducible and reliable. The adaptive entropy mechanism exhibits stable convergence behavior across runs without any signs of divergence, even when extended to recurrent actor networks.

Conclusion & Next Steps

Across twelve experimental runs (three seeds per method), the results consistently demonstrated that adaptive entropy mechanisms outperform fixed or scheduled entropy settings in both convergence stability and cooperative reward outcomes. The baseline MAPPO exhibited entropy collapse and early convergence, leading to suboptimal coordination.

In contrast, entropy scheduling delayed this collapse and improved average reward by approximately 40%.

The Auto-Tuned Entropy MAPPO achieved the most significant improvement, sustaining higher exploration levels ($H\approx1.55$) and achieving an average step reward of approximately -2.3, compared to -5.0 in the baseline.

The recurrent variant produced slightly lower peak performance (-2.6) but delivered greater temporal stability and smoother long-horizon learning curves, validating its robustness under partial observability.

From a theoretical standpoint, these findings reinforce the central role of entropy modulation in preventing deterministic policy collapse in MARL. The adaptive mechanism's ability to self-regulate the entropy coefficient $\beta t\beta\_t\beta t$ according to policy behavior establishes a closed feedback loop that automatically balances exploration and exploitation — a property that manual schedules cannot guarantee. This demonstrates that entropy can serve as an implicit form of dynamic regularization, improving both training stability and the diversity of emergent cooperative strategies.

Next Steps

Future work will extend the current investigation in several directions:

1. **Statistical Significance Testing:** Conduct multi-seed (≥10) experiments and compute confidence intervals for reward and entropy metrics to establish statistical validity of performance differences.
2. **Recurrent Policy Refinement:** Improve the R-MAPPO stability using *sequence normalization*, *burn-in unrolling*, and *state-space gating* to further enhance temporal credit assignment.
3. **Environment Generalization:** Evaluate the proposed methods on additional MPE scenarios such as `simple_reference_v3` and `simple_adversary_v3`, as well as on more complex benchmarks like **SMAC** and **Hanabi** to test scalability and generalization.
4. **Adaptive KL Penalty Integration:** Extend the auto-tuning mechanism to jointly regulate both entropy and KL-divergence thresholds, enabling self-adjusting trust regions for improved convergence safety.
5. **Multi-Agent Credit Assignment Analysis:** Investigate how entropy regulation interacts with decentralized value functions and attention-based critics, potentially leading to hybrid adaptive-regularization frameworks.

In conclusion, the experimental and theoretical results presented here confirm that Auto-Tuned Entropy MAPPO provides a robust improvement over state-of-the-art MAPPO baselines by introducing *adaptive exploration stability*.

The work lays a foundation for future MARL algorithms that can autonomously control their exploration dynamics, paving the way toward more general, self-regulating, and scalable multi-agent learning systems.

Reference

[1] H. Yu, M. Xu, S. Chen, J. Ma, and X. Zhao. (2021). The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Environments. arXiv:2103.01955.