Towards Cost-Efficient Federated Multi-Agent RL with Learnable Aggregation

Yi Zhang[1,2], Sen Wang[1], Zhi Chen[1], Xuwei Xu[1,2],
Stano Funiak[2], and Jiajun Liu[2,1]

[1] School of Electrical Engineering and Computer Science,
The University of Queensland, St Lucia, QLD 4066, Australia
{uqyzha91, sen.wang, zhi.chen, xuwei.xu}@uq.edu.au
[2] DATA61, Commonwealth Scientific and Industrial Research Organisation (CSIRO),
Pullenvale, QLD 4069, Australia
{stano.funiak, ryan.liu}@data61.csiro.au

Abstract. Multi-agent reinforcement learning (MARL) often adopts centralized training with a decentralized execution (CTDE) framework to facilitate cooperation among agents. When it comes to deploying MARL algorithms in real-world scenarios, CTDE requires gradient transmission and parameter synchronization for each training step, which can incur disastrous communication overhead. To enhance communication efficiency, federated MARL is proposed to average the gradients periodically during communication. However, such straightforward averaging leads to poor coordination and slow convergence arising from the non-*i.i.d.* problem which is evidenced by our theoretical analysis. To address the two challenges, we propose a federated MARL framework, termed cost-efficient federated multi-agent reinforcement learning with learnable aggregation (FMRL-LA). Specifically, we use asynchronous critics to optimize communication efficiency by filtering out redundant local updates based on the estimation of agent utilities. A centralized aggregator rectifies these estimations conditioned on global information to improve cooperation and reduce non-*i.i.d.* impact by maximizing the composite system objectives. For a comprehensive evaluation, we extend a challenging multi-agent autonomous driving environment to the federated learning paradigm, comparing our method to competitive MARL baselines. Our findings indicate that FMRL-LA can adeptly balance performance and efficiency. Code and appendix can be found in https://github.com/ArronDZhang/FMRL_LA.

Keywords: Multi-agent reinforcement learning · Federated learning.

1 Introduction

Federated reinforcement learning (RL) [5, 12, 10] has exhibited immense potential in integrating deep reinforcement learning models into a client-server paradigm. It has been proven effective in balancing communication efficiency and privacy preservation. With the burgeoning rise of the Internet of Things [22] that requires agent cooperation, and the prevalent use of multi-agent systems (MAS) [17, 19], it is desirable to develop federated multi-agent reinforcement learning (MARL) frameworks.

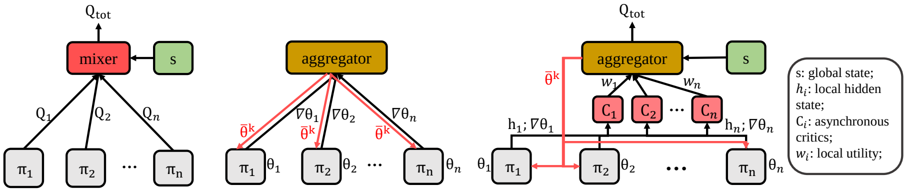

Fig. 1. Framework Comparison. π_i: local policy whose parameters are θ_i; Q_i: local value function; Q_{tot}: joint value function; $\nabla\theta_i$: local gradients; $\bar{\theta}^k$: global model's parameters at round k; h_i: local hidden state; w_i: local utility. The left subplot indicates the CTDE framework where the agents share parameters. The middle one represents the current vanilla FMARL where the agents maintain a global model through periodical averaging and the right figure refers to our FMRL-LA framework which selectively chooses and weights the involved agents to maximize the system utility.

In MARL, centralized training with decentralized execution (CTDE) [23, 13, 26] is a conventional learning regime lying between independent learning [30] and fully centralized learning [29]. This middle-ground strategy can not only mitigate the non-stationarity caused by agents' simultaneous decision-making but also prevent state and action spaces from expanding exponentially with agent number. Nevertheless, the training phase of CTDE requires continual communication between agents and servers. Thus, simply incorporating CTDE into federated learning (FL) [18] will lead to intractable communication overhead and bandwidth burdens. On the other hand, agent interactions with their local environments make their experiences non-independent and identically distributed (non-*i.i.d.*).

While recent efforts like FMARL [31] and Fed-MADRL [25] have marked advances in federated MARL [5], they typically assume an implicit *i.i.d.* in agent interactions and lack server-side coordination. Furthermore, these methods tend to optimize singular, task-oriented objectives, *e.g.,* the average speed in multi-vehicle autonomous systems [31] and the system throughput in wireless communications [25]. Such settings may be impractical for complex real-world settings with composite objectives. For instance, in autonomous driving [14, 21, 7], apart from communication efficiency, we also consider factors like success rate in reaching destinations, overall safety, and average vehicle speed.

In response to these challenges, we introduce Cost-Efficient Federated MARL with Learnable Aggregation (FMRL-LA). It decouples the CTDE by separating the training steps of the server and the client. On the server side, we propose two components for learnable aggregation: 1. Asynchronous critics evaluate the utility of learning agents, guiding selection for optimal system communication. 2. A centralized aggregator integrates global information with agent utilities to periodically update the global model, thus maximizing composite system targets. This design facilitates FMRL-LA to improve coordination under the federated paradigm. Delving deeper into the non-*i.i.d.* challenge posed by federated MARL, we theoretically delineate its adverse effects, providing a convergence upper bound. We further prove that the proposed learnable aggregation can mitigate the challenge. The comparison of different frameworks is exhibited in Fig. 1.

To conduct experiments with FMRL-LA, we resort to real-world multi-agent environment simulations based on MetaDrive [14], an intricate autonomous driving benchmark out of its flexibility across diverse scenarios. We extend it to support a client-server learning paradigm, incorporating communication efficiency. To further enhance the practicality, in addition to the existing navigation tasks, we design a multi-vehicle cooperative exploration task. Notably, we have integrated baselines from the representative methods of cooperative MARL [30] and communication-inclusive MARL [8], as well as the state-of-the-art method [21] using MetaDrive. Our experimental evaluations in navigation and exploration tasks underscore that FMRL-LA can optimize system performance and efficiency simultaneously, delivering a balanced performance across the metrics corresponding to composite objectives.

2 Preliminary

Cooperative MARL Cooperative MARL can be formulated as Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs) [13, 33], described by a tuple $G = \langle n, S, O, A, P, r, Z, \gamma \rangle$, where n is the number of agents, and S, O denote the state and observation spaces. A, the joint action space, is the product of all agents' action spaces, *i.e.,* $A = \prod_{i=1}^{n} A_i$, where i is the agent index. We use lowercase s, o, a to represent an element in the corresponding space. The environments' dynamics are characterized by the transition function $P(s'|s, a) : S \times A \times S \to [0, 1]$. The system has a shared team reward function $r(s, a) : S \times A \to \mathbb{R}$. In the aspect of each agent, due to the partially observable setting, at time step t, its observation o_t is drawn by applying the function Z to the current state s_t. Thus, $o_i^t = Z_i(s^t) : S \to O$. γ is the discount factor. The solution of a Dec-POMDP is a joint policy $\bar{\pi} = (\pi_1, \pi_2, ..., \pi_n)$, where π_i stands for the policy of agent i and we use θ_i, $\bar{\theta}$ to represent the parameters of agent i and the joint policy, respectively. Each agent policy is trained with the agent's experience comprised of a collection of agent observation-action history denoted as $\xi_i = \{(o^t, a^t, o^{t+1})\}_{t=0}^{T}$, where T denotes the time horizon. In addition, we use $\xi = \{(s^t, a^t, s^{t+1}, r^t)\}_{t=0}^{T}$to represent one global team episode. The goal of MARL is to learn a joint policy that can maximize the expected cumulative reward, *i.e.,* $\pi^* = \arg\max_{\bar{\pi}} \mathbb{E}_{\tau \sim \bar{\pi}}[R_T(\tau)], \text{where} R_T(\tau) = \sum_{t=0}^{T} \gamma^t r^{(t)}$.

Federated MARL We use τ to represent the number of local updates. K is the termination condition of the training process, which is usually set as maximum communication rounds [5]. ψ denotes the system communication efficiency. We use the parameter θ to represent policy π. $F(\cdot)$ represents the global objective function, while $F_i(\cdot)$ stands for the local objective function for each agent i. Their relationship between the global objective and the locals in [25, 31, 5] are the same: $F(x) = \frac{1}{n} \sum_{i=1}^{n} F_i(x)$. In round k, all agent policies are synchronized as $\bar{\theta}^k$, which is drawn from the server. Then, each agent interacts with the environment concurrently to accumulate local experience for updating the local policy indicated by $\{\theta_i^{k,\tau_i}\}_{i=1}^{n}$. Next, the parameters $\{\theta_i^{k,\tau_i}\}_{i=1}^{n}$ or stochastic gradients $\{g(\theta_i^{k,j}; \xi_i^{k,j})\}_{j=1}^{\tau_i}$ for $i \in 1, 2, \cdots, n$ will be uploaded to the server. To sum up, the update rules on the server and client side are:

$$\bar{\theta}^{k+1} = \bar{\theta}^{k} - \eta \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{\tau_i^k} g(\theta_i^{k,j}), \quad \theta_i^{k+1,j} = \begin{cases} \bar{\theta}^{k+1}, & j \bmod \tau_i = 0, \\ \theta_i^{k,j} - \eta g(\theta_i^{k,j}), & \text{otherwise.} \end{cases} \tag{1}$$

To indicate the convergence of the algorithm, we use the expected averaged gradient norm to guarantee convergence to a stationary point [28, 3, 27]:

$$\mathbb{E}[\frac{1}{K} \sum_{k=0}^{K-1} ||\nabla F(\overline{\theta}^k)||^2] \leq \epsilon, \tag{2}$$

where $|| \cdot ||$ is the ℓ_2-norm and ϵ is used to describe the sub-optimality. When the above condition holds, we say the algorithm achieves an ϵ-suboptimal solution.

3 Federated MARL with Learnable Aggregation

Server Side When federated MARL [31, 25] adopts Eq. (1) as the update rule for the server, it implicitly assumes that the agents are homogeneous. However, in real-world environments, the agents are diverse in various aspects such as computing capability, network connection, and local observation distributions, which results in heterogeneous agents with non-*i.i.d.* experience distribution.

