OG-MARL: Optimistic Gossiping Multi-Agent Reinforcement Learning

Manuel Wendl
ETH Zurich
Zurich, Switzerland
mwendl@ethz.ch

***Abstract*—Optimism, which encourages exploration by favoring uncertain but potentially rewarding actions, has demonstrated improvements in sample efficiency for single-agent reinforcement learning (RL). We propose a fully decentralized framework for sample-efficient model-based multi-agent reinforcement learning (MARL). Our Optimistic Gossiping Multi-Agent Reinforcement Learning algorithm, OG-MARL, can efficiently balance learning the environment by exploring and achieving high cooperative performances in exploitation. OG-MARL learns a Bayesian uncertainty-aware world model for each agent, approximating the unknown environment dynamics and quantifying epistemic uncertainty. Nearby agents exchange model updates and converge over time to a consensus of the global model. During the learning process, the agents construct upper confidence bounds on their learned world models in a decentralized manner and act optimistically, thereby achieving a bound on their simple regret. We evaluate the efficacy of our OG-MARL implementation[1] on a cooperative reward collection task with action penalties.**

***Index Terms*—Distributed Systems, Multi-Agent RL, Optimism**

I. Introduction

For tasks and applications with multiple interacting agents, multi-agent reinforcement learning (MARL) has shown promising results [3, 13, 14, 21]. Most of this notable progress relies on the use of good simulators because, in many cases, RL methods require large amounts of data for learning. For online learning tasks in real-life applications, the RL algorithms have to learn efficiently, balancing exploration of the environment and exploitation of the learned reward and transition structure, to achieve good performance in as few episodes as possible. For single-agent reinforcement learning, various model-based techniques using optimism lead to theoretical and empirical faster learning behavior, reaching near-optimal policies [4, 5, 18, 20]. These algorithms learn an uncertainty-aware, well-calibrated world model to exploit certain parts of the environment and efficiently explore uncertain regions. In contrast, there has been little progress for MARL, and the existing related work [17] uses a centralized world model, which is a rather limiting choice for a distributed MARL setting. Other related work on optimistic MARL that uses upper confidence bounds in a distributed setting [11, 12, 15, 22, 23] considers discrete state and action spaces following the multi-armed bandit problem formulation. Since most real-world applications have continuous state representations, we focus on general Markov games in continuous state action spaces. As the main contribution of this work, we:

- Introduce OG-MARL, a decentralized model-based multi-agent reinforcement learning algorithm using optimism for optimal exploration and exploitation trade-off.
- Prove convergence of the distributed world model via gossiping on a communication graph.
- Derive a simple regret bound for cooperative tasks for provable convergence.
- Demonstrate the capabilities of OG-MARL on a cooperative reward collection task.

[1]GitHub: https://github.com/ManuelWendl/OG-MARL

II. Problem Statement

A. Multi-Agent Markov Game (MG)

Let us consider a multi-agent reinforcement learning problem formulated as an MG with N individual agents acting for an infinite horizon. At each timestep t agent i selects an action $a_t^i \in \mathcal{A}^i \subseteq \mathbb{R}^{d_{\mathcal{A}^i}}$ observing the environment state $s_t \in \mathcal{S} \subseteq \mathbb{R}^{d_{\mathcal{S}}}$. Each agent i plays the actions a_t^i according to the policy $\pi^i : \mathcal{S} \to \mathcal{A}^i$ and obtains its individual reward according to $r^i : \mathcal{S} \times \prod_{i=1}^N \mathcal{A}^i \to \mathbb{R}_+$. The environment evolves according to the unknown dynamics f and additive i.i.d. zero-mean Gaussian noise ω_t with variance σ:

$$s_{t+1} = f(s_t, a_t^1, \dots, a_t^N) + \omega_t. \tag{1}$$

In this work, we have the following regularity assumptions on the reward and the true system dynamics:

Assumption 1 (Lipschitz Dynamics and Reward)**.** *There exist constants $L_f, L_r > 0$ such that for all $s, s' \in \mathcal{S}$, and all $a, a' \in \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$,*

$$\|f(s,a) - f(s',a')\|_2 \leq L_f(\|s - s'\|_2 + \|a - a'\|_2),$$
$$\|r^i(s,a) - r^i(s',a')\|_2 \leq L_r(\|s - s'\|_2 + \|a - a'\|_2).$$

Additionally, we have the following requirements on the state and action spaces as well as the reward:

Assumption 2 (Bounded State and Action Spaces)**.** *The state space $\mathcal{S} \subseteq \mathbb{R}^{d_{\mathcal{S}}}$ and each action space $\mathcal{A}^i \subseteq \mathbb{R}^{d_{\mathcal{A}^i}}$ are compact and bounded. All rewards $r^i(s, a^1, \dots, a^N)$ are uniformly bounded in $[R_{min}, R_{max}]$, where $R_{min}, R_{max} \in \mathbb{R}$.*

B. Cooperative Objective

We define the cooperative reward at time t by

$$r(s_t,\, a_t^1, \ldots, a_t^N) = \sum_{i=1}^{N} r^i\big(s_t, a_t^1, \ldots, a_t^N\big). \tag{2}$$

The cooperative return under the joint policy π is the accumulated and discounted reward with discount factor $\gamma \in (0,1)$

$$J(\pi) = \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t^1, \ldots, a_t^N)\Big]. \tag{3}$$

Since interacting with the real environment is costly, our goal is to minimize the simple cooperative regret, defined over H learning episodes as

$$R(H) = \sum_{h=1}^{H}\Big(\max_{\pi\in\Pi} J(\pi) - J\big(\pi_h\big)\Big), \tag{4}$$

where $\Pi = \prod_{i=1}^{N} \Pi^i$ is the joint policy space and $\pi_h \in \Pi$ is the joint policy chosen in episode h. Sublinear growth of $R(H)$ implies that the agents collectively approach the optimal cooperative policy.

C. Model-Based Planning with Consensus

In our distributed model-based learning scheme, each agent i maintains a local world-model in the form of a sparse inducing points-based Gaussian Process (GP) [16] via zero-padding, and runs K rounds of gossiping on a communication graph to agree on a common model of f with all other agents. Based on this converged consensus model, the agents update their policies π_h^i in each episode h using optimistic model-based policy optimization (MBPO)[9] to maximize the cooperative return $J(\pi)$. While the theory in this work builds on GPs, it can also be extended to other classes of calibrated models. For the well-calibrated GPs, we assume boundedness of f in the reproducing Kernel Hilbert space.

