SCALABLE MULTI-AGENT REINFORCEMENT LEARNING OVER HYPERGRAPHIC ENVIRONMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Scalability is a fundamental challenge in cooperative multi-agent reinforcement learning (MARL) for large-scale networked systems. Existing graph-based methods address this issue by sparsifying global coordination into pairwise edges and local neighborhoods, but always overlook the intrinsic overlapping higher-order interdependencies inherent in complex systems. In this paper, we propose HygRL, a novel MARL framework for networked systems with hypergraphic structure. We utilize hypergraphs to model the evolution of the multi-agent environment, where each hyperedge represents a group of interdependent agents. HygRL decomposes the global task into hyperedge-localized sub-tasks trained in parallel, while a mean policy aggregation resolves conflicts among overlapping hyperedges during decentralized execution. In particular, we theoretically derive the first performance error bound for hypergraph decomposition, quantifying the effects of approximate decomposition and aggregated policy sampling on training performance. As a special case, when the hypergraph degenerates to a single hyperedge, HygRL reduces to a CTDE algorithm and recovers the corresponding theoretical guarantees as the decomposition error vanishes. Experiments across three domains with up to 55 agents demonstrate that HygRL consistently outperforms CTDE, fully decentralized, and graph-based baselines.

1 INTRODUCTION

Scaling cooperative Multi-Agent Reinforcement Learning (MARL) to large-scale real-world systems remains a fundamental challenge. Although Centralized Training with Decentralized Execution (CTDE) has achieved remarkable success in application areas including infrastructure and industrial systems (Nie et al., 2025; Wang et al., 2025; van der Sar et al., 2023), internet security (Li et al., 2022), and network management (Selmonaj et al., 2023; Boussaoud et al., 2023), its reliance on aggregating data from all agents during centralized training becomes progressively less tractable as the number of agents scales to the tens or hundreds (Anand et al., 2025; Oroojlooy & Hajinezhad, 2023; Ma et al., 2024; Qu et al., 2022). At such scales, bandwidth and privacy constraints render the collection of global state-action information impractical (Zhang et al., 2024).

To improve scalability, existing research has pivoted towards localized learning schemes. One main approach, networked MARL, connects agents via a fixed communication graph and restricts policy updates to neighborhood information (Hu et al., 2021; Qu et al., 2022; Ma et al., 2024; Zhang et al., 2024). This reduces communication overhead and accommodates heterogeneity, but it has two limitations. First, the neighborhood and the interaction pattern can be mismatched: what enters an agent's update is its neighbors' position in the graph rather than whether they interact (Qu et al., 2022; Zhang et al., 2018). Second, a graph admits only pairwise relations, so projecting the system onto edges loses higher-order dependencies with no pairwise counterpart (Wang & Kleinberg, 2024; Lucas et al., 2026).

Hypergraphs offer a natural solution: by letting a hyperedge connect an arbitrary number of vertices, they model higher-order relationships natively (Jain et al., 2026; Zhu et al., 2024; Han et al., 2026; Liu & Li, 2025; Zhou et al., 2026). Recent hypergraph-based CTDE algorithms show promise by grouping agents according to a constructed hypergraph (Zhu et al., 2024), but they use this grouping for coordination rather than scalability: their centralized critics still scale with the agent count, and they give no theoretical account of the error introduced by the hypergraph. When hyperedges

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

overlap, they also stay representation-oriented, treating overlapped agents as feature-fusion nodes (Zhou et al., 2026) and leaving the arbitration among groups implicit.

In this paper, we propose HygRL, a hybrid semi-decentralized MARL framework that uses hypergraph structure for scalable coordination. Its distinctive idea is to localize CTDE within hyperedges, replacing the global centralized critic by a local one, so that the training cost does not scale with the total number of agents. At execution, each agent aggregates the policies it holds for its hyperedges into a single policy, so no communication is needed. Our main contributions are as follows.

(i) A scalable hybrid semi-decentralized framework. We design a hypergraph-based MARL framework that decomposes the global task into hyperedge-localized sub-tasks. Each hyperedge keeps a local joint policy, trained in parallel on shared data, so that its cost is independent of the total number of agents. The resulting architecture is hybrid and parallel: training is centralized within each hyperedge, hyperedges are trained in parallel, and execution is fully decentralized. To handle overlapping hyperedges, a multi-policy aggregation reconciles the preferences of the groups an agent belongs to at execution time, so that each agent acts on a single policy. This differs from the hypergraph-based CTDE methods above, which keep a global centralized critic and improve how the hypergraph is built. HygRL instead takes the hypergraph as a pluggable input and replaces that critic with hyperedge-localized training.

(ii) Error estimation for hypergraph decomposition. Our main theoretical contribution is to estimate the error incurred by this decomposition: locally, by comparing the true sampling distribution of a hyperedge with its surrogate model (Theorem 1), and globally, by comparing the true objective with the sum of the hyperedge-local objectives (Theorem 2). Combined with HATRL (Zhong et al., 2024), they yield the first performance error bound for hypergraph decomposition (Corollary 2); as the argument applies to any CTDE method embedded in our framework, it also gives a template for analyzing semi-decentralized training. In the degenerate case of a single hyperedge the estimates vanish and the guarantee reduces to that of the embedded algorithm, so HygRL is a provably safe generalization of CTDE.

(iii) An empirical study across three domains. We evaluate HygRL on wireless communication, power grid load balancing, and traffic signal control with up to 55 agents. Against CTDE (HAPPO, MAPPO), decentralized IPPO, and graph-based NPPO, it outperforms CTDE from 20 agents onward in final performance and sample efficiency, with the largest gains where higher-order cooperation is present.

2 Model Formulation

Higher-order interactions in networked systems give rise to group-level behavior that pairwise interactions alone do not reproduce (Boccaletti et al., 2023; Benson et al., 2016; Iacopini et al., 2019). Since the joint effect of a group is not determined by the pairwise interactions within it (Jain et al., 2026), the locality governing each agent's decision is group-wise rather than pairwise, which makes a hypergraph the more faithful model: a hyperedge represents an interaction among an arbitrary number of agents, whereas a graph represents only pairwise relations.

Motivated by the above observation, we consider a multi-agent Markov decision process (MAMDP) with hypergraphic structure, denoted HygMAMDP. Formally, n agents coexist in a hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{E})$, where each element of $\mathcal{E}$ is a non-empty subset of $\mathcal{N}$, and we define the HygMAMDP as

$$G = \langle \mathcal{H}, \{\mathcal{S}_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mu, P, \{r_i\}_{i\in\mathcal{N}}, \gamma \rangle. \tag{1}$$

Each agent i has its local state space $\mathcal{S}_i$ and local action space $\mathcal{A}_i$. The global state and action spaces are $\mathcal{S} = \prod_{i\in\mathcal{N}} \mathcal{S}_i$ and $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}_i$. The transition kernel is $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$. Each agent maintains a localized behavior policy $\pi_i(a_i|s_i)$, and the joint policy is $\pi(a|s) = \prod_{i\in\mathcal{N}} \pi_i(a_i|s_i)$. The reward function of each agent is $r_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and the global reward is $r(s,a) = \sum_{i\in\mathcal{N}} r_i(s,a)$. We assume each r_i to be bounded, i.e., $\sup_{z\in\mathcal{Z}} |r_i(z)| \le R_i$ for some constant $R_i \in \mathbb{R}$. The initial state distribution is μ and the discount factor is $\gamma \in [0,1)$. We define the global state-action space $\mathcal{S} \times \mathcal{A}$ as $\mathcal{Z}$ and $z := (s,a)$ as a state-action pair. The goal of reinforcement

learning is to find a joint policy π that maximizes the expected discounted return:

$$\pi = \arg\max_{\pi} J(\pi) = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]. \tag{2}$$

Discussion. This setup follows the networked MDP of Ma et al. (2024), but differs in the unit of factorization. Where the networked MDP factorizes the transition per agent, $p(s'|s,a) = \prod_i p_i(s'_i|s_{\mathcal{N}_i^\kappa}, a_i)$, we use a hypergraph in place of a graph and factorize per hyperedge, so that the agents of a hyperedge may remain coupled.

Using $\mathcal{H}$ to factorize the global problem yields a family of local MAMDPs, one per hyperedge, and we quantify the resulting error as follows. For each hyperedge e, the local state-action space is $\mathcal{Z}_e = \mathcal{S}_e \times \mathcal{A}_e$ with $\mathcal{S}_e = \prod_{i\in e} \mathcal{S}_i$ and $\mathcal{A}_e = \prod_{i\in e} \mathcal{A}_i$. We write $e = \{i_1, \dots, i_{m_e}\}$ with $m_e = |e|$, $n_i = |\mathcal{E}_i|$ for $\mathcal{E}_i = \{e \in \mathcal{E} : i \in e\}$, and $-\mathcal{X}$ for the complement of $\mathcal{X} \subseteq \mathcal{N}$. The local reward is $\tilde{r}_e(s,a) = \sum_{i\in e} w_{i,e}\, r_i(s,a)$, with weights $w_{i,e}$ satisfying $\sum_{e\in\mathcal{E}_i} w_{i,e} = 1$ and $R_e = \sum_{i\in e} |w_{i,e}| R_i$. We take the local initial state distribution to be the marginal of μ onto $\mathcal{S}_e$, so that only the transition kernel and the reward are approximated. The following definition quantifies these two approximations.

Definition 1 (Hypergraphic Environment)**.** *An HygMAMDP is a hypergraphic environment with error profile $\{(\epsilon_p(e), \epsilon_r(e))\}_{e\in\mathcal{E}}$ if the following two conditions are satisfied:*

***(i):** For each hyperedge $e \in \mathcal{E}$, there exists a **surrogate transition kernel** $P_e : \mathcal{Z}_e \to \Delta(\mathcal{S}_e)$ such that*

$$\sup_{z\in\mathcal{Z}} \left\| \int_{\mathcal{S}_{-e}} P(\cdot, \mathrm{d}s'_{-e} \mid z) - P_e(\cdot \mid z_e) \right\|_{\mathrm{TV}} \leq \epsilon_p(e), \tag{3}$$

where $\epsilon_p(e) \in [0,1]$ is a constant.

***(ii):** For each hyperedge $e \in \mathcal{E}$, there exists a **surrogate reward function** $r_e : \mathcal{Z}_e \to \mathbb{R}$ such that*

$$\sup_{z\in\mathcal{Z}} |\tilde{r}_e(z) - r_e(z_e)| \leq \epsilon_r(e) R_e, \tag{4}$$

where $\epsilon_r(e) \in [0, \infty)$ is a constant.

Condition (i) states that the next local state is determined, up to $\epsilon_p(e)$, by the state-action pair inside e alone. The surrogate kernel P_e is a function of z_e only, and therefore cannot reproduce any dependence of the next local state on the complement; $\epsilon_p(e)$ measures precisely that residual dependence, and vanishes when there is none.

Discussion. The hypergraphic environment generalizes the independent networked system (INS) (Ma et al., 2024; Qu et al., 2022) in a different direction: what separates the two is the granularity of the decomposition. Under this choice of the local initial distribution, each hyperedge inherits a well-defined sub-problem, with the transition kernel and the reward as its two approximations. INS instead adopts node-level conditional independence and assigns the entire system a single error constant ξ, whereas the errors here are per-hyperedge and listed as $(\epsilon_p(e), \epsilon_r(e))$ for each e.

Remark. The quantities $\epsilon_p(e)$ and $\epsilon_r(e)$ accompany the decomposition rather than the algorithm: for a given $\mathcal{H}$ they are the worst-case errors of the best local model, they vanish for $e = \mathcal{N}$, and they enter the bounds of Theorems 1 and 2 explicitly, so the guarantees remain valid for any error profile and tighten as the errors shrink. In particular, $\epsilon_p(e)$ measures only the influence of the complement on the local dynamics, since the local initial distribution is exact by construction. Their role is thus to make the cost of decomposing the system measurable *in principle*; Appendix B gives constructions under which they are small.

3 Our Method

In this section, we introduce HygRL, a hypergraph-based hybrid semi-decentralized MARL framework. We begin with a high-level overview of the algorithmic architecture, then detail its two core mechanisms, hyperedge-localized training and multi-policy aggregation. Practical hypergraph construction is discussed in Appendix B.

Algorithm 1: Hypergraph Reinforcement Learning (HygRL)

Require: HygMAMDP G, hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{E})$, CTDE algorithm $\mathcal{A}$, iterations K
Ensure: Aggregated policies $\{\bar{\pi}_i\}_{i\in\mathcal{N}}$
1: **for** each hyperedge $e \in \mathcal{E}$ **do**
2: Initialize local policies $\{\pi_{i,e}\}_{i\in e}$ and critic C_e via $\mathcal{A}$
3: Let $\bar{\pi}_i(a_i|s_i) = \frac{1}{n_i}\sum_{e\in\mathcal{E}_i}\pi_{i,e}(a_i|s_i)$
4: **end for**
5: **for** iteration $k = 1$ to K **do**
6: Sample trajectories using aggregated policies $\{\bar{\pi}_i\}_{i\in\mathcal{N}}$
7: Each agent stores (s_i, a_i, r_i, s_i') in local buffer
8: **for** each hyperedge $e \in \mathcal{E}$ **in parallel do**
9: Gather data $\{(s_i, a_i, r_i, s_i')\}_{i\in e}$
10: Construct $s_e = (s_i)_{i\in e}$, $a_e = (a_i)_{i\in e}$, $s_e' = (s_i')_{i\in e}$
11: Compute local reward $\tilde{r}_e = \sum_{i\in e} w_{i,e} r_i$ {Here $w_{i,e} = \frac{1}{n_i}$}
12: Update $\{\pi_{i,e}\}_{i\in e}$ and critic C_e using $\mathcal{A}$ with transition $(s_e, a_e, \tilde{r}_e, s_e')$
13: **end for**
14: **for** each agent $i \in \mathcal{N}$ **do**
15: Update $\bar{\pi}_i(a_i|s_i) = \frac{1}{n_i}\sum_{e\in\mathcal{E}_i}\pi_{i,e}(a_i|s_i)$
16: **end for**
17: **end for**
18: **return** $\{\bar{\pi}_i\}_{i\in\mathcal{N}}$

3.1 Algorithmic Overview

The central challenge we address is how to scale cooperative MARL to large systems without sacrificing coordination quality. Our key insight is that many real-world multi-agent systems exhibit *higher-order cooperative structures*: groups of agents naturally form collaborative units that transcend pairwise interactions. By decomposing the global system into such units (hyperedges), we can localize centralized training to each unit while preserving global coherence through careful policy aggregation.

Given a hypergraphic environment, HygRL operates as follows. Instead of training a single joint policy over all agents, we assign to each hyperedge $e \in \mathcal{E}$ an exclusive local joint policy $\pi_e = \prod_{i\in e}\pi_{i,e}$ over its constituent agents. Each agent i thus maintains a set of policies $\{\pi_{i,e}\}_{e\in\mathcal{E}_i}$, one for each hyperedge it participates in. The overall workflow is shown in Algorithm 1, and its two phases are separated in time: trajectories are always collected with the aggregated policies, whereas each $\pi_{i,e}$ is updated from data that the aggregated policies have already produced.

Execution phase (lines 6–7). Agents act independently and in parallel using their aggregated behavior policies $\bar{\pi}_i$, defined as the average of their hyperedge-specific policies:

$$\bar{\pi}_i(a_i|s_i) = \frac{1}{n_i}\sum_{e\in\mathcal{E}_i}\pi_{i,e}(a_i|s_i). \tag{5}$$

This aggregation mechanism, detailed in Section 3.3, blends the policies only at execution time.

Off-Policy Training phase (lines 8–16). For each hyperedge e, we gather the transitions of its constituent agents, which were collected under the aggregated policies $\bar{\pi}$, and construct a local joint transition $(s_e, a_e, \tilde{r}_e, s_e')$; the local policy π_e is then updated using an arbitrary CTDE algorithm (e.g., MAPPO or HATRL). Crucially, because policy ownership is disjoint across hyperedges, each $\pi_{i,e}$ belongs to exactly one hyperedge, these updates can be executed in a *fully parallelized fashion*.

Remark. We call this scheme semi-decentralized because training is centralized only within each hyperedge, where a local critic and a local joint policy are learned, while execution is fully decentralized: each agent acts on its local observation using the aggregated policy $\bar{\pi}_i$. In this sense the architecture is hybrid, combining CTDE within each hyperedge with fully decentralized execution across hyperedges.

Discussion. Three further aspects of the architecture warrant emphasis. First, it reduces the training unit: for clique-based hyperedges, where each hyperedge is a clique of an underlying graph $\mathcal{G}$, we have $|e| \le \min_{i\in e}\deg(i)+1$, and on Erdős–Rényi graphs $G(n,p)$ with fixed p the maximum clique

size is $\Theta(\log n)$ (Bollobás, 2001, Corollary 11.2) while the maximum degree is $\Theta(n)$ (Bollobás, 2001, Corollary 3.14). Constructing one hyperedge per maximum clique therefore yields training units of size $\Theta(\log n)$ in expectation, in contrast to the $\Theta(n)$ neighbors that graph-based methods involve per agent. Second, it is flexible: the framework imposes no requirements on the embedded CTDE algorithm, which can be value-based (e.g., QMIX (Rashid et al., 2018)) or policy-based (e.g., MADDPG (Lowe et al., 2017), MAPPO (Yu et al., 2022)), making HygRL a *meta-algorithm* that inherits the strengths of its base algorithm while adding scalability. Third, it keeps the communication requirement of graph-based methods: for clique-based hyperedges the graph serves directly as the communication graph, so the agents of a hyperedge are exactly those within direct communication range of one another. Training therefore runs under the same topology-dependent communication constraints as graph-based methods: any setting in which the latter can train, HygRL can train as well.

3.2 Hyperedge-Localized Training

The first pillar of our approach is the decomposition of the global problem into the hyperedge-localized sub-tasks defined in Definition 1. Recall that, for each hyperedge $e \in \mathcal{E}$, the local reward is defined as $\tilde{r}_e(s,a) = \sum_{i \in e} w_{i,e} r_i(s,a)$, where $\sum_{e \in \mathcal{E}_i} w_{i,e} = 1$. With this weighting, the local rewards sum to the global reward exactly,

$$\sum_{e \in \mathcal{E}} \tilde{r}_e(s,a) = \sum_{e \in \mathcal{E}} \sum_{i \in e} w_{i,e} r_i(s,a) = \sum_{i \in \mathcal{N}} r_i(s,a) = r(s,a),$$

which allows the global objective to be decomposed into per-hyperedge terms.

In Algorithm 1, we choose $w_{i,e} = \frac{1}{n_i}$, which reflects a principled neutrality: since agent i participates in n_i hyperedges, it allocates equal importance to each. Adaptive weighting schemes, such as learned attention weights, could further improve performance and are left to future work.

3.3 Mean Policy Aggregation

The second pillar addresses a subtle but crucial challenge: each agent belongs to multiple hyperedges, and these hyperedges may have conflicting preferences over the agent's actions. For instance, extending the main-road green phase is best for the arterial-coordination group covering a signal's incoming links but worst for the downstream-evacuation group covering its outgoing links. Since a single policy must commit to one action distribution at every state, it can represent only one of the two preferences.

Our solution is to equip each agent with *multiple policies*, one per incident hyperedge, and aggregate them *only at execution time* via Eq. (5). This design has two important properties:

(i) Preservation of specialized behavior. Each policy $\pi_{i,e}$ can specialize to the cooperative objective of hyperedge e without interference from other hyperedges, because training occurs independently per hyperedge. This allows the agent to simultaneously learn distinct behavioral modes tailored to different cooperative contexts.

(ii) Implicit conflict resolution. The mean aggregation $\bar{\pi}_i$ is exactly the distribution that minimizes the total divergence from the group-specific policies,

$$\bar{\pi}_i = \arg\min_q \sum_{e \in \mathcal{E}_i} \mathrm{D}_{\mathrm{KL}}\left(\pi_{i,e} \| q\right), \tag{6}$$

so it acts as a soft vote: a group can neither impose an action the others assign low probability nor by itself drive an action to high probability. The disagreement is thus preserved during training and resolved only at execution.

In general, this aggregation mechanism is the execution-side counterpart of semi-decentralized training: policies are specialized within hyperedges during training, and reconciled only at execution time without centralized coordination. This differs from hypergraph methods that handle overlapping memberships in value space, such as HIVE (Zhou et al., 2026), which requires an allocation rule over the joint value function rather than a composition of policies. The resulting aggregation depends only on the hypergraph, so no additional structure needs to be imposed on the system to apply it.

4 THEORETICAL ANALYSIS

In this section we establish two error-estimation theorems and give an application example. The first bounds, for each hyperedge, the gap between its true marginal and the local MAMDP. The second bounds the gap between the joint objective and the sum of the local objectives, which is the error incurred by splitting the global problem into local MAMDPs. The analysis is non-trivial because each local MAMDP is trained on offline samples collected by the aggregated behavior policy rather than by its own policy, so the standard on-policy guarantees are not available. We therefore relax the notion of an occupancy measure to one that only carries the normalization mass $1/(1-\gamma)$ and need not be generated by any Markov decision process, which we term a *format occupancy measure* (Definition 10 in Appendix A.2); the hyperedge marginals of the global MAMDP are of this form, which allows us to construct a local MAMDP for each hyperedge. As an example, we then apply the two theorems to one specific algorithm, HATRL (Zhong et al., 2024), obtaining the corresponding performance bounds.

To conduct the analysis, we need one additional assumption.

Assumption 1. *For each agent i, there exists a constant δ_i such that for all hyperedges $e \ni i$,*

$$\mathrm{D}_{\mathrm{TV}}^{\max}\left(\pi_{i,e}, \bar{\pi}_i\right) \leq \delta_i, \tag{7}$$

where $\mathrm{D}_{\mathrm{TV}}^{\max}\left(\pi, \pi'\right) = \sup_{s\in\mathcal{S}} \|\pi(\cdot|s) - \pi'(\cdot|s)\|_{\mathrm{TV}}$.

Assumption 1 requires that the policies an agent maintains across different hyperedges remain close to their average. By Proposition 8 in Appendix A, $\mathrm{D}_{\mathrm{TV}}^{\max}\left(\pi_{i,e}, \bar{\pi}_i\right) \leq \frac{n_i-1}{n_i}$; therefore δ_i always exists and is less than $\frac{n_i-1}{n_i}$. In particular, δ_i vanishes when $n_i = 1$, i.e., when agent i belongs to a single hyperedge.

Together, Definition 1 and Assumption 1 quantify both sources of error. Because the true marginal of a hyperedge is not generated by any local MAMDP, and each hyperedge is trained off-policy, the two are separated by the mismatch of the surrogate model and the divergence between $\pi_{i,e}$ and $\bar{\pi}_i$.

Theorem 1. *Under the conditions of Definition 1 and Assumption 1, with bounded rewards as assumed in the model formulation, for each hyperedge $e \in \mathcal{E}$, the following bounds hold:*

$$\begin{aligned}
\left\|\eta^e_{\mu,P,\bar{\pi}} - \eta_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} &\leq \frac{\gamma}{(1-\gamma)^2}\mathcal{F}\left(\phi_e\right), \\
\left\|\rho^e_{\mu,P,\bar{\pi}} - \rho_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} &\leq \frac{1}{(1-\gamma)^2}\mathcal{F}_{1+\epsilon_p(e)}\left(\phi_e\right) - \frac{\epsilon_p(e)}{1-\gamma}, \\
\left\|\rho^e_{\mu,P,\bar{\pi}}(\cdot|z'_e) - \rho_{\mu_e,P_e,\pi_e}(\cdot|z'_e)\right\|_{\mathrm{TV}} &\leq \frac{\gamma}{(1-\gamma)^2}\mathcal{F}(\phi_e), \\
|r_e(z_e) - r^e_{\mu,P,\bar{\pi}}(z_e)| &\leq \epsilon_r(e)R_e, \\
|r^e_{\mu,P,\bar{\pi}}(z_e)| &\leq R_e,
\end{aligned}$$

where the symbols on the left-hand side carrying the superscript e describe the marginals of the global MAMDP on hyperedge e, and the remaining ones on the left-hand side describe the quantities of the local MAMDP on e. Here $\phi_e = \epsilon_p(e) + \sum_{i\in e}\delta_i$ is a constant. $\mathcal{F}$ is an auxiliary function. These notations are defined in detail in Appendix A.

Theorem 1 bounds, inside each hyperedge, the gap between the true marginal and the local MAMDP. The remaining question is whether these local gaps accumulate when the hyperedges are recombined, since the joint objective is what the algorithm ultimately optimizes. In what follows, $J(\pi|\mu, P, r)$ denotes the expected discounted return of policy π under initial distribution μ, transition kernel P and reward r. The following theorem answers this question.

Theorem 2. *Under the same conditions and notations as Theorem 1, which are defined in Appendix A, the global expected return satisfies*

$$\left|J(\bar{\pi}|\mu, P, r) - \sum_{e\in\mathcal{E}} J(\pi_e|\mu_e, P_e, r_e)\right| \leq \sum_{e\in\mathcal{E}}\left(\frac{2\mathcal{F}_{1+\epsilon_p(e)}(\phi_e)}{(1-\gamma)^2} + \frac{3\epsilon_r(e) - 2\epsilon_p(e)}{1-\gamma}\right)R_e. \tag{8}$$

Discussion. The two theorems bound the two gaps into which the error of the local MAMDPs decomposes. Theorem 1 concerns a single hyperedge: the gap between the true marginal and the

local MAMDP is governed by the divergence introduced by offline training, δ_i, together with the approximation errors introduced by the split, $\epsilon_p(e)$ and $\epsilon_r(e)$, and the four approximation bounds vanish as these errors go to zero. Theorem 2 concerns the recombination: the gap between the joint objective and the sum of the local objectives decomposes additively over hyperedges, with the dynamics and policy errors entering through the same ϕ_e and the reward error entering linearly. The two estimates therefore cover the two sources of error separately, which is what allows them to be combined into a single performance bound.

These estimates are statements about approximation, not yet about performance. We now give the example announced above, applying them to the HATRL update rule (Zhong et al., 2024) and starting from a single hyperedge.

Corollary 1. *Under the same conditions, if each hyperedge e updates using HATRL with an update order $\langle i_1, \dots, i_{m_e} \rangle \in Sym(e)$, i.e., a permutation of the agents in e, then*

$$J(\pi'_e|\mu_e, P_e, r_e) \geq J(\pi_e|\mu_e, P_e, r_e) - D_e R_e,$$

where D_e is a constant defined in the restated corollary in Appendix A.5.

Corollary 1 bounds the effect at one hyperedge; combined with Theorem 2, it yields the corresponding bound on the global update.

Corollary 2. *Under the same conditions, when HygRL is equipped with HATRL, the global policy update satisfies*

$$J(\bar{\pi}'|\mu, P, r) \geq J(\bar{\pi}|\mu, P, r) - \sum_{e \in \mathcal{E}} M_e R_e,$$

where M_e is a constant defined in the restated corollary in Appendix A.6.

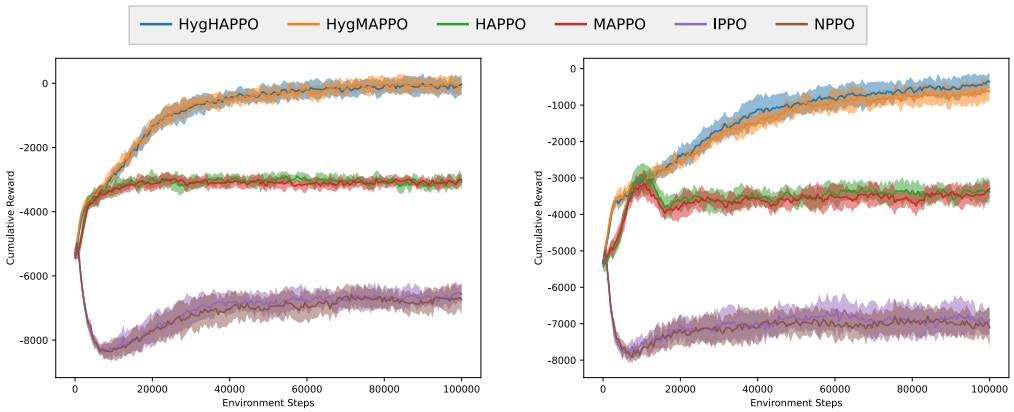
Discussion. Corollary 1 quantifies the degradation at a single hyperedge: the constant D_e is determined by $\epsilon_p(e)$, $\epsilon_r(e)$, δ_i and R_i, and grows with the hyperedge size m_e. Corollary 2 aggregates this into the global bound $\sum_{e \in \mathcal{E}} M_e R_e$, whose every term depends only on quantities attached to the hyperedge itself, namely its size m_e, the discount factor γ, and the error profile $(\epsilon_p(e), \epsilon_r(e))$, or on quantities attached to each agent $i \in e$, namely its policy divergence δ_i, its reward bound R_i, and the number n_i of hyperedges it belongs to. Writing $n_{\max,e} = \max_{i \in e} n_i$ for the largest number of hyperedges that share an agent, the explicit form of M_e in Appendix A.6 gives $M_e = O(m_e \log n_{\max,e})$ at a fixed error profile, so the bound tightens as the hypergraph becomes sparser, but never depends on the total number of agents n. When $\mathcal{H}$ degenerates to $(\mathcal{N}, \{\mathcal{N}\})$, the surrogate is exact and $\pi_{i,\mathcal{N}} = \bar{\pi}_i$, so all error terms vanish and $M_e = 0$. Corollary 2 therefore recovers the HATRL guarantee (Theorem 7 of Zhong et al. (2024)); the same holds for any CTDE algorithm embedded into HygRL, since a single hyperedge makes the decomposition trivial. HygRL is thus a provably safe generalization of CTDE, and the same procedure provides a template for analyzing semi-decentralized training. Complete proofs are given in Appendix A.