To deal with these issues, we introduce **Asynchronous Critics** to dynamically evaluate the agent utilities in each communication round. Each critic corresponds to one learning agent. Its goal is to maximize the return of the current agent. The inputs are hidden information h_i^k, accumulated rewards in recent communication round $r^k := \sum_{j=\tau_i^{k-1}}^{\tau_i^k} r_i^j$ and in agent history $\sum_{j=0}^{\tau_i^k} r_i^j$. The output is a prediction of the agent's local utility:

$$w_i^k = C_i \left(h_i^k, r^k, \sum_k r^k \right), \tag{3}$$

where C_i is the asynchronous critic network of agent i. The output w_i^k can be zero, which means the corresponding agent does not need to upload its training parameters to the server in the current communication round to implement client selection.

Next, the agent utilities are passed through a **Centralized Aggregator** to facilitate coordination. It works similarly to the mixing network in value decomposition methods such as [23, 26], which takes the local utility function as input and facilitates agent coordination by maximizing the system utility condition on the global state. The RL loss is back-propagated to the critics for improving the local utility estimation:

$$Q_{tot} = Mix\left(w_1^k, w_2^k, \cdots, w_n^k | s\right), \tag{4}$$

where Q_{tot} denotes the system utility, which is reflected by the composite objectives. The server aggregates the gradients based on the local utilities to update

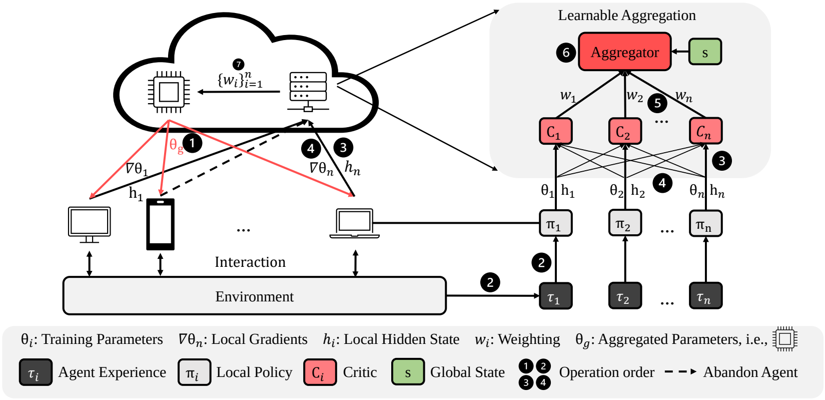

Fig. 2. The workflow of our proposed framework.

the global policy. Thus, the update rule of the server is:

$$\bar{\theta}^{k+1} = \bar{\theta}^{k} - \eta \sum_{i=1}^{n} w_i^k \sum_{j=1}^{\tau_i^k} g(\theta_i^{k,j}), \tag{5}$$

while the update rule for the clients remains the same as the client side of Eq. 1.

Client Side Considering the generalization of our method, we choose an independent reinforcement learning algorithm and take the hidden states as additional outputs. During the communication, the upload process of the clients can be divided into two stages. In the first stage, the agents upload their rewards and hidden information to the asynchronous critics for local utility estimation and agent selection, optimizing communication efficiency. In the second stage, the selected agents upload their gradients to the server for aggregation.

Framework Design Compared to the update rule for FMARL [31, 25] represented by Eq. (1), we adopt a weighted aggregation for global policy update implemented by the learnable aggregation module. The workflow of FMRL-LA is illustrated by Fig. 2. Specifically, 1. the server broadcasts the global model $\bar{\theta}^k$ to each agent; 2. The agents learn local behavior policies $\{\pi_i\}_{i=1}^n$ by interacting with the environment and maintaining hidden states; 3 and 4. During client-server communication, agents conduct the two-stage upload described in the above subsection; 5. The centralized aggregator maximizes the composite system objectives condition on the global states to facilitate coordination; 6. The global model is updated based on the local utilities. The corresponding pseudo code is in Appendix 1.

4 Convergence Analysis

In FL theory, a substantial body of research is devoted to exploring convergence properties under diverse settings. These settings predominantly fall into two categories, *i.i.d.*[18, 31, 25, 5], and non-*i.i.d.* [15, 28, 11, 20]. While *i.i.d.* settings facilitate robust theoretical results, non-*i.i.d.* settings are more realistic about data distribution. Despite the theoretical progresses in FL schemes in supervised learning, the influence of non-*i.i.d.* in federated MARL remains uncharted. To address this issue, we conduct our theoretical analysis in the following paragraphs.

We begin with showing the convergence under the ideal *i.i.d.* setting. To do that, we first list out the following assumptions:

Assumption 1 *(Lipschitz continuity) The local loss functions at the client side are Lipschitz continuous, which means* $||\nabla F_i(\theta_1) - \nabla F_i(\theta_2)|| \le L||\theta_1 - \theta_2||, \forall i \in \{1, 2, ..., n\}$.

Assumption 2 *(Unbiased gradients and bounded variance under i.i.d.) The stochastic gradients at the client side are unbiased estimators of the global gradient, i.e.,* $\mathbb{E}_\xi[g_i(\theta)] = \nabla F(\theta)$ *and* $\mathbb{E}_\xi[||g_i(\theta) - \nabla F(\theta)||^2] \le \mu||\nabla F(\theta)||^2 + \sigma^2, \forall i \in \{1, 2, ..., n\}$, μ *and* σ^2 *are non-negative.*

Assumption 3 *(Unbiased local gradients and bounded variance under non-i.i.d.) The stochastic gradient at each client is an unbiased estimator of the local gradient, i.e.,* $\mathbb{E}_\xi[g_i(\theta)] = \nabla F_i(\theta)$ *and* $\mathbb{E}_\xi[||g_i(\theta|\xi) - \nabla F_i(\theta)||^2] \le \mu||\nabla F_i(\theta)||^2 + \sigma^2, \forall i \in \{1, 2, ..., n\}$, μ *and* σ^2 *are non-negative.*

Assumption 4 *(Bounded Dissimilarity) For any sets of weights* $\{w_i^k \ge 0\}_{i=1}^n$, $\sum_{i=1}^n w_i^k \le M^k, M^k$*is finite,*$\forall k \in [0, K]$*, there exist constants* $\beta^2 \ge 1, \kappa^2 \ge 0$ *such that* $\sum_{i=1}^n w_i^k||\nabla F_i(\theta)||^2 \le \beta^2\{||\sum_{i=1}^n w_i^k \nabla F_i(\theta)||^2, ||\sum_{i=1}^n \frac{1}{n}\nabla F_i(\theta)||^2\}_{min} + \kappa^2, \forall k \in [0, K]$. *If local loss functions are identical to each other, then we have* $\beta^2 = 1$, $\kappa^2 = 0$.

Assumption 1 is Lipschitz continuity, a common assumption in the convergence analysis in FL theory. Assumption 2 states that the local stochastic gradient is an unbiased estimation of the local gradient and the variance of the deviation is bounded to support our exploration under a *i.i.d.* setting. Assumption 3, on the other hand, is the gradient bias and variance assumption under non-*i.i.d.* setting. Assumption 4 is inspired by FedNova [28], which bounds the dissimilarities on the weighted norm of local gradients.

We provide the convergence bound under the *i.i.d.* and non-*i.i.d.* settings as Theorem 1 and Theorem 2, respectively. We further provide a special case to show that there is room for tightening the bound with the learnable aggregation mechanism as Theorem 3 in the Appendix A.4. In addition, the proof of these theorems as well as more theoretical details are provided in the Appendix A.4.

Theorem 1. *Suppose* $\{\theta_i^{k,j}\}$ *and* $\{\overline{\theta}^k\}$ *are parameters' sequences generated by Eq.*(1)*. The federated MARL is conducted under Assumptions 1 and 2. If the total communication rounds K is large enough, which can be divided by* τ*, and the learning rate* η *satisfies:*

$$\left\{L\eta < 1, 2L^2\eta^2\tau(2\mu + 1 + \tau) < 1\right\}, \tag{6}$$

then the expected gradient norm after K iterations is bounded by:

$$\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}||\nabla F(\bar{\theta}^k)||^2] \le \frac{2[F(\bar{\theta}^1) - F(\bar{\theta}^K)]}{\eta K} + \frac{\eta L\sigma^2}{n} + \eta^2L^2\sigma^2(\tau + 1), \tag{7}$$

where $\bar{\theta}^1$ stands for one lower bound of the objective function.

Theorem 2. *Suppose $\{\overline{\theta}^k\}$ are parameters' sequences generated by the weighted gradients Eq.(5), while the $\{\theta_i^{k,j}\}$ remains the same. The federated MARL is conducted under Assumptions 1, 3 and 4. If the total communication rounds K is large enough, and the learning rate η satisfies (6), then the expected gradient norm after K iterations is bounded by:*

$$\begin{aligned}\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}||\nabla F(\bar{\theta}^k)||^2] \leq \frac{4\left(E\left[F\left(\bar{\theta}^1\right)-E\left[F\left(\bar{\theta}^K\right)\right]\right)\right.}{K\eta} \\ + 4\left(C+D+E+F+\mu\eta C\sum_{k=0}^{K}\frac{1}{K}\sum_{i=1}^{n}w_i^2\tau_i^k\right),\end{aligned} \tag{8}$$

where $\bar{A} = \frac{1}{K}\sum_{i=1}^{K} A$, $B = 2L^2\eta^2\tau(2\mu+1+\tau)$, $C = \frac{\eta^2\sigma^2L^2}{\mu L\eta\tau\beta^2+2B\beta^2}$, $D = \frac{\left(1-2\mu L\eta\tau\beta^2\right)\kappa^2}{(2\mu L\eta\tau\beta^2+4B\beta^2)(1+4\beta^2)}$, $E = \frac{\mu L\eta\tau\kappa^2}{2\mu L\eta\tau\beta^2+4B\beta^2}$, *and* $F = \frac{L\eta\sigma^2}{2K}\sum_{k=1}^{K}\left(M^k\right)^2$.

Discussion The result of Theorem 1 is an ideal upper bound where the distribution of each client is *i.i.d.* More generally, in Theorem 2, we provide another upper bound to illustrate the impact introduced by the non-*i.i.d.* issue. Comparing the bounds in these two theorems, it is obvious that the convergence bound in the non-*i.i.d.* setting is greater than that in the *i.i.d.* setting.

Special Cases When $w_i^k \equiv \frac{1}{n}$, the convergence upper bound degenerates to the same as FMARL [31], which derives the same upper bound as in its Theorem 2. When $w_i^k \equiv \frac{1}{n}$ and $\tau_i^k \equiv 1$, the convergence upper bound further degenerates to the same as PASGD, which coincides with the conclusions drawn from [27].