Assumption 3 (Reproducing Kernel Hilbert Space (RKHS))**.** *The unknown dynamics f lie in an RKHS of a kernel k and therefore have a bounded norm $||f||_k \leq B$ with a known and finite constant B.*

Definition 1 (Calibrated Model)**.** *The uncertainty aware model $\mathcal{F}_{h'} := \{\tilde{f} \mid |\tilde{f} - \mu_{h'}| \leq \beta_{h'}\sigma_{h'}\}$ with nominal model $\mu_{h'} : \mathcal{S} \times \prod_{i=1}^{N} \mathcal{A}^i \to \mathbb{R}^{d_\mathcal{S}}$ and uncertainty $\sigma_{h'} : \mathcal{S} \times \prod_{i=1}^{N} \mathcal{A}^i \to \mathbb{R}^{d_\mathcal{S}}$ is well-calibrated over all episodes $h' \in \{1, \ldots, h\}$ if there exist $\beta_{h'}$, so that with probability of at least $1-\delta$ the confidence intervals enclose the true dynamics $\forall (s, a^1, \ldots, a^N) \in \mathcal{S} \times \prod_{i=1}^{N} \mathcal{A}^i$ that $f(s, a^1, \ldots, a^N) \in \mathcal{F}_h := \bigcap_{h'=1}^{h} \mathcal{F}_h$.*

For the distributed GP model, agents communicate with each other and exchange knowledge about their models. The agents are connected via a communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertices $\mathcal{V}$ being the agents and their connecting edges $\mathcal{E}$. The communication network is fully defined via the doubly stochastic matrix $W \in \mathbb{R}^{N\times N}$ and has $W_{i,j} \in (0,1]$ if agents i and j are connected and have a common edge $(i,j) \in \mathcal{E}$.

Definition 2 (Doubly Stochastic Matrix)**.** *A matrix $W \in \mathbb{R}^{N\times N}$ is doubly stochastic if*

$$\mathbf{1}^\top W = \mathbf{1}^\top \wedge W\mathbf{1} = \mathbf{1}.$$

For learning the world-model fully decentralized, we have the following additivity assumption on the global dynamics given the individual states and actions of all agents:

Assumption 4 (Additive Dynamics)**.** *The global dynamics decompose into the individual contributions as*

$$f(s, a^1, \ldots, a^N) = \sum_{i=1}^{N} f^i(s, a^i)$$

given the agents $i = 1, \ldots, N$ with the individual actions a^i.

III. OG-MARL: Optimistic Gossiping MARL

We now detail the OG-MARL method. First, we describe how each agent builds a sparse GP model of the dynamics section III-A, then how agents perform consensus section III-B, and finally the overall algorithm flow section III-C.

A. Sparse GP with Shared Inducing Points

1) Global Feature Mapping: To learn a distributed world-model, each agent needs to have the same representation of the global state. Therefore, we first embed the state and joint-action into a shared, D-dimensional feature space of the state $s \in \mathcal{S}$, observed by all agents and the individual actions $a^i \in \mathcal{A}^i$, that are only visible to agent i:

$$\varphi :\, \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N \;\longrightarrow\; \mathbb{R}^D\,. \tag{5}$$

To allow each agent to update its world-model only on its action, agent i forms input x_i by placing the global state s and its own action a_i in the appropriate slots, with zeros for other agents' actions; this ensures each agent's input lies in the full joint-action space

$$x^i = \big(s,\; 0, \ldots, a^i, \ldots, 0\big) \in \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N, \tag{6}$$

and defines its own local and zero-padded feature

$$\varphi^i(s, a^i) \;=\; \varphi(x^i) \;\in\; \mathbb{R}^D. \tag{7}$$

All the agents have to agree on the same mapping φ and the same embedding dimension D, to learn a distributed model.

Further, we require for sufficient expressivity of a learned distributed world model that:

Assumption 5. *Each agent's zero-padded mapping $\varphi^i(s, a^i)$ is sufficiently expressive to approximate its dynamics component $f^i(s, a^i)$ of the decomposed global dynamics f.*

Typical choices for such mappings are random Fourier features or radial basis function (RBF) embeddings.

2) Shared Inducing Points: The prior of the world model is constructed from a set of M common joint-action inducing points $Z = \{\, z_j = (s^{(j)}, a^{(j)})\}_{j=1}^{M} \subset \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$, that are shared by all agents and determined once, before the optimization starts. Further, a suitable kernel is chosen to fulfill assumption 3: $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ and we precompute once:

$$K_{Z,Z} \in \mathbb{R}^{M\times M}, \quad [K_{Z,Z}]_{jk} = k\big(\varphi(z_j),\, \varphi(z_k)\big). \tag{8}$$

3) Local Sparse-GP Posterior: In every episode h, each agent i observes the transitions $\{(s_t, a_t^i, s_{t+1})\}_{t=1}^{T_i}$, with the global state s_t, the next state s_{t+1} and its own action a_t^i. Let us define the state difference as $\Delta s_t = s_{t+1} - s_t$ and determine the local, zero-padded features $\varphi_t^i = \varphi^i(s_t, a_t^i)$. Then the local GP posterior over the inducing outputs $f(Z)$ of each agent i is parametrized by the mean and variance [16, eqs. 8.23,8.24]

$$m_h^i \in \mathbb{R}^{M\,d_S}, \quad S_h^i \in \mathbb{R}^{M\,d_S \times M\,d_S}, \tag{9}$$

$$m_h^i = \tfrac{1}{\sigma^2}\, S_h^i\, K_{Z,X^i}\, Y^i, \tag{10}$$

$$S_h^i = \Big(K_{Z,Z} + \tfrac{1}{\sigma^2} K_{Z,X^i} K_{X^i,Z}\Big)^{-1}, \tag{11}$$

with the cross-covariance matrix K_{Z,X^i} between the inducing points Z and the agent data X^i, as well as the target Y^i of the state differences Δs_t computed for all $t = 1, \dots, T_i$:

$$K_{Z,X^i} = \big[k(\varphi(z_j), \varphi^i(s_t, a_t^i))\big]_{j=1,\dots,M}^{t=1,\dots,T_i}, \tag{12}$$

$$Y^i = [\Delta s_1, \dots, \Delta s_{T_i}]^\top. \tag{13}$$

The local one-step predictive mean and variance of agent i for state s and action a^i are hence computed by:

$$\mu_h^i(s, a^i) = K_{x^i,Z}\,(K_{Z,Z} + \sigma^2 I)^{-1}\, m_h^i, \tag{14}$$

$$\sigma_h^i(s, a^i)^2 = k\big(\varphi^i(s, a^i), \varphi^i(s, a^i)\big) \tag{15}$$

$$- K_{x^i,Z}\,(K_{Z,Z} + \sigma^2 I)^{-1}\, K_{Z,x^i}, \tag{16}$$

where $K_{Z,x^i} = K_{x^i,Z}^\top$. However, this predictive distribution with mean $\mu_h^i(s, a_i)$ and variance $\sigma_h^i(s, a_i)^2$ is not incorporating any knowledge of the other agents and does not correspond to the predictive distribution one would obtain from a global world model that has access to the actions a_i of all the agents.

