Multi-Agent Probabilistic Ensembles With Trajectory Sampling for Connected Autonomous Vehicles

Ruoqi Wen , *Student Member, IEEE*, Jiahao Huang, Rongpeng Li , *Senior Member, IEEE*, Guoru Ding ,
and Zhifeng Zhao , *Member, IEEE*

Abstract—**Connected Autonomous Vehicles (CAVs) have attracted significant attention in recent years and Reinforcement Learning (RL) has shown remarkable performance in improving the autonomy of vehicles. In that regard, Model-Based RL (MBRL) manifests itself in sample-efficient learning, but the asymptotic performance of MBRL might lag behind the state-of-the-art Model-Free RL (MFRL) algorithms. Furthermore, most studies for CAVs are limited to the decision-making of a single vehicle only, thus underscoring the performance due to the absence of communications. In this study, we try to address the decision-making problem of multiple CAVs with limited communications and propose a decentralized Multi-Agent Probabilistic Ensembles (PEs) with Trajectory Sampling (TS) algorithm namely `MA-PETS`. In particular, to better capture the uncertainty of the unknown environment, `MA-PETS` leverages PE neural networks to learn from communicated samples among neighboring CAVs. Afterward, `MA-PETS` capably develops TS-based model-predictive control for decision-making. On this basis, we derive the multi-agent group regret bound affected by the number of agents within the communication range and mathematically validate that incorporating effective information exchange among agents into the multi-agent learning scheme contributes to reducing the group regret bound in the worst case. Finally, we empirically demonstrate the superiority of `MA-PETS` in terms of the sample efficiency comparable to MFRL.**

Index Terms—**Autonomous vehicle control, multi-agent model-based reinforcement learning, probabilistic ensembles with trajectory sampling.**

I. INTRODUCTION

RECENTLY, there has emerged significant research interest towards Connected Autonomous Vehicles (CAVs) with a particular emphasis on developing suitable Reinforcement Learning (RL)-driven controlling algorithms [2], [3] for the optimization of intelligent traffic flows [4], decision-making [5],

and control of AVs [6]. Notably, these RL methods capably learn complex tasks for CAVs through effective interaction between agents and the environment, and existing RL works for CAVs can be classified as Model-Free Reinforcement Learning (MFRL) and Model-Based Reinforcement Learning (MBRL), the key differences of which lie in whether agents estimate an explicit environment model for the policy learning [7].

Conventionally, MFRL, which relies on collected rewards on recorded state-action transition pairs, has been widely applied to model complex mixed urban traffic systems in multi-vehicle scenarios, showing excellent performance in various situations of autonomous driving [8], [9]. Typical examples of MFRL include MADDPG [10], COMA [11], QMIX [12], SVMIX [13]. However, the computational complexity of most MFRL algorithms grows exponentially with the number of agents. To solve training data scarcity-induced out-of-distribution (OOD) problems, MFRL is typically required to repetitively interact with the real world to collect a sufficient amount of training data, which might be infeasible in practice and possibly lead to learning instability and huge overhead. On the contrary, due to the impressive sample efficiency, MBRL promises to solve CAVs issues [14] more capably and starts to attain some research interest [15], [16].

Typically, MBRL is contingent on learning an accurate probabilistic dynamics model that can clearly distinguish between *aleatoric* and *epistemic* uncertainty [17], where the former is inherent to the system noise, while the latter stems from sample scarcity and contributes to solving the OOD problem to a certain extent. Afterward, based on the learned dynamics model from the collected data, MBRL undergoes a planning and control phase by simulating model-consistent transitions and optimizing the policy accordingly. However, the asymptotic performance of MBRL algorithms has lagged behind state-of-the-art MFRL methods in common benchmark tasks, especially as the environmental complexity increases. In other words, although MBRL algorithms tend to learn faster, they often converge to poorer results [18], [19]. Consequently, the deployment of model-based strategies to intricate tasks often encounters substantial compounded errors, which may impede trajectory learning. However, by extending single-agent RL [20] to multi-agent contexts through efficient communication protocols, one can compensate for the learning deficiency to some extent [21], [22]. In that regard, despite the simplicity of assuming the existence of a central controller, it might be practically infeasible or cost-effective to install such a controller in many real-world

scenarios. Meanwhile, it is often challenging to establish fully connected communication between all agents, where the required communication overhead can scale exponentially [8], [10]. Instead, a more complex communication-limited decentralized multi-agent MBRL for CAVs becomes feasible, and specifying a suitable protocol for cooperation between agents turns crucial.

On the other hand, to theoretically characterize the sampling efficiency of RL, the concept of the regret bound emerges, which targets to theoretically measure the T-time-step difference between an agent's accumulated rewards and the total reward that an optimal policy (for that agent) would have achieved. Without loss of generality, for an H-episodic RL environment with S states and A actions, contingent on Hoeffding inequality and Bernstein inequality, tabular upper-confidence bound (UCB) algorithms can lead to a regret bound of $\tilde{O}(\sqrt{H^4 SAT})$ and $\tilde{O}(\sqrt{H^3 SAT})$ respectively [23], [24], where $\tilde{O}(\cdot)$ hides the logarithmic factors. On the other hand, for any communicating Markov decision process (MDP) with the diameter D, Jaksch et al. propose a classical confidence upper bound algorithm `UCRL2` algorithm (abbreviation of *Upper Confidence bound for RL*) to achieve the regret bound $\tilde{O}(DS\sqrt{AT})$. In the literature, there only sheds little light on the regret bound of collaborative Multi-Agent Reinforcement Learning (MARL). In this paper, inspired by the regret bound of `UCRL2`, we investigate the regret bound of decentralized communication-limited MBRL, and demonstrate how communication among the multi-agents can be used to reduce the regret bound and boost the learning performance.

In this paper, targeted at addressing the sample efficiency issue in a communication-limited multi-agent scenario, we propose a fully decentralized Multi-Agent Probabilistic Ensembles with Trajectory Sampling (`MA-PETS`) algorithm. In particular, `MA-PETS` could effectively exchange collected samples from individual agents to neighboring vehicles within a certain range, and extend the widely adopted single-agent Probabilistic Ensemble (PE) technique to competently reduce both aleatoric and epistemic uncertainty while fitting the multi-agent environmental transition. Furthermore, `MA-PETS` employs Model Predictive Control (MPC) [25], [26], due to its excellence in scenarios that necessitate the joint prediction and optimization of future behavior, to generate appropriate control actions from a learned-model-based Trajectory Sampling (TS) approach. Compared to the existing literature, the contribution of the paper can be summarized as follows:

- We formulate the decentralized multi-agent decision-making issue for CAVs as a parallel time-homogeneous Markov Decision Process (MDP). On this basis, we devise a sample-efficient MBRL solution `MA-PETS` which utilizes a multi-agent PE to learn the unknown environmental transition dynamics model from data exchanged within a limited range, and calibrate TS-based MPC for model-based decision-making.
- We analyze the group regret bound of `MA-PETS`, which is based on `UCRL2` and fundamentally different from the approach in [22] that utilizes a mere superposition of

TABLE I
THE KEY PARAMETERS FOR THE ALGORITHMS

Notation	Parameters Description
I	Number of RL agents
d	Communication range
v_t	Speed per vehicle at time-step t
z_t	Position per vehicle at t
$\bar{v}_t$	Target velocity per vehicle at t
$v_{t,a}, v_{t,e}$	Speed of vehicles ahead and behind
$l_{t,a}, l_{t,e}$	Distance of vehicles ahead and behind
Λx_t	Travel distance per vehicle at t
K	No. of episodes
H	Length per episode
B	No. of Ensembles for dynamics Model
w	Horizon of MPC
Q	Candidate action sequences in CEM
X	Elite candidate action sequences in CEM
P	No. of particles
Y	Max iteration of CEM
$u_k(s,a)$	Number of state-action (s,a) visits in the episode k
$N_k(s,a)$	Number of state-action (s,a) visits before episode k
$\mathbf{C}_d$	Cliques cover of $\mathcal{G}_d$
$\bar{\chi}(\mathcal{G}_d)$	Minimum number of cliques cover $\mathbf{C}_d$

single-agent regret bounds. By incorporating the concept of clique cover [27] for a limited-communication, undirected graph, we theoretically demonstrate that, even in the worst case, augmenting a multi-agent algorithm with additional communication information still results in a sub-linear group regret relative to the number of agents, and accelerates convergence. This validates that multiple agents jointly exploring the state-action space in similar environments could communicate to discover the optimal policy faster than individual agents operating in parallel.

- We further illustrate our approach experimentally on a CAV simulation platform SMARTS [28] and validate the superiority of our proposed algorithm over other MARL methods in terms of sample efficiency. Besides, we evaluate the impact of communication ranges on `MA-PETS` and demonstrate the contributing effects of information exchange.

The remainder of the paper is organized as follows. Section II briefly introduces the related works. In Section III, we introduce the preliminaries of MDPs and formulate the system model. In Section IV and Section V, we present the details of `MA-PETS` and unveil the effect of communication range on the convergence of the MARL via the group regret bound, respectively. Finally, Section VI demonstrates the effectiveness and superiority `MA-PETS` through extensive simulations. We conclude the paper in Section VII.

II. RELATED WORKS

A. Multi-Agent Reinforcement Learning of CAVs

Decision-making and planning are crucial components of CAV systems, with significant implications for enhancing the safety, driving experience, and efficiency of CAVs [3], [29]. The decision-making of CAVs in high-density mixed-traffic intersections belongs to one of the most challenging tasks and attracts

significant research interest [30], [31], [32], towards improving traffic efficiency and safety. In this regard, Centralized Training and Decentralized Execution (CTDE) has become one of the most popular paradigms. For instance, within the actor-critic MFRL framework, [10], [11] proposes using a central critic for centralized training coupled with multiple actors for distributed execution. However, as the number of agents increases or the action space expands, the computational complexity of MFRL algorithms like `MA-DDPG` can rise exponentially, presenting a significant challenge. To mitigate this computational burden, [12] proposes `QMIX` to involve the value decomposition of the joint value function into separate individual value functions for each agent. In addition, [13] proposes to adopt a stochastic graph neural network to capture dynamic topological features of time-varying graphs while decomposing the value function. To avoid the large number of communication overheads that frequent information exchange incurs, [33] develops a consensus-based optimization scheme on top of the periodic averaging method, which introduces the consensus algorithm into Federated Learning (FL) for the exchange of a model's local gradients.

Although these studies showcase the robust post-training performance of RL in real-time applications, the MFRL algorithm is still plagued by high overhead, stemming from the extensive computational and sampling requirements. This issue is particularly pronounced in scenarios such as CAVs [34], where data acquisition is challenging, interactions between agents and the environment are inefficient, and each iteration yields only a limited amount of data. Therefore, as data samples become scarce, the performance of MFRL can quickly degrade and become unstable.

B. Model-Based Reinforcement Learning

To address the sampling efficiency and communication overhead issues in MFRL, MBRL naturally emerges as an alternative solution. Unfortunately, MBRL suffers from performance deficiency, as it might fail to accurately estimate the uncertainty in the environment and characterize the dynamics model, which belongs to a critical research component in MBRL. For example, [35] proposes a DNN-based method that is, to some extent, qualified to separate aleatoric and epistemic uncertainty while maintaining appropriate generalization capabilities, while PILCO [36] marginalizes aleatoric and epistemic uncertainty of a learned dynamics model to optimize the expected performance. [19] takes advantage of MPC to optimize the RL agent's behavior policy by predicting and planning within the modeled virtual environment at each training step. Another type of MBRL falls within the scope of Dyna-style methods [37], where additional data is generated from interactions between the RL and the virtual environment, thus improving the efficiency of the RL. [15] proposes an uncertainty-aware MBRL and verifies that it has competitive performance as the state-of-the-art MFRL. [8] proposes an RL training algorithm `MAPPO` that incorporates a prior model into `PPO` algorithm to speed up the learning process using current centralized coordination methods. However, centralized coordination methods face issues

such as high resource-intensiveness, inflexibility, and annoying delays, since the central node must process a large amount of information and handle decision-making for the entire system. In addition, there is still little light shed on the multi-agent MBRL scenario, especially in the communication-limited case [38]. Considering the substantial communication overhead associated with centralized coordination methods, we mainly focus on the fully decentralized multi-agent MBRL algorithm under communication constraints.

