Submodular Policy Learning for Distributed Task Allocation in Open Multi-Agent Systems

Jing Liu, Luca Ballotta, *Member, IEEE*, Yangyang Yang, Fangfei Li, *Member, IEEE*, Yang Tang, *Fellow, IEEE*, and Ruggero Carli, *Member, IEEE*

***Abstract*— This paper studies policy learning for distributed task allocation in open multi-agent systems, where agents may join and leave in a time-varying fashion, with submodular stage team utilities. At each time, the active agents select actions from local categorical policies such that the feasible joint agent-action pairs form a partition matroid. Standard continuous relaxations of submodular set functions are based on independent Bernoulli sampling, making them inconsistent with agents' policies. To solve this mismatch, we propose the *partition multilinear extension* (PME), a policy-based relaxation whose continuous support matches feasible actions under categorical policies. We prove that the marginal gains of the stage utility provide an unbiased estimator of the gradient of the PME and that maximizing the PME over action distributions is equivalent to maximizing the stage utilities over agent actions, which are critical to devise principled policy gradient. Building on this, we design *SubMAPL*, a centralized-training decentralized-execution KL-mirror policy-learning method that uses local marginal gains as stochastic PME gradients during training. KL-mirror updates preserve categorical feasibility without Euclidean projection. In the case where agents run tabular-softmax policies, we introduce open policy migration and an open-system KL tracking variation to handle agent arrivals and departures. Using dynamic regret analysis, we establish a lower bound on the cumulative utility which accounts for the openness of the environment and for the gap between optimal stage-wise and global utilities. Simulations on multi-agent coverage demonstrate that SubMAPL outperforms policy-gradient and online-learning baselines.**

***Index Terms*— Open multi-agent systems, submodular optimization, multi-agent reinforcement learning, distributed task allocation, policy learning.**

This work is supported by the Program of China Scholarship Council under Grant 202506740033, the National Key R&D Program of China under Grant 2025YFA1016504, the National Natural Science Foundation of China under Grants 62233005, 62573198, U2441245, U25B6002, 62403141, the National Key Laboratory of Space Target Awareness under Grant STA2025ZCB0208, in part by the Shanghai Institute for Mathematics and Interdisciplinary Sciences (SIMIS) under Grant SIMIS-ID-2025-SP. *(Corresponding authors: Fangfei Li and Yang Tang.)*

Jing Liu and Yangyang Yang are with the School of Mathematics, East China University of Science and Technology, Shanghai 200237, China (e-mail: y20220091@mail.ecust.edu.cn; y20250091@mail.ecust.edu.cn).

Luca Ballotta and Ruggero Carli are with the Department of Information Engineering, University of Padova, 35131 Padova, Italy (e-mail: luca.ballotta@unipd.it; ruggero.carli@unipd.it).

Fangfei Li is with the School of Mathematics and the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China (e-mail: li_fangfei@163.com, lifangfei@ecust.edu.cn).

Yang Tang is with the Key Laboratory of Smart Manufacturing in Energy Chemical Process, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China (e-mail: yangtang@ecust.edu.cn).

I. Introduction

MULTI-agent task allocation in open multi-agent systems (OMAS) arises in dynamic target tracking, active sensing, and coverage, where agents must coordinate local decisions under time-varying participation, limited communication, and non-additive team utilities. In long-duration deployments, agents may enter, leave, fail, or recover due to hardware malfunctions, battery recharging, or mission-level reconfiguration [1], [2], [3].

A central feature of such tasks is that utility functions are rarely additive across agents. For instance, overlapping field-of-views capture redundant information in coverage tasks. This so-called diminishing-returns structure is naturally modeled by monotone submodular set functions [4], [5], [6]. Submodularity has been extensively used in combinatorial optimization because it enables approximation guarantees for greedy selection under matroid constraints [2], [7], [8], [9] and the use of continuous relaxations such as the multilinear extension (MLE) with rounding [10], [11]. While these approaches are mostly tailored to static centralized optimization and task allocation and require dense communication graphs for consensus, the benefits of submodularity in learning data-driven multi-agent policies are underexplored.

Multi-agent reinforcement learning (MARL) is widespread to learn decentralized agent policies under partial observability and dynamical environment [1], [12], [13]. Although the interplay between MARL and submodularity has been studied, existing results are limited to centralized settings [14], [15], [16]. On the other hand, OMASs have garnered attention in the online learning literature, where submodularity-based methods enable suboptimality quantification of online optimization algorithms rather than learnable policies [2], [11], [17].

Incorporating submodular utilities into the learning of distributed policies is challenging. Static submodular problems are NP-hard in general and admit approximation guarantees rather than exact polynomial-time solutions [7], [8], [10]. The data-driven and dynamic nature of multi-agent policy learning introduces additional challenges compared to static optimization. First, the gradient feedback signal used in training must be consistent with decentralized categorical policies. However, classical multilinear relaxations rely on independent Bernoulli sampling and rounding, and therefore

do not represent factorized categorical policies executed by agents [10], [11]. Moreover, the combinatorial difficulty is coupled with sequential state evolution, decentralized information, and exponentially growing joint action space which demands careful credit assignment [1], [12], [13]. Existing submodular RL formulations target centralized policies rather than fully distributed decision-making [14], [15], [16].

Contribution: Building on the PME and its gradient characterization introduced in our preliminary work [18], we propose three main contributions to bridge the above mentioned gaps.

(1) We formulate distributed online submodular coordination as Markov decision process (MDP) in an OMAS and develop *SubMAPL*, a KL-mirror policy-learning method for decentralized submodular task allocation. Unlike sampled-action difference-reward policy gradients and counterfactual credit-assignment methods [18], [19], [20], SubMAPL uses a stochastic local gradient that evaluates every feasible local action of each active agent under the same sampled actions of the other agents. We show that this vector is an unbiased estimator of the PME coordinate gradient, thereby providing a principled first-order interpretation of marginal-contribution feedback.

(2) We introduce an open policy-migration mechanism to handle agent arrivals and departures, and define an open-system KL tracking variation that measures changes in both the stagewise optimal comparator and the time-varying policy domain. Our preliminary work [18] establishes guarantees for idealized Euclidean projected dynamics that directly update the PME marginals. However, the standard softmax policy-gradient implementation updates the logits, and the resulting marginal change depends on the softmax Jacobian and agrees with PME ascent only to first order. SubMAPL admits an exact equivalence between its KL-mirror policy update and an additive tabular-logit update. This removes the first-order mismatch between the analyzed policy-space dynamics and the implemented parameter update, allowing us to establish open-system guarantees directly for the policy sequence generated under agent arrivals and departures.

(3) By leveraging dynamic regret analysis in conjunction with DR-submodularity and monotonicity of the PME, we establish a lower bound on the cumulative utility within a $1/2$-approximation factor from the optimum. We further establish a $1/3$-factor bound in the case where the stage utilities are themselves marginal gains of a utility function capturing performance through the whole horizon, including scenarios such as coverage of static maps. These results complement previous analysis on stability and adaptability of open systems [3], [21]. Our analysis significantly improves our preliminary work [18], which derives a stagewise bound on the PME and a regret bound on the cumulative PME without formal guarantees on the original finite-horizon cumulative utility.

Organization: Section II formulates distributed submodular coordination as an MDP in OMAS. Section III introduces the PME and its properties. Section IV presents SubMAPL and Section V derives its suboptimality bound. Section VI reports simulations, and Section VII concludes the paper.

TABLE I: Main notation in the paper.

Symbol	Meaning
$\mathcal{M}$	Markov decision process in OMAS.
$\mathcal{N}_t$	Set of active agents at time t.
$\mathcal{R}_t$	Set of agents active at both times $t-1$ and t.
$\mathcal{N}_t^{\text{in}}$	Set of agents arriving at time t.
$\mathcal{N}_t^{\text{out}}$	Set of agents departing at time t.
$\mathcal{S}$	Global state space.
s_t	Global state at time t.
$\mathcal{A}_i$	Local action space of agent i.
$o_{i,t}$	Local observation of agent i at time t.
$a_{i,t}$	Local action selected by agent i at time t.
$\mathbf{a}_t$	Joint action tuple $(a_{i,t})_{i\in\mathcal{N}_t}$.
Ω_t	Active agent-action space at time t.
$\Omega_{i,t}$	Local agent-action space of agent i at time t.
$\mathcal{I}_t$	Partition-matroid feasible family on Ω_t.
A_t	Set of agent-action pairs executed at time t.
$F_t(A_t; s_t)$	Stage utility at time t.
π^i	Local categorical policy of agent i.
π	Sequence of factorized policies with active agents.
p^π	Distribution of state-action trajectory under π.
$J(\pi)$	Expected cumulative utility under π.
$x_t^\pi(s_t)$	Policy-induced action marginal vector at state s_t.
$\mathcal{P}(\mathcal{I}_t)$	Partition-matroid polytope at time t.
$\mathcal{F}_t$	Categorical-policy face of $\mathcal{P}(\mathcal{I}_t)$.
$\tilde{f}_t(\cdot; s_t)$	Partition multilinear extension of $F_t(\cdot; s_t)$.
$\mathcal{D}_t(\cdot \mid x)$	Partition sampling distribution induced by x.
$\mathcal{D}_t^{-i}(\cdot \mid x)$	Sampling distribution over all agents except i.
$\widehat{g}_t$	Stochastic estimator of $\nabla\tilde{f}_t(x_t; s_t)$.
$\mathcal{H}_t$	Pre-sampling history at time t.
$x_t^\star$	PME maximizer at time t.
$\mathcal{V}_T^{\text{open}}$	Open-system KL tracking variation.
$\text{Reg}_T^{1/2}$	Dynamic $1/2$-approximation regret.

Notation: We use $\mathbb{N}$ to denote the set of positive integers and define $[T] = \{1, \dots, T\}$ for any $T \in \mathbb{N}$. For a finite set S, $|S|$ denotes its cardinality and 2^S denotes its power set. For sets A and B, $A \setminus B$, $A \cup B$, and $A \cap B$ denote set difference, union, and intersection, respectively. For a scalar z, $[z]_+ = \max\{z, 0\}$. For vectors x and y, $x \le y$ denotes componentwise inequality and $\langle x, y\rangle$ denotes the Euclidean inner product. We use $\nabla f(x)$ and $\partial f/\partial x_i$ for gradients and partial derivatives, respectively. Expectations and probabilities are denoted by $\mathbb{E}[\cdot]$ and $\Pr(\cdot)$. The Kullback–Leibler divergence is denoted by $D_{\text{KL}}(\cdot\|\cdot)$. Standard asymptotic notation is denoted by $O(\cdot)$ and $o(\cdot)$. Table I summarizes the main paper-specific notation.

II. System Setup and Problem Formulation

We study decentralized task allocation in open MDPs [3], [14]. At each time step, the active agents locally observe the state of the environment and execute actions aiming to maximize a non-additive submodular global utility. Active agents and utilities may change over time, reflecting agent arrivals and departures, dynamic tasks, or an evolving environment [22], as shown in Figure 1. Our goal is to efficiently learn effective decentralized policies by leveraging the submodularity of team utilities.

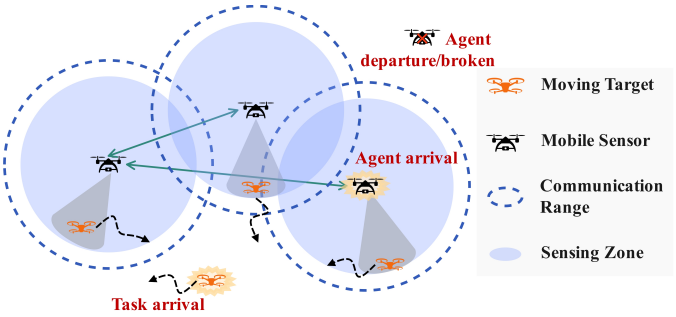

Fig. 1: Open multi-agent task allocation with time-varying agent participation, dynamic tasks, and limited communication and sensing.

A. Markov Decision Process in OMAS

Let $t \in [T] = \{1, \dots, T\}$ denote the time step and $\mathcal{N}_t$ the set of active agents at time t. We model the task-allocation process as a finite-horizon partially observed open multi-agent Markov decision process $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, F, \mu_1 \rangle$, where $\mathcal{S}$ is the global state space, $\mathcal{A}$ denotes the collection of feasible local action sets, $\mathcal{O} = \{O_t\}_{t\in[T]}$ is the sequence of local observation maps, $\mathcal{P} = \{P_t\}_{t\in[T]}$ is the sequence of controlled transition kernels, $F = \{F_t\}_{t\in[T]}$ is the sequence of stage-utility set functions, and μ_1 is the initial-state distribution on $\mathcal{S}$. For each agent i that may be active during the horizon, let $\mathcal{O}_i$ denote its finite local observation space. Here, $O_t = (O_{i,t})_{i\in\mathcal{N}_t}$, where $O_{i,t} : \mathcal{S} \to \mathcal{O}_i$ and $o_{i,t} = O_{i,t}(s_t)$. The initial state is sampled as $s_1 \sim \mu_1$ [3], [16], [23].

Each active agent $i \in \mathcal{N}_t$ receives a local observation $o_{i,t}$ and selects an action $a_{i,t}$ from a finite nonempty feasible action set $\mathcal{A}_i$. The global state $s_t \in \mathcal{S}$ characterizes the environment, including the tasks to be completed. Given the joint action $\mathbf{a}_t = (a_{i,t})_{i\in\mathcal{N}_t}$, the transition kernel P_t governs the state evolution as $s_{t+1} \sim P_t(\cdot \mid s_t, \mathbf{a}_t)$.

The ground set of agent-action pairs available at time t is $\Omega_t = \{(i,a) : i \in \mathcal{N}_t,\ a \in \mathcal{A}_i\}$. For each active agent, define the local actions $\Omega_{i,t} = \{(i,a) : a \in \mathcal{A}_i\},\ i \in \mathcal{N}_t$. The feasible subsets of agent-action pairs are described by the partition matroid $(\Omega_t, \mathcal{I}_t)$ with $\mathcal{I}_t = \{A \subseteq \Omega_t : |A \cap \Omega_{i,t}| \le 1,\ \forall i \in \mathcal{N}_t\}$. The executed joint action induced by the local choices is $A_t = \{(i, a_{i,t}) : i \in \mathcal{N}_t\} \in \mathcal{I}_t$ [10].