B. Sparse-GP Consensus

Since the local predictive distribution of each agent does not know anything about the actions of neighboring agents, the agents use multiple communication rounds to obtain an arbitrarily close approximation of the global world-model prediction. The agents communicate over a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the agents being represented as vertices $\mathcal{V}$ and their communication connections as edges $\mathcal{E}$. The communication matrix of all edges has doubly-stochastic weights W. At the end of each episode h, each agent computes its local sparse-GP posterior of the inducing inputs with eqs. (10) and (11) and initializes the communication round with

$$\mu_h^{i,(0)} = \mu_h^i, \quad \Sigma_h^{i,(0)} = \Sigma_h^i. \tag{17}$$

Then the agents communicate and update their initial believes of their local sparse-GP posteriors based on the other agents for $k = 0, \dots, K-1$ gossiping steps. We therefore define the stacked block vectors $\eta_h^{(k)} \in \mathbb{R}^{NMd_S}, \Lambda_h^{(k)} \in \mathbb{R}^{NMd_S \times Md_S}$

$$\eta_h^{(k)} = \begin{bmatrix} \mu_h^{1,(k)} \\ \vdots \\ \mu_h^{N,(k)} \end{bmatrix}, \quad \Lambda_h^{(k)} = \begin{bmatrix} \Sigma_h^{1,(k)} \\ \vdots \\ \Sigma_h^{N,(k)} \end{bmatrix}, \tag{18}$$

and perform the following gossiping step iteratively

$$\eta_h^{(k+1)} = (W \otimes I)\eta_h^{(k)}, \tag{19}$$

$$\Lambda_h^{(k+1)} = (W \otimes I)\Lambda_h^{(k)}, \tag{20}$$

in matrix notation using the Kronecker product $\otimes$. After K iterations all agent agree on $\bar{\mu}_h = \mu_h^{i,(K)}$ and $\bar{\Sigma}_h = \Sigma_h^{i,(K)}$.

C. OG-MARL Algorithm

We summarize OG-MARL in Algorithm 1. In each episode h, the agents perform four main steps of (a) data collection in (line 4): each agent executes its current policy and collects local transitions. Next we perform the (b) local sparse-GP update (line 5): agents fit their GP world models to the newly collected data in eqs. (10) and (11). Further in (c) the communication is performed reaching Consensus (line 6): agents run K rounds of gossip averaging on their GP parameters eqs. (19) and (20). After achieving consensus we (d) update the current policy optimistically (line 7): each agent uses the converged consensus model given by the natural parameters

$$\mu_h = \mu_0 + N\bar{\mu}_h, \quad \Sigma_h = \Sigma_0 + N\bar{\Sigma}_h, \tag{21}$$

with the priors (μ_0, Σ_0) and the consensus parameters $(\bar{\mu}_h, \bar{\Sigma}_h)$ after gossiping. This converged consensus model is used to compute the predictive distribution $(\mu_h^{cons}, \sigma_h^{cons})$ with eqs. (14) and (15) and the moment-form parameters

$$m_h = S_h \mu_h, \quad S_h = \Sigma_h^{-1}. \tag{22}$$

The policy is then optimistically updates using the model prediction and model-based policy optimization.

Algorithm 1 OG-MARL: Optimistic Gossiping MARL

1: Initialize the world-models with joint inducing points Z.
2: Choose global mapping φ, kernel, k and compute prior Gramm-matrix $K_{Z,Z}$. Initialize policies π_0^i.
3: **for** episode $h = 1, \dots, H$ **do**
4: **(a)** agents execute π_h^i, and collect $\{(s_t, a_t^i, s_{t+1})\}_{t=1}^{T_i}$.
5: **(b)** compute μ_h^i, Σ_h^i from data $\{(\varphi^i(s_t, a_t^i), \Delta s_t)\}$.
6: **(c)** run K gossip rounds on $\{\mu_h^i, \Sigma_h^i\}$ over $\mathcal{G}$.
7: **(d)** compute $\mu_h^{\text{cons}}, \sigma_h^{\text{cons}}$ from μ_h, Σ_h and update π_h^i.
8: **end for**

IV. Convergence of Distributed World-Model

According to our problem definition, we have the following assumption for the communication graph of the agents

Assumption 6. *Let the communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be connected, aperiodic, and let the weight matrix W of the communication graph be doubly-stochastic.*

Lemma 1 (Consensus)**.** *Given assumption 6 is satisfied, the gossiping round with a sufficiently large number of K updates*

$$\mu^{(k+1)} = W\,\mu^{(k)}, \qquad \mu^{(0)} = \big[\mu_h^1; \dots; \mu_h^N\big]$$

converges to the consensus vector

$$\bar{\mu}_h = \mu_h^{(K)} = \tfrac{1}{N} \sum_{i=1}^{N} \mu_h^i. \tag{23}$$

Proof. Given the block-vector notation of eq. (18) for the gossiping step in eqs. (19) and (20) and that assumption 6 holds, [2, theorem 2.13], and perform communication for K rounds until convergence, we obtain

$$\eta^{(K)} = (W \otimes I_{Md_s})^K \eta^{(0)} = (W^K \otimes I_{Md_s})\eta^{(0)}, \tag{24}$$

where the K-th power of W converges to [2, theorem 2.13]

$$W^K = \mathbf{1}w^\top, \tag{25}$$

where the left dominant eigenvector is $w = \frac{1}{N}\mathbf{1}$ due to W being doubly-stochastic [2, Corr. 13.2].

$$\eta^{(K)} = \left(\tfrac{1}{N}\,\mathbf{1}\mathbf{1}^\top \otimes I_{Md_S}\right)\eta^{(0)} = \begin{bmatrix}\bar{\mu}_h \\ \vdots \\ \bar{\mu}_h\end{bmatrix}. \tag{26}$$

Then $\eta^{(K)}$ is a fixed point of the communication update:

$$(W \otimes I)\eta^{(K)} = \eta^{(K)}. \tag{27}$$

Let us define the error term $E^{(k)} = \eta^{(k)} - \eta^{(K)}$. And since $\mathbf{1}^\top W = \mathbf{1}^\top$ holds due to doubly-stochastic W,

$$\begin{aligned} E^{(k+1)} &= \eta^{(k+1)} - \eta^{(K)} && (28)\\ &= (W \otimes I)\eta^{(k)} - (W \otimes I)\eta^{(K)} && (29)\\ &\overset{(27)}{=} (W \otimes I)\eta^{(k)} - \eta^{(K)} && (30)\\ &= (W \otimes I)\, E^{(k)}. && (31)\end{aligned}$$

The operator norm of the Kronecker product satisfies [2, E.8.1] $\|W \otimes I\|_2 = \|W\|_2\|I\|_2 = \|W\|_2$ and from double-stochasticity, we obtain [2, Lemma 2.9] $\|W\|_2 = \rho(W) = 1$. On the subspace $\{x : \mathbf{1}^\top x = 0\}$, we have $|\lambda_2(W)| \le 1$, thus

$$\|E^{(k+1)}\|_2 \;\le\; \lambda_2(W)\,\|E^{(k)}\|_2, \tag{32}$$

so by Banach's Fixed-Point Theorem in [1] the iterations converge geometrically to the fixed point $M^{(K)}$:

$$\|\eta^{(k)} - \eta^{(K)}\| \le \lambda_2(W)^k\,\|\eta^{(0)} - \eta^{(K)}\|. \tag{33}$$

Since each block of $\eta^{(K)}$ equals μ_h^K, every agent's local $\mu_h^{i,(K)}$ converges to the average $\mu_h^K = \bar{\mu}_h$. □

The same argument as in lemma 1 applies to the block-vector of covariances $\left[\Sigma_h^{1,(k)}; \ldots; \Sigma_h^{N,(k)}\right]$. Hence after K rounds, we get for the post-communication posterior estimates that

To obtain a meaningful predictive distribution parametrized by mean and variance μ_h^{cons} and σ_h^{cons}, computed with eqs. (14) and (15) from the converged consensus model defined in eq. (21) we show that theorem 1 holds:

Theorem 1 (Global World-Model Consensus)**.** *Given assumptions 4 to 6 are satisfied and a sufficiently large number K of communication rounds have taken place, each agent's sparse GP posterior estimates (m_h, S_h) have converged to the posterior of the global world model with access to the entire pooled data $\bigcup_{i=1}^N x_h^i$ of all agents $i = 1, \ldots, N$.*

Proof. We prove theorem 1 by (i) writing both local and global GP updates in their *natural-parameter* form, and (ii) exploiting that consensus averages the distributed natural parameters.

a) Global Computation: Recall our sparse GP prior on the inducing outputs $u = f(Z) \in \mathbb{R}^{Md_S}$:

$$p(u) \;=\; \mathcal{N}\big(u\,;\,0,\,K_{Z,Z}\big) \Longleftrightarrow (\Sigma_0,\,\mu_0) \tag{34}$$

with the natural parameters mean and covariance

$$\mu_0 = 0, \qquad \Sigma_0 = K_{Z,Z}^{-1}. \tag{35}$$

Agent i observes local data X_h^i of size T_i, with features $\Phi_h^i = [\varphi_t^i]_{t=1}^{T_i} \in \mathbb{R}^{T_i \times D}$ and targets $Y_i \in \mathbb{R}^{T_i \times d_S}$. Under aleatoric Gaussian noise $\mathcal{N}(0, \sigma^2 I)$, we obtain the log-likelihood

$$\log p(\{\Delta s_t\}|u) = \prod_{i=1}^N \log p(\{\Delta s_t\}|u; \Phi_h^i), \tag{36}$$

with local natural parameters

$$\Sigma_h^i \;=\; \tfrac{1}{\sigma^2}\,\Phi_h^{i\top}\Phi_h^i\,, \qquad \mu^i \;=\; \tfrac{1}{\sigma^2}\,\Phi_h^{i\top} Y_i. \tag{37}$$

Hence, the *global* posterior after pooling all data has consequently under assumption 4 the natural parameters :

$$\Sigma_h = \Sigma_0 + \sum_{i=1}^N \Sigma_h^i = K_{Z,Z}^{-1} + \frac{1}{\sigma^2}\sum_{i=1}^N K_{Z,X^i}K_{X^i,Z}, \tag{38}$$

$$\mu_h = \mu_0 \;+\; \sum_{i=1}^N \mu_h^i = \frac{1}{\sigma^2}\sum_{i=1}^N K_{Z,X^i}Y^i. \tag{39}$$

b) Local Computation and Consensus: Each agent i can compute its *local* parameters (Σ_h^i, μ_h^i) from its own data alone. We stack these into a block-vector given by

$$\Lambda_h^{i(0)} = \begin{bmatrix}\Sigma_h^i \\ \vdots \\ \Sigma_h^i\end{bmatrix}, \quad \eta_h^{i(0)} = \begin{bmatrix}\mu_h^i \\ \vdots \\ \mu_h^i\end{bmatrix} \tag{40}$$

We then run K-step gossip (with W doubly-stochastic) independently on the Σ-blocks and μ-blocks. By lemma 1, the gossiping updates

$$\Lambda_h^{i(k+1)} = (W \otimes I)\Lambda_h^{i(k)}, \quad \eta_h^{i(k+1)} = (W \otimes I)\eta_h^{i(k)} \tag{41}$$

converge to the block averages

$$\bar{\Sigma}_h = \frac{1}{N}\sum_{i=1}^N \Sigma_h^i, \quad \bar{\mu}_h = \frac{1}{N}\sum_{i=1}^N \mu_h^i. \tag{42}$$

c) Rescaling: Since $\bar{\Sigma}_h$ and $\bar{\mu}_h$ are just averages, every agent recovers the true global sum by multiplying by N:

$$\Sigma_h = \Sigma_0 + \sum_{i=1}^N \Sigma_h^i = \Sigma_0 + N\bar{\Sigma}_h \tag{43}$$

$$\mu_h = \mu_0 + \sum_{i=1}^N \mu_h^i = \mu_0 + N\bar{\mu}_h. \tag{44}$$

Agent i already knows the prior Σ_0 and μ_0, so it can form $(\Sigma_h,\, h_h)$ exactly after gossiping converged to consensus.

d) Recovery of (m_h, S_h)*:* Finally, all agents compute

$$m_h = S_h\,\mu_h, \quad S_h = \Sigma_h^{-1}. \tag{45}$$

By construction, these match the posterior mean and covariance of the sparse GP on the pooled data. Hence the predictive $\mu(s,a)$ and $\sigma^2(s,a)$ built from (m_h, S_h) are the same as in the centralized case if the communication rounds have converged.
e) Uniqueness and Contraction: The posterior parameter map $(\Sigma_h, \mu_h) \mapsto (\Sigma_0 + \sum \Sigma_h^i,\ \mu_0 + \sum \mu_h^i)$ is affine and has a unique fixed point. The consensus averaging operator is a Banach contraction on the error subspace with a spectral radius < 1, so any initial $X^{(0)}$ converges to $\bar{X}$. Rescaling the consensus parameters by N and adding the prior parameters (Σ_0, h_0) preserves uniqueness. Thus, for a large enough number K of communications, all agents converge to consensus, so that every agent recovers the global posterior parameters of the distributed world-model and hence the predictive distribution, despite only ever observing its own actions. □

V. Regret Analysis

In this section, we derive a regret bound for the *fully cooperative* setting, in which all agents jointly seek to maximize the team return

$$J(\pi) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t^1, \ldots, a_t^N)\Big], \tag{46}$$

$$r(s, a^1, \ldots, a^N) = \sum_{i=1}^{N} r^i(s, a^1, \ldots, a^N). \tag{47}$$

with the optimistically chosen joint policy

$$\pi_h = \arg\max_{\pi \in \Pi} \sum_i UCB_h^i(\pi), \tag{48}$$

where we optimize for the upper confidence bound

$$UCB_h^i(\pi) = \max_{\eta \in [-1,1]^p} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^i(s_t, a_t^1, \ldots, a_t^N)\right] \tag{49}$$

$$s.t.\ a_t = \pi(s_t) \tag{50}$$

$$s_t = \mu_h^{cons}(s_t, a_t) + \beta_h \sigma_h^{cons}(s_t, a_t)\eta(s_t, a_t) + \omega_t. \tag{51}$$

In contrast to solving this with hallucinating control of the uncertainty with $\eta(s_t, a_t)$ as proposed in [6], we add an intrinsic exploration bonus to the reward, that is dependent on the scaled model uncertainty $\|\sigma_h^{cons}(s_t, a_t)\|$, but plan on the nominal dynamics following the mean $\mu_h^{cons}(s_t, a_t)$ of the predictive distribution:

$$\pi_h^i = \arg\max_{\pi^i \in \Pi^i} \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t (r^i(s_t, a_t^1, \ldots, a_t^N) \tag{52}$$

$$+ \lambda_h \|\sigma_h^{cons}(s_t, a_t^1, \ldots a_t^N))\|\Big] \tag{53}$$

$$s.t.\ a_t = \pi(s_t), \quad s_t = \mu_h^{cons}(s_t, a_t) + \omega_t. \tag{54}$$

A. Confidence lemma

To show that the regret is bounded, we first need to establish an upper bound on the value of each agent using the UCB.