C. Regret Bounds of Reinforcement Learning

Understanding the regret bound of online single-agent RL-based approaches to a time-homogeneous MDP has received considerable research interest. For example, [39] discusses the performance guarantees of a learned policy with polynomial scaling in the size of the state and action spaces. [40] introduces a `UCRL` algorithm and shows that its expected online regret in unichain MDPs is $O(\log T)$ after T steps. Furthermore, [41] proposes the a `UCRL2` algorithm which is capable of identifying an optimal policy through Extended Value Iteration (EVI), by conjecturing a set of plausible MDPs formed within the confidence intervals dictated by the Hoeffding inequality. Moreover, [41] demonstrates that the total regret for an optimal policy can be effectively bounded by $\tilde{O}(DS\sqrt{AT})$. Afterward, many variants of `UCRL2` have been proposed for the generation of tighter regret bounds. For instance, [42] proposes a `UCRL` algorithm that derives sub-linear regret bounds $\tilde{O}(T^{\frac{2+\alpha}{2+2\alpha}})$ with a parameter α for un-discounted RL in continuous state space. By using more efficient posterior sampling for episodic RL, [43] achieves the expected regret bound $\tilde{O}(\iota S\sqrt{AT})$ with an episode length ι. [44] introduces a non-parametric tailored multiplier bootstrap method, which significantly reduces regret in a variety of linear multi-armed bandit challenges. Similar to the single-agent setting, agents in MARL attempt to maximize their cumulative reward by estimating value functions, and the regret bound can be analyzed as well. For example, [45] proposes that a specific class of online, episodic, tabular multi-agent Q-learning problems with UCB-Hoeffding exploration through communication yields a regret $\mathcal{O}(\sqrt{IH^4SAT\iota})$, where I is the number of RL agents and $\iota := \log(SATI/p)$. Nevertheless, despite the progress on single-agent regret bound and the group regret bound for fully connected multi-agent cases [22], the group regret of MBRL in communication-limited multi-agent scenarios remains an unexplored area of research. Consequently, we integrate dynamic graph theory, akin to that is found in [45], with an analysis of small blocks to explore the group regret bounds within a distinct MARL algorithm.

III. PRELIMINARIES AND SYSTEM MODEL

In this section, we briefly introduce some fundamentals and necessary assumptions of the underlying MDPs and the framework of MBRL. On this basis, towards the decision-making issue for CAVs, we highlight how to formulate MBRL-based problems.

A. Preliminaries

1) Parallel Markov Decision Process: The decision-making problem of CAVs can be formulated as a collection of *parallel time-homogeneous stochastic MDPs* $\mathcal{M} := \{\text{MDP} (\mathcal{S}^{(i)}, \mathcal{A}^{(i)}, \mathcal{P}^{(i)}, \mathcal{R}^{(i)}, H)\}_{i=1}^{I}$ among agents $i \in [I] := \{1, 2, \ldots, I\}$ for H-length episodes [46]. Despite its restrictiveness, parallel MDPs provide a valuable baseline for generalizing to more complex environments, such as heterogeneous MDPs. Notably, agents have access to identical state and action space (i.e., $\mathcal{S}^{(i)} = \mathcal{S}^{(j)}$ and $\mathcal{A}^{(i)} = \mathcal{A}^{(j)}$, $\forall i, j \in [I]$). We assume bounded rewards, where all rewards are contained in the interval $[0, r_{\max}]$ with mean $\bar{r}(s, a)$. Besides, we assume the transition functions $\mathcal{P}^{(i)}$ and reward functions $\mathcal{R}^{(i)}$ only depend on the current state and chosen action of agent i, and conditional independence applies between the transition function and the reward function. Furthermore, we focus exclusively on stationary policies, denoted as $\pi^{(i)} : \mathcal{S}^{(i)} \rightarrow \mathcal{A}^{(i)}$, which indicates the taken action $a_t^{(i)} \in \mathcal{A}^{(i)}$ for an agent $i \in [I]$ after observing a state $s_t^{(i)} \in \mathcal{S}^{(i)}$ at time-step t. Based on the taken action, the agent i receives a reward $r_t^{(i)} = \mathcal{R}^{(i)}(s_t^{(i)}, a_t^{(i)})$, and the environment transitions to the next state $s_{t+1}^{(i)}$ according to an unknown dynamics function $\mathcal{P}^{(i)}(s_{t+1}^{(i)} \mid s_t^{(i)}, a_t^{(i)}) : \mathcal{S}^{(i)} \times \mathcal{A}^{(i)} \rightarrow \mathcal{S}^{(i)}$. In other words, each agent interacts with the corresponding homogeneous MDP and calculates an expectation over trajectories where action $a_{t+1}^{(i)}$ follows the distribution $\pi^{(i)}(s_t^{(i)})$.

On the other hand, an MDP is called *communicating*, if for any two states $s, s' \in \mathcal{S}$, there always exists a policy π that guarantees a finite number of steps $L_{\pi}(s, s')$ to transition from state s to state s' when implementing policy π. In such a communicating MDP, the opportunity for recovery remains viable even if an incorrect action is executed. As previously mentioned, the diameter D of the communicating MDP $M^{(i)}$, which measures the maximal distance between any two states in the communicating MDP, can be defined as

$$D(M^{(i)}) := \max_{s, s' \in \mathcal{S}} \min_{\pi^{(i)}} L_{\pi^{(i)}}(s, s'). \tag{1}$$

Moreover, unlike cumulative rewards over T steps, we take account of average rewards, which can be optimized through a stationary policy $\pi^{*(i)}$ [47]. The objective is for each agent to learn a policy $\pi^{*(i)}$ that maximizes its individual average reward after any T steps, i.e,

$$\pi^{*(i)} = \arg\sup_{\pi \in \Pi^{(i)}} \left\{ \liminf_{T \rightarrow +\infty} \mathbb{E}_{\pi} \left[\frac{1}{T} \sum_{t=1}^{T} r_t^{(i)} \right] \right\}, \tag{2}$$

where $\Pi^{(i)}$ is the set of the plausible stationary randomized policies.

Lemma 1 (Lemma 10 of [48]): Consider $M^{(i)}$ as a time-homogeneous and communicating MDP with a diameter of D. Let $\mathcal{R}_T^{*(i)}(M^{(i)})$ represent the optimal T-step reward, and $\rho^{*(i)}$ denote the optimal average reward under the reward function $\mathcal{R}^{(i)}$. It can be asserted that for any MDP $M^{(i)}$, the disparity between the optimal T-step reward and the optimal average reward

is minor, capped at a maximum of order D. Mathematically,

$$\mathcal{R}_T^{*(i)}(M^{(i)}) \leq T\rho^{*(i)}(M^{(i)}) + D \cdot r_{\max}^{(i)}. \tag{3}$$

Hence, by Lemma 1, the optimal average reward $\rho^{*(i)}$ serves as an effective approximation for the expected optimal reward over T-steps. To evaluate the convergence of the RL algorithms for each agent i, we consider its regret after an arbitrary number of steps T, defined as

$$\text{Regret}^{(i)}(T) := T\rho^{*(i)} - \sum_{t=1}^{T} r_t^{(i)}. \tag{4}$$

In the CAV scenario, as it is challenging to consider each agent's regret bound independently, we introduce the concept of multi-agent group regret, which is defined as

$$\text{Regret}_G(T) = \sum_{i=1}^{I} \text{Regret}^{(i)}(T). \tag{5}$$

2) Model-Based Reinforcement Learning: An MBRL framework typically involves two phases (i.e., dynamics model learning, and planning & control). In the dynamics model learning phase, each agent $i \in [I]$ estimates the dynamics model $\tilde{\mathcal{P}}^{(i)}$ from collected environmental transition samples by a continuous model (e.g., DNNs). Afterward, based on the approximated dynamics model $\tilde{\mathcal{P}}^{(i)}$, the agent simulates the environment and makes predictions for subsequent action selection. In that regard, we can evaluate w-length action sequences $\mathbf{A}_{t:t+w-1}^{(i)} = (a_t^{(i)}, \ldots, a_{t+w-1}^{(i)})$ by computing the expected reward over possible state-action trajectories $\tau_{t:t+w-1}^{(i)} = (s_t^{(i)}, a_t^{(i)}, \ldots, s_{t+w-1}^{(i)}, a_{t+w-1}^{(i)}, s_{t+w}^{(i)})$, and optimize the policy accordingly. Mathematically, this planning & control phase can be formulated as

$$\mathbf{A}_{t:t+w-1}^{*(i)} = \arg\max_{\mathbf{A}_{t:t+w-1}^{(i)}} \sum_{t'=t}^{t+w-1} r_{t'}^{(i)}$$

$$s.t. \quad s_{t+1}^{(i)} \sim \tilde{\mathcal{P}}^{(i)} \left(s_t^{(i)}, a_t^{(i)} \right). \tag{6}$$

Limited by the non-linearity of the dynamics model, it is usually difficult to calculate the exact optimal solution of (6). However, many methods exist to obtain an approximate solution to the finite-level control problem and competently complete the desired task. Common methods include the traversal method, Monte Carlo Tree Search (MCTS) [49], Iterative Linear Quadratic Regulator (ILQR) [50], etc.

B. System Model

As illustrated in Fig. 1, we consider a mixed autonomy traffic system model with I CAVs and some human-driving vehicles (HVs). Consistent with the terminology of parallel MDPs, the state $s_t^{(i)} \in \mathcal{S}^{(i)} = (v_t^{(i)}, x_t^{(i)}, y_t^{(i)}, v_{t,a}^{(i)}, v_{t,e}^{(i)}, l_{t,a}^{(i)}, l_{t,e}^{(i)})$ for agent $i \in [I]$ at time-step t encompasses information like its velocity $v_t^{(i)} \in \mathbb{R}$, its position $z_t^{(i)} = (x_t^{(i)}, y_t^{(i)}) \in \mathbb{R}^2$, the velocity and relative distance of the vehicles ahead and behind (i.e., $v_{t,a}^{(i)}, v_{t,e}^{(i)}, l_{t,a}^{(i)}, l_{t,e}^{(i)} \in \mathbb{R}$). Meanwhile, each vehicle i is controlled by an adjustable target velocity $\bar{v}_t^{(i)} \in \mathbb{R}$ (i.e., $a_t^{(i)} = (\bar{v}_t^{(i)})$). Furthermore, we assume the dynamics model shall be learned via interactions with the environment. At the same time, the
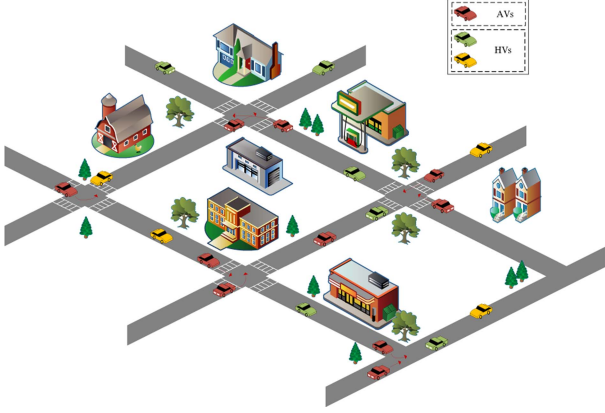
Fig. 1. The complex mixed urban traffic control scenario with unsignalized lane merging and intersections.

reward function can be calibrated in terms of the velocity and collision-induced penalty, thus being known beforehand.

For simplicity, let t_k be the start time of episode k. Within the framework of decentralized multi-agent MBRL, we assume the availability of state transition dataset $\mathcal{D}_{0:t}^{(i)} = \{s_{t'}^{(i)}, a_{t'}^{(i)}, s_{t'+1}^{(i)}\}_{t'=0}^{t}$ and each agent approximates $\tilde{\mathcal{P}}^{(i)}, \forall i \in [I]$ by DNNs parameterized by θ based on historical dataset and exchanged samples from neighbors. In particular, for a time-varying undirected graph $\mathcal{G}_k$ constituted by I CAVs, each CAV can exchange its latest H-length[1] dataset $\mathcal{D}_{t_k-H:t_k}^{(i)}$ with neighboring CAVs within the communication range $d \in [0, D(\mathcal{G}_k)]$ at the end of an episode k, where $D(\mathcal{G}_k)$ is the dynamic diameter[2] of graph $\mathcal{G}_k$. Therefore, when $d = 0$, the multi-agent MBRL problem reverts to I parallel single-agent cases. Mathematically, the dataset for dynamics modeling at the end of an episode k could be denoted as $\mathcal{D}_{0:t_k}^{(i)+} := \mathcal{D}_{0:t_k-H-1}^{(i)+} \cup \{\mathcal{D}_{t_k-H:t_k}^{(j)}\}$ for all CAV $j \in \mathcal{F}_{d,k}^{(i)}$, where the set $\mathcal{F}_{d,k}^{(i)}$ encompasses all CAVs satisfying $\mathrm{dis}_t(i,j) \leq d$ with $\mathrm{dis}_t(i,j)$ computing the Euclidean distance between CAV i and j at the last time-step $t_k + H$. Furthermore, the planning and control objective for multi-agent MBRL in a communication-limited scenario can be re-written as

$$J\left(\tilde{\mathcal{P}}^{(i)}, d\right) = \mathbb{E}_{\boldsymbol{\pi} \sim \tilde{\mathcal{P}}^{(i)}}\left[\sum_{t'=0}^{H} r_{t'}^{(i)} \mid s_0^{(i)}\right]$$

$$s.t. \quad \tilde{\mathcal{P}}^{(i)} \propto \mathcal{D}_{0:t}^{(i)+}, 0 \leq d \leq D(\mathcal{G}_k)$$

$$s_0^{(i)} \sim p(s_0), s_{t+1}^{(i)} \sim \tilde{\mathcal{P}}^{(i)}(s_t^{(i)}, a_t^{(i)}), \forall t \in [H], \quad (7)$$

[1]Notably, the applied length could be episode-dependent but is limited by the maximum value H, as will be discussed in Section IV.