Discussion with Federated Learning in supervised learning We compare our method with FedNova [28] – a general federated method targeting supervised learning. It induces several federated learning methods in a general form and targets the problem of an unbalanced number of local updates by regularizing the weights for one-period local gradients with the number of local updates. However, in MARL, the different number of policy iterations may not be a more significant reason than the diversity of local environments to the non-*i.i.d.* issue. In other words, this issue cannot be rectified by simply regularizing the weights of local gradients by the number of local policy iterations, which necessitates the importance of our learnable aggregation mechanism.

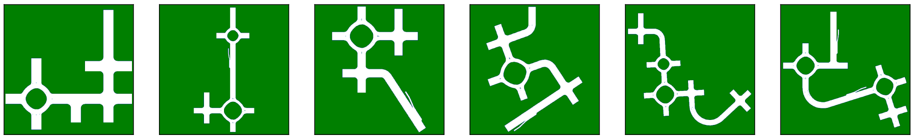

Fig. 3. The six extended scenarios used in our evaluation.

5 Experiments

Baseline Methods We present a comparative analysis of our proposed method alongside strong baselines covering a wide range of related fields, namely conventional MARL (**IPPO** [30]), communication-based MARL (**RIAL and DIAL** [8]), the state-of-the-art method **CoPO** [21] and **FMARL** [31] in a multi-agent autonomous driving simulation environment, MetaDrive [14]. We provide a detailed introduction and adaptation of these methods in the Appendix A.6.

Implementation by Extending MetaDrive The MetaDrive benchmark represents a flexible and lightweight simulation benchmark for autonomous driving, encompassing a variety of tasks that serve as a reasonable abstraction of real-world environments. In this paper, we focus on its multi-agent tasks. The agents adopt the conventional MARL suite, including parameter sharing and disregarding communication overhead. We add six challenging scenarios whose maps are depicted in Fig 3. Originally, MetaDrive used parameter sharing for all the methods, so we first expand this benchmark into a client-server learning setting by adopting a non-parameter sharing scheme and simulating a virtual server. This server only periodically collects local gradients and hidden states from the clients, aggregates the gradients for the update of the global model, and then sends it back to the agents. When an existing vehicle terminates and a new vehicle spawns, it accepts the latest global model from the server to prevent the "cold start" problem. To enrich the testing bed and serve as a real-world simulation, in addition to the existing navigation task, we extend a cooperative exploration task where vehicles cooperatively explore the specified destinations.

Evaluation Metrics MetaDrive provides three evaluation metrics: success rate, efficiency, and safety, which respectively reflect navigation capability (the success ratio of vehicles relative to the total number of vehicles), navigation efficiency (the differences between the successes and failures within a unit of time), and safety driving (the number of crushes within an episode). In our cooperative exploration task, we adapt the navigation success rate to the exploration success rate. For realistic concerns, we also record the communication overhead which is reflected by the number of parameters exchanged between the agents and the server. The **system utility** is derived from the weighted average of these metrics, where the weights reflect the specific requirements of particular scenarios. In our experiments, we employ a simple average to gauge overall performance.

In summary, in the **cooperative navigation** task, our evaluation metrics including the navigation success rate (*Success*), safety (*Safety*), overall agents' speed (*Speed*), and communication efficiency (*Comm-efficiency*). As for the **cooperative exploration**, our evaluation metrics including the exploration success rate (*Explore*), safety (*Safety*), overall agents' speed (*Speed*), communication efficiency (*Comm-efficiency*).

Main Results Analysis The experiment results on cooperative navigation and exploration across six scenarios can be found in Fig. 4 and Fig. 5, respectively. More detailed results related to the performance on two tasks can be found in Tab. 2, 3, 4, and 5 in the Appendix A.7.

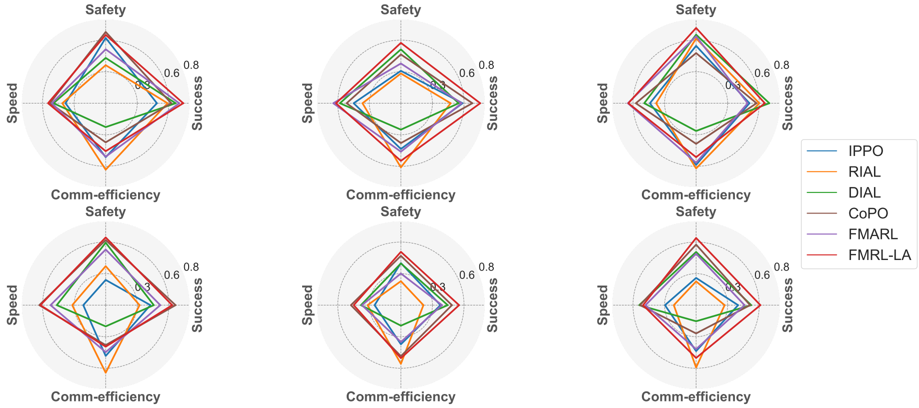

Fig. 4. The system performance and efficiency comparison with baselines in six scenarios of the cooperative navigation tasks.

Our Performance. In both tasks, FMRL-LA achieves or is comparable to the best success, speed, safety, and system utility. From the perspective of communication efficiency, since RIAL [8] uses a simple actor-critic algorithm with few parameters, its communication efficiency serves as a reference of the upper bound for the methods with a relatively complex algorithm on the client side. Actually, more than half of the baselines adapt PPO[24] as the clients' algorithm. Among them, our method exhibits the capacity to dynamically harmonize the system performance and communication efficiency.

In detail, we focus on the performance of IPPO, FMARL, and our FMRL-LA in both tasks. The three methods have nearly the same client-side algorithms but differ from each other on the server side. IPPO only conducts direct training parameter averaging, while the FMARL adds a weight decay mechanism during the averaging. And FMRL-LA dynamically learns the aggregation weights. From IPPO to FMARL, then to FMRL-LA, the performance of success, safety, speed, and system utility follow an ascending manner. We believe that it is because the performance of IPPO is bound by the averaging capability of all learning agents while FMARL can enlarge the bound to some extent by weight decay. Nonetheless, the potential performance of FMRL-LA is bound by the best agent, which improves the generalization of our method since we can deploy a suitable behavior model for the clients in advance if we can make use of prior knowledge about the environments.

Task Comparison. Comparing the overall performance of all methods on cooperative navigation and cooperative exploration on six scenarios, we can find navigation is more difficult than exploration, especially in relatively complex scenarios, *e.g.*, scenario 4, 5, and 6. We notice that in the navigation task, each agent has its own destination. Therefore, we believe the different performance on the two tasks may be because the relationship between the local utilities and the system utility is easier to capture in exploration than in navigation.

Scenario Difficulty. In both tasks, if we compare the performance pair by pair such as scenario 1 and scenario 5, we observe that generally, the more building blocks involved in a scenario, the more challenging it is. Then, looking into the performance of scenario 1 and 2, both of them consists of four building blocks while scenario 2 contains one more roundabout than scenario 1. If we compare the safety of CoPO, FMARL, and our FMRL-LA, three robust methods in the two tasks on these six scenarios, we can find that roundabout tends to result in more crushes. Further, if we compare the performance of scenarios 1 and 4 on two tasks, we observe that though the two scenarios both contain one roundabout and the same number of building blocks, the performance of our method and other baselines is generally better in scenario 1. Considering the difference in these two scenarios, we hypothesize that it is due to the influence of wide turn. For intuitive methods like IPPO and RIAL, it is difficult for them to avoid crushing or driving out of the roads during the wide turns. On the other hand, the safety of CoPO in scenarios 3 and 4 is relatively high, it may benefit from its explicit modeling of the surrounding agents.

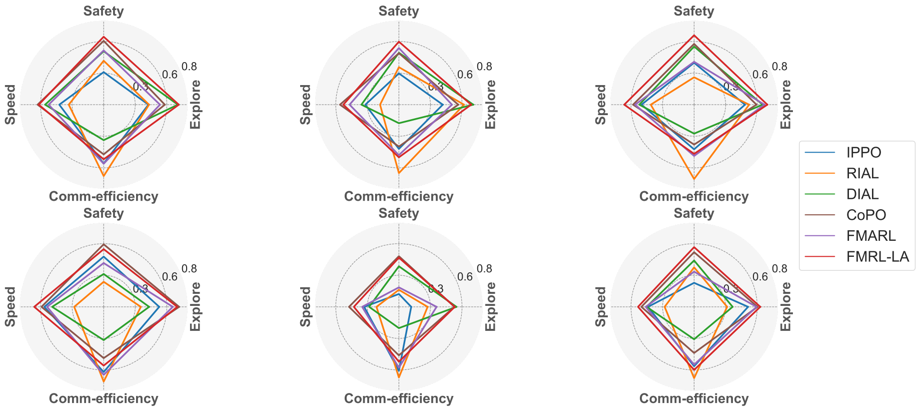

Fig. 5. The system performance and efficiency comparison with baselines on six scenarios of the cooperative exploration tasks.

Ablation Study To investigate the effectiveness of our design and components in FMRL-LA, we conduct an ablation study about the usage of asynchronous critics and a centralized aggregator as well as an alternative design for the centralized aggregator. We intuitively use the average of multiple metrics as the system utility. From Tab. 1 we can observe that if we directly use critics without the centralized aggregator, the performance is unstable, resulting in large standard errors. In scenario 4, the performance w.r.t. system utility is worse than federated IPPO. We believe that without the coordination of the centralized aggregator, the server cannot filter out less valuable agents, so their parameters can depreciate the update of the global model in the current round. Meanwhile, the asynchronous critics are useful in our method since the variant that only uses an aggregator performs worse than the full model. We believe that accepting information from all involved agents can stagnate the learning of an aggregator due to redundant information. When we change the centralized aggregator to a VDN-based [26] one, it yields an inferior performance compared to our QMIX-based [23] design, which suggests the non-linear modeling of the relationship between the agents and the server is more suitable for complex realistic environments than a simple sum as in VDN.

Table 1. Ablation study on the effectiveness of our critical components on navigation task. The **system utility** is provided.

