GIFAIR-FL: A Framework for Group and Individual Fairness in Federated Learning

Xubo Yue[1], Maher Nouiehed[2], and Raed Al Kontar[1]

[1] *Industrial and Operations Engineering, University of Michigan, Ann Arbor*

[2] *Industrial Engineering and Management, American University of Beirut, Beirut, Lebanon*

1 Appendix

In Sec. 2, we restate our main assumptions. In Sec. 3, we provide the detailed proofs of Lemmas and Theorems in our main paper. Finally, in Sec. 4, we present some additional empirical results.

2 Assumptions

We make the following assumptions.

Assumption 1. *F_k is L-smooth and μ-strongly convex for all $k \in [K]$.*

Assumption 2. *Denote by $\zeta_k^{(t)}$ the batched data from client k and $g_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)})$ the stochastic gradient calculated on this batched data. The variance of stochastic gradients are bounded. Specifically,*

$$\mathbb{E}\Big\{ \Big\| g_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) - \nabla F_k(\boldsymbol{\theta}_k^{(t)}) \Big\|^2 \Big\} \leq \sigma_k^2, \forall k \in [K].$$

It can be shown that, at local iteration t during communication round c,

$$\begin{aligned}
&\mathbb{E}\Big\{ \Big\| \nabla H_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) - \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \Big\|^2 \Big\} \\
&= \mathbb{E}\Big\{ \Big\| (1 + \frac{\lambda r_k^c}{p_k |\mathcal{A}_{s_k}|}) g_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) - (1 + \frac{\lambda r_k^c}{p_k |\mathcal{A}_{s_k}|}) \nabla F_k(\boldsymbol{\theta}_k^{(t)}) \Big\|^2 \Big\} \\
&\leq (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k^c)^2 \sigma_k^2, \forall k \in [K].
\end{aligned}$$

Here, $\nabla H_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)})$ denotes the stochastic gradient of H_k evaluated on the batched data $\zeta_k^{(t)}$.

Assumption 3. *The expected squared norm of stochastic gradient is bounded. Specifically,*

$$\mathbb{E}\left\{ \left\| g_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) \right\|^2 \right\} \leq G^2, \forall k \in [K].$$

It can be shown that, at local iteration t during communication round c,

$$\mathbb{E}\left\{ \left\| \nabla H_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) \right\| \right\} = \mathbb{E}\left\{ \left\| (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k^c) g_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) \right\|^2 \right\}$$
$$\leq (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k^c)^2 G^2, \forall k \in [K].$$

For the non-convex setting, we replace Assumption 1 by the following assumption.

Assumption 4. *F_k is L-smooth for all $k \in [K]$.*

In our proof, for the sake of neatness, we drop the superscript of r_k^c.

We use the definition in Li et al. (2019) to roughly quantify the degree of non-*i.i.d.*-ness. Specifically,

$$\Gamma_K = H^* - \sum_{k=1}^{K} p_k H_k^* = \sum_{k=1}^{K} p_k (H^* - H_k^*).$$

If data from all sensitive attributes are *i.i.d.*, then $\Gamma_K = 0$ as number of clients grows. Otherwise, $\Gamma_K \neq 0$ (Li et al., 2019).

3 Detailed Proof

3.1 Proof of Lemma

Lemma 1. *For any given $\boldsymbol{\theta}$, the global objective function $H(\boldsymbol{\theta})$ defined in the main paper can be expressed as*

$$H(\boldsymbol{\theta}) = \sum_{k=1}^{K} \left(p_k + \frac{\lambda}{|\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}) \right) F_k(\boldsymbol{\theta}),$$

where

$$r_k(\boldsymbol{\theta}) \triangleq \sum_{1 \leq j \neq s_k \leq d} \text{sign}(L_{s_k}(\boldsymbol{\theta}) - L_j(\boldsymbol{\theta}))$$

and $s_k \in [d]$ is the group index of device k. Consequently,

$$H(\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k H_k(\boldsymbol{\theta}).$$