5 Experiments

We evaluate HygRL across three domains: wireless communication, power grid load balancing (reported in Appendix C.2), and traffic signal control. We embed two representative CTDE algorithms, HAPPO (Zhong et al., 2024) and MAPPO (Yu et al., 2022), into HygRL, yielding HygHAPPO and HygMAPPO respectively. Baselines include the original CTDE counterparts (HAPPO, MAPPO), fully decentralized IPPO (Tan, 1993), and graph-based NPPO (Neighbor PPO), a re-implementation of the networked MARL framework of Qu et al. (2022) in which we replace its actor-critic with PPO so that all methods share one backbone and the comparison isolates the decomposition; this is the only deviation from the original method and favors the baseline, since every other method uses the same optimizer. All experiments are run with 10 random seeds; shaded regions in figures represent mean $\pm$ standard deviation. Complete results, statistical significance, hyperparameters and detailed environment descriptions are provided in Appendices C, D.1 and D.2.

5.1 Wireless Communication

Following Qu et al. (2022), we consider a modified wireless communication problem with n users and m access points. We evaluate on 5, 20, 35, and 50 agents with both MLP and GRU policy parameterizations. Environment details and complete results are given in Appendix C.1.



(a) 35 agents, MLP

(b) 35 agents, GRU

Figure 1: Training curves on Wireless Communication with 35 agents under both parameterizations. Each curve reports the cumulative reward over training steps; shaded regions represent the standard deviation over 10 random seeds.

Setup	HygHAPPO	HAPPO	HygMAPPO	MAPPO	IPPO	NPPO
5 agents, GRU	**257.46** ±**2.26**	**257.27** ±**8.64**	**257.50** ±**0.79**	**257.88** ±**8.56**	94.77 ±44.25	82.29 ±24.21
20 agents, GRU	**729.45** ±**33.94**	−389.89 ±46.08	**705.58** ±**23.32**	−339.66 ±153.29	28.49 ±88.35	78.89 ±98.81
35 agents, GRU	**−75.80** ±**42.90**	−654.01 ±44.45	−123.78 ±47.26	−669.38 ±50.37	−1375.09 ±68.06	−1414.99 ±85.14
50 agents, GRU	**1211.25** ±**165.82**	174.68 ±281.26	**1198.69** ±**117.83**	429.05 ±401.44	−1034.18 ±132.13	−927.97 ±182.90

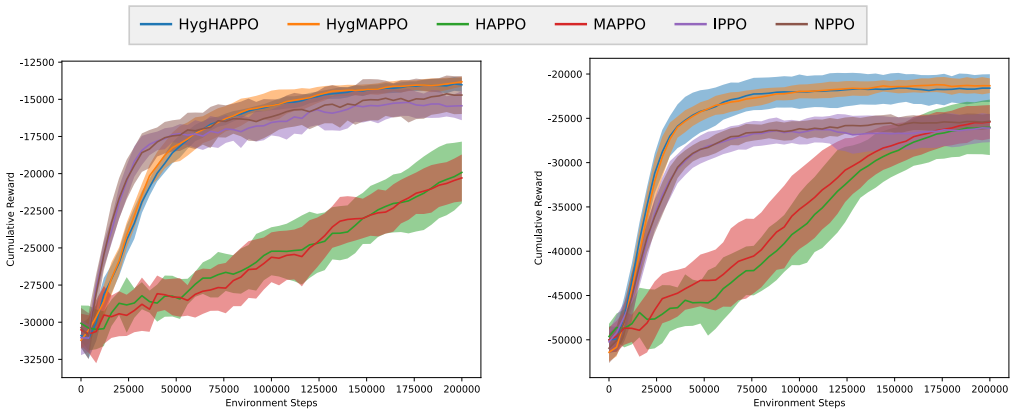
Table 1: Average performance on Wireless Communication (mean ± std). Best results are underlined; bold indicates within one standard deviation of the best result.

Figure 1 and Table 1 show that for 5 agents all CTDE-based methods perform similarly. However, as the agent count increases, HygHAPPO and HygMAPPO significantly outperform their CTDE counterparts. For example, in the 20-agent GRU setting, HAPPO and MAPPO show virtually no effective learning, while HygRL maintains robust performance. IPPO and NPPO fail beyond 20 agents, and already underperform markedly at 5, the latter confirming that graph-based neighbor aggregation is insufficient here.

5.2 Traffic Signal Control with LightSim

We test our algorithm on LightSim (Su & Deng, 2026), a lightweight traffic simulator. Agents optimize traffic flow by controlling traffic light phases; we adopt the built-in "pressure" objective, which minimizes the imbalance between upstream and downstream densities at each intersection. We evaluate on grid networks (4×2, 4×4, 4×6) and four real-world OpenStreetMap networks (Beijing, 53 agents; Manhattan, 33 agents; San Francisco, 47 agents; Los Angeles, 55 agents). We evaluate both MLP and GRU parameterizations on the grid networks. Since MLP and GRU yielded similar trends across all grid configurations (Figure 6 in Appendix C.3), we omit the real-world networks under the GRU parameterization. Complete results are given in Appendix C.3.

As shown in Figure 2 and Table 2, in small grids (4×2), HygRL performs comparably to CTDE methods. As the grid size increases to 4×4 and 4×6, HygHAPPO and HygMAPPO demonstrate increasingly clear advantages in both final performance and sample efficiency. In real-world networks (up to 55 agents), HygRL achieves faster convergence and higher final returns than CTDE in most cases. NPPO underperforms HygRL in all but the smallest grid, confirming that clique-based decomposition captures the traffic network structure better than pairwise neighborhoods.



(a) grid-4x6, (GRU, 24 agents)

(b) Beijing, (MLP, 53 agents)

Figure 2: Training curves on LightSim for a synthetic grid and a real-world network, with both parameterizations given in the subcaptions. Each curve reports the cumulative reward over training steps; shaded regions represent the standard deviation over 10 random seeds.

Setup	HygHAPPO	HAPPO	HygMAPPO	MAPPO	IPPO	NPPO
grid-4x2, GRU	-1797.35 ± 190.00	-1938.15 ± 137.48	-1783.87 ± 171.54	-1971.46 ± 128.35	$\mathbf{-1626.56}$ $\mathbf{\pm 52.03}$	$\underline{\mathbf{-1609.56}}$ $\underline{\mathbf{\pm 20.37}}$
grid-4x4, GRU	$\underline{\mathbf{-2136.09}}$ $\underline{\mathbf{\pm 50.84}}$	-2947.78 ± 287.82	$\mathbf{-2176.69}$ $\mathbf{\pm 123.45}$	-2839.86 ± 125.37	-2400.20 ± 239.24	-2479.12 ± 137.91
grid-4x6, GRU	$\mathbf{-2802.04}$ $\mathbf{\pm 90.36}$	-3986.09 ± 414.72	$\underline{\mathbf{-2771.94}}$ $\underline{\mathbf{\pm 72.56}}$	-4056.21 ± 306.40	-3076.33 ± 167.69	-2933.97 ± 256.47
Beijing, MLP	$\mathbf{-4305.76}$ $\mathbf{\pm 308.04}$	-5180.00 ± 624.51	$\underline{\mathbf{-4248.85}}$ $\underline{\mathbf{\pm 165.76}}$	-5087.61 ± 389.47	-5224.31 ± 315.22	-5064.95 ± 171.80
Manhattan, MLP	$\mathbf{-3421.48}$ $\mathbf{\pm 274.61}$	-3693.36 ± 266.12	$\underline{\mathbf{-3276.76}}$ $\underline{\mathbf{\pm 211.76}}$	-3596.09 ± 409.17	-3585.38 ± 158.70	$\mathbf{-3460.44}$ $\mathbf{\pm 173.11}$

Table 2: Average performance on LightSim (mean $\pm$ std). Best results are underlined; bold indicates within one standard deviation of the best result.

5.3 Summary

Across the domains reported above and the power grid domain of Appendix C.2, up to 55 agents, HygRL outperforms CTDE from 20 agents onward, and outperforms NPPO where higher-order cooperation is present. The trend is consistent: CTDE methods match HygRL at 5 agents but degrade rapidly beyond 20, whereas HygRL remains stable. The advantage stems from its *hybrid semi-decentralized design*: (i) hyperedge-localized training keeps the per-hyperedge training scale constant regardless of the total agent count, preserving sample efficiency; and (ii) embedding a CTDE algorithm within each hyperedge retains the coordinated behavior that joint optimization provides, so the gains are preserved rather than traded away for scalability. The multi-policy aggregation further reconciles the policies an agent holds across overlapping hyperedges, which lets that coordinated behavior survive when groups overlap.

6 Conclusion

We proposed HygRL, a hypergraph-based hybrid semi-decentralized MARL framework that decomposes global cooperative tasks into hyperedge-localized sub-tasks trained in parallel, with mean policy aggregation resolving conflicts across overlapping groups. We derived performance bounds quantifying the degradation induced by semi-decentralized training, and proved that HygRL recovers the standard CTDE guarantees when the hypergraph degenerates to the trivial case. Experiments across three domains with up to 55 agents show that HygRL consistently outperforms CTDE, fully decentralized, and graph-based baselines, leaving adaptive hypergraph evolution and reward weighting to future work.

AI USE STATEMENT

In this work, we used generative AI tools for drafting and editing portions of the paper manuscript, assisting with translation. We have not used generative AI tools for formulating mathematical claims, proposing or refining hypotheses, designing the research methodology or experiments, interpreting results, or implementing the core algorithm. The remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for proofreading and grammar correction, formatting references, and sourcing/searching for relevant literature. We reviewed all AI-generated text for technical accuracy and all AI-suggested references were verified against the original sources. LLM-assisted code generation was confined to auxiliary tooling and scripting; the methods, the theoretical results, and the experimental pipeline were implemented by the authors. All AI-assisted code was manually reviewed, tested, and corrected where necessary. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, an anonymized version of the source code is provided as supplementary material, covering the environments, all baselines, and the scripts that reproduce every figure and table. The code will be released under an open-source license upon publication.

For each experiment, we conducted 10 independent runs with different randomly drawn seeds, rather than fixing a specific seed. This design accounts for the inherent stochasticity in environment initialization, transition dynamics, and policy sampling. The reported results (mean and standard deviation over the 10 runs) reflect this variability. Complete training curves, performance tables, and statistical significance tests (Cohen's d and p-values) are provided in Appendix C.

The experiment implementation details and hyperparameter configurations are described in Appendix D.1 and D.2, respectively. Additional details, including graph construction procedures and environment design, are provided in Appendix B and Appendix C, respectively.

REFERENCES

Emile Anand, Ishani Karmarkar, and Guannan Qu. Mean-field sampling for cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 8462–8521, 2025.

Austin R. Benson, David F. Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353:163–166, 2016.

Andrey Bernstein and Emiliano Dall'Anese. Linear power-flow models in multiphase distribution networks. In *2017 IEEE PES Innovative Smart Grid Technologies Conference Europe*, pp. 1–6, 2017.

Olivier Binette. A note on reverse Pinsker inequalities. *IEEE Transactions on Information Theory*, 65(7):4094–4096, 2019.

S. Boccaletti, P. De Lellis, C. I. del Genio, K. Alfaro-Bittner, R. Criado, S. Jalan, and M. Romance. The structure and dynamics of networks with higher order interactions. *Physics Reports*, 1018: 1–64, 2023.

Béla Bollobás. *Random Graphs*. Cambridge University Press, 2nd edition, 2001.

Kaoutar Boussaoud, Meryeme Ayache, and Abdeslam En-Nouaary. A multi-agent reinforcement learning approach for congestion control in network based-SDN. In *10th International Conference on Future Internet of Things and Cloud*, pp. 250–255, 2023.

Chen Chu, Yong Li, Jinzhuo Liu, Shuyue Hu, Xuelong Li, and Zhen Wang. A formal model for multiagent q-learning dynamics on regular graphs. In Luc De Raedt (ed.), *Proceedings of the 31st International Joint Conference on Artificial Intelligence*, pp. 194–200, 2022.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, revised edition, 1977.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.

Yi Han, Xiaojun Wang, Ji Qiao, Zhao Liu, Jiaming Dou, Minjie Jin, and Pei Zhang. A hypergraph-enhanced multi-agent reinforcement learning for coordinated operation of networked microgrids under contingency. *CSEE Journal of Power and Energy Systems*, pp. 1–14, 2026.

Jueming Hu, Zhe Xu, Weichang Wang, Guannan Qu, Yutian Pang, and Yongming Liu. Decentralized graph-based multi-agent reinforcement learning using reward machines, 2021. arXiv:2110.00096.

Iacopo Iacopini, Giovanni Petri, Alain Barrat, and Vito Latora. Simplicial models of social contagion. *Nature Communications*, 10(1):2485, 2019.

Rishabh Jain, Keisuke Okumura, Michael Amir, Pietro Lio, and Amanda Prorok. Pairwise is not enough: Hypergraph neural networks for multi-agent pathfinding, 2026. arXiv:2602.06733.

David A Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2nd edition, 2017.

Tianxu Li, Kun Zhu, Nguyen Cong Luong, Dusit Niyato, Qihui Wu, Yang Zhang, and Bing Chen. Applications of multi-agent reinforcement learning in future internet: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 24(2):1240–1279, 2022.

Chiqiang Liu and Dazi Li. HYGMA: Hypergraph coordination networks with dynamic grouping for multi-agent reinforcement learning. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Maxime Lucas, Luca Gallo, Arsham Ghavasieh, Federico Battiston, and Manlio De Domenico. Reducibility of higher-order networks from dynamics. *Nature Communications*, 17(1):1551, 2026.

Chengdong Ma, Aming Li, Yali Du, Hao Dong, and Yaodong Yang. Efficient and scalable reinforcement learning for large-scale network control. *Nature Machine Intelligence*, 6(9):1006–1020, 2024.

Lei Nie, Dandan Qi, Bingyi Liu, Peng Li, Haizhou Bao, and Heng He. CMRM: Collaborative multi-agent reinforcement learning for multi-objective traffic signal control. *IEEE Transactions on Consumer Electronics*, 71(2):2793–2805, 2025.

Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning for multiagent networked systems. *Operations Research*, 70(6):3601–3628, 2022.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4295–4304, 2018.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016.

Ardian Selmonaj, Oleg Szehr, Giacomo Del Rio, Alessandro Antonucci, Adrian Schneider, and Michael Rüegsegger. Hierarchical multi-agent reinforcement learning for air combat maneuvering. In *International Conference on Machine Learning and Applications*, pp. 1031–1038, 2023.

Haoran Su and Hanxiao Deng. LightSim: A lightweight cell transmission model simulator for traffic signal control research, 2026. arXiv:2602.21852.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning*, pp. 330–337, 1993.

Erica van der Sar, Alessandro Zocca, and Sandjai Bhulai. Multi-agent reinforcement learning for power grid topology optimization, 2023. arXiv:2310.02605.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Ming Wang, Jie Zhang, Peng Zhang, and Mengyu Jin. Cooperative multi-agent reinforcement learning for multi-area integrated scheduling in wafer fabs. *International Journal of Production Research*, 63(8):2871–2888, 2025.

Yanbang Wang and Jon Kleinberg. From graphs to hypergraphs: Hypergraph projection and its reconstruction. In *International Conference on Learning Representations*, volume 2024, pp. 52745–52775, 2024.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, 2022.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5872–5881, 2018.

Lijun Zhang, Lin Li, Wei Wei, Huizhong Song, Yaodong Yang, and Jiye Liang. Scalable constrained policy optimization for safe multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 138698–138730, 2024.

Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32): 1–67, 2024.

Changdong Zhou, Ruofan Hu, Naiyao Wang, Ke Song, and Hongbo Liu. HIVE: A hypergraph-based game-theoretic interactive value decomposition engine for multi-lateral agents collaboration. *Neural Networks*, 203:109103, 2026.

Tianyu Zhu, Xinli Shi, Xiangping Xu, Jie Gui, and Jinde Cao. HyperComm: Hypergraph-based communication in multi-agent reinforcement learning. *Neural Networks*, 178:106432, 2024.

A Proofs of Theoretical Results

In this section, we present complete mathematical proofs and discussions. We first introduce some essential concepts, and then use HATRL (Zhong et al., 2024) as an example to illustrate the HygRL analytical framework.

A.1 Definitions of Basic Concepts

In this subsection we fix the notation for MDPs, MAMDPs and hypergraphs, define the occupancy measures on which the analysis rests, and express the value functions through them.

Given a measurable space $(\Omega, \mathcal{F})$, let $\mathcal{M}_+(\Omega)$ denote the set of positive measures and $\Delta(\Omega)$ the space of probability measures. Unless otherwise specified, the reference measure is taken to be Lebesgue.

Definition 2 (Markov Decision Process)**.** *A Markov Decision Process (MDP) is defined as a tuple*

$$G = \langle \mathcal{S}, \mathcal{A}, \mu, P, r, \gamma \rangle,$$

where $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces, $\mu \in \Delta(\mathcal{S})$ is the initial state distribution, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps each state to a distribution over actions.

Definition 3 (Multi-Agent Markov Decision Process)**.** *A Multi-Agent Markov Decision Process (MAMDP) is defined as a tuple*

$$G = \langle \mathcal{N}, \{\mathcal{S}_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mu, P, \{r_i\}_{i\in\mathcal{N}}, \gamma \rangle,$$

where $\mathcal{N} = \{1, \ldots, n\}$ is the agent set. For each agent i, $\mathcal{S}_i$ and $\mathcal{A}_i$ are state and action spaces. The joint state and action spaces are $\mathcal{S} = \prod_{i\in\mathcal{N}} \mathcal{S}_i$ and $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}_i$. The transition kernel is $P : \mathcal{S}\times\mathcal{A} \to \Delta(\mathcal{S})$, and $r_i : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is the local reward for agent i. A joint policy $\pi = \prod_{i\in\mathcal{N}} \pi_i$ factorizes over agents, where each $\pi_i : \mathcal{S}_i \to \Delta(\mathcal{A}_i)$.

We assume throughout that all state and action spaces are Polish. This standard regularity condition ensures the existence of regular conditional probabilities and supports the measure-theoretic arguments in our analysis. We adopt the cooperative setup, i.e. the global reward $r = \sum_{i\in\mathcal{N}} r_i$. With this setup, MAMDP can be rewritten as an MDP.

In the following definitions, δ denotes the Dirac distribution.

Definition 4. *Given an MDP $G = \langle \mathcal{S}, \mathcal{A}, \mu, P, r, \gamma \rangle$ and a policy π, the discounted state-action occupancy measure $\rho_{\mu,P,\pi}(\cdot|z_0)$ is defined as:*

$$\rho_{\mu,P,\pi}(\mathrm{d}z|z_0) = \sum_{t=0}^{\infty} \gamma^t \varrho^t_{\mu,P,\pi}(\mathrm{d}z|z_0),$$

where

$$\varrho^0_{\mu,P,\pi}(\mathrm{d}z|z_0) = \delta_{z_0},$$
$$\varrho^{t+1}_{\mu,P,\pi}(\mathrm{d}z|z_0) = \pi(\mathrm{d}a|s) \int_{z'\in\mathcal{Z}} P(\mathrm{d}s|z') \varrho_{\mu,P,\pi}(\mathrm{d}z'|z_0).$$

Besides, define $\rho_{\mu,P,\pi}(\cdot|s_0)$ and $\rho_{\mu,P,\pi}$ as:

$$\rho_{\mu,P,\pi}(\mathrm{d}z|s_0) = \int_{a'\in\mathcal{A}} \pi(\mathrm{d}a'|s_0) \rho_{\mu,P,\pi}(\mathrm{d}z|s_0, a'),$$

$$\rho_{\mu,P,\pi}(\mathrm{d}z) = \int_{s'\in\mathcal{S}} \mu(\mathrm{d}s') \rho_{\mu,P,\pi}(\mathrm{d}z|s').$$

Definition 5. *Given an MDP $G = \langle \mathcal{S}, \mathcal{A}, \mu, P, r, \gamma \rangle$ and a policy π, the discounted state occupancy measure $\eta_{\mu,P,\pi}(\cdot|s_0)$ is defined as:*

$$\eta_{\mu,P,\pi}(\mathrm{d}s|s_0) = \sum_{t=0}^{\infty} \gamma^t \zeta^t_{\mu,P,\pi}(\mathrm{d}s|s_0),$$

where

$$\zeta^0_{\mu,P,\pi}(\mathrm{d}s|s_0) = \delta_{s_0}, \tag{9}$$
$$\zeta^{t+1}_{\mu,P,\pi}(\mathrm{d}s|s_0) = \int_{(s',a)\in\mathcal{S}\times\mathcal{A}} P(\mathrm{d}s|s', a) \pi(\mathrm{d}a|s') \zeta_{\mu,P,\pi}(\mathrm{d}s'|s_0). \tag{10}$$

Besides, define $\eta_{\mu,P,\pi}$ as:

$$\eta_{\mu,P,\pi}(\mathrm{d}s) = \int_{s'\in\mathcal{S}} \mu(\mathrm{d}s') \eta_{\mu,P,\pi}(\mathrm{d}s|s').$$

With the definition of MDP and occupancy measures, we can rewrite the state-action value function Q_π, state value function V_π, advantage function A_π and expected discounted return J as follows:

Definition 6. *Given an MDP $G = \langle \mathcal{S}, \mathcal{A}, \mu, P, r, \gamma \rangle$ and a policy π, the value functions are defined as:*

$$Q_\pi(z) = \mathbb{E}_{s_{1:\infty}\sim P, a_{1:\infty}\sim\pi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a \right] = \int_{z'\in\mathcal{Z}} r(z') \rho_{\mu,P,\pi}(\mathrm{d}z'|z),$$

$$V_\pi(s) = \mathbb{E}_{s_{1:\infty}\sim P, a_{0:\infty}\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)|s_0=s\right] = \int_{z'\in\mathcal{Z}} r(z')\rho_{\mu,P,\pi}(\mathrm{d}z'|s),$$
$$A_\pi(z) = Q_\pi(z) - V_\pi(s),$$

and

$$J(\pi) = \mathbb{E}_{s0\sim\mu, s_{1:\infty}\sim P, a_{0:\infty}\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(z_t)\right] = \int_{z'\in\mathcal{Z}} r(z')\rho_{\mu,P,\pi}(\mathrm{d}z').$$

The advantage of the occupancy measure is that it decouples the value function from the MDP, thereby extending the definition of the value function and facilitating the handling of off-policy scenarios. With regard to HATRL, we adopt the definition from Zhong et al. (2024) and present the following value functions for several multi-agent scenarios.

Definition 7. *Given an MAMDP* $\langle\mathcal{N}, \{\mathcal{S}_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mu, P, \{r_i\}_{i\in\mathcal{N}}, \gamma\rangle$. *Let* $\mathcal{I}\subseteq\mathcal{N}$. *The corresponding multi-agent state-action value function is defined as:*

$$Q_\pi^{\mathcal{I}}(s, a_{\mathcal{I}}) = \mathbb{E}_{\mathbf{a}_{-\mathcal{I}}\sim\pi_{-\mathcal{I}}}\left[Q_\pi(s, a_{\mathcal{I}}, a_{-\mathcal{I}})\right],$$

Let $\mathcal{I},\mathcal{J}\subseteq\mathcal{N}$ *and* $\mathcal{I}\cap\mathcal{J}=\emptyset$. *The corresponding multi-agent advantage function is defined as:*

$$A_\pi^{\mathcal{I}}(s, a_{\mathcal{J}}, a_{\mathcal{I}}) = Q_\pi^{\mathcal{I}\cup\mathcal{J}}(s, a_{\mathcal{I}\cup\mathcal{J}}) - Q_\pi^{\mathcal{J}}(s, a_{\mathcal{J}}).$$

Let $\pi = \prod_{i=1}^n \pi_i$ *be a joint policy,* $\pi'_{\mathcal{I}} = \prod_{i\in\mathcal{I}}^n \pi'_i$ *be a joint policy over* $\mathcal{I}\subseteq\mathcal{N}$ *and* $\hat{\pi}_j$ *be some other policy for each* $j\in\mathcal{N}$. *Define*

$$L_\pi^{\mathcal{J}\cup\{i\}}\left(\pi'_{\mathcal{J}}, \hat{\pi}_i\right) = \mathbb{E}_{s\sim\rho_{\mu,P,\pi}, a_{\mathcal{J}}\sim\pi'_{\mathcal{J}}, a_i\sim\hat{\pi}_i}\left[A_\pi^{\{i\}}(s, a_{\mathcal{J}}, a_i)\right].$$

The key idea of our approach is the use of hypergraphs to model higher-order cooperative structures.

Definition 8 (Hypergraph)**.** *A hypergraph is a pair* $\mathcal{H} = (\mathcal{N}, \mathcal{E})$, *where* $\mathcal{N}$ *is the set of vertices (agents) and* $\mathcal{E}\subseteq 2^{\mathcal{N}}$ *is a set of hyperedges. Each hyperedge* $e\in\mathcal{E}$ *is an arbitrary non-empty subset of vertices.*

Combining the definitions of MAMDP and hypergraph, we obtain the following object.

Definition 9 (Hypergraphic MAMDP)**.** *A hypergraphic MAMDP (HygMAMDP) augments a MAMDP with a hypergraph* $\mathcal{H} = (\mathcal{N}, \mathcal{E})$ *over the agent set, so that the object of study is*

$$G = \langle\mathcal{H}, \{\mathcal{S}_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \mu, P, \{r_i\}_{i\in\mathcal{N}}, \gamma\rangle,$$

as introduced in Eq. equation 1. Each hyperedge $e\in\mathcal{E}$ *induces a local model on the coordinates of* e.

Besides, we define an additional auxiliary function $\mathcal{F}$ to simplify the expression. Assume $a\geq 0$, let

$$\mathcal{F}_a(\delta) = \begin{cases} 0 \text{ if } \delta = 0 \\ \left(1-\gamma^{\frac{a}{\delta}}\right)\delta \text{ if } 0<\delta \end{cases}.$$

Notice that $\lim_{\delta\to 0^+}\mathcal{F}_a(\delta) = 0$. Thus $\mathcal{F}_a$ is continuous on $[0,+\infty)$. Furthermore, $\lim_{\delta\to+\infty}\mathcal{F}_a(\delta) = -a\log\gamma$, and it is easy to verify that $\mathcal{F}_a$ is a monotonically increasing function; therefore, $\mathcal{F}_a\leq -a\log\gamma$ always holds. In particular, $\mathcal{F}_1$ is abbreviated as $\mathcal{F}_a$.

A.2 Format Occupancy Measure and Format Value Function

In this subsection, we first explain the rationale for defining the format occupancy measure and the format value function, and then provide their definitions.

Examining a hyperedge e, we observe that the local transitions assembled in Algorithm 1 are not drawn from a self-contained local multi-agent Markov decision process, but are instead projections of the global multi-agent Markov decision process onto the coordinate subspace of e. This marginalization yields the following occupancy measures and reward function.

$$\eta^e_{\mu,P,\bar{\pi}}(\mathrm{d}s_e) = \int_{s_{-e}\in\mathcal{S}_{-e}} \eta_{\mu,P,\bar{\pi}}(\mathrm{d}s_e, \mathrm{d}s_{-e}).$$

$$\rho^e_{\mu,P,\bar{\pi}}(\mathrm{d}z_e) = \int_{z_{-e}\in\mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}).$$

$$\rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e}) = \int_{z_e\in\mathcal{Z}_e} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}).$$

$$\rho^e_{\mu,P,\bar{\pi}}(\mathrm{d}z_e|z'_e) = (1-\gamma)\int_{z'_{-e}\in\mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z'_{-e}) \int_{z_{-e}\in\mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}|z'_e, z'_{-e}).$$

$$r^e_{\mu,P,\bar{\pi}}(z_e) = (1-\gamma)\int_{z_{-e}\in\mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e})\tilde{r}_e(z_e, z_{-e}).$$

$\eta^e_{\mu,P,\bar{\pi}}$, $\rho^e_{\mu,P,\bar{\pi}}$ and $\rho^e_{\mu,P,\bar{\pi}}(\cdot|z'_e)$ are not standard occupancy measures of any single Markov decision process, as they embed averaging over the agents outside e and over the aggregated behavior policy $\bar{\pi}$. To enable a theoretical analysis under this distributional mismatch, we introduce the following generalized notions of occupancy measures and value functions, which we term format occupancy measures and format value functions.

Definition 10 (Format Occupancy Measure)**.** *A format state-action occupancy measure with a starting point of $\mathcal{Z}$ is a function $\rho : \mathcal{Z} \to \mathcal{M}_+(\mathcal{Z})$ such that for each $z \in \mathcal{Z}$*

$$\int_{z'\in\mathcal{Z}} \rho(\mathrm{d}z'|z) = \frac{1}{1-\gamma}.$$

A format state-action occupancy measure of $\mathcal{Z}$ is a measure $\rho \in \mathcal{M}_+(\mathcal{Z})$ such that

$$\int_{z\in\mathcal{Z}} \rho(\mathrm{d}z) = \frac{1}{1-\gamma}.$$

A format state occupancy measure of $\mathcal{S}$ is a measure $\eta \in \mathcal{M}_+(\mathcal{S})$ such that

$$\int_{s\in\mathcal{S}} \eta(\mathrm{d}s) = \frac{1}{1-\gamma}.$$

We can see that $\eta^e_{\mu,P,\bar{\pi}}$, $\rho^e_{\mu,P,\bar{\pi}}$ and $\rho^e_{\mu,P,\bar{\pi}}(\cdot|z'_e)$ satisfy the definition of a format occupancy measure.

