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 θ_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]:

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

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.

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

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:*

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

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,

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

After minor rearranging, we derive

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

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

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

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

Take (37) back to (30), then

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

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

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

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**

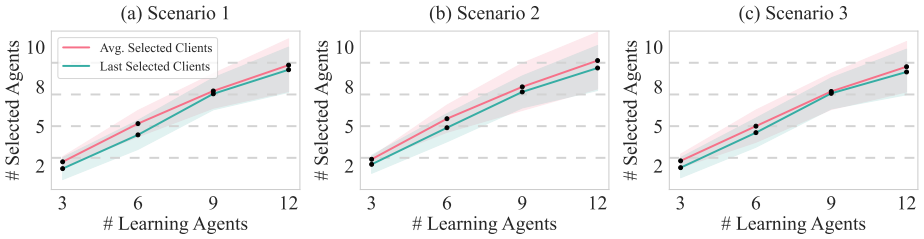

Fig. 6. Number of selected clients under different numbers of learning agents for the three scenarios

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