Lemma 2 (Confidence)**.** *Given assumptions 1 to 3 hold and the model is calibrated according to definition 1, then for every joint policy π and episode h,*

$$\sum_{i=1}^{N} UCB_h^i(\pi) \geq J(\pi).$$

Proof. [17, Lemma 1]. □

Thus the aggregated UCB is an optimistic estimate of the true cooperative performance. Further, we upper-bound the gap of the upper confidence bound to the value and demonstrate that the intrinsic reward is sufficient do so:

Lemma 3 (Optimism Gap Bound)**.** *Given assumptions 1 to 3 hold and the model is calibrated according to definition 1, there exists a constant $\lambda_h = \frac{R_{max}\gamma}{1-\gamma}\frac{(1+\sqrt{d_x})\beta_h}{\sigma}$ so that for any joint policy π of the N agents,*

$$\Big|\sum_i UCB_h^i(\pi) - J(\pi)\Big| \leq N\lambda_h \mathbb{E}_\pi\Big[\sum_{t=0}^{\infty} \gamma^t \|\sigma_h^{cons}(s_t, \pi(s_t))\|\Big].$$

Proof. Let $\{s_t\}_{t=0}^{\infty}$ be the trajectory under the true dynamics f and noise ω, and let $\{\tilde{s}_t\}_{h=0}^{H-1}$ be the "hallucinated" trajectory under $\tilde{f} = \mu_h^{cons} + \beta_h\,\sigma_h^{cons}\,\eta^\star$, with the same noise ω. Then

$$V^i(\pi) = \mathbb{E}_\omega\Big[\sum_{t=0}^{\infty} \gamma^t r^i(s_t, \pi(s_t))\Big], \tag{55}$$

$$UCB_h^i(\pi) = \mathbb{E}_\omega\Big[\sum_{t=0}^{\infty} \gamma^t r^i(\tilde{s}_t, \pi(\tilde{s}_t))\Big]. \tag{56}$$

Under assumption 1, we bound

$$\big|UCB_h^i(\pi) - V^i(\pi)\big| \tag{57}$$

$$= \sum_{t=0}^{\infty} \gamma_r \mathbb{E}\Bigg[\sqrt{\max\{\mathbb{E}_{\omega_t}[R(\tilde{s}_{t+1})], \mathbb{E}_{\omega_t}[R(s_{t+1})]\}} \tag{58}$$

$$\times\, \gamma_r^t \min\left\{\frac{\|s_{t+1} - \tilde{s}_{t+1}\|}{\sigma}, 1\right\}\Bigg]. \tag{59}$$

$$\leq \sum_{t=0}^{\infty} \gamma_r \mathbb{E}\Bigg[\sqrt{\max\{\mathbb{E}_{\omega_t}[R(\tilde{s}_{t+1})], \mathbb{E}_{\omega_t}[R(s_{t+1})]\}} \tag{60}$$

$$\times\, \gamma_r^t \min\left\{\frac{\|f(\tilde{s}_t, \pi(\tilde{s}_t)) - \mu_h^{cons}(\tilde{s}_t, \pi(\tilde{s}_t))\|}{\sigma}, 1\right\}\Bigg] \tag{61}$$

$$\leq \frac{R_{max}\gamma_r}{1-\gamma_r}\frac{(1+\sqrt{d_s})\beta_h}{\sigma}\sum_{t=0}^{\infty} \mathbb{E}\left[\gamma_r^t \|\sigma_h^{cons}(\tilde{s}_t, \pi(\tilde{s}_t))\|\right] \tag{62}$$

$$= \lambda_h \mathbb{E}\left[\gamma_r^t \|\sigma_h^{cons}(\tilde{s}_t, \pi(\tilde{s}_t))\|\right] \tag{63}$$

for $\frac{(1+\sqrt{d_s})\beta_h}{\sigma} R(s) = {V_{r,f}^{\pi}}^2(s)$ and with $R(s) \leq \lambda_h$, given the simulation lemma [10, lemma 3.9] and the additive noise standard deviation σ. We obtain the resulting lemma by summing over all N agents. □

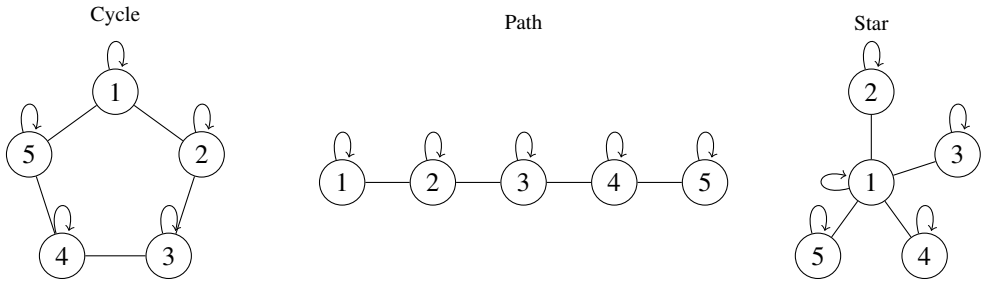

Fig. 1. Three connected, symmetric graphs on 5 nodes with nonzero self-weights (loops) and Metropolis weights: (left) cycle, (center) path, (right) star.

B. Regret Bound

The regret bound of OG-MARL depends on the maximum information gain of the chosen kernel k [19], defined as

$$\Gamma_H(k) = \max_{\mathcal{U}\subset\mathcal{S}\times\prod_i \mathcal{A}^i;|\mathcal{U}|\leq H} \frac{1}{2}\log|I+\sigma^{-2}K_H|. \tag{64}$$

Theorem 2 (Cooperative Simple Regret)**.** *Assume we determine the policy π_h in each episode planning with the intrinsic reward in eq.* (52)*, solving for $\pi_h = \arg\max_{\pi\in\Pi}\sum_i UCB_h^i(\pi)$. Then, under assumptions 1 to 3 and definition 1 it holds with high probability $1-\delta$,*

$$R(H) = \sum_{h=1}^{H}\Big(\max_{\pi} J(\pi) - J(\pi_h)\Big) \leq \mathcal{O}\left(N\Gamma^{3/2}_{H\log(H)}\sqrt{H}\right).$$

Proof. Fix any episode h. By lemma 2 and by the definition of our optimistically chosen policy in eq. (52)

$$\max_{\pi} J(\pi) \leq \max_{\pi}\sum_i UCB_h^i(\pi) \tag{65}$$

$$\overset{(52)}{=} \sum_i UCB_h^i(\pi_h). \tag{66}$$

Thus, the single-episode regret satisfies

$$\max_{\pi} J(\pi) - J(\pi_h) \leq \sum_i UCB_h^i(\pi_h) - J(\pi_h). \tag{67}$$

Applying lemma 3 and summing over $h = 1,\dots,H$ gives

$$R(H) \leq N\lambda_h \sum_{h=1}^{H}\mathbb{E}_{\pi_h}\Big[\sum_{t=0}^{\infty}\gamma^t\|\sigma_h^{cons}(s_t,\pi_h(s_t))\|\Big]. \tag{68}$$