Proof. By definition, at communication round c,

$$\begin{aligned}
H(\boldsymbol{\theta}) &= \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}) + \lambda \sum_{1\le i<j\le d} |L_i(\boldsymbol{\theta}) - L_j(\boldsymbol{\theta})| \\
&= \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}) + \lambda \sum_{1\le i<j\le d} \left| \frac{1}{|\mathcal{A}_i|} \sum_{k\in\mathcal{A}_i} F_k(\boldsymbol{\theta}) - \frac{1}{|\mathcal{A}_j|} \sum_{k\in\mathcal{A}_j} F_k(\boldsymbol{\theta}) \right| \\
&= \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}) + \lambda \sum_{1\le i<j\le d} \operatorname{sign}(L_i(\boldsymbol{\theta}) - L_j(\boldsymbol{\theta})) \left(\frac{1}{|\mathcal{A}_i|} \sum_{k\in\mathcal{A}_i} F_k(\boldsymbol{\theta}) - \frac{1}{|\mathcal{A}_j|} \sum_{k\in\mathcal{A}_j} F_k(\boldsymbol{\theta}) \right) \\
&= \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}) + \lambda \sum_{u=1}^{d-1} \sum_{u<j\le d} \operatorname{sign}(L_u(\boldsymbol{\theta}) - L_j(\boldsymbol{\theta})) \left(\frac{1}{|\mathcal{A}_u|} \sum_{k\in\mathcal{A}_u} F_k(\boldsymbol{\theta}) - \frac{1}{|\mathcal{A}_j|} \sum_{k\in\mathcal{A}_j} F_k(\boldsymbol{\theta}) \right) \\
&= \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}) + \lambda \sum_{u=1}^{d} \sum_{k\in\mathcal{A}_u} \sum_{u\ne j\le d} \operatorname{sign}(L_u(\boldsymbol{\theta}) - L_j(\boldsymbol{\theta})) \frac{F_k(\boldsymbol{\theta})}{|\mathcal{A}_u|} \\
&= \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}) + \sum_{k=1}^{K} \frac{\lambda}{|\mathcal{A}_{s_k}|} \sum_{1\le j\ne s_k\le d} \operatorname{sign}(L_{s_k}(\boldsymbol{\theta}) - L_j(\boldsymbol{\theta})) F_k(\boldsymbol{\theta}) \\
&= \sum_{k=1}^{K} \left(p_k + \frac{\lambda}{|\mathcal{A}_{s_k}|} r_k^c(\boldsymbol{\theta}) \right) F_k(\boldsymbol{\theta}).
\end{aligned}$$

The fifth equality is achieved by rearranging the equation and merging items with the same group label. By definition of H_k, we thus proved

$$H(\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k H_k(\boldsymbol{\theta}).$$

□

3.2 Learning Bound

We present a generalization bound for our learning model. Denote by $\mathcal{G}$ the family of the losses associated to a hypothesis set $\mathcal{H}$: $\mathcal{G} = \{(x, y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$. The weighted Rademacher complexity (Mohri et al., 2019) is defined as

$$\mathfrak{R}_{\boldsymbol{m}}(\mathcal{G}, \boldsymbol{p}) := \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h\in\mathcal{H}} \sum_{k=1}^{K} \frac{p_k}{N_k} \sum_{n=1}^{N_k} \sigma_{k,n} \ell(h(x_{k,n}), y_{k,n}) \right]$$

where $\boldsymbol{m} = (N_1, N_2, \dots, N_k)$, $\boldsymbol{p} = (p_1, \dots, p_K)$ and $\boldsymbol{\sigma} = (\sigma_{k,n})_{k\in[K], n\in[N_k]}$ is a collection of Rademacher variables taking values in $\{-1, +1\}$. Denote by $\mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)$ the expected loss according to our fairness formulation. Denote by $\hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)$ the expected empirical loss (See Appendix for a detailed expression).

Theorem 1. *Assume that the loss ℓ is bounded above by $M > 0$. Fix $\epsilon_0 > 0$ and $\boldsymbol{m}$. Then, for any $\delta_0 > 0$,* *with probability at least $1-\delta_0$ over samples $D_k \sim \mathcal{D}_k$, the following holds for all $h \in \mathcal{H}$:*

$$\mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) \leq \hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) + \sqrt{\frac{1}{2}\sum_{k=1}^{K}(\frac{p_k}{N_k}M + \lambda\frac{d(d-1)}{2}M)^2 \log\frac{1}{\delta_0}} + 2\mathfrak{R}_{\boldsymbol{m}}(\mathcal{G}, \boldsymbol{p}) + \lambda\frac{d(d-1)}{2}M.$$

It can be seen that, given a sample of data, we can bound the generalization error $\mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)$ with high probability. When $\lambda = 0$, the bound is same as the generalization bound in `FedAvg` (Mohri et al., 2018). When we consider the worst combination of p_k by taking the supremum of the upper bound in Theorem 1 and let $\lambda = 0$, then our generalization bound is same as the one in `AFL` (Mohri et al., 2019).