[2]We slightly abuse the notation of the diameter for an MDP and that for a time-varying undirected graph.

where $\tilde{\mathcal{P}}^{(i)}$ is the learned dynamics model based on $\mathcal{D}_{0:t}^{(i)+}$. In this paper, we resort to a model-based PETS solution for solving (7) and calculating the group regret bound.

IV. THE MA-PETS ALGORITHM

In this section, we discuss how to extend PETS [19] to a multi-agent case and present the dynamics model learning and planning & control phases in MA-PETS, which can be depicted as in Fig. 2.

A. Learning the Dynamics Model

In MA-PETS, we leverage an ensemble of bootstrapped probabilistic DNNs to reduce both aleatoric and epistemic uncertainty. In particular, in order to combat the aleatoric uncertainty, we approximate the dynamics model at time-step t (i.e., $\tilde{\mathcal{P}}(s_{t+1}^{(i)} \mid s_t^{(i)}, a_t^{(i)}), \forall i \in [I]$) by a probabilistic DNN, by assuming that the conditioned probability distribution of s_{t+1} follows a Gaussian distribution $\mathcal{N}(\mu_\theta, \sigma_\theta)$ with mean μ_θ and diagonal covariance σ_θ parameterized by θ. In other words,

$$\tilde{\mathcal{P}}_\theta^{(i)} = \arg\min_\theta \frac{1}{\left|\hat{\mathcal{D}}^{(i)+}\right|} \mathrm{loss}_{\mathrm{Gauss}}(\theta), \quad (8)$$

where $\hat{\mathcal{D}}^{(i)+}$ denotes a sub-set sampled from $\mathcal{D}^{(i)+}$ and $\mathrm{loss}_{\mathrm{Gauss}}(\theta)$ is defined as in (9) shown at the bottom of the page. Consistent with the definition of states and actions in Section III-B, $\mu_\theta(s,a) \in \mathsf{R}^7$ and $\sigma_\theta(s,a) \in \mathsf{R}^{7 \times 7}$.

In addition, in order to mitigate the epistemic uncertainty, which arises primarily from the lack of sufficient data, a PE method is further adopted. Specifically, MA-PETS consists of B bootstrap models in the ensemble, each of which is an independent and identically distributed probabilistic DNN $\tilde{\mathcal{P}}_{\theta_b}^{(i)}, b \in [B]$ from a unique dataset $\hat{\mathcal{D}}_b^{(i)+}$ uniformly sampled from $\mathcal{D}^{(i)+}$ with the same size. Typically, [19] points out that $B = 5$ yields satisfactory results.

B. Planning and Control With Learned Dynamics

Based on a learned dynamics model $\tilde{\mathcal{P}}^{(i)}$, MA-PETS tries to obtain a solution to (6) by resorting to a sample-efficiency controller MPC [25]. Generally speaking, MPC excels in providing high-precision trajectory planning, thus reducing the risk of collisions during vehicle operation. Moreover, it can predict both aleatoric and epistemic uncertainty accurately within the learned dynamics model [32]. Without loss of generality, assume that for a time-step t, agent i observes the state $s_t^{(i)}$. Afterward, agent i leverages the Cross-Entropy Method (CEM) [51] to generate Q candidate action sequences within a w-horizon of MPC. Initially, each candidate action sequence $\mathbf{A}_{t:t+w-1}^{(q,i)}, \forall q \in$

$$\mathrm{loss}_{\mathrm{Gauss}}(\theta) = -\sum_{(s_t, a_t, s_{t+1}) \in \hat{\mathcal{D}}^{(i)+}} \log \tilde{\mathcal{P}}_\theta^{(i)}(s_{t+1} \mid s_t, a_t)$$

$$= \sum_{(s_t, a_t, s_{t+1}) \in \hat{\mathcal{D}}^{(i)+}} \left\{ [\mu_\theta(s_t, a_t) - s_{t+1}]^\top \sigma_\theta^{-1}(s_t, a_t) [\mu_\theta(s_t, a_t) - s_{t+1}] + \log \det \sigma_\theta(s_t, a_t) \right\} \quad (9)$$
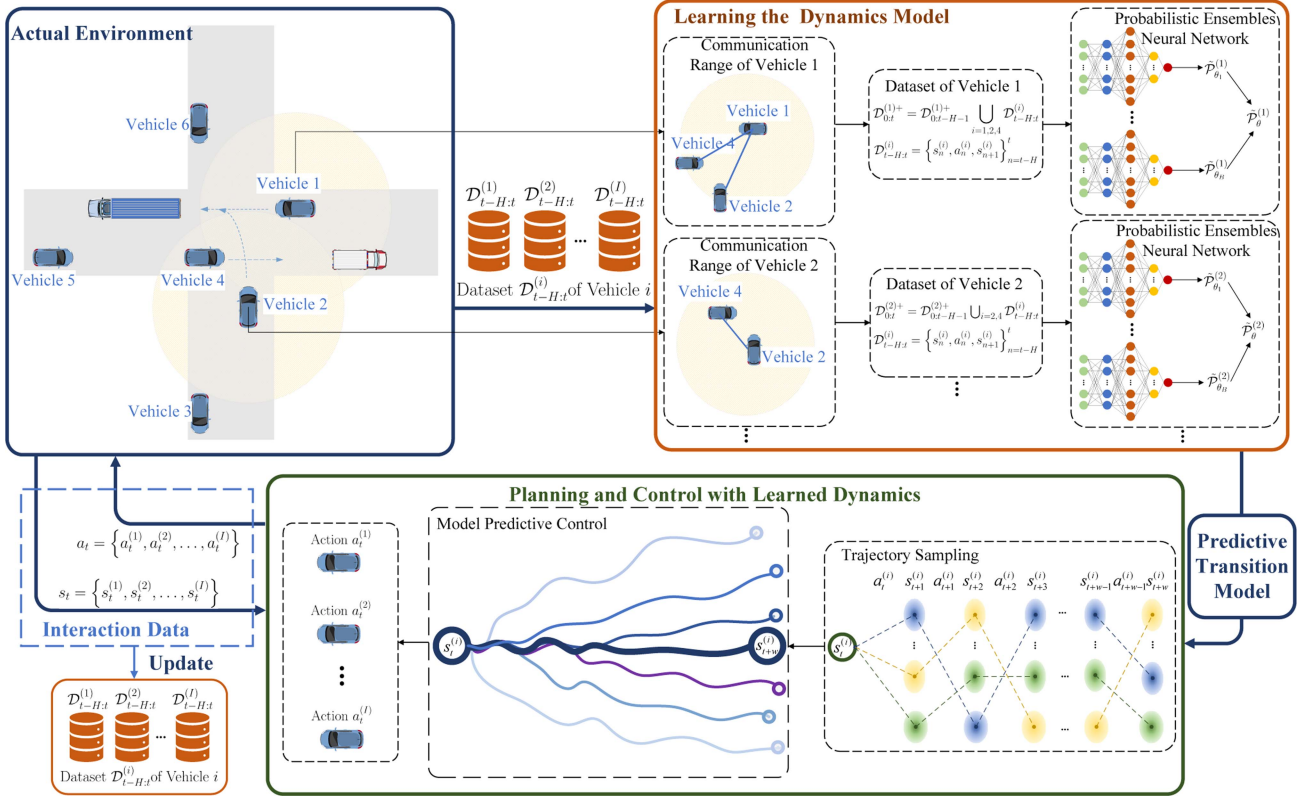
Fig. 2. The illustration of the `MA-PETS` algorithm for CAVs.

$[Q]$ is sampled following a Gaussian distribution $\mathcal{N}(\mu_\phi, \sigma_\phi)$ parameterized by ϕ. Meanwhile, agent i complements possible next-time-step states from the dynamics model ensembles, by adopting particle-based TS [52]. In other words, for each candidate action sequence $\mathbf{A}_{t:t+w-1}^{(q,i)}$, TS predicts plausible state-action trajectories by simultaneously creating P particles that propagate a set of Monte Carlo samples from the state $s_t^{(i)}$. Due to the randomness in the learned and assumed time-invariant dynamics model, each particle $p \in [P]$ can be propagated by $s_{t'+1,p}^{(i)} = \tilde{\mathcal{P}}_{\theta_b}^{(i)}(s_{t',p}^{(i)}, where a_{t'}^{(i)})$ ($t' \in [t, t+w-2]$), and $\tilde{\mathcal{P}}_{\theta_b}^{(i)}, b \in [B]$ is a randomly selected dynamics model in the ensemble model. Therefore, the set of plausible state-action trajectories for each candidate action sequence $\mathbf{A}_{t:t+w-1}^{(q,i)} = (a_t^{(q,i)}, \ldots, a_{t+w-1}^{(q,i)})$ ($q \in [Q]$) consists of P parallel propagated states with the same action sequence $\mathbf{A}_{t:t+w-1}^{(q,i)}$. Afterward, given the calibrated known reward function $\mathcal{R}^{(i)}$, the evaluation of a candidate action sequence can be derived from the average cumulative reward of P different parallel action-state trajectories, that is,

$$ J(\mathbf{A}_{t:t+w-1}^{(q,i)}) = \mathbb{E}\left[\frac{1}{P} \sum_{p=1}^{P} \sum_{t'=t}^{t+w-1} \mathcal{R}^{(i)}\left(s_{p,t'}^{(i)}, a_{t'}^{(q,i)} \right) \right]. \quad (10) $$

After sorting Q candidate action sequences in terms of the evaluation $J(\mathbf{A}_{t:t+w-1}^{(i)})$, elite candidate sequences $X \leq Q$ can be selected to update the Gaussian-distributed sequential decision-making function $\mathcal{N}(\mu_\phi, \sigma_\phi)$. Such a TS-based CEM procedure

can repeat until convergence of ϕ. Therefore, it can yield the desired optimal action sequence satisfying (6) and the MPC controller executes only the first action $a_t^{*(i)}$, transitions the actual environment into the state $s_{t+1}^{(i)}$, and re-calculates the optimal action sequence at the next time-step.

Finally, we summarize our model-based MARL method `MA-PETS` in Algorithm 1.

V. REGRET BOUND OF `MA-PETS`

Next, we investigate its group regret bound of `MA-PETS`, by first introducing how to construct an optimistic MDP from a set of plausible MDPs through EVI. Based on that, we derive the group regret bound facilitated by communications within a range of d.

A. EVI for an Optimistic MDP

Beforehand, in order to better exemplify the sample efficiency, we make the following assumptions.

Assumption 1: The continuous state and action space (i.e., $\mathcal{S}^{(i)}$ and $\mathcal{A}^{(i)}$) can be quantized. Correspondingly, $S = |\mathcal{S}^{(i)}|$ and $A = |\mathcal{A}^{(i)}|$ denote the size of the discrete state space and action space, respectively.

Assumption 2: All agents in `MA-PETS` could enter into the next episode in a "simultaneous" manner, while the episodic length can be tailored to meet some pre-defined conditions.

Assumption 3: For simplicity of representation, we can scale the reward r from $[0, r_{\max}]$ to $[0,1]$.

Algorithm 1: The `MA-PETS` algorithms.

Input: communication range d, initial state $\mathcal{N}(\mu_{\phi_0}, \sigma_{\phi_0})$, rarity parameters α, max iteration of CEM Y, accuracy ϵ.