To describe changes in the open-system domain, let $\mathcal{N}_0 = \emptyset$ and define $\mathcal{R}_t = \mathcal{N}_t \cap \mathcal{N}_{t-1}, \mathcal{N}_t^{\text{in}} = \mathcal{N}_t \setminus \mathcal{N}_{t-1}, \mathcal{N}_t^{\text{out}} = \mathcal{N}_{t-1} \setminus \mathcal{N}_t$. Here $\mathcal{R}_t$, $\mathcal{N}_t^{\text{in}}$, and $\mathcal{N}_t^{\text{out}}$ denote the remaining, arriving, and departing agents at time t, respectively. Moreover, we assume the agent arrivals and departures are independent of the agents' local policies and the state transitions [3].

B. Decentralized Categorical Policies

Each active agent $i \in \mathcal{N}_t$ processes a local observation $o_{i,t} = O_{i,t}(s_t)$, where $O_{i,t} : \mathcal{S} \to \mathcal{O}_i$ is a deterministic observation map modeling local sensing and communication with neighbors, and selects an action $a_{i,t} \in \mathcal{A}_i$. A policy π^i assigns a categorical distribution over $\mathcal{A}_i$ given $o \in \mathcal{O}_i$ such that, at time t, active agent $i \in \mathcal{N}_t$ computes its action as a realization $a_{i,t} \sim \pi^i(\cdot \mid o_{i,t})$. Let $\mathbf{o}_t = (o_{i,t})_{i\in\mathcal{N}_t}$ denote the joint observation vector. We consider factorized categorical policies [1], [24]

$$\pi_t(\mathbf{a}_t \mid \mathbf{o}_t) = \prod_{i\in\mathcal{N}_t} \pi^i(a_{i,t} \mid o_{i,t}). \tag{1}$$

The factorized policy π_t is time dependent as a consequence of time-varying active agent sets $\mathcal{N}_t$. The factorization in (1) governs decentralized execution during deployment; since the distributions $\{\pi^i(\cdot \mid o_{i,t})\}_{i\in\mathcal{N}_t}$ are conditioned on local observations $o_{i,t}$, the agents sample their actions independently.

C. Submodular Team Utility and Problem Statement

At each time step t, the team performance is measured by a set function $F_t(\cdot\,; s_t) : 2^{\Omega_t} \to \mathbb{R}_{\ge 0}$ defined over subsets of the current agent-action ground set Ω_t.

Definition 1 (Normalized monotone submodular set function [25])**.** Let Ω be a finite ground set. For any $A \subseteq \Omega$ and $e \in \Omega \setminus A$, define the marginal gain of adding e to A as $F(e \mid A) = F(A \cup \{e\}) - F(A)$. A set function $F : 2^{\Omega} \to \mathbb{R}_{\ge 0}$ is normalized, monotone, and submodular if it respectively satisfies the following properties:

1) $F(\emptyset) = 0$.
2) for any sets $A \subseteq A' \subseteq \Omega$, $F(A) \le F(A')$.
3) for any sets $A \subseteq A' \subseteq \Omega$ and element $e \in \Omega \setminus A'$, $F(e \mid A) \ge F(e \mid A')$.

Assumption 1 (Utility and Feasibility Conditions)**.** For every time step $t \in [T]$, the following holds.

(i) For every state s_t, the team utility $F_t(\cdot\,; s_t) : 2^{\Omega_t} \to \mathbb{R}_{\ge 0}$ is normalized, monotone, and submodular.
(ii) Each active agent has a nonempty local feasible action set, i.e., $\mathcal{A}_i \neq \emptyset$ for all $i \in \mathcal{N}_t$.

Assumption 1 means that adding an agent-action pair can increase the team utility but its marginal gain decreases when more agent-action pairs have already been selected. This diminishing-returns structure captures redundancy among agent contributions [2].

Problem 1 (Open-System Submodular Policy Learning)**.** Given the open multi-agent MDP, find a decentralized categorical policy $\pi = \{\pi_t\}_{t=1}^T$, with $\pi_t = \prod_{i\in\mathcal{N}_t} \pi^i$, that maximizes the expected cumulative submodular utility

$$\begin{aligned} \max_{\pi=\{\pi_t\}_{t=1}^T} \quad & J(\pi) = \mathbb{E}_{\tau\sim p^\pi}\left[\sum_{t=1}^{T} F_t(A_t; s_t)\right] \\ \text{s.t.} \quad & a_{i,t} \sim \pi^i(\cdot \mid o_{i,t}),\ \forall i \in \mathcal{N}_t,\ \forall t \in [T], \\ & A_t \in \mathcal{I}_t,\ \forall t \in [T]. \end{aligned} \tag{2}$$

Here, $\tau = (s_1, \mathbf{a}_1, s_2, \dots, s_T, \mathbf{a}_T, s_{T+1})$ and p^π denotes the trajectory distribution induced by π and the transition kernels $\{P_t\}_{t=1}^T$.

Even for a single time step, Problem 1 reduces to monotone submodular maximization under a partition matroid constraint, which is NP-hard in general [7], [10]. Over a finite horizon, this combinatorial difficulty is exacerbated by action-dependent state evolution and time-varying agent participation. Since directly optimizing the cumulative utility (i.e., the

return) is intractable [14], we adopt a stagewise approach and leverage the submodularity of stage utilities to learn the decentralized policies. The next section develops a continuous relaxation of stage utilities that formally supports our stagewise policy learning framework.

III. Partition Multilinear Extension

A standard approach to combinatorial submodular maximization is to replace the discrete set problem with a continuous relaxation that is amenable to continuous optimization methods, particularly under matroid constraints [10], [26]. The standard multilinear extension is based on independent Bernoulli sampling over ground-set elements [27]. Under the partition constraint induced by $\mathcal{I}_t$, this sampling rule may select multiple elements from the same agent action subset $\Omega_{i,t}$, creating a formal mismatch with the policy in (1) which selects one feasible action per agent. This motivates the continuous relaxation defined in the following to support training of factorized categorical policies in a principled fashion.

Building on our preliminary work [18], we use the partition multilinear extension (PME), a continuous relaxation of utilities F_t whose sampling distribution matches the factorized categorical execution under the partition matroid. At each time step t, policy (1) at observations $\{o_{i,t}\}_{i\in\mathcal{N}_t}$ induces a marginal vector $x_t^\pi(s_t) \in [0,1]^{|\Omega_t|}$ over agent-action pairs, with coordinates

$$x^\pi_{(i,a),t}(s_t) = \pi^i(a \mid o_{i,t}),\ \ (i,a) \in \Omega_t. \tag{3}$$

For readability, we denote a generic marginal vector by x. Each coordinate $x^\pi_{(i,a),t}(s_t)$ is the probability that agent $i \in \mathcal{N}_t$ executes action a at the current state s_t under policy π^i.

A. Definition and Policy Equivalence

The partition matroid polytope associated with $(\Omega_t, \mathcal{I}_t)$ is [10]

$$\mathcal{P}(\mathcal{I}_t) = \left\{ x \in [0,1]^{|\Omega_t|} : \sum_{a\in\mathcal{A}_i} x_{(i,a)} \le 1,\ \forall i \in \mathcal{N}_t \right\}. \tag{4}$$

For $x \in \mathcal{P}(\mathcal{I}_t)$, let $\mathcal{D}_t(\cdot \mid x)$ denote the partition sampling distribution over $\mathcal{I}_t$. Under this distribution, each agent $i \in \mathcal{N}_t$ selects element (i,a) with probability $x_{(i,a)}$, and selects no element with probability $1 - \sum_{a\in\mathcal{A}_i} x_{(i,a)}$. The choices are independent across agents.

Definition 2 (Partition Multilinear Extension [18])**.** Given the current state s_t and open-system domain, the partition multilinear extension of $F_t(\cdot; s_t)$ is the function $\tilde{f}_t(\cdot; s_t) : \mathcal{P}(\mathcal{I}_t) \to \mathbb{R}_{\ge 0}$ defined by

$$\tilde{f}_t(x; s_t) = \mathbb{E}_{A\sim\mathcal{D}_t(\cdot|x)}\left[F_t(A; s_t)\right] \tag{5}$$

$$= \sum_{A\in\mathcal{I}_t} F_t(A; s_t) \prod_{i\in\mathcal{N}_t} p_i(A; x), \tag{6}$$

where

$$p_i(A; x) = \begin{cases} x_{(i,a)}, & \text{if } A \cap \Omega_{i,t} = \{(i,a)\}, \\ 1 - \sum_{a'\in\mathcal{A}_i} x_{(i,a')}, & \text{if } A \cap \Omega_{i,t} = \emptyset. \end{cases} \tag{7}$$

The categorical-policy face of $\mathcal{P}(\mathcal{I}_t)$ is

$$\mathcal{F}_t = \left\{ x \in \mathcal{P}(\mathcal{I}_t) : \sum_{a\in\mathcal{A}_i} x_{(i,a)} = 1,\ \forall i \in \mathcal{N}_t \right\}. \tag{8}$$

For each $i \in \mathcal{N}_t$, define the local categorical simplex $\mathcal{F}_i = \left\{x_i \in [0,1]^{|\mathcal{A}_i|} : \sum_{a\in\mathcal{A}_i} x_{(i,a)} = 1\right\}$. Then $\mathcal{F}_t = \prod_{i\in\mathcal{N}_t} \mathcal{F}_i$. For any $x \in \mathcal{F}_t$, the probability of selecting no action is zero in (7). Hence, the distribution induced by the PME respects the partition structure of Ω_t since it selects exactly one agent-action pair from each $\Omega_{i,t}$ w.p.1, matching the categorical action-selection model in (1).

Remark 1. At a fixed time step, the PME is closely related to the policy-based continuous extension in [17] and the Multinoulli Extension in [28]. The former considers online coordination over fixed agent and action sets, whereas the latter addresses static subset selection under fixed partition constraints. Since these formulations do not model controlled Markovian dynamics or changes in the active-agent set, they cannot capture the coupling among current decisions, future states, and utilities. Our formulation extends the PME to finite-horizon decentralized policy learning for an OMAS and establishes exact equivalence with both the expected stage utility and the finite-horizon policy objective.

Lemma 1 (Policy-PME Objective Equivalence [18])**.** *Let $\pi_t = \prod_{i\in\mathcal{N}_t} \pi^i$ be a factorized categorical policy, and let $x_t^\pi(s_t)$ be defined by (3). Then, for any fixed s_t,*

$$\mathbb{E}_{A_t\sim\pi_t}\left[F_t(A_t; s_t) \mid s_t\right] = \tilde{f}_t\big(x_t^\pi(s_t); s_t\big). \tag{9}$$

Proof. Consider the current state s_t and the current open-system domain. Then the local observations $\{o_{i,t}\}_{i\in\mathcal{N}_t}$ are fixed, and each agent samples independently from $\pi^i(\cdot \mid o_{i,t})$. Hence, the executed set is $A_t = \{(i, a_{i,t}) : i \in \mathcal{N}_t\}$, $a_{i,t} \sim \pi^i(\cdot \mid o_{i,t})$. By definition of $x_t^\pi(s_t)$, the per-agent factor in (7) satisfies $p_i(A; x_t^\pi(s_t)) = \pi^i(a \mid o_{i,t})$ if $A \cap \Omega_{i,t} = \{(i,a)\}$. Moreover, since $\pi^i(\cdot \mid o_{i,t})$ is a categorical distribution, $\sum_{a\in\mathcal{A}_i} x^\pi_{(i,a),t}(s_t) = 1$, the probability of selecting no action in (7) is zero for every active agent. Therefore, under $\mathcal{D}_t(\cdot \mid x_t^\pi(s_t))$, positive probability is assigned only to feasible sets that select exactly one action per active agent, and this probability equals the product of the corresponding local categorical probabilities. Thus, for any feasible $A \in \mathcal{I}_t$, $\Pr(A_t = A \mid s_t) = \prod_{i\in\mathcal{N}_t} p_i(A; x_t^\pi(s_t))$. Substituting this identity into the conditional expectation gives

$$\begin{aligned} \mathbb{E}_{A_t\sim\pi_t}\left[F_t(A_t; s_t) \mid s_t\right] &= \sum_{A\in\mathcal{I}_t} F_t(A; s_t) \Pr(A_t = A \mid s_t) \\ &= \sum_{A\in\mathcal{I}_t} F_t(A; s_t) \prod_{i\in\mathcal{N}_t} p_i(A; x_t^\pi(s_t)) \\ &= \tilde{f}_t\big(x_t^\pi(s_t); s_t\big), \end{aligned}$$

which proves (9). □

Let $\mathcal{D}_t^{-i}(\cdot \mid x)$ denote the joint distribution induced by x over the local action blocks of all agents except i, namely $\{\Omega_{j,t}\}_{j\in\mathcal{N}_t\setminus\{i\}}$.

Lemma 2 (PME Coordinate Gradient [18])**.** *For any $(i,a) \in$*

Ω_t *and* $x \in \mathcal{P}(\mathcal{I}_t)$,

$$\frac{\partial \tilde{f}_t}{\partial x_{(i,a)}}(x; s_t) = \mathbb{E}_{A^{-i}\sim\mathcal{D}_t^{-i}(\cdot|x)}\left[F_t\big((i,a) \mid A^{-i}; s_t\big)\right], \quad (10)$$

where $F_t\big((i,a) \mid A^{-i}; s_t\big) = F_t\big(A^{-i} \cup \{(i,a)\}; s_t\big) - F_t(A^{-i}; s_t)$ *denotes the marginal gain of adding the agent-action pair* (i,a) *to* A^{-i} *at state* s_t.