Proof. Define

$$\Phi(D_1, \ldots, D_K) = \sup_{h\in\mathcal{H}} \Big(\mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)\Big).$$

Let $D' = (D'_1, \ldots, D'_K)$ be a sample differing from $D = (D_1, \ldots, D_K)$ only by one point $x'_{k,n}$. Therefore, we have

$$\begin{aligned}
\Phi(D') - \Phi(D) &= \sup_{h\in\mathcal{H}} \Big(\mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \hat{\mathcal{L}}_{\mathcal{D'}_{\boldsymbol{p}}^{\lambda}}(h)\Big) - \sup_{h\in\mathcal{H}} \Big(\mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)\Big) \\
&\leq \sup_{h\in\mathcal{H}} \Big(\mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \hat{\mathcal{L}}_{\mathcal{D'}_{\boldsymbol{p}}^{\lambda}}(h)\Big) - \Big(\mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)\Big) \\
&\leq \sup_{h\in\mathcal{H}} \big\{ \sup_{h\in\mathcal{H}} \mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \sup_{h\in\mathcal{H}} \hat{\mathcal{L}}_{\mathcal{D'}_{\boldsymbol{p}}^{\lambda}}(h) - \mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) + \hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)\big\} \\
&= \sup_{h\in\mathcal{H}} \big\{\hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \hat{\mathcal{L}}_{\mathcal{D'}_{\boldsymbol{p}}^{\lambda}}(h)\big\}
\end{aligned}$$

By definition,

$$\hat{\mathcal{L}}_{\mathcal{D'}_{\boldsymbol{p}}^{\lambda}}(h) = \sum_{k=1}^{K}\frac{p_k}{N_k}\sum_{n=1}^{N_k}\ell(h(x'_{k,n}), y'_{k,n}) +$$
$$\lambda \sum_{1\leq i<j\leq d} \Big|\frac{\sum_{k\in\mathcal{A}_i}\frac{1}{N_k}\sum_{n=1}^{N_k}\ell(h(x'_{k,n}), y'_{k,n})}{|\mathcal{A}_i|} - \frac{\sum_{k\in\mathcal{A}_j}\frac{1}{N_k}\sum_{n=1}^{N_k}\ell(h(x'_{k,n}), y'_{k,n})}{|\mathcal{A}_j|}\Big|.$$

Therefore,

$$\begin{aligned}
&\sup_{h\in\mathcal{H}} \big\{\hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \hat{\mathcal{L}}_{\mathcal{D'}_{\boldsymbol{p}}^{\lambda}}(h)\big\} \\
&\leq \sup_{h\in\mathcal{H}} \left[\frac{p_k}{N_k}(\ell(h(x'_{k,n}), y'_{k,n}) - \ell(h(x_{k,n}), y_{k,n})) + \lambda\frac{d(d-1)}{2}M\right] \\
&\leq \frac{p_k}{N_k}M + \lambda\frac{d(d-1)}{2}M.
\end{aligned}$$

By McDiarmid's inequality, for $\delta_0 = \exp\left(\frac{-2\epsilon_0^2}{\sum_{k=1}^{K}(\frac{p_k}{N_k}M+\lambda\frac{d(d-1)}{2}M)^2}\right)$, the following holds with probability at least $1-\delta_0$

$$\Phi(D) - \mathbb{E}_D[\Phi(D)] \leq \epsilon_0 = \sqrt{\frac{1}{2}\sum_{k=1}^{K}(\frac{p_k}{N_k}M + \lambda\frac{d(d-1)}{2}M)^2 \log\frac{1}{\delta_0}}.$$