1:   Initialize vehicle' dataset $\mathcal{D}^{(i)}, \mathcal{D}^{(i)+} = \emptyset, \forall i \in [I]$;
2:   Randomly sampling action in the initial episode and update $\mathcal{D}^{(i)+}_{0:H}, \forall i \in [I]$;
3:   **for** Vehicle i and episode k, $\forall i \in [I], \forall k \in [K]$ **do**
4:     Train the virtual environment dynamics ensemble model $\tilde{\mathcal{P}}^i$ given $\mathcal{D}^{(i)+}$ for each vehicle i;
5:     **for** step $t = 1, 2, \dots H$ do **do**
6:       **while** $j \leq Y$ and $\sigma_{\phi_j} - \sigma_{\phi_{j-1}} \leq \epsilon$ **do**
7:         Each vehicle samples Q candidate action sequences $\mathbf{A}^{(1,i)}_{t:t+H-1}, \dots, \mathbf{A}^{(Q,i)}_{t:t+H-1} \sim_{iid} \mathcal{N}(\mu_{\phi_j}, \sigma_{\phi_j})$;
8:         Propagate state particles $s^{(q,i)}_{t,p}$ by using ensemble model and TS method;
9:         Evaluate action sequences by (10);
10:        Select the top $\alpha\%$ of action sequences with the high rewards;
11:        Update $\mathcal{N}(\mu_{\phi_{j+1}}, \sigma_{\phi_{j+1}})$ distribution by high reward action sequences;
12:       **end while**
13:       Only execute first action $a^{*(i)}_t, \forall i \in [I]$ from optimal actions $\mathbf{A}^{*(i)}_{t:t+w-1}$;
14:       Obtain reward $r^{(i)}_t$, and observe next-step state $s^{(i)}_{t+1}$;
15:     **end for**
16:     Determine neighboring vehicle i', $\forall \text{dis}(i, i') \leq d$ (i.e., within the communication range d of vehicle i) to exchange information;
17:     Update dataset $\mathcal{D}^{(i)+} \leftarrow \mathcal{D}^{(i)+} \cup \{\mathcal{D}^{(i')}_{t-H:t}\}$.
18:   **end for**

Notably, Assumption 1 holds naturally, if we neglect the possible discretization error of DNN in `MA-PETS`, as it has a trivial impact on understanding the sample efficiency incurred by information exchange. The quantization discussed here involves discretizing continuous state and action spaces. Common discretization methods [53] include uniform quantization, which divides each dimension of the continuous state space into intervals of equal width (bins). These intervals define the precision of discretization, with each interval representing a discrete state. Besides, non-uniform quantization may discretize based on the distribution or importance of states, and clustering algorithms (such as k-means or DBSCAN [54], [55]) can group points in the state space, with each cluster center representing a discrete state. Meanwhile, text encoding (e.g., Tile Coding [56]), which involves overlaying different tiles on the states and assigning one bit to each tile, belongs to an alternative solution. Therefore, if an agent is in a certain position, the corresponding discrete state can be represented by one bit-vector, with activated positions set to 1 and all other positions set to 0. On the other hand, Assumption 2 could be easily met by intentionally ignoring the experienced visits of some "diligent" agents. Moreover,

Assumption 3 implies unanimously scaling the group regret bound by $r_{\max}$, which does not affect learning the contributing impact from inter-agent communications.

Based on these assumptions, we incorporate the concept of a classical MBRL algorithm UCRL2 into the analyses of `MA-PETS`. In particular, UCRL2 primarily implements the "optimism in the face of uncertainty", and performs an EVI through episodes $k = 1, 2, \dots$, each of which consists of multiple time steps. The term $N^{(i)}_k(s, a)$ denotes the number of visits to the state-action pair (s, a) by agent i before episode k. Next, UCRL2 determines the optimal policy for an optimistic MDP, choosing from a collection of plausible MDPs $\mathcal{M}^i$ constructed based on the agents' estimates and their respective confidence intervals, as commonly governed by the Hoeffding inequality for an agent [57].

Consistent with UCRL2, we allow each agent i to enter into a new episode $k + 1$ once there exists a state-action pair (s, a) that has just been played and satisfies $u^{(i)}_k(s, a) = N^{(i)+}_k(s, a)$. Here $u^{(i)}_k(s, a)$ denotes the number of state-action (s, a) visits by agent i in the episode k and $N^{(i)+}_k(s, a) = \max\{1, N^{(i)}_k(s, a)\}$. By Assumption 2, all agents could enter into the next episode in a "simultaneous" manner, by intentionally ignoring the experienced visits of a "diligent" agent i after meeting $u^{(i)}_k(s, a) = N^{(i)+}_k(s, a)$ in episode k. Therefore, at the very beginning of episode $k + 1$, for each agent i, $N^{(i)}_{k+1}(s, a) = N^{(i)}_k(s, a) + u^{(i)}_k(s, a)$ for all state-action pairs (s, a). This setting is similar to the doubling criterion in single-agent UCRL2 and ensures that each episode is long enough to allow sufficient learning. Furthermore, different from the single-agent UCRL2, at the end of an episode k, consistent with in `MA-PETS`, each agent i has access to all the set of neighboring CAVs $\mathcal{F}^{(i)}_{d,k}$, so as to obtain the latest dataset $\mathcal{D}^{(j)}_{t_k-H:t_k}, \forall j \in \mathcal{F}^{(i)}_{d,k}$ and constitute $\mathcal{D}^{(i)+}_{0:t_k}$. Again, according to Assumption 2, $H = \sum_{s,a} u^{(i)}_k(s, a)$. Afterward, the EVI proceeds to estimate the transition probability function $\mathcal{P}^{(i)}(\cdot \mid s, a)$. In particular, various high-probability confidence intervals of the true MDP $M^{(i)}$ can be conjectured according to different concentration inequalities. For example, based on Hoeffding's inequality [57], the confidence interval for the estimated transition probabilities can be given as

$$\left\| \tilde{\mathcal{P}}^{(i)}(\cdot \mid s, a) - \hat{\mathcal{P}}^{(i)}_k(\cdot \mid s, a) \right\|_1 \leq \sqrt{\frac{14 S \log(2A t_k / \delta)}{N^+_k(s, a)}}, \tag{11}$$

and it bounds the gap between the EVI-conjectured (estimated) transition matrix $\tilde{\mathcal{P}}^{(i)}(\cdot \mid s, a)$ derived from the computed policy $\tilde{\pi}^{(i)}_k$ and the empirical average one $\hat{\mathcal{P}}^{(i)}_k(\cdot \mid s, a)$ in episode k. Besides, $\delta \in [0, 1]$ is a pre-defined constant.

Furthermore, a set $\mathcal{M}^{(i)}_k$ of plausible MDPs (12) is computed for each agent i as

$$\mathcal{M}^{(i)}_k \overset{\text{def}}{=} \left\{ \tilde{M}^{(i)} = \langle \mathcal{S}^{(i)}, \mathcal{A}^{(i)}, \mathcal{R}^{(i)}, \tilde{\mathcal{P}}^{(i)} \rangle \right\}. \tag{12}$$

Correspondingly, a policy for an optimistic MDP can be attained from learning in the conjectured plausible MDPs, and then the

impact of the communication range d on the performance of `MA-PETS` can be analyzed, in terms of the multi-agent group regret in (5).

B. Analyses of Group Regret Bound

The exchange of information in `MA-PETS` enhances the sample efficiency, which will be further shown in the derived multi-agent group regret bound. In other words, the difference between single-agent and multi-agent group regret bounds manifests the usefulness of information sharing among agents.

1) Results for Single-Agent Regret Bound: Before delving into the detailed analyses of multi-agent group regret bound, we introduce the following useful pertinent results derived from single-agent regret bound [41].

Consider $M^{(i)}$ as a time-homogeneous and communicating MDP with a policy $\pi'^{*(i)}$. Meanwhile, by Lemma 1, the optimal T-step reward $\mathcal{R}_T^{*(i)}(M^{(i)})$ can be approximated by the optimistic average reward $\rho^{*(i)}$. Therefore, if $\pi'^*(i)$ is implemented in $M^{(i)}$ for T steps, and we define the regret in a single episode k as $\Delta_k^{(i)} := \sum_{(s,a)} u_k^{(i)}(s,a)(T\rho^{*(i)} - \bar{r}^{(i)}(s,a))$, which captures the difference between the T-step optimal reward $\rho^{*(i)}$ and the mean reward $\bar{r}^{(i)}(s,a)$, with $\sum_k u_k^{(i)}(s,a) = N_{k+1}^{(i)}(s,a)$ and $\sum_{(s,a)} N_k^{(i)}(s,a) = T$, we have the following lemma.

Lemma 2 (Equations (7)– (8) of [41]): Under Lemma 1, the differences between the observed rewards $r_t^{(i)}$, acquired when agent i follows the policy $\pi'^{*(i)}$ to choose action $a_t^{(i)}$ in state $s_t^{(i)}$ at step t, and the aforementioned optimistic average reward $\rho^{*(i)}$ forms a martingale difference sequence. Besides, based on Hoeffding's inequality, with a probability at least $1 - \frac{\delta}{12(T)^{5/4}}$, the single-agent bound can be bounded as

$$\text{Regret}^{(i)}(T) = T\mathcal{R}_T^{*(i)}(M^{(i)}) - \sum_{t=1}^{T} r_t^{(i)}(s_t, a_t)$$
$$\leq \sum_k \Delta_k^{(i)} + \sqrt{C_1 \, T \log(8\,T/\delta)}, \quad (13)$$

where $C_1 = \frac{5}{8}$ is a constant.

Lemma 2 effectively transforms the cumulative T-step regrets into individual episodes of regret. Moreover, the regret bound $\sum_k \Delta_k^{(i)}$ in Lemma 2 can be classified into two categories, depending on whether the true MDP $M^{(i)}$ falls into the scope of plausible MDPs $\mathcal{M}_k^{(i)}$. As for the case where the true MDP $M^{(i)}$ is not included in the set of plausible MDPs $\mathcal{M}_k^{(i)}$, the corresponding regret $\Delta_{k,M^{(i)} \notin \mathcal{M}_k^{(i)}}^{(i)}$ can be bounded by the following lemma.

Lemma 3 (Regret with Failing Confidence Intervals, Equation (13) of [42]): As detailed in Sec. 5.1 of [42], at each step t, the probability of the true MDP not being encompassed within the set of plausible MDPs is given by $P\{M^{(i)} \notin \mathcal{M}_k^{(i)}\} \leq \frac{\delta}{15t^6}$. Correspondingly, the regret attributable to the failure of confidence intervals is

$$\sum_k \Delta_{k,M^{(i)} \notin \mathcal{M}_k^{(i)}}^{(i)} \leq \sqrt{T}. \quad (14)$$

On the other hand, under the assumption that $M^{(i)}$ falls within the set $\mathcal{M}_k^{(i)}$, in light of (8) from [58], we ascertain that due to the approximation error of the EVI,

$$\tilde{\rho}_k^{(i)} \geq \rho^{*(i)} - \frac{1}{\sqrt{t_k}}, \quad (15)$$

where $\tilde{\rho}_k^{(i)}$ represents the optimistic average reward obtained by the optimistically chosen policy $\tilde{\pi}_k^{(i)}$, while $\rho^{*(i)}$ signifies the actual optimal average reward.

Define $\Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{r(i)} := \sum_{s,a} u_k^{(i)}(s,a)(\tilde{r}_k^{(i)}(s,a) - \bar{r}^{(i)}(s,a))$ as the accumulated regret in reward estimation over all state-action pairs. $\tilde{r}_k^{(i)}(s,a)$ denotes the maximal possible reward according to a confidence interval similar to (11). Meanwhile, $\Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{p(i)} := (\boldsymbol{u}_k^{(i)})^\top \left((\tilde{\mathcal{P}}_k^{(i)}(\cdot \mid s, \tilde{\pi}_k(s)))_s - \mathcal{I} \right) \boldsymbol{w}_k^{(i)}$ quantifies the regret from the estimation of transition probabilities. Additionally, $\boldsymbol{u}_k^{(i)} := (u_k^{(i)}(s, \tilde{\pi}_k(s)))_s$ denotes the vector representing the final count of state visits by agent i in episode k, as determined under the policy $\tilde{\pi}_k(s)$, where the operator $(\cdot)_s$ concatenates the values for all $s \in \mathcal{S}$. $(\tilde{\mathcal{P}}_k^{(i)}(\cdot \mid s, \tilde{\pi}_k(s)))_s$ is the transition matrix of the policy $\tilde{\pi}_k^{(i)}$ in the optimistic MDP $\tilde{M}_k^{(i)}$ computed via EVI and $\mathcal{I}$ typically represents the identity matrix. Furthermore, $\boldsymbol{w}_k^{(i)}$ denotes a modified value vector to indicate the state value range of EVI-iterated episode k specific to agent i, where $w_k^{(i)}(s) := \eta_k^{(i)}(s) - \frac{1}{2}(\min_{s \in \mathcal{S}^{(i)}} \eta_k^{(i)}(s) + \max_{s \in \mathcal{S}^{(i)}} \eta_k^{(i)}(s))$ and $\eta_k^{(i)}(s)$ is the state value given by EVI. We can then decompose the aggregate regret considering $M^{(i)} \in \mathcal{M}_k^{(i)}$, which leads to the result in Lemma 4.