Definition 11 (Format Value Functions)**.** *Given a format state-action occupancy measure with a starting point $\rho(\cdot|z')$, a reward function $r(z)$, the format state-action value is defined as*

$$Q_{\rho,r}(z) = \int_{z'\in\mathcal{Z}} r(z')\rho(\mathrm{d}z'|z).$$

Given a finite set $\mathcal{N}$ and $\mathcal{I} \subseteq \mathcal{N}$. Let $\rho(\cdot|z')$ be a format state-action occupancy measure with a starting point of $\prod_{i\in\mathcal{N}} \mathcal{S}_i \times \prod_{i\in\mathcal{N}} \mathcal{A}_i$, η be a format state occupancy measure of $\prod_{i\in\mathcal{N}} \mathcal{S}_i$, $\pi = \prod_{i\in\mathcal{N}} \pi_i$ be a joint policy. The format multi-agent state-action value function is defined as

$$Q^{\mathcal{I}}_{\rho,r,\pi}(s, a_{\mathcal{I}}) = \mathbb{E}_{a_{-\mathcal{I}}\sim\pi_{-\mathcal{I}}} Q_{\rho,r}(s, a_{\mathcal{I}}, a_{-\mathcal{I}}).$$

Let $\mathcal{J} \subseteq \mathcal{N}$ and $\mathcal{I} \cap \mathcal{J} = \emptyset$. The format multi-agent advantage function is defined as

$$A^{\mathcal{I}}_{\rho,r,\pi}(s, a_{\mathcal{J}}, a_{\mathcal{I}}) = Q^{\mathcal{I}\cup\mathcal{J}}_{\rho,r,\pi}(s, a_{\mathcal{I}\cup\mathcal{J}}) - Q^{\mathcal{J}}_{\rho,r,\pi}(s, a_{\mathcal{J}}).$$

Let $\tilde{\pi}$ and $\hat{\pi}$ be two joint policies. The format surrogate target is defined as

$$L^{\mathcal{J}\cup\mathcal{I}}_{\eta,\rho,r,\pi}(\tilde{\pi}_{\mathcal{J}}, \hat{\pi}_{\mathcal{I}}) = \mathbb{E}_{s\sim\eta, a_{\mathcal{J}}\sim\tilde{\pi}_{\mathcal{J}}, a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}} \left[A^{\mathcal{I}}_{\rho,r,\pi}(s, a_{\mathcal{J}}, a_{\mathcal{I}})\right].$$

A key property of the format value functions defined above is that they generalize their standard counterparts. Whenever the format occupancy measure $\rho(\cdot|z')$ coincides with the occupancy measure $\rho_{\mu,P,\pi}(\cdot|z')$ of a concrete MDP, the format state-action value function $Q_{\rho,r}(z)$ reduces to the standard $Q_\pi(z)$; likewise, the format multi-agent value functions reduce to the standard multi-agent value functions of Definition 6 and 7. This ensures that our analysis subsumes the standard setting as a special case.

A.3 TECHNICAL LEMMAS

To illustrate the following proofs, we first state several lemmas and propositions.

Lemma 1. *Given $\gamma \in [0, 1)$ and $a, \delta \geq 0$, the following two inequalities hold.*

$$\sum_{t=0}^{\infty} \gamma^t \min(t\delta, a) \leq \frac{\gamma}{(1-\gamma)^2} \mathcal{F}_a(\delta), \tag{11}$$

$$\sum_{t=0}^{\infty} \gamma^t \min((t+1)\delta, a) \leq \frac{1}{(1-\gamma)^2} \mathcal{F}_a(\delta). \tag{12}$$

Proof. Firstly, we prove Ineq. (11).

If $\delta = 0$, then Ineq. (11) is trivial.

When $0 < \delta$, let $T = \left\lfloor \frac{a}{\delta} \right\rfloor$.

$$\begin{aligned}
\sum_{t=0}^{\infty} \gamma^t \min(t\delta, a) &= \delta \sum_{t=0}^{T} t\gamma^t + a \sum_{t=T+1}^{\infty} \gamma^t \\
&= \frac{\gamma}{(1-\gamma)^2} \left(1 - \gamma^T (T+1) + \gamma^{T+1} T\right) \delta + \frac{a}{1-\gamma} \gamma^{T+1}
\end{aligned}$$

Let $f(x) = \gamma^{x+1} x - \gamma^x (x+1)$. When $x \geq 0$, we have

$$\begin{aligned}
f'(x) &= \gamma^{x+1} + x\gamma^{x+1} \log \gamma - \gamma^x - (x+1)\gamma^x \log \gamma \\
&= \gamma^x \left(\gamma + x\gamma \log \gamma - 1 - (x+1) \log \gamma\right) \\
&= \gamma^x \left(\left((\gamma - 1) \log \gamma\right) x + (\gamma - 1 - \log \gamma)\right).
\end{aligned}$$

$$(\gamma - 1) \log \gamma \geq 0,$$

$$\gamma - 1 - \log \gamma \geq \gamma - 1 - \gamma + 1 = 0.$$

Thus, when $x \geq 0$,

$$\left((\gamma - 1) \log \gamma\right) x + (\gamma - 1 - \log \gamma) \geq 0,$$

and

$$f'(x) \geq 0.$$

Hence,

$$f(T) \leq f\left(\frac{a}{\delta}\right).$$

Besides,

$$\gamma^T \leq \gamma^{\frac{a}{\delta}}.$$

Therefore,

$$\begin{aligned}
\sum_{t=0}^{\infty} \gamma^t \min(t\delta, a) &\leq \frac{\gamma}{(1-\gamma)^2} \left(1 - \gamma^{\frac{a}{\delta}} \left(\frac{a}{\delta} + 1\right) + \gamma^{\frac{a}{\delta}+1} \frac{a}{\delta}\right) \delta + \frac{a}{1-\gamma} \gamma^{\frac{a}{\delta}+1} \\
&= \frac{\gamma}{(1-\gamma)^2} \mathcal{F}_a(\delta).
\end{aligned}$$

Now, we prove Ineq. (12). If $\gamma = 0$, then

$$\sum_{t=0}^{\infty} \gamma^t \min((t+1)\delta, a) = \min(\delta, a) \leq \delta = \frac{1}{(1-\gamma)^2} \mathcal{F}_a(\delta).$$

When $0 < \gamma \leq 1$,

$$\begin{aligned}
\sum_{t=0}^{\infty} \gamma^t \min((t+1)\delta, a) &= \frac{1}{\gamma} \sum_{t=0}^{\infty} \gamma^{t+1} \min((t+1)\delta, a) \\
&= \frac{1}{\gamma} \sum_{t=0}^{\infty} \gamma^t \min(t\delta, a) - \gamma^0 \min(0, a) \\
&\leq \frac{1}{(1-\gamma)^2} \mathcal{F}_a(\delta).
\end{aligned}$$

□

Corollary 3. *Given $\gamma \in [0,1)$ and $0 \le \epsilon \le 1$, the following inequality holds.*

$$\sum_{t=0}^{\infty} \gamma^t \min(t\delta + (t+1)\epsilon, 1) \le \frac{1}{(1-\gamma)^2} \mathcal{F}_{1+\delta}(\delta + \epsilon) - \frac{\delta}{1-\gamma}.$$

Proof. Notice that

$$\min(t\delta + (t+1)\epsilon, 1) = \min((t+1)\delta + (t+1)\epsilon, 1+\delta) - \delta,$$

we have

$$\begin{aligned}\sum_{t=0}^{\infty} \gamma^t \min(t\delta + (t+1)\epsilon, 1) &= \sum_{t=0}^{\infty} \gamma^t \min((t+1)\delta + (t+1)\epsilon, 1+\delta) - \frac{\delta}{1-\gamma} \\ &\overset{\text{Lemma 1}}{\le} \frac{1}{(1-\gamma)^2} \mathcal{F}_{1+\delta}(\delta + \epsilon) - \frac{\delta}{1-\gamma}.\end{aligned}$$

□

Given two measures μ and ν, let $(\mu \otimes \nu)(A, B) = \mu(A)\nu(B)$. Given a kernel K, let $(\mu \otimes K)(\mathrm{d}x, \mathrm{d}y) = \mu(\mathrm{d}x)K(\mathrm{d}y|x)$, $(\mu K)(A) = \int_{x \in X} K(A|x)\mu(\mathrm{d}x)$. The following properties regarding total variation are well known.

Proposition 1. *Let $(\Omega, \mathcal{F})$ be a measurable space, given two measure μ, ν and a bounded function $f : \Omega \to \mathbb{R}$, we have*

$$\left| \int_{x \in \Omega} f(x)\mu(\mathrm{d}x) - \int_{x \in \Omega} f(x)\nu(\mathrm{d}x) \right| \le 2\|\mu - \nu\|_{\mathrm{TV}} \sup_{x \in \Omega} |f(x)|.$$

Proposition 1 follows from Proposition 4.5 of (Levin & Peres, 2017).

Proposition 2. *Suppose for each $i \in \{1, \dots, n\}$, $(\Omega_i, \mathcal{F}_i)$ be a measurable space, $p_i, q_i \in \mathcal{M}_+(\Omega_i)$. Then in the product space $(\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n)$, the product distributions $P = \otimes_{i=1}^n p_i$ and $Q = \otimes_{i=1}^n q_i$ satisfy*

$$\|P - Q\|_{\mathrm{TV}} \le \sum_{i=1}^{n} \|p_i - q_i\|_{\mathrm{TV}}.$$

Proposition 2 follows from Lemma 3.6.5 of (Durrett, 2019).

Proposition 3. *Let $(\Omega, \mathcal{F})$ be a measurable space, measure sequence $\{\mu_t\}_{t=0}^{\infty}$ and $\{\nu_t\}_{t=0}^{\infty}$, and factor $\gamma \in [0,1)$. If total mass of $\{\mu_t\}_{t=0}^{\infty}$ and $\{\nu_t\}_{t=0}^{\infty}$ are bounded, then*

$$\left\| \sum_{t=0}^{\infty} \gamma^t \mu_t - \sum_{t=0}^{\infty} \gamma^t \nu_t \right\|_{\mathrm{TV}} \le \sum_{t=0}^{\infty} \gamma^t \|\mu_t - \nu_t\|_{\mathrm{TV}}.$$

Proof.

$$\left\| \sum_{t=0}^{\infty} \gamma^t \mu_t - \sum_{t=0}^{\infty} \gamma^t \nu_t \right\|_{\mathrm{TV}} = \left\| \sum_{t=0}^{\infty} \gamma^t (\mu_t - \nu_t) \right\|_{\mathrm{TV}} \le \sum_{t=0}^{\infty} \gamma^t \|\mu_t - \nu_t\|_{\mathrm{TV}}.$$

□

Proposition 4. *Let $(\Omega, \mathcal{F})$ be a measurable space, $\eta \in \mathcal{M}_+(\Omega)$. Let $(X, \mathcal{B})$ be another measurable space, for each $\omega \in \Omega$, $\mu_\omega, \nu_\omega \in \mathcal{M}_+(X)$, and both $\omega \mapsto \mu_\omega$ and $\omega \mapsto \nu_\omega$ are measurable. Then,*

$$\left\| \int_{\omega \in \Omega} \mu_\omega \eta(\mathrm{d}\omega) - \int_{\omega \in \Omega} \nu_\omega \eta(\mathrm{d}\omega) \right\|_{\mathrm{TV}} \le \int_{\omega \in \Omega} \|\mu_\omega - \nu_\omega\|_{\mathrm{TV}} \eta(\mathrm{d}\omega).$$

Proof. For any measurable set $A \in \mathcal{B}$, we have

$$\begin{aligned}\left| \int_{\Omega} \mu_\omega(A)\, \eta(\mathrm{d}\omega) - \int_{\Omega} \nu_\omega(A)\, \eta(\mathrm{d}\omega) \right| &\le \int_{\Omega} |\mu_\omega(A) - \nu_\omega(A)|\, \eta(\mathrm{d}\omega) \\ &\le \int_{\Omega} \|\mu_\omega - \nu_\omega\|_{\mathrm{TV}}\, \eta(\mathrm{d}\omega).\end{aligned}$$

Taking the supremum over all measurable $A \in \mathcal{B}$ yields the desired inequality. □

Proposition 5. *Let $(X, \mathcal{F}_X)$, $(Y, \mathcal{F}_Y)$ be two measurable spaces, $\mu, \nu \in \mathcal{M}_+(X)$. For any Markov kernel K from $(X, \mathcal{F}_X)$ to $(Y, \mathcal{F}_Y)$, we have*

$$\|\mu K - \nu K\|_{\mathrm{TV}} \leq \|\mu - \nu\|_{\mathrm{TV}}.$$

Proof. By the definition of the total variation distance,

$$\begin{aligned}
\|\mu K - \nu K\|_{\mathrm{TV}} &= \sup_{A \in \mathcal{F}} \left| \int_X K(A|x)\, \mu(\mathrm{d}x) - \int_X K(A|x)\, \nu(\mathrm{d}x) \right| \\
&= \sup_{A \in \mathcal{F}} \left| \int_X K(A|x)\, (\mu - \nu)(\mathrm{d}x) \right| \\
&\leq \sup_{A \in \mathcal{F}} \int_X K(A|x)\, |(\mu - \nu)|(\mathrm{d}x) \\
&\leq \|\mu - \nu\|_{\mathrm{TV}}.
\end{aligned}$$

The last inequality follows since $0 \leq K(A|x) \leq 1$ for all $x \in X$. □

Proposition 6. *Let $(X, \mathcal{F}_X)$, $(Y, \mathcal{F}_Y)$ be two measurable spaces, $\mu, \nu \in \mathcal{M}_+(X)$. For any Markov kernel K from $(X, \mathcal{F}_X)$ to $(Y, \mathcal{F}_Y)$, we have*

$$\|\mu \otimes K - \nu \otimes K\|_{\mathrm{TV}} = \|\mu - \nu\|_{\mathrm{TV}}.$$

Proof. Let $\lambda = \mu + \nu \in \mathcal{M}_+(X)$. Then μ and ν are absolutely continuous with respect to λ. Let $p \coloneqq \mathrm{d}\mu/\mathrm{d}\lambda$ and $q \coloneqq \mathrm{d}\nu/\mathrm{d}\lambda$ be their Radon-Nikodym derivatives.

Consider the product measure $\lambda \otimes K$ on $X \times Y$. Since $\mu \ll \lambda$ and $\nu \ll \lambda$, we have $\mu \otimes K \ll \lambda \otimes K$ and $\nu \otimes K \ll \lambda \otimes K$, with densities

$$\frac{\mathrm{d}(\mu \otimes K)}{\mathrm{d}(\lambda \otimes K)}(x, y) = p(x), \qquad \frac{\mathrm{d}(\nu \otimes K)}{\mathrm{d}(\lambda \otimes K)}(x, y) = q(x),$$

where the densities are independent of y because K is the same kernel in both measures.

Therefore,

$$\begin{aligned}
\|\mu \otimes K - \nu \otimes K\|_{\mathrm{TV}} &= \frac{1}{2} \int_{X \times Y} |p(x) - q(x)|\, (\lambda \otimes K)(\mathrm{d}x, \mathrm{d}y) \\
&= \frac{1}{2} \int_X |p(x) - q(x)| \left(\int_Y K(\mathrm{d}y|x) \right) \lambda(\mathrm{d}x) \\
&= \frac{1}{2} \int_X |p(x) - q(x)|\, \lambda(\mathrm{d}x) \\
&= \|\mu - \nu\|_{\mathrm{TV}},
\end{aligned}$$

where the last equality follows from the definition of total variation via the L^1 norm of the density with respect to λ. □

Proposition 7. *Let $(X, \mathcal{F}_X)$, $(Y, \mathcal{F}_Y)$ be two measurable spaces, $\mu \in \mathcal{M}_+(X)$, for any Markov kernel K and L from $(X, \mathcal{F}_X)$ to $(Y, \mathcal{F}_Y)$,*

$$\|\mu \otimes K - \mu \otimes L\|_{\mathrm{TV}} = \int_{x \in X} \|K(\cdot|x) - L(\cdot|x)\|_{\mathrm{TV}} \mu(\mathrm{d}x).$$

Proof. For each fixed $x \in X$, let $\lambda_x \coloneqq K(\cdot|x) + L(\cdot|x) \in \mathcal{M}_+(Y)$. Since both $K(\cdot|x)$ and $L(\cdot|x)$ are probability measures, they are absolutely continuous with respect to λ_x. Define their Radon-Nikodym derivatives

$$p_x(y) \coloneqq \frac{\mathrm{d}K(\cdot|x)}{\mathrm{d}\lambda_x}(y), \qquad q_x(y) \coloneqq \frac{\mathrm{d}L(\cdot|x)}{\mathrm{d}\lambda_x}(y).$$

Then by the definition of total variation distance,

$$\|K(\cdot|x) - L(\cdot|x)\|_{\mathrm{TV}} = \frac{1}{2} \int_Y |p_x(y) - q_x(y)|\, \lambda_x(\mathrm{d}y).$$

Both $\mu \otimes K$ and $\mu \otimes L$ are absolutely continuous with respect to $\mu \otimes \lambda_x$, with densities

$$\frac{\mathrm{d}(\mu \otimes K)}{\mathrm{d}(\mu \otimes \lambda_x)}(x, y) = p_x(y), \qquad \frac{\mathrm{d}(\mu \otimes L)}{\mathrm{d}(\mu \otimes \lambda_x)}(x, y) = q_x(y).$$

Therefore, applying the total variation formula for signed measures with density,

$$\begin{aligned}
\|\mu \otimes K - \mu \otimes L\|_{\mathrm{TV}} &= \frac{1}{2} \int_{X \times Y} |p_x(y) - q_x(y)| (\mu \otimes \lambda_x)(\mathrm{d}x, \mathrm{d}y) \\
&= \frac{1}{2} \int_X \left(\int_Y |p_x(y) - q_x(y)| \, \lambda_x(\mathrm{d}y) \right) \mu(\mathrm{d}x) \\
&= \int_X \|K(\cdot|x) - L(\cdot|x)\|_{\mathrm{TV}} \, \mu(\mathrm{d}x),
\end{aligned}$$

which proves the proposition. □

Proposition 8. *Given n policies $\pi_1, \ldots, \pi_n$ of a Markov Decision Process, let $\bar{\pi} = \frac{1}{n} \sum_{i=1}^n \pi_i$. Then, for each i, we have*

$$\mathrm{D}_{\mathrm{TV}}^{\max}(\pi_i, \bar{\pi}) \leq \frac{n-1}{n},$$

and

$$\mathrm{D}_{\mathrm{KL}}^{\max}(\pi, \bar{\pi}) \leq \frac{n \log n}{n-1} \mathrm{D}_{\mathrm{TV}}^{\max}(\pi, \bar{\pi}).$$

Proposition 8 follows from (Binette, 2019).

A.4 Offline Sample Update

As noted in the previous subsection, for a hyperedge e, the local sample distribution is not obtained by sampling from a specific MDP via π_e; therefore, we first characterize the performance of HATRL when using offline samples through the following proposition.

Proposition 9. *Given a multi-agent Markov decision process $\langle \mathcal{N}, \{\mathcal{S}_i\}_{i \in \mathcal{N}}, \{\mathcal{A}_i\}_{i \in \mathcal{N}}, \mu, P, r \rangle$, let $\mathcal{N} = \{1, \ldots, n\}$ and π, π' and $\tilde{\pi}$ be three joint policies.*

Assume there are a format state-action occupancy measure with starting point $\rho'(\cdot|z')$, a reward function $r'(z)$, such that r and r' are bounded by R and for each $z \in \mathcal{Z}$

$$\|\rho_{\mu,P,\pi}(\cdot|z) - \rho'(\cdot|z)\|_{\mathrm{TV}} \leq \epsilon_\rho,$$

$$|r(z) - r'(z)| \leq \epsilon_r R,$$

respectively. Assume there is a format state occupancy measure η', such that

$$\|\eta_{\mu,P,\pi} - \eta'\|_{\mathrm{TV}} \leq \epsilon_\eta.$$

Besides, assume π and $\tilde{\pi}$ satisfy

$$\forall i \in \mathcal{N} \exists \delta_i \geq 0 \left(\mathrm{D}_{\mathrm{TV}}^{\max}(\pi_i, \tilde{\pi}_i) \leq \delta_i \right).$$

Given an update order $\langle i_1, i_2, \ldots, i_n \rangle$ of $\mathcal{N}$, the performance difference between π and π' can be bounded as:

$$\begin{aligned}
J(\pi'|\mu, P, r) \geq J(\pi|\mu, P, r) &+ \sum_{m=1}^n \left[L_{\eta', \rho', r', \tilde{\pi}}^{i_{1:m}} \left(\pi'_{i_{1:m-1}}, \pi'_{i_m} \right) - C \mathrm{D}_{\mathrm{KL}}^{\max} \left(\pi_{i_m}, \pi'_{i_m} \right) \right] \\
&- \left[\frac{2n\epsilon_r}{(1-\gamma)^2} + \frac{4n}{1-\gamma} (\epsilon_\eta + \epsilon_\rho) + \frac{2}{(1-\gamma)^2} \sum_{j=1}^n (2j-1) \delta_{i_j} \right] R,
\end{aligned}$$

where

$$C = \frac{8\gamma}{(1-\gamma)^3} R.$$

To prove the proposition, we introduce the following lemmas.

Lemma 2. *Assume $\mathcal{S}$ is a state space and $\mathcal{A}$ is an action space. Consider format state-action occupancy measures with starting point $\rho(\cdot|z)$ and $\rho'(\cdot|z)$ and two rewards $r(z)$ and $r'(z)$ such that r and r' are bounded by R and for each $z \in \mathcal{Z}$,*

$$\|\rho(\cdot|z) - \rho'(\cdot|z)\|_{\mathrm{TV}} \leq \epsilon_\rho,$$
$$|r(z) - r'(z)| \leq \epsilon_r R,$$

respectively, then for each state-action pair z, we have

$$|Q_{\rho,r}(z) - Q_{\rho',r'}(z)| \leq \left(2\epsilon_\rho + \frac{\epsilon_r}{1-\gamma}\right) R.$$

Proof. According to the Definition 11, we have

$$\begin{aligned}
& |Q_{\rho,r}(z) - Q_{\rho',r'}(z)| \\
= \; & \left| \int_{z' \in \mathcal{Z}} \rho(\mathrm{d}z'|z) r(z') - \int_{z' \in \mathcal{Z}} \rho'(\mathrm{d}z'|z) r'(z') \right| \\
\leq \; & \left| \int_{z' \in \mathcal{Z}} \rho(\mathrm{d}z'|z) r(z') - \int_{z' \in \mathcal{Z}} \rho'(\mathrm{d}z'|z) r(z') \right| \\
& + \left| \int_{z' \in \mathcal{Z}} \rho'(\mathrm{d}z'|z) r(z') - \int_{z' \in \mathcal{Z}} \rho'(\mathrm{d}z'|z) r'(z') \right| \\
\overset{\text{Proposition 1}}{\leq} \; & 2R\|\rho(\cdot|z) - \rho'(\cdot|z)\|_{\mathrm{TV}} + \left| \int_{z' \in \mathcal{Z}} (r(z') - r'(z')) \rho'(\mathrm{d}z'|z) \right| \\
\leq \; & 2R\epsilon_\rho + \int_{z' \in \mathcal{Z}} |r(z') - r'(z')| \, \rho'(\mathrm{d}z'|z) \\
\leq \; & 2R\epsilon_\rho + \epsilon_r R \int_{z' \in \mathcal{Z}} \rho'(\mathrm{d}z'|z) \\
\leq \; & 2\epsilon_\rho R + \frac{\epsilon_r R}{1-\gamma},
\end{aligned}$$

which completes the proof. □

Lemma 3. *Assume $\mathcal{N}$ is a finite set, and for each $i \in \mathcal{N}$, $\mathcal{S}_i$ is a state space and $\mathcal{A}_i$ is an action space. In $\prod_{i\in\mathcal{N}} \mathcal{S}_i \times \prod_{i\in\mathcal{N}} \mathcal{A}_i$, consider two format state-action occupancy measures with starting point $\rho(\cdot|z)$ and $\rho'(\cdot|z)$ and two rewards $r(z)$ and $r'(z)$ such that r and r' are bounded by R and for each $z \in \mathcal{Z}$,*

$$\|\rho(\cdot|z) - \rho'(\cdot|z)\|_{\mathrm{TV}} \leq \epsilon_\rho,$$
$$|r(z) - r'(z)| \leq \epsilon_r R,$$

respectively.

Let $\pi = \prod_{i\in\mathcal{N}} \pi_i, \pi' = \prod_{i\in\mathcal{N}} \pi'_i$ be two policies and $\mathcal{I} \subseteq \mathcal{N}$, such that

$$\forall i \in \mathcal{I} \exists \delta_i \in \mathbb{R} \left(\mathrm{D}_{\mathrm{TV}}^{\max}(\pi_i, \pi'_i) \leq \delta_i \right).$$

Then

$$\left| Q^{\mathcal{I}}_{\rho,r,\pi}(s, a_{\mathcal{I}}) - Q^{\mathcal{I}}_{\rho',r',\pi'}(s, a_{\mathcal{I}}) \right| \leq \left(2\epsilon_\rho + \frac{2}{1-\gamma} \sum_{i \in -\mathcal{I}} \delta_i + \frac{\epsilon_r}{1-\gamma}\right) R.$$

Proof. Proposition 2 shows

$$\mathrm{D}_{\mathrm{TV}}^{\max}(\pi, \pi') \leq \sum_{i\in\mathcal{N}} \mathrm{D}_{\mathrm{TV}}^{\max}(\pi_i, \pi'_i) \leq \sum_{i\in\mathcal{N}} \delta_i.$$

Notice that $Q_{\rho,r}(z)$ is bounded, when r is bound by R i.e.,

$$|Q_{\rho,r}(z)| \leq R \int_{z' \in \mathcal{Z}} \rho(\mathrm{d}z'|z) = \frac{1}{1-\gamma} R.$$

Therefore $\forall \mathcal{I} \subseteq \mathcal{N}$, we have $Q^{\mathcal{I}}_{\rho,r,\pi}(s,a) \leq \frac{1}{1-\gamma} R$.

When $\mathcal{I} = \mathcal{N}$, we have

$$\begin{aligned}\left|Q^{\mathcal{N}}_{\rho,r,\pi}(s,a_{\mathcal{N}}) - Q^{\mathcal{N}}_{\rho',r',\pi'}(s,a_{\mathcal{N}})\right| &= \left|Q_{\rho,r}(s,a) - Q_{\rho',r'}(s,a)\right| \\ &\overset{\text{Lemma 2}}{\leq} \left(2\epsilon_\rho + \frac{\epsilon_r}{1-\gamma}\right) R. \end{aligned} \tag{13}$$

When $\mathcal{I} \neq \mathcal{N}$. Let $i \in -\mathcal{I}$, $\mathcal{J} = \mathcal{I} \cup \{i\}$. Suppose Ineq. (3) holds for $\mathcal{J}$. According to the definition of $Q^{\mathcal{I}}_{\rho,r,\pi}$, we have

$$\begin{aligned}
& \left|Q^{\mathcal{I}}_{\rho,r,\pi}(s,a_{\mathcal{I}}) - Q^{\mathcal{I}}_{\rho',r',\pi'}(s,a_{\mathcal{I}})\right| \\
= \; & \left|\int_{a_i \in \mathcal{A}_i} \pi_i(\mathrm{d}a_i|s_i)\, Q^{\mathcal{J}}_{\rho,r,\pi}(s,a_{\mathcal{J}}) - \int_{a_i \in \mathcal{A}_i} \pi'_i(\mathrm{d}a_i|s_i)\, Q^{\mathcal{J}}_{\rho',r',\pi'}(s,a_{\mathcal{J}})\right| \\
\leq \; & \left|\int_{a_i \in \mathcal{A}_i} \pi_i(\mathrm{d}a_i|s_i)\, Q^{\mathcal{J}}_{\rho,r,\pi}(s,a_{\mathcal{J}}) - \int_{a_i \in \mathcal{A}_i} \pi'_i(\mathrm{d}a_i|s_i)\, Q^{\mathcal{J}}_{\rho,r,\pi}(s,a_{\mathcal{J}})\right| \\
& + \left|\int_{a_i \in \mathcal{A}_i} \pi'_i(\mathrm{d}a_i|s_i)\, Q^{\mathcal{J}}_{\rho,r,\pi}(s,a_{\mathcal{J}}) - \int_{a_i \in \mathcal{A}_i} \pi'_i(\mathrm{d}a_i|s_i)\, Q^{\mathcal{J}}_{\rho',r',\pi'}(s,a_{\mathcal{J}})\right| \\
\overset{\text{Proposition 1}}{\leq} \; & \frac{2R}{1-\gamma}\|\pi_i(\cdot|s_i) - \pi'_i(\cdot|s_i)\|_{\mathrm{TV}} + \int_{a_i \in \mathcal{A}_i} \pi'_i(\mathrm{d}a_i|s_i) \left|Q^{\mathcal{J}}_{\rho,r,\pi}(s,a_{\mathcal{J}}) - Q^{\mathcal{J}}_{\rho',r',\pi'}(s,a_{\mathcal{J}})\right| \\
\overset{\text{I.H.}}{\leq} \; & \frac{2R}{1-\gamma}\delta_i + \left(2\epsilon_\rho + \frac{2}{1-\gamma}\sum_{i \in -\mathcal{J}} \delta_i + \frac{\epsilon_r}{1-\gamma}\right) R \int_{a_i \in \mathcal{A}_i} \pi'_i(\mathrm{d}a_i|s_i) \\
= \; & \left(2\epsilon_\rho + \frac{2}{1-\gamma}\sum_{i \in -\mathcal{I}} \delta_i + \frac{\epsilon_r}{1-\gamma}\right) R.
\end{aligned} \tag{14}$$

With Ineq. (13) and (14), applying mathematical induction, we can conclude that Ineq. (3) holds. □

With Lemma 3, we have the following corollary.