Our next goal is to bound $\mathbb{E}[\Phi(D)]$. We have

$$\begin{aligned}
\mathbb{E}_D[\Phi(D)] &= \mathbb{E}_D\left[\sup_{h\in\mathcal{H}}\left(\mathcal{L}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h) - \hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)\right)\right] \\
&= \mathbb{E}_D\left[\sup_{h\in\mathcal{H}}\mathbb{E}_{D'}\left(\hat{\mathcal{L}}_{\mathcal{D}'^{\lambda}_{\boldsymbol{p}}}(h) - \hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)\right)\right] \\
&\leq \mathbb{E}_D\mathbb{E}_{D'}\sup_{h\in\mathcal{H}}\left(\hat{\mathcal{L}}_{\mathcal{D}'^{\lambda}_{\boldsymbol{p}}}(h) - \hat{\mathcal{L}}_{\mathcal{D}_{\boldsymbol{p}}^{\lambda}}(h)\right) \\
&\leq \mathbb{E}_D\mathbb{E}_{D'}\sup_{h\in\mathcal{H}}\left[\sum_{k=1}^{K}\frac{p_k}{N_k}\sum_{n=1}^{N_k}\ell(h(x'_{k,n}),y'_{k,n}) - \sum_{k=1}^{K}\frac{p_k}{N_k}\sum_{n=1}^{N_k}\ell(h(x_{k,n}),y_{k,n}) + \lambda\frac{d(d-1)}{2}M\right] \\
&\leq \mathbb{E}_D\mathbb{E}_{D'}\mathbb{E}_{\boldsymbol{\sigma}}\sup_{h\in\mathcal{H}}\left[\sum_{k=1}^{K}\frac{p_k}{N_k}\sum_{n=1}^{N_k}\sigma_{k,n}\ell(h(x'_{k,n}),y'_{k,n}) - \sum_{k=1}^{K}\frac{p_k}{N_k}\sum_{n=1}^{N_k}\sigma_{k,n}\ell(h(x_{k,n}),y_{k,n}) + \lambda\frac{d(d-1)}{2}M\right] \\
&\leq 2\mathfrak{R}_{\boldsymbol{m}}(\mathcal{G},\boldsymbol{p}) + \lambda\frac{d(d-1)}{2}M.
\end{aligned}$$

Therefore,

$$\Phi(D) \leq \sqrt{\frac{1}{2}\sum_{k=1}^{K}(\frac{p_k}{N_k}M + \lambda\frac{d(d-1)}{2}M)^2\log\frac{1}{\delta_0}} + 2\mathfrak{R}_{\boldsymbol{m}}(\mathcal{G},\boldsymbol{p}) + \lambda\frac{d(d-1)}{2}M.$$

□

3.3 Convergence (Strongly Convex)

Our proof is based on the convergence result of `FedAvg` (Li et al., 2019).

Theorem 2. *Assume Assumptions in the main paper hold and $|\mathcal{S}_c| = K$. For $\gamma,\mu > 0$ and $\eta^{(t)}$ is decreasing in a rate of $\mathcal{O}(\frac{1}{t})$. If $\eta^{(t)} \leq \mathcal{O}(\frac{1}{L})$, we have*

$$\mathbb{E}\left\{H(\bar{\boldsymbol{\theta}}^{(T)})\right\} - H^* \leq \frac{L}{2}\frac{1}{\gamma+T}\left\{\frac{4\xi}{\epsilon^2\mu^2} + (\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\right\|^2\right\},$$

where $\xi = 8(E-1)^2G^2 + 4L\Gamma_K + 2\frac{\Gamma_{max}}{\eta^{(t)}} + 4\sum_{k=1}^{K}p_k^2\sigma_k^2$ *and* $\Gamma_{max} \coloneqq \sum_{k=1}^{K}p_k|(H^* - H_k^*)| \geq |\sum_{k=1}^{K}p_k(H^* - H_k^*)| = |\Gamma_K|$.