Experiment	Scenario1	Scenario2	Scenario3	Scenario4	Scenario5	Scenario6
IPPO	49.96±3.06	45.13±3.68	51.49±2.73	33.12±5.29	34.41±6.17	34.33±5.97
FMRL-LA w/o aggregator	54.75±6.00	49.28±5.23	52.06±4.71	40.43±7.02	32.85±8.92	40.22±7.74
FMRL-LA w/o critics	52.69±3.55	48.46±4.65	55.49±3.18	45.35±5.26	38.49±6.19	48.07±5.57
FMRL-LA w/ vdn-aggregator	57.98±2.97	55.84±4.39	57.26±3.69	52.50±5.80	46.98±6.81	50.86±5.09
FMRL-LA	**59.84**±2.82	**62.30**±4.29	**63.16**±3.33	**57.27**±4.70	**50.28**±6.67	**56.42**±5.63

Client Selection Analysis To verify that FMRL-LA can effectively select involved agents to save communication costs, we also conduct experiments with varying numbers of agents to show the client selection effect. The experiment results and elaboration can be found in the Appendix A.7.

6 Related Work

Cooperative MARL Cooperative MARL has widespread applications in real-world scenarios [32, 14]. Current methods are mainly developed in game scenarios [23, 17, 4] where the methods can focus on technical design rather than practical details. These environments support parameter sharing (PS) [6] and CTDE regime [17, 9] to enable multiple agents to be trained on one device and facilitate cooperation, respectively. However, when it is the stage to consider practical MARL in realistic environments[21, 1], either PS or CTDE cannot be simply applied due to privacy concerns and communication overhead.

Federated MARL Federated MARL [25] appears to be a feasible way towards realistic MARL. Most of these methods enable agents to learn individual behavior policies and set a virtual server to maintain a global policy. The agents' policies are aggregated and updated periodically through communication with the server [5]. In this way, the communication overhead is reduced, and the majority of them aggregate the local gradients by direct averaging [31] or weighted by the relative mini-batch size [25]. Although this oversimplified update may work well under *i.i.d.* setting, the MASs are naturally non-*i.i.d.* due to the interaction among agents. The notorious non-*i.i.d.* issue can stagnate convergence [15, 28]. Besides, without centralized training, it is hard for MARL to learn coordination [13].

7 Conclusion

We aim to adapt MARL for real-world applications by introducing a hybrid distributed, client-server learning framework that takes into account communication and computation overhead. Our framework offers theoretical guarantees even under the influence of non-*i.i.d.* distribution of agents in local environments. To empirically validate the efficacy of our proposed Cost-Efficient Federated Multi-Agent Reinforcement Learning with Learnable Aggregation (FMRL-LA) method, we modify an existing multi-agent autonomous driving simulation environment to conform to a client-server scheme. Experimental results emphasize the superior performance against baseline methods.

8 Acknowledgements

This work is supported by project DE200101610 funded by Australian Research Council and CSIRO's Science Leader project R-91559.

References

1. Abegaz, M., Erbad, A., Nahom, H., Albaseer, A., Abdallah, M., Guizani, M.: Multi-agent federated reinforcement learning for resource allocation in uav-enabled internet of medical things networks. IoT-J (2023)
2. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: COMPSTAT (2010)
3. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM (2018)
4. Chaudhuri, R., Mukherjee, K., Narayanam, R., Vallam, R.D.: Collaborative reinforcement learning framework to model evolution of cooperation in sequential social dilemmas. In: PAKDD (2021)
5. Chen, T., Zhang, K., Giannakis, G.B., Başar, T.: Communication-efficient policy gradient methods for distributed reinforcement learning. TCNS (2021)
6. Christianos, F., Papoudakis, G., Rahman, A., Albrecht, S.V.: Scaling multi-agent reinforcement learning with selective parameter sharing. In: ICML (2021)
7. Du, X., Wang, J., Chen, S.: Multi-agent meta-reinforcement learning with coordination and reward shaping for traffic signal control. In: PAKDD (2023)
8. Foerster, J., Assael, I.A., De Freitas, N., Whiteson, S.: Learning to communicate with deep multi-agent reinforcement learning. In: NeurIPS (2016)
9. Hu, S., Zhu, F., Chang, X., Liang, X.: Updet: Universal multi-agent reinforcement learning via policy decoupling with transformers. In: ICLR (2021)
10. Jin, H., Peng, Y., Yang, W., Wang, S., Zhang, Z.: Federated reinforcement learning with environment heterogeneity. In: AISTATS (2022)
11. Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S., Suresh, A.T.: Scaffold: Stochastic controlled averaging for federated learning. In: ICML (2020)
12. Khodadadian, S., Sharma, P., Joshi, G., Maguluri, S.T.: Federated reinforcement learning: Linear speedup under markovian sampling. In: ICML (2022)
13. Kuba, J.G., Chen, R., Wen, M., Wen, Y., Sun, F., Wang, J., Yang, Y.: Trust region policy optimisation in multi-agent reinforcement learning. In: ICLR (2022)
14. Li, Q., Peng, Z., Feng, L., Zhang, Q., Xue, Z., Zhou, B.: Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning. TPAMI (2022)
15. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. In: MLSys (2020)
16. Lian, X., Huang, Y., Li, Y., Liu, J.: Asynchronous parallel stochastic gradient for nonconvex optimization. In: NeurIPS (2015)
17. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I.: Multi-agent actor-critic for mixed cooperative-competitive environments. In: NeurIPS (2017)
18. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: AISTATS (2017)
19. Mo, J., Xie, H.: A multi-player mab approach for distributed selection problems. In: PAKDD (2023)

20. Pang, Y., Zhang, H., Deng, J.D., Peng, L., Teng, F.: Rule-based collaborative learning with heterogeneous local learning models. In: PAKDD (2022)
21. Peng, Z., Hui, K.M., Liu, C., Zhou, B.: Learning to simulate self-driven particles system with coordinated policy optimization. In: NeurIPS (2021)
22. Pinto Neto, E.C., Sadeghi, S., Zhang, X., Dadkhah, S.: Federated reinforcement learning in iot: Applications, opportunities and open challenges. Appl. Sci. (2023)
23. Rashid, T., Samvelyan, M., De Witt, C.S., Farquhar, G., Foerster, J., Whiteson, S.: Monotonic value function factorisation for deep multi-agent reinforcement learning. JMLR (2020)
24. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv:1707.06347 (2017)
25. Song, Y., Chang, H.H., Liu, L.: Federated dynamic spectrum access through multi-agent deep reinforcement learning. In: GLOBECOM (2022)
26. Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W.M., Zambaldi, V.F., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J.Z., Tuyls, K., Graepel, T.: Value-decomposition networks for cooperative multi-agent learning. arXiv:1706.05296 (2017)
27. Wang, J., Joshi, G.: Cooperative sgd: A unified framework for the design and analysis of local-update sgd algorithms. JMLR (2021)
28. Wang, J., Liu, Q., Liang, H., Joshi, G., Poor, H.V.: Tackling the objective inconsistency problem in heterogeneous federated optimization. In: NeurIPS (2020)
29. Wen, M., Kuba, J.G., Lin, R., Zhang, W., Wen, Y., Wang, J., Yang, Y.: Multi-agent reinforcement learning is a sequence modeling problem. Front Comput. Sci. (2022)
30. de Witt, C.S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P.H., Sun, M., Whiteson, S.: Is independent learning all you need in the starcraft multi-agent challenge? arXiv:2011.09533 (2020)
31. Xu, X., Li, R., Zhao, Z., Zhang, H.: The gradient convergence bound of federated multi-agent reinforcement learning with efficient communication. TWC (2023)
32. Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., Wu, Y.: The surprising effectiveness of PPO in cooperative multi-agent games. In: NeurIPS (2022)
33. Zhou, X., Matsubara, S., Liu, Y., Liu, Q.: Bribery in rating systems: A game-theoretic perspective. In: PAKDD (2022)

A Appendix

A.1 Pseudo Code

Algorithm 1 Cost-efficient Federated MARL with Learnable Aggregation.

```
Input: The Environment $E$, the aggregation weight threshold $w_m$;
Initialize the policy network $\{\theta_i\}$ and weights for aggregation $\{w_i\}$;
repeat
    Client Side:
    for $i = 1$ to $n$, simultaneously do
        Interact with $E$ to collect data;
        Collect rewards $r$, hidden information $h_i$ and update $\theta_i$ by the right formula
        of Eq. (1);
        Upload $r$ and $h_i$ to the Server;
    end for
    Server Side:
        Calculate $w_i$ by Eq. (3);
        Select clients with $w_i \geq w_m$;
        Accept selected clients' $\{\theta_i^{\tau_i}\}$;
        Obtain $\bar{\theta}$ by Eq. (5);
        Broadcast $\bar{\theta}$ to the Clients;
        Derive the system utility by Eq. (4);
        Calculate the RL loss and update the critics;
until Converge or reach the terminate conditions.
Output: The policy $\bar{\theta}$ and weights $\{w_i\}$
```

A.2 Federated MARL

Though some work has implemented their federated MARL (FMARL) methods [31, 25] towards different environments, the specialty of FMARL with respect to conventional MARL is not fully demonstrated. In this part, we induce a general formulation of FMARL which integrates G with several new elements to derive $\Lambda =< G, \tau, K, \psi >$. Here τ indicates the number of local updates within each communication round while K is the termination condition of the training process which is usually set as maximal communication rounds [5, 12]. In addition, ψ denotes the system communication efficiency. We use the parameter θ to represent the policy π for simplicity. $F(\cdot)$ is used to represent the global objective function of the system whose minimization is equivalent to the maximization of the expected return. $F_i(\cdot)$ stands for the local objective function for each agent i. Their relationship between the global objective and the locals in [25, 31, 5] are the same: $F(x) = \frac{1}{n}\sum_{i=1}^{n} F_i(x)$.

The learning protocol is similar to federated learning in a supervised setting: in round k, all agents' policies are synchronized as $\bar{\theta}^k$ which is drawn from the server agent. Then, each agent interacts with the environment concurrently to accumulate local experience used for updating the local policy indicated by $\{\theta_i^{k,\tau_i}\}_{i=1}^n$ with SGD [2]: $g(\theta_i^{k,j};\xi_i^{k,j}) = \frac{1}{|\xi_i^{k,j}|}\sum_{\phi\in\xi_i^{k,j}}\nabla F_i(\phi)$, where ϕ stands for a transition in mini-batch $\xi_i^{k,j}$ and j is the index of local updates. Next, the parameters $\{\theta_i^{k,\tau_i}\}_{i=1}^n$ or stochastic gradients $\{g(\theta_i^{k,j};\xi_i^{k,j})\}_{j=1}^{\tau_i}$ for $i \in 1, 2, \cdots, n$ will be uploaded to the server agent.