Corollary 4. *With the same condition as Lemma 3, assume $\mathcal{I}, \mathcal{J} \subseteq \mathcal{N}$, $\mathcal{I} \cap \mathcal{J} = \emptyset$, we have*

$$\begin{aligned}
& \left|A^{\mathcal{I}}_{\rho,r,\pi}(s,a_{\mathcal{J}},a_{\mathcal{I}}) - A^{\mathcal{I}}_{\rho',r',\pi'}(s,a_{\mathcal{J}},a_{\mathcal{I}})\right| \\
\leq & \left(4\epsilon_\rho + \frac{4}{1-\gamma}\sum_{i \in -\mathcal{I} \cap -\mathcal{J}} \delta_i + \frac{2}{1-\gamma}\sum_{i \in \mathcal{I}} \delta_i + \frac{2\epsilon_r}{1-\gamma}\right) R.
\end{aligned}$$

Proof.

$$\begin{aligned}
& \left|A^{\mathcal{I}}_{\rho,r,\pi}(s,a_{\mathcal{J}},a_{\mathcal{I}}) - A^{\mathcal{I}}_{\rho',r',\pi'}(s,a_{\mathcal{J}},a_{\mathcal{I}})\right| \\
= \; & \left|Q^{\mathcal{I}\cup\mathcal{J}}_{\rho,r,\pi}(s,a_{\mathcal{I}\cup\mathcal{J}}) - Q^{\mathcal{J}}_{\rho,r,\pi}(s,a_{\mathcal{J}}) - Q^{\mathcal{I}\cup\mathcal{J}}_{\rho',r',\pi'}(s,a_{\mathcal{I}\cup\mathcal{J}}) + Q^{\mathcal{J}}_{\rho',r',\pi'}(s,a_{\mathcal{J}})\right| \\
\leq \; & \left|Q^{\mathcal{I}\cup\mathcal{J}}_{\rho,r,\pi}(s,a_{\mathcal{I}\cup\mathcal{J}}) - Q^{\mathcal{I}\cup\mathcal{J}}_{\rho',r',\pi'}(s,a_{\mathcal{I}\cup\mathcal{J}})\right| + \left|Q^{\mathcal{J}}_{\rho,r,\pi}(s,a_{\mathcal{J}}) - Q^{\mathcal{J}}_{\rho',r',\pi'}(s,a_{\mathcal{J}})\right| \\
\overset{\text{Lemma 3}}{\leq} \; & \left(2\epsilon_\rho + \frac{2}{1-\gamma}\sum_{i \in -\mathcal{I} \cap -\mathcal{J}} \delta_i + \frac{\epsilon_r}{1-\gamma}\right) R + \left(2\epsilon_\rho + \frac{2}{1-\gamma}\sum_{i \in -\mathcal{J}} \delta_i + \frac{\epsilon_r}{1-\gamma}\right) R \\
= \; & \left(4\epsilon_\rho + \frac{4}{1-\gamma}\sum_{i \in -\mathcal{I} \cap -\mathcal{J}} \delta_i + \frac{2}{1-\gamma}\sum_{i \in \mathcal{I}} \delta_i + \frac{2\epsilon_r}{1-\gamma}\right) R.
\end{aligned}$$

□

The following lemma is Lemma 6 from (Zhong et al., 2024).

Lemma 4. *Given a multi-agent Markov decision process $\langle \mathcal{N}, \{\mathcal{S}\}_{i\in\mathcal{N}}, \{\mathcal{A}\}_{i\in\mathcal{N}}, \mu, P, r\rangle$, let π be a joint policy. Then, for any joint policy $\bar{\pi}$, we have*

$$J(\bar{\pi}) \geq J(\pi) + \sum_{m=1}^{n} \left[L_{\pi}^{i_{1:m}} \left(\bar{\pi}_{i_{1:m-1}}, \bar{\pi}_{i_m} \right) - C \mathrm{D}_{\mathrm{KL}}^{\max} \left(\pi_{i_m}, \bar{\pi}_{i_m} \right) \right],$$

where

$$C = \frac{4\gamma \max_{s,a} \left| A_{\pi}(s,a) \right|}{(1-\gamma)^2}.$$

Lemma 5. *Assume $\mathcal{N}$ is a finite set, and for each $i \in \mathcal{N}$, $\mathcal{S}_i$ is a state space and $\mathcal{A}_i$ is an action space. For $\prod_{i\in\mathcal{N}} \mathcal{S}_i \times \prod_{i\in\mathcal{N}} \mathcal{A}_i$, consider format state-action occupancy measures with starting point $\rho(\cdot|z)$ and $\rho'(\cdot|z)$ and two rewards $r(z)$ and $r'(z)$ such that r and r' are bounded by R and for each $z \in \mathcal{Z}$,*

$$\left\| \rho(\cdot|z) - \rho'(\cdot|z) \right\|_{\mathrm{TV}} \leq \epsilon_{\rho},$$

$$\left| r(z) - r'(z) \right| \leq \epsilon_r R,$$

respectively. For $\prod_{i\in\mathcal{N}} \mathcal{S}_i$, consider two format state occupancy measures η and η', such that

$$\left\| \eta - \eta' \right\|_{\mathrm{TV}} \leq \epsilon_{\eta}.$$

Let $\pi = \prod_{i\in\mathcal{N}} \pi_i, \pi' = \prod_{i\in\mathcal{N}} \pi_i'$ be two policies such that

$$\forall i \in \mathcal{N} \exists \delta_i \in \mathbb{R} \left(\mathrm{D}_{\mathrm{TV}}^{\max} \left(\pi_i, \pi_i' \right) \leq \delta_i \right),$$

and $\bar{\pi}_i, \hat{\pi}_i$ are another policies of agent i.

Besides, let $\mathcal{I}, \mathcal{J} \subset \mathcal{N}$ and $\mathcal{I} \cap \mathcal{J} = \emptyset$, then we have

$$\begin{aligned}
&\left| L_{\eta,\rho,r,\pi}^{\mathcal{J}\cup\mathcal{I}} \left(\bar{\pi}_{\mathcal{J}}, \hat{\pi}_{\mathcal{I}} \right) - L_{\eta',\rho',r',\pi'}^{\mathcal{J}\cup\mathcal{I}} \left(\bar{\pi}_{\mathcal{J}}, \hat{\pi}_{\mathcal{I}} \right) \right| \\
&\leq \left(\frac{4}{1-\gamma} (\epsilon_{\eta} + \epsilon_{\rho}) + \frac{4}{(1-\gamma)^2} \sum_{i \in -\mathcal{I}\cap-\mathcal{J}} \delta_i + \frac{2}{(1-\gamma)^2} \sum_{i\in\mathcal{I}} \delta_i + \frac{2\epsilon_r}{(1-\gamma)^2} \right) R.
\end{aligned}$$

Proof. Recall the proof of Lemma 3, notice that $Q_{\rho,r,\pi}^{\mathcal{I}}$ is bounded by $\frac{1}{1-\gamma} R$, when r is bound by R, then we have

$$\begin{aligned}
\left| A_{\rho,r,\pi}^{\mathcal{I}} \left(s, a_{\mathcal{J}}, a_{\mathcal{I}} \right) \right| &= \left| Q_{\rho,r,\pi}^{\mathcal{I}\cup\mathcal{J}} \left(s, a_{\mathcal{I}\cup\mathcal{J}} \right) - Q_{\rho,r,\pi}^{\mathcal{J}} \left(s, a_{\mathcal{J}} \right) \right| \\
&\leq \left| Q_{\rho,r,\pi}^{\mathcal{I}\cup\mathcal{J}} \left(s, a_{\mathcal{I}\cup\mathcal{J}} \right) \right| + \left| Q_{\rho,r,\pi}^{\mathcal{J}} \left(s, a_{\mathcal{J}} \right) \right| \\
&\leq \frac{2}{1-\gamma} R.
\end{aligned}$$

Besides, we have

$$\begin{aligned}
& \left| \mathbb{E}_{a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}}, a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}} \left[A_{\rho,r,\pi}^{\mathcal{I}} \left(s, a_{\mathcal{J}}, a_{\mathcal{I}} \right) \right] - \mathbb{E}_{a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}}, a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}} \left[A_{\rho',r',\pi'}^{\mathcal{I}} \left(s, a_{\mathcal{J}}, a_{\mathcal{I}} \right) \right] \right| \\
\leq \quad & \mathbb{E}_{a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}}, a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}} \left[\left| A_{\rho,r,\pi}^{\mathcal{I}} \left(s, a_{\mathcal{J}}, a_{\mathcal{I}} \right) - A_{\rho',r',\pi'}^{\mathcal{I}} \left(s, a_{\mathcal{J}}, a_{\mathcal{I}} \right) \right| \right] \\
\overset{\text{Corollary 4}}{\leq} \quad & \left(4\epsilon_{\rho} + \frac{4}{1-\gamma} \sum_{i \in -\mathcal{I}\cap-\mathcal{J}} \delta_i + \frac{2}{1-\gamma} \sum_{i\in\mathcal{I}} \delta_i + \frac{2\epsilon_r}{1-\gamma} \right) R.
\end{aligned}$$

Recall the definition of L, we have

$$
\begin{aligned}
& \left|L^{\mathcal{J}\cup\mathcal{I}}_{\eta,\rho,r,\pi}\left(\bar{\pi}_{\mathcal{J}},\hat{\pi}_{\mathcal{I}}\right)-L^{\mathcal{J}\cup\mathcal{I}}_{\eta',\rho',r',\pi'}\left(\bar{\pi}_{\mathcal{J}},\hat{\pi}_{\mathcal{I}}\right)\right| \\
= & \left|\mathbb{E}_{s\sim\eta,a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}},a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}}\left[A^{\mathcal{I}}_{\rho,r,\pi}\left(s,a_{\mathcal{J}},a_{\mathcal{I}}\right)\right]-\mathbb{E}_{s\sim\eta',a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}},a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}}\left[A^{\mathcal{I}}_{\rho',r',\pi'}\left(s,a_{\mathcal{J}},a_{\mathcal{I}}\right)\right]\right| \\
\leq & \left|\int_{s\in\mathcal{S}}\eta(\mathrm{d}s)\mathbb{E}_{a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}},a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}}\left[A^{\mathcal{I}}_{\rho,r,\pi}\left(s,a_{\mathcal{J}},a_{\mathcal{I}}\right)\right]\right. \\
& \left.-\int_{s\in\mathcal{S}}\eta'(\mathrm{d}s)\mathbb{E}_{a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}},a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}}\left[A^{\mathcal{I}}_{\rho,r,\pi}\left(s,a_{\mathcal{J}},a_{\mathcal{I}}\right)\right]\right| \\
& +\left|\int_{s\in\mathcal{S}}\eta'(\mathrm{d}s)\mathbb{E}_{a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}},a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}}\left[A^{\mathcal{I}}_{\rho,r,\pi}\left(s,a_{\mathcal{J}},a_{\mathcal{I}}\right)\right]\right. \\
& \left.-\int_{s\in\mathcal{S}}\eta'(\mathrm{d}s)\mathbb{E}_{a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}},a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}}\left[A^{\mathcal{I}}_{\rho',r',\pi'}\left(s,a_{\mathcal{J}},a_{\mathcal{I}}\right)\right]\right| \\
\overset{\text{Proposition 1}}{\leq} & \frac{4R}{1-\gamma}\|\eta-\eta'\|_{\mathrm{TV}}+\int_{s\in\mathcal{S}}\eta'(\mathrm{d}s)\left|\mathbb{E}_{a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}},a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}}\left[A^{\mathcal{I}}_{\rho,r,\pi}\left(s,a_{\mathcal{J}},a_{\mathcal{I}}\right)\right]\right. \\
& \left.-\mathbb{E}_{s\sim\eta',a_{\mathcal{J}}\sim\bar{\pi}_{\mathcal{J}},a_{\mathcal{I}}\sim\hat{\pi}_{\mathcal{I}}}\left[A^{\mathcal{I}}_{\rho',r',\pi'}\left(s,a_{\mathcal{J}},a_{\mathcal{I}}\right)\right]\right| \\
\leq & \frac{4R}{1-\gamma}\epsilon_\eta+\left(4\epsilon_\rho+\frac{4}{1-\gamma}\sum_{i\in-\mathcal{I}\cap-\mathcal{J}}\delta_i+\frac{2}{1-\gamma}\sum_{i\in\mathcal{I}}\delta_i+\frac{2\epsilon_r}{1-\gamma}\right)R\int_{s\in\mathcal{S}}\eta'(\mathrm{d}s) \\
\leq & \frac{4R}{1-\gamma}\epsilon_\eta+\frac{1}{1-\gamma}\left(4\epsilon_\rho+\frac{4}{1-\gamma}\sum_{i\in-\mathcal{I}\cap-\mathcal{J}}\delta_i+\frac{2}{1-\gamma}\sum_{i\in\mathcal{I}}\delta_i+\frac{2\epsilon_r}{1-\gamma}\right)R \\
= & \left(\frac{4}{1-\gamma}(\epsilon_\eta+\epsilon_\rho)+\frac{4}{(1-\gamma)^2}\sum_{i\in-\mathcal{I}\cap-\mathcal{J}}\delta_i+\frac{2}{(1-\gamma)^2}\sum_{i\in\mathcal{I}}\delta_i+\frac{2\epsilon_r}{(1-\gamma)^2}\right)R.
\end{aligned}
$$

□

Combining Lemmas 4 and 5, we obtain the following proof of Proposition 9.

Proof of Proposition 9.

$$
\begin{aligned}
& \left|L^{i_{1:m}}_{\eta_\mu,P,\pi,\rho_\mu,P,\pi,r,\pi}\left(\pi'_{i_{1:m-1}},\pi'_{i_m}\right)-L^{i_{1:m}}_{\eta',\rho',r',\tilde{\pi}}\left(\pi'_{i_{1:m-1}},\pi'_{i_m}\right)\right| \\
\overset{\text{Lemma 5}}{\leq} & \left[\frac{4}{1-\gamma}(\epsilon_\eta+\epsilon_\rho)+\frac{4}{(1-\gamma)^2}\sum_{j=m+1}^{n}\delta_{i_j}+\frac{2}{(1-\gamma)^2}\delta_{i_m}+\frac{2\epsilon_r}{(1-\gamma)^2}\right]R.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& J(\pi'|\mu,P,r) \\
\overset{\text{Lemma 4}}{\geq} \; & J(\pi|\mu,P,r) + \sum_{m=1}^{n}\left[L^{i_{1:m}}_{\eta_{\mu,P,\pi},\rho_{\mu,P,\pi},r,\pi}\left(\pi'_{i_{1:m-1}},\pi'_{i_m}\right) - C\mathrm{D}^{\max}_{\mathrm{KL}}\left(\pi_{i_m},\pi'_{i_m}\right)\right] \\
= \; & J(\pi|\mu,P,r) + \sum_{m=1}^{n}\left[L^{i_{1:m}}_{\eta',\rho',r',\tilde{\pi}}\left(\pi'_{i_{1:m-1}},\pi'_{i_m}\right) - C\mathrm{D}^{\max}_{\mathrm{KL}}\left(\pi_{i_m},\pi'_{i_m}\right)\right] \\
& -\frac{2n\epsilon_r R}{(1-\gamma)^2} - \sum_{m=1}^{n}\left[\frac{4}{1-\gamma}(\epsilon_\eta+\epsilon_\rho) + \frac{4}{(1-\gamma)^2}\sum_{j=m+1}^{n}\delta_{i_j} + \frac{2}{(1-\gamma)^2}\delta_{i_m}\right]R \\
= \; & J(\pi|\mu,P,r) + \sum_{m=1}^{n}\left[L^{i_{1:m}}_{\eta',\rho',r',\tilde{\pi}}\left(\pi'_{i_{1:m-1}},\pi'_{i_m}\right) - C\mathrm{D}^{\max}_{\mathrm{KL}}\left(\pi_{i_m},\pi'_{i_m}\right)\right] \\
& -\left[\frac{2n\epsilon_r}{(1-\gamma)^2} + \frac{4n}{1-\gamma}(\epsilon_\eta+\epsilon_\rho) + \frac{2}{(1-\gamma)^2}\sum_{j=1}^{n}(2j-1)\delta_{i_j}\right]R,
\end{aligned}
$$

where

$$
C = \frac{4\gamma \max_{s,a}|A_\pi(s,a|\mu,P,r)|}{(1-\gamma)^2}.
$$

Recall the proof of Lemma 5, notice that A is bounded by $\frac{2}{1-\gamma}R$, thus we have

$$
C \leq \frac{8\gamma}{(1-\gamma)^3}R.
$$

Besides, for each i_m,

$$
\mathrm{D}^{\max}_{\mathrm{TV}}\left(\pi_{i_m},\pi'_{i_m}\right) \geq 0.
$$

Therefore, we can select

$$
C = \frac{8\gamma}{(1-\gamma)^3}R
$$

as the constant. □

A.5 Theorem 1 and Corollary 1

In this subsection, we provide the complete proofs of Theorem 1 and Corollary 1. We begin with restating Theorem 1 and Corollary 1 in a more complete form.

Theorem 3 (Theorem 1 (Restated))**.** *Under the conditions of Definition 1 and Assumption 1, with bounded rewards as assumed in the model formulation, for each hyperedge $e \in \mathcal{E}$, the following bounds hold:*

$$
\begin{aligned}
\left\|\eta^e_{\mu,P,\bar{\pi}} - \eta_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} &\leq \frac{\gamma}{(1-\gamma)^2}\mathcal{F}\left(\phi_e\right), \\
\left\|\rho^e_{\mu,P,\bar{\pi}} - \rho_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} &\leq \frac{1}{(1-\gamma)^2}\mathcal{F}_{1+\epsilon_p(e)}\left(\phi_e\right) - \frac{\epsilon_p(e)}{1-\gamma}, \\
\left\|\rho^e_{\mu,P,\bar{\pi}}(\cdot|z'_e) - \rho_{\mu_e,P_e,\pi_e}(\cdot|z'_e)\right\|_{\mathrm{TV}} &\leq \frac{\gamma}{(1-\gamma)^2}\mathcal{F}(\phi_e), \\
|r_e(z_e) - r^e_{\mu,P,\bar{\pi}}(z_e)| &\leq \epsilon_r(e)R_e, \\
|r^e_{\mu,P,\bar{\pi}}(z_e)| &\leq R_e,
\end{aligned}
$$

where $\eta^e_{\mu,P,\bar{\pi}}$, $\rho^e_{\mu,P,\bar{\pi}}$, $\rho^e_{\mu,P,\bar{\pi}}(\cdot|z'_e)$, $r^e_{\mu,P,\bar{\pi}}$ are defined in Appendix A.2 and η_{μ_e,P_e,π_e}, ρ_{μ_e,P_e,π_e}, $\rho_{\mu_e,P_e,\pi_e}(\cdot|z'_e)$, $r_e(z_e)$ are defined by Definitions 1, 4 and 5. $R_e = \sum_{i\in e}|w_{i,e}|R_i$ and $\phi_e = \epsilon_p(e) + \sum_{i\in e}\delta_i$ are constants. $\mathcal{F}$ is defined in Appendix A.1.

Corollary 5. *(Corollary 1 (Restate)) Under the same conditions and notation as in Theorem 3, for each hyperedge $e \in \mathcal{E}$, if $\{\pi_{i,e}\}_{i\in e}$ is updated by Algorithm 1 with HATRL embedded, using an update order $\langle i_1, \ldots, i_{m_e}\rangle \in Sym(e)$, i.e., a permutation of the agents in e, for each $j = 1, \ldots, m_e$,*

$$\pi'_{i_j,e} = \arg\max_{\pi} \left[L^{i_{1:j}}_{\eta^e_{\mu,P,\bar{\pi}},\rho^e_{\mu,P,\bar{\pi}},r^e_{\mu,P,\bar{\pi}},\bar{\pi}}(\pi'_{i_{1:j-1},e}, \pi) - C_e \mathrm{D}^{\max}_{\mathrm{KL}}\left(\pi_{i_j,e}, \pi\right) \right].$$

Then the following performance inequality holds:

$$J(\pi'_e|\mu_e, P_e, r_e) \geq J(\pi_e|\mu_e, P_e, r_e) - D_e R_e,$$

where $D_e = A_e + B_e$,

$$A_e = \frac{8\gamma}{(1-\gamma)^3}\left(m_e \mathcal{F}(\phi_e) + \sum_{i\in e} \frac{n_i \log n_i}{n_i - 1}\delta_i\right),$$

and

$$B_e = \frac{2}{(1-\gamma)^2}\left(m_e \epsilon_r(e) + \sum_{j=1}^{m_e}(2j-1)\delta_{i_j}\right).$$

Proof of Theorem 1 is divided into Lemmas 6–9.

Lemma 6. *With Definition 1 and Assumption 1, the following inequality holds.*

$$\left\|\eta^e_{\mu,P,\bar{\pi}} - \eta_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} \leq \frac{\gamma}{(1-\gamma)^2}\mathcal{F}\left(\phi_e\right),$$

where

$$\phi_e = \sum_{i\in e}\delta_i + \epsilon_p(e).$$

Proof. Recall the definition 5, we have

$$\begin{aligned}
\eta_{\mu,P,\pi}(\mathrm{d}s) &= \int_{s'\in\mathcal{S}} \mu(\mathrm{d}s')\rho_{\mu,P,\pi}(\mathrm{d}s|s') \\
&= \int_{s'\in\mathcal{S}} \mu(\mathrm{d}s')\sum_{t=0}^{\infty}\gamma^t \zeta^t_{\mu,P,\pi}(\mathrm{d}s|s') \\
&= \sum_{t=0}^{\infty}\gamma^t \int_{s'\in\mathcal{S}} \mu(\mathrm{d}s')\zeta^t_{\mu,P,\pi}(\mathrm{d}s|s').
\end{aligned}$$

Let

$$\zeta^t_{\mu,P,\pi}(\mathrm{d}s) = \int_{s'\in\mathcal{S}} \mu(\mathrm{d}s')\zeta^t_{\mu,P,\pi}(\mathrm{d}s|s'),$$

then

$$\zeta^0_{\mu,P,\pi}(\mathrm{d}s) = \int_{s'\in\mathcal{S}} \mu(\mathrm{d}s')\delta_{s'}(\mathrm{d}s) = \mu(\mathrm{d}s),$$

$$\begin{aligned}
\zeta^{t+1}_{\mu,P,\pi}(\mathrm{d}s) &= \int_{s'\in\mathcal{S}} \mu(\mathrm{d}s') \int_{(s'',a)\in\mathcal{S}\times\mathcal{A}} P(\mathrm{d}s|s'',a)\pi(\mathrm{d}a|s'')\zeta^t_{\mu,P,\pi}(\mathrm{d}s''|s') \\
&= \int_{(s'',a)\in\mathcal{S}\times\mathcal{A}} P(\mathrm{d}s|s'',a)\pi(\mathrm{d}a|s'')\zeta^t_{\mu,P,\pi}(\mathrm{d}s''),
\end{aligned}$$

therefore

$$\begin{aligned}
\zeta^0_{\mu,P,\pi} &= \mu, \\
\zeta^{t+1}_{\mu,P,\pi} &= (\zeta^t_{\mu,P,\pi} \otimes \pi)P.
\end{aligned}$$

Let $\Pi_e : \mathcal{M}_+(\mathcal{S}) \to \mathcal{M}_+(\mathcal{S}_e)$ such that,

$$(\Pi_e \nu)(\mathrm{d}s_e) = \int_{s_{-e}\in\mathcal{S}_{-e}} \nu(\mathrm{d}s_e, \mathrm{d}s_{-e}).$$

$$\eta^e_{\mu,P,\bar{\pi}} = \Pi_e \eta_{\mu,P,\bar{\pi}} = \Pi_e \left(\sum_{t=0}^{\infty} \gamma^t \zeta^t_{\mu,P,\bar{\pi}} \right) = \sum_{t=0}^{\infty} \gamma^t \Pi_e \zeta^t_{\mu,P,\bar{\pi}}.$$

With Proposition 3, we have

$$\left\| \eta^e_{\mu,P,\bar{\pi}} - \eta_{\mu_e,P_e,\pi_e} \right\|_{\mathrm{TV}} \leq \sum_{t=0}^{\infty} \gamma^t \left\| \zeta^t_{\mu_e,P_e,\pi_e} - \Pi_e \zeta^t_{\mu,P,\bar{\pi}} \right\|_{\mathrm{TV}}$$

Let

$$\Xi^t = \left\| \zeta^t_{\mu_e,P_e,\pi_e} - \Pi_e \zeta^t_{\mu,P,\bar{\pi}} \right\|_{\mathrm{TV}} \leq 1.$$
$$\Xi^0 = \left\| \mu_e - \Pi_e \mu \right\|_{\mathrm{TV}} = 0.$$

$$\begin{aligned}
\Pi_e(\zeta^t_{\mu,P,\bar{\pi}} \otimes \bar{\pi}) P(\mathrm{d}s_e) &= \int_{s_{-e} \in \mathcal{S}_{-e}} \int_{(s',a) \in \mathcal{S} \times \mathcal{A}} P(\mathrm{d}s_e, \mathrm{d}s_{-e} | s', a) \bar{\pi}(\mathrm{d}a | s') \zeta^t_{\mu,P,\bar{\pi}}(\mathrm{d}s') \\
&= \int_{(s',a) \in \mathcal{S} \times \mathcal{A}} \left(\int_{s_{-e} \in \mathcal{S}_{-e}} P(\mathrm{d}s_e, \mathrm{d}s_{-e} | s', a) \right) \bar{\pi}(\mathrm{d}a | s') \zeta^t_{\mu,P,\bar{\pi}}(\mathrm{d}s'),
\end{aligned}$$

therefore

$$\Pi_e(\zeta^t_{\mu,P,\bar{\pi}} \otimes \bar{\pi}) P = \left(\zeta^t_{\mu,P,\bar{\pi}} \otimes \bar{\pi}\right) (\Pi_e P)$$

$$\begin{aligned}
\Xi^{t+1} &= \left\| (\zeta^t_{\mu_e,P_e,\pi_e} \otimes \pi_e) P_e - \left(\zeta^t_{\mu,P,\bar{\pi}} \otimes \bar{\pi}\right) (\Pi_e P) \right\|_{\mathrm{TV}} \\
&\leq \underbrace{\left\| (\zeta^t_{\mu_e,P_e,\pi_e} \otimes \pi_e) P_e - (\Pi_e \zeta^t_{\mu,P,\bar{\pi}} \otimes \pi_e) P_e \right\|_{\mathrm{TV}}}_{I_1} \\
&\quad + \underbrace{\left\| (\Pi_e \zeta^t_{\mu,P,\bar{\pi}} \otimes \pi_e) P_e - (\Pi_e \zeta^t_{\mu,P,\bar{\pi}} \otimes \bar{\pi}_e) P_e \right\|_{\mathrm{TV}}}_{I_2} \\
&\quad + \underbrace{\left\| (\Pi_e \zeta^t_{\mu,P,\bar{\pi}} \otimes \bar{\pi}_e) P_e - \left(\zeta^t_{\mu,P,\bar{\pi}} \otimes \bar{\pi}\right) (\Pi_e P) \right\|_{\mathrm{TV}}}_{I_3}.
\end{aligned}$$

Recall that $\bar{\pi} = \prod_{i \in \mathcal{N}} \bar{\pi}_i$,

$$\begin{aligned}
I_1 &\overset{\text{Proposition 5}}{\leq} \left\| \zeta^t_{\mu_e,P_e,\pi_e} \otimes \pi_e - \Pi_e \zeta^t_{\mu,P,\bar{\pi}} \otimes \pi_e \right\|_{\mathrm{TV}} \\
&\overset{\text{Proposition 6}}{\leq} \left\| \zeta^t_{\mu_e,P_e,\pi_e} - \Pi_e \zeta^t_{\mu,P,\bar{\pi}} \right\|_{\mathrm{TV}} \\
&= \Xi^t,
\end{aligned}$$

$$\begin{aligned}
I_2 &\overset{\text{Proposition 5}}{\leq} \left\| (\Pi_e \zeta^t_{\mu,P,\bar{\pi}}) \otimes \pi_e - (\Pi_e \zeta^t_{\mu,P,\bar{\pi}}) \otimes \bar{\pi}_e \right\|_{\mathrm{TV}} \\
&\overset{\text{Proposition 7}}{\leq} \int_{s_e \in \mathcal{S}_e} \left\| \pi_e(\cdot | s'_e) - \bar{\pi}_e(\cdot | s'_e) \right\|_{\mathrm{TV}} \Pi_e \zeta^t_{\mu,P,\bar{\pi}}(\mathrm{d}s_e) \\
&\overset{\text{Assumption 1}}{\leq} \sum_{i \in e} \delta_i \int_{s_e \in \mathcal{S}_e} \Pi_e \zeta^t_{\mu,P,\bar{\pi}}(\mathrm{d}s_e) \\
&= \sum_{i \in e} \delta_i,
\end{aligned}$$

$$\begin{aligned}
I_3 &= \left\| \int_{(s',a) \in \mathcal{S} \times \mathcal{A}} P_e(\cdot | s'_e, a_e) \bar{\pi}(\mathrm{d}a | s') \zeta^t_{\mu,P,\bar{\pi}}(\mathrm{d}s') \right. \\
&\quad \left. - \int_{(s',a) \in \mathcal{S} \times \mathcal{A}} (\Pi_e P)(\cdot | s', a) \bar{\pi}(\mathrm{d}a | s') \zeta^t_{\mu,P,\bar{\pi}}(\mathrm{d}s') \right\|_{\mathrm{TV}} \\
&\overset{\text{Proposition 4}}{\leq} \int_{(s',a) \in \mathcal{S} \times \mathcal{A}} \left\| P_e(\cdot | s'_e, a_e) - \Pi_e P(\cdot | s', a) \right\|_{\mathrm{TV}} \bar{\pi}(\mathrm{d}a | s') \zeta^t_{\mu,P,\bar{\pi}}(\mathrm{d}s') \\
&\overset{\text{Definition 1}}{\leq} \epsilon_p(e) \int_{(s',a) \in \mathcal{S} \times \mathcal{A}} \bar{\pi}(\mathrm{d}a | s') \zeta^t_{\mu,P,\bar{\pi}}(\mathrm{d}s') \\
&= \epsilon_p(e).
\end{aligned}$$