Proof. For each device k, we introduce an intermediate model parameter $\boldsymbol{w}_k^{(t+1)} = \boldsymbol{\theta}_k^{(t)} - \eta^{(t)} \nabla H_k(\boldsymbol{\theta}_k^{(t)})$. If iteration $t+1$ is in the communication round, then $\boldsymbol{\theta}_k^{(t+1)} = \sum_{k=1}^K p_k \boldsymbol{w}_k^{(t+1)}$ (i.e., aggregation). Otherwise, $\boldsymbol{\theta}_k^{(t+1)} = \boldsymbol{w}_k^{(t+1)}$. Define $\bar{\boldsymbol{w}}^{(t)} = \sum_{k=1}^K p_k \boldsymbol{w}_k^{(t)}$ and $\bar{\boldsymbol{\theta}}^{(t)} = \sum_{k=1}^K p_k \boldsymbol{\theta}_k^{(t)}$. Also, define $\boldsymbol{g}^{(t)} = \sum_{k=1}^K p_k \nabla H_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)})$ and $\bar{\boldsymbol{g}}^{(t)} = \mathbb{E}(\boldsymbol{g}^{(t)}) = \sum_{k=1}^K p_k \nabla H_k(\boldsymbol{\theta}_k^{(t)})$.

Denote by $\boldsymbol{\theta}^*$ the optimal model parameter of the global objective function $H(\cdot)$. At iteration t, we have

$$
\begin{aligned}
&\mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^* \right\|^2 \right\} = \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)} \boldsymbol{g}^{(t)} - \boldsymbol{\theta}^* - \eta^{(t)} \bar{\boldsymbol{g}}^{(t)} + \eta^{(t)} \bar{\boldsymbol{g}}^{(t)} \right\|^2 \right\} \\
&= \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* - \eta^{(t)} \bar{\boldsymbol{g}}^{(t)} \right\|^2 \right\} + \mathbb{E}\left\{ 2\eta^{(t)} \langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* - \eta^{(t)} \bar{\boldsymbol{g}}^{(t)}, \bar{\boldsymbol{g}}^{(t)} - \boldsymbol{g}^{(t)} \rangle \right\} + \mathbb{E}\left\{ \eta^{(t)2} \left\| \boldsymbol{g}^{(t)} - \bar{\boldsymbol{g}}^{(t)} \right\|^2 \right\} \\
&= \underbrace{\mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* - \eta^{(t)} \bar{\boldsymbol{g}}^{(t)} \right\|^2 \right\}}_{A} + \underbrace{\mathbb{E}\left\{ \eta^{(t)2} \left\| \boldsymbol{g}^{(t)} - \bar{\boldsymbol{g}}^{(t)} \right\|^2 \right\}}_{B},
\end{aligned}
$$

since $\mathbb{E}\left\{ 2\eta^{(t)} \langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* - \eta^{(t)} \bar{\boldsymbol{g}}^{(t)}, \bar{\boldsymbol{g}}^{(t)} - \boldsymbol{g}^{(t)} \rangle \right\} = 0$. Our remaining work is to bound term A and term B.

Part I: Bounding Term A We can split term A above into three parts:

$$
\mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* - \eta^{(t)} \bar{\boldsymbol{g}}^{(t)} \right\|^2 \right\} = \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\} \underbrace{- 2\eta^{(t)} \mathbb{E}\left\{ \langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}}^{(t)} \rangle \right\}}_{\mathrm{C}} + \underbrace{\eta^{(t)2} \mathbb{E}\left\{ \left\| \bar{\boldsymbol{g}}^{(t)} \right\|^2 \right\}}_{\mathrm{D}}.
$$

For part C, We have

$$
\begin{aligned}
\mathrm{C} &= -2\eta^{(t)} \mathbb{E}\left\{ \langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}}^{(t)} \rangle \right\} = -2\eta^{(t)} \mathbb{E}\left\{ \sum_{k=1}^K p_k \langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \rangle \right\} \\
&= -2\eta^{(t)} \mathbb{E}\left\{ \sum_{k=1}^K p_k \langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}, \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \rangle \right\} - 2\eta^{(t)} \mathbb{E}\left\{ \sum_{k=1}^K p_k \langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \rangle \right\}
\end{aligned}
$$

To bound C, we need to use Cauchy-Schwarz inequality, inequality of arithmetic and geometric means. Specifically, the Cauchy-Schwarz inequality indicates that

$$
\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}, \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \rangle \geq - \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\| \left\| \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \right\|
$$

and inequality of arithmetic and geometric means further implies

$$
- \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\| \left\| \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \right\| \geq - \frac{\left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|^2 + \left\| \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \right\|^2}{2}.
$$