Proof. Consider an arbitrary $(i,a) \in \Omega_t$. In the sum representation in (6), the coordinate $x_{(i,a)}$ appears only in the local factor $p_i(A; x)$. There are three cases. If $A \cap \Omega_{i,t} = \{(i,a)\}$, then $\partial p_i(A; x)/\partial x_{(i,a)} = 1$. If $A \cap \Omega_{i,t} = \{(i,a')\}$ for some $a' \neq a$, then $\partial p_i(A; x)/\partial x_{(i,a)} = 0$. If $A \cap \Omega_{i,t} = \emptyset$, then $\partial p_i(A; x)/\partial x_{(i,a)} = -1$. Applying the product rule to (6) gives

$$\frac{\partial \tilde{f}_t}{\partial x_{(i,a)}}(x; s_t) = \sum_{A:(i,a)\in A} F_t(A; s_t) \prod_{k\neq i} p_k(A; x) - \sum_{A:A\cap\Omega_{i,t}=\emptyset} F_t(A; s_t) \prod_{k\neq i} p_k(A; x).$$

Every set in the first sum can be written uniquely as $A^{-i} \cup \{(i,a)\}$, while every set in the second sum can be identified with the same partial joint action A^{-i} of all agents except i. Moreover, $\prod_{k\neq i} p_k(A; x)$ is exactly the probability of A^{-i} under $\mathcal{D}_t^{-i}(\cdot \mid x)$. Hence

$$\begin{aligned}&\frac{\partial \tilde{f}_t}{\partial x_{(i,a)}}(x; s_t)\\ &= \sum_{A^{-i}} \Pr_{\mathcal{D}_t^{-i}(\cdot|x)}(A^{-i}) \left[F_t(A^{-i} \cup \{(i,a)\}; s_t) - F_t(A^{-i}; s_t)\right]\\ &= \mathbb{E}_{A^{-i}\sim\mathcal{D}_t^{-i}(\cdot|x)}\left[F_t\big((i,a) \mid A^{-i}; s_t\big)\right].\end{aligned}$$

This proves (10). □

Lemma 2 shows that partial derivative of PME is the expected marginal gain of the corresponding agent-action pair under the current categorical sampling distribution of the other agents. This identity provides the first-order information used by the counterfactual marginal-gain estimator in Section IV.

Proposition 1 (Properties of the PME [18])**.** *Under Assumption 1,* $\tilde{f}_t(\cdot; s_t)$ *satisfies the following properties on* $\mathcal{P}(\mathcal{I}_t)$*:*

1) $\tilde{f}_t(0; s_t) = 0$.
2) $\tilde{f}_t(x; s_t) \geq 0$ *for all* $x \in \mathcal{P}(\mathcal{I}_t)$.
3) $\tilde{f}_t(\cdot; s_t)$ *is a multilinear polynomial and hence is smooth.*
4) $\nabla \tilde{f}_t(x; s_t) \geq 0$ *componentwise for all* $x \in \mathcal{P}(\mathcal{I}_t)$.
5) *For any* $x, y \in \mathcal{P}(\mathcal{I}_t)$ *with* $x \leq y$ *componentwise,* $\nabla \tilde{f}_t(x; s_t) \geq \nabla \tilde{f}_t(y; s_t)$ *componentwise.*

Consequently, $\tilde{f}_t(\cdot; s_t)$ *is monotone and DR-submodular on* $\mathcal{P}(\mathcal{I}_t)$.

Proof. We prove the properties in order.

1) When $x = 0$, each agent selects no element with probability one under $\mathcal{D}_t(\cdot \mid 0)$. Therefore, the only set with nonzero probability is $\emptyset$, and $\tilde{f}_t(0; s_t) = F_t(\emptyset; s_t) = 0$.

2) By nonnegativity of F_t and by the fact that all factors $p_i(A; x)$ are valid probabilities for $x \in \mathcal{P}(\mathcal{I}_t)$, every term in (6) is nonnegative. Thus, $\tilde{f}_t(x; s_t) \geq 0$.

3) For each fixed set $A \in \mathcal{I}_t$, every factor $p_i(A; x)$ is affine in the coordinates $\{x_{(i,a)} : a \in \mathcal{A}_i\}$. Since these coordinate sets are disjoint across agents, the product $\prod_{i\in\mathcal{N}_t} p_i(A; x)$ is multilinear in the variables associated with different agents. Hence $\tilde{f}_t(\cdot; s_t)$ is a finite sum of multilinear polynomials and is smooth.

4) By Lemma 2, for any $(i,a) \in \Omega_t$, $\frac{\partial \tilde{f}_t}{\partial x_{(i,a)}}(x; s_t) = \mathbb{E}_{A^{-i}\sim\mathcal{D}_t^{-i}(\cdot|x)}\left[F_t\big((i,a) \mid A^{-i}; s_t\big)\right]$.

For every realization A^{-i} in the support of $\mathcal{D}_t^{-i}(\cdot \mid x)$, we have $A^{-i} \subseteq A^{-i} \cup \{(i,a)\}$. Since $F_t(\cdot; s_t)$ is monotone, $F_t\big((i,a) \mid A^{-i}; s_t\big) = F_t\big(A^{-i} \cup \{(i,a)\}; s_t\big) - F_t(A^{-i}; s_t) \geq 0$. Taking expectation preserves the inequality, and hence $\frac{\partial \tilde{f}_t}{\partial x_{(i,a)}}(x; s_t) \geq 0$, $(i,a) \in \Omega_t$. Therefore, $\nabla \tilde{f}_t(x; s_t) \geq 0$ componentwise.

5) We prove the DR property by showing that the gradient is componentwise nonincreasing. For each agent i, the PME is affine in the local coordinates $\{x_{(i,a)} : a \in \mathcal{A}_i\}$, because the per-agent factor $p_i(A; x)$ in (7) is affine in these coordinates and no product contains two coordinates associated with the same agent. Hence, all second partial derivatives with respect to two coordinates of the same agent are zero:

$$\frac{\partial^2 \tilde{f}_t}{\partial x_{(i,a)} \partial x_{(i,b)}}(x; s_t) = 0, \ a, b \in \mathcal{A}_i. \quad (11)$$

Consider two distinct agents $i \neq u$ and actions $a \in \mathcal{A}_i$ and $b \in \mathcal{A}_u$. Let $\mathcal{D}_t^{-\{i,u\}}(\cdot \mid x)$ denote the product distribution induced by x over all agents in $\mathcal{N}_t \setminus \{i,u\}$. By Lemma 2,

$$\frac{\partial \tilde{f}_t}{\partial x_{(i,a)}}(x; s_t) = \mathbb{E}_{A^{-i}\sim\mathcal{D}_t^{-i}(\cdot|x)}\left[F_t\big((i,a) \mid A^{-i}; s_t\big)\right]. \quad (12)$$

To differentiate this expression with respect to $x_{(u,b)}$, separate the random choice of agent u from the choices of the other agents. For any fixed realization $A^{-\{i,u\}}$ of the agents other than i and u, the contribution of agent u to (12) is

$$\left(1 - \sum_{c\in\mathcal{A}_u} x_{(u,c)}\right) F_t\big((i,a) \mid A^{-\{i,u\}}; s_t\big) + \sum_{c\in\mathcal{A}_u} x_{(u,c)} F_t\big((i,a) \mid A^{-\{i,u\}} \cup \{(u,c)\}; s_t\big). \quad (13)$$

Differentiating (13) with respect to $x_{(u,b)}$ gives

$$\begin{aligned}&\frac{\partial^2 \tilde{f}_t}{\partial x_{(u,b)} \partial x_{(i,a)}}(x; s_t)\\ &= \mathbb{E}_{A^{-\{i,u\}}\sim\mathcal{D}_t^{-\{i,u\}}(\cdot|x)}\Big[F_t\big((i,a) \mid A^{-\{i,u\}} \cup \{(u,b)\}; s_t\big) - F_t\big((i,a) \mid A^{-\{i,u\}}; s_t\big)\Big].\end{aligned} \quad (14)$$

Since $A^{-\{i,u\}} \subseteq A^{-\{i,u\}} \cup \{(u,b)\}$, submodularity of $F_t(\cdot; s_t)$ implies $F_t\big((i,a) \mid A^{-\{i,u\}} \cup \{(u,b)\}; s_t\big) - F_t\big((i,a) \mid A^{-\{i,u\}}; s_t\big) \leq 0$. Taking expectation in (14) yields

$$\frac{\partial^2 \tilde{f}_t}{\partial x_{(u,b)} \partial x_{(i,a)}}(x; s_t) \leq 0, \ i \neq u. \quad (15)$$

Equations (11) and (15) show that each coordinate derivative of $\tilde{f}_t$ is nonincreasing in every coordinate. Therefore, for any $x, y \in \mathcal{P}(\mathcal{I}_t)$ with $x \leq y$ componentwise,

$\nabla \tilde{f}_t(x; s_t) \geq \nabla \tilde{f}_t(y; s_t)$ componentwise. This establishes DR-submodularity on $\mathcal{P}(\mathcal{I}_t)$. □

Proposition 1 shows that the PME is a smooth, monotone, DR-submodular objective on $\mathcal{P}(\mathcal{I}_t)$ while preserving the per-agent categorical sampling structure required for decentralized execution [29], [30].

B. Problem Equivalence

In this section, we relate the PME to Problem 1. First, we relate the continuous PME optimum to the discrete stagewise submodular optimum. Then, we extend the equivalence to the finite-horizon objective $J(\pi)$.

Although $\tilde{f}_t(\cdot; s_t)$ is defined over the partition-matroid polytope $\mathcal{P}(\mathcal{I}_t)$, monotonicity implies that an optimal solution can be chosen on the categorical-policy face $\mathcal{F}_t$.

Lemma 3 (PME Exactness on the Categorical Face)**.** *Under Assumption 1, the following stagewise identity holds at each time step t and state s_t:*

$$\max_{x \in \mathcal{P}(\mathcal{I}_t)} \tilde{f}_t(x; s_t) = \max_{x \in \mathcal{F}_t} \tilde{f}_t(x; s_t) = \max_{A \in \mathcal{I}_t} F_t(A; s_t). \tag{16}$$

Proof. We first show that the optimum of $f_t(\cdot; s_t)$ over $\mathcal{P}(\mathcal{I}_t)$ is attained on $\mathcal{F}_t$. Let $x^\star \in \arg\max_{x \in \mathcal{P}(\mathcal{I}_t)} \tilde{f}_t(x; s_t)$. If $x^\star \in \mathcal{F}_t$, the claim is immediate. Otherwise, there exists an agent $i \in \mathcal{N}_t$ for which the corresponding partition constraint is nonactive: $\rho_i = 1 - \sum_{a \in \mathcal{A}_i} x^\star_{(i,a)} > 0$. Choose any action $a_i \in \mathcal{A}_i$ and define $x' = x^\star + \rho_i e_{(i,a_i)}$, where $e_{(i,a_i)}$ is the canonical vector with nonzero coordinate (i, a_i). Then $x' \in \mathcal{P}(\mathcal{I}_t)$ and the i-th partition constraint becomes tight. By monotonicity of $\tilde{f}_t$ from Proposition 1, $\tilde{f}_t(x'; s_t) \geq \tilde{f}_t(x^\star; s_t)$. Repeating this operation for every agent i such that $\sum_{a \in \mathcal{A}_i} x^\star_{(i,a)} < 1$ yields a point $\bar{x} \in \mathcal{F}_t$ with $\tilde{f}_t(\bar{x}; s_t) \geq \tilde{f}_t(x^\star; s_t)$. Since $x^\star$ is optimal over $\mathcal{P}(\mathcal{I}_t)$, equality holds. Thus, $\max_{x \in \mathcal{P}(\mathcal{I}_t)} \tilde{f}_t(x; s_t) = \max_{x \in \mathcal{F}_t} \tilde{f}_t(x; s_t)$.

It remains to relate the optimum over $\mathcal{F}_t$ to the discrete optimum over $\mathcal{I}_t$. The set $\mathcal{F}_t$ is a product of local categorical simplices. Fix the local marginals of all agents except i. Then the PME is affine in the local vector $x_i \in \mathcal{F}_i$. Therefore, optimizing $\tilde{f}_t$ over x_i admits an optimal solution at an extreme point of $\mathcal{F}_i$. Applying this argument successively to all active agents, there exists a maximizer of $\tilde{f}_t$ over $\mathcal{F}_t$ that is an extreme point of every local categorical simplex. Such a point assigns probability one to exactly one action for each active agent and therefore corresponds to a deterministic feasible set in $\mathcal{I}_t$.

Conversely, let $A^\star \in \arg\max_{A \in \mathcal{I}_t} F_t(A; s_t)$. If $A^\star$ omits an active agent, monotonicity of F_t allows us to add any action from that agent's local action set without decreasing the utility. Repeating this step gives a feasible set $\bar{A}^\star \in \mathcal{I}_t$ that selects exactly one action per active agent and satisfies $F_t(\bar{A}^\star; s_t) \geq F_t(A^\star; s_t)$. The indicator vector $\mathbf{1}_{\bar{A}^\star}$ lies in $\mathcal{F}_t$, and by the PME definition at deterministic vertices, $\tilde{f}_t(\mathbf{1}_{\bar{A}^\star}; s_t) = F_t(\bar{A}^\star; s_t)$.

Therefore the continuous optimum over $\mathcal{F}_t$ and the discrete optimum over $\mathcal{I}_t$ have the same value. □

Lemma 3 shows that, for monotone utilities, the PME admits an optimal solution on the categorical-policy face $\mathcal{F}_t$. Therefore, restricting to factorized categorical policies does not reduce the optimal value.

The stagewise equivalence in Lemma 1 yields an exact reformulation of the finite-horizon objective in Problem 1. The following corollary shows that the original expected cumulative submodular utility is equal to the cumulative PME objective evaluated along the trajectory induced by the factorized categorical policy sequence.

Corollary 1 (Problem Equivalence)**.** *Under the factorized categorical policy sequence $\pi = \{\pi_t\}_{t=1}^T$, the finite-horizon objective in Problem 1 is exactly equal to the expected cumulative PME objective evaluated along the induced state trajectory:*

$$J(\pi) = \mathbb{E}_{\tau \sim p^\pi}\left[\sum_{t=1}^{T} \tilde{f}_t\big(x_t^\pi(s_t); s_t\big)\right], \tag{17}$$

where p^π is the trajectory distribution induced by π and the transition kernels $\{P_t\}_{t=1}^T$.

