

A APPENDIX

A.1 FEDERATED MARL

Though some work has implemented their federated MARL (FMARL) methods (Xu et al., 2023; Kumar et al., 2017; Song et al., 2022) 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 $\Lambda = \langle G, \tau, K, \psi \rangle$. 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 (Chen et al., 2021; Khodadadian et al., 2022; Kumar et al., 2017). 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 (Song et al., 2022; Xu et al., 2023; Chen et al., 2021) 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 (Bottou, 2010): $g(\theta_i^{k,j}; \xi_i^{k,j}) = \frac{1}{|\xi_i^{k,j}|} \sum_{\phi \in \xi_i^{k,j}} \nabla F_i(\phi)$, 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 $i \in 1, 2, \dots, n$ will be uploaded to the server agent.

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

$$\bar{\theta}^{k+1} = \bar{\theta}^k - \eta \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{\tau_i} g(\theta_i^{k,j}). \quad (10)$$

And the update rule for clients i is:

$$\theta_i^{k+1,j} = \begin{cases} \bar{\theta}^{k+1}, & j \bmod \tau_i = 0, \\ \theta_i^{k,j} - \eta g(\theta_i^{k,j}), & \text{otherwise,} \end{cases} \quad (11)$$

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 (Wang et al., 2020; Bottou et al., 2018; Wang & Joshi, 2021; Xu et al., 2023; Lian et al., 2015):

$$\mathbb{E}[\frac{1}{K} \sum_{k=0}^{K-1} \|\nabla F(\bar{\theta}^k)\|^2] \leq \epsilon, \quad (12)$$

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.2 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 $\sum_{i=1}^n w_i^k = M^k \leq n, \forall i \in [1, 2, \dots, n]$. Besides, we assume $\tau_i^k \in [1, \tau], \forall i \in \{1, \dots, n\}, k \in [0, K]$. To avoid being overly complicated, we omit superscripts or subscripts for some expressions.

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

$$\|Z\|_F^2 = |\text{Tr}(ZZ^\top)| = \sum_{i=1}^p \sum_{j=1}^q |z_{i,j}|^2 = \sum_{j=1}^q \|\vec{Z}_j\|^2, \quad (13)$$

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

$$\|Z\|_{op} = \max_{\|x\|=1} \|Zx\| = \sqrt{\lambda_{\max}(Z^\top Z)} \quad (14)$$

where λ_{\max} is the maximal eigenvalue of Z . From Lemma 7 of (Wang & Joshi, 2021), we have the following conclusion: suppose $Z_{p \times q}, D_{q \times q}$ are real matrices and D is symmetric, then we have:

$$\|ZD\|_F \leq \|Z\|_F \|D\|_{op} \quad (15)$$

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

$$\begin{aligned} E \left[\left\| \sum_{i=1}^n w_i X_i^k - E \left[\sum_{i=1}^n w_i X_i^k \right] \right\|^2 \right] &= E \left[\left\| \sum_{i=1}^n w_i X_i^k \right\|^2 \right] + \left(E \left[\sum_{i=1}^n w_i X_i^k \right] \right)^2 \\ &\quad - 2E \left[\left(\sum_{i=1}^n w_i X_i^k \right) E \left[\sum_{i=1}^n w_i X_i^k \right] \right] \\ &= E \left[\left\| \sum_{i=1}^n w_i X_i^k \right\|^2 \right] - \left(E \left[\sum_{i=1}^n w_i X_i^k \right] \right)^2. \end{aligned} \quad (16)$$

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

$$\mathbb{E} [\mathbf{X}^k] = \mathbb{E} [\mathbf{Y}^k] = \mathbf{Y}^k, \quad (17)$$

and

$$\mathbb{E}_{p \neq q} \langle g_p(\theta_p) - \nabla F_p(\theta_p), g_q(\theta_q) - \nabla F_q(\theta_q) \rangle = 0, \quad (18)$$

Further, we can derive

$$\mathbb{E}_{p \neq q} [\langle \mathbf{X}_p^k - \mathbf{Y}_p^k, \mathbf{X}_q^k - \mathbf{Y}_q^k \rangle] = 0. \quad (19)$$