Therefore, by applying mathematical induction, we have

$$\Xi^t \leq \min\left(t\sum_{i\in e}\delta_i + t\epsilon_p(e), 1\right)$$

$$\left\|\eta^e_{\mu,P,\bar{\pi}} - \eta_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} \leq \sum_{t=0}^{\infty}\gamma^t\Xi^t \overset{\text{Lemma 1}}{\leq} \frac{\gamma}{(1-\gamma)^2}\mathcal{F}\left(\sum_{i\in e}\delta_i + \epsilon_p(e)\right).$$

□

Lemma 7. *With Definition 1 and Assumption 1, the following inequality holds.*

$$\left\|\rho^e_{\mu,P,\bar{\pi}} - \rho_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} \leq \frac{1}{(1-\gamma)^2}\mathcal{F}_{1+\epsilon_p(e)}\left(\phi_e\right) - \frac{\epsilon_p(e)}{1-\gamma},$$

where

$$\phi_e = \sum_{i\in e}\delta_i + \epsilon_p(e).$$

Proof. Recall the definition 4, we have

$$\begin{aligned}
\rho_{\mu,P,\pi}(\mathrm{d}z) &= \int_{s'\in\mathcal{S}}\mu(\mathrm{d}s')\rho_{\mu,P,\pi}(\mathrm{d}z|s') \\
&= \int_{s'\in\mathcal{S}}\int_{a'\in\mathcal{A}}\pi(\mathrm{d}a'|s')\mu(\mathrm{d}s')\rho_{\mu,P,\pi}(\mathrm{d}z|s',a') \\
&= \int_{(s',a')\in\mathcal{S}\times\mathcal{A}}\mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')\rho_{\mu,P,\pi}(\mathrm{d}z|s',a') \\
&= \int_{(s',a')\in\mathcal{S}\times\mathcal{A}}\mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')\sum_{t=0}^{\infty}\gamma^t\varrho^t_{\mu,P,\pi}(\mathrm{d}z|s',a') \\
&= \sum_{t=0}^{\infty}\gamma^t\int_{(s',a')\in\mathcal{S}\times\mathcal{A}}\mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')\varrho^t_{\mu,P,\pi}(\mathrm{d}z|s',a')
\end{aligned}$$

Let

$$\varrho^t_{\mu,P,\pi}(\mathrm{d}z) = \int_{(s',a')\in\mathcal{S}\times\mathcal{A}}\mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')\varrho^t_{\mu,P,\pi}(\mathrm{d}z|s',a'),$$

then

$$\rho_{\mu,P,\pi} = \sum_{t=0}^{\infty}\gamma^t\varrho^t_{\mu,P,\pi}.$$

$$\varrho^0_{\mu,P,\pi}(\mathrm{d}z) = \int_{(s',a')\in\mathcal{S}\times\mathcal{A}}\mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')\delta_{s,a}(\mathrm{d}s',\mathrm{d}a') = \mu(\mathrm{d}s)\pi(\mathrm{d}a|s),$$

and

$$\begin{aligned}
&\varrho^{t+1}_{\mu,P,\pi}(\mathrm{d}z) \\
&= \int_{(s',a')\in\mathcal{S}\times\mathcal{A}}\mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')\varrho^{t+1}_{\mu,P,\pi}(\mathrm{d}z|s',a') \\
&= \int_{(s',a')\in\mathcal{S}\times\mathcal{A}}\mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')\pi(\mathrm{d}a|s)\int_{z''\in\mathcal{Z}}P(\mathrm{d}s|z'')\varrho^t_{\mu,P,\pi}(\mathrm{d}z''|s',a') \\
&= \pi(\mathrm{d}a|s)\int_{z''\in\mathcal{Z}}P(\mathrm{d}s|z'')\left(\int_{(s',a')\in\mathcal{S}\times\mathcal{A}}\mu(\mathrm{d}s')\pi(\mathrm{d}a'|s')\varrho^t_{\mu,P,\pi}(\mathrm{d}z''|s',a')\right) \\
&= \pi(\mathrm{d}a|s)\int_{z''\in\mathcal{Z}}P(\mathrm{d}s|z'')\varrho^t_{\mu,P,\pi}(\mathrm{d}z''),
\end{aligned}$$

therefore,

$$\varrho^0_{\mu,P,\pi} = \mu\otimes\pi,$$

$$\varrho^{t+1}_{\mu,P,\pi} = (\varrho^t_{\mu,P,\pi} P) \otimes \pi. \tag{15}$$

Let $\Pi_e : \mathcal{M}_+(\mathcal{Z}) \to \mathcal{M}_+(\mathcal{Z}_e)$ such that,

$$(\Pi_e \nu)(\mathrm{d}z_e) = \int_{z_{-e} \in \mathcal{Z}_{-e}} \nu(\mathrm{d}z_e, \mathrm{d}z_{-e}).$$

Let $\Phi_e : \mathcal{M}_+(\mathcal{S}) \to \mathcal{M}_+(\mathcal{S}_e)$ such that,

$$(\Phi_e \nu)(\mathrm{d}s_e) = \int_{s_{-e} \in \mathcal{S}_{-e}} \nu(\mathrm{d}s_e, \mathrm{d}s_{-e}).$$

$$\begin{aligned}
\rho^e_{\mu,P,\bar{\pi}} &= \Pi_e \rho_{\mu,P,\bar{\pi}} \\
&= \Pi_e \left(\sum_{t=0}^{\infty} \gamma^t \varrho^t_{\mu,P,\bar{\pi}} \right) \\
&= \sum_{t=0}^{\infty} \gamma^t \Pi_e \varrho^t_{\mu,P,\bar{\pi}}.
\end{aligned}$$

With Proposition 3, we have

$$\left\| \rho^e_{\mu,P,\bar{\pi}} - \rho_{\mu_e,P_e,\pi_e} \right\|_{\mathrm{TV}} \leq \sum_{t=0}^{\infty} \gamma^t \left\| \varrho^t_{\mu_e,P_e,\pi_e} - \Pi_e \varrho^t_{\mu,P,\bar{\pi}} \right\|_{\mathrm{TV}}$$

Let

$$\Xi^t = \left\| \varrho^t_{\mu_e,P_e,\pi_e} - \Pi_e \varrho^t_{\mu,P,\bar{\pi}} \right\|_{\mathrm{TV}} \leq 1.$$

Recall that $\bar{\pi} = \prod_{i \in \mathcal{N}} \bar{\pi}_i$,

$$\begin{aligned}
\Pi_e(\mu \otimes \bar{\pi})(\mathrm{d}s_e, \mathrm{d}a_e) &= \int_{(s_{-e}, a_{-e}) \in \mathcal{S}_{-e} \times \mathcal{A}_{-e}} \mu(\mathrm{d}s_e, \mathrm{d}s_{-e}) \bar{\pi}(\mathrm{d}a_e, \mathrm{d}a_{-e} | s_e, s_{-e}) \\
&= \bar{\pi}_e(\mathrm{d}a_e | s_e) \int_{s_{-e} \in \mathcal{S}_{-e}} \mu(\mathrm{d}s_e, \mathrm{d}s_{-e}),
\end{aligned}$$

therefore,

$$\Pi_e(\mu \otimes \bar{\pi}) = (\Phi_e \mu) \otimes \bar{\pi}_e.$$

$$\begin{aligned}
\Xi^0 &= \left\| \mu_e \otimes \pi_e - \Pi_e(\mu \otimes \bar{\pi}) \right\|_{\mathrm{TV}} \\
&= \left\| \mu_e \otimes \pi_e - (\Phi_e \mu) \otimes \bar{\pi}_e \right\|_{\mathrm{TV}} \\
&\leq \left\| \mu_e \otimes \pi_e - (\Phi_e \mu) \otimes \pi_e \right\|_{\mathrm{TV}} + \left\| (\Phi_e \mu) \otimes \pi_e - (\Phi_e \mu) \otimes \bar{\pi}_e \right\|_{\mathrm{TV}} \\
&\overset{\substack{\text{Proposition 6} \\ \text{Proposition 7}}}{=} \left\| \mu_e - \Phi_e \mu \right\|_{\mathrm{TV}} + \int_{s_e \in \mathcal{S}_e} \left\| \pi_e(\cdot | s_e) - \bar{\pi}(\cdot | s_e) \right\|_{\mathrm{TV}} \Phi_e \mu(\mathrm{d}s_e) \\
&\overset{\text{Definition 1}}{=} \int_{s_e \in \mathcal{S}_e} \left\| \pi_e(\cdot | s_e) - \bar{\pi}(\cdot | s_e) \right\|_{\mathrm{TV}} \Phi_e \mu(\mathrm{d}s_e) \\
&\overset{\text{Assumption 1}}{\leq} \sum_{i \in e} \delta_i \int_{s_e \in \mathcal{S}_e} \Phi_e \mu(\mathrm{d}s_e) \\
&= \sum_{i \in e} \delta_i.
\end{aligned}$$

$$\begin{aligned}
\left(\Pi_e \varrho^{t+1}_{\mu,P,\bar{\pi}} \right) (\mathrm{d}s_e, \mathrm{d}a_e) &= \int_{(s_{-e}, a_{-e}) \in \mathcal{S}_{-e} \times \mathcal{A}_{-e}} \bar{\pi}(\mathrm{d}a_e, \mathrm{d}a_{-e} | s) \int_{z' \in \mathcal{Z}} P(\mathrm{d}s_e, \mathrm{d}s_{-e} | z') \varrho^t_{\mu,P,\bar{\pi}}(\mathrm{d}z') \\
&= \int_{z' \in \mathcal{Z}} \varrho^t_{\mu,P,\bar{\pi}}(\mathrm{d}z') \int_{(s_{-e}, a_{-e}) \in \mathcal{S}_{-e} \times \mathcal{A}_{-e}} \bar{\pi}(\mathrm{d}a_e, \mathrm{d}a_{-e} | s) P(\mathrm{d}s_e, \mathrm{d}s_{-e} | z') \\
&= \bar{\pi}_e(\mathrm{d}a_e | s_e) \int_{z' \in \mathcal{Z}} \varrho^t_{\mu,P,\bar{\pi}}(\mathrm{d}z') \int_{s_{-e} \in \mathcal{S}_{-e}} P(\mathrm{d}s_e, \mathrm{d}s_{-e} | z'),
\end{aligned}$$

therefore

$$\Pi_e \varrho^{t+1}_{\mu,P,\bar{\pi}} = \left(\varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right)\right) \otimes \bar{\pi}_e. \tag{16}$$

$$\begin{aligned}
\Xi^{t+1} &= \left\| \varrho^{t+1}_{\mu_e,P_e,\pi_e} - \Pi_e \varrho^{t+1}_{\mu,P,\bar{\pi}} \right\|_{\mathrm{TV}} \\
&= \left\| (\varrho^t_{\mu_e,P_e,\pi_e} P_e) \otimes \pi_e - \left(\varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right)\right) \otimes \bar{\pi}_e \right\|_{\mathrm{TV}} \\
&\leq \left\| (\varrho^t_{\mu_e,P_e,\pi_e} P_e) \otimes \pi_e - \left(\varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right)\right) \otimes \pi_e \right\|_{\mathrm{TV}} \\
&\quad + \left\| \left(\varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right)\right) \otimes \pi_e - \left(\varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right)\right) \otimes \bar{\pi}_e \right\|_{\mathrm{TV}} \\
&\overset{\text{Proposition 6}}{\underset{\text{Proposition 7}}{=}} \left\| \varrho^t_{\mu_e,P_e,\pi_e} P_e - \varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right) \right\|_{\mathrm{TV}} \\
&\quad + \int_{s_e \in \mathcal{S}_e} \left\| \pi_e(\cdot|s_e) - \bar{\pi}(\cdot|s_e) \right\|_{\mathrm{TV}} \left(\varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right)\right)(\mathrm{d}s_e) \\
&\leq \underbrace{\left\| \varrho^t_{\mu_e,P_e,\pi_e} P_e - \left(\Pi_e \varrho^t_{\mu,P,\bar{\pi}}\right) P_e \right\|_{\mathrm{TV}}}_{I_1} + \underbrace{\left\| \left(\Pi_e \varrho^t_{\mu,P,\bar{\pi}}\right) P_e - \varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right) \right\|_{\mathrm{TV}}}_{I_2} \\
&\quad + \underbrace{\int_{s_e \in \mathcal{S}_e} \left\| \pi_e(\cdot|s_e) - \bar{\pi}(\cdot|s_e) \right\|_{\mathrm{TV}} \left(\varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right)\right)(\mathrm{d}s_e)}_{I_3}.
\end{aligned}$$

$$\begin{aligned}
I_1 &\overset{\text{Proposition 5}}{\leq} \left\| \varrho^t_{\mu_e,P_e,\pi_e} - \Pi_e \varrho^t_{\mu,P,\bar{\pi}} \right\|_{\mathrm{TV}} \\
&= \Xi^t.
\end{aligned}$$

$$\begin{aligned}
I_2 &= \left\| \left(\Pi_e \varrho^t_{\mu,P,\bar{\pi}}\right) P_e - \varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right) \right\|_{\mathrm{TV}} \\
&= \left\| \int_{z' \in \mathcal{Z}} P_e(\cdot|z'_e) \varrho^t_{\mu,P,\bar{\pi}}(\mathrm{d}z) - \int_{z' \in \mathcal{Z}} \left(\Phi_e P\right)(\cdot|z') \varrho^t_{\mu,P,\bar{\pi}}(\mathrm{d}z) \right\|_{\mathrm{TV}} \\
&\overset{\text{Proposition 4}}{\leq} \int_{z_{-e} \in \mathcal{Z}_{-e}} \left\| P_e(\cdot|z'_e) - \left(\Phi_e P\right)(\cdot|z') \right\|_{\mathrm{TV}} \varrho^t_{\mu,P,\bar{\pi}}(\mathrm{d}z) \\
&\overset{\text{Definition 1}}{\leq} \epsilon_p(e) \int_{z_{-e} \in \mathcal{Z}_{-e}} \varrho^t_{\mu,P,\bar{\pi}}(\mathrm{d}z) \\
&= \epsilon_p(e),
\end{aligned}$$

where $P(\cdot|z')_{-e}$ denotes the marginal distribution of $P(\cdot|z')$ on $\mathcal{S}_{-e}$.

$$I_3 \overset{\substack{\text{Proposition 2}\\ \text{Assumption 1}}}{\leq} \sum_{i \in e} \delta_i \int_{s_e \in \mathcal{S}_e} \left(\varrho^t_{\mu,P,\bar{\pi}}\left(\Phi_e P\right)\right)(\mathrm{d}s_e) = \sum_{i \in e} \delta_i.$$

Therefore, by applying mathematical induction, we have

$$\Xi^t \leq \min(t\epsilon_p(e) + (t+1)\sum_{i \in e} \delta_i, 1).$$

$$\begin{aligned}
\left\| \rho^e_{\mu,P,\bar{\pi}} - \rho_{\mu_e,P_e,\pi_e} \right\|_{\mathrm{TV}} &= \sum_{t=0}^{\infty} \gamma^t \Xi^t \\
&\overset{\text{Corollary 3}}{\leq} \frac{1}{(1-\gamma)^2} \mathcal{F}_{1+\epsilon_p(e)}\left(\epsilon_p(e) + \sum_{i \in e} \delta_i\right) - \frac{\epsilon_p(e)}{1-\gamma}.
\end{aligned}$$

□

Lemma 8. *With Definition 1 and Assumption 1, for each $z'_e \in \mathcal{S}_e \times \mathcal{A}_e$, the following inequality holds.*

$$\left\| \rho^e_{\mu,P,\bar{\pi}}(\cdot|z'_e) - \rho_{\mu_e,P_e,\pi_e}(\cdot|z'_e) \right\|_{\mathrm{TV}} \leq \frac{\gamma}{(1-\gamma)^2} \mathcal{F}(\phi_e),$$

where

$$\phi_e = \sum_{i \in e} \delta_i + \epsilon_p(e).$$

Proof. Arbitrarily select a $z'_e \in \mathcal{S}_e \times \mathcal{A}_e$.

$$\begin{aligned}
&\rho^e_{\mu,P,\bar{\pi}}(\mathrm{d}z_e|z'_e) \\
&= (1-\gamma)\int_{z'_{-e}\in\mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z'_{-e}) \int_{z_{-e}\in\mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}|z'_e, z'_{-e}) \\
&= (1-\gamma)\int_{z'_{-e}\in\mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z'_{-e}) \int_{z_{-e}\in\mathcal{Z}_{-e}} \sum_{t=0}^{\infty}\gamma^t \varrho^t_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}|z'_e, z'_{-e}) \\
&= \sum_{t=0}^{\infty}\gamma^t(1-\gamma)\int_{z_{-e}\in\mathcal{Z}_{-e}}\int_{z'_{-e}\in\mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z'_{-e})\varrho^t_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}|z'_e, z'_{-e}).
\end{aligned}$$

$$\rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e|z'_e) = \sum_{t=0}^{\infty}\gamma^t \varrho^t_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e|z'_e).$$

As z'_e is fixed, we denote $\rho^e_{\mu,P,\bar{\pi}}(\cdot|z'_e)$ as $\theta^e_{\mu,P,\bar{\pi}}$, $\varrho^t_{\mu,P,\bar{\pi}}(\cdot|z'_e, z'_{-e})$ as $\vartheta^t_{\mu,P,\bar{\pi}}(\cdot|z'_{-e})$, $\varrho^t_{\mu_e,P_e,\pi_e}(\cdot|z'_e)$ as $\vartheta^t_{\mu_e,P_e,\pi_e}$ and $(1-\gamma)\rho^{-e}_{\mu,P,\bar{\pi}}$ as μ.

Let $\Pi_e : \mathcal{M}_+(\mathcal{Z}) \to \mathcal{M}_+(\mathcal{Z}_e)$ such that,

$$(\Pi_e\nu)(\mathrm{d}z_e) = \int_{z_{-e}\in\mathcal{Z}_{-e}} \nu(\mathrm{d}z_e, \mathrm{d}z_{-e}).$$

Let $\Phi_e : \mathcal{M}_+(\mathcal{S}) \to \mathcal{M}_+(\mathcal{S}_e)$ such that,

$$(\Phi_e\nu)(\mathrm{d}s_e) = \int_{s_{-e}\in\mathcal{S}_{-e}} \nu(\mathrm{d}s_e, \mathrm{d}s_{-e}).$$

With Proposition 3, we have

$$\left\|\theta^e_{\mu,P,\bar{\pi}} - \theta_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} \le \sum_{t=0}^{\infty}\gamma^t \left\|\left(\Pi_e\left(\mu\vartheta^t_{\mu,P,\bar{\pi}}\right)\right) - \vartheta^t_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}}.$$

Let

$$\Xi^t = \left\|\left(\Pi_e\left(\mu\vartheta^t_{\mu,P,\bar{\pi}}\right)\right) - \vartheta^t_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} \le 1.$$

$$\begin{aligned}
\left(\Pi_e\left(\mu\vartheta^0_{\mu,P,\bar{\pi}}\right)\right)(\mathrm{d}z_e) &= \int_{z_{-e}\in\mathcal{Z}_{-e}}\int_{z'_{-e}\in\mathcal{Z}_{-e}} \mu(\mathrm{d}z'_{-e})\delta_{z'_e,z'_{-e}}(\mathrm{d}z_e, \mathrm{d}z_{-e}) \\
&= \delta_{z'_e}(\mathrm{d}z_e)\int_{z'_{-e}\in\mathcal{Z}_{-e}}\int_{z_{-e}\in\mathcal{Z}_{-e}} \mu(\mathrm{d}z'_{-e})\delta_{z'_{-e}}(\mathrm{d}z_{-e}) \\
&= \delta_{z'_e}(\mathrm{d}z_e)\int_{z'_{-e}\in\mathcal{Z}_{-e}} \mu(\mathrm{d}z'_{-e}) \\
&= \delta_{z'_e}(\mathrm{d}z_e) = \vartheta^0_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e).
\end{aligned}$$

Therefore,

$$\Xi^0(s'_e, a'_e) = 0.$$

Recall the definition 4, we have

$$\vartheta^{t+1}_{\mu,P,\bar{\pi}}(\mathrm{d}s, \mathrm{d}a|z'_{-e}) = \bar{\pi}(\mathrm{d}a|s)\int_{z''\in\mathcal{Z}} P(\mathrm{d}s|z'')\vartheta^t_{\mu,P,\bar{\pi}}(\mathrm{d}z''|z'_{-e}),$$

and

$$\vartheta^{t+1}_{\mu_e,P_e,\pi_e}(\mathrm{d}s_e, \mathrm{d}a_e) = \pi_e(\mathrm{d}a_e|s_e)\int_{z''_e\in\mathcal{Z}_e} P_e(\mathrm{d}s_e|z''_e)\vartheta^t_{\mu_e,P_e,\pi_e}(\mathrm{d}z''_e).$$

Therefore,

$$\begin{aligned}
\Pi_e\left(\mu\vartheta^{t+1}_{\mu,P,\bar{\pi}}\right) &= \Pi_e\left(\mu\left(\vartheta^t_{\mu,P,\bar{\pi}}P\otimes\bar{\pi}\right)\right) \\
&= \Pi_e\left(\left(\mu\vartheta^t_{\mu,P,\bar{\pi}}\right)P\otimes\bar{\pi}\right) \\
&= \left(\mu\vartheta^t_{\mu,P,\bar{\pi}}\right)(\Phi_eP)\otimes\bar{\pi}_e, \qquad (17)
\end{aligned}$$

$$\vartheta^{t+1}_{\mu_e,P_e,\pi_e} = \left(\vartheta^{t}_{\mu_e,P_e,\pi_e} P_e\right) \otimes \pi_e. \tag{18}$$

The third line of Eq. (16) follows from the fact that $\mu\vartheta^t_{\mu,P,\bar{\pi}}$ is a measure on $\mathcal{Z}_e$, while Π_e only affects on $\mathcal{Z}_{-e}$. It can be observed that the recurrence relation shown in Eq. (17) and (18) are identical in form to Eq. (16) and (15) from the proof of Lemma 7; therefore, the following result can be derived using the same procedure, the details of which will not be reiterated here.

$$\Xi^t(s'_e, a'_e) \leq \min(t\phi_e, 1).$$

$$\left\|\theta^e_{\mu,P,\bar{\pi}} - \theta_{\mu_e,P_e,\pi_e}\right\|_{\mathrm{TV}} \leq \sum_{t=0}^{\infty} \gamma^t \Xi^t(s'_e, a'_e) \overset{\text{Lemma 1}}{\leq} \frac{\gamma}{(1-\gamma)^2}\mathcal{F}(\phi_e).$$

□

Lemma 9. *With bounded rewards and Definition 1, for each $z_e \in \mathcal{Z}_e$, the following inequality holds:*

$$|r_e(z_e) - r^e_{\mu,P,\bar{\pi}}(z_e)| \leq \epsilon_r(e) R_e.$$

Besides, $r^e_{\mu,P,\bar{\pi}}(z_e)$ is bound by R_e.

Proof.

$$\begin{aligned}
& |r_e(z_e) - r^e_{\mu,P,\bar{\pi}}(z_e)| \\
= \; & \left|(1-\gamma)\int_{z_{-e}\in\mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e}) r_e(z_e) - (1-\gamma)\int_{z_{-e}\in\mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e})\tilde{r}_e(z_e, z_{-e})\right| \\
\leq \; & (1-\gamma)\int_{z_{-e}\in\mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e}) |r_e(z_e) - \tilde{r}_e(z_e, z_{-e})| \\
\overset{\text{Definition 1}}{\leq} \; & \epsilon_r(e) R_e (1-\gamma)\int_{z_{-e}\in\mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e}) \\
\leq \; & \epsilon_r(e) R_e.
\end{aligned}$$

For each $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$|\tilde{r}_e(s,a)| \leq \sum_{i\in e} |w_{i,e}|\,|r_i(s,a)| \overset{\text{bounded rewards}}{\leq} R_e,$$

thus

$$|r^e_{\mu,P,\bar{\pi}}(z_e)| \leq R_e.$$

□

Collecting the results, the first inequality of Theorem 1 follows from Lemma 6, the second from Lemma 7, the third from Lemma 8, and the last two from Lemma 9. Thus all the inequalities of Theorem 1 are proven. Corollary 1 can be obtained directly by applying Theorem 1 to Proposition 9.

Proof of Corollary 1.

$$\begin{aligned}
& J(\pi'_e|\mu_e, P_e, r_e) \\
& \geq J(\pi_e|\mu_e, P_e, r_e) + \sum_{j=1}^{m_e}\left[L^{i_{1:j}}_{\eta^e_{\mu,P,\bar{\pi}},\rho^e_{\mu,P,\bar{\pi}},r^e_{\mu,P,\bar{\pi}},\bar{\pi}}\left(\pi'_{i_{1:j-1}}, \pi'_{i_j}\right) - C_e \mathrm{D}^{\max}_{\mathrm{KL}}\left(\pi_{i_j}, \bar{\pi}_{i_j}\right)\right] \\
& \quad - \left[\frac{2m_e\epsilon_r(e)}{(1-\gamma)^2} + \frac{8m_e\gamma}{(1-\gamma)^3}\mathcal{F}(\phi_e) + \frac{2}{(1-\gamma)^2}\sum_{j=1}^{m_e}(2j-1)\delta_{i_j}\right] R_e,
\end{aligned}$$

where

$$C_e = \frac{8\gamma}{(1-\gamma)^3} R_e$$

Notice that

$$
\begin{aligned}
&L^{i_{1:j}}_{\eta^e_{\mu,P,\bar{\pi}},\rho^e_{\mu,P,\bar{\pi}},r^e_{\mu,P,\bar{\pi}},\bar{\pi}}\left(\pi'_{i_{1:j-1}},\bar{\pi}_{i_j}\right)\\
&=\mathbb{E}_{s\sim\eta^e_{\mu,P,\bar{\pi}},a_{i_{1:j-1}}\sim\pi'_{i_{1:j-1}},a_{i_j}\sim\bar{\pi}_{i_j}}\left[A^{i_{1:j-1}}_{\rho^e_{\mu,P,\bar{\pi}},r^e_{\mu,P,\bar{\pi}},\bar{\pi}}(s,a_{i_{1:j-1}},a_{i_j})\right]\\
&=\mathbb{E}_{s\sim\eta^e_{\mu,P,\bar{\pi}},a_{i_{1:j-1}}\sim\pi'_{i_{1:j-1}},a_{i_j}\sim\bar{\pi}_{i_j}}\left[Q^{i_{1:j}}_{\rho^e_{\mu,P,\bar{\pi}},r^e_{\mu,P,\bar{\pi}},\bar{\pi}}(s,a_{i_{1:j-1}},a_{i_j})-Q^{i_{1:j-1}}_{\rho^e_{\mu,P,\bar{\pi}},r^e_{\mu,P,\bar{\pi}},\bar{\pi}}(s,a_{i_{1:j-1}})\right]\\
&=\mathbb{E}_{s\sim\eta^e_{\mu,P,\bar{\pi}},a_{i_{1:j-1}}\sim\pi'_{i_{1:j-1}}}\left[Q^{i_{1:j-1}}_{\rho^e_{\mu,P,\bar{\pi}},r^e_{\mu,P,\bar{\pi}},\bar{\pi}}(s,a_{i_{1:j-1}})-Q^{i_{1:j-1}}_{\rho^e_{\mu,P,\bar{\pi}},r^e_{\mu,P,\bar{\pi}},\bar{\pi}}(s,a_{i_{1:j-1}})\right]\\
&=0,
\end{aligned}
$$

and Proposition 8, we have

$$
L^{i_{1:j}}_{\eta^e_{\mu,P,\bar{\pi}},\rho^e_{\mu,P,\bar{\pi}},r^e_{\mu,P,\bar{\pi}},\bar{\pi}}\left(\pi'_{i_{1:j-1}},\pi'_{i_j}\right)-C_e\mathrm{D}^{\max}_{\mathrm{KL}}\left(\pi_{i_j},\bar{\pi}_{i_j}\right)\geq -C_e\frac{n_{i_j}\log n_{i_j}}{n_{i_j}-1}\delta_{i_j}.
$$

Hence,

$$
J(\pi'_e|\mu_e,P_e,r_e)\geq J(\pi_e|\mu_e,P_e,r_e)-(A_e+B_e)R_e,
$$

where,

$$
A_e=\frac{8\gamma}{(1-\gamma)^3}\left(m_e\mathcal{F}(\phi_e)+\sum_{i\in e}\frac{n_i\log n_i}{n_i-1}\delta_i\right),
$$

and

$$
B_e=\frac{2}{(1-\gamma)^2}\left(m_e\epsilon_r(e)+\sum_{j=1}^{m_e}(2j-1)\delta_{i_j}\right).
$$

□

A.6 Theorem 2 and Corollary 2

In this subsection, we provide the complete proofs of Theorem 2 and Corollary 2. We begin with restating Theorem 2 and Corollary 2 in a more complete form.