Lemma 4 (Regret with $M^{(i)} \in \mathcal{M}_k^{(i)}$, Sec. 5.2 of [58]): By the assumption $M^{(i)} \in \mathcal{M}_k^{(i)}$ and the gap derived from (15), the regret $\Delta_k^{(i)}$ of agent i accumulated in episode k could be upper bounded by

$$\Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{(i)} \leq \underbrace{\Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{r(i)}}_{(a)} + \underbrace{\Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{p(i)}}_{(b)}$$
$$+ 2\sum_{s,a} \frac{u_{k(M(i) \in \mathcal{M}_k^{(i)})}^{(i)}(s,a)}{\sqrt{t_k}}, \quad (16)$$

As previously noted in Section III-B, we posit that the agents' reward functions $\mathcal{R}^i$ are predetermined, often designed based on task objectives. In line with `MA-PETS`, the regret contributed by the term (a) in (4) can be disregarded. Therefore, we have the following corollary.

Corollary 1: By the assumption $M^{(i)} \in \mathcal{M}_k^{(i)}$ and the gap derived from (15), the regret $\Delta_k^{(i)}$ of agent i accumulated in episode k could be upper bounded by

$$\Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{(i)} \leq \underbrace{\Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{p(i)}}_{(b)} + 2\sum_{s,a} \frac{u_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{(i)}(s,a)}{\sqrt{t_k}}.$$

$$(17)$$

On the other hand, the item (b) (i.e., $\Delta^{(i)}_{k,M^{(i)}\in\mathcal{M}^{(i)}_k}$) in Lemma 4 (as well as Corollary 1) can be further decomposed and bounded as (18) shown at the bottom of this page, where, as demonstrated in (23) of [58], the modified value vector $\boldsymbol{w}^{(i)}_k$ satisfies $\|\boldsymbol{w}^{(i)}_k\|_\infty \leq \frac{D(M^{(i)})}{2}$. In other words, the state value range is bounded by the diameter D of the MDP at any EVI iteration. By slight rearrangement of $\Delta^{(i)}_{k,M^{(i)}\in\mathcal{M}^{(i)}_k}$, we have (18), which can be bounded by the following lemma.

Lemma 5 (Equations (16)– (17) of [41]): Contingent on the confidence interval (11), based on our assumption that both $\tilde{M}^{(i)}_k$ and $M^{(i)}$ belong to the set of plausible MDPs $\mathcal{M}^{(i)}_k$, the term $\Delta^{p_1(i)}_k$ in (18) can be bounded as

$$\Delta^{p_1(i)}_k \leq D\sqrt{C_2 S \log(2AT/\delta)} \sum_k \sum_{s,a} \frac{u^{(i)}_k(s,a)}{\sqrt{N^+_{t_k}(s,a)}}, \quad (19)$$

where C_1 denotes a constant. Moreover, by applying the Azuma-Hoeffding inequality, the term $\Delta^{p_2(i)}_k$ in (18) can be bounded as

$$\Delta^{p_2(i)}_k \leq D\sqrt{4C_1 T \log(8T/\delta)} + DSA\log_2(8T/SA), \quad (20)$$

where C_2 denotes a constant.

2) Analysis for Group Regret With Limited-Range Communications: Before analyzing the group regret in the multi-agent system within a constrained communication range, we start with some necessary notations and assume all CAVs constitute a graph. Considering the power graph $\mathcal{G}_{d,k}$, which is derived by selecting nodes from the original graph and connecting all node pairs in the new graph that are at a distance less than or equal to a specified value d in the original graph. The neighborhood graph $\mathcal{G}^{(i)}_{d,k} = (\mathcal{F}^{(i)}_{d,k}, \mathcal{E}^{(i)}_{d,k})$ represents a sub-graph of $\mathcal{G}_{d,k}$, where $\mathcal{E}^{(i)}_{d,k}$ comprises the set of communication links connecting the agents i and $i' \in \mathcal{F}^{(i)}_{d,k}$. We also define the total number of state-action observations for agent i, which is calculated as the sum of observations from its neighbors within the communication range $d \in [0, D(\mathcal{G}_k)]$ before episode k, denoted as $N^{(i)+}_{d,k}(s,a) := N^{(i)+}_{d,k-1}(s,a) + \sum_{i'\in\mathcal{G}^{(i)}_{d,k}} u^{(i')+}_k(s,a)$. Given the complication to directly estimate $N^{(i)+}_{d,k}(s,a)$, we introduce the concept of clique cover [27] for $\mathcal{F}^{(i)}_{d,k}$, which is a collection of cliques that can cover all vertices of a power graph. Further,

let $\mathbf{C}_{d,k}$ denote a clique cover of $\mathcal{G}_{d,k}$ and a clique $\mathcal{C} \in \mathbf{C}_{d,k}$. Meanwhile, $|\mathcal{C}|$ represents the size of the clique $\mathcal{C}$. Additionally, the clique covering number $\bar{\chi}(\mathcal{G}_{d,k})$ signifies the minimum number of cliques to cover the power graph $\mathcal{G}_{d,k}$ within the communication. Furthermore, characterized by a clique covering number $\bar{\chi}(\mathcal{G}_d) := \max_{k\in[K]}\{\bar{\chi}(\mathcal{G}_{d,k})\}$, $\mathbf{C}_d$ constitutes the clique cover for the graph $\mathcal{G}_d$.

Lemma 6 ((6) of [45]): Let $\mathbf{C}_{d,k}$ represent the clique covering of the graph $\mathcal{G}_{d,k}$, where the graph $\mathcal{G}_{d,k}$ consists of I nodes, and the cliques within $\mathbf{C}_{d,k}$ have node counts $|\mathcal{C}_{\mathcal{C}\in\mathbf{C}_{d,k}}|$. The minimum clique cover $\bar{\chi}(\mathcal{G}_{d,k})$ maintains a consistent relationship $\sum_{\mathcal{C}\in\mathbf{C}_{d,k}} \sqrt{|\mathcal{C}|} = \sqrt{\bar{\chi}(\mathcal{G}_d)I}$.

In this setup, all agents within a clique $\mathcal{C}$ explore in proportion to the clique's size and share the collected samples among themselves. For simplicity, we assume that $\mathcal{G}_{d,k}$ is connected, meaning that all cliques can communicate with one another. Meanwhile, we take $N^+_{\mathcal{C},k}(s,a)$ to be the number of samples exchanged within the clique $\mathcal{C}$ before an episode k that are available for all agents $i \in \mathcal{C}$. In the most challenging scenario, characterized by uniformly random exploration, the algorithm fails to capitalize on any inherent structures or patterns within the environment to potentially refine its learning approach. Consequently, this necessitates a maximum quantity of samples to ascertain an optimal policy, thus giving the worst case. Under these conditions, we present the following lemma.

Lemma 7 (Theorem 1 of [45]): In the worst case, where the exploration is uniformly random, the number of samples within each clique can be approximated as $N_{\mathcal{C},k}(s,a) = |\mathcal{C}|Hk/SA$, where H is the maximum horizon, K is the number of episodes, S and A represents the state and action spaces, respectively.

Based on the results of the single-agent regret bound described in Section V-B-1, it is ready to analyze the upper bound on the group regret of the multi-agent setting under limited communication range, as shown in Theorem 1.

Theorem 1: (Hoeffding Regret Bound for Parallel MDP with Limited-Range Communications) With probability at least $1 - \delta$, it holds that for all initial state distributions and after any T steps, the group regret with limited communication range is upper bounded by (21) shown at the bottom of the next page.

Proof: The proof of Theorem 1 stands consistently with that of Lemma 2. However, it contains significant differences, due to the information exchange-induced distinctive empirical sample size for the transition probabilities.

In particular, the T-step group regret in the MDP settings could be bounded based on (5) and (13), that is, by Lemma 2

$$\sum_k \Delta^{p(i)}_{k,M^{(i)}\in\mathcal{M}^{(i)}_k} \leq \underbrace{\sum_k \sum_s \left\| u^{(i)}_k(s,\tilde{\pi}_k(s)) \left(\left(\tilde{\mathcal{P}}^{(i)}_k(\cdot \mid s, \tilde{\pi}_k(s)) \right)_s - \left(\mathcal{P}^{(i)}_k(\cdot \mid s, \pi_k(s)) \right)_s \right) \right\|_1 \left\| \boldsymbol{w}^{(i)}_k \right\|_\infty}_{\Delta^{p_1(i)}_k}$$

$$+ \underbrace{\sum_k \sum_s \sum_{s'} u^{(i)}_k(s,\tilde{\pi}_k(s)) \left(\mathcal{P}^{(i)}_k(s' \mid s, \pi_k(s)) - \mathcal{I}_{s,s'} \right) w^{(i)}_k(s')}_{\Delta^{p_2(i)}_k} \quad (18)$$

with a high probability at least $1 - \frac{\delta}{12(IT)^{5/4}}$,

$$\text{Regret}_G(T) = \sum_i \left[T\rho^{*(i)} - \sum_{t=1}^{T} r_t^{(i)}(s_t, a_t) \right]$$

$$\leq \sum_i^I \sum_k^K \left[\Delta_{k,M^{(i)} \notin \mathcal{M}_k^{(i)}}^{(i)} + \Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{(i)} \right]$$

$$+ \sqrt{C_1 IT \log(8IT/\delta)} \tag{22}$$

As elucidated in Lemma 3, the portion of group regret $\sum_i^I \sum_k \Delta_{k,M^{(i)} \notin \mathcal{M}_k^{(i)}}^{(i)}$ resulting from failed confidence regions can be bounded by $I\sqrt{T}$ with a probability of at least $\frac{\delta}{15t^6}$.

On the other hand, for each episode k, under the premise that the true MDP is encompassed within the set of plausible MDPs ($\mathcal{M}_k^i, \forall k$), the term $\Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{(i)}$ is further dissected according to Corollary 1. With Lemma 5, we have

$$\sum_i^I \sum_k^K \Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{(i)}$$

$$\leq \sum_i^I \sum_{k=1}^K \Delta_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{p(i)} + 2 \sum_i^I \sum_k^K \sum_{s,a} \frac{u_k^{(i)}(s,a)}{\sqrt{N_k^{(i)+}(s,a)}}$$

$$\leq D\sqrt{C_2 S \log(2AT/\delta)} \sum_{k=1}^K \sum_{\mathcal{C} \in \mathbf{C}_{d,k}} \sum_{i \in \mathcal{C}} \sum_{s,a} \frac{u_k^{(i)}(s,a)}{\sqrt{N_{\mathcal{C},k}^+(s,a)}}$$

$$+ DSAI \log_2(8T/SA) + D\sqrt{4C_1 TI \log(8TI/\delta)}$$

$$+ 2 \sum_i^I \sum_{k=1}^K \sum_{s,a} \frac{u_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{(i)}(s,a)}{\sqrt{N_k^{(i)+}(s,a)}}. \tag{23}$$

Leveraging the definition that $N_{k+1}^{(i)+}(s,a) = \max\{1, \sum_{i=1}^k u_k^{(i)}(s,a)\}$, and applying Jensen's inequality and the inequality in Lemma 8 in Appendix, we obtain the following results

$$\sum_i^I \sum_{k=1}^K \sum_{s,a} \frac{u_{k,M^{(i)} \in \mathcal{M}_k^{(i)}}^{(i)}(s,a)}{\sqrt{N_k^{(i)+}(s,a)}} \leq (\sqrt{2}+1)I\sqrt{SAT}. \tag{24}$$

Furthermore, by Lemma 7, $\sum_{(s,a)} N_{\mathcal{C},k}^+(s,a) = |\mathcal{C}|Hk$. Thus, we have

$$\sum_{k=1}^K \sum_{\mathcal{C} \in \mathbf{C}_{d,k}} \sum_{s,a} \frac{\sum_{i \in \mathcal{C}} u_k^{(i)}(s,a)}{\sqrt{N_{\mathcal{C},k}^+(s,a)}}$$

$$\leq \sum_{k=1}^K \sum_{\mathcal{C} \in \mathbf{C}_{d,k}} \sum_{s,a} \frac{u_{\mathcal{C},k}(s,a)}{\sqrt{N_{\mathcal{C},k}^+(s,a)}}$$

$$\leq \sum_{\mathcal{C} \in \mathbf{C}_d} (\sqrt{2}+1)\sqrt{|\mathcal{C}|SAKH}$$

$$= (\sqrt{2}+1)\sqrt{\bar{\chi}(\mathcal{G}_d) ISAT}, \tag{25}$$

where $u_{\mathcal{C},k}(s,a) := \sum_{i \in \mathcal{C}} u_k^{(i)}(s,a)$ represents the frequency of visits within the clique $\mathcal{C}$ to a state-action pair after the communications at the end of episode k.

In summary, we conclude the proof of Theorem 1 and establish that the total regret is bounded by (21) with a probability of $1 - \frac{\delta}{4T^{5/4}}$. When considering all values of $T = 2, \ldots$, it is evident that this bound holds simultaneously for all $T \geq 2$ with a probability of at least $1 - \delta$. ∎

Theorem 1 unveils the group regret bound and demonstrates that concerning T, a sub-linear increase in the group regret bound is attained in (21). On the contrary, in the standard non-communicative reinforcement learning setting for the same type of algorithm, the group regret is essentially equal to the sum of the single-agent regrets, as implied in Lemma 2.