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

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

$$\begin{aligned}
&= \sum_{i=1}^n w_i^2 E \left[\sum_{j=1}^{\tau_i^k} \left(g_i \left(\theta_i^{k,j} \right) - \nabla F_i \left(\theta_i^{k,j} \right) \right)^2 \right. \\
&\quad \left. + \sum_{p \neq q} \left\langle g_i \left(\theta_i^{k,p} \right) - \nabla F_i \left(\theta_i^{k,p} \right), g_i \left(\theta_i^{k,q} \right) - \nabla F_i \left(\theta_i^{k,q} \right) \right\rangle \right] \\
&= \sum_{i=1}^n w_i^2 E \left[\sum_{j=1}^{\tau_i^k} \left(g_i \left(\theta_i^{k,j} \right) - \nabla F_i \left(\theta_i^{k,j} \right) \right)^2 \right] \\
&\leq \sum_{i=1}^n w_i^2 \sum_{j=1}^{\tau_i^k} \left[\mu \left\| \nabla F_i \left(\theta_i^{k,j} \right) \right\|^2 + \sigma^2 \right] \\
&= \mu \sum_{i=1}^n w_i^2 \sum_{j=1}^{\tau_i^k} \left\| \nabla F_i \left(\theta_i^{k,j} \right) \right\|^2 + \sigma^2 \sum_{i=1}^n w_i^2 \\
&\leq \mu \sum_{i=1}^n w_i^2 \sum_{j=1}^{\tau_i^k} \left\| \nabla F_i \left(\theta_i^{k,j} \right) \right\|^2 + \sigma^2 (M^k)^2.
\end{aligned}$$

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

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

According to equation (16), (17) and the definition of $\mathbf{X}^{(k)}$, we have

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

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

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

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

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

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

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

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

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

A.3 PROOF OF THEOREMS

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

$$\frac{4(E[F(\bar{\theta}^1)] - E[F(\bar{\theta}^k)])}{K\eta} + 4 \left(\bar{A} + C + D + E + F + \mu\eta C \sum_{k=0}^K \frac{1}{K} \sum_{i=1}^n w_i^2 \tau_i^k \right) \quad (24)$$

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

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

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

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

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

Based on the above expressions, we can obtain

$$E \left\| \bar{\theta}^k - \theta_i^{k,j} \right\|^2 \leq 2\eta^2 \sigma^2 + (2\mu \eta^2 + 2j \eta^2) \sum_{j=1}^{\tau_i^k} E \left\| \nabla F_i \left(\theta_i^{k,j} \right) \right\|^2, \quad (25)$$

In addition,

$$\begin{aligned}
\left\| \nabla F_i \left(\theta_i^{k,j} \right) \right\|^2 & \leq 2 \left\| \nabla F_i \left(\theta_i^{k,j} \right) - \nabla F_i \left(\bar{\theta}^k \right) \right\|^2 + 2 \left\| \nabla F_i \left(\bar{\theta}^k \right) \right\|^2 \\
& \leq 2L^2 \left\| \theta_i^{k,j} - \bar{\theta}^k \right\|^2 + 2 \left\| \nabla F_i \left(\bar{\theta}^k \right) \right\|^2
\end{aligned} \quad (26)$$

Take (32) back to (31), then

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

Take the sum within two communication rounds,

$$\begin{aligned} \sum_{j=1}^{\tau_i^k} E \left\| \bar{\theta}^k - \theta_i^{k,j} \right\|^2 &\leq 2\tau_i^k \eta^2 \sigma^2 + 2\eta^2 \left(\mu \tau_i^k + \frac{\tau_i^k (1 + \tau_i^k)}{2} \right) \sum_{j=1}^{\tau_i^k} E \left[2L^2 \left\| \theta_i^{k,j} - \bar{\theta}^k \right\|^2 + 2 \left\| \nabla F_i (\bar{\theta}^k) \right\|^2 \right] \\ &= 2\tau_i^k \eta^2 \sigma^2 + 2L^2 \eta^2 [2\mu \tau_i^k + \tau_i^k (1 + \tau_i^k)] \sum_{j=1}^{\tau_i^k} E \left\| \theta_i^{k,j} - \bar{\theta}^k \right\|^2 \\ &\quad + 2\eta^2 [2\mu \tau_i^k + \tau_i^k (1 + \tau_i^k)] \cdot \tau_i^k \left\| \nabla F_i (\bar{\theta}^k) \right\|^2. \end{aligned} \quad (28)$$

After minor rearranging, we derive

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

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

$$(1 - B_i^k) \sum_{j=1}^{\tau_i^k} E \left\| \bar{\theta}^k - \theta_i^{k,j} \right\|^2 \leq 2\tau_i^k \eta^2 \sigma^2 + \frac{\tau_i^k}{L^2} \cdot B_i^k \left\| \nabla F_i (\bar{\theta}^k) \right\|^2, \quad (30)$$