Theorem 4 (Theorem 2 (Restate))**.** *Under the same conditions and notations as Theorem 3, the global expected return satisfies*

$$
\left|J(\bar{\pi}|\mu,P,r)-\sum_{e\in\mathcal{E}}J(\pi_e|\mu_e,P_e,r_e)\right|\leq\sum_{e\in\mathcal{E}}\left(\frac{2\mathcal{F}_{1+\epsilon_p(e)}(\phi_e)}{(1-\gamma)^2}+\frac{3\epsilon_r(e)-2\epsilon_p(e)}{1-\gamma}\right)R_e.
$$

Corollary 6 (Corollary 2 (Restate))**.** *Suppose the update order of every hyperedge is drawn uniformly from Sym(e), as required by HATRL. Then HygRL with HATRL embedded yields*

$$
J(\bar{\pi}'|\mu,P,r)\geq J(\bar{\pi}|\mu,P,r)-\sum_{e\in\mathcal{E}}M_eR_e,
$$

where

$$
M_e=\frac{3\epsilon_r(e)-2\epsilon_p(e)}{1-\gamma}+\frac{2}{(1-\gamma)^2}(m_e\psi_e+\mathcal{F}_{1+\epsilon_p(e)}(\phi_e))+\frac{8\gamma}{(1-\gamma)^3}\left(m_e\mathcal{F}(\phi_e)+\hat{\delta}_e\right),
$$

and

$$
R_e=\sum_{i\in e}|w_{i,e}|R_i,\quad \phi_e=\epsilon_p(e)+\sum_{i\in e}\delta_i,\quad \psi_e=\epsilon_r(e)+\sum_{i\in e}\delta_i,\quad \hat{\delta}_e:=\sum_{i\in e}\frac{n_i\log n_i}{n_i-1}\delta_i.
$$

Scaling of M_e. Let $n_{\max,e}=\max_{i\in e}n_i$ denote the largest number of hyperedges sharing an agent of e. Each of the three terms of M_e is then of a simple order. The first, $\frac{3\epsilon_r(e)-2\epsilon_p(e)}{1-\gamma}$, does not involve m_e at all. In the second, the factor $m_e\psi_e$ is $O(m_e)$ because $\psi_e=\epsilon_r(e)+\sum_{i\in e}\delta_i$ collects

one term per agent of e, while $\mathcal{F}_{1+\epsilon_p(e)}(\phi_e) \le -(1+\epsilon_p(e))\log\gamma$ is bounded. The third contains $m_e\mathcal{F}(\phi_e) = O(m_e)$ together with

$$\hat{\delta}_e = \sum_{i\in e} \frac{n_i \log n_i}{n_i - 1}\delta_i \le m_e \max_{i\in e} \frac{n_i \log n_i}{n_i - 1}\delta_i = O(m_e \log n_{\max,e}),$$

where we used $\delta_i \le 1$ and $n_i \le n_{\max,e}$. Hence, at a fixed error profile,

$$M_e = O(m_e \log n_{\max,e}).$$

Thus the guarantee degrades only logarithmically with the overlap of hyperedges, and not at all with the total number of agents.

Proof of Theorem 4. Notice that

$$r(s,a) = \sum_{e\in\mathcal{E}} \tilde{r}_e(s,a).$$

We have

$$\begin{aligned} J(\bar{\pi}|\mu,P,r) &= \int_{z\in\mathcal{Z}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z) r(z) \\ &= \int_{z\in\mathcal{Z}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z) \sum_{e\in\mathcal{E}} \tilde{r}_e(z) \\ &= \sum_{e\in\mathcal{E}} \int_{z\in\mathcal{Z}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z)\tilde{r}_e(z) \\ &= \sum_{e\in\mathcal{E}} \int_{z_e\in\mathcal{Z}_e} \int_{z_{-e}\in\mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e,\mathrm{d}z_{-e})\tilde{r}_e(z_e,z_{-e}), \end{aligned}$$

$$J(\pi_e|\mu_e,P_e,r_e) = \int_{z_e\in\mathcal{Z}_e} \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) r_e(z_e).$$

Let

$$\begin{aligned} \Xi^e &= \left| \int_{z_e\in\mathcal{Z}_e} \int_{z_{-e}\in\mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e,\mathrm{d}z_{-e})\tilde{r}_e(z_e,z_{-e}) - \int_{z_e\in\mathcal{Z}_e} \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) r_e(z_e) \right| \\ &\le \underbrace{\left| \int_{z_e\in\mathcal{Z}_e} \int_{z_{-e}\in\mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e,\mathrm{d}z_{-e})\tilde{r}_e(z_e,z_{-e}) - \int_{z_e\in\mathcal{Z}_e} \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) r^e_{\mu,P,\bar{\pi}}(z_e) \right|}_{I_1} \\ &\quad + \underbrace{\left| \int_{z_e\in\mathcal{Z}_e} \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) r^e_{\mu,P,\bar{\pi}}(z_e) - \int_{z_e\in\mathcal{Z}_e} \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) r_e(z_e) \right|}_{I_2} \end{aligned}$$

$$
\begin{aligned}
I_1 \quad = \quad & \left| \int_{z_e \in \mathcal{Z}_e} \int_{z_{-e} \in \mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}) \tilde{r}_e(z_e, z_{-e}) \right. \\
& \left. - \int_{z_e \in \mathcal{Z}_e} (1-\gamma) \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) \int_{z_{-e} \in \mathcal{Z}_{-e}} \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e}) \tilde{r}_e(z_e, z_{-e}) \right| \\
\leq \quad & \int_{z_e \in \mathcal{Z}_e} \int_{z_{-e} \in \mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}) \left| \tilde{r}_e(z_e, z_{-e}) - r_e(z_e) \right| \\
& + \left| \int_{z_e \in \mathcal{Z}_e} \int_{z_{-e} \in \mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}) r_e(z_e) \right. \\
& \left. - \int_{z_e \in \mathcal{Z}_e} \int_{z_{-e} \in \mathcal{Z}_{-e}} (1-\gamma) \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e}) r_e(z_e) \right| \\
& + \int_{z_e \in \mathcal{Z}_e} \int_{z_{-e} \in \mathcal{Z}_{-e}} (1-\gamma) \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e}) \left| r_e(z_e) - \tilde{r}_e(z_e, z_{-e}) \right| \\
\overset{\text{Definition 1}}{\leq} \quad & \epsilon_r(e) \int_{z_e \in \mathcal{Z}_e} \int_{z_{-e} \in \mathcal{Z}_{-e}} \rho_{\mu,P,\bar{\pi}}(\mathrm{d}z_e, \mathrm{d}z_{-e}) \\
& + \epsilon_r(e) \int_{z_e \in \mathcal{Z}_e} \int_{z_{-e} \in \mathcal{Z}_{-e}} (1-\gamma) \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) \rho^{-e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_{-e}) \\
& + \left| \int_{z_e \in \mathcal{Z}_e} \rho^{e}_{\mu,P,\bar{\pi}}(\mathrm{d}z_e) r_e(z_e) - \int_{z_e \in \mathcal{Z}_e} \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) r_e(z_e) \right| \\
\overset{\text{Proposition 1}}{\leq} \quad & \frac{2\epsilon_r(e) R_e}{1-\gamma} + 2R_e \left\| \rho^{e}_{\mu,P,\bar{\pi}} - \rho_{\mu_e,P_e,\pi_e} \right\|_{\mathrm{TV}} \\
\overset{\text{Lemma 7}}{\leq} \quad & \left(\frac{2\epsilon_r(e)}{1-\gamma} + \frac{2}{(1-\gamma)^2} \mathcal{F}_{1+\epsilon_p(e)}(\phi_e) - \frac{2\epsilon_p(e)}{1-\gamma} \right) R_e
\end{aligned}
$$

$$
\begin{aligned}
I_2 \quad \leq \quad & \int_{z_e \in \mathcal{Z}_e} \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) \left| r_e(z_e) - r^{e}_{\mu,P,\bar{\pi}}(z_e) \right| \\
\overset{\text{Lemma 9}}{\leq} \quad & \epsilon_r(e) R_e \int_{z_e \in \mathcal{Z}_e} \rho_{\mu_e,P_e,\pi_e}(\mathrm{d}z_e) \\
= \quad & \frac{1}{1-\gamma} \epsilon_r(e) R_e.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
& \left| J(\bar{\pi}|\mu, P, r) - \sum_{e \in \mathcal{E}} J(\pi_e|\mu_e, P_e, r_e) \right| \\
& \leq \sum_{e \in \mathcal{E}} \int_{z_e \in \mathcal{Z}_e} |\Xi^e(z_e)| \\
& \leq \sum_{e \in \mathcal{E}} \left(\frac{2\mathcal{F}_{1+\epsilon_p(e)}(\phi_e)}{(1-\gamma)^2} + \frac{3\epsilon_r(e) - 2\epsilon_p(e)}{1-\gamma} \right) R_e.
\end{aligned}
$$

□

Since the update order is drawn uniformly, Corollary 5 can be averaged over $\mathrm{Sym}(e)$ as follows.

Corollary 7. *Under the conditions of Definition 1 and Assumption 1, for each hyperedge $e \in \mathcal{E}$, if $\{\pi_{i,e}\}_{i \in e}$ is updated by HygRL with HATRL embedded, using an update order drawn uniformly from $\mathrm{Sym}(e)$, then, **in expectation**, the following performance inequality holds:*

$$
J(\pi'_e|\mu_e, P_e, r_e) \geq J(\pi_e|\mu_e, P_e, r_e) - R_e D_e,
$$

where $D_e = A_e + B_e$,

$$A_e = \frac{8\gamma}{(1-\gamma)^3}\left(m_e\mathcal{F}(\phi_e) + \sum_{i\in e}\frac{n_i \log n_i}{n_i - 1}\delta_i\right),$$

$$B_e = \frac{2m_e}{(1-\gamma)^2}\psi_e,$$

and

$$R_e = \sum_{i\in e}|w_{i,e}|R_i, \quad \phi_e = \epsilon_p(e) + \sum_{i\in e}\delta_i, \quad \psi_e = \epsilon_r(e) + \sum_{i\in e}\delta_i.$$

Proof. The only term affected by the update order in Ineq. (5) is $\sum_{j=1}^{m_e}(2j-1)\delta_{i_j}$. When randomly select $\langle i_1, ..., i_{m_e}\rangle \in \text{Sym}(e)$, the expectation of this term is

$$\mathbb{E}\left[\sum_{j=1}^{m_e}(2j-1)\delta_{i_j}\right] = \sum_{j=1}^{m_e}(2j-1)\mathbb{E}\left[\delta_{i_j}\right].$$

Notice that for each $k \in \{1, \ldots, m_e\}$, $\Pr(\delta_{i_j} = \delta_k) = \frac{1}{m_e}$. Thus,

$$\mathbb{E}\left[\delta_{i_j}\right] = \frac{1}{m_e}\sum_{i\in e}\delta_i,$$

and

$$\mathbb{E}\left[\sum_{j=1}^{m_e}(2j-1)\delta_{i_j}\right] = m_e\sum_{i\in e}\delta_i.$$

Hence, the expected local performance bound is

$$\begin{aligned} J(\pi_e'|\mu_e, P_e, r_e) &\geq J(\pi_e|\mu_e, P_e, r_e) - A_eR_e - \frac{2}{(1-\gamma)^2}\left(m_e\epsilon_r(e) + m_e\sum_{i\in e}\delta_i\right)R_e \\ &= J(\pi_e|\mu_e, P_e, r_e) - A_eR_e - \frac{2m_e}{(1-\gamma)^2}\psi_eR_e. \end{aligned}$$

□

Corollary 6 can be obtained by combining Theorem 4 and Corollary 5.

Proof of Corollary 6.

$$\begin{aligned} J(\bar{\pi}'|\mu, P, r) - J(\bar{\pi}|\mu, P, r) = &\underbrace{J(\bar{\pi}'|\mu, P, r) - \sum_{e\in\mathcal{E}}J(\pi_e'|\mu_e, P_e, r_e)}_{I_1} \\ &+ \sum_{e\in\mathcal{E}}\underbrace{\left(J(\pi_e'|\mu_e, P_e, r_e) - J(\pi_e|\mu_e, P_e, r_e)\right)}_{I_e} \\ &+ \underbrace{\sum_{e\in\mathcal{E}}J(\pi_e|\mu_e, P_e, r_e) - J(\bar{\pi}|\mu, P, r)}_{I_2} \end{aligned}$$

Applying Theorem 4, we have

$$I_1, I_2 \geq -R_e\sum_{e\in\mathcal{E}}\left(\frac{2\mathcal{F}_{1+\epsilon_p(e)}(\phi_e)}{(1-\gamma)^2} + \frac{3\epsilon_r(e) - 2\epsilon_p(e)}{1-\gamma}\right).$$

For each $e \in \mathcal{E}$, Corollary 7 shows

$$I_e \geq -(A_e + B_e)R_e,$$

where

$$J(\pi_e'|\mu_e, P_e, r_e) \geq J(\pi_e|\mu_e, P_e, r_e) - (A_e + B_e)R_e,$$

where,

$$A_e = \frac{8\gamma}{(1-\gamma)^3}\left(m_e\mathcal{F}(\phi_e) + \sum_{i\in e}\frac{n_i \log n_i}{n_i - 1}\delta_i\right),$$

$$B_e = \frac{2m_e}{(1-\gamma)^2}\psi_e.$$

Let $\hat{\delta}_e := \sum_{i\in e}\frac{n_i \log n_i}{n_i - 1}\delta_i$,

$$\begin{aligned} M_e :=& \frac{4\mathcal{F}_{1+\epsilon_p(e)}(\phi_e)}{(1-\gamma)^2} + \frac{6\epsilon_r(e) - 4\epsilon_p(e)}{1-\gamma} + A_e + B_e \\ =& \frac{3\epsilon_r(e) - 2\epsilon_p(e)}{1-\gamma} + \frac{2}{(1-\gamma)^2}(m_e\psi_e + \mathcal{F}_{1+\epsilon_p(e)}(\phi_e)) + \frac{8\gamma}{(1-\gamma)^3}\left(m_e\mathcal{F}(\phi_e) + \hat{\delta}_e\right), \end{aligned}$$

then

$$J(\bar{\pi}'|\mu, P, r) \geq J(\bar{\pi}|\mu, P, r) - R_e\sum_{e\in\mathcal{E}} M_e.$$

□

Recovering CTDE in the trivial case. When $\mathcal{E} = \{\mathcal{N}\}$, for the unique hyperedge $e = \mathcal{N}$, Definition 1 holds with $\mu_e = \mu$, $P_e = P$ and $r_e = r$, yielding $\epsilon_p(e) = 0$ and $\epsilon_r(e) = 0$. Assumption 1 holds with $\pi_{i,e} = \bar{\pi}_i$, yielding $\delta_i = 0$ for all $i \in \mathcal{N}$. Substituting these into the definition of M_e in Corollary 6 gives $M_e = 0$. Hence $\sum_{e\in\mathcal{E}} M_e = 0$, and Corollary 6 becomes exactly the HATRL guarantee.

B Hypergraph Construction Details

All the procedures discussed below are concerned only with the choice of $\mathcal{H}$: they are not part of the algorithm of Section 3, which runs on any hypergraph and does not depend on how that hypergraph was obtained. Their role is purely to indicate how a hypergraph with a small error profile can be found with respect to a given system. In this section, we briefly discuss methods for constructing hypergraphs. The goal is to construct a hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{E})$, together with a surrogate kernel P_e and a surrogate reward r_e for each hyperedge $e \in \mathcal{E}$, such that the associated errors $\epsilon_p(e)$ and $\epsilon_r(e)$ of Definition 1 are small. We present two complementary approaches: a theoretical optimization-based method (Algorithm 2) that guarantees minimal approximation error, and a practical method based on Markov random field decomposition and structured mean-field inference that is computationally tractable for large-scale systems. Since the local initial distribution is the exact marginal of μ, only the kernel and the reward require approximation, and the two are handled separately: the dynamics error $\epsilon_p(e)$ arises from approximating a distribution, whereas the reward error $\epsilon_r(e)$ arises from approximating a deterministic function.

B.1 Theoretical Construction via Optimization

The first approach is purely theoretical and serves as a proof of concept: for any transition kernel over the agent set, there exists a hypergraph decomposition with bounded error.

Definition 1 (i) constrains, for every state-action pair z and every hyperedge e, the integral of the next-state distribution $P(\cdot \mid z)$ over the coordinates outside e. For a fixed z_e, the object to be approximated is therefore the family

$$\left\{\int_{s'_{-e}\in\mathcal{S}_{-e}} P(\cdot, \mathrm{d}s'_{-e} \mid z_e, z_{-e})\right\}_{z_{-e}\in\mathcal{Z}_{-e}} \tag{19}$$

of distributions on $\mathcal{S}_e$, parameterized by z_{-e}. Since the surrogate may depend on z_e but not on z_{-e}, we seek a kernel on $\mathcal{S}_e$: for each fixed z_e, one distribution must cover the whole family at once.

Define the worst-case approximation error of a candidate distribution $D \in \Delta(\mathcal{S}_e)$ at z_e as

$$F_e(z_e, D) = \sup_{z_{-e} \in \mathcal{Z}_{-e}} \left\| D - \int_{s'_{-e} \in \mathcal{S}_{-e}} P(\cdot, \mathrm{d}s'_{-e} \mid z_e, z_{-e}) \right\|_{\mathrm{TV}}, \tag{20}$$

which measures how well D approximates the projection of the transition kernel onto the coordinates of e, uniformly over the complement. By construction F_e depends on z only through z_e, because the surrogate is allowed to depend on z_e but not on z_{-e}.

Since the state spaces are compact Polish, $\Delta(\mathcal{S}_e)$ is compact under the weak-* topology by Prokhorov's theorem. For fixed z_e, the functional $D \mapsto F_e(z_e, D)$ is lower semicontinuous, hence by Weierstrass's theorem the minimizer

$$Q_{e,z_e} \in \arg\min_{D \in \Delta(\mathcal{S}_e)} F_e(z_e, D) \tag{21}$$

exists. Solving this problem separately at each z_e defines the surrogate kernel $P_e(\cdot \mid z_e) := Q_{e,z_e}$ required by Definition 1, whose approximation error is the worst case over z_e:

$$\varepsilon_e = \sup_{z_e \in \mathcal{Z}_e} F_e\big(z_e, Q_{e,z_e}\big). \tag{22}$$

Comparing Eq. equation 22 with Definition 1 (i), the minimizer Q_{e,z_e} is the surrogate kernel $P_e(\cdot \mid z_e)$ and ε_e is the corresponding $\epsilon_p(e)$.

With these definitions, we can construct a hypergraph by selecting a cover of $\mathcal{N}$ from subsets with small ϵ. Algorithm 2 formalizes this procedure.

Algorithm 2: Theoretical Hypergraph Construction

Require: Agent set $\mathcal{N}$, transition kernel P, threshold $C \in (0, 1)$
Ensure: Hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{E})$

1: $\mathcal{L} \leftarrow \emptyset$
2: **for** each $e \subseteq \mathcal{N}$ **do**
3: **if** $\varepsilon_e \leq C$ **then**
4: Add e to $\mathcal{L}$
5: **end if**
6: **end for**
7: $\mathcal{R} \leftarrow \emptyset$
8: **for** each $\mathcal{U} \subseteq \mathcal{L}$ **do**
9: **if** $\bigcup_{e \in \mathcal{U}} e = \mathcal{N}$ **then**
10: Add $\mathcal{U}$ to $\mathcal{R}$
11: **end if**
12: **end for**
13: $\mathcal{E} \leftarrow \arg\min_{\mathcal{U} \in \mathcal{R}} \left(\max_{e \in \mathcal{U}} |e|,\ |\mathcal{U}| \right)$
14: **return** $\mathcal{H} = (\mathcal{N}, \mathcal{E})$

Algorithm 2 first collects all subsets $e \subseteq \mathcal{N}$ whose approximation error ϵ_e is within the prescribed threshold C (lines 2-6). It then finds all subfamilies of these subsets that cover $\mathcal{N}$ (lines 8-12), and finally selects, among those covers, one whose largest member is as small as possible, breaking ties by the number of hyperedges (line 13). Since $\epsilon_{\mathcal{N}} = 0$ (the full set has zero approximation error), the feasible set $\mathcal{R}$ is nonempty, and the algorithm always returns a valid hypergraph. By construction, for each hyperedge $e \in \mathcal{E}$ we have $\epsilon_p(e) \leq C$, so Definition 1 (i) holds with uniform error bound C.

The threshold C controls the resulting decomposition: a smaller C restricts the admissible subsets and therefore forces the cover to use larger or more numerous hyperedges, while a larger C admits smaller hyperedges at the price of a larger per-hyperedge error.

Two properties of this construction are worth recording. First, the objective is minimized separately at each z_e, so the resulting P_e is a genuine kernel on $\mathcal{Z}_e$ rather than a single distribution that must serve all z_e. Second, the hypergraph produced by the algorithm is obtained by thresholding ϵ_e at a common level C; since Definition 1 permits an error profile that varies across hyperedges, the threshold may also be chosen per subset, which allows the approximation budget to be spent where it is cheapest.

While Algorithm 2 provides a clean theoretical guarantee, its computational cost grows exponentially with the number of agents, making it impractical for large-scale systems. The following section presents a more practical alternative.

B.2 Practical Construction via Markov Random Field Decomposition

For real-world applications, we often have prior knowledge about the dependency structure among agents—for instance, a communication graph or physical network topology. In such cases, we can exploit this structure through Markov Random Field (MRF) decomposition.

For each state-action pair z, assume that the next-state distribution $P(\cdot \mid z)$ admits an MRF representation with graph structure G:

$$p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \phi_C(x_C), \qquad Z = \sum_x \prod_{C \in \mathcal{C}} \phi_C(x_C), \tag{23}$$

where $\mathcal{C}$ is the set of cliques of G, and ϕ_C are positive potential functions defined on each clique.

For a clique C, consider an approximating distribution $\tilde{q}_C(x) = q_C(x_C) q_{-C}(x_{-C})$ that factorizes as the product of a distribution on the clique and a distribution on its complement; here q_{-C} is an auxiliary factor, introduced only to make the objective tractable, and the quantity of interest is the marginal q_C. Define the variational free energy

$$\mathcal{L}(\tilde{q}_C) = \sum_x \tilde{q}_C(x) \log \frac{\tilde{p}(x)}{\tilde{q}_C(x)} = \mathbb{E}_{x \sim \tilde{q}_C}[\log \tilde{p}(x)] + H(q_C) + H(q_{-C}), \tag{24}$$

where $\tilde{p}(x) = \prod_{C' \in \mathcal{C}} \phi_{C'}(x_{C'})$ is the unnormalized density and $\tilde{q}_C(x) = q_C(x_C) q_{-C}(x_{-C})$. The KL divergence between $\tilde{q}_C$ and p satisfies

$$\log Z = \mathcal{L}(\tilde{q}_C) + \mathrm{D}_{\mathrm{KL}}(\tilde{q}_C \| p). \tag{25}$$

Maximizing $\mathcal{L}(\tilde{q}_C)$ over q_C and q_{-C} yields the best factorized approximation to p. For fixed q_{-C}, the optimal q_C is given by the structured mean-field update

$$q_C(x_C) \propto \phi_C(x_C) \prod_{C' \cap C \neq \emptyset,\ C' \neq C} \exp\left(\mathbb{E}_{q_{-C}}[\log \phi_{C'} \mid x_{C \cap C'}]\right). \tag{26}$$

This equation is derived from standard variational inference results (see Section 5.5 of Wainwright & Jordan (2008)) and reveals a direct proportionality between the optimal q_C and the local factor ϕ_C, coupled with neighboring factors through expectations with respect to the complement distribution.

By solving for q_C for each clique C, typically via fixed-point iteration of Eq. (26), we obtain a set of local distributions $\{q_C\}_{C \in \mathcal{C}}$. Each q_C approximates the true marginal p_C with the following TV distance bound:

$$\|q_C - p_C\|_{\mathrm{TV}} \leq \sqrt{1 - \exp(-\mathrm{D}_{\mathrm{KL}}(q_C \| p_C))} \leq \sqrt{1 - \exp(-\mathrm{D}_{\mathrm{KL}}(\tilde{q}_C \| p))} \leq \sqrt{1 - \frac{\exp(\mathcal{L}(\tilde{q}_C))}{Z}}, \tag{27}$$

where the first inequality is the Bretagnolle-Huber inequality, the second follows from the chain rule of KL divergence, and the third uses $\log Z = \mathcal{L}(\tilde{q}_C) + \mathrm{D}_{\mathrm{KL}}(\tilde{q}_C \| p)$.

Applying this decomposition procedure to each transition kernel $P(z)$ (for all $z \in \mathcal{Z}$) yields a collection of local kernels that approximate the global dynamics, while the local initial distribution is obtained exactly as the marginal $\mu_e = \int_{s_{-e} \in \mathcal{S}_{-e}} \mu(\cdot, \mathrm{d}s_{-e})$. The construction is carried out separately at each z: solving Eq. (26) for $P(\cdot \mid z)$ produces $q_C^{(z)}$, and setting $P_e(\cdot \mid z_e) := q_e^{(z)}$ defines a surrogate kernel in the sense of Definition 1, whose approximation error is the worst case over z. The cliques $\mathcal{C}$ of the MRF graph directly define the hyperedge set $\mathcal{E} = \mathcal{C}$, and the bounds in Eq. (27) provide, for each z, a bound on the deviation between the surrogate kernel P_e and the marginal of $P(\cdot|z)$ on $\mathcal{S}_e$; taking the worst case over z yields the approximation errors $\epsilon_p(e)$ required by Definition 1.

B.3 Constructing the Surrogate Reward

The dynamics are approximated in the preceding subsection by solving for a best local distribution. The reward requires a different treatment, because $\tilde{r}_e$ is not a distribution and cannot be handled by mean-field inference. We first present a general fact about best uniform approximations, and then specialize it to Definition 1 (ii).

Lemma 10. *Given functions $h_1(x,y),\dots,h_n(x,y)$ and weights $a_1,\dots,a_n$, each h_i is bounded by M_i. Consider*

$$g(x) = \arg\min_{p} \sup_{y} \left| p - \sum_{i=1}^{n} a_i h_i(x,y) \right|.$$

Then $g = \frac{1}{2}\left(\inf_y \sum_{i=1}^n a_i h_i(x,y) + \sup_y \sum_{i=1}^n a_i h_i(x,y)\right)$ is given by the midpoint of the range of the sum, and is bounded by $\sum_{i=1}^n |a_i| M_i$.

Proof. Select a fixed x, and let $H(y) = \sum_{i=1}^n a_i h_i(x,y)$. Then H is bounded by $\sum_{i=1}^n |a_i| M_i$; let $L_x = \inf_y H(y)$ and $U_x = \sup_y H(y)$. For any c,

$$\sup_y |c - H(y)| \geq \frac{U_x - L_x}{2},$$

with equality at $c = \frac{U_x + L_x}{2}$, so

$$g(x) = \frac{U_x + L_x}{2},$$

which attains the best error. Since $|H(y)| \leq \sum_{i=1}^n |a_i| M_i$, we obtain

$$|g(x)| \leq \frac{|U_x| + |L_x|}{2} \leq \sum_{i=1}^{n} |a_i| M_i.$$

□

Applying Lemma 10 to the local reward is now immediate. Take $x = z_e$, $y = z_{-e}$, $h_i = r_i$ and $a_i = w_{i,e}$, so that the sum in the lemma is $\tilde{r}_e(z_e, z_{-e}) = \sum_{i \in e} w_{i,e} r_i(z_e, z_{-e})$ and the bound $M_i = R_i$ holds by the boundedness of the rewards. The objective is exactly the worst-case deviation in Definition 1 (ii), so the lemma yields the optimal surrogate together with its error. Writing

$$L_e(z_e) = \inf_{z_{-e} \in \mathcal{Z}_{-e}} \tilde{r}_e(z_e, z_{-e}), \qquad U_e(z_e) = \sup_{z_{-e} \in \mathcal{Z}_{-e}} \tilde{r}_e(z_e, z_{-e}), \tag{28}$$

so that $[L_e(z_e), U_e(z_e)]$ is the range of the local reward over all configurations of the complement compatible with z_e, the surrogate produced by the lemma and the error it attains are

$$r_e(z_e) = \frac{U_e(z_e) + L_e(z_e)}{2}, \qquad \sup_{z_{-e} \in \mathcal{Z}_{-e}} \left| \tilde{r}_e(z_e, z_{-e}) - r_e(z_e) \right| = \frac{U_e(z_e) - L_e(z_e)}{2}, \tag{29}$$

so that r_e attains the smallest possible relative error

$$\epsilon_r(e) = \frac{1}{2R_e} \sup_{z_e \in \mathcal{Z}_e} \left(U_e(z_e) - L_e(z_e) \right) \tag{30}$$

in the sense of Definition 1 (ii). This error is exactly half the spread of $\tilde{r}_e$ over the complement, normalized by R_e: it vanishes when the reward depends on z_e alone and grows as the agents in e become less able to determine $\tilde{r}_e$ on their own. Finally, the lemma gives $|r_e(z_e)| \leq \sum_{i \in e} |w_{i,e}| R_i = R_e$, which is finite by the boundedness of the rewards.

Illustrative Example. To concretely illustrate the mean-field update in Eq. (26), consider a simple graphical model with two overlapping cliques: $C_1 = \{1,2\}$ and $C_2 = \{2,3\}$. The unnormalized density is

$$\tilde{p}(x) = \phi_{12}(x_1, x_2)\, \phi_{23}(x_2, x_3). \tag{31}$$

Suppose we wish to approximate the marginal on clique $C_1 = \{1, 2\}$. Fixing $q_{-C_1} = q_{C_2}$, the structured mean-field update in Eq. (26) gives

$$q_{C_1}(x_1, x_2) \propto \phi_{12}(x_1, x_2) \exp\left(\mathbb{E}_{q_{C_2}(x_2', x_3)}\left[\log \phi_{23}(x_2', x_3) \,\middle|\, x_2\right]\right), \tag{32}$$

where the conditional expectation is taken with respect to x_2' and x_3 under q_{C_2}, conditioned on $x_2' = x_2$. This update has a clear interpretation: the exponential term acts as an *effective potential* on x_2 that summarizes the influence of the neighboring clique C_2 through the shared variable x_2. Thus, rather than having to account for the full joint distribution over $\{1, 2, 3\}$, the local update on C_1 only needs to incorporate a summarized message from C_2, which is analogous to belief propagation but applied in the variational inference setting. Repeating this procedure for C_2 symmetrically yields a pair of local distributions that approximate the global joint distribution with bounded error as in Eq. (27).

B.4 Discussion and Practical Considerations

The two methods presented above serve complementary roles. Algorithm 2 is theoretically appealing: it guarantees the existence of a hypergraph with controlled error without requiring any structural assumptions, but is computationally prohibitive for large n. The MRF-based method is practical when the system admits a graphical model structure, which is common in networked multi-agent systems where interactions are spatially or functionally localized.