Therefore, we obtain

$$
\begin{aligned}
\mathrm{C} &= -2\eta^{(t)}\mathbb{E}\bigg\{\langle\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, \bar{\boldsymbol{g}}^{(t)}\rangle\bigg\} = -2\eta^{(t)}\mathbb{E}\bigg\{\sum_{k=1}^{K} p_k\langle\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, \nabla H_k(\boldsymbol{\theta}_k^{(t)})\rangle\bigg\} \\
&= -2\eta^{(t)}\mathbb{E}\bigg\{\sum_{k=1}^{K} p_k\langle\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}, \nabla H_k(\boldsymbol{\theta}_k^{(t)})\rangle\bigg\} - 2\eta^{(t)}\mathbb{E}\bigg\{\sum_{k=1}^{K} p_k\langle\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, \nabla H_k(\boldsymbol{\theta}_k^{(t)})\rangle\bigg\} \\
&\leq \mathbb{E}\bigg\{\eta^{(t)}\sum_{k=1}^{K} p_k\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 + {\eta^{(t)}}^2\sum_{k=1}^{K} p_k\left\|\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2 \\
&\qquad - 2\eta^{(t)}\sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k(\boldsymbol{\theta}^*)) - 2\eta^{(t)}\sum_{k=1}^{K} p_k\frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))\mu}{2}\left\|\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*\right\|^2\bigg\},
\end{aligned}
$$

where $-2\eta^{(t)}\mathbb{E}\bigg\{\sum_{k=1}^{K} p_k\langle\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, \nabla H_k(\boldsymbol{\theta}_k^{(t)})\rangle\bigg\}$ is bounded by the property of strong convexity of H_k.

Since H_k is $(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L$-smooth, we know

$$
\left\|\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2 \leq 2(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*)
$$

and therefore

$$
\begin{aligned}
\mathrm{D} &= \eta^{(t)2}\mathbb{E}\bigg\{\left\|\bar{\boldsymbol{g}}^{(t)}\right\|^2\bigg\} \leq \eta^{(t)2}\mathbb{E}\bigg\{\sum_{k=1}^{K} p_k\left\|\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2\bigg\} \\
&\leq 2\eta^{(t)2}\mathbb{E}\bigg\{\sum_{k=1}^{K} p_k(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*)\bigg\}
\end{aligned}
$$

by convexity of norm.

Therefore, combining C and D, we have

$$
\begin{aligned}
A &= \mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* - \eta^{(t)}\bar{\boldsymbol{g}}^{(t)}\right\|^2 \Big\} \\
&\le \mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2 \Big\} + 2\eta^{(t)2}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*)\Big\} \\
&\quad + \eta^{(t)}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k \frac{1}{\eta^{(t)}} \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 \Big\} + \eta^{(t)^2}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k \left\|\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2 \Big\} \\
&\quad - 2\eta^{(t)}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k(\boldsymbol{\theta}^*))\Big\} - 2\eta^{(t)}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))\mu}{2} \left\|\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*\right\|^2 \Big\} \\
&\le \mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2 \Big\} - \eta^{(t)}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))\mu \left\|\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*\right\|^2 \Big\} + \sum_{k=1}^{K} p_k \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 \\
&\quad + \underbrace{4\eta^{(t)2}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*)\Big\} - 2\eta^{(t)}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k(\boldsymbol{\theta}^*))\Big\}}_{\mathrm{E}}.
\end{aligned}
$$

In the last inequality, we simply rearrange other terms and use the fact that $\left\|\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2 \le 2(1 + \frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*)$ as aforementioned.

To bound E, we define $\gamma_k^{(t)} = 2\eta^{(t)}(1 - 2(1 + \frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L\eta^{(t)})$. Assume $\eta^{(t)} \le \frac{1}{4(1+\frac{(d-1)}{\min\{p_k|\mathcal{A}_{s_k}|\}}\lambda)L}$, then we know $\eta^{(t)} \le \gamma_k^{(t)} \le 2\eta^{(t)}$.