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

Take (37) back to (30), then

$$\begin{aligned} \frac{1}{\tau_i^k} \sum_{j=1}^{\tau_i^k} E \left\| \nabla F_i (\bar{\theta}^k) - \nabla F_i (\theta_i^{k,j}) \right\|^2 &\leq \frac{L^2}{\tau_i^k} \sum_{j=1}^{\tau_i^k} E \left\| \bar{\theta}^k - \theta_i^{k,j} \right\|^2 \\ &\leq \frac{2\eta^2 \sigma^2 L^2}{1 - B} + \frac{B}{1 - B} \left\| \nabla F_i (\bar{\theta}^k) \right\|^2. \end{aligned} \quad (32)$$

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

$$\begin{aligned} \sum_{j=1}^{\tau_i^k} \left\| \nabla F_i (\theta_i^{k,j}) \right\|^2 &\leq 2L^2 \sum_{j=1}^{\tau_i^k} \left\| \theta_i^{k,j} - \bar{\theta}^k \right\|^2 + 2\tau_i^k \left\| \nabla F_i (\bar{\theta}^k) \right\|^2 \\ &\leq 2\tau_i^k \left(\frac{2\eta^2 \sigma^2 L^2}{1 - B} + \frac{B}{1 - B} \left\| \nabla F_i (\bar{\theta}^k) \right\|^2 \right) + 2\tau_i^k \left\| \nabla F_i (\bar{\theta}^k) \right\|^2 \\ &= \frac{4\tau_i^k L \eta^2 \sigma^2}{1 - B^2} + \frac{2\tau_i^k}{1 - B} \left\| \nabla F_i (\bar{\theta}^k) \right\|^2 \end{aligned} \quad (33)$$

Further,

$$\begin{aligned} \frac{\mu L \eta}{2} \sum_{i=1}^n w_i^2 \sum_{j=1}^{\tau_i^k} \left\| \nabla F_i (\theta_i^{k,j}) \right\|^2 &\leq \frac{\mu L \eta}{2} \sum_{i=1}^n w_i^2 \left(\frac{4\tau_i^k L \eta^2 \sigma^2}{1 - B} + \frac{2\tau_i^k}{1 - B} \left\| \nabla F_i (\bar{\theta}^k) \right\|^2 \right) \\ &\leq \frac{2\mu L^2 \sigma^2 \eta^3}{1 - B} \sum_{i=1}^n w_i^2 \tau_i^k + \frac{\mu L \eta \tau}{1 - B} \sum_{i=1}^n w_i^2 \left\| \nabla F_i (\bar{\theta}^k) \right\|^2 \end{aligned} \quad (34)$$

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

$$\begin{aligned}
\frac{E[F(\bar{\theta}^{k+1})] - E[F(\bar{\theta}^k)]}{\eta} &\leq -\frac{1}{2} \|\nabla F(\bar{\theta}^k)\|^2 + \frac{1}{n} \sum_{i=1}^n \left[\frac{2\eta^2 \sigma^2 L^2}{1-B} + \frac{B}{1-B} \|\nabla F_i(\bar{\theta}^k)\|^2 \right] \\
&\quad + E \left\| \sum_{i=1}^n \left(\frac{1}{n\tau_i^k} - w_i \right) Y_i^k \right\|^2 + \frac{2\mu L^2 \sigma^2 \eta^3}{1-B} \sum_{i=1}^n w_i^2 \tau_i^k \\
&\quad + \frac{\mu L \eta \tau}{1-B} \sum_{i=1}^n w_i^2 \|\nabla F_i(\bar{\theta}^k)\|^2 + \frac{L \eta \sigma^2}{2} \sum_{i=1}^n w_i^2 \\
&= -\frac{1}{2} \|\nabla F(\bar{\theta}^k)\|^2 + \frac{2\eta^2 \sigma^2 L^2}{1-B} + \frac{B}{1-B} \sum_{i=1}^n \frac{1}{n} \|\nabla F_i(\bar{\theta}^k)\|^2 \\
&\quad + E \left\| \sum_{i=1}^n \left(\frac{1}{n\tau_i^k} - w_i \right) Y_i^k \right\|^2 + \frac{\mu L \eta \tau}{1-B} \sum_{i=1}^n w_i^2 \|\nabla F_i(\bar{\theta}^k)\|^2 \\
&\quad + \frac{2\mu L^2 \sigma^2 \eta^3}{1-B} \sum_{i=1}^n w_i^2 \tau_i^k + \frac{L \eta \sigma^2}{2} \sum_{i=1}^n w_i^2
\end{aligned} \tag{35}$$