To sum up, the update rule on the server side is:

$$\bar{\theta}^{k+1} = \bar{\theta}^k - \eta\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{\tau_i^k} g(\theta_i^{k,j}). \tag{9}$$

And the update rule for clients i is:

$$\theta_i^{k+1,j} = \begin{cases}\bar{\theta}^{k+1}, & j \bmod \tau_i = 0,\\ \theta_i^{k,j} - \eta g(\theta_i^{k,j}), & \text{otherwise,}\end{cases} \tag{10}$$

Since in real-world environments, agents with diverse devices may spend different time in interaction and policy iteration, we enable τ_i^k to be different times across agents.

In real-world settings, the objective functions or loss functions are usually non-convex, so the global policy optimized by SGD may fall into a local minimum or saddle point. To indicate the convergence of the algorithm, we use the expected averaged gradient norm to guarantee convergence to a stationary point [28, 3, 27, 31, 16]:

$$\mathbb{E}[\frac{1}{K}\sum_{k=0}^{K-1}||\nabla F(\overline{\theta}^k)||^2] \leq \epsilon, \tag{11}$$

where $||\cdot||$ is the ℓ_2-norm and ϵ is used to describe the sub-optimality. When the above condition holds, we say the algorithm achieves an ϵ-suboptimal solution.

A.3 Proof Preliminaries

In this subsection, we introduce some notations to facilitate reading. Then, some key lemmas as well as their proof will be provided.

To begin with, we define the sum of stochastic gradients and the full batch gradients at round k as: $\mathbf{X}_i^k := \sum_{j=1}^{\tau_i^k} g_i(\theta_i^{k,j})$ and $\mathbf{Y}_i^k := \sum_{j=1}^{\tau_i^k}\nabla F_i(\theta_i^{k,j})$, respectively. Recall that $w_i \in [0, 1]$ and we denote $\sum_{i=1}^n w_i^k = M^k \leq n, \forall i \in [1, 2, \cdots, n]$. Besides, we assume $\tau_i^k \in [1, \tau], \forall i \in \{1, ..., n\}, k \in [0, K]$. To avoid being overly complicated, we omit superscripts or subscripts for some expressions.

The Frobenius norm for matrix $Z_{p\times q}$ is:

$$\|Z\|_F^2 = \left|Tr(ZZ^\top)\right| = \sum_{i=1}^{p}\sum_{j=1}^{q}|z_{i,j}|^2 = \sum_{j=1}^{q}\|\boldsymbol{Z_j}\|^2, \tag{12}$$

where $\boldsymbol{Z_j}$ is the j-th column vector of matrix Z. And the operator norm for $Z_{p\times q}$ is:

$$\|Z\|_{op} = \max_{\|x\|=1} \|Zx\| = \sqrt{\lambda_{\max}(Z^\top Z)} \tag{13}$$

where $\lambda_{\max}$ is the maximal eigenvalue of Z. From Lemma 7 of [27], we have the following conclusion: suppose $Z_{p\times q}$, $D_{q\times q}$ are real matrices and D is symmetric, then we have:

$$\|ZD\|_F \leq \|Z\|_F \, \|D\|_{op} \tag{14}$$

Besides, we can directly derive some intuitive equations that can simplify the subsequent proofs.

$$\begin{aligned} E\left[\left\|\sum_{i=1}^{n} w_i X_i^k - E\left[\sum_{i=1}^{n} w_i X_i^k\right]\right\|^2\right] &= E\left\|\sum_{i=1}^{n} w_i X_i^k\right\|^2 + \left(E\left[\sum_{i=1}^{n} w_i X_i^k\right]\right)^2 \\ &\quad - 2E\left[\left(\sum_{i=1}^{n} w_i X_i^k\right) E\left[\sum_{i=1}^{n} w_i X_i^k\right]\right] \\ &= E\left[\left\|\sum_{i=1}^{n} w_i X_i^k\right\|^2\right] - \left(E\left[\sum_{i=1}^{n} w_i^k X_i^k\right]\right)^2 . \end{aligned} \tag{15}$$

Based on the definition of $\mathbf{X}^k$ and $\mathbf{Y}^k$, under assumption 2 or assumption 3, we have

$$\mathbb{E}\left[\mathbf{X}^k\right] = \mathbb{E}\left[\mathbf{Y}^k\right] = \mathbf{Y}^k, \tag{16}$$

and

$$\mathbb{E}_{p\neq q}\langle g_p(\theta_p) - \nabla F_p(\theta_p), g_q(\theta_q) - \nabla F_q(\theta_q)\rangle = 0, \tag{17}$$

Further, we can derive

$$\mathbb{E}_{p\neq q}\left[\langle \mathbf{X}_p^k - \mathbf{Y}_p^k, \mathbf{X}_q^k - \mathbf{Y}_q^k \rangle\right] = 0. \tag{18}$$

Lemma 1 bounds the variance of weighted sum stochastic gradients w.r.t. weighted sum full batch gradients at round k.

Lemma 1. Under Assumptions 1, 3 and 4 in the non-*i.i.d.* setting, the variance of the weighted sum of mini-batch gradients is bounded by

$$E\left[\left\|\sum_{i=1}^{n} w_i X_i^k - \sum_{i=1}^{n} w_i Y_i^k\right\|^2\right] \leq \mu \sum_{i=1}^{n} w_i^2 \sum_{j=1}^{\tau_i^k} \left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2 + \sigma^2 \sum_{i=1}^{n} w_i^2. \tag{19}$$

Proof.

$$
\begin{aligned}
& E\left[\left\|\sum_{i=1}^{n} w_i X_i^k-\sum_{i=1}^{n} w_i Y_i^k\right\|^2\right] \\
= & E\left[\sum_{i=1}^{n} w_i^2\left(X_i^k-Y_i^k\right)^2+\sum_{p \neq q} w_p w_q\left\langle X_p^k-Y_p^k, X_q^k-Y_q^k\right\rangle\right] \\
= & \sum_{i=1}^{n} w_i^2 E\left\|X_i^k-Y_i^k\right\|^2 \\
= & \sum_{i=1}^{n} w_i^2 E\left\|\sum_{j=1}^{\tau_i^k}\left(g_i\left(\theta_i^{k, j}\right)-\nabla F_i\left(\theta_i^{k, j}\right)\right)\right\|^2 \\
= & \sum_{i=1}^{n} w_i^2 E\left[\sum_{j=1}^{\tau_i^k}\left(g_i\left(\theta_i^{k, j}\right)-\nabla F_i\left(\theta_i^{k, j}\right)\right)^2\right. \\
& \qquad \left.+\sum_{p \neq q}\left\langle g_i\left(\theta_i^{k, p}\right)-\nabla F_i\left(\theta_i^{k, p}\right), g_i\left(\theta_i^{k, q}\right)-\nabla F_i\left(\theta_i^{k, q}\right)\right\rangle\right] \\
= & \sum_{i=1}^{n} w_i^2 E\left[\sum_{j=1}^{\tau_i^k}\left(g_i\left(\theta_i^{k, j}\right)-\nabla F_i\left(\theta_i^{k, j}\right)\right)^2\right] \\
\leq & \sum_{i=1}^{n} w_i^2 \sum_{j=1}^{\tau_i^k}\left[\mu\left\|\nabla F_i\left(\theta_i^{k, j}\right)\right\|^2+\sigma^2\right] \\
= & \mu \sum_{i=1}^{n} w_i^2 \sum_{j=1}^{\tau_i^k}\left\|\nabla F_i\left(\theta_i^{k, j}\right)\right\|^2+\sigma^2 \sum_{i=1}^{n} w_i^2 \\
\leq & \mu \sum_{i=1}^{n} w_i^2 \sum_{j=1}^{\tau_i^k}\left\|\nabla F_i\left(\theta_i^{k, j}\right)\right\|^2+\sigma^2\left(M^k\right)^2 .
\end{aligned}
\tag{20}
$$

Lemma 2. Under assumption 1, 3 and 4 in the non-$i.i.d.$ setting, the expected weighted sum of mini-batch gradients is bounded by

$$
E\left\|\sum_{i=1}^{n} w_i X_i^k\right\|^2 \leq \mu \sum_{i=1}^{n} w_i^2 \sum_{i=1}^{\tau_i^k}\left\|\nabla F_i\left(\theta_i^{k, j}\right)\right\|^2+\sigma^2 \sum_{i=1}^{n} w_i^2+\left\|\sum_{i=1}^{n} w_i Y_i^k\right\|^2 \tag{21}
$$

According to equation (15), (16) and the definition of $\mathbf{X}^{(k)}$, we have

$$\begin{aligned}
&E\left\|\sum_{i=1}^{n} w_i X_i^k\right\|^2 \\
=&E\left[\left\|\sum_{i=1}^{n} w_i X_i^k - E\left[\sum_{i=1}^{n} w_i X_i^k\right]\right\|^2\right] + \left(E\left[\sum_{i=1}^{n} w_i X_i^k\right]\right)^2 \\
=&E\left[\left\|\sum_{i=1}^{n} w_i X_i^k - \sum_{i=1}^{n} w_i Y_i^k\right\|^2\right] + \left\|\sum_{i=1}^{n} w_i Y_i^k\right\|^2 \\
\leqslant&\mu \sum_{i=1}^{n} w_i^2 \sum_{i=1}^{\tau_i^k} \left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2 + \sigma^2 \sum_{i=1}^{n} w_i^2 + \left\|\sum_{i=1}^{n} w_i Y_i^k\right\|^2 \\
\leq&\mu \sum_{i=1}^{n} w_i^2 \sum_{j=1}^{\tau_i^k} \left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2 + \sigma^2 \left(M^k\right)^2 + \left\|\sum_{i=1}^{n} w_i Y_i^k\right\|^2 .
\end{aligned}$$

Proposition 1. Under assumption 2 in the *i.i.d.* setting and assumption 3 in the non-*i.i.d.* setting, we can obtain the same expected inner product between the weighted sum of stochastic gradients and the full-batch gradients as

$$\begin{aligned}
E\left[\left\langle \nabla F\left(\bar{\theta}^k\right), \sum_{i=1}^{n} w_i X_i^k\right\rangle\right] &= E\left[\sum_{i=1}^{n} w_i \left\langle \nabla F\left(\bar{\theta}^k\right), X_i^k\right\rangle\right] \\
&= E\left[\left\langle \nabla F\left(\bar{\theta}^k\right), \sum_{i=1}^{n} w_i Y_i^k\right\rangle\right] \\
&= \frac{1}{2}\left\|\nabla F\left(\bar{\theta}^k\right)\right\|^2 + \frac{1}{2}\left\|\sum_{i=1}^{n} w_i Y_i^k\right\|^2 \\
&\quad - \frac{1}{2} E\left\|\nabla F\left(\bar{\theta}^k\right) - \sum_{i=1}^{n} w_i Y_i^k\right\|^2 .
\end{aligned} \tag{22}$$

The last equation is due to $2 < a, b >= ||a||^2 + ||b||^2 - ||a-b||^2$.