Proof. By the definition of the finite-horizon objective, $J(\pi) = \mathbb{E}_{\tau \sim p^\pi}\left[\sum_{t=1}^T F_t(A_t; s_t)\right]$, where the trajectory distribution p^π is induced by the policy sequence π and the transition kernels $\{P_t\}_{t=1}^T$. By linearity of expectation, $J(\pi) = \sum_{t=1}^T \mathbb{E}_{\tau \sim p^\pi}[F_t(A_t; s_t)]$. For each time step t, applying the tower property of conditional expectation with respect to the state s_t gives $\mathbb{E}_{\tau \sim p^\pi}[F_t(A_t; s_t)] = \mathbb{E}_{\tau \sim p^\pi}[\mathbb{E}_{A_t \sim \pi_t}[F_t(A_t; s_t) \mid s_t]]$. Conditional on s_t, the local observations $\{o_{i,t}\}_{i \in \mathcal{N}_t}$ are fixed, and the action set A_t is sampled from the factorized categorical policy π_t. Therefore, by Lemma 1, $\mathbb{E}_{A_t \sim \pi_t}[F_t(A_t; s_t) \mid s_t] = \tilde{f}_t\big(x_t^\pi(s_t); s_t\big)$. Substituting this identity into the preceding display yields $\mathbb{E}_{\tau \sim p^\pi}[F_t(A_t; s_t)] = \mathbb{E}_{\tau \sim p^\pi}\left[\tilde{f}_t\big(x_t^\pi(s_t); s_t\big)\right]$. Summing over $t = 1, \ldots, T$ gives

$$J(\pi) = \mathbb{E}_{\tau \sim p^\pi}\left[\sum_{t=1}^{T} \tilde{f}_t\big(x_t^\pi(s_t); s_t\big)\right],$$

which proves (17). □

IV. Submodular Multi-Agent Policy Learning

This section describes our policy-learning method, Submodular Multi-Agent Policy Learning (SubMAPL). By Corollary 1, policy learning for Problem 1 can be carried out through the PME objectives evaluated along the induced state trajectory. The main challenge is to construct tractable first-order feedback and update the local categorical policies without violating their simplex constraints.

Mirror ascent is a standard first-order method for constrained optimization that replaces Euclidean projection with a Bregman-divergence regularization [31], [32]. For categorical distributions, the KL divergence is particularly suitable because the resulting mirror step admits a multiplicative update that preserves nonnegativity and normalization without Euclidean projection [11], [33]. Motivated by this geometry, and in contrast to the Euclidean projected-gradient and

policy-gradient framework studied in our preliminary work [18], SubMAPL uses this PME representation to construct local marginal-gain feedback during training and updates local categorical policies through a KL-mirror step.

SubMAPL follows the centralized training-decentralized execution (CTDE) paradigm [12], [19]. In the training phase, the stage utility $F_t(\cdot\,; s_t)$ is evaluated on feasible sets that differ only in one agent's action to construct the stochastic local gradient estimator used in the KL-mirror update. In execution, each active agent applies its learned local policy using observations $o_{i,t}$.

A. Tabular Categorical Policies

For each active agent $i \in \mathcal{N}_t$, SubMAPL uses a tabular softmax parameterization of the local categorical policy. Here, "tabular" means that a separate trainable logit is maintained for each local observation-action entry [34], [35], [36]. Let $\theta^i(o,a) \in \mathbb{R}$ denote the logit of agent i for observation $o \in \mathcal{O}_i$ and action $a \in \mathcal{A}_i$. Given the local observation $o_{i,t}$, the policy is

$$\pi^i(a \mid o_{i,t}) = \frac{\exp\big(\theta^i(o_{i,t},a)\big)}{\sum_{b\in\mathcal{A}_i}\exp\big(\theta^i(o_{i,t},b)\big)},\ a \in \mathcal{A}_i. \tag{18}$$

The logits $\{\theta^i(o,a)\}_{o\in\mathcal{O}_i,\,a\in\mathcal{A}_i}$ parameterize the tabular policy π^i. Equation (18) evaluates this policy at the current observation $o_{i,t}$. The induced marginal vector satisfies $x^{\pi}_{(i,a),t}(s_t) = \pi^i\big(a \mid o_{i,t}\big),\ (i,a) \in \Omega_t$.

In open systems, the active-agent set $\mathcal{N}_t$ may vary over time. The logits of remaining agents are inherited across consecutive open-system domains. Each arriving agent is initialized using finite logits. Departing agents are excluded from the active policy domain. Hence, every feasible action of each active agent has strictly positive probability under the local softmax policy.

B. Marginal-Gain Feedback and KL-Mirror Update

At time t, condition on the current state s_t. For each active agent $i \in \mathcal{N}_t$ and $a \in \mathcal{A}_i$, define the local PME gradient coordinate by

$$g_{i,a,t}(\pi_t; s_t) = \frac{\partial \tilde{f}_t}{\partial x_{(i,a)}}(x^{\pi}_t(s_t); s_t),\ (i,a) \in \Omega_t. \tag{19}$$

By Lemma 2, this coordinate admits the marginal-gain representation $g_{i,a,t}(\pi_t; s_t) = \mathbb{E}\left[F_t((i,a) \mid A^{-i}; s_t)\right]$, where the actions in A^{-i} are sampled independently according to $\pi^j(\cdot \mid o_{j,t})$ for $j \in \mathcal{N}_t \setminus \{i\}$.

After the joint action $A_t = \{(i, a_{i,t}) : i \in \mathcal{N}_t\}$ is sampled, let $A_t^{-i} = A_t \setminus \{(i, a_{i,t})\}$. SubMAPL constructs the local marginal-gain estimator

$$\widehat{g}_{i,a,t} = F_t((i,a) \mid A_t^{-i}; s_t),\ a \in \mathcal{A}_i. \tag{20}$$

The vector $\widehat{g}_{i,\cdot,t}$ evaluates every feasible action of agent i under the same sampled actions of the other agents. Conditional on π_t and s_t, the actions in A_t^{-i} are sampled independently according to $\pi^j(\cdot \mid o_{j,t})$ for $j \in \mathcal{N}_t \setminus \{i\}$. Therefore,

$$\mathbb{E}\left[\widehat{g}_{i,a,t} \mid \pi_t, s_t\right] = g_{i,a,t}(\pi_t; s_t) = \frac{\partial \tilde{f}_t}{\partial x_{(i,a)}}(x^{\pi}_t(s_t); s_t). \tag{21}$$

Thus, $\widehat{g}_{i,\cdot,t}$ is an unbiased stochastic estimator of the local PME gradient evaluated at the current policy.

The estimator in (20) provides first-order information with respect to the local action probabilities. A direct additive probability update would require projection onto the probability simplex [31], [32]. SubMAPL instead applies KL-regularized mirror ascent, which preserves the simplex constraint without Euclidean projection and can be implemented through tabular softmax logits.

Let $\pi^{i,+}$ denote the tabular policy of agent i after the KL-mirror update at time t and before open-system migration. Given the current local policy $\pi^i(\cdot \mid o_{i,t})$ and the stochastic local gradient estimator $\widehat{g}_{i,\cdot,t}$, the local policy distribution at the current observation is updated as [31]

$$\pi^{i,+}(\cdot \mid o_{i,t}) \in \arg\max_{p\in\mathcal{F}_i}\left\{\eta\langle \widehat{g}_{i,\cdot,t}, p\rangle - D_{\mathrm{KL}}\big(p\|\pi^i(\cdot \mid o_{i,t})\big)\right\}, \tag{22}$$

where $\eta > 0$ is the step size. For two local categorical distributions $p, q \in \mathcal{F}_i$, define

$$D_{\mathrm{KL}}(p\|q) = \sum_{a\in\mathcal{A}_i} p_a \log\frac{p_a}{q_a}, \tag{23}$$

with the convention $0\log(0/q_a) = 0$. For two tabular policies π and π' defined on the current open-system domain, define

$$D_{\mathrm{KL},t}(\pi\|\pi') = \sum_{i\in\mathcal{N}_t}\sum_{o\in\mathcal{O}_i} D_{\mathrm{KL}}\left(\pi^i(\cdot \mid o)\|\pi'^i(\cdot \mid o)\right). \tag{24}$$

The first-order optimality condition yields

$$\pi^{i,+}(a \mid o_{i,t}) = \frac{\pi^i(a \mid o_{i,t})\exp(\eta\widehat{g}_{i,a,t})}{\sum_{b\in\mathcal{A}_i}\pi^i(b \mid o_{i,t})\exp(\eta\widehat{g}_{i,b,t})},\ a \in \mathcal{A}_i. \tag{25}$$

For all other observations, the local policy distributions remain unchanged:

$$\pi^{i,+}(\cdot \mid o) = \pi^i(\cdot \mid o),\ o \in \mathcal{O}_i \setminus \{o_{i,t}\}. \tag{26}$$

Hence, $\pi^{i,+}(\cdot \mid o_{i,t}) \in \mathcal{F}_i$. Since the KL-mirror step changes only the row associated with $o_{i,t}$, all other terms in the tabular-policy KL divergence remain unchanged.

Under the tabular softmax parameterization (18), (25) is implemented by the additive logit update

$$\theta^{i,+}(o_{i,t},a) = \theta^i(o_{i,t},a) + \eta\widehat{g}_{i,a,t},\ a \in \mathcal{A}_i, \tag{27}$$

where all other logit entries remain unchanged:

$$\theta^{i,+}(o,a) = \theta^i(o,a),\ o \in \mathcal{O}_i \setminus \{o_{i,t}\},\ a \in \mathcal{A}_i. \tag{28}$$

Substituting (27) into the softmax parameterization yields

$$\begin{aligned}\pi^{i,+}(a \mid o_{i,t}) &= \frac{\exp\left(\theta^i(o_{i,t},a) + \eta\widehat{g}_{i,a,t}\right)}{\sum_{b\in\mathcal{A}_i}\exp\left(\theta^i(o_{i,t},b) + \eta\widehat{g}_{i,b,t}\right)} \\ &= \frac{\exp\left(\theta^i(o_{i,t},a)\right)\exp(\eta\widehat{g}_{i,a,t})}{\sum_{b\in\mathcal{A}_i}\left(\exp\left(\theta^i(o_{i,t},b)\right)\exp(\eta\widehat{g}_{i,b,t})\right)} \\ &\overset{(i)}{=} \frac{\pi^i(a \mid o_{i,t})\exp(\eta\widehat{g}_{i,a,t})}{\sum_{b\in\mathcal{A}_i}\pi^i(b \mid o_{i,t})\exp(\eta\widehat{g}_{i,b,t})},\ a \in \mathcal{A}_i,\end{aligned} \tag{29}$$

where (i) follows by dividing the numerator and denominator by $\sum_{b\in\mathcal{A}_i}\exp(\theta^i(o_{i,t},b))$. Thus, the additive logit update exactly implements the KL-mirror update in (25). The post-update active joint policy before migration is defined by

$$\pi_t^+(\mathbf{a}_t \mid \mathbf{o}_t) = \prod_{i\in\mathcal{N}_t} \pi^{i,+}(a_i \mid o_{i,t}). \tag{30}$$

Therefore, SubMAPL performs a local full-action KL-mirror step on the current local observation of each active agent's tabular policy and implements this policy-space update through an additive logit update.

C. Open Policy Migration

The transition from time step t to $t+1$ involves the remaining agents $\mathcal{R}_{t+1}$, the arriving agents $\mathcal{N}_{t+1}^{\text{in}}$, and the departing agents $\mathcal{N}_{t+1}^{\text{out}}$. After the KL-mirror update at time t, let π_t^+ denote the post-update active joint policy on the current agent domain $\mathcal{N}_t$. Open policy migration retains the updated local policies of the remaining agents, assigns policies to the arriving agents, and excludes the departing agents from the next active policy domain.

Definition 3 (Open Policy Migration)**.** Given the post-update active joint policy π_t^+, the open policy migration map $\mathcal{M}_t$ constructs $\pi_{t+1} = \mathcal{M}_t(\pi_t^+)$ according to

$$(\pi_{t+1})^i(\cdot \mid o) = \begin{cases} \pi^{i,+}(\cdot \mid o), & i \in \mathcal{R}_{t+1}, \\ \pi^{i,0}(\cdot \mid o), & i \in \mathcal{N}_{t+1}^{\text{in}}, \end{cases} \tag{31}$$

where $\pi^{i,0}$ is the categorical policy assigned to an arriving agent i. It is represented by finite logits $\theta^{i,0} \in \mathbb{R}^{|\mathcal{O}_i|\times|\mathcal{A}_i|}$, so that $\pi^{i,0}(a \mid o) > 0, o \in \mathcal{O}_i, a \in \mathcal{A}_i$. Agents in $\mathcal{N}_{t+1}^{\text{out}}$ do not appear in the next active joint-policy domain.

The complete training procedure is summarized in Algorithm 1. At each time step, SubMAPL evaluates the marginal gain of every feasible action of each active agent while keeping the sampled actions of the other agents fixed. The learned policy remains decentralized at execution time, because each active agent samples from its own categorical policy using only its local observation.

V. Performance Analysis

This section establishes performance guarantees for SubMAPL as a PME-based KL-mirror policy-learning method in OMAS. The analysis is conducted along the state sequence generated by Algorithm 1. At each time step, the current state and open-system domain define a conditional stagewise PME objective over the current categorical-policy face. We first collect the basic gradient and mirror-ascent inequalities used in the analysis, then derive a bound on the finite-horizon objective in Problem 1 through regret analysis, and finally specialize the general bound to a tighter one when the finite-horizon objective $J(\pi)$ is itself submodular.