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

$$\begin{aligned}
\frac{E[F(\bar{\theta}^{k+1})] - E[F(\bar{\theta}^k)]}{\eta} &\leq \frac{2\mu L \eta \tau \beta^2 + 2B\beta^2 - (1-B)}{2(1-B)} \|\nabla F(\bar{\theta}^k)\|^2 + \frac{2\eta^2 \sigma^2 L^2}{1-B} \\
&\quad + \frac{BK^2 + \mu L \eta \tau \kappa^2}{1-B} + E \left\| \sum_{i=1}^n \left(\frac{1}{n\tau_i^k} - w_i \right) Y_i^k \right\|^2 \\
&\quad + \frac{2\mu L^2 \sigma^2 \eta^3}{1-B} \sum_{i=1}^n w_i^2 \tau_i^k + \frac{L \eta \sigma^2}{2} \sum_{i=1}^n w_i^2
\end{aligned} \tag{36}$$

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

$$\begin{aligned}
\frac{1}{K} \sum_{k=1}^K \|\nabla F(\bar{\theta}^k)\|^2 &\leq \frac{4(E[F(\bar{\theta}^1)] - E[F(\bar{\theta}^K)])}{K\eta} \\
&\quad + 4 \left(\bar{A} + C + D + E + F + \mu \eta C \sum_{k=0}^K \frac{1}{K} \sum_{i=1}^n w_i^2 \tau_i^k \right)
\end{aligned} \tag{37}$$

where

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

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

A.4 EXTENDED RELATED WORK

Cooperative MARL Cooperative MARL is one of the most prevalent categories of MARL research whose applications range from multi-player games (Samvelyan et al., 2019; Mordatch & Abbeel,

2018; Bard et al., 2020) to industrial simulations (Vinitsky et al., 2018; Li et al., 2022a; Zhou et al., 2020; Peng et al., 2021a). Considering the complexity of cooperative tasks and partial observability of learning agents, much research (Lowe et al., 2017; Rashid et al., 2020; Sunehag et al., 2017; Qiu et al., 2021; Yu et al., 2022a; Hu et al., 2021) adopts some simplified practices to facilitate policy learning. The two most frequent common practices are parameter sharing (PS) and centralized training with decentralized execution (CTDE). PS means all agents share the same network architectures as well as network parameters. CTDE treats the training phase and the evaluation phase differently: in the training phase, a centralized structure can accept global state information rather than local observation information to improve cooperation while in the evaluation phase, the agents make decisions only based on their local observations. Although these practices work well in game environments, when it comes to deploying MARL methods into real-world environments where 1) multiple devices are available to necessitate **scalability**, 2) the information transmission relies on **cross-device communication**, and 3) **user privacy** are critical, PS is no longer practical. An intuitive thought is letting each agent learn its policy. Actually, some research (de Witt et al., 2020; Kuba et al., 2022; Yu et al., 2022a; Christianos et al., 2021) have considered non-PS settings. MAPPO (Yu et al., 2022a), HATRPO and HAPPO (Kuba et al., 2022) find that for some tasks on SMAC (Samvelyan et al., 2019) micromanagement environment and multi-agent MuJoCo (Peng et al., 2021a), non-PS can help to learn more agent-specialized policies to achieve better cooperation. SePS (Christianos et al., 2021) details the influence of PS on training efficiency and performance. However, most of these methods are still developed for game environments aiming at improving task-oriented performance. Though they pose a potential way to protect user privacy and support scalability on multiple devices, the high-frequency information transmission causes unaffordable bandwidth overload and communication costs in the perspective of realistic scenarios.

Distributed MARL Aiming at bringing MARL into real-world applications, an increasing number of research (Yu et al., 2022b; Raffard et al., 2004; Xu et al., 2023; 2022; Li et al., 2022b; Khodadadian et al., 2022; Chen et al., 2021; Zhuo et al., 2019; Kumar et al., 2017; Wang et al., 2022) explore diverse paradigms of distributed MARL. They mainly form two categories. One is a fully decentralized setting where agents learn policies associated with the information transferred from other agents (Yu et al., 2022b; Raffard et al., 2004; Wang et al., 2022). This category usually requires predefined or learned communication protocols drawn from prior knowledge. Thus, intricate synchronization mechanisms need to be introduced, which can impact the communication efficiency and be unstable due to the fluctuations in agents’ presence.