Due to the monotonicity of λ_h we upper bound $\lambda_h \leq \lambda_H$ and use Cochy-Schwarz to obtain

$$\leq N\lambda_H\sqrt{H}\sqrt{\sum_{h=1}^{H}\mathbb{E}_{\pi_h}\Big[\sum_{t=0}^{\infty}\gamma^t\|\sigma_h^{cons}(s_t,\pi_h(s_t))\|^2\Big]^2} \tag{69}$$

$$\leq N\lambda_H\sqrt{\frac{H}{1-\gamma}}\sqrt{\sum_{h=1}^{H}\mathbb{E}_{\pi_h}\Big[\sum_{t=0}^{\infty}\gamma^t\|\sigma_h^{cons}(s_t,\pi_h(s_t))\|^2\Big]} \tag{70}$$

Using the maximum information gain of kernel k $\Gamma_H(k)$, we obtain [20]:

$$R(H) \leq N\lambda_H\sqrt{\frac{R_\gamma H\Gamma_{H\log(H)}}{1-\gamma} + \frac{R_\gamma\sigma^2_{max}H\log(H)}{1-\gamma^2}}, \tag{71}$$

where $R_\gamma = \frac{s_{max}}{\log(1+s_{max})}$, with $s_{max} = \frac{\sigma^{-2}d_x\sigma^2_{max}}{1-\gamma}$ and since $\lambda_H \propto \frac{\beta_H}{1-\gamma_r}$, we can derive the regret bound

$$R(H) \leq \mathcal{O}\left(N\Gamma^{3/2}_{H\log(H)}\sqrt{H}\right). \tag{72}$$

□

VI. Experiments

A. Distributed GP Consensus

We first evaluate the convergence behavior of our distributed sparse-GP world model isolated from the reinforcement learning pipeline to gain insight into how the underlying communication topology affects consensus of the model.

a) Experimental Setup: We consider $N = 5$ agents and fix a common set of $M = 50$ inducing inputs. Every agent i updates its local natural-parameter pair (Σ_h^i, μ_h^i) from $T_i = 500$ new $(s, a^i, \Delta s)$ samples, and then performs K rounds of gossip on the doubly-stochastic weight matrix W. We measure the *consensus error* of the first agent with the Frobenius norm

$$\epsilon_K = \max_i \left\|\Sigma_h - \Sigma_h^*\right\|_F \tag{73}$$

where Σ_h^* is the centralized posterior precision matrix computed on the union of the data of all agents.

b) Communication Graphs: We test three connected, symmetric graphs satisfying assumption 6 (see fig. 1):

- **Cycle:** each agent connects to two neighbors in a ring.
- **Path:** agents form a linear chain.
- **Star:** one central node connects to all others.

c) Consensus Error vs. Communication Rounds: Figure 3 plots ϵ_K against the number of gossip iterations K. As expected, the star graph achieves zero error in a single step (equivalent to a centralized model with a hub node). Both the cycle and path graphs require multiple rounds to converge, with the cycle consistently exhibiting faster error decay than the path, owing to its smaller diameter and larger spectral gap.

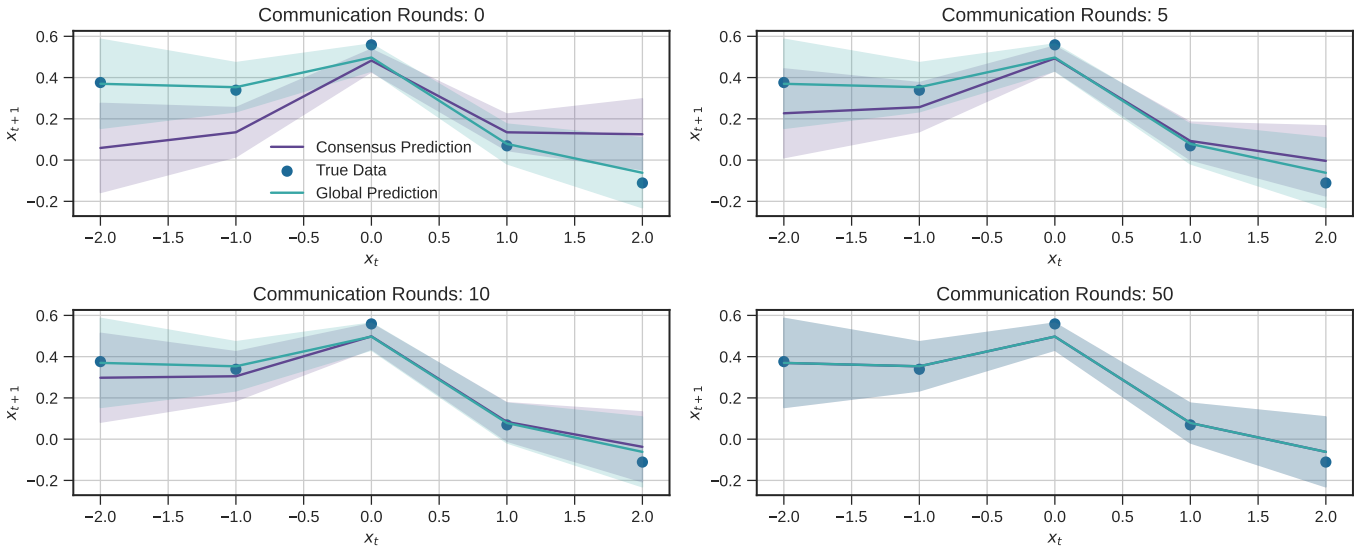

Fig. 2. Visualizations of the convergence of the predictive distribution of agent $i = 1$ for a path graph after $\{0, 5, 10, 50\}$ gossip rounds.

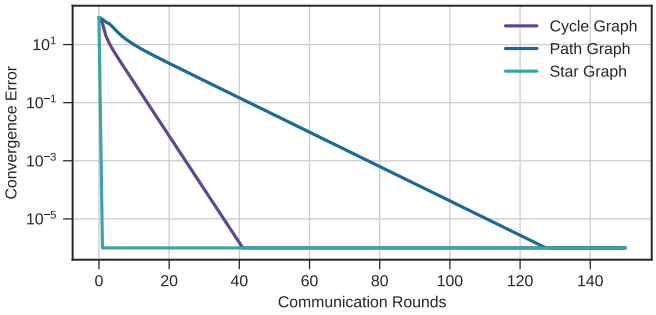

Fig. 3. Convergence Error Analysis for different graph structures.

d) Predictive Distribution Convergence: To illustrate how consensus improves the predictive world model, fig. 2 shows the posterior mean contribution for the first dimension of Δs for the path graph after $K \in \{0, 5, 10, 50\}$ rounds. At $K = 0$ (no communication), each agent's prediction is highly variable; by $K = 50$ the curves coincide almost exactly with the centralized predictor, demonstrating that parameter consensus directly translates into correct predictions of the model.

B. Optimistic Reinforcement Learning

We evaluate OG-MARL on a cooperative reward-collection task in which N agents jointly navigate to accumulate as much reward as possible. At each timestep, an agent receives a positive reward for occupying a goal location; the goals yield rewards drawn from $\{0.1, 0.5, 1\}$ and is optionally penalized for large action magnitudes. Each agent observes the global state s (the stacked positions of all agents), its action a^i, and its immediate reward. We compare:

1) **SAC:** model-free Soft Actor-Critic agents that learn independently without sharing data or model information.
2) **MB-MARL:** centralized model-based multi-agent RL using a global world model, trained on all agents' data.
3) **OG-MARL:** our optimistic gossiping multi-agent RL approach using a decentralized world model.