Lemma 3. Under assumption 3 and 4 in the non-*i.i.d.* setting, we can obtain the variance upper bound between the global gradient and the weighted sum of local gradients as

$$\begin{aligned}
E\left\|\nabla F\left(\bar{\theta}^k\right) - \sum_{i=1}^{n} w_i Y_i^k\right\|^2 =& \frac{2}{n} \sum_{i=1}^{n} \frac{1}{\tau_i^k} \sum_{j=1}^{\tau_i^k} E\left\|\nabla F_i\left(\bar{\theta}^k\right) - \nabla F_i\left(\theta_i^{k,j}\right)\right\|^2 \\
&+ 2A\left(\left\|Y_i^k\right\|^2; n, \tau_i^k, w_i\right),
\end{aligned} \tag{23}$$

where $A\left(\left\|Y_i^k\right\|^2; n, \tau_i^k, w_i\right) = 2E\left\|\sum_{i=1}^{n}\left[\frac{1}{n}\frac{1}{\tau_i^k} - w_i\right] Y_i^k\right\|^2$. When w_i closes to $\frac{1}{n\tau_i^k}$, A can be minimized.

$$
\begin{aligned}
& E\left\|\nabla F\left(\bar{\theta}^k\right)-\sum_{i=1}^n w_i Y_i^k\right\|^2 \\
= & E[\sum_{i=1}^n \frac{1}{n} \frac{1}{\tau_i^k} \sum_{j=1}^{\tau_j^k} \nabla F_i\left(\bar{\theta^k}\right)-\sum_{i=1}^n \frac{1}{n} \frac{1}{\tau_i^k} \sum_{j=1}^{n_i^k} \nabla F_i\left(\theta_i^k\right)+ \\
& \sum_{i=1}^n \frac{1}{n} \frac{1}{\tau_i^k} Y_i^k-\sum_{i=1}^n w_i Y_i^k]^2. \\
\leq & 2 E\left\|\sum_{i=1}^n \frac{1}{n} \frac{1}{\tau_i^k} \sum_{j=1}^{\tau_i^k}\left[\nabla F_i\left(\bar{\theta^k}\right)-\nabla F_i\left(\theta_i^{k, j}\right)\right]\right\|^2+ \\
& 2 E\left\|\sum_{i=1}^n\left[\frac{1}{n} \frac{1}{\tau_i^k}-w_i\right] Y_i^k\right\|^2 \\
\leq & 2 E\left[\sum_{i=1}^n \frac{1}{n}\left\|\frac{1}{\tau_i^k} \sum_{j=1}^{\tau_i^k}\left[\nabla F_i\left(\bar{\theta^k}\right)-\nabla F_i\left(\theta_i^{k, j}\right)\right]\right\|^2\right]+ \\
& 2 A\left(\left\|Y_i^k\right\|^2 ; n, \tau_i^k, w_i\right) \\
\leq & \frac{2}{n} \sum_{i=1}^n \frac{1}{\tau_i^k} \sum_{j=1}^{\tau_i^k} E\left\|\nabla F_i\left(\bar{\theta^k}\right)-\nabla F_i\left(\theta_i^{k, j}\right)\right\|^2+2 A\left(\left\|Y_i^k\right\|^2 ; n, \tau_i^k, w_i\right) .
\end{aligned}
\tag{24}
$$

The first inequality is obtained by $< a, b > \leq ||a||^2 + ||b||^2$, while the last two inequalities is derived by Jensen's inequality.

A.4 Proof of Theorems

Here we first state Theorem 3.

Theorem 3. *Suppose the same condition as Theorem 2, we can reduce the convergence upper bound by tuning the aggregation weights. If we define $w_i \to \frac{1}{n\tau_i^k}$, then the expected gradient norm after K iterations is bounded by:*

$$
\begin{aligned}
\mathbb{E}[\frac{1}{K}\sum_{k=1}^K ||\nabla F(\bar{\theta}^k)||^2] \leq & \frac{4\left(E\left[F\left(\bar{\theta}^1\right)-E\left[F\left(\bar{\theta}^K\right)\right]\right)\right.}{K\eta} \\
& +4\mathcal{O}\left(\bar{A}+C+D+E+F+\mu\eta C\sum_{k=0}^K \frac{1}{K}\sum_{i=1}^n w_i^2\tau_i^k\right).
\end{aligned}
\tag{25}
$$

Theorem 2 Under Assumptions 1, 3 and 4 in the non-*i.i.d.* setting, the expected weighted sum of mini-batch gradients is bounded by

$$\frac{4\left(E\left[F\left(\bar{\theta}^{1}\right)-E\left[F\left(\bar{\theta}^{k}\right)\right]\right)}{K\eta}+4\left(\bar{A}+C+D+E+F+\mu\eta C\sum_{k=0}^{K}\frac{1}{K}\sum_{i=1}^{n}w_i^2\tau_i^k\right) \tag{26}$$

Based on the Lipschitz smoothness, we can obtain an intermediate result

$$\begin{aligned}
&E\left[F\left(\bar{\theta}^{k+1}\right)\right]-E\left[F\left(\bar{\theta}^{k}\right)\right]\\
\leq&E\left[<\nabla F\left(\bar{\theta}^{k}\right),\bar{\theta}^{k+1}-\bar{\theta}^{k}>\right]+\frac{L}{2}E\left\|\bar{\theta}^{k+1}-\bar{\theta}^{k}\right\|^2\\
\leq&-\eta E\left[\left\langle\nabla F\left(\bar{\theta}^{k}\right),\sum_{i=1}^{n}w_iX_i^k\right\rangle\right]+\frac{L}{2}\eta^2E\left\|\sum_{i=1}^{n}w_iX_i^k\right\|^2\\
\leq&-\frac{\eta}{2}\left\|\nabla F\left(\bar{\theta}^{k}\right)\right\|^2-\frac{\eta}{2}\left\|\sum_{i=1}^{n}w_iY_i^k\right\|^2+\frac{\eta}{2}E\left\|\nabla F\left(\bar{\theta}^{k}\right)-\sum_{i=1}^{n}w_iY_i^k\right\|^2\\
&+\frac{\mu L}{2}\eta^2\sum_{i=1}^{n}w_i^2\sum_{j=1}^{\tau_i^k}\left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2+\frac{L}{2}\eta^2\sigma^2\sum_{i=1}^{n}w_i^2+\frac{L\eta^2}{2}\left\|\sum_{i=1}^{n}w_iY_i^k\right\|^2\\
\leq&-\frac{\eta}{2}\left\|\nabla F\left(\bar{\theta}^{k}\right)\right\|^2+\frac{\eta}{2}(L\eta-1)\left\|\sum_{i=1}^{n}w_iY_i^k\right\|^2\\
&+\eta\cdot\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\tau_i^k}\sum_{j=1}^{\tau_i^k}E\left\|\nabla F_i\left(\bar{\theta}^{k}\right)-\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2+\eta E\left\|\sum_{i=1}^{n}\left[\frac{1}{n\tau_i^k}-w_i\right]Y_i^k\right\|^2\\
&+\frac{\mu L}{2}\eta^2\sum_{i=1}^{n}w_i^2\sum_{j=1}^{\tau_i^k}\left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2+\frac{L}{2}\eta^2\sigma^2\sum_{i=1}^{n}w_i^2
\end{aligned}$$

As for $E\left\|\nabla F_i\left(\bar{\theta}^{k}\right)-\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2$, due to the Lipschitz smoothness again, we have

$$\begin{aligned}
&E\left\|\nabla F_i\left(\bar{\theta}^{k}\right)-\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2\\
\leq&L^2E\left\|\bar{\theta}^{k}-\theta_i^{k,j}\right\|^2\\
=&L^2\eta^2E\left\|\sum_{s=1}^{j}g_i\left(\theta_i^{k,s}\right)\right\|^2\\
\leq&2L^2\eta^2E\left\|\sum_{s=1}^{j}\left[g_i\left(\theta_i^{k,s}\right)-\nabla F_i\left(\theta_i^{k,s}\right)\right]\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&+ 2L^2\eta^2 E\left\|\sum_{s=1}^{j} \nabla F_i\left(\theta_i^{k,s}\right)\right\|^2 \\
&= 2L^2\eta^2 E\left[\sum_{s=1}^{j}\left[g_i\left(\theta_i^{k,s}\right) - \nabla F_i\left(\theta_i^{k,s}\right)\right]^2\right] \\
&+ 2L^2\eta^2 E\left\|\sum_{s=1}^{j} \nabla F_i\left(\theta_i^{k,s}\right)\right\|^2 \\
&\leq 2L^2\eta^2 \sum_{s=1}^{j}\left[\mu\left\|\nabla F_i\left(\theta_i^{k,s}\right)\right\|^2 + \sigma^2\right] \\
&+ 2L^2\eta^2 jE\left[\sum_{s=1}^{j}\left\|\nabla F_i\left(\theta_i^{k,s}\right)\right\|^2\right] \\
&\leq 2L^2\eta^2\sigma^2 + \left(2\mu L^2\eta^2 + 2jL^2\eta^2\right)\sum_{j=1}^{\tau_i^k} E\left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2
\end{aligned}
$$

Based on the above expressions, we can obtain

$$
E\left\|\bar{\theta}^k - \theta_i^{k,j}\right\|^2 \leq 2\eta^2\sigma^2 + \left(2\mu\eta^2 + 2j\eta^2\right)\sum_{j=1}^{\tau_i^k} E\left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2, \tag{27}
$$

In addition,

$$
\begin{aligned}
\left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2 &\leq 2\left\|\nabla F_i\left(\theta_i^{k,j}\right) - \nabla F_i\left(\bar{\theta}^k\right)\right\|^2 + 2\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2 \\
&\leq 2L^2\left\|\theta_i^{k,j} - \bar{\theta}^k\right\|^2 + 2\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2
\end{aligned} \tag{28}
$$

Take (32) back to (31), then

$$
E\left\|\bar{\theta}^k - \theta_i^{k,j}\right\|^2 \leq 2\eta^2\sigma^2 + 2\eta^2(\mu + j)\sum_{j=1}^{\tau_i^k} E\left[2L^2\left\|\theta_i^{k,j} - \bar{\theta}^k\right\|^2 + 2\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2\right] \tag{29}
$$