Algorithm 1 Submodular Multi-Agent Policy Learning (SubMAPL)

Input: horizon T, step size η, initial active parameter collection $\Theta_1 = \{\theta^{i,0} : i \in \mathcal{N}_1\}$, finite-logit migration rule for arriving agents, and stage-utility evaluators $\{F_t\}_{t=1}^T$
for $t = 1, \dots, T$ **do**
  for each active agent $i \in \mathcal{N}_t$ **do**
    Receive the local observation $o_{i,t}$
    Form $\pi^i(\cdot \mid o_{i,t})$ using (18)
    Sample $a_{i,t} \sim \pi^i(\cdot \mid o_{i,t})$
  end for
  Form $A_t = \{(i, a_{i,t}) : i \in \mathcal{N}_t\}$
  for each active agent $i \in \mathcal{N}_t$ **do**
    Set $A_t^{-i} = A_t \setminus \{(i, a_{i,t})\}$
    for each feasible action $a \in \mathcal{A}_i$ **do**
      Evaluate $\widehat{g}_{i,a,t}$ using (20) at the pre-transition state s_t
      Update $\theta^i(o_{i,t}, a) \leftarrow \theta^i(o_{i,t}, a) + \eta\widehat{g}_{i,a,t}$
    end for
  end for
  Execute A_t
  Sample the next state $s_{t+1} \sim P_t(\cdot \mid s_t, \mathbf{a}_t)$
  if $t < T$ **then**
    Observe the next active-agent set $\mathcal{N}_{t+1}$
    Apply the open policy migration in Definition 3
  end if
end for
Output: the online policy sequence $\{\pi_t\}_{t=1}^T$ and the final post-update policy π_T^+

A. Analysis of KL-Mirror and Stagewise Bound

Assumption 2 (Bounded Marginal Gains [37])**.** There exists a constant $B < \infty$ such that, for every $t \in [T]$ and state s_t,

$$0 \le F_t(e \mid A; s_t) \le B, \ \forall A \subseteq \Omega_t, \ e \in \Omega_t \setminus A. \tag{32}$$

Let $\widehat{g}_t = (\widehat{g}_{i,a,t})_{(i,a)\in\Omega_t} \in \mathbb{R}^{|\Omega_t|}$ denote the stochastic marginal-gain vector used by SubMAPL, where $\widehat{g}_{i,a,t}$ is defined in (20). Let $\mathcal{H}_t$ denote the information available before action sampling at time t, including the past trajectory, the current state s_t, the current open-system domain $(\mathcal{N}_t, \{\mathcal{A}_i\}_{i\in\mathcal{N}_t})$, and the current tabular policy π_t. Since $x_t = x_t^\pi(s_t)$, conditional on $\mathcal{H}_t$, the quantities π_t, x_t, s_t, and the current open-system domain are fixed. By (21), we have

$$\mathbb{E}\left[\widehat{g}_{i,a,t} \mid \mathcal{H}_t\right] = \frac{\partial \tilde{f}_t}{\partial x_{(i,a)}}(x_t; s_t), \ (i,a) \in \Omega_t. \tag{33}$$

Equivalently,

$$\mathbb{E}\left[\widehat{g}_t \mid \mathcal{H}_t\right] = \nabla \tilde{f}_t(x_t; s_t). \tag{34}$$

Assumption 2 also gives a uniform bound on the stochastic feedback. Indeed, for every $(i,a) \in \Omega_t$, we have $(i,a) \notin A_t^{-i}$. Hence Assumption 2 can be applied with $e = (i,a)$ and $A = A_t^{-i}$, which yields $0 \le \widehat{g}_{i,a,t} = F_t((i,a) \mid A_t^{-i}; s_t) \le B, \ (i,a) \in \Omega_t$. Consequently, for each active agent $i \in \mathcal{N}_t$, $\max_{a\in\mathcal{A}_i} |\widehat{g}_{i,a,t}| \le B$.

For any vector $g \in \mathbb{R}^{|\Omega_t|}$ indexed by the current agent-action ground set Ω_t, define the per-agent mixed norm by

$$\|g\|_{\infty,2}^2 = \sum_{i\in\mathcal{N}_t} \left(\max_{a\in\mathcal{A}_i} |g_{i,a}|\right)^2. \tag{35}$$

The norm $\|\cdot\|_{\infty,2}$ is always understood with respect to the current open-system domain $(\mathcal{N}_t, \{\mathcal{A}_i\}_{i\in\mathcal{N}_t})$. Then

$$\|\widehat{g}_t\|_{\infty,2}^2 = \sum_{i\in\mathcal{N}_t}\left(\max_{a\in\mathcal{A}_i}|\widehat{g}_{i,a,t}|\right)^2 \le \sum_{i\in\mathcal{N}_t} B^2 = |\mathcal{N}_t|B^2. \quad (36)$$

Let $N_{\max} = \max_{t\in[T]}|\mathcal{N}_t|$. Then

$$\|\widehat{g}_t\|_{\infty,2}^2 \le N_{\max}B^2,\ t\in[T]. \quad (37)$$

Lemma 4 (Auxiliary Inequalities)**.** *Consider a time step t, a state s_t, and the open-system domain at time t. Let $\tilde{f}_t(\cdot\,;s_t)$ be the PME of a normalized, monotone, submodular utility $F_t(\cdot\,;s_t)$. Then the following inequalities hold.*

1) ***Categorical-face first-order bound [29]:*** *For any $x, y \in \mathcal{F}_t$,*

$$\frac{1}{2}\tilde{f}_t(y;s_t) - \tilde{f}_t(x;s_t) \le \frac{1}{2}\left\langle \nabla\tilde{f}_t(x;s_t), y-x\right\rangle. \quad (38)$$

2) ***KL mirror-ascent inequality [31]:*** *Suppose $x \in \mathcal{F}_t$ satisfies $x_{(i,a)} > 0$ for all $i \in \mathcal{N}_t$ and $a \in \mathcal{A}_i$. Let $x^+ \in \mathcal{F}_t$ be generated by*

$$x^+ \in \arg\max_{z\in\mathcal{F}_t}\{\eta\langle g, z\rangle - D_{\mathrm{KL}}(z\|x)\}, \quad (39)$$

where $\eta > 0$ and

$$D_{\mathrm{KL}}(z\|x) = \sum_{i\in\mathcal{N}_t}\sum_{a\in\mathcal{A}_i} z_{(i,a)}\log\frac{z_{(i,a)}}{x_{(i,a)}}.$$

Then, for any $u \in \mathcal{F}_t$,

$$\langle g, u-x\rangle \le \frac{D_{\mathrm{KL}}(u\|x) - D_{\mathrm{KL}}(u\|x^+)}{\eta} + \frac{\eta}{2}\|g\|_{\infty,2}^2. \quad (40)$$

Proof. (1) Let $A \sim \mathcal{D}_t(\cdot \mid x)$ and $Y \sim \mathcal{D}_t(\cdot \mid y)$ be sampled independently. Since $x, y \in \mathcal{F}_t$, both A and Y select exactly one agent-action pair for each active agent. Define $A = \{e_i : i \in \mathcal{N}_t\}$ and $Y = \{u_i : i \in \mathcal{N}_t\}$, where $e_i, u_i \in \Omega_{i,t}$. Let $A^{-i} = A \setminus \{e_i\}$. By monotonicity, $F_t(Y;s_t) - F_t(A;s_t) \le F_t(A\cup Y;s_t) - F_t(A;s_t)$.

Fix an arbitrary ordering $\{i_1,\dots,i_m\}$ of $\mathcal{N}_t$, where $m = |\mathcal{N}_t|$. By telescoping, $F_t(A\cup Y;s_t) - F_t(A;s_t) = \sum_{\ell=1}^m F_t\left(u_{i_\ell} \mid A\cup\{u_{i_1},\dots,u_{i_{\ell-1}}\};s_t\right)$. Since $A^{-i_\ell} \subseteq A\cup\{u_{i_1},\dots,u_{i_{\ell-1}}\}$, submodularity implies $F_t\left(u_{i_\ell} \mid A\cup\{u_{i_1},\dots,u_{i_{\ell-1}}\};s_t\right) \le F_t(u_{i_\ell} \mid A^{-i_\ell};s_t)$. Therefore,

$$F_t(Y;s_t) - F_t(A;s_t) \le \sum_{i\in\mathcal{N}_t} F_t(u_i \mid A^{-i};s_t). \quad (41)$$

Next, by normalization and telescoping along the same ordering, $F_t(A;s_t) = \sum_{\ell=1}^m F_t\left(e_{i_\ell} \mid \{e_{i_1},\dots,e_{i_{\ell-1}}\};s_t\right)$. Since $\{e_{i_1},\dots,e_{i_{\ell-1}}\} \subseteq A^{-i_\ell}$, submodularity gives $F_t\left(e_{i_\ell} \mid \{e_{i_1},\dots,e_{i_{\ell-1}}\};s_t\right) \ge F_t(e_{i_\ell} \mid A^{-i_\ell};s_t)$. Hence,

$$\sum_{i\in\mathcal{N}_t} F_t(e_i \mid A^{-i};s_t) \le F_t(A;s_t). \quad (42)$$

Combining (41) and (42) yields

$$\begin{aligned} &F_t(Y;s_t) - 2F_t(A;s_t) \\ &\le \sum_{i\in\mathcal{N}_t}\Big[F_t(u_i \mid A^{-i};s_t) - F_t(e_i \mid A^{-i};s_t)\Big]. \end{aligned} \quad (43)$$

Taking expectations in (43) over the independent samples $A \sim \mathcal{D}_t(\cdot \mid x)$ and $Y \sim \mathcal{D}_t(\cdot \mid y)$ gives

$$\begin{aligned} &\tilde{f}_t(y;s_t) - 2\tilde{f}_t(x;s_t) \\ &\le \sum_{i\in\mathcal{N}_t}\sum_{a\in\mathcal{A}_i} y_{(i,a)}\mathbb{E}_{A^{-i}\sim\mathcal{D}_t^{-i}(\cdot|x)}\left[F_t((i,a) \mid A^{-i};s_t)\right] \\ &\quad - \sum_{i\in\mathcal{N}_t}\sum_{a\in\mathcal{A}_i} x_{(i,a)}\mathbb{E}_{A^{-i}\sim\mathcal{D}_t^{-i}(\cdot|x)}\left[F_t((i,a) \mid A^{-i};s_t)\right]. \end{aligned}$$

Here, under $\mathcal{D}_t(\cdot \mid x)$, the selected action of agent i is independent of A^{-i} and has marginal distribution $(x_{(i,a)})_{a\in\mathcal{A}_i}$; similarly, under $\mathcal{D}_t(\cdot \mid y)$, the selected action of agent i has marginal distribution $(y_{(i,a)})_{a\in\mathcal{A}_i}$ and is independent of A. By Lemma 2, the expected marginal gain is the corresponding PME coordinate derivative. Therefore,

$$\tilde{f}_t(y;s_t) - 2\tilde{f}_t(x;s_t) \le \left\langle \nabla\tilde{f}_t(x;s_t), y-x\right\rangle.$$

Dividing by 2 proves (38).

(2) The optimality condition of (39) gives the three-point inequality

$$\eta\langle g, u-x^+\rangle \le D_{\mathrm{KL}}(u\|x) - D_{\mathrm{KL}}(u\|x^+) - D_{\mathrm{KL}}(x^+\|x). \quad (44)$$

For each active agent $i \in \mathcal{N}_t$, Hölder's inequality gives [31], [32]

$$\begin{aligned} &\eta\sum_{a\in\mathcal{A}_i} g_{i,a}\left(x^+_{(i,a)} - x_{(i,a)}\right) \\ &\le \eta\left(\max_{a\in\mathcal{A}_i}|g_{i,a}|\right)\sum_{a\in\mathcal{A}_i}\left|x^+_{(i,a)} - x_{(i,a)}\right|. \end{aligned} \quad (45)$$

By Young's inequality,

$$\begin{aligned} &\eta\left(\max_{a\in\mathcal{A}_i}|g_{i,a}|\right)\sum_{a\in\mathcal{A}_i}\left|x^+_{(i,a)} - x_{(i,a)}\right| \\ &\le \frac{\eta^2}{2}\left(\max_{a\in\mathcal{A}_i}|g_{i,a}|\right)^2 + \frac{1}{2}\left(\sum_{a\in\mathcal{A}_i}\left|x^+_{(i,a)} - x_{(i,a)}\right|\right)^2. \end{aligned}$$

Pinsker's inequality, applied to the two categorical distributions over $\mathcal{A}_i$, gives

$$\frac{1}{2}\left(\sum_{a\in\mathcal{A}_i}\left|x^+_{(i,a)} - x_{(i,a)}\right|\right)^2 \le \sum_{a\in\mathcal{A}_i} x^+_{(i,a)}\log\frac{x^+_{(i,a)}}{x_{(i,a)}}.$$

Thus,

$$\begin{aligned} &\eta\sum_{a\in\mathcal{A}_i} g_{i,a}\left(x^+_{(i,a)} - x_{(i,a)}\right) \\ &\le \frac{\eta^2}{2}\left(\max_{a\in\mathcal{A}_i}|g_{i,a}|\right)^2 + \sum_{a\in\mathcal{A}_i} x^+_{(i,a)}\log\frac{x^+_{(i,a)}}{x_{(i,a)}}. \end{aligned}$$

Summing over $i \in \mathcal{N}_t$ yields

$$\eta\langle g, x^+ - x\rangle \le D_{\mathrm{KL}}(x^+\|x) + \frac{\eta^2}{2}\|g\|_{\infty,2}^2. \quad (46)$$

Combining (44) and (46), we obtain

$$\begin{aligned} \eta\langle g, u-x\rangle &= \eta\langle g, u-x^+\rangle + \eta\langle g, x^+ - x\rangle \\ &\le D_{\mathrm{KL}}(u\|x) - D_{\mathrm{KL}}(u\|x^+) + \frac{\eta^2}{2}\|g\|_{\infty,2}^2. \end{aligned}$$

Dividing by η proves (40). □

B. Suboptimality Gap in the General Case

Dynamic approximation regret compares the cumulative utility attained by an algorithm with that of a sequence of stagewise optimal allocations [11], [17]. Existing formulations for multi-agent online submodular coordination consider a fixed agent-action domain. However, the consecutive comparator policies in OMAS may belong to different policy domains as agents arrive and depart. We introduce an open-system variant that measures comparator variation after mapping the previous comparator to the new domain through policy migration.