In our experiments, we adopt the MRF-based philosophy. For wireless communication, hyperedges are defined by access points (users sharing the same access point form a natural factor). For power grid load balancing, hyperedges are constructed from maximal cliques of the physical network graph. For traffic signal control, hyperedges are constructed from maximal cliques of the augmented road network graph. These constructions align with the MRF decomposition principle, as the physical topology directly induces the dependency structure among agents. The experimental results in Sections 5 and C demonstrate that these practically constructed hypergraphs yield strong empirical performance, validating the effectiveness of the MRF-based approach even when full variational inference is not explicitly performed.

We emphasize that the theoretical analysis in Section 4 holds for *any* hypergraphic environment in the sense of Definition 1, with the approximation error entering the bounds of Theorems 1 and 2 explicitly. This separation between construction and analysis gives practitioners the flexibility to incorporate domain knowledge when designing the hypergraph for their specific application.

C Complete Experimental Results

In this section, we describe the experimental setup and present the complete experimental results. To systematically evaluate scalability, we adopt a suite of synthetic benchmark topologies commonly used in networked multi-agent systems, including regular graphs (e.g., cube and dodecahedral graphs) (Chu et al., 2022), grid graphs (Qu et al., 2022), and random graphs. These are complemented by real-world road networks from LightSim (Su & Deng, 2026). The synthetic topologies span a range of graph properties,degree, diameter, clustering, and planarity,to test algorithm performance under diverse structural conditions.

C.1 Complete Results for Wireless Communication

Environment details. We adopt the wireless communication environment from Qu et al. (2022) with a custom modification. The system consists of n users and m access points (APs). Each user i has access to a subset of APs $Y_i \subseteq \{1, \ldots, m\}$. Users i and j are defined as neighbors if they share at least one accessible AP, i.e., $Y_i \cap Y_j \neq \emptyset$; this neighbor relation defines the interaction graph $\mathcal{G}$.

Each user i maintains a packet queue, represented by a d_i-dimensional binary tuple $s_i = (e_1, \ldots, e_{d_i}) \in \{0, 1\}^{d_i}$, where e_ℓ indicates whether the user has a packet with remaining deadline ℓ. The action space is $\mathcal{A}_i = \{\text{null}\} \cup Y_i$, where "null" means not sending.

At each time step, with probability p_i, user i receives a new packet with an initial deadline d_i. The user may then choose to transmit the earliest packet in its queue to one of its accessible APs, or

Table 3: Complete average performance on Wireless Communication (mean ± std). Best results are underlined; bold indicates within one standard deviation of the best result.

Setup	HygHAPPO	HAPPO	HygMAPPO	MAPPO	IPPO	NPPO
5 agents, MLP	**257.33** ±1.78	**259.62** ±4.92	**257.02** ±1.68	247.42 ±14.41	106.05 ±33.52	69.12 ±33.02
5 agents, GRU	**257.46** ±2.26	**257.27** ±8.64	**257.50** ±0.79	**257.88** ±8.56	94.77 ±44.25	82.29 ±24.21
20 agents, MLP	**690.04** ±29.76	463.10 ±40.56	647.69 ±78.15	515.25 ±47.90	37.39 ±40.07	40.48 ±74.61
20 agents, GRU	**729.45** ±33.94	−389.89 ±46.08	**705.58** ±23.32	−339.66 ±153.29	28.49 ±88.35	78.89 ±98.81
35 agents, MLP	**−17.80** ±57.29	−606.29 ±38.39	**−22.30** ±44.98	−608.31 ±27.06	−1311.70 ±61.68	−1351.10 ±93.99
35 agents, GRU	**−75.80** ±42.90	−654.01 ±44.45	−123.78 ±47.26	−669.38 ±50.37	−1375.09 ±68.06	−1414.99 ±85.14
50 agents, MLP	**1211.31** ±107.17	511.38 ±80.63	**1251.62** ±139.75	488.86 ±50.27	−1131.70 ±122.21	−1036.03 ±86.71
50 agents, GRU	**1211.25** ±165.82	174.68 ±281.26	**1198.69** ±117.83	429.05 ±401.44	−1034.18 ±132.13	−927.97 ±182.90

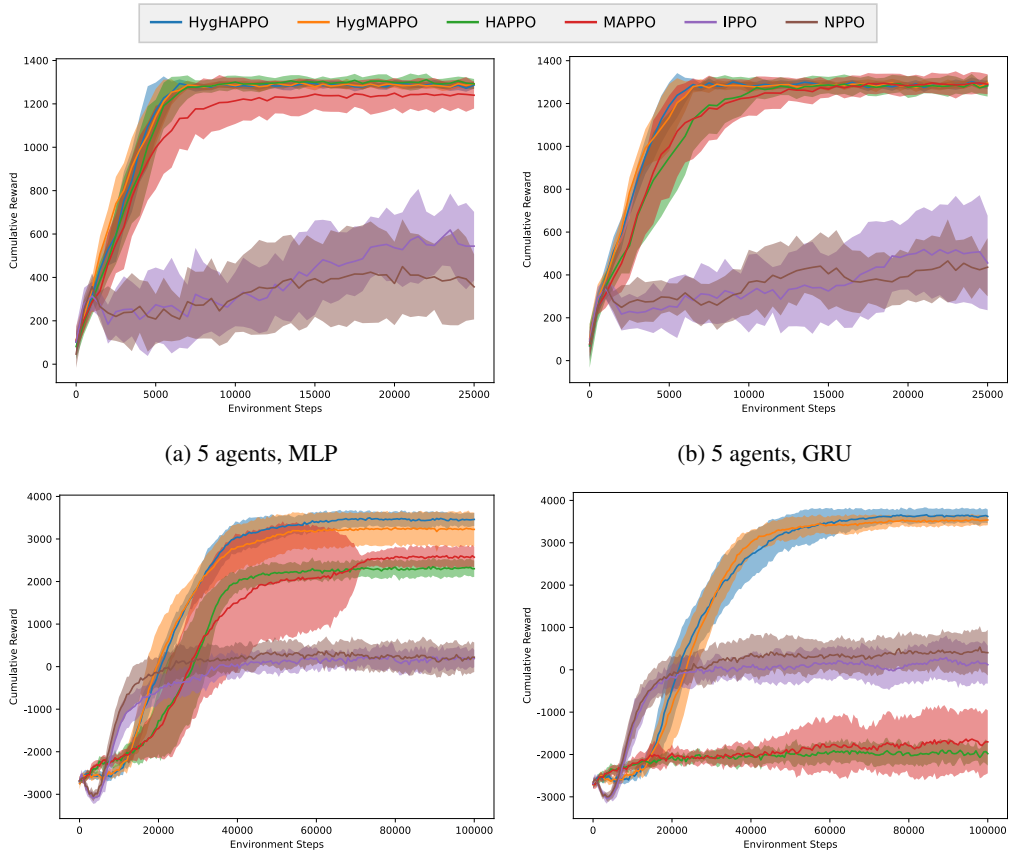
refrain from transmitting. Each AP j has a capacity w_j, specifying the maximum number of packets it can successfully receive at each time step. If user i transmits to AP $y_k \in Y_i$, and the total number of users transmitting to y_k at that time does not exceed its capacity w_k, then the packet is successfully transmitted, yielding a reward of $+1$ and removing the packet from the queue. If the total number of users transmitting to y_k exceeds w_k, a collision occurs and no packet is transmitted. Packets that remain in the queue beyond their deadline are discarded, incurring a reward of -1 for the user.

Graph and hypergraph construction. NPPO uses the interaction graph $\mathcal{G}$ defined above. For HygRL, we construct the hypergraph $\mathcal{H} = (\mathcal{N}, \mathcal{E})$ by taking all maximal cliques of $\mathcal{G}$ as hyperedges. Since each maximal clique corresponds exactly to the set of users sharing a common AP, this construction yields one hyperedge per AP: $\mathcal{E} = \{e_j | j = 1, \ldots, m\}$, where $e_j = \{i \in \mathcal{N} | j \in Y_i\}$. This ensures that the communication requirement of HygRL during training is consistent with that of NPPO: agents within the same hyperedge are exactly those within direct communication range (one-hop neighbors in $\mathcal{G}$). Thus, both algorithms operate under the same communication constraints. Moreover, the hyperedge structure naturally captures the cooperative-competitive relationship among users competing for the same AP, providing a more faithful representation of the underlying coordination structure than the pairwise graph. The specific network topologies used in the experiments are as follows.

- *5 users, 2 APs:* AP_0 covers $\{0, 1, 2\}$, AP_1 covers $\{2, 3, 4\}$, forming a chain of two overlapping cliques.
- *20 users, 12 APs:* Users are placed at the 20 vertices of a regular dodecahedron; each of the 12 pentagonal faces defines an AP covering its 5 vertices. The resulting conflict graph is the dodecahedral graph.
- *35 users, 9 APs and 50 users, 14 APs:* Users are sampled from a planar hexagonal grid via Poisson disk sampling to ensure spatial uniformity; APs are placed at the centers of hexagonal cells, and each AP covers users within a circle. This construction yields locally clustered conflict graphs.

Figure 3 and 4 present the complete set of training curves across all 8 configurations (5/20/35/50 agents × MLP/GRU), and Table 3 reports the corresponding average performance.

Summary of complete results. The complete results confirm the conclusions drawn from the representative results in the main text: (i) for 5 agents, all CTDE-based methods (HygHAPPO, HAPPO, HygMAPPO, MAPPO) perform similarly; (ii) as the number of agents increases to 20, 35, and 50, HygHAPPO and HygMAPPO significantly outperform their CTDE counterparts HAPPO and MAPPO; (iii) IPPO and NPPO converge to suboptimal policies beyond 20 agents, with NPPO showing only marginal improvement over IPPO despite using graph information; and (iv) these



(a) 5 agents, MLP

(b) 5 agents, GRU

(c) 20 agents, MLP

(d) 20 agents, GRU

Figure 3: Training curves on Wireless Communication. Each subfigure shows the cumulative reward over training steps for 5 and 20 agents with MLP and GRU parameterizations. Shaded regions represent twice the standard deviation over 10 random seeds.

trends are consistent across both MLP and GRU parameterizations, demonstrating that the advantage of HygRL is robust to the choice of network architecture.

C.2 Complete Results for Power Grid Load Balancing

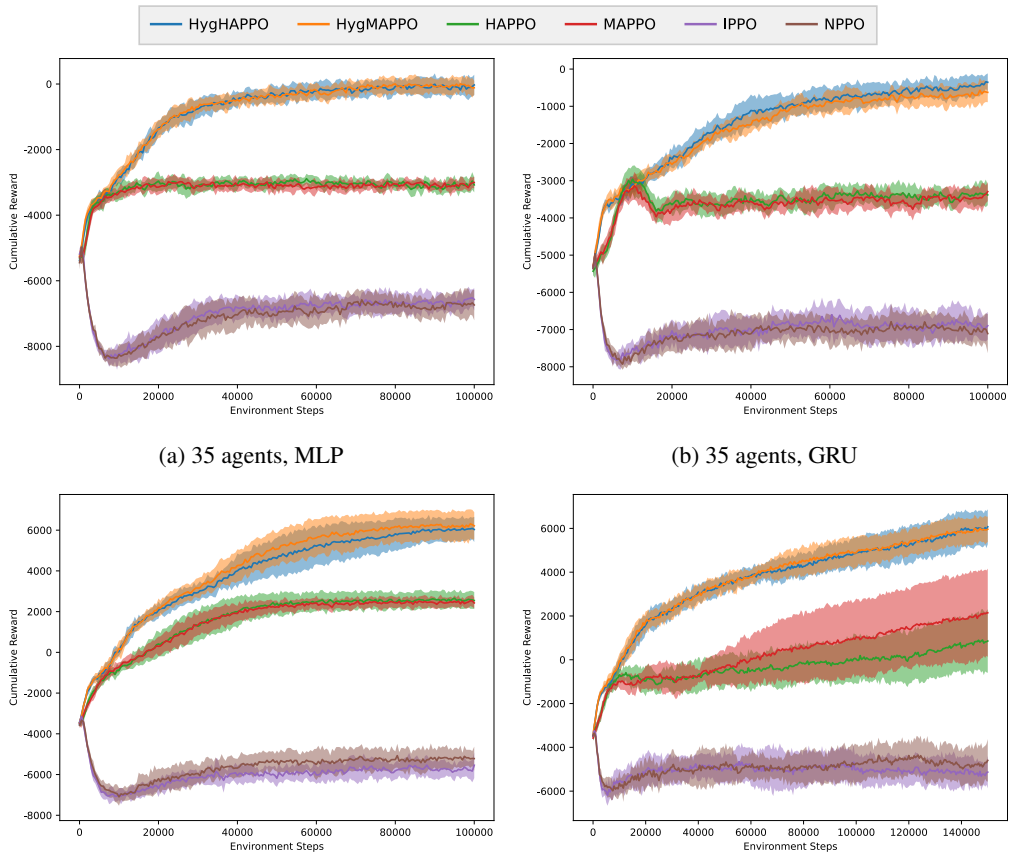
Environment details. We design a power grid load balancing scenario in a power distribution network. The network consists of three types of nodes: `Gate` (grid gateway), `Inverter` (solar inverter), and `Load` (power load). Each node has distinct state and action spaces, making this a heterogeneous multi-agent system. The interaction graph is defined by the physical connectivity of the network.

At each time step:

- `Gate` nodes choose their real power P and reactive power Q outputs, incurring a generation cost proportional to P when $P > 0$.
- `Inverter` nodes choose a reactive-to-active power ratio $r = Q/P$ (i.e., $\tan\varphi$), which determines the conversion from solar input power P_{in} (modeled by an Ornstein–Uhlenbeck process) to real output P_{out} and reactive output Q:

$$P_{\text{out}} = \frac{P_{\text{in}} - P_{\text{loss},0}}{1 + k|r|}, \qquad Q = P_{\text{out}} \cdot r,$$

  where $P_{\text{loss},0}$ and k are node-specific constants. A reward proportional to $P_{\text{out}}/P_{\text{in}}$ is given to incentivize conversion efficiency.



(a) 35 agents, MLP

(b) 35 agents, GRU

(c) 50 agents, MLP

(d) 50 agents, GRU

Figure 4: Training curves on Wireless Communication. Each subfigure shows the cumulative reward over training steps for 35 and 50 agents with MLP and GRU parameterizations. Shaded regions represent twice the standard deviation over 10 random seeds.

- `Load` nodes consume real power according to a sinusoidal daily profile with additive Gaussian noise, and can send advice $(-1, 0, 1)$ to neighboring `Gate` and `Inverter` nodes to influence their decisions. Their reactive power is determined by a time-varying factor derived from a piecewise-constant mean with added noise.

Power and load dynamics. The environment simulates a distribution network with solar inverters, grid gateways, and loads. The power generation and consumption are driven by time-varying stochastic processes: solar input follows an Ornstein–Uhlenbeck process with diurnal stage-dependent parameters, while load real power follows a sinusoidal pattern with additive noise.

Solar generation (PanelPower). The solar input power for each inverter follows a diurnal profile modulated by an OU process. Let x_t denote the cloud cover fraction (between 0 and 1). At each time step, the OU update is:

$$x_{t+\Delta t} = \mathrm{clip}\Big(x_t + \theta\big(\mu(t) - x_t\big)\Delta t + \sigma\sqrt{\Delta t}\,\mathcal{N}(0,1),\ 0,\ 1\Big),$$

where $\mu(t)$ is a piecewise-linear function interpolating between stage-dependent mean values over the day, θ is the mean-reversion rate, and σ is the volatility. The solar altitude angle h is computed from latitude ϕ, solar declination δ, day length T_{day}, and phase offset t_0:

$$\sin h(t) = \sin\phi\sin\delta + \cos\phi\cos\delta\cos\left(\frac{2\pi}{T_{\text{day}}}(t - t_0)\right).$$

The irradiance $I(t)$ is then given by:

$$I(t) = \begin{cases} \sin h(t) \cdot \left(P_{\text{clear}} - C \cdot x_t\right)^{1/\sin h(t)}, & \sin h(t) > 0, \\ 0, & \text{otherwise}, \end{cases}$$

where P_{clear} is the clear-sky transmission factor and C is a cloud attenuation coefficient. Note that x_t is the current cloud cover after the OU update, directly reducing the clear-sky irradiance. The solar input power to the inverter is:

$$P_{\text{in}} = I(t) \cdot S_0 \cdot A + P_{\text{base}},$$

with S_0 the base solar radiance, A the panel area, and P_{base} a constant offset.

Load consumption (LoadPower). Each load node consumes real power following a sinusoidal daily pattern with additive Gaussian noise:

$$P_{\text{load}}(t) = a_0 + a_1 \sin(\omega_1 t - \phi_1) + a_2 \sin(\omega_2 t - \phi_2) + \mathcal{N}(0, \sigma^2),$$

where $\omega_1 = 2\pi/T_{\text{day}}$, $\omega_2 = 2\omega_1$, and $a_0, a_1, a_2, \phi_1, \phi_2, \sigma$ are node-specific constants. The load's reactive power is obtained via a factor $r_{\text{load}} = Q/P_{\text{load}}$. This factor is derived from a stage-dependent mean cosine value $\mu_{\text{pf}}(t)$ (piecewise constant across day stages) with additive Gaussian noise: $\theta \sim \mu_{\text{pf}}(t) + \mathcal{N}(0, \sigma_{\text{pf}}^2)$, clipped to $[-1, 1]$, and then converted to:

$$r_{\text{load}} = \frac{\sqrt{1-\theta^2}}{\theta}.$$

Thus $Q_{\text{load}} = P_{\text{load}} \cdot r_{\text{load}}$.

Power conversion. Given solar input P_{in} and the chosen ratio r (for inverter) or the derived r_{load} (for load), the inverter outputs:

$$P_{\text{out}} = \frac{P_{\text{in}} - P_{\text{loss,0}}}{1 + k|r|}, \qquad Q_{\text{out}} = P_{\text{out}} \cdot r,$$

while loads simply consume:

$$Q_{\text{load}} = P_{\text{load}} \cdot r_{\text{load}}.$$

Gate nodes directly set their real and reactive power outputs (P, Q).

Power flow. Given the net real and reactive power injections $P, Q \in \mathbb{R}^n$ at all nodes, the steady-state voltages are computed via a linearized power flow:

$$V = RP + XQ, \quad \text{where } Z = Y^{-1},\ R = \Re(Z),\ X = \Im(Z),$$

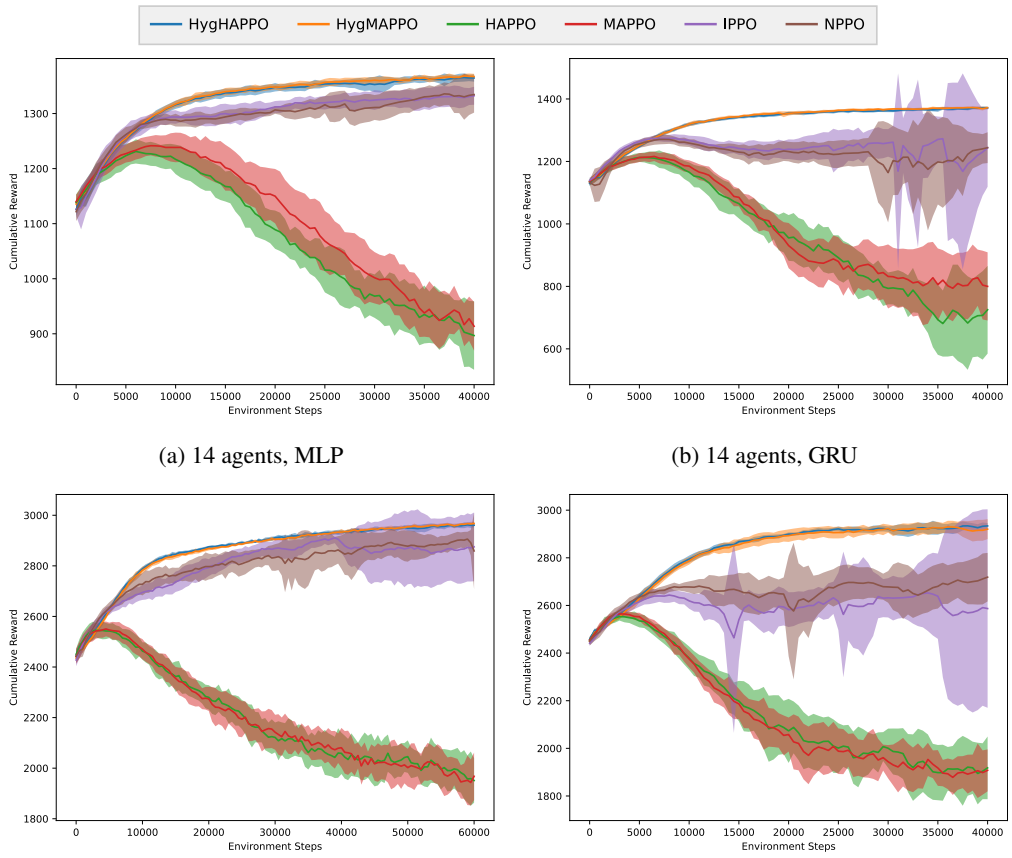
and Y is the network admittance matrix constructed from the physical topology and line impedances. This linearized model is common in distribution network studies (Bernstein & Dall'Anese, 2017).

All OU process parameters, stage means, base profiles, and network impedances are specified in the configuration files in the supplementary material.

The reward function for each node is designed to capture load balancing objectives:

- Voltage regulation: a penalty is applied when $|V|$ exceeds a pre-specified bound V_{bound}: $r_V = -c_V \cdot \max(0, |V| - V_{\text{bound}})$.
- Voltage flattening: a penalty proportional to the deviation between a node's voltage and the average voltage of its neighbors is applied: $r_{\text{flat}} = -c_{\text{flat}} \cdot |V - \bar{V}_{\text{neighbor}}|$.
- Generation cost: `Gate` nodes incur a cost $c_{\text{gate}} P$ when generating real power.
- Efficiency bonus: `Inverter` nodes receive a reward $r_{\text{eff}} = \eta \cdot P_{\text{out}}/P_{\text{in}}$ for high conversion efficiency.

The global reward is the sum of all node-wise rewards.



(a) 14 agents, MLP

(b) 14 agents, GRU

(c) 32 agents, MLP

(d) 32 agents, GRU

Figure 5: Training curves on Power Grid Load Balancing. Each subfigure shows the cumulative reward over training steps for 14 and 32 agents with MLP and GRU parameterizations. Shaded regions represent twice the standard deviation over 10 random seeds.

Graph and hypergraph construction. For HygRL, hyperedges are constructed from maximal cliques of the physical network graph. For the NPPO baseline, agents use their one-hop neighbors in the same physical graph, so that both methods share the same communication requirement. The specific network topologies used in the experiments are as follows.

- *14 nodes (2* `Gate` *nodes, 4* `Inverter` *nodes, 8* `Load` *nodes):* Nodes are placed at the 8 vertices and 6 face centers of a cube. A pair of opposite faces are `Gate` nodes; the remaining 4 faces are `Inverter` nodes; the 8 vertices are `Load` nodes. Edges connect each face center to its four bounding vertices, and connect the 8 vertices along the cube's 12 edges.
- *32 nodes (2* `Gate` *nodes, 10* `Inverter` *nodes, 20* `Load` *nodes):* Nodes are placed at the 20 vertices and 12 face centers of a regular dodecahedron. A pair of opposite faces are `Gate` nodes; the remaining 10 faces are `Inverter` nodes; the 20 vertices are `Load` nodes. Edges connect each face center to its five bounding vertices, and connect the 20 vertices along the dodecahedron's 30 edges.

Figure 5 presents the complete set of training curves across all 4 configurations (14/32 agents $\times$ MLP/GRU), and Table 4 reports the corresponding average performance.

Summary of complete results. The complete results confirm that HygRL's advantages extend beyond the homogeneous wireless communication setting to heterogeneous multi-agent systems with complex physical constraints. In both 14- and 32-agent configurations, HygHAPPO and

Table 4: Complete average performance on Power Grid Load Balancing (mean ± std). Best results are underlined; bold indicates within one standard deviation of the best result.

Setup	HygHAPPO	HAPPO	HygMAPPO	MAPPO	IPPO	NPPO
14 agents, MLP	**272.93** ±1.19	178.12 ±13.19	**273.47** ±0.88	182.88 ±9.04	266.27 ±3.33	266.94 ±6.65
14 agents, GRU	**274.23** ±0.88	144.18 ±29.98	**274.28** ±0.55	160.94 ±19.67	248.70 ±24.79	249.10 ±9.57
32 agents, MLP	**592.15** ±1.48	390.72 ±17.66	**593.31** ±0.88	390.97 ±16.65	574.85 ±26.78	572.07 ±29.96
32 agents, GRU	**586.51** ±2.66	382.21 ±24.26	**583.92** ±8.30	380.59 ±16.04	517.65 ±82.21	543.89 ±21.06

HygMAPPO consistently outperform their CTDE counterparts HAPPO and MAPPO, with performance gaps exceeding roughly 100 points in the 14-agent GRU setting and exceeding roughly 200 points in the 32-agent settings. IPPO and NPPO achieve competitive performance in smaller configurations (14 agents, MLP), but become unstable as the system grows, whereas HygRL remained stable. Notably, NPPO shows only marginal improvement over IPPO despite using graph information, confirming that pairwise graph neighborhoods do not adequately capture the higher-order dependencies among heterogeneous nodes. These trends mirror those observed in the wireless communication domain, demonstrating that HygRL's hypergraph-based decomposition generalizes across both homogeneous-role systems with individual state-action variations and fully heterogeneous-role systems.

C.3 Complete Results for Traffic Signal Control

Environment details. We use LightSim (Su & Deng, 2026), a lightweight traffic simulator based on the Cell Transmission Model (CTM). Agents control traffic signal phases to optimize traffic flow; we adopt the built-in "pressure" objective, which optimizes the difference between upstream and downstream densities at each intersection. The road network is modeled as an undirected graph where agents (intersections) are connected if they are adjacent on the road network. We evaluate on both synthetic grid networks and real-world OpenStreetMap networks. For grid networks, we consider three sizes: 4×2 (8 agents), 4×4 (16 agents), and 4×6 (24 agents), with edges connecting orthogonally adjacent intersections. For real-world networks, we use four scenarios from LightSim: Beijing (53 agents), Manhattan (33 agents), San Francisco (47 agents), and Los Angeles (55 agents).

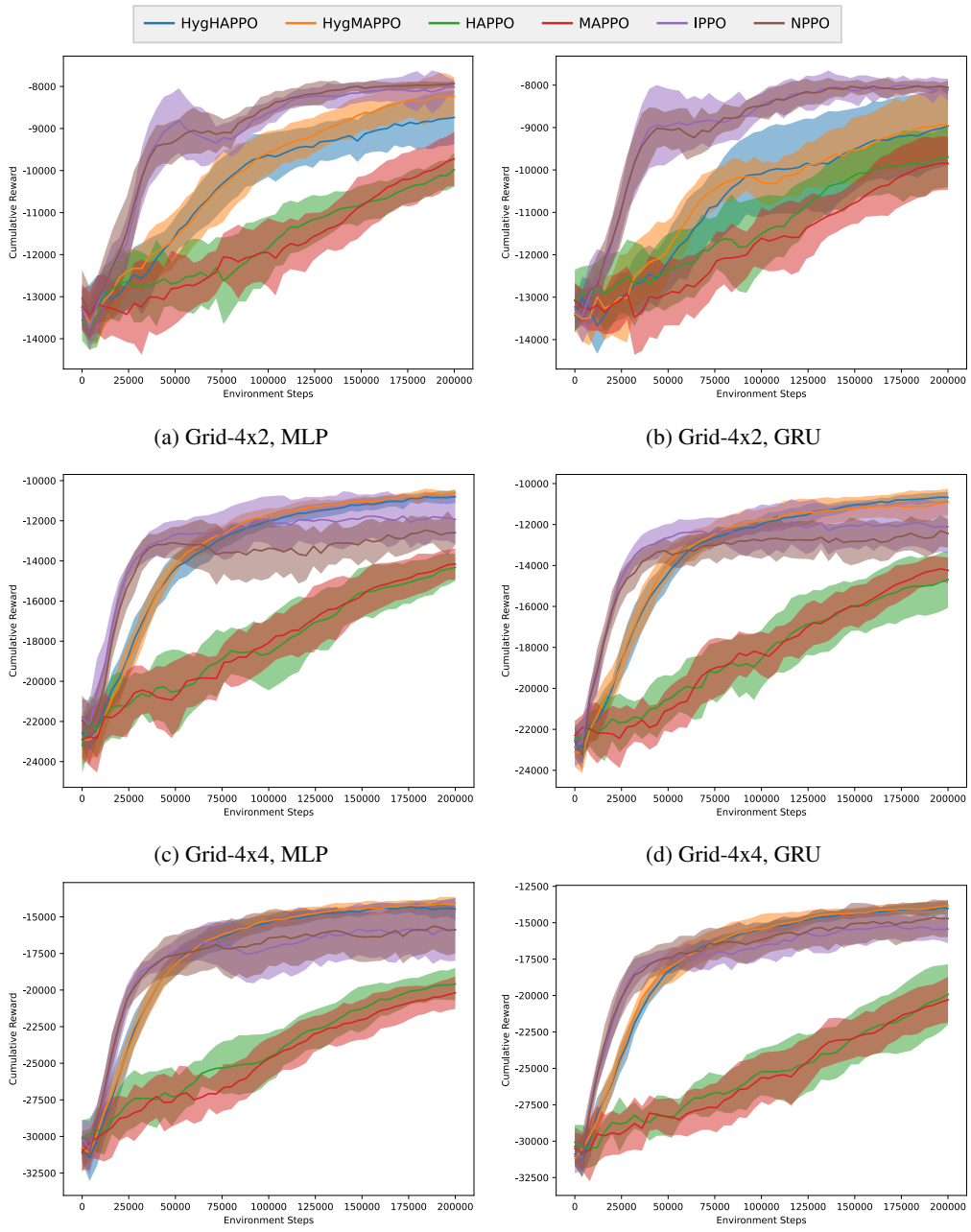
We note that the real-world networks contain intersections whose action spaces consist of a single point (fixed-phase intersections). These agents are not trained but still provide observations to neighboring agents, ensuring that the communication and information flow structure remains intact for all baselines.