We also perform an ablation study on the optimism scaling factor. All methods use the hyper-parameters listed in table I and have been evaluated on 5 random seeds.

a) Sample Efficiency: Figures 4 and 5 shows the cumulative reward of all agents versus environment steps. Both model-based approaches (Global MB-MARL and OG-MARL) dramatically outperform isolated SAC in sample efficiency: by sharing data through their world models and performing multiple synthetic roll-out updates per environment step, they reach near-optimal policies with significantly fewer interactions.

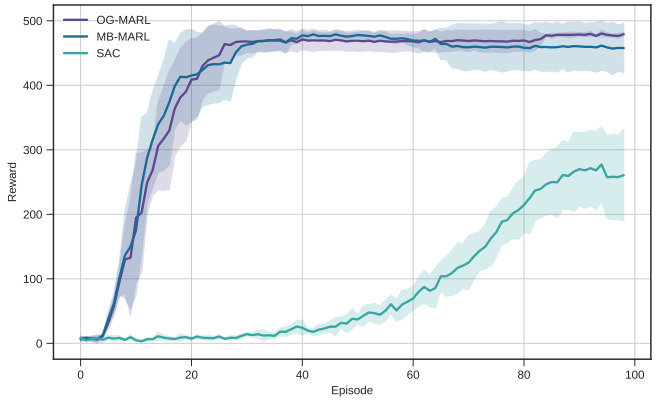

Fig. 4. Comparison of achieved cooperative reward for the goal collection task without action penalties.

b) Cooperative Behavior and Shared Knowledge: In SAC, individual agents often fail to discover the highest-reward goals, as each agent must explore the environment independently.

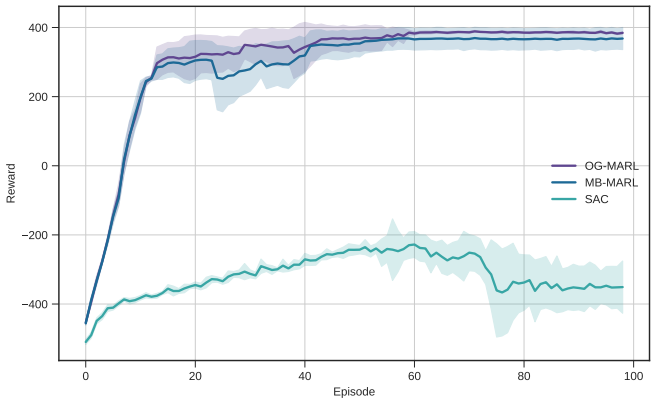

Fig. 5. Comparison of achieved cooperative reward in the goal collection task with action penalties.

This leads to a heterogeneous performance and suboptimal collective reward (see SAC video[2], and appendix C). By contrast, both model-based methods consistently guide every agent to the $r = 1$ goals and maintain position there for the remainder of the episode. OG-MARL's decentralized consensus allows agents to rapidly share learned dynamics, resulting in optimal goal allocation (see OG-MARL video[3]) and therefore lower regret as visualized in fig. 6.

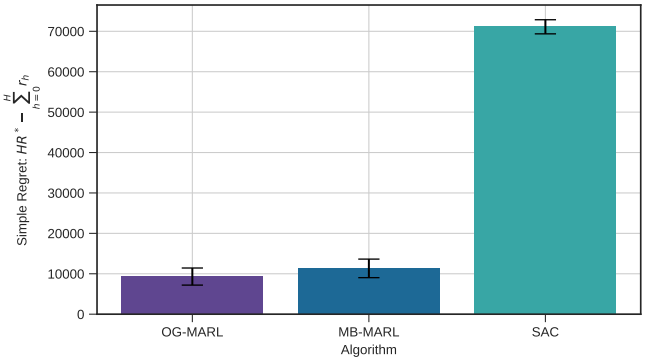

Fig. 6. Comparison of simple regret among the different algorithms in the goal collection task with action penalties.

c) Optimism: Encouraging Exploration: OG-MARL augments each agent's reward with an intrinsic exploration bonus proportional to the model uncertainty, implementing an upper-confidence bound strategy. This intrinsic optimism drives agents to explore rarely visited regions rather than greedily exploiting known rewards. The standard Global MB-MARL approach, which lacks such a bonus, occasionally converges to suboptimal goal combinations that are not necessarily the closest (see MB-MARL video[4] and appendix C). In contrast, OG-MARL reliably discovers and occupies the highest-value goals nearest each agent. Figure 7 presents an ablation over different optimism-scale values, demonstrating that an appropriately tuned optimism level is critical for balancing exploration and exploitation in OG-MARL. Too little optimism does not explore potentially better policies, while overly optimistic levels explore too much and delay convergence.

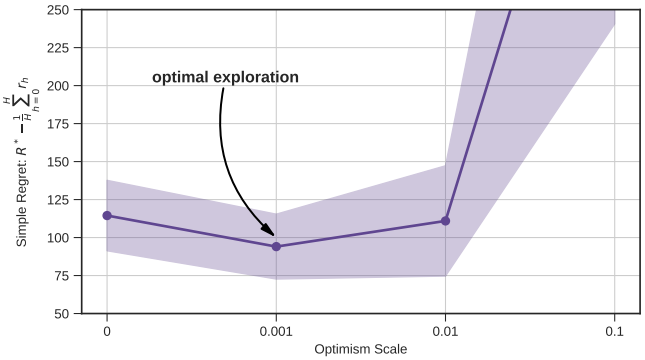

Fig. 7. Simple regret ablation for different optimism levels of OG-MARL.

VII. Limitations

OG-MARL and other GP-based algorithms naturally scale with $\mathcal{O}(N^3)$ due to matrix inversions in the Kernel matrix computation. There exists work to reduce the computational complexity [7]; however, scalability remains a challenge for high-dimensional state space systems with many agents. Different approaches that use Bayesian neural networks or other function approximations do not guarantee formal sampling bounds. Further, OG-MARL is limited to the additive environment structure from assumption 4, which is not necessarily applicable to any multi-agent objectives and systems.

VIII. Conclusion

In this work, we introduce OG-MARL, an optimistic, distributed model-based MARL algorithm in which each agent maintains its own sparse GP world model and fuses these models via a consensus protocol to a global distributed world model. We prove that under mild connectivity and additivity assumptions, the consensus step recovers the same posterior mean and uncertainty estimates as a centralized sparse GP trained on all unified agent data. By leveraging these shared models in an optimism-in-the-face-of-uncertainty planning scheme, OG-MARL guarantees sublinear simple regret and provably converges to near-optimal cooperative policies, despite never requiring any agent to observe or communicate full joint actions. Our approach combines the data-efficiency and exploration guarantees of model-based RL with the advantages of decentralized learning. Empirically, we demonstrate enhanced problem-solving capabilities in cooperative multi-agent tasks, as each agent continually enriches its local model by exchanging the experience with its neighboring agents.

[2]Video SAC: https://t1p.de/9hbzv
[3]Video OG-MARL: https://t1p.de/cqal5
[4]Video MB-MARL: https://t1p.de/1h5pr

REFERENCES

[1] Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3, 1922.