Given the current state s_t and open-system domain, define the stagewise discrete optimum $F_t^\star(s_t) = \max_{A\in\mathcal{I}_t} F_t(A; s_t)$. Let $x_t^\star \in \arg\max_{x\in\mathcal{F}_t} \tilde{f}_t(x; s_t)$. By Lemma 3, $\tilde{f}_t(x_t^\star; s_t) = F_t^\star(s_t)$. The sequence $\{x_t^\star\}_{t=1}^T$ may vary with the state and the active-agent domain. Define the tabular policy $\pi_t^\star$ associated with $x_t^\star$ by

$$(\pi_t^\star)^i(a \mid o) = \begin{cases} x^\star_{(i,a),t}, & o = o_{i,t}, \\ \pi^i(a \mid o), & o \neq o_{i,t}, \end{cases} \; i \in \mathcal{N}_t,\; o \in \mathcal{O}_i,\; a \in \mathcal{A}_i. \tag{47}$$

By (24), $D_{\mathrm{KL},t}(\pi_t^\star \| \pi_t) = D_{\mathrm{KL}}(x_t^\star \| x_t)$.

The transition from time step t to $t+1$ consists of remaining agents $\mathcal{R}_{t+1}$, arriving agents $\mathcal{N}_{t+1}^{\mathrm{in}}$, and departing agents $\mathcal{N}_{t+1}^{\mathrm{out}}$. SubMAPL first obtains the post-update tabular policy π_t^+ through the KL-mirror update in (25)-(26). It then applies the open policy migration in Definition 3: $\pi_{t+1} = \mathcal{M}_t(\pi_t^+)$. Let $\widehat{\pi} = \{\pi_t\}_{t=1}^T$ denote the online policy sequence generated by SubMAPL, and let $p^{\widehat{\pi}}$ denote the trajectory distribution induced by $\widehat{\pi}$ and the transition kernels $\{P_t\}_{t=1}^T$. Following KL-potential tracking analyses on a fixed domain [33], [38], we define the open-system KL tracking variation as

$$\mathcal{V}_T^{\mathrm{open}} = \mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\Bigg[\sum_{t=1}^{T-1}\Big(D_{\mathrm{KL},t+1}(\pi_{t+1}^\star \| \pi_{t+1}) - D_{\mathrm{KL},t}(\pi_t^\star \| \pi_t^+)\Big)_+\Bigg]. \tag{48}$$

This variation records the increase in the KL tracking potential between consecutive time steps while allowing both the stagewise benchmark and the open-system policy domain to change. In particular, it incorporates agent arrivals and departures through $\mathcal{R}_{t+1}$, $\mathcal{N}_{t+1}^{\mathrm{in}}$, and $\mathcal{N}_{t+1}^{\mathrm{out}}$. When the active-agent set is fixed, $\mathcal{M}_t$ is the identity map, and $\mathcal{V}_T^{\mathrm{open}}$ reduces to the corresponding KL tracking variation on a fixed tabular-policy domain.

Definition 4 (Open-System Dynamic Approximation-Regret)**.** The open-system dynamic $1/2$-approximation regret is defined as

$$\mathrm{Reg}_T^{1/2} = \sum_{t=1}^T \mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[\frac{1}{2}F_t^\star(s_t) - \tilde{f}_t(x_t; s_t)\right]. \tag{49}$$

By Lemma 1, $\tilde{f}_t(x_t; s_t)$ is the conditional expected stage utility of the factorized categorical policy π_t. Thus, $\mathrm{Reg}_T^{1/2}$ measures its cumulative approximation gap relative to the time-varying stagewise optimal values.

Lemma 5 (Open-System KL-Mirror Approximation-Regret Bound)**.** *Let $\{\pi_t\}_{t=1}^T$ be the policy sequence generated by Algorithm 1. Under Assumption 1 and Assumption 2, for any constant step size $\eta > 0$,*

$$\mathrm{Reg}_T^{1/2} \le \frac{\mathbb{E}_{s_1\sim\mu_1}\left[D_{\mathrm{KL},1}\left(\pi_1^\star \| \pi_1\right)\right] + \mathcal{V}_T^{\mathrm{open}}}{2\eta} + \frac{\eta T N_{\max} B^2}{4}. \tag{50}$$

Proof. For each time step t, the categorical-face first-order bound in Lemma 4 gives

$$\frac{1}{2}F_t^\star(s_t) - \tilde{f}_t(x_t; s_t) \le \frac{1}{2}\left\langle \nabla\tilde{f}_t(x_t; s_t), x_t^\star - x_t\right\rangle. \tag{51}$$

Since x_t and $x_t^\star$ are measurable with respect to the pre-sampling history $\mathcal{H}_t$, taking expectations and using (34) yields

$$\begin{aligned}\mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[\frac{1}{2}F_t^\star(s_t) - \tilde{f}_t(x_t; s_t)\right] \\ \le \frac{1}{2}\mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[\langle \widehat{g}_t, x_t^\star - x_t\rangle\right].\end{aligned} \tag{52}$$

Since $\mathcal{F}_t = \prod_{i\in\mathcal{N}_t}\mathcal{F}_i$, the local KL-mirror updates in (22) jointly realize the product-space update in Lemma 4. Apply the KL mirror-ascent inequality with $x = x_t$, $u = x_t^\star$, and $g = \widehat{g}_t$, where the post-update vector x_t^+ has coordinates

$$x^+_{(i,a),t} = \pi^{i,+}(a \mid o_{i,t}),\;\; (i,a) \in \Omega_t.$$

By (24) and (47), together with the fact that π_t^+ and π_t coincide at all noncurrent observations,

$$\begin{aligned} D_{\mathrm{KL}}(x_t^\star \| x_t) &= D_{\mathrm{KL},t}\left(\pi_t^\star \| \pi_t\right), \\ D_{\mathrm{KL}}(x_t^\star \| x_t^+) &= D_{\mathrm{KL},t}\left(\pi_t^\star \| \pi_t^+\right). \end{aligned} \tag{53}$$

Therefore,

$$\langle \widehat{g}_t, x_t^\star - x_t\rangle \le \frac{D_{\mathrm{KL},t}\left(\pi_t^\star \| \pi_t\right) - D_{\mathrm{KL},t}\left(\pi_t^\star \| \pi_t^+\right)}{\eta} + \frac{\eta}{2}\|\widehat{g}_t\|_{\infty,2}^2. \tag{54}$$

By the uniform gradient bound in (37),

$$\|\widehat{g}_t\|_{\infty,2}^2 \le N_{\max} B^2. \tag{55}$$

Combining (52), (54), and (55) gives

$$\begin{aligned}\mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[\frac{1}{2}F_t^\star(s_t) - \tilde{f}_t(x_t; s_t)\right] \\ \le \frac{\mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[D_{\mathrm{KL},t}\left(\pi_t^\star \| \pi_t\right) - D_{\mathrm{KL},t}\left(\pi_t^\star \| \pi_t^+\right)\right]}{2\eta} + \frac{\eta N_{\max} B^2}{4}.\end{aligned} \tag{56}$$

For compactness, define $\Phi_t = D_{\mathrm{KL},t}\left(\pi_t^\star \| \pi_t\right)$, and $\Gamma_t = D_{\mathrm{KL},t}\left(\pi_t^\star \| \pi_t^+\right)$. For $t = 1, \ldots, T-1$, we have

$$\begin{aligned}\Phi_t - \Gamma_t &= \Phi_t - \Phi_{t+1} + \Phi_{t+1} - \Gamma_t \\ &\le \Phi_t - \Phi_{t+1} + [\Phi_{t+1} - \Gamma_t]_+.\end{aligned} \tag{57}$$

By the definition of the two potentials,

$$[\Phi_{t+1} - \Gamma_t]_+ = \left[D_{\mathrm{KL},t+1}\left(\pi_{t+1}^\star \| \pi_{t+1}\right) - D_{\mathrm{KL},t}\left(\pi_t^\star \| \pi_t^+\right)\right]_+. \tag{58}$$

Summing (57) over $t = 1, \dots, T-1$ gives

$$\sum_{t=1}^{T-1}(\Phi_t - \Gamma_t) \le \Phi_1 - \Phi_T + \sum_{t=1}^{T-1}[\Phi_{t+1} - \Gamma_t]_+ . \tag{59}$$

Adding the terminal term $\Phi_T - \Gamma_T$ to both sides and using $\Gamma_T \ge 0$ yields

$$\sum_{t=1}^{T}(\Phi_t - \Gamma_t) \le \Phi_1 + \sum_{t=1}^{T-1}[\Phi_{t+1} - \Gamma_t]_+ . \tag{60}$$

Taking expectations with respect to $\tau \sim p^{\widehat{\pi}}$ and using the definition of $\mathcal{V}_T^{\text{open}}$ in (48), we obtain

$$\sum_{t=1}^{T}\mathbb{E}_{\tau\sim p^{\widehat{\pi}}}[\Phi_t - \Gamma_t] \le \mathbb{E}_{\tau\sim p^{\widehat{\pi}}}[\Phi_1] + \mathcal{V}_T^{\text{open}}. \tag{61}$$

Summing (56) over $t = 1, \dots, T$ and applying (61) gives

$$\sum_{t=1}^{T}\mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[\frac{1}{2}F_t^\star(s_t) - \tilde{f}_t(x_t; s_t)\right] \le \frac{\mathbb{E}_{s_1\sim\mu_1}\left[D_{\mathrm{KL},1}\left(\pi_1^\star \| \pi_1\right)\right] + \mathcal{V}_T^{\text{open}}}{2\eta} + \frac{\eta T N_{\max} B^2}{4}. \tag{62}$$

The left-hand side equals $\mathrm{Reg}_T^{1/2}$ by Definition 4, which proves (50). □

Lemma 5 controls the cumulative approximation gap relative to the time-varying stagewise optimal values. However, Problem 1 evaluates the cumulative policy value induced over the full trajectory. We next connect the stagewise approximation-regret guarantee in Lemma 5 to the finite-horizon objective of Problem 1. By Corollary 1,

$$J(\widehat{\pi}) = \mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[\sum_{t=1}^{T}\tilde{f}_t(x_t; s_t)\right]. \tag{63}$$

Let $J^\star = \max_\pi J(\pi)$ be the optimal finite-horizon value over the considered class of decentralized factorized policy sequences. For readability, denote the bound in (50) by

$$\mathcal{C}_T = \frac{\mathbb{E}_{s_1\sim\mu_1}\left[D_{\mathrm{KL},1}\left(\pi_1^\star \| \pi_1\right)\right] + \mathcal{V}_T^{\text{open}}}{2\eta} + \frac{\eta T N_{\max} B^2}{4}. \tag{64}$$

Moreover, define the finite-horizon benchmark mismatch

$$\Delta_T = \left[J^\star - \mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[\sum_{t=1}^{T}F_t^\star(s_t)\right]\right]_+ . \tag{65}$$

Theorem 1 (Finite-Horizon Performance Bound)**.** *Let $\widehat{\pi} = \{\pi_t\}_{t=1}^T$ be the policy sequence generated by Algorithm 1. Under Assumption 1 and Assumption 2, we have*

$$\frac{1}{2}J^\star - J(\widehat{\pi}) \le \mathcal{C}_T + \frac{1}{2}\Delta_T, \tag{66}$$

where $\mathcal{C}_T$ and Δ_T are defined in (64) and (65), respectively.

Proof. By Definition 4 and Lemma 5,

$$J(\widehat{\pi}) \ge \frac{1}{2}\mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[\sum_{t=1}^{T}F_t^\star(s_t)\right] - \mathcal{C}_T. \tag{67}$$

By the definition of Δ_T in (65),

$$\mathbb{E}_{\tau\sim p^{\widehat{\pi}}}\left[\sum_{t=1}^{T}F_t^\star(s_t)\right] \ge J^\star - \Delta_T. \tag{68}$$

Substituting (68) into (67) and rearranging yields (66). □

The mismatch term Δ_T captures the discrepancy between the optimal finite-horizon value and the cumulative stagewise optimal values along the trajectory induced by SubMAPL. Since the stagewise KL-mirror analysis does not control this term without additional temporal structure, Theorem 1 provides a finite-horizon decomposition rather than a uniform approximation guarantee.

C. Suboptimality Gap With Submodular Finite-Horizon Utility

We next identify a structural setting in which the stagewise approximation-regret bound yields a finite-horizon approximation guarantee. In coverage and information-collection problems, the stage utility can often be represented as the marginal gain of a trajectory-level submodular function with respect to the accumulated history [14]. Under such a time-expanded representation, if the ground set and the feasible families are independent of the executed policy, submodularity connects the cumulative stagewise oracle values along the SubMAPL trajectory with the finite-horizon optimum $J^\star$.

Assumption 3 (Time-Expanded Submodular Utility [14])**.** There exist a policy-independent time-expanded ground set $\bar{\Omega}_{1:T}$, policy-independent feasible families $\{\bar{\mathcal{I}}_t\}_{t=1}^T$, and a normalized, monotone, submodular function $\bar{F} : 2^{\bar{\Omega}_{1:T}} \to \mathbb{R}_{\ge 0}$. For every feasible execution, each $A_t \in \mathcal{I}_t$ corresponds to a time-expanded action set $\bar{A}_t \in \bar{\mathcal{I}}_t$, and the state s_t contains the accumulated history $\bar{A}_{1:t-1} = \bigcup_{\tau=1}^{t-1}\bar{A}_\tau$. The stage utility satisfies $F_t(A_t; s_t) = \bar{F}(\bar{A}_t \mid \bar{A}_{1:t-1}) = \bar{F}(\bar{A}_{1:t-1} \cup \bar{A}_t) - \bar{F}(\bar{A}_{1:t-1})$. Moreover, every $\bar{A}'_t \in \bar{\mathcal{I}}_t$ corresponds to some feasible $A'_t \in \mathcal{I}_t$.