VI. EXPERIMENTAL SETTINGS AND NUMERICAL RESULTS

A. Experimental Settings

In this section, we evaluate the performance of MA-PETS within the domain of autonomous vehicle control and demonstrate the superiority of our proposed algorithm over several other state-of-the-art RL methods, including FIRL [33], SVMIX [13], MAPPO [8], MADDPG [10], DQN [59], SAC [60]. Specifically, we run our experiments using the CAV simulation platform SMARTS [28] and select the "Unprotected signal-free Intersection" scenario in the closed single-lane and multi-lane "Figure Eight" loop, which is a typical mixed autonomy traffic scenario as illustrated in Fig. 3, for the evaluation. In these experiments, we deploy a fixed number (i.e., I) CAVs respectively, managed by our MA-PETS algorithm, and a random number I_{hv} of human-driven vehicles (HVs) controlled by the environment within the SMARTS framework. All CAVs will start from a one-way lane with an intersection, and drive circularly by passing through the intersection while avoiding collisions and congestion. Following the MDP defined in Section III, the corresponding MDP for CAVs in "Unprotected Intersection" is defined as below.

- *State and Observation:* Except for information about the vehicle itself, each CAV can only observe the information of the vehicle ahead and behind. Hence, it has the information about current state containing the velocity $v_t \in \mathbb{R}$, position $z_t = (x_t, y_t) \in \mathbb{R}^2$ of itself, the speed and distance

$$\text{Regret}_G(T) \leq \sqrt{C_1 IT \log(8IT/\delta)} + I\sqrt{T}\left[1 + (1+\sqrt{2})\sqrt{SA}\right] + D\sqrt{4C_1 IT \log(8IT/\delta)}$$

$$+ DSAI \log_2(8T/SA) + (1+\sqrt{2})DS\sqrt{C_2 \bar{\chi}(\mathcal{G}_d) IAT \log(2AT/\delta)} \tag{21}$$
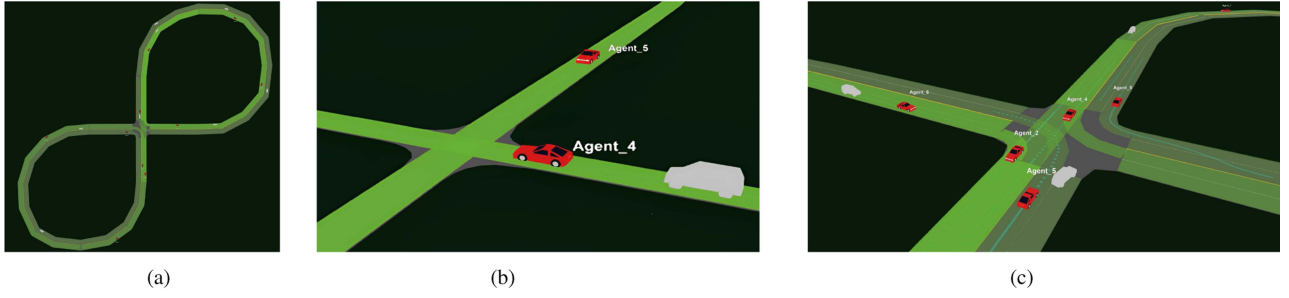
Fig. 3.   The "Unprotected Intersection" scenario in the closed single-lane and multi-lane "Figure Eight" loop for simulations. (a) presents an aerial view of the "Figure Eight" loop, while (b) and (c) provide the regional enlarged view of the single-lane and multi-lane "Unprotected Intersection", respectively. (a) "Figure Eight" loop. (b) Single-lane "Unprotected Intersection". (c) Multi-lane "Unprotected Intersection".

of the vehicles ahead and behind $v_{t,a}, v_{t,e}, l_{t,a}, l_{t,e} \in \mathbb{R}$. Hence, as mentioned earlier, the state of vehicle i can be represented as $s_t^{(i)} = (v_t^{(i)}, x_t^{(i)}, y_t^{(i)}, v_{t,a}^{(i)}, v_{t,e}^{(i)}, l_{t,a}^{(i)}, l_{t,e}^{(i)})$.

- *Action:* As each CAV is controlled via the target velocity $v \in \mathbb{R}$ and the special actions indicating whether to make a lane change $c \in \{0, 1\}$, decided by itself, the action of vehicle i can be represented as $a_t^{(i)} = \{v_t^{(i)}, c_t^{(i)}\}$. Besides, for the single-lane scenario, c nulls.

- *Reward:* CAVs aim to maintain a maximum velocity based on no collision. Accordingly, the reward function $\mathcal{R}^{(i)}$ can be defined as

$$r_t^{(i)} = \mathcal{R}^{(i)}\left(s_t^{(i)}, a_t^{(i)}\right) = v_t^{(i)} + v_{t,a}^{(i)} + v_{t,e}^{(i)} + \beta, \quad (26)$$

where the extra term β imposes a penalty term on collision, that is, $\beta = -10$ if a collision occurs; and it nulls otherwise.

Notably, given the definition of the reward function in (26), a direct summation of the rewards corresponding to all agents could lead to the computation duplication of some vehicles, thus misleading the evaluation. Therefore, we develop two evaluation metrics from the perspective of system agility and safety. In particular, taking account of the cumulative travel distance $\Lambda x_t^{(i)}$ of vehicle i at time-step t and the lasting-time t_c before the collision, we define

$$\text{Agility}(K) = \mathbb{E}_{\boldsymbol{\pi} \sim \tilde{\mathcal{P}}_k} \frac{1}{I} \sum_{i=1}^{I} \frac{1}{t_c^{(i)}} \sum_{t=0}^{t_c^{(i)}} \Lambda x_t^{(i)} \quad (27)$$

and

$$\text{Safety}(K) = \mathbb{E}_{\boldsymbol{\pi} \sim \tilde{\mathcal{P}}_k} \frac{1}{I} \sum_{i=1}^{I} \frac{t_c^{(i)}}{H}, \quad (28)$$

where $\tilde{\mathcal{P}}_k$ ($\forall k \in [K]$) refers to a learned dynamics model until the episode k.

In our configuration, the number of episodes is $K = 15$, and in the event of no collisions, the length of each episode can reach up to $H = 200$ for single-lane scenarios and $H = 400$ for multi-lane scenarios, respectively. All of our experiments are conducted on the NVIDIA GeForce RTX 4090 with 5 independent simulations. Moreover, typical parameters, including the simulation environment setup, deep neural network, and `MA-PETS`, are summarized in Table II.

TABLE II
LIST OF KEY PARAMETER SETTINGS FOR THE SIMULATION

Parameters Description	Value
Simulator	
Discrete-time step	$\tau = 0.1$ s
Vehicle numbers	$I = 8$ or 22
Human-driving Vehicles	$I_{\text{hv}} \in [8, 20]$
Distance of "Figure Eight" scenario	480 m
Maximum velocity of "Figure Eight" scenario	$v_{\max} = 13.89$ m/s
Minimum velocity of "Figure Eight" scenario	$v_{\min} = 0$ m/s
MA-PETS	
Speed per vehicle at time-step t	$v_t \in [0, 13.89]$ m/s
Position per vehicle at t	$z_t \in \mathbb{R}^2$
Target velocity per vehicle at t	$\bar{v}_t \in [0, 13.89]$ m
Speed of vehicles ahead and behind	$v_{t,a}, v_{t,e} \in [0, 13.89]$ m/s
Distance of vehicles ahead and behind	$l_{t,a}, l_{t,e} \in [0, 75]$ m
Travel distance per vehicle at t	$\Lambda x_t^{(i)} \in \mathbb{R}$
No. of episodes	$K = 15$
Maximal length per episode	$H = 200$ or 400
Communication range	$d \in [0, 200]$ m
Horizon of MPC	$w = 25$
Candidate action sequences in CEM	$Q = 400$
Elite candidate action sequences in CEM	$X = 40$
No. of particles	$P = 20$
Max iteration of CEM	$Y = 5$
Proportion of elite candidate action sequences	$\alpha = 10$
Accuracy of CEM Optimizer	$\epsilon = 0.001$
Seed number	5
Probabilistic Ensemble Neural Network	
No. of Ensembles for dynamics Model	$B = 5$
Hidden layer number	3
Hidden units number	300 or 400
Buffer length	$2,048$
Batch size	128
Epochs	5
Learning rate	$lr = 0.001$
Optimizer	Adam

∗ The parameters setting used only by `MA-PETS`.

B. Numerical Results

Firstly, we evaluate the performance of `MA-PETS` for CAVs with $I = 8$ and communication range $d = 100$ in the single-lane scenario, and present the performance comparison with MFRL algorithms in Fig. 4. Besides, we testify the real-time performance for each step of the 10-th episode in the simulation process in Fig. 5. It can be observed from Figs. 4 and 5 that our
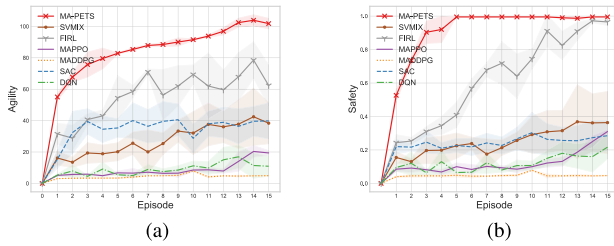
Fig. 4. Comparison of utility in the single-lane "Unprotected Intersection" scenario. (a) Agility. (b) Safety.
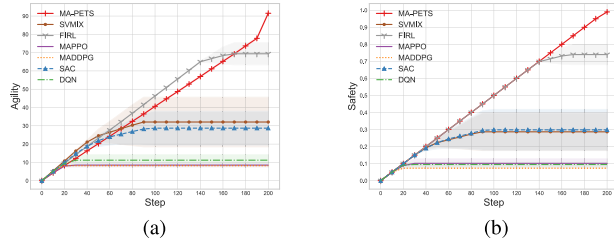

Fig. 5. Real-time utility comparison for each step of the 10-th episode under the single-lane "Unprotected Intersection" scenario. (a) Agility. (b) Safety.
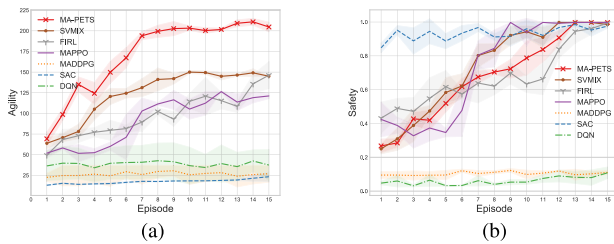

Fig. 6. Comparison of utility in the multi-lane "Unprotected Intersection" scenario. (a) Agility. (b) Safety.

Fig. 7. Performance comparison of the MA-PETS with continuousness and discretization at the single-lane "Unprotected Intersection" scenario.

Fig. 8. Performance comparison under different communication range d. (a) Agility. (b) Safety.

TABLE III
THE COMMUNICATION OVERHEADS OF DIFFERENT COMMUNICATION RANGE d

Name	Value				
Communication Distance d	0	50	100	150	200
Communication Overheads	0	330.13	1,866	6,139.7	10,961.06

algorithm MA-PETS significantly yields superior outcomes than the others in terms of both agility and safety. More importantly, MA-PETS converges at a faster pace.

Building upon these single-lane results, we further increase the complexity of the simulation scenarios by establishing a multi-lane intersection without traffic lights and utilizing $I = 22$ CAVs controlled by our algorithm MA-PETS, as depicted in Fig. 6. In this setting, we perform a comparative evaluation of our algorithm against a suite of state-of-the-art RL methods with a communication range $d = 100$. As can be observed from Fig. 6, while our algorithm may initially exhibit a marginally reduced efficacy in terms of collision avoidance relative to MAPPO, SVMIX, and SAC, it nearly achieves a collision-free state by the 13-th episode. Simultaneously, our algorithm MA-PETS notably excels in agility and achieves convergence at a significantly accelerated pace compared to other algorithms.

