Levering Organizational Hiearchy to Simplify Reward Design
in Cooperative Multi-agent Reinforcement Learning

Lixing Liu, Volkan Ustun, Rajay Kumar
Institute for Creative Technologies, University of Southern California
12015 E Waterfront Dr, Los Angeles, CA 90094, USA
{lxliu, ustun, kumar}@ict.usc.edu

Abstract
The effectiveness of multi-agent reinforcement learning (MARL) hinges largely on the meticulous arrangement of objectives. Yet, conventional MARL methods might not completely harness the inherent structures present in environmental states and agent relationships for goal organization. This study is conducted within the domain of military training simulations, which are typically characterized by complexity, heterogeneity, non-stationary and doctrine-driven environments with a clear organizational hierarchy and a top-down chain of command. This research investigates the approximation and integration of the organizational hierarchy into MARL for cooperative training scenarios, with the goal of streamlining the processes of reward engineering and enhancing team coordination. In the preliminary experiments, we employed two-tiered commander-subordinate feudal hierarchical (CSFH) networks to separate the prioritized team goal and individual goals. The empirical results demonstrate that the proposed framework enhances learning efficiency. It guarantees the learning of a prioritized policy for the commander agent and encourages subordinate agents to explore areas of interest more frequently, guided by appropriate soft constraints imposed by the commander.

Introduction
Recently, the practical use of cooperative Multiagent Reinforcement Learning (MARL) techniques has flourished, expanding into various real-world tasks, including robot control (Hadfield-Menell et al. 2016; Gupta, Egorov, and Kochenderfer 2017), financial operations (Fischer 2018), traffic planning (Chu et al. 2019), and professional-level game-play (Vinyals et al. 2019). Despite these strides, the transition from single-agent learning algorithms to cooperative multi-agent systems presents numerous open challenges due to environmental heterogeneity, the dynamics of agent interaction, and the need for multi-objective optimization (Yang and Wang 2020; Zhang, Yang, and Başar 2021; Oroojlooy and Hajinezhad 2023).

Organizational hierarchy refers to the arrangement of individuals and teams within corporations and institutions.

Related works
One of the key challenge in MARL is how to distribute the reward signal to individual agents, which is known as the credit assignment problem. The process of designing and learning such intricate rewards or intermediary models is consistently challenging and domain-specific. Common methods involve breaking down the global reward function into a linear combination of local reward functions, such as in CIRL (Hadfield-Menell et al. 2016), and linearly decomposing the value function as in QMIX (Rashid et al. 2020), or non-linearly as in QTRAN (Son et al. 2019). Hierarchical Reinforcement Learning (HRL) methods provide benefits over flattened MARL in terms of aggregating potentially conflicting objectives and concurrently optimizing policies across various agent levels and roles (Pateria et al. 2021). They have demonstrated their capacity to learn transferable abstractions within the same problem setting and offer reusable lower-level policies for similar scenarios without prior coordination mechanisms (Nachum et al. 2019).
We aim to investigate the appropriate approximation of inherent organizational structures in multi-agent systems for reward assignment and goal separation. Feudal Networks (FuN) (Vinyals et al. 2019) offer a versatile framework for goal splitting and reward hiding in a manager-worker setup, enabling us to effectively represent the commander-subordinate hierarchy.

Figure 1: A snapshot of the graph represented Scout-Mission training simulation environment scenarios.

Simulation Setup
In order to obtain general cooperative behaviors and arbitrary tactics, we create our customized graph-based topologically grounded environment (Figure 1) for a turn-based two-team (Red v.s. Blue) “Scout-Mission” competitive simulation scenario (Ustun et al. 2022). There are 116 major way-points, chosen by human experts, evenly distributed on the terrain graph, and each node has four immediate neighbors at most. The standard action space is composed of three action branches: a four-way movement action, a four-way facing action and a posturing action (i.e. Standing or Squatting). The graph environment manages state variables and other attributes such as damage and health. Agents on both sides are anticipated to develop cooperative strategies incorporating structured rewards, including damage, time, and distance components (Liu et al. 2021). To simplify the interpolation and validation of the acquired behavior policies, we limit the learning capabilities to just one team (the Reds) in this study. This configuration enabled the simulation environment in a setting of mixed-sum cooperation. Meanwhile, the opponents follow pre-defined heuristics to navigate towards their assigned target nodes. They will move at a quarter of their normal speed when they encounter threats.

Consider a simulation task where the primary team objective is to delay the opposing forces capturing key terrains until reinforcements arrive. To maintain control over prioritized team goals, we employ a two-level commander-subordinate feudal hierarchy (CSFH). To shape the problem as a sequential decision-making task, it’s essential to set up an intermediate task between these two decision levels. The commander’s objective is to postpone the advancement of the opposing forces for as long as possible by assigning a target node to each subordinate agent. The assignment is based on their individual positions in relation to the opponents, as well as the learned inductive bias and geo-specific preferences from previous training cycles. The subordinate agents utilize a common policy network, and their goals are to arrive at their designated targets as quickly as possible while interacting with the opponents.

Figure 2: The blue and orange curves show the normalized team rewards for the commander-subordinate policies and the baseline policies, respectively.

Experiment Results
We employ the PyTorch (Paszke et al. 2019) based RLLib (Liang et al. 2018) as the test bed to run experiments, and utilize the multiagent variant of the Proximal Policy Optimization method (Schulman et al. 2017) as the baseline algorithm. We evaluate and compare the performance of learning with and without the organizational hierarchy in the baseline test. Meanwhile, we attempted to determine how closely the subordinates should obey the commander’s order by adjusting reward distributions in ablation tests. To ensure a fair comparison, we report the mean of the delaying reward components across 20 random seeds for each test case.

Results show that the commander-subordinate feudal hierarchical framework significantly reduced the time to learn a stable policy with near-optimal team rewards (Figure 2). The optimal settings for the learning rate are $5 \times 10^{-4}$ for the CSFH model and $1 \times 10^{-4}$ for the baseline model. The CSFH policies with the best configurations have an average delaying reward that is 89.27% of the maximum possible reward, outperform the 76.37% reward of the best baseline policies. We argue that the benefits of leveraging this feudal hierarchy are largely due to the commander’s policy can notably reduce the exploration spaces of subordinates and steer them towards exploiting our areas of interest.

Conclusions
The primary goal of this work is to simplify reward design in MARL through the use of organizational hierarchy. We have observed the empirical success of CSFH in consistently enhancing the learning efficiency of prioritized tactical behaviors and acquiring robust, adaptable operational behaviors.
References