Assumption 3 is used to convert the stagewise approximation-regret bound into a finite-horizon guarantee. It holds for trajectory-level coverage and information-collection objectives whose cumulative utility is represented by a monotone submodular function of the selected time-expanded items [14], [39], [40].

Example 1 (Static Sensing Coverage Utility [41])**.** Consider a multi-agent coverage problem over a finite target set $\mathcal{Q}$. At time step t, each agent-action pair $(i, a) \in \Omega_t$ covers a sensing region $C_t(i, a) \subseteq \mathcal{Q}$, that does not depend on the current state. Assume that Ω_t and the sensing regions $\{C_t(i, a)\}_{(i,a)\in\Omega_t}$ are determined independently of the executed policy.

Define the time-expanded ground set $\bar{\Omega}_{1:T} = \{(i, a, t) : t \in [T],\ (i, a) \in \Omega_t\}$. For each feasible $A_t \in \mathcal{I}_t$, let $\bar{A}_t = \{(i, a, t) : (i, a) \in A_t\}$, and let $\bar{\mathcal{I}}_t = \{\bar{A}_t : A_t \in \mathcal{I}_t\}$. Define $\bar{A}_{1:t} = \bigcup_{\tau=1}^{t}\bar{A}_\tau$ and $\bar{A}_{1:0} = \emptyset$. Given target weights $w_q \ge 0$, define

$$\bar{F}(S) = \sum_{q\in\mathcal{Q}} w_q \mathbf{1}\left\{q \in \bigcup_{(i,a,t)\in S} C_t(i, a)\right\},\ S \subseteq \bar{\Omega}_{1:T}.$$

Then $\bar{F}$ is normalized, monotone, and submodular. If s_t records the targets covered before time t, the stage utility satisfies $F_t(A_t; s_t) = \bar{F}(\bar{A}_t \mid \bar{A}_{1:t-1})$, and therefore

$$\sum_{t=1}^{T} F_t(A_t; s_t) = \bar{F}(\bar{A}_{1:T}).$$

Thus, Assumption 3 holds.

Theorem 2 (Finite-Horizon Approximation under Time-Expanded Submodularity)**.** *Let $\widehat{\pi}$ be the policy sequence generated by Algorithm 1. Suppose Assumption 1, Assumption 2, and Assumption 3 hold. Then*

$$J(\widehat{\pi}) \geq \frac{1}{3} J^\star - \frac{2}{3} \mathcal{C}_T. \tag{69}$$

Proof. Consider an arbitrary trajectory generated by $\widehat{\pi}$. Let $\{\bar{A}_t\}_{t=1}^T$ denote the corresponding time-expanded action sets, and let $\{s_t\}_{t=1}^T$ denote the state sequence. Let $\{\bar{A}'_t\}_{t=1}^T$ be any fixed feasible reference sequence with $\bar{A}'_t \in \bar{\mathcal{I}}_t$ for each t, and define $\bar{A}'_{1:t} = \bigcup_{\tau=1}^{t} \bar{A}'_\tau$ with $\bar{A}'_{1:0} = \emptyset$. By Assumption 3, for each $\bar{A}'_t \in \bar{\mathcal{I}}_t$, there exists $A'_t \in \mathcal{I}_t$ whose corresponding time-expanded action set is $\bar{A}'_t$.

By monotonicity of $\bar{F}$,

$$\bar{F}(\bar{A}'_{1:T}) \leq \bar{F}(\bar{A}_{1:T}) + \bar{F}(\bar{A}'_{1:T} \mid \bar{A}_{1:T}). \tag{70}$$

Using the chain rule for marginal gains, we have

$$\begin{aligned} \bar{F}(\bar{A}'_{1:T} \mid \bar{A}_{1:T}) &= \sum_{t=1}^{T} \bar{F}\left(\bar{A}'_t \mid \bar{A}_{1:T} \cup \bar{A}'_{1:t-1}\right) \\ &\leq \sum_{t=1}^{T} \bar{F}\left(\bar{A}'_t \mid \bar{A}_{1:t-1}\right), \end{aligned} \tag{71}$$

where the inequality follows from $\bar{A}_{1:t-1} \subseteq \bar{A}_{1:T} \cup \bar{A}'_{1:t-1}$ and the diminishing-returns property of the submodular function $\bar{F}$.

By Assumption 3, the marginal gain of the reference time-expanded action set $\bar{A}'_t$ along the SubMAPL history is exactly the corresponding stage utility:

$$\bar{F}(\bar{A}'_t \mid \bar{A}_{1:t-1}) = F_t(A'_t; s_t) \leq F_t^\star(s_t). \tag{72}$$

Combining (70)–(72) gives, for the trajectory under consideration,

$$\bar{F}(\bar{A}'_{1:T}) \leq \bar{F}(\bar{A}_{1:T}) + \sum_{t=1}^{T} F_t^\star(s_t). \tag{73}$$

We next take expectation with respect to the trajectory distribution induced by SubMAPL. By Assumption 3, the cumulative utility along the SubMAPL trajectory telescopes:

$$\begin{aligned} \sum_{t=1}^{T} F_t(A_t; s_t) &= \sum_{t=1}^{T} \bar{F}(\bar{A}_t \mid \bar{A}_{1:t-1}) \\ &= \sum_{t=1}^{T} \left[\bar{F}(\bar{A}_{1:t}) - \bar{F}(\bar{A}_{1:t-1})\right] \\ &= \bar{F}(\bar{A}_{1:T}), \end{aligned}$$

where $\bar{F}(\emptyset) = 0$ because $\bar{F}$ is normalized. Therefore,

$$J(\widehat{\pi}) = \mathbb{E}_{\tau \sim p^{\widehat{\pi}}}\left[\bar{F}(\bar{A}_{1:T})\right].$$

Taking expectation in (73) over $\tau \sim p^{\widehat{\pi}}$ gives

$$\bar{F}(\bar{A}'_{1:T}) \leq J(\widehat{\pi}) + \mathbb{E}_{\tau \sim p^{\widehat{\pi}}}\left[\sum_{t=1}^{T} F_t^\star(s_t)\right]. \tag{74}$$

Since (74) holds for every feasible reference sequence $\{\bar{A}'_t\}_{t=1}^T$ with $\bar{A}'_t \in \bar{\mathcal{I}}_t$, it also holds after maximizing the left-hand side over all such reference sequences:

$$\max_{\bar{A}'_t \in \bar{\mathcal{I}}_t,\ t \in [T]} \bar{F}(\bar{A}'_{1:T}) \leq J(\widehat{\pi}) + \mathbb{E}_{\tau \sim p^{\widehat{\pi}}}\left[\sum_{t=1}^{T} F_t^\star(s_t)\right]. \tag{75}$$

Since the time-expanded feasible families are policy-independent, every realization generated by an admissible policy π satisfies $\bar{A}_t \in \bar{\mathcal{I}}_t$ for all $t \in [T]$. Therefore, for any admissible policy π,

$$J(\pi) = \mathbb{E}_{\tau \sim p^{\pi}}\left[\bar{F}(\bar{A}_{1:T})\right] \leq \max_{\bar{A}'_t \in \bar{\mathcal{I}}_t,\ t \in [T]} \bar{F}(\bar{A}'_{1:T}).$$

Taking the maximum over π gives

$$J^\star \leq \max_{\bar{A}'_t \in \bar{\mathcal{I}}_t,\ t \in [T]} \bar{F}(\bar{A}'_{1:T}).$$

Combining this inequality with (75) yields

$$J^\star \leq J(\widehat{\pi}) + \mathbb{E}_{\tau \sim p^{\widehat{\pi}}}\left[\sum_{t=1}^{T} F_t^\star(s_t)\right]. \tag{76}$$

It remains to use the stagewise approximation-regret bound. From (67), we have

$$\mathbb{E}_{\tau \sim p^{\widehat{\pi}}}\left[\sum_{t=1}^{T} F_t^\star(s_t)\right] \leq 2J(\widehat{\pi}) + 2\mathcal{C}_T. \tag{77}$$

Substituting (77) into (76) gives

$$J^\star \leq 3J(\widehat{\pi}) + 2\mathcal{C}_T.$$

Rearranging proves (69). □

Lemma 5 establishes an open-system dynamic $1/2$-approximation-regret bound for SubMAPL relative to the time-varying stagewise PME maximizers. These results characterize the stagewise tracking performance of SubMAPL along its induced state trajectory. However, the objective in Problem 1 is the optimal finite-horizon policy value, which need not coincide with the cumulative stagewise optimal values. Thus, we quantify this discrepancy through the mismatch term Δ_T in (66). Under Assumption 3, the policy-independent time-expanded submodular representation connects the stagewise benchmark to the finite-horizon optimum, yielding the approximation guarantee in Theorem 2.

VI. Numerical Experiments

We evaluate SubMAPL on a multi-agent coverage task [14]. SubMAPL is trained with a fixed (time-invariant) agent set and evaluated on open systems with a time-varying agent set.

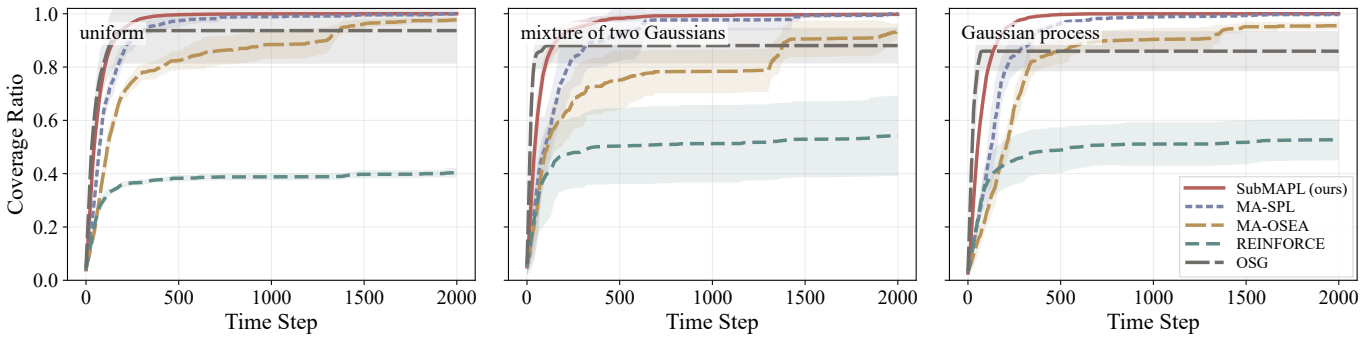

Fig. 2: Coverage ratio under the controlled $5 \to 2 \to 5$ active-agent profile. From left to right: uniform, mixture of two Gaussians, and Gaussian process fields. Curves show the means, and shaded regions denote 95% confidence intervals over shared evaluation scenarios.

A. Experimental Setup

The workspace $\mathcal{V}$ is a 30×30 grid with $N_{\max} = 5$ agents. To model spatially heterogeneous information, we weigh grid cells according to a probability distribution. We consider a uniform field, a mixture of two isotropic Gaussian components, and a positive log-Gaussian process field generated using a separable radial basis function (RBF) kernel. Cell weights remain constant throughout each experiment. All methods are evaluated with the same cell weights and initial agent positions.

The global state s_t contains the active-agent set $\mathcal{N}_t$, the current agent positions, and the accumulated coverage set $\mathcal{C}_{t-1}^{\text{hist}} = \bigcup_{k=1}^{t-1} \mathcal{C}_k(A_k; s_k)$, which records all cells sensed before the action at time t. The complete coverage history is maintained by the environment for state transitions and utility evaluation but is not directly available to individual agents during decentralized execution. Each agent observes only a local encoding of the coverage status described below.

Each agent receives a discrete local observation $o_{i,t}$. It contains: 1) the index of the agent's current 6×6 region in a partition of the workspace into 25 regions; 2) the relative positions of at most two nearest active neighbors within the communication radius $r_{\text{com}} = 2$; 3) the states of the agent's current cell and its four axially adjacent cells, where each cell is classified as already covered or outside the workspace, uncovered with low information density, or uncovered with high information density; and 4) the agent's previous action. Thus, an agent observes only nearby coverage and neighbor information, rather than the complete coverage history or the positions of all agents.

At each step t, active agent $i \in \mathcal{N}_t$ selects a motion action $a_{i,t} \in \mathcal{A}_i = \{\text{idle}, \text{left}, \text{right}, \text{up}, \text{down}\}$, which determines its next position. Let $D_i(a_{i,t}; s_t)$ denote the sensing region of agent i after applying action $a_{i,t}$ at state s_t, which includes all cells within Chebyshev distance $r_{\text{cov}} = 1$ from the agent's new position. Hence, $D_i(a_{i,t}; s_t)$ forms a 3×3 neighborhood and determines the information collected by agent i.

The cells sensed as a result of executing A_t at state s_t are

$$\mathcal{C}_t(A_t; s_t) = \bigcup_{i \in \mathcal{N}_t} D_i(a_{i,t}; s_t). \tag{78}$$

Let $\mathcal{C}_{t-1}^{\text{hist}} = \bigcup_{k=1}^{t-1} \mathcal{C}_k(A_k; s_k)$ denote all cells covered before the action at time t. We define the stage utility as the weighted information collected from uncovered cells:

$$F_t(A_t; s_t) = \sum_{v \in \mathcal{C}_t(A_t; s_t) \setminus \mathcal{C}_{t-1}^{\text{hist}}} \rho(v), \tag{79}$$

where $\rho(v) \geq 0$ is the weight of cell $v \in \mathcal{V}$. Utility $F_t(\cdot; s_t)$ is a normalized, monotone, and submodular set function over the active agent-action ground set Ω_t, satisfying Assumption 1 [25], [41]. We measure performance as the weighted coverage ratio

$$\text{CovRatio}_t = \frac{\sum_{v \in \mathcal{C}_t^{\text{hist}}} \rho(v)}{\sum_{v \in \mathcal{V}} \rho(v)}. \tag{80}$$

We compare our proposed algorithm SubMAPL (Algorithm 1) with four benchmarks.

***MA-SPL* [17]:** a policy-based continuous-extension method that employs a submodular surrogate gradient, a minimum-marginal-gain correction, neighbor consensus, and Euclidean projection.