[2] F. Bullo. *Lectures on Network Systems*. Kindle Direct Publishing, 1.7 edition, 2024.

[3] Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 2008.

[4] Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[5] Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, 2019.

[6] Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[7] Elad Gilboa, Yunus Saatçi, John Cunningham, and Elad Gilboa. Scaling multidimensional Gaussian processes using projected additive approximations. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[8] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning,ICML 2018*, Proceedings of Machine Learning Research, 2018.

[9] Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. *When to trust your model: model-based policy optimization*. 2019.

[10] Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.

[11] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision-making in multiarmed bandits: Frequentist and bayesian algorithms. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 2016.

[12] Peter Landgren, Vaibhav Srivastava, and Naomi Ehrich Leonard. Distributed cooperative decision making in multi-agent multi-armed bandits. *Automatica*, 125, 2021.

[13] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pages 157–163. Morgan Kaufmann, 1994.

[14] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[15] David Martínez-Rubio, Varun Kanade, and Patrick Rebeschini. Decentralized cooperative stochastic bandits. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, 2019.

[16] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[17] Pier Giuseppe Sessa, Maryam Kamgarpour, and Andreas Krause. Efficient model-based multi-agent reinforcement learning via optimistic equilibrium computation. In *International Conference on Machine Learning*, 2022.

[18] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 2010.

[19] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 2012.

[20] Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Florian Dorfler, Pieter Abbeel, and Andreas Krause. Optimism via intrinsic rewards: Scalable and principled exploration for model-based reinforcement learning. In *7th Robot Learning Workshop: Towards Robots with Human-Level Abilities*, 2025.

[21] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. *Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms*. Springer International Publishing, 2021.

[22] Jingxuan Zhu, Romeil Sandhu, and Ji Liu. A distributed algorithm for sequential decision making in multi-armed bandit with homogeneous rewards. In *2020 59th IEEE Conference on Decision and Control (CDC)*, 2020.

[23] Zhaowei Zhu, Jingxuan Zhu, Ji Liu, and Yang Liu. Federated bandit: A gossiping approach. *Proc. ACM Meas. Anal. Comput. Syst.*, 2021.

APPENDIX

A. Practical Implementation

Our implementation of OG-MARL[5] is based on the model-based policy optimization algorithm MBPO [9], which uses a modified version of soft actor-critic SAC [8] for policy optimization. In this implementation, we fix the optimistic scale throughout learning and treat it as a hyper-parameter, which is ablated in fig. 7.

B. Hyper-paramters

We use for the cooperative goal collection task the following hyper-parameters for the different algorithms:

Parameters	**SAC**	**OG-MARL**	**MB-MARL**
Training steps	10000	10000	10000
Number of agents	5	5	5
Buffer size	10000	10000	10000
Batch size	256	256	256
γ	0.99	0.99	0.99
τ (target updates)	0.005	0.005	0.005
Learning rates	3×10^{-4}	3×10^{-4}	3×10^{-4}
Model train freq.	250	250	250
Updates per env step	1	10	10
Roll-out length	–	2	2
Optimism scale	–	0.01	0.0
Consensus rounds	–	100	–

TABLE I
HYPERPARAMETERS FOR THE COOPERATIVE REWARD COLLECTION TASK.

C. Visual Experiment Evaluation

We also visually analyze the impact of the optimism and cooperation in the cooperative reward collection task. We compare no optimism with optimism scale 0.01 of OG-MARL on the same random seed for training. We observe that agent $i = 2$ does not discover the nearby goal and moves from fig. 8 to fig. 9 to the centered highest goal instead. In contrast, when using optimism scale 0.01, we observe that agent $i = 2$ discovers its nearby goal from fig. 8 to fig. 10. Comparing the performances of OG-MARL with the non-cooperating SAC agents, we can see that only the agents that are very close to a goal discover their nearby goal form fig. 8 to fig. 11 but do not share the information of the other agents.

[5]GitHub: https://github.com/ManuelWendl/OG-MARL

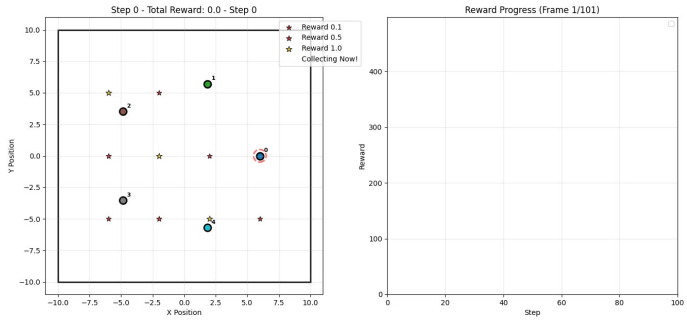

Fig. 8. Initial configuration of cooperative reward collection task.

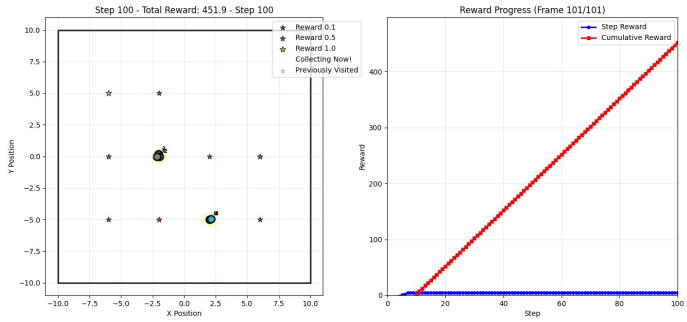

Fig. 9. Final agent positions for OG-MARL without optimism (with action penalty).

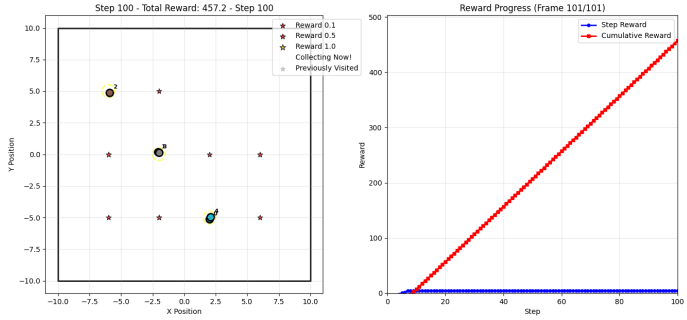

Fig. 10. Final agent positions for OG-MARL with 0.01 optimism (with action penalty).

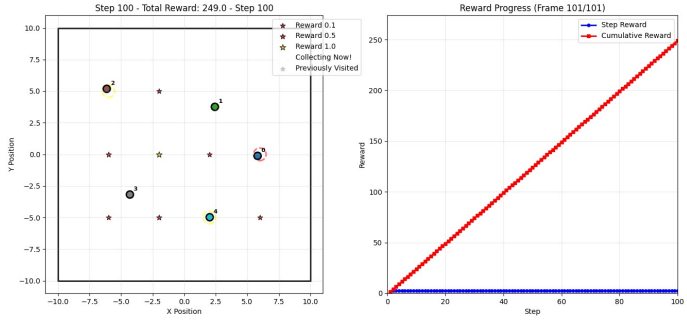

Fig. 11. Final positions of individual SAC agents (without action penalty).