Take the sum within two communication rounds,

$$\begin{aligned}
&\sum_{j=1}^{\tau_i^k} E\left\|\bar{\theta}^k-\theta_i^{k,j}\right\|^2 \\
\leq & \tau_i^k \eta^2 \sigma^2+2 \eta^2\left(\mu \tau_i^k+\frac{\tau_i^k\left(1+\tau_i^k\right)}{2}\right) \sum_{j=1}^{\tau_i^k} E\left[2 L^2\left\|\theta_i^{k,j}-\bar{\theta}^k\right\|^2+2\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2\right] \\
= & 2 \tau_i^k \eta^2 \sigma^2+2 L^2 \eta^2\left[2 \mu \tau_i^k+\tau_i^k\left(1+\tau_i^k\right)\right] \sum_{j=1}^{\tau_i^k} E\left\|\theta_i^{k,j}-\bar{\theta}^k\right\|^2 \\
+ & 2 \eta^2\left[2 \mu \tau_i^k+\tau_i^k\left(1+\tau_i^k\right)\right] \cdot \tau_i^k\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2 .
\end{aligned} \tag{30}$$

After minor rearranging, we derive

$$\begin{aligned}
\left[1-2 L^2 \eta^2 \tau_i^k\left(2 \mu+1+\tau_i^k\right)\right] \sum_{j=1}^{\tau_i^k} E\left\|\bar{\theta}^k-\theta_i^k\right\|^2 \leq & \quad 2 \tau_i^k \eta^2 \sigma^2 \\
+2 \eta^2\left(\tau_i^k\right)^2\left(2 \mu+1+\tau_i^k\right)\left\|\nabla F_i\left(\theta^k\right)\right\|^2 . &
\end{aligned} \tag{31}$$

If we define $B_i^k=2 L^2 \eta^2 \tau_i^k\left[2 \mu+1+\tau_i^k\right] \leqslant 2 L^2 \eta^2 \tau(2 \mu+1+\tau):=B$, then it follows that,

$$(1-B_i^k) \sum_{j=1}^{\tau_i^k} E\left\|\bar{\theta}^k-\theta_i^{k,j}\right\|^2 \leqslant 2 \tau_i^k \eta^2 \sigma^2+\frac{\tau_i^k}{L^2} \cdot B_i^k\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2, \tag{32}$$

$$\begin{aligned}
\frac{L^2}{\tau_i^k} \sum_{j=1}^{\tau_i^k} E\left\|\bar{\theta}^k-\theta_i^{k,j}\right\|^2 & \leqslant \frac{2 \eta^2 \sigma^2 L^2}{1-B_i^k}+\frac{B_i^k}{1-B_i^k}\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2 \\
& \leqslant \frac{2 \eta^2 \sigma^2 L^2}{1-B}+\frac{B}{1-B}\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2 .
\end{aligned} \tag{33}$$

Take (37) back to (30), then

$$\begin{aligned}
\frac{1}{\tau_i^k} \sum_{j=1}^{\tau_i^k} E\left\|\nabla F_i\left(\bar{\theta}^k\right)-\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2 & \leq \frac{L^2}{\tau_i^k} \sum_{j=1}^{\tau_i^k} E\left\|\bar{\theta}^k-\theta_i^{k,j}\right\|^2 \\
& \leq \frac{2 \eta^2 \sigma^2 L^2}{1-B}+\frac{B}{1-B}\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2 .
\end{aligned} \tag{34}$$

With the help of (32) and (37), we can further obtain

$$\begin{aligned}
\sum_{j=1}^{\tau_i^k}\left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2 &\leqslant 2L^2\sum_{j=1}^{\tau_i^k}\left\|\theta_i^{k,j}-\bar{\theta}^k\right\|^2+2\tau_i^k\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2 \\
&\leqslant 2\tau_i^k\left(\frac{2\eta^2\sigma^2L}{1-B}+\frac{B}{1-B}\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2\right)+2\tau_i^k\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2 \\
&=\frac{4\tau_i^kL\eta^2\sigma^2}{1-B^2}+\frac{2\tau_i^k}{1-B}\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2
\end{aligned} \tag{35}$$

Further,

$$\begin{aligned}
&\frac{\mu L\eta}{2}\sum_{i=1}^{n}w_i^2\sum_{j=1}^{\tau_i^k}\left\|\nabla F_i\left(\theta_i^{k,j}\right)\right\|^2 \\
\leq&\frac{\mu L\eta}{2}\sum_{i=1}^{n}w_i^2\left(\frac{4\tau_i^kL\eta^2\sigma^2}{1-B}+\frac{2\tau_i^k}{1-B}\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2\right) \\
\leq&\frac{2\mu L^2\sigma^2\eta^3}{1-B}\sum_{i=1}^{n}w_i^2\tau_i^k+\frac{\mu L\eta\tau}{1-B}\sum_{i=1}^{n}w_i^2\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2
\end{aligned} \tag{36}$$

Take (38) and (40) to the intermediate result (29), and if $L\eta \leq 1$,

$$\begin{aligned}
&\frac{E\left[F\left(\bar{\theta}^{k+1}\right)\right]-E\left[F\left(\bar{\theta}^k\right)\right]}{\eta} \\
\leq&-\frac{1}{2}\left\|\nabla F\left(\bar{\theta}^k\right)\right\|^2+\frac{1}{n}\sum_{i=1}^{n}\left[\frac{2\eta^2\sigma^2L^2}{1-B}+\frac{B}{1-B}\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2\right] \\
&+E\left\|\sum_{i=1}^{n}\left(\frac{1}{n\tau_i^k}-w_i\right)Y_i^k\right\|^2+\frac{2\mu L^2\sigma^2\eta^3}{1-B}\sum_{i=1}^{n}w_i^2\tau_i^k \\
&+\frac{\mu L\eta\tau}{1-B}\sum_{i=1}^{n}w_i^2\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2+\frac{L\eta\sigma^2}{2}\sum_{i=1}^{n}w_i^2 \\
=&-\frac{1}{2}\left\|\nabla F\left(\bar{\theta}^k\right)\right\|^2+\frac{2\eta^2\sigma^2L^2}{1-B}+\frac{B}{1-B}\sum_{i=1}^{n}\frac{1}{n}\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2 \\
&+E\left\|\sum_{i=1}^{n}\left(\frac{1}{n\tau_i^k}-w_i\right)Y_i^k\right\|^2+\frac{\mu L\eta\tau}{1-B}\sum_{i=1}^{n}w_i^2\left\|\nabla F_i\left(\bar{\theta}^k\right)\right\|^2 \\
&+\frac{2\mu L^2\sigma^2\eta^3}{1-B}\sum_{i=1}^{n}w_i^2\tau_i^k+\frac{L\eta\sigma^2}{2}\sum_{i=1}^{n}w_i^2
\end{aligned} \tag{37}$$

With the assumption 4 of bounded dissimilarity, we can further simplify the above expression

$$\begin{aligned}
&\frac{E\left[F\left(\bar{\theta}^{k+1}\right)\right]-E\left[F\left(\bar{\theta}^{k}\right)\right]}{\eta} \\
\leq&\frac{2\mu L\eta\tau\beta^2+2B\beta^2-(1-B)}{2(1-B)}\left\|\nabla F\left(\bar{\theta}^{k}\right)\right\|^2+\frac{2\eta^2\sigma^2L^2}{1-B} \\
+&\frac{BK^2+\mu L\eta\tau\kappa^2}{1-B}+E\left\|\sum_{i=1}^{n}\left(\frac{1}{n\tau_i^k}-w_i\right)Y_i^k\right\|^2 \\
+&\frac{2\mu L^2\sigma^2\eta^3}{1-B}\sum_{i=1}^{n}w_i^2\tau_i^k+\frac{L\eta\sigma^2}{2}\sum_{i=1}^{n}w_i^2
\end{aligned} \tag{38}$$

If $\frac{\mu L\eta\tau\beta^2+2B\beta^2}{1-B}\leqslant\frac{1}{2}$, then $2\left(\mu L\eta\tau\beta^2+2B\beta^2\right)\leq 1-B$. Next, we take the average across all communication rounds, we obtain

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\left\|\nabla F\left(\bar{\theta}^{k}\right)\right\|^2\leq&\frac{4\left(E\left[F\left(\bar{\theta}^{1}\right)-E\left[F\left(\bar{\theta}^{k}\right)\right]\right)\right.}{K\eta} \\
&+4\left(\bar{A}+C+D+E+F+\mu\eta C\sum_{k=0}^{K}\frac{1}{K}\sum_{i=1}^{n}w_i^2\tau_i^k\right)
\end{aligned} \tag{39}$$

where

$$\begin{aligned}
&\bar{A}=\frac{1}{K}\sum_{i=1}^{K}A,\quad C=\frac{\eta^2\sigma^2L^2}{\mu L\eta\tau\beta^2+2B\beta^2},\quad D=\frac{\left(1-2\mu L\eta\tau\beta^2\right)\kappa^2}{\left(2\mu L\eta\tau\beta^2+4B\beta^2\right)\left(1+4\beta^2\right)}, \\
&E=\frac{\mu L\eta\tau\kappa^2}{2\mu L\eta\tau\beta^2+4B\beta^2},\quad F=\frac{L\eta\sigma^2}{2K}\sum_{k=1}^{K}\left(M^k\right)^2
\end{aligned}.$$

Intuitively, we can find that the convergence upper bound increases along with the value β^2, L, σ^2, κ^2. They are all parameters related to the quality of local objectives, local gradients, and local stochastic gradients. In addition, by applying $F_1(\cdot)=F_2(\cdot)=\cdots=F_n(\cdot)=F(\cdot)$ and $w_i=\frac{1}{n\tau_i^k}$, we can easily obtain the result in Theorem 1 and Theorem 3, respectively.

A.5 Discussion about System Utility

System Utility To provide a comprehensive evaluation for fair comparisons, we propose a general utility function that reconciles both theoretical and practical considerations. Specifically, from the perspective of numerical performance, both task-oriented performance and system efficiency are crucial metrics. We denote them as Q_s and ψ, respectively. For instance, in a multi-agent navigation environment, Q_s can be a composite objective comprised of navigation success rate, average speed, and safety while ψ refers to the system communication and computation cost. As for the convergence property, we consider an ideal setting where the agents are *i.i.d.* It serves as the optimal convergence upper bound ϵ_m. By comparing the convergence bound with it, we can tell the tightness of the federated MARL method. Thus, we derive our system utility function as $Q_{tot} = \frac{Q_s - \lambda\psi}{e^{\|\epsilon - \epsilon_m\|^2}}$, where λ is a positive constant used to balance the importance of system cost and performance.