To validate the rationality of Assumption 1 and the feasibility of MA-PETS after quantization, we conduct supplementary simulation experiments. We compare the performance of MA-PETS using the tile coding method [53] to discretize our six-dimensional continuous state space and two-dimensional
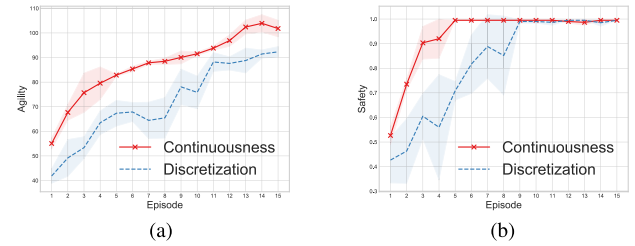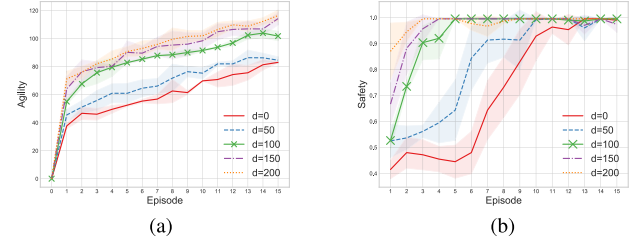
continuous action space with MA-PETS operating in continuous state and action spaces within a single-lane "Figure Eight" loop scenario, depicted in Fig. 7. It can be observed that although the discretized MA-PETS algorithm after quantization exhibits larger fluctuations during the learning process, with the increase in the number of training episodes (e.g., post-15 episodes), the MA-PETS with discretization can also achieve performance similar to the continuous MA-PETS in terms of agility and safety. This demonstrates that MA-PETS remains effective even in a discretized setting after quantization.

Furthermore, we show the performance of MA-PETS concerning different values of communication range d by varying from 0 to 200 in Fig. 8 and Table III. It can be observed from Table III that consistent with our intuition, the increase in communication range leads to a significant boost of the communications overheads. Meanwhile, as depicted in Fig. 8, it also benefits the learning efficiency of the CAVs, thus greatly upgrading agility and safety. On the other hand, Fig. 8 also unveils that when the communication range increases to a certain extent, a further increase of d contributes trivially to the agility and safety of learned policies. In contrast, Table III indicates an exponential increase in the average communications overhead. In other words, it implies a certain trade-off between the learning performance and communication overheads. Consistent with the group regret bound detailed in Section V, we further investigate
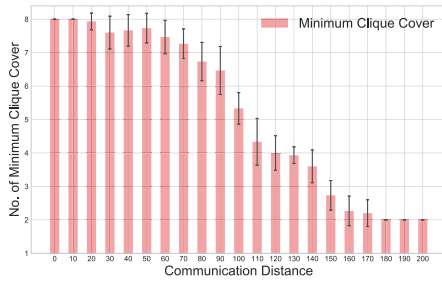
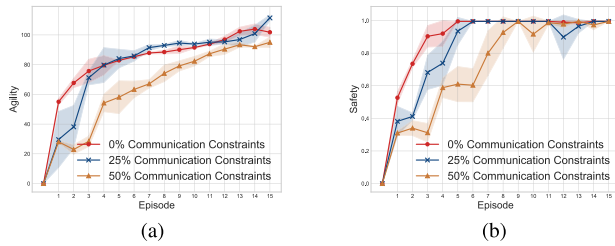Fig. 9. Correlation between number of minimum clique cover $\bar{\chi}(\mathcal{G}_{d,k})$ and communication distance d.

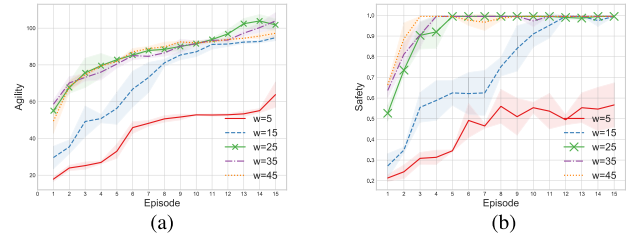Fig. 10. Comparison of utility for different communication constraints. (a) Agility. (b) Safety.

Fig. 11. Performance comparison under different horizon w of MPC. (a) Agility. (b) Safety.

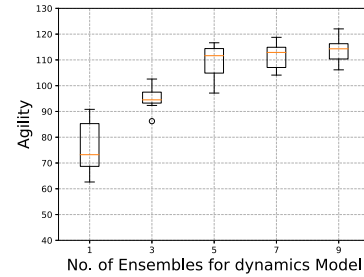
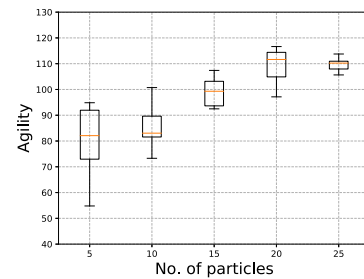Fig. 12. Performance comparison under different ensembles for dynamics model.

Fig. 13. Performance comparison under different particles for TS.

the numerical interplay between the minimum number of clique covers $\bar{\chi}(\mathcal{G}_{d,k})$ and communication distance d in 15 independent trials. It can be observed from Fig. 9 that while the minimum number of clique covers remains constant beyond a specific communication range threshold, a generally inverse correlation exists between the minimum clique cover number and the communication radius, which aligns with both our theoretical proofs and the experimental findings presented in Fig. 8.

To further illustrate the feasibility and robustness of MA-PETS under more complex and variable real-world communication scenarios, such as vehicles experiencing communication hindrances while passing through tunnels or other challenging environments, we conduct simulations under dynamic and unpredictable communication constraints. In our simulations, we assume the possibilities of each CAV controlled by the MA-PETS to potentially encounter communication blockages or failures as 0%, 25%, and 50% when attempting to communicate with other vehicles within a communication range of $d = 100$ at the end of each episode. The results of these simulations are presented in Fig. 10. From Fig. 10, it is evident that dynamic and unpredictable communication constraints at varying probabilities can cause fluctuations in the MA-PETS's safety and agility to some extent. However, the algorithm can effectively learn within 15 episodes, demonstrating its feasibility and robustness in real-world autonomous vehicle control environments.

To clarify the effect of the MPC horizon w in MA-PETS, we perform supplementary experiments. As depicted in Fig. 11, from a security point of view, extending the MPC horizon w gradually enhances security. However, beyond a horizon length of $w = 25$, further increases yield only marginal security improvements. In terms of agility, the choice of the MPC horizon w is crucial for the performance of the system. In particular, a "too

short" horizon introduces a substantial bias in the performance of MA-PETS, as it hampers the ability to make accurate long-term predictions due to the scarcity of time steps. Conversely, a "too long" horizon also results in a noticeable bias. This is due to the increased divergence of particles over extended periods, which reduces the correlation between the currently chosen action and the long-term expected reward. This bias escalates as the horizon w increases. Hence choosing an excessively long horizon adversely affects performance.

Our research also investigates the influence of the number of ensembles B and the number of particles P in MA-PETS. In terms of agility, Figs. 12 and 13 show the corresponding results respectively, which are derived from 15 training episodes through 15 independent simulation runs. It can be observed from Fig. 12 that along with the increase in B, the learning becomes more regularized and the performance improves. However, the performance improvement is no longer apparent for a sufficiently large B, and this improvement is more pronounced in more challenging and complex environments that require the learning of intricate dynamical models, leaving more scope for effective exploitation of the strategy without the use of model

TABLE IV
THE COMPARISON OF MEANS AND STANDARD DEVIATIONS OF THE AVERAGE DECISION-MAKING TIME PER STEP IN TESTING MARL
METHODS IN "FIGURE EIGHT"

Scenarios	The Different MARL Methods			
	MA-PETS	FIRL	MAPPO	SVMIX
Single-lane loop	8.851 ± 0.3105	7.604 ± 0.3034	8.007 ± 0.5486	10.28 ± 0.2668
Multi-lane loop	9.726 ± 0.2615	8.104 ± 0.4519	8.779 ± 0.4607	11.06 ± 0.3025

Unit: 1×10^{-2} seconds.

integration. In Fig. 13, superior performance can be reaped for a larger number of particles, because more particles allow for a more accurate estimation of the reward for state-action trajectories influenced by transition probabilities.

To conduct a more in-depth numerical analysis of the complexity and computational time of MA-PETS, we compare the average decision-making time and its standard deviation of different MARL methods after learning 15 episodes, in both single-lane loop and multi-lane loop scenarios as shown in Table IV. The results reveal that the average decision time of MA-PETS is slightly higher than that of FIRL and MAPPO in both scenarios, but faster than SVMIX. Overall, compared to FIRL and MAPPO, MA-PETS has smaller variance and lower variability in decision times, demonstrating greater stability.

VII. CONCLUSION AND DISCUSSION

In this paper, we have studied a decentralized MBRL-based control solution for CAVs with a limited communication range. In particular, we have proposed the MA-PETS algorithm with a significant performance improvement in terms of sample efficiency. Specifically, MA-PETS learns the environmental dynamics model from samples communicated between neighboring CAVs via PE-DNNs. Subsequently, MA-PETS efficiently develops TS-based MPC for decision-making. Afterward, we have derived UCRL2-based group regret bounds, which theoretically manifests that in the worst case, limited-range communications in multiple agents still benefit the learning. We have validated the superiority of MA-PETS over classical MFRL algorithms and demonstrated the contribution of communication to multi-agent MBRL.

There are many interesting directions to be explored in the future. For example, MA-PETS encounters significant challenges in more realistic scenarios. As environmental complexity escalates, scenarios with numerous variables or agents heighten the learning and computational requirements of MA-PETS. Moreover, if the dynamics of the environment are unpredictable, MA-PETS may face difficulties due to the probabilistic nature of its ensemble predictions. Another critical factor is the availability of computational resources. The computational power at hand can significantly influence the performance of MA-PETS, especially as more intricate environments or accurate ensemble methods demand robust processing capabilities. The key next steps involve optimizing the algorithm, reducing its reliance on computational resources, minimizing its decision delays, and improving its adaptability to rapidly changing and complex environments. Furthermore, MA-PETS confronts significant OOD challenges resulting from scarce training data, which can lead to learning instability and substantial overhead. Consequently, designing more accurate models to mitigate OOD issues is a critical priority that requires urgent attention. Finally, we have only derived the worst case of group regret bound, and the results for more general cases can be explored.

Lemma 8 (Appendix C.3 of [41]): For any sequence of numbers $x_1, \ldots, x_n$ with $0 \le x_k \le X_{k-1} := \max\{1, \sum_{i=1}^{k-1} x_i\}$, we have

$$\sum_{k=1}^{n} \frac{x_k}{\sqrt{X_{k-1}}} \le (\sqrt{2} + 1)\sqrt{X_n}. \tag{29}$$

REFERENCES

[1] R. Wen, J. Huang, and Z. Zhao, "Multi-agent probabilistic ensembles with trajectory sampling for connected autonomous vehicles," in *Proc. IEEE Globecom (Intelligent6GArch Workshop)*, Kuala Lumpur, Malaysia, 2023, pp. 2025–2030.

[2] B. R. Kiran et al., "Deep reinforcement learning for autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 4909–4926, Jun. 2022.

[3] M. Hua et al., "Multi-agent reinforcement learning for connected and automated vehicles control: Recent advancements and future prospects," Dec. 2023, *arXiv:2312.11084*.

[4] X. Liang, X. Du, G. Wang, and Z. Han, "A deep reinforcement learning network for traffic light cycle control," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1243–1253, Sep. 2019.

[5] J. Wu, Z. Huang, W. Huang, and C. Lv, "Prioritized experience-based reinforcement learning with human guidance for autonomous driving," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 855–869, Jan. 2024, doi: 10.1109/TNNLS.2022.3177685.

[6] D. González, J. Pérez, V. Milanés, and F. Nashashibi, "A review of motion planning techniques for automated vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 4, pp. 1135–1145, Apr. 2016.

[7] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: A survey," *Artif. Intell. Rev.*, vol. 55, no. 2, pp. 895–943, Feb. 2022.

[8] Y. Guan, Y. Ren, S. E. Li, Q. Sun, L. Luo, and K. Li, "Centralized cooperation for connected and automated vehicles at intersections by proximal policy optimization," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12597–12608, Nov. 2020.

[9] J. Zhang, C. Chang, X. Zeng, and L. Li, "Multi-agent DRL-based lane change with right-of-way collaboration awareness," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 854–869, Jan. 2023.

[10] R. Lowe et al., "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, California, USA, 2017, pp. 6382–6393.

[11] J. N. Foerster et al., "Counterfactual multi-agent policy gradients," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, Louisiana, USA, 2018, pp. 2974–2982.

[12] T. Rashid, M. Samvelyan, C. S. d. Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 7234–7284, Jan. 2020.

[13] B. Xiao et al., "Stochastic graph neural network-based value decomposition for marl in internet of vehicles," *IEEE Trans. Veh. Technol.*, vol. 73, no. 2, pp. 1582–1596, Feb. 2024.

[14] Z. Huang, J. Wu, and C. Lv, "Efficient deep reinforcement learning with imitative expert priors for autonomous driving," *IEEE Trans. Neural Net. Learn. Syst.*, vol. 34, no. 10, pp. 7391–7403, Oct. 2023.

[15] J. Wu, Z. Huang, and C. Lv, "Uncertainty-aware model-based reinforcement learning: Methodology and application in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 8, no. 1, pp. 194–203, Jan. 2023.