***MA-OSEA* [11]:** a multilinear-extension method based on curvature-aware surrogate gradients, neighbor consensus, uniform mixing, and a projection-free KL update.

***REINFORCE* [42]:** a tabular return-based policy-gradient method trained using the global newly covered information as the stage reward.

***OSG* [2]:** a centralized online greedy method that sequentially selects one action for each agent according to their current marginal coverage gains following a fixed order.

SubMAPL and REINFORCE are trained in a closed-system setting for 3000 episodes with horizon $T = 100$ steps. Each evaluation trajectory has length $T = 2000$. MA-SPL and MA-OSEA are initialized with uniform policies and perform their native online updates throughout the evaluation horizon. At each time step, OSG processes the active agents in a fixed order, and each active agent greedily maximizes F_t over the active agent set at each step t. A complete communication graph is used for MA-SPL and MA-OSEA. SubMAPL does not perform inter-agent communication or consensus during decentralized execution. Each agent instead locally observes the relative positions of at most the two nearest active agents within communication radius of $r_{\text{com}} = 2$.

For the methods trained in the closed-system setting, each agent retains its finite-logit policy table while inactive and restores it upon reactivation. We use five training seeds and five

TABLE II: Normalized areas under the coverage curves in the open-system evaluation. Higher values indicate faster overall information acquisition.

Method	Uniform	Mixture of two Gaussians	Gaussian process	Average
SubMAPL	**0.966**	**0.960**	**0.963**	**0.963**
MA-SPL	0.939	0.914	0.918	0.924
MA-OSEA	0.843	0.775	0.831	0.816
REINFORCE	0.377	0.499	0.487	0.454
OSG	0.917	0.873	0.850	0.880

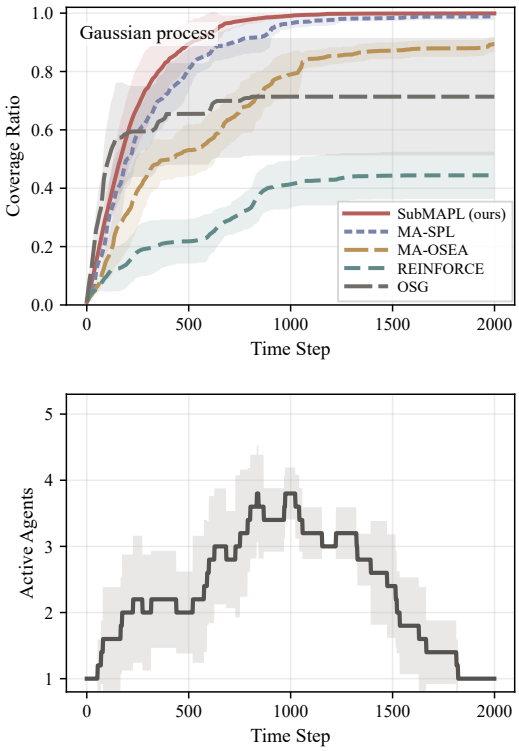

Fig. 3: Performance under random agent participation on the Gaussian-process field. Top: coverage ratios of the five methods. Bottom: the active-agent profile shared by all methods, reported as the mean with 95% confidence intervals over five shared evaluation scenarios.

evaluation scenarios. For SubMAPL and REINFORCE, results are first averaged over training seeds within each evaluation scenario. The curves report averages across the five evaluation scenarios, and the shaded regions delimit 95% confidence intervals.

B. Results

We first evaluate performance under mild changes in the active agent set $\mathcal{N}_t$, and then conduct a stress test with strongly time-varying agent participation.

1) Main Comparison: In the main comparison, all five agents are active for $t = 0, \ldots, 699$, three agents are temporarily unavailable for $t = 700, \ldots, 1299$, and all five are active again from $t = 1300$. Fig. 2 compares the five methods

TABLE III: Performance under the random agent participation on the Gaussian-process information field.

Method	Normalized area	Final coverage	95% coverage successes	Mean $T_{0.95}$
SubMAPL	**0.887**	**0.999**	**5/5**	**580**
MA-SPL	0.844	0.989	5/5	904
MA-OSEA	0.683	0.894	0/5	–
REINFORCE	0.338	0.444	0/5	–
OSG	0.665	0.714	0/5	–

The normalized area is the area under the coverage-ratio curve normalized by the evaluation horizon. $T_{0.95}$ denotes the first time step at which the coverage ratio reaches 0.95 and is averaged only over successful scenarios. The mean normalized-area advantage of SubMAPL over MA-SPL is 0.043, with a 95% confidence interval of $[0.009, 0.078]$.

under the same open-system profile. SubMAPL achieves the fastest overall information acquisition in all three landscapes. Table II reports the normalized areas under the coverage curves for all five methods. SubMAPL achieves the largest area in every information-density landscape, with an average normalized area of 0.963, compared with 0.924 for the strongest competing method, MA-SPL. This advantage indicates that SubMAPL acquires information more rapidly throughout the evaluation horizon.

The online optimization baselines exhibit different transient behavior. OSG initially acquires information rapidly by selecting actions from current marginal gains, but its sequential myopic decisions eventually plateau below complete coverage. MA-OSEA improves after the active population is restored at $t = 1300$, especially in the nonuniform fields, but remains slower than MA-SPL and SubMAPL. REINFORCE is substantially slower and shows the clearest loss of progress during the interval with only two active agents. The results support the benefit of combining this categorical relaxation with full local-action marginal feedback and KL-mirror policy updates in the tested open-system coverage task.

2) Open-system Robustness: We next evaluate the transferability of the closed-system trained policies under stronger stochastic variation in agent participation. The experiment uses the Gaussian process field. One agent remains active throughout the evaluation horizon, whereas each of the other four agents is active over a randomly generated contiguous time interval. All methods are evaluated under the same active-agent schedule within each evaluation scenario.

Fig. 3 shows that SubMAPL achieves the fastest cumulative information acquisition despite substantial changes in both the size and composition of the active-agent set. Table III shows that SubMAPL achieves the largest normalized area and the highest final coverage. Both SubMAPL and MA-SPL reach 0.95 coverage in all five scenarios, but SubMAPL reaches this level considerably earlier, with a mean hitting time of 580 steps compared with 904 steps for MA-SPL.

SubMAPL reuses observation-conditioned local policy tables learned before deployment and restores the corresponding policy whenever an agent becomes active. Therefore, it avoids restarting the decision process after each change in

participation. MA-SPL continues to adapt online, but its slower transient response reduces the information accumulated early in the finite horizon. These results indicate that the learned SubMAPL policies remain effective under substantial random changes in the size and composition of the active-agent set, without deployment-time policy updates.

VII. Conclusion

This paper studied decentralized policy learning for open multi-agent task allocation with monotone submodular team utilities. We introduced the partition multilinear extension (PME) as a policy-based continuous representation of the expected stage utility induced by factorized categorical policies and developed SubMAPL, a KL-mirror policy-learning method using the stochastic local gradient as feedback over each agent's feasible action set. We showed that this feedback provides unbiased PME coordinate-gradient information and that the KL-mirror update is exactly equivalent to an additive logit update under tabular softmax parameterization. We established open-system approximation-regret guarantees and finite-horizon performance bounds. Numerical experiments on information coverage showed that SubMAPL policies remain effective under random changes in agent participation and outperform online submodular-coordination and policy-gradient baselines. Future work will consider richer communication constraints, neural policy approximation, and long-horizon value-aware objectives.

References

[1] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 80. PMLR, 2018, pp. 5872–5881.

[2] Z. Xu, H. Zhou, and V. Tzoumas, "Online submodular coordination with bounded tracking regret: Theory, algorithm, and applications to multi-robot coordination," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2261–2268, 2023.

[3] D. Deplano, N. Bastianello, M. Franceschelli, and K. H. Johansson, "Optimization and learning in open multi-agent systems," *IEEE Transactions on Automatic Control*, vol. 71, no. 6, pp. 3864–3879, 2026.

[4] S. Fujishige, *Submodular functions and optimization*. Elsevier, 2005, vol. 58.

[5] S. Welikala, C. G. Cassandras, H. Lin, and P. J. Antsaklis, "A new performance bound for submodular maximization problems and its application to multi-agent optimal coverage problems," *Automatica*, vol. 144, p. 110493, 2022.

[6] Z. Xu, S. S. Garimella, and V. Tzoumas, "Communication-and computation-efficient distributed submodular optimization in robot mesh networks," *IEEE Transactions on Robotics*, vol. 41, pp. 3480–3499, 2025.

[7] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Mathematical programming*, vol. 14, pp. 265–294, 1978.

[8] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey, "An analysis of approximations for maximizing submodular set functions—II," in *Polyhedral Combinatorics: Dedicated to the memory of DR Fulkerson*. Springer, 2009, pp. 73–87.

[9] J. Liu, F. Li, X. Jin, and Y. Tang, "Distributed task allocation for multi-agent systems: A submodular optimization approach," *IEEE Transactions on Automatic Control*, 2026.

[10] G. Calinescu, C. Chekuri, M. Pal, and J. Vondrák, "Maximizing a monotone submodular function subject to a matroid constraint," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1740–1766, 2011.

[11] Q. Zhang, Z. Wan, Y. Yang, L. Shen, and D. Tao, "Near-optimal online learning for multi-agent submodular coordination: Tight approximation and communication efficiency," in *International Conference on Learning Representations*, 2025.

[12] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[13] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 611–24 624, 2022.

[14] M. Prajapat, M. Mutnỳ, M. N. Zeilinger, and A. Krause, "Submodular reinforcement learning," in *International Conference on Learning Representations*, 2024.

[15] R. De Santi, M. Prajapat, and A. Krause, "Global reinforcement learning: beyond linear and convex rewards via submodular semi-gradient methods," in *Proceedings of the 41st International Conference on Machine Learning*, vol. 235. PMLR, 2024, pp. 10 235–10 266.

[16] W. Chen, C. Qian, S. Xing, Y. Zhou, and V. Crawford, "Multi-agent reinforcement learning with submodular reward," *arXiv preprint arXiv:2603.06810*, 2026.

[17] Q. Zhang, Y. Sun, C. Jin, X. Zhang, Y. Shu, P. Zhao, L. Shen, and D. Tao, "Effective policy learning for multi-agent online coordination beyond submodular objectives," *Advances in Neural Information Processing Systems*, vol. 38, pp. 164 278–164 339, 2025.

[18] J. Liu, Y. Yang, L. Ballotta, F. Li, Y. Tang, and R. Carli, "Submodular multi-agent policy learning for online distributed task allocation in open multi-agent systems," *arXiv preprint arXiv:2605.13269*, 2026.

[19] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018, pp. 2974–2982.

[20] J. Castellini, S. Devlin, F. A. Oliehoek, and R. Savani, "Difference rewards policy gradients," *Neural Computing and Applications*, vol. 37, no. 19, pp. 13 163–13 186, 2025.

[21] G. Anil, P. Doshi, D. A. Redder, A. Eck, and L.-K. Soh, "Mohito: Multi-agent reinforcement learning using hypergraphs for task-open systems," in *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

[22] L. Su, X. Li, Y. Hua, X. Dong, and J. Lü, "Optimal tracking of time-varying formation in open nonlinear multiagent systems," *IEEE Transactions on Automatic Control*, 2026.

[23] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[24] F. A. Oliehoek, C. Amato *et al.*, *A concise introduction to decentralized POMDPs*. Springer, 2016.

[25] A. Krause and D. Golovin, "Submodular function maximization," in *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, 2014.

[26] J. Vondrák, "Optimal approximation for the submodular welfare problem in the value oracle model," in *Proceedings of the fortieth annual ACM symposium on Theory of computing*, 2008, pp. 67–74.

[27] C. Chekuri, J. Vondrák, and R. Zenklusen, "Submodular function maximization via the multilinear relaxation and contention resolution schemes," *SIAM Journal on Computing*, vol. 43, no. 6, pp. 1831–1879, 2014.

[28] Q. Zhang, W. Huang, C. Jin, P. Zhao, Y. Shu, L. Shen, and D. Tao, "Multinoulli extension: A lossless yet effective probabilistic framework for subset selection over partition constraints," in *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025, pp. 75 030–75 066.

[29] A. A. Bian, B. Mirzasoleiman, J. Buhmann, and A. Krause, "Guaranteed non-convex optimization: Submodular maximization over continuous domains," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, vol. 54. PMLR, 2017, pp. 111–120.

[30] H. Hassani, M. Soltanolkotabi, and A. Karbasi, "Gradient methods for submodular maximization," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[31] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[32] E. Hazan, "Introduction to online convex optimization," *Foundations and Trends in Optimization*, vol. 2, no. 3-4, pp. 157–325, 2016.

[33] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Transactions on Automatic Control*, vol. 63, no. 3, pp. 714–725, 2017.

[34] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in Neural Information Processing Systems*, vol. 12, 1999.

[35] J. Peters and S. Schaal, "Natural actor-critic," *Neurocomputing*, vol. 71, no. 7-9, pp. 1180–1190, 2008.

[36] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "On the theory of policy gradient methods: Optimality, approximation, and distribution shift," *Journal of Machine Learning Research*, vol. 22, no. 98, pp. 1–76, 2021.
[37] M. Zhang, L. Chen, H. Hassani, and A. Karbasi, "Online continuous submodular maximization: From full-information to bandit feedback," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
[38] E. Hall and R. Willett, "Dynamical models and tracking regret in online convex programming," in *International Conference on Machine Learning*. PMLR, 2013, pp. 579–587.
[39] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies," in *Journal of Machine Learning Research*, vol. 9, no. 8, 2008, pp. 235–284.
[40] D. Golovin and A. Krause, "Adaptive submodularity: Theory and applications in active learning and stochastic optimization," *Journal of Artificial Intelligence Research*, vol. 42, pp. 427–486, 2011.
[41] X. Sun, C. G. Cassandras, and X. Meng, "Exploiting submodularity to quantify near-optimality in multi-agent coverage problems," *Automatica*, vol. 100, pp. 349–359, 2019.
[42] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3, pp. 229–256, 1992.