A.6 Baseline Methods

IPPO [30] demonstrates an intuitive application of single-agent RL methods into multi-agent systems by sharing the parameters of agents' actors and critics. In the federated learning setting, the sharing of these parameters happens during the communication between the server and the clients. We also notice a similar potential baseline, MAPPO [32], the main difference between IPPO and MAPPO is the input to the critics. While IPPO only takes the local observation as input, MAPPO additionally requires the global state. It is hard to adapt to a federated learning setting.

RIAL and DIAL [8] are strong baselines for MARL communication. They both incorporate the last communication signals from other agents as additional observation to assist policy learning and produce communication messages to maintain the scheme. Thus, the communication messages are step-wise. RIAL can work in CTDE or a fully decentralized manner, and in our implementation, we choose the latter serving as a decentralized baseline with communication. As for DIAL, it enables differentiable communication by maintaining the gradients with respect to the network parameters and the communication signals, which means that each local update requires the other agents' accumulated gradients.

CoPO [21] achieves state-of-the-art performance on the original MetaDrive multi-agent tasks. It employs a meta-learning approach to bundle local policy learning with global reward optimization. Owing to its unique characteristics, we only changed it into a non-parameter sharing scheme. It totally has four networks in its original version.

FMARL [31] is a strong baseline in the domain of federated MARL, which is also experimentally evaluated in a multi-vehicle autonomous driving simulation benchmark. We notice that there are two methods proposed in [31]. The first one can be regarded as federated IPPO with weight decay while the second method requires agent-to-agent communication, which is not compatible with our pure client-server setting, so we implement the first method as our baseline.

A.7 More Experiments

The cooperative navigation results are exhibited in Tab. 2 and Tab. 3, while the cooperative exploration results are reported in Tab. 4 and Tab. 5.

In the first three scenarios in Tab. 2, FMRL-LA has achieved the best system utility. RIAL, due to its fully decentralized training paradigm and unique communication mechanism, reaches the highest communication efficiency. However, the

Table 2. For cooperative navigation tasks, the detailed system performance and efficiency on the first three scenarios

scenarios	scenario1					scenario2					scenario3				
methods\metrics	success	safety	speed	ψ_1	utility	success	safety	speed	ψ_1	utility	success	safety	speed	ψ_1	utility
IPPO	0.487	0.621	0.382	0.513	0.501	0.586	0.308	0.449	0.435	0.445	0.507	0.546	0.439	0.586	0.520
RIAL	0.588	0.362	0.412	**0.632**	0.499	0.473	0.283	0.365	**0.610**	0.433	0.601	0.616	0.380	**0.620**	0.554
DIAL	0.665	0.431	0.496	0.227	0.455	0.549	0.512	0.580	0.251	0.473	**0.696**	0.652	0.493	0.264	0.526
CoPO	0.623	**0.678**	0.530	0.372	0.551	0.679	0.464	0.516	0.379	0.510	0.574	0.477	0.573	0.386	0.503
FMARL	0.699	0.513	0.511	0.508	0.558	0.577	0.378	**0.643**	0.461	0.514	0.490	0.637	**0.649**	0.566	0.586
FMRL-LA	**0.737**	0.650	**0.548**	0.456	**0.598**	**0.754**	**0.574**	0.617	0.548	**0.623**	0.653	**0.716**	0.644	0.513	**0.632**

success rate, safety, and speed of RIAL in the three scenarios are not very high. It proves that currently, fully decentralized training in MAS is extra difficult.

Table 3. For cooperative navigation tasks, the detailed system performance and efficiency on the last three scenarios

scenarios	scenario4					scenario5					scenario6				
methods\metrics	success	safety	speed	ψ_1	utility	success	safety	speed	ψ_1	utility	success	safety	speed	ψ_1	utility
IPPO	0.425	0.240	0.213	0.483	0.340	0.376	0.400	0.250	0.372	0.350	0.399	0.259	0.296	0.437	0.348
RIAL	0.322	0.371	0.318	**0.643**	0.414	0.214	0.228	0.317	**0.558**	0.329	0.271	0.226	0.212	**0.593**	0.326
DIAL	0.457	0.594	0.465	0.202	0.430	0.445	0.397	0.380	0.195	0.354	0.524	0.506	**0.541**	0.153	0.431
CoPO	**0.664**	0.619	**0.631**	0.380	**0.574**	0.485	0.469	**0.475**	0.485	0.479	0.513	0.576	0.533	0.269	0.473
FMARL	0.519	0.529	0.524	0.447	0.505	0.391	0.301	0.370	0.350	0.353	0.460	0.489	0.481	0.418	0.462
FMRL-LA	0.632	**0.644**	0.621	0.395	0.573	**0.553**	**0.507**	0.448	0.504	**0.503**	**0.611**	**0.639**	0.504	0.502	**0.564**

In Tab. 3, these three scenarios are more difficult than the former three. We can tell it from the performance of the methods. In particular, the average speed of all methods in scenarios 5 and 6 is relatively low.

Table 4. For cooperative exploration tasks, the detailed system performance and efficiency on the first three scenarios

scenarios	scenario1					scenario2					scenario3				
methods\metrics	explore	safety	speed	ψ_1	utility	explore	safety	speed	ψ_1	utility	explore	safety	speed	ψ_1	utility
IPPO	0.425	0.309	0.422	0.559	0.429	0.417	0.297	0.319	0.420	0.363	0.484	0.398	0.500	0.424	0.452
RIAL	0.431	0.417	0.332	**0.678**	0.465	0.619	0.356	0.179	**0.650**	0.451	0.523	0.260	0.412	**0.706**	0.475
DIAL	0.705	0.507	0.557	0.337	0.523	**0.707**	0.493	0.356	0.176	0.433	0.651	0.552	0.521	0.275	0.500
CoPO	0.581	0.606	**0.628**	0.472	0.572	0.564	0.489	**0.560**	0.398	0.503	0.593	0.577	0.580	0.379	0.532
FMARL	0.535	0.515	0.523	0.561	0.534	0.504	0.538	0.469	0.479	0.498	0.669	0.407	0.561	0.487	0.531
FMRL-LA	**0.714**	**0.645**	0.617	0.519	**0.624**	0.687	**0.598**	0.528	0.500	**0.578**	**0.696**	**0.659**	**0.664**	0.467	**0.622**

In Tab. 4, by comparing the performance with cooperative navigation on the same scenarios in Tab. 2, we observe that on the two tasks, our method performs robustly with respect to all evaluation metrics.

In Tab. 5, comparing the performance of our method with other baselines, we find that FMRL-LA suffers less from the complexity of the maps. In addition, we find the baseline CoPO, which is the state-of-the-art on the original MetaDrive also performs well in these complex scenarios. We believe it is because of the explicit modeling of the neighbor agents.

Table 5. For cooperative exploration tasks, the detailed system performance and efficiency on the last three scenarios

scenarios	scenario4					scenario5					scenario6				
methods\metrics	explore	safety	speed	ψ_1	utility	explore	safety	speed	ψ_1	utility	explore	safety	speed	ψ_1	utility
IPPO	0.528	0.477	0.564	0.619	0.547	0.117	0.122	0.330	0.608	0.294	0.513	0.228	0.438	0.564	0.436
RIAL	0.353	0.238	0.280	**0.710**	0.395	0.273	0.162	0.213	**0.671**	0.330	0.314	0.375	0.280	**0.677**	0.412
DIAL	0.432	0.312	0.490	0.316	0.388	**0.547**	0.385	0.283	0.202	0.354	0.366	0.440	0.451	0.308	0.391
CoPO	**0.717**	**0.594**	0.591	0.487	0.597	0.528	**0.481**	**0.474**	0.461	0.486	0.603	0.519	0.501	0.438	0.515
FMARL	0.657	0.416	0.538	0.646	0.564	0.358	0.185	0.349	0.578	0.368	0.593	0.334	0.460	0.549	0.484
FMRL-LA	0.699	0.548	**0.660**	0.558	**0.616**	0.531	0.467	0.428	0.521	**0.487**	**0.629**	**0.567**	**0.533**	0.599	**0.582**

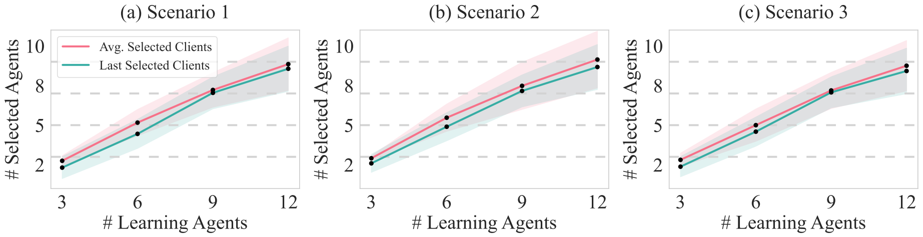

Fig. 6. Number of selected clients under different numbers of learning agents for the three scenarios

Client Selection Analysis To investigate the effectiveness of the learnable aggregation on the perspective of client selection during client-server communication, we conduct experiments on cooperative navigation tasks in scenarios 1, 2, and 3 with different numbers of learning agents. The results are shown in Fig. 6. The average number of selected agents is calculated during the evaluation intervals, which is consistent with the calculation of all the evaluation metrics. Thus, it reflects the number of agents selected for the overall training phase. The less the agents are selected, the higher the communication efficiency we achieve. When there are only 3 agents in the scenarios, due to the partial observability and the complexity of the tasks, the agents require a longer time to learn stable policies. Besides, each agent's gradients are important to the system. Thus, the learnable aggregation module does not select agents' gradients frequently and may pay more attention to weighting the gradients toward a better collective policy learning. However, as the number of agents increases in the same tasks, the effect of selection becomes more significant.

Table 6. Hyperparameter settings for experiments

Hyperparameters	Value	Hyperparameters	Value
Critic lr	5e-4	Hidden layer	1
Activation	ReLU	Hideen layer dim	32
GAE lambda	0.95	Number of random seeds	5
Gamma	0.99	Network initialization	Orthogonal
Hypernet embed	32	Maximal environment steps for each trial	[1M, 5M]
Batch size	1024	Maximal local updates τ	[5, 10]
Mini batch size	512	Maximal communication round K	maximal environment steps / maximal local updates
Optimizer	Adam	Optimizer episilon	1e-5