[16] T. Pan, R. Guo, W. H. K. Lam, R. Zhong, W. Wang, and B. He, "Integrated optimal control strategies for freeway traffic mixed with connected automated vehicles: A model-based reinforcement learning approach," *Transp. Res. C-Emerg. Technol.*, vol. 123, Feb. 2021, Art. no. 102987.

[17] A. D. Kiureghian and O. Ditlevsen, "Aleatory or epistemic? Does it matter?," *Struct. Saf.*, vol. 31, no. 2, pp. 105–112, Mar. 2009.

[18] A. Nagabandi, G. Kahn, R. S. Fearing, and S. Levine, "Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning," in *Proc. IEEE Int. Conf. Robot. Autom*, Long Beach, CA, USA, 2018, pp. 7559–7566.

[19] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," in *Proc. Adv. Neural Inf. Proces. Syst. (NIPS)*, Montréal, Canada, 2018, pp. 4754–4765.

[20] A. S. Polydoros and L. Nalpantidis, "Survey of model-based reinforcement learning: Applications on robotics," *J. Intell. Robot. Syst.*, vol. 86, no. 2, pp. 153–173, May 2017.

[21] P. G. Sessa, M. Kamgarpour, and A. Krause, "Efficient model-based multi-agent reinforcement learning via optimistic equilibrium computation," in *Proc. 39th Int. Conf. Mach. Learn.*, Baltimore, Maryland, USA, 2022, pp. 19580–19597. [Online]. Available: https://proceedings.mlr.press/v162/sessa22a.html

[22] A. Tuynman and R. Ortner, "Transfer in reinforcement learning via regret bounds for learning agents," Feb. 2022, *arXiv: 2202.01182*.

[23] C. Jin et al., "Is Q-learning provably efficient?," in *Proc. Adv. Neural Inf. Proces. Syst.*, Montreal, QC, Canada, 2018, pp. 4863–4873.

[24] Z. Zhang, Y. Zhou, and X. Ji, "Almost optimal model-free reinforcement learningvia reference-advantage decomposition," in *Proc. Adv. Neural Inf. Proces. Syst.*, Vancouver, BC, Canada, 2020, pp. 15198–15207.

[25] J. B. Rawlings, "Tutorial overview of model predictive control," *IEEE Control Syst. Mag.*, vol. 20, no. 3, pp. 38–52, Jun. 2000.

[26] J. Rawlings, D. Mayne, and M. Diehl, *Model Predictive Control: Theory, Computation, and Design*. San Francisco, CA, USA: Nob Hill Publishing, 2017.

[27] D. Corneil and J. Fonlupt, "The complexity of generalized clique covering," *Discrete Appl. Math.*, vol. 22, no. 2, pp. 109–118, 1988. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0166218X88900868

[28] M. Zhou et al., "SMARTS: An open-source scalable multi-agent RL training school for autonomous driving," in *Proc. Mach. Learn. Res.*, Cambridge, MA, USA, 2021, pp. 264–285.

[29] L. Wei, Z. Li, J. Gong, C. Gong, and J. Li, "Autonomous driving strategies at intersections: Scenarios, state-of-the-art, and future outlooks," in *Proc. IEEE Conf. Intell. Transport. Syst. Proc. (ITSC)*, Indianapolis, IN, United states, 2021, pp. 44–51.

[30] G.-P. Antonio and C. Maria-Dolores, "Multi-agent deep reinforcement learning to manage connected autonomous vehicles at tomorrow's intersections," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7033–7043, Jul. 2022.

[31] Z. Guo, Y. Wu, L. Wang, and J. Zhang, "Coordination for connected and automated vehicles at non-signalized intersections: A value decomposition-based multiagent deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 72, no. 3, pp. 3025–3034, Mar. 2023.

[32] S. Chen, X. Hu, J. Zhao, R. Wang, and M. Qiao, "A review of decision-making and planning for autonomous vehicles in intersection environments," *World Electr. Veh. J.*, vol. 15, no. 3, 2024, Art. no. 99.

[33] X. Xu, R. Li, Z. Zhao, and H. Zhang, "Communication-efficient consensus mechanism for federated reinforcement learning," in *Proc. IEEE Int. Conf. Commun.*, Virtual Edition, 2022, pp. 80–85.

[34] S. Feng et al., "Dense reinforcement learning for safety validation of autonomous vehicles," *Nature*, vol. 615, pp. 620–627, Mar. 2023.

[35] D. Huseljic, B. Sick, M. Herde, and D. Kottke, "Separation of aleatoric and epistemic uncertainty in deterministic deep neural networks," in *Proc. Int. Conf. Pattern Recognit.*, Milan, Italy, 2021, pp. 9172–9179.

[36] M. Deisenroth and C. E. Rasmussen, "PILCO: A model-based and data-efficient approach to policy search," in *Proc. Int. Conf. Mach. Learn.*, Bellevue, Washington, USA, 2011, pp. 465–472.

[37] S. Gu, T. Lillicrap, I. Sutskever, and S. Levine, "Continuous deep Q-learning with model-based acceleration," in *Proc. Int. Conf. Mach. Learn.*, New York, USA, 2016, pp. 2829–2838.

[38] F.-M. Luo, T. Xu, H. Lai, X.-H. Chen, W. Zhang, and Y. Yu, "A survey on model-based reinforcement learning," *Sci. China Inform. Sci.*, vol. 67, no. 2, pp. 1–26, Jan. 2024.

[39] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," *Mach. Learn.*, vol. 49, no. 2, pp. 209–232, Nov. 2002.

[40] P. Auer and R. Ortner, "Logarithmic online regret bounds for undiscounted reinforcement learning," in *Proc. Adv. Neural Inf. Proces. Syst.*, Vancouver, BC, Canada, 2006, pp. 49–56.

[41] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, pp. 1563–1600, Aug. 2010.

[42] R. Ortner and D. Ryabko, "Online regret bounds for undiscounted continuous reinforcement learning," in *Proc. 25th Int. Conf. Neural Inf. Proces. Syst. (NIPS)*, Stateline, NV, USA, 2012, pp. 1763–1771.

[43] I. Osband, D. Russo, and B. Van Roy, "(More) efficient reinforcement learning via posterior sampling," in *Proc. 26th Int. Conf. Neural Inf. Proces. Syst. (NIPS)*, Lake Tahoe, Nevada, USA, 2013, pp. 3003–3011.

[44] B. Hao, Y. Abbasi-Yadkori, Z. Wen, and G. Cheng, "Bootstrapping upper confidence bound," in *Proc. 3rd Int. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, Canada, 2019, pp. 12123–12133.

[45] J. Lidard, U. Madhushani, and N. E. Leonard, "Provably efficient multi-agent reinforcement learning with fully decentralized communication," in *Proc. Amer. Control Conf. (ACC)*, Atlanta, GA, USA, 2022, pp. 3311–3316.

[46] D. S. Bernstein, S. Zilberstein, and N. Immerman, "The complexity of decentralized control of markov decision processes," *Math. Oper. Res.*, vol. 27, no. 4, pp. 819–840, Nov. 2002.

[47] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, Apr. 1994.

[48] R. Ortner, P. Gajane, and P. Auer, "Variational regret bounds for reinforcement learning," in *Proc. 35th Uncertainty Artif. Intell. Conf.*, Tel Aviv-Yafo, Israel, 2019, pp. 81–90.

[49] C. B. Browne et al., "A survey of Monte Carlo tree search methods," *IEEE Trans. Comp. Intel. AI*, vol. 4, no. 1, pp. 1–43, Mar. 2012.

[50] W. Li and E. Todorov, "Iterative linear quadratic regulator design for nonlinear biological movement systems," in *Proc. 1st Int. Conf. Inform. Control, Automat. Robot.*, Setúbal, Portugal, 2004, pp. 222–229.

[51] Z. I. Botev et al., "The cross-entropy method for optimization," in *Handbook of Statistics*, New York, NY, USA: Elsevier, 2013, vol. 31, pp. 35–59.

[52] A. Girard et al., "Multiple-step ahead prediction for non linear dynamic systems–a Gaussian process treatment with propagation of the uncertainty," in *Proc. Adv. Neural Inf. Proces. Syst. (NIPS)*, Vancouver, British Columbia, Canada, 2002, pp. 529–536.

[53] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* 2nd ed. Cambridge, MA, USA: MIT Press, 2018.

[54] N. Ratliff, D. Silver, and J. A. D. Bagnell, "Learning to search: Functional gradient techniques for imitation learning," *Auton. Robots*, vol. 27, no. 1, pp. 25–53, Jul. 2009.

[55] K. Azizzadenesheli, E. Brunskill, and A. Anandkumar, "Efficient exploration through Bayesian deep q-networks," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, United States, 2018, pp. 1–9.

[56] A. G. Barto and S. Mahadevan, "Recent advances in hierarchical reinforcement learning," *Discrete Event Dyn. Syst.*, vol. 13, pp. 341–379, 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:386824

[57] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, 1994.

[58] R. Ortner, O.-A. Maillard, and D. Ryabko, "Selecting near-optimal approximate state representations in reinforcement learning," in *Lect. Notes Comput. Sci.*, Bled, Slovenia, 2014, pp. 140–154.

[59] V. Mnih et al., "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, Feb. 2015.

[60] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Mach. Learn. Res.*, Stockholm, Sweden, 2018, pp. 1861–1870.

Ruoqi Wen (Student Member, IEEE) received the B.Sc. degree in mathematics and applied mathematics from the Hefei University of Technology, Hefei, China, in June 2021. She is currently working toward the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. Her research interests include multi-agent reinforcement learning, regret bound, endogenous and exogenous uncertainty analysis, wireless networks, and autonomous vehicle control.
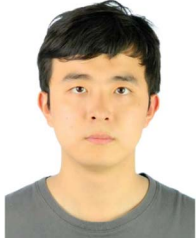
Jiahao Huang received the B.E. degree in information engineering from Zhejiang University, Hangzhou, China, in 2023. He is currently working toward the master's degree with the Department of Information Science and Electronic Engineering, Zhejiang University, Hangzhou. His research interests include machine learning, signal detection, vehicular communications, and radio resource management.

Rongpeng Li (Senior Member, IEEE) is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. From 2015 to 2016, he was a Research Engineer with the Wireless Communication Laboratory, Huawei Technologies Company, Ltd., Shanghai, China. In 2020, he was a Visiting Scholar with the Department of Computer Science and Technology, University of Cambridge, Cambridge, U.K. His research focuses on networked intelligence for communications evolving (NICE). He was the recipient of the Wu Wenjun Artificial Intelligence Excellent Youth Award in 2021. He is as an Editor for *China Communications*.

Guoru Ding received the B.S. (Hons.) degree in electrical engineering from Xidian University, Xi'an, China, in 2008, and the Ph.D. (Hons.) degree in communications and information systems from the College of Communications Engineering, Nanjing, China, in 2014. From 2015 to 2018, he was a Postdoctoral Research Associate with the National Mobile Communications Research Laboratory, Southeast University, Nanjing. He is currently a Professor with the College of Communications Engineering, Nanjing. His research interests include cognitive radio networks, massive MIMO, machine learning, and data analytics over wireless networks. He was the recipient of the Excellent Doctoral Thesis Award of the China Institute of Communications in 2016, Alexander von Humboldt Fellowship in 2017, and 14th IEEE COMSOC Aisa-Pacific Outstanding Young Researcher Award in 2019, Natural Science Foundation for Distinguished Young Scholars of Jiangsu Province, China, and six best paper awards from international conferences such as the IEEE VTC-FALL 2014. He was the Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS (special issue on spectrum sharing and aggregation in future wireless networks) and the *Chinese Journal of Aeronautics* (special issue on when aeronautics meets 6 G and AI).

Zhifeng Zhao (Member, IEEE) received the B.E. degree in computer science, the M.E. degree in communication and information systems, and the Ph.D. degree in communication and information systems from the PLA University of Science and Technology, Nanjing, China, in 1996, 1999, and 2002, respectively. From 2002 to 2004, he acted as a Posdoctoral Researcher with Zhejiang University, Hangzhou, China, where his researches were focused on multimedia next-generation networks (NGNs) and soft switch technology for energy efficiency. From 2005 to 2006, he acted as a Senior Researcher with the PLA University of Science and Technology, where he performed research and development on advanced energy-efficient wireless routers, ad-hoc network simulators, and cognitive mesh networking test-bed. From 2006 to 2019, he was an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University. He is currently with the Zhejiang Lab, Hangzhou, as the Chief Engineering Officer. His research interests include software-defined networks (SDNs), wireless networks in 6 G, computing networks, and collective intelligence. He is the Symposium Co-Chair of ChinaCom 2009 and 2010. He is the Technical Program Committee (TPC) Co-Chair of the 10th IEEE International Symposium on Communication and Information Technology (ISCIT 2010).