Alternatively, other distributed MARL approaches (Xu et al., 2023; Song et al., 2022; Li et al., 2022b; Chen et al., 2021; Kumar et al., 2017) employ federated learning, instituting a client-server paradigm to develop Federated MARL. FCRL (Kumar et al., 2017) combines hierarchical RL with federated learning. It uses a centralized controller for subtask assignment and allows direct agent-to-agent communication, which is not the general paradigm we considered in this paper. LAPG (Chen et al., 2021) proposes a distributed policy gradient method that can be applied in distributed RL and MARL. It uses a condition that controls the parameter uploading process to enhance communication efficiency. LAPG focuses more on distributed RL since the condition only utilizes information from the current agent rather than combining it with information from other agents. FMARL (Xu et al., 2023), Fed-MADRL (Song et al., 2022) and Fed-MARL (Li et al., 2022b) all adopt the conventional federated learning scheme that the training parameters, *i.e.*, stochastic gradients for FMARL, model parameters for Fed-MARL and quantified model parameters for Fed-MADRL, are uploaded to the server for aggregation during client-server communication. Then, the updated global model will be broadcast to the agents. The difference is that Fed-MARL’s aggregation is weighted by relative mini-batch size while Fed-MADRL utilizes a momentum aggregation mechanism. FMARL provides two optimization methods: one is weight decay during aggregation, and the other is through agent-to-agent communication. However, their aggregation does not take the diversity of learning agents’ local environments into account. To be specific, each agent has different local observation in real-world settings like cooperative exploration and navigation, which leads to **non-i.i.d.** problem and necessitates the importance of the **centralized coordination** in the centralized training phase in MARL. In addition, current methods mostly focus on optimizing a singular target rather than a **composite** one to ensure simultaneous optimization of task-oriented performance and system efficiency.

A.5 BASELINE METHODS

IPPO (de Witt et al., 2020) 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 (Yu et al., 2022a), 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 (Foerster et al., 2016) 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 (Peng et al., 2021b) 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 (Xu et al., 2023) is a strong baseline in the domain of federated MARL, which is also experimentally evaluated in a multi-vehicle autonomous driving simulation benchmark (Vinitsky et al., 2018). We notice that there are two methods proposed in (Xu et al., 2023). 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.6 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.

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

scenarios methods/metrics	scenario1					scenario2					scenario3				
	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	<u>0.650</u>	0.548	0.456	0.598	0.754	0.574	<u>0.617</u>	<u>0.548</u>	0.623	<u>0.653</u>	0.716	<u>0.644</u>	0.513	0.632

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 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 methods/metrics	scenario4					scenario5					scenario6				
	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	<u>0.632</u>	0.644	<u>0.621</u>	0.395	<u>0.573</u>	0.553	0.507	<u>0.448</u>	<u>0.504</u>	0.503	0.611	0.639	0.504	<u>0.502</u>	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.

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

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 neighbour agents.

Table 6: Average client selection effectiveness in navigation task on scenarios 1, 2, and 3 with varying number of agents.

Scenarios	Scenario1				Scenario2				Scenario3			
total agent number	3	6	9	12	3	6	9	12	3	6	9	12
average selected	2.75 \pm 0.32	5.16 \pm 0.83	7.22 \pm 1.09	8.85 \pm 1.66	2.91 \pm 0.15	5.47 \pm 0.86	7.48 \pm 1.49	9.14 \pm 1.79	2.81 \pm 0.42	5.00 \pm 1.02	7.19 \pm 1.14	8.75 \pm 1.59
average selected at last	2.32 \pm 0.69	4.45 \pm 0.92	7.04 \pm 1.01	8.56 \pm 1.43	2.59 \pm 0.57	4.90 \pm 0.89	7.16 \pm 1.01	8.67 \pm 1.42	2.38 \pm 0.64	4.59 \pm 0.92	7.07 \pm 0.98	8.42 \pm 1.44

Table 7: Hyperparameter settings for experiments

Hyperparameters	Value	Hyperparameters	Value
Critic lr	5e-4	Hidden layer	1
Activation	ReLU	Hidden 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 epsilon	1e-5

In Tab. 6, we show the client selection effect. We can find an obvious trend that the selected agents' ratio is decreasing with the number of initial agents. It is consistent with our intuition since the more learning agents involved in a scenario, the higher the probability that agents can make similar decisions.