Graph and hypergraph construction. The raw road network graph provided by LightSim is sparse and consists largely of grid-like cycles. To better capture the higher-order coordination structure among intersections,e.g., vehicles circulating around a city block interact with all intersections on that block, not just adjacent pairs,we augment the graph as follows.

Given a raw undirected graph $\mathcal{G}_0 = (\mathcal{N}, \mathcal{E}_0)$ and a threshold $C \in \mathbb{N}$, for every simple cycle in $\mathcal{G}_0$ of length at most C, we add edges between all pairs of vertices in that cycle, turning it into a clique. The resulting graph $\mathcal{G}$ is used by NPPO for one-hop neighbor information; the hypergraph $\mathcal{H}$ for HygRL is then constructed by taking all maximal cliques of $\mathcal{G}$. The specific values of C are listed below.

- Grid networks: $C = 4$;
- Real-world networks: Manhattan, San Francisco: $C = 4$; Beijing, Los Angeles: $C = 5$.

Figure 6 presents the complete training curves for grid networks, Figure 7 for real-world networks, and Table 5 reports the corresponding average performance.



(a) Grid-4x2, MLP

(b) Grid-4x2, GRU

(c) Grid-4x4, MLP

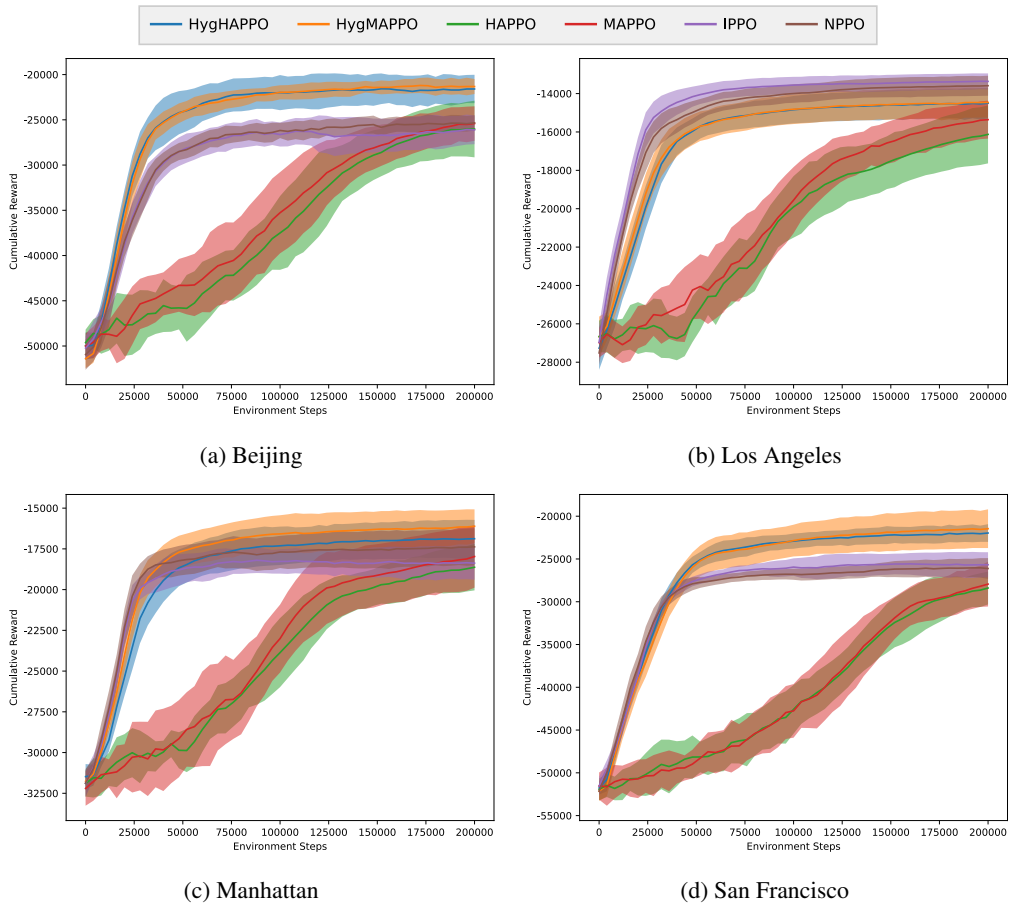
(d) Grid-4x4, GRU

(e) Grid-4x6, MLP

(f) Grid-4x6, GRU

Figure 6: Training curves on grid networks. Each subfigure shows the cumulative reward over training steps for 3 scenarios with MLP and GRU parameterizations. Shaded regions represent twice the standard deviation over 10 random seeds.

Summary of complete results. The complete results corroborate the conclusions drawn from the representative results in the main text. In grid networks, HygHAPPO and HygMAPPO perform comparably to CTDE methods in the smallest grid (4×2), but their advantage becomes increasingly evident as the grid size grows to 4×4 and 4×6. In the 4×6 grid, HygHAPPO and HygMAPPO outperform HAPPO and MAPPO by approximately $+1100$ and $+1200$ points, respectively. IPPO and NPPO converge to suboptimal solutions when the grid is 4×4 or 4×6, with NPPO showing almost no improvement over IPPO despite incorporating graph information.



(a) Beijing (b) Los Angeles

(c) Manhattan (d) San Francisco

Figure 7: Training curves on real-world networks. Each subfigure shows the cumulative reward over training steps for 4 real-world scenarios with MLP parameterization.

Table 5: Complete average performance on Traffic Signal Control (mean ± std). Best results are underlined; bold indicates within one standard deviation of the best result.

Setup	HygHAPPO	HAPPO	HygMAPPO	MAPPO	IPPO	NPPO
Grid-4x2, MLP	−1747.73 ±122.29	−1998.55 ±83.55	−1641.60 ±92.15	−1944.74 ±125.43	**−1600.67 ±38.69**	**−1593.16 ±16.72**
Grid-4x2, GRU	−1797.35 ±190.00	−1938.15 ±137.48	−1783.87 ±171.54	−1971.46 ±128.35	**−1626.56 ±52.03**	**−1609.56 ±20.37**
Grid-4x4, MLP	−2163.72 ±60.73	−2872.60 ±127.43	**−2115.69 ±34.60**	−2833.71 ±159.30	−2386.54 ±256.87	−2502.50 ±124.69
Grid-4x4, GRU	**−2136.09 ±50.84**	−2947.78 ±287.82	**−2176.69 ±123.45**	−2839.86 ±125.37	−2400.20 ±239.24	−2479.12 ±137.91
Grid-4x6, MLP	−2897.36 ±145.71	−3903.65 ±223.07	**−2809.65 ±85.58**	−4032.32 ±212.05	−3198.92 ±447.35	−3191.78 ±339.88
Grid-4x6, GRU	**−2802.04 ±90.36**	−3986.09 ±414.72	**−2771.94 ±72.56**	−4056.21 ±306.40	−3076.33 ±167.69	−2933.97 ±256.47
Beijing, MLP	**−4305.76 ±308.04**	−5180.00 ±624.51	**−4248.85 ±165.76**	−5087.61 ±389.47	−5224.31 ±315.22	−5064.95 ±171.80
Los Angeles, MLP	−2958.05 ±171.37	−3309.15 ±298.11	−2995.50 ±158.78	−3164.99 ±176.12	**−2719.32 ±115.80**	**−2779.73 ±135.04**
Manhattan, MLP	**−3421.48 ±274.61**	−3693.36 ±266.12	**−3276.76 ±211.76**	−3596.09 ±409.17	−3585.38 ±158.70	**−3460.44 ±173.11**
San Francisco, MLP	**−4451.42 ±212.08**	−5769.78 ±419.55	**−4358.16 ±442.16**	−5615.16 ±520.00	−5155.10 ±265.80	−5231.99 ±268.14

In real-world networks, HygRL consistently achieves higher final returns and faster convergence than CTDE methods in all tested scenarios. Notably, training curves are smoother than in grid scenarios, likely because real networks contain locally simple regions where policies converge quickly, stabilizing the surrounding environment. The graph-based NPPO baseline underperforms HygRL in most real-world networks, confirming that clique-based hypergraph decomposition captures traffic

Setup	Cohen's d	p-value
Wireless Communication, 5 agents, MLP	-0.59	2.05×10^{-01}
Wireless Communication, 5 agents, GRU	0.03	9.49×10^{-01}
Wireless Communication, 20 agents, MLP	6.05	7.11×10^{-11}
Wireless Communication, 20 agents, GRU	26.24	5.17×10^{-22}
Wireless Communication, 35 agents, MLP	11.45	1.31×10^{-15}
Wireless Communication, 35 agents, GRU	12.56	2.58×10^{-16}
Wireless Communication, 50 agents, MLP	7.00	6.28×10^{-12}
Wireless Communication, 50 agents, GRU	4.26	1.88×10^{-08}
Power Grid Load Balancing, 14 agents, MLP	9.61	2.79×10^{-14}
Power Grid Load Balancing, 14 agents, GRU	5.82	1.36×10^{-10}
Power Grid Load Balancing, 32 agents, MLP	15.25	8.29×10^{-18}
Power Grid Load Balancing, 32 agents, GRU	11.23	1.83×10^{-15}
LightSim, Grid-4x2 (8 agents), MLP	2.27	7.80×10^{-05}
LightSim, Grid-4x2 (8 agents), GRU	0.81	8.85×10^{-02}
LightSim, Grid-4x4 (16 agents), MLP	6.74	1.20×10^{-11}
LightSim, Grid-4x4 (16 agents), GRU	3.73	1.37×10^{-07}
LightSim, Grid-4x6 (24 agents), MLP	5.07	1.26×10^{-09}
LightSim, Grid-4x6 (24 agents), GRU	3.74	1.28×10^{-07}
LightSim, Manhattan (33 agents), MLP	0.95	4.69×10^{-02}
LightSim, San Francisco (47 agents), MLP	3.76	1.19×10^{-07}
LightSim, Beijing (53 agents), MLP	1.68	1.41×10^{-03}
LightSim, Los Angeles (55 agents), MLP	1.37	6.70×10^{-03}

Table 6: Cohen's d effect sizes and p-values from two-sample t-tests (Welch's t-test) for HygHAPPO vs HAPPO.

network structure more effectively than pairwise graph neighborhoods. Together with the wireless communication and load balancing results, these findings establish that HygRL's advantage is consistent across homogeneous competitive systems, heterogeneous physical networks, and large-scale traffic networks.

C.4 Statistical Significance

To quantify the reliability of the performance improvements and to demonstrate the scalability property of HygRL, we report Cohen's d effect sizes and p-values from two-sample t-tests for the primary comparisons: HygHAPPO vs. HAPPO and HygMAPPO vs. MAPPO. Cohen's d is computed as $d = (\bar{x}_1 - \bar{x}_2)/s_{\text{pooled}}$, where s_{pooled} is the pooled standard deviation. Following convention (Cohen, 1977), $|d| \geq 0.8$ indicates a large effect. All p-values are two-tailed and computed from two-sample t-tests assuming unequal variances (Welch's t-test) over the 10 independent random seeds for each algorithm.

The effect sizes and p-values are reported in Tables 6 and 7. Across most configurations, the effect sizes increase with the number of agents within the same problem domain, indicating that HygRL's advantage becomes more pronounced as the system scales. Notable exceptions, particularly in the wireless communication domain, can be attributed to changes in the underlying graph topology across different agent counts (e.g., regular polyhedral graphs at small scales versus random geometric graphs at large scales). In contrast, when the topology is held fixed (as in the LightSim grid networks), the effect sizes grow almost monotonically with the number of agents. Overall, the data consistently show large effect sizes ($|d| \geq 0.8$) and highly significant results ($p \leq 1 \times 10^{-2}$) in the vast majority of configurations, confirming the practical significance of HygRL's improvements.

Setup	Cohen's d	p-value
Wireless Communication, 5 agents, MLP	0.89	6.28×10^{-02}
Wireless Communication, 5 agents, GRU	-0.06	8.95×10^{-01}
Wireless Communication, 20 agents, MLP	1.94	3.99×10^{-04}
Wireless Communication, 20 agents, GRU	9.04	7.95×10^{-14}
Wireless Communication, 35 agents, MLP	14.98	1.14×10^{-17}
Wireless Communication, 35 agents, GRU	10.60	5.06×10^{-15}
Wireless Communication, 50 agents, MLP	6.89	8.23×10^{-12}
Wireless Communication, 50 agents, GRU	2.47	3.06×10^{-05}
Power Grid Load Balancing, 14 agents, MLP	13.38	8.39×10^{-17}
Power Grid Load Balancing, 14 agents, GRU	7.73	1.18×10^{-12}
Power Grid Load Balancing, 32 agents, MLP	16.28	2.59×10^{-18}
Power Grid Load Balancing, 32 agents, GRU	15.10	9.88×10^{-18}
LightSim, Grid-4x2 (8 agents), MLP	2.61	1.56×10^{-05}
LightSim, Grid-4x2 (8 agents), GRU	1.17	1.71×10^{-02}
LightSim, Grid-4x4 (16 agents), MLP	5.91	1.05×10^{-10}
LightSim, Grid-4x4 (16 agents), GRU	5.06	1.31×10^{-09}
LightSim, Grid-4x6 (24 agents), MLP	7.17	4.17×10^{-12}
LightSim, Grid-4x6 (24 agents), GRU	5.47	3.69×10^{-10}
LightSim, Manhattan (33 agents), MLP	0.93	5.22×10^{-02}
LightSim, San Francisco (47 agents), MLP	2.47	3.03×10^{-05}
LightSim, Beijing (53 agents), MLP	2.66	1.26×10^{-05}
LightSim, Los Angeles (55 agents), MLP	0.96	4.59×10^{-02}

Table 7: Cohen's d effect sizes and p-values from two-sample t-tests (Welch's t-test) for HygMAPPO vs MAPPO.

Table 8: MLP actor and critic structure.

Layer	Module	Width
0	Input	W_{input}
1	Linear	Output width: $W_{\text{MLP, actor}}$/$W_{\text{MLP, critic}}$
2	Tanh	-
3	Linear	Output width: $W_{\text{MLP, actor}}$/$W_{\text{MLP, critic}}$
4	Tanh	-
5	Linear	Output width: W_{output}

D Hyperparameters and Implementation Details

D.1 Implementation Details

Neural Network Structure. For all experiments, we use policy and value networks with two hidden layers and `Tanh` activations. The hidden layer widths are determined by the input dimension: we use $W_{\text{MLP, actor}} = \min(128, \max(32, 0.5W_{\text{input}}))$ and $W_{\text{MLP, critic}} = \min(512, \max(32, 0.5W_{\text{input}}))$ for MLP-based policies, and $W_{\text{GRU, actor}} = \min(128, \max(32, 0.25W_{\text{input}}))$ and $W_{\text{GRU, critic}} = \min(512, \max(32, 0.5W_{\text{input}}))$ for the GRU layer. This scaling strategy ensures that the networks remain sufficiently expressive for high-dimensional inputs while avoiding excessive parameter counts that could impair sample efficiency or lead to overfitting, particularly in small-scale settings. The same scaling rule is applied consistently across all domains to maintain a fair comparison. The specific network structures are presented in Tables 8 and 9.

Network input. Both the policy network and the value network use the same vectorization scheme. For discrete spaces, one-hot encoding is used; for $\mathbb{R}^n$, the space is described using sample

Table 9: GRU actor and critic structure.

Layer	Module	Width
0	Input	W_{input}
1	GRU	Hidden size: $W_{\text{GRU, actor}}$/$W_{\text{GRU, critic}}$
2	Linear	Output width: $W_{\text{MLP, actor}}$/$W_{\text{MLP, critic}}$
3	Tanh	-
4	Linear	Output width: $W_{\text{MLP, actor}}$/$W_{\text{MLP, critic}}$
5	Tanh	-
6	Linear	Output width: W_{output}

vectors; all encodings are then concatenated to obtain the final vectorized form. For the critic, the input consists of a direct concatenation of the vectors of the relevant agents.

Policy network output. The policy network outputs a probability distribution over actions conditioned on the observation. The output parameterization varies by action space type:

- **Discrete action spaces.** The network outputs logits over the finite set of actions; a categorical distribution is constructed via softmax, from which actions are sampled during training and the mode is selected during evaluation.
- **Unbounded continuous actions.** The network outputs the mean $\mu \in \mathbb{R}$ and standard deviation $\sigma \in \mathbb{R}$ of a Gaussian distribution. Actions are sampled as $a \sim \mathcal{N}(\mu, \sigma^2)$.

For action spaces with multiple dimensions (e.g., $\mathbb{R}^d$), the network independently outputs parameters for each dimension; actions are sampled per dimension and then concatenated.

Observation and Action Space.

- **Wireless communication.** The observation of user i is its packet queue, encoded as a binary tuple $s_i = (e_1, \ldots, e_{d_i}) \in \{0, 1\}^{d_i}$, where $e_\ell = 1$ indicates that the user has a packet with remaining deadline ℓ. The 2^{d_i} possible queue states are treated as 2^{d_i} independent categorical states in the environment; these states constitute the state space $\mathcal{S}_i$. The action space is $\mathcal{A}_i = \{\text{null}\} \cup Y_i$, where $Y_i \subseteq \{1, \ldots, m\}$ is the set of access points user i can access.
- **Load balancing.** The observation of each node includes the current time t, its local voltage V_i, real power P_i, and reactive power Q_i. For `Gate` and `Inverter` nodes, the observation additionally includes advice received from neighboring `Load` nodes; for `Inverter` nodes, it also includes the solar input power P_{in}. The action spaces are: $\mathcal{A}_{\text{Gate}} = \mathbb{R}^2$ (real and reactive power outputs), $\mathcal{A}_{\text{Inverter}} = \mathbb{R}$ (dissipation factor), and $\mathcal{A}_{\text{Load}} = \{-1, 0, 1\}^k$ (discrete advice to each neighboring generation node), where k is the number of neighboring `Gate` and `Inverter` nodes.
- **Traffic signal control.** We use LightSim (Su & Deng, 2026) with its default observation space. The action space is the set of available phases for each intersection, which is discrete. For a detailed description, we refer the reader to the LightSim documentation.

Other Details. All algorithms used in our experiments are variants of PPO and share a unified implementation backbone. Critically, within each experimental scenario, we use exactly the same set of hyperparameters for all methods (HygHAPPO, HAPPO, HygMAPPO, MAPPO, IPPO, and NPPO), without per-algorithm tuning. This is feasible because all baselines share the same PPO core; it also ensures that performance differences are attributed to the algorithmic structure, specifically, the use of hypergraph-based decomposition and policy aggregation, rather than to per-method hyperparameter optimization. We use GAE (Schulman et al., 2016) for advantage computation, as in the standard PPO implementation, and disable entropy regularization and advantage normalization across all variants to isolate the effect of our contributions. All networks are optimized with Adam using PyTorch's default parameters except for the learning rate, which is reported in Appendix D.2.

Table 10: Wireless Communication environment hyperparameters.

Hyperparameter	5 agents	20 agents	35 agents	50 agents
Number of APs	2	12	9	14
Queue length d_i	RandInt$(3, 6)$	RandInt$(3, 6)$	RandInt$(3, 8)$	RandInt$(3, 8)$
Packet receiving probability p_i	Uniform$(0.75, 1)$	Uniform$(0.75, 1)$	Uniform$(0.6, 1)$	Uniform$(0.6, 1)$
AP capacity w_j	RandInt$(1, 2)$	RandInt$(1, 2)$	RandInt$(1, 3)$	RandInt$(1, 4)$
Truncated length T (steps)	100	100	100	100

Table 11: Wireless Communication experiment training hyperparameters with MLP parameterization.

Hyperparameter	5 agents	20 agents	35 agents	50 agents
Discount factor γ	0.9	0.9	0.9	0.9
GAE λ	0.95	0.95	0.95	0.95
Clip range ϵ	0.2	0.2	0.2	0.2
Sample size N (episodes)	5	5	5	5
Episode length T (steps)	100	100	100	100
Batch size	500	500	500	500
Actor epochs	10	10	10	10
Actor learning rate	0.01	0.01	0.01	0.01
Critic epochs	10	10	10	10
Critic learning rate	0.01	0.01	0.01	0.01
Total Environment Steps	2.5e4	1e5	1e5	1e5

Compute resources. All runs were executed on a heterogeneous pool of NVIDIA GPUs, with three to five independent runs sharing each device; this concurrency is possible because the runs are not device-bound, so the per-run wall-clock times reported below are unaffected by the placement. The cost of a run varies with the configuration: the smallest takes about five minutes and the largest about five hours, averaging roughly three hours. With 22 configurations, 6 methods and 10 seeds per method, this corresponds to about 1.3×10^3 runs, i.e. on the order of 4.0×10^3 run-hours, or roughly 2.0×10^3 GPU-hours of device occupancy. Total wall-clock time and per-device allocations were not recorded, since runs were distributed across machines rather than scheduled as a single batch.

D.2 Hyperparameters

Wireless Communication Basic environmental hyperparameters are listed in Table 10. Training hyperparameters are listed in Tables 11 and 12. For graph and hypergraph construction, refer to the description in Appendix C.1.

Table 12: Wireless Communication experiment training hyperparameters with GRU parameterization.

Hyperparameter	5 agents	20 agents	35 agents	50 agents
Discount factor γ	0.9	0.9	0.9	0.9
GAE λ	0.95	0.95	0.95	0.95
Clip range ϵ	0.2	0.2	0.2	0.2
Sample size N (episodes)	5	5	5	5
Episode length T (steps)	100	100	100	100
Batch size	500	500	500	500
Actor epochs	10	10	10	10
Actor learning rate	0.01	0.01	0.01	0.01
Critic epochs	10	10	10	10
Critic learning rate	0.01	0.01	0.01	0.01
Total Environment Steps	2.5e4	1e5	1e5	1.5e5

Table 13: Power Grid Load Balancing experiment global hyperparameters.

Hyperparameter	14 agents	32 agents
Number of `Gate` nodes	2	2
Number of `Inverter` nodes	4	10
Number of `Load` nodes	8	20
V_{base} (V)	400	400
S_{base} (W)	1e5	1e5
V_{bound} (p.u.)	Uniform$(0.03, 0.05)$	Uniform$(0.03, 0.05)$
Day length T_{day}	24	24
Δt	0.24	0.24
c_V (p.u.$^{-1}$)	Uniform$(9, 11)$	Uniform$(9, 11)$
c_{flat} (p.u.$^{-1}$)	0.1	0.1
R (Ω)	Uniform$(0.15, 0.3)$	Uniform$(0.15, 0.3)$
$\frac{R}{X}$	Uniform$(2.5, 3.5)$	Uniform$(2.5, 3.5)$
Truncated length T (steps)	100	100

Table 14: `Gate` node hyperparameters.

Hyperparameter	14 agents	32 agents
c_{gate}	Uniform$(0.5, 2)$	Uniform$(0.5, 2)$

Power Grid Load Balancing Basic environmental hyperparameters are listed in Tables 13-18. Training hyperparameters are listed in Table 19. For graph and hypergraph construction, refer to the description in Appendix C.2.

All $\mu(t)$ are generated as follows: First sample t_i and $\mu(t_i)$ independently from the distributions in Table 17 for each stage $i = 1, \ldots, 3$. The sampled time points are then sorted to obtain an ordered sequence $t_1 < t_2 < t_3$. The function is then defined as $\mu(t) = \mu(t_i) + \frac{\mu(t_{i+1}) - \mu(t_i)}{t_{i+1} - t_i}(t - t_i)$ for $t_i \leq (t \bmod T_{\text{day}}) < t_{i+1}$, with $t_4 = t_1 + T_{\text{day}}$ and $\mu(t_4) = \mu(t_1)$.

All $\mu_{\text{pf}}(t)$ are generated as follows: First sample t_i and $\mu_{\text{pf}}(t_i)$ independently from the distributions in Table 18 for each stage $i = 1, \ldots, 5$. The sampled time points are then sorted to obtain an ordered sequence $t_1 < \cdots < t_5$. The function is then defined as $\mu_{\text{pf}}(t) = \mu_{\text{pf}}(t_i)$ for $t_i \leq (t \bmod T_{\text{day}}) < t_{i+1}$, with $t_6 = t_1 + T_{\text{day}}$.

Traffic Signal Control All scenarios in this experiment use the same set of environmental hyperparameters, which are listed in Table 20; parameters not listed use their default values. For details, refer to the LightSim implementation. All training tasks use the same set of hyperparameters, which

Table 15: `Inverter` node hyperparameters.

Hyperparameter	14 agents	32 agents
Solar Radiance S_0 (W $\cdot$ m^{-2})	1361	1361
θ	Uniform$(0.3, 0.8)$	Uniform$(0.3, 0.8)$
σ	Uniform$(0.05, 0.15)$	Uniform$(0.05, 0.15)$
Latitude ϕ	Uniform $\left(0, \frac{\pi}{3}\right)$	Uniform $\left(0, \frac{\pi}{3}\right)$
P_{clear}	Uniform$(0.75, 0.85)$	Uniform$(0.75, 0.85)$
C	Uniform$(0.5, 0.7)$	Uniform$(0.5, 0.7)$
A (m^2)	Uniform$(100, 200)$	Uniform$(100, 200)$
$P_{\text{loss},0}$ (W)	Uniform$(100, 200)$	Uniform$(100, 200)$
Loss coefficient k	Uniform$(0.2, 0.5)$	Uniform$(0.2, 0.5)$
P_{base}	Normal$(2000, 500^2)$	Normal$(1000, 200^2)$
η	Normal$(0.9, 0.1^2)$	Normal$(1, 0.1^2)$

Table 16: `Load` node hyperparameters.

Hyperparameter	14 agents	32 agents
a_0 (W)	Normal$(40000, 1000^2)$	Normal$(40000, 1000^2)$
a_1 (W)	Normal$(800, 400^2)$	Normal$(1200, 400^2)$
a_2 (W)	Normal$(4000, 1000^2)$	Normal$(3600, 1200^2)$
ϕ_1	Normal $\left(\frac{\pi}{2}, 0.5^2\right)$	Normal $\left(\frac{\pi}{2}, 0.5^2\right)$
ϕ_2	Normal $\left(\pi, 0.5^2\right)$	Normal $\left(\pi, 0.5^2\right)$
σ (W)	Normal$(2000, 400^2)$	Normal$(2000, 400^2)$
σ_{pf}	Normal$(0, 0.01^2)$	Normal$(0, 0.01^2)$

Table 17: Distributions for the random generation of each stage of $\mu(t)$

Stage	Time point	μ
1	$X_1 \bmod T_{\text{day}}, \quad X_1 \sim \text{Uniform}(5, 8)$	$\text{clip}(Y_1, 0, 1), \quad Y_1 \sim \text{Normal}(0.9, 0.3^2)$
2	$X_2 \bmod T_{\text{day}}, \quad X_2 \sim t_1 + \text{Normal}(6, 1^2)$	$\text{clip}(Y_2, 0, 1), \quad Y_2 \sim \text{Normal}(0.2, 0.2^2)$
3	$X_3 \bmod T_{\text{day}}, \quad X_3 \sim t_2 + \text{Normal}(6, 1^2)$	$\text{clip}(Y_3, 0, 1), \quad Y_3 \sim \text{Normal}(0.7, 0.3^2)$

are listed in Table 21. For graph and hypergraph construction, refer to the description in Appendix C.3.

Table 18: Distributions for the random generation of each stage of $\mu_{\text{pf}}(t)$

Stage	Time point	μ_{pf}
1	Uniform$(5, 7)$	$\text{clip}(Y_1, -1, 1), \quad Y_1 \sim \text{Normal}(0.89, 0.002^2)$
2	Uniform$(11, 13)$	$\text{clip}(Y_2, -1, 1), \quad Y_2 \sim \text{Normal}(0.93, 0.002^2)$
3	Uniform$(13, 15)$	$\text{clip}(Y_3, -1, 1), \quad Y_3 \sim \text{Normal}(0.87, 0.002^2)$
4	Uniform$(17, 19)$	$\text{clip}(Y_4, -1, 1), \quad Y_4 \sim \text{Normal}(0.95, 0.002^2)$
5	Uniform$(21, 23)$	$\text{clip}(Y_5, -1, 1), \quad Y_5 \sim \text{Normal}(0.94, 0.002^2)$

Table 19: Power Grid Load Balancing environment training hyperparameters with MLP and GRU parameterizations.

Hyperparameter	14 agents, MLP	14 agents, GRU	32 agents, MLP	32 agents, GRU
Discount factor γ	0.9	0.9	0.9	0.9
GAE λ	0.95	0.95	0.95	0.95
Clip range ϵ	0.2	0.2	0.2	0.2
Sample size N (episodes)	5	5	5	5
Episode length T (steps)	100	100	100	100
Batch size	500	500	500	500
Actor epochs	10	10	10	10
Actor learning rate	0.01	0.01	0.01	0.01
Critic epochs	10	10	10	10
Critic learning rate	0.01	0.01	0.01	0.01
Total Environment Steps	4e4	4e4	6e4	4e4

Table 20: LightSim environment parameters.

Parameter	Value
`max_steps`	800
`dt`	1
`reward_fn`	"pressure"

Table 21: LightSim scenarios experiment training hyperparameters.

Hyperparameter	Value
Discount factor γ	0.95
GAE λ	0.95
Clip range ϵ	0.2
Sample size N (episodes)	5
Episode length T (steps)	800
Batch size	4000
Actor epochs	10
Actor learning rate	0.005
Critic epochs	10
Critic learning rate	0.005
Total Environment Steps	2e5