Therefore, we have

$$
\begin{aligned}
\mathrm{E} &= 4\eta^{(t)2}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*)\Big\} - 2\eta^{(t)}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k(\boldsymbol{\theta}^*))\Big\} \\
&= 4\eta^{(t)2}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*)\Big\} - 2\eta^{(t)}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^* + H_k^* - H_k(\boldsymbol{\theta}^*))\Big\} \\
&= -2\eta^{(t)}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(1 - 2(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L\eta^{(t)})(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*)\Big\} + 2\eta^{(t)}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}^*) - H_k^*)\Big\} \\
&= -\mathbb{E}\Big\{ \sum_{k=1}^{K} \gamma_k^{(t)} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H^* + H^* - H_k^*)\Big\} + 2\eta^{(t)}\mathbb{E}\Big\{ H^* - \sum_{k=1}^{K} p_k H_k^*\Big\} \\
&= \underbrace{-\mathbb{E}\Big\{ \sum_{k=1}^{K} \gamma_k^{(t)} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H^*)\Big\}}_{\mathrm{F}} + \underbrace{\mathbb{E}\Big\{ \sum_{k=1}^{K} (2\eta^{(t)} - \gamma_k^{(t)}) p_k(H^* - H_k^*)\Big\}}_{\mathrm{G}}.
\end{aligned}
$$

If $H^* - H_k^* \ge 0$ for some k, then $(2\eta^{(t)} - \gamma_k^{(t)})p_k(H^* - H_k^*) \le 2\eta^{(t)}p_k(H^* - H_k^*)$. If $H^* - H_k^* < 0$ otherwise, then $(2\eta^{(t)} - \gamma_k^{(t)})p_k(H^* - H_k^*)$ is negative and $(2\eta^{(t)} - \gamma_k^{(t)})p_k(H^* - H_k^*) \le -2\eta^{(t)}p_k(H^* - H_k^*)$. Therefore, by definition of Γ_{max},

$$\mathrm{G} \leq 2\eta^{(t)}\mathbb{E}\bigg\{ \sum_{k=1}^{K} p_k |H^* - H_k^*| \bigg\} = 2\eta^{(t)}\Gamma_{max}.$$

The remaining goal of Part I is to bound term F. Note that

$$\begin{aligned}
\mathrm{F} &= -\mathbb{E}\bigg\{ \sum_{k=1}^{K} \gamma_k^{(t)} p_k (H_k(\boldsymbol{\theta}_k^{(t)}) - H^*) \bigg\} \\
&= -\mathbb{E}\bigg\{ \bigg(\sum_{k=1}^{K} p_k \gamma_k^{(t)} (H_k(\boldsymbol{\theta}_k^{(t)}) - H_k(\bar{\boldsymbol{\theta}}^{(t)})) + \sum_{k=1}^{K} p_k \gamma_k^{(t)} (H_k(\bar{\boldsymbol{\theta}}^{(t)}) - H^*) \bigg) \bigg\} \\
&\leq -\mathbb{E}\bigg\{ \bigg(\sum_{k=1}^{K} p_k \gamma_k^{(t)} \langle \nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}), \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \rangle + \sum_{k=1}^{K} p_k \gamma_k^{(t)} (H_k(\bar{\boldsymbol{\theta}}^{(t)}) - H^*) \bigg) \bigg\} \\
&\leq \mathbb{E}\bigg\{ \sum_{k=1}^{K} \frac{1}{2} \gamma_k^{(t)} p_k \bigg[\eta^{(t)} \left\| \nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}) \right\|^2 + \frac{1}{\eta^{(t)}} \left\| \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \right\|^2 \bigg] - \sum_{k=1}^{K} p_k \gamma_k^{(t)} (H_k(\bar{\boldsymbol{\theta}}^{(t)}) - H^*) \bigg\} \\
&\leq \mathbb{E}\bigg\{ \sum_{k=1}^{K} \gamma_k^{(t)} p_k \bigg[\eta^{(t)} (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta})) L (H_k(\bar{\boldsymbol{\theta}}^{(t)}) - H_k^*) + \frac{1}{2\eta^{(t)}} \left\| \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \right\|^2 \bigg] \\
&\qquad - \sum_{k=1}^{K} p_k \gamma_k^{(t)} (H_k(\bar{\boldsymbol{\theta}}^{(t)}) - H^*) \bigg\}.
\end{aligned}$$

In the second inequality, we again use the Cauchy–Schwarz inequality and Inequality of arithmetic and geometric means. In the last inequality, we use the fact that $\left\| \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \right\|^2 \leq 2(1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta})) L (H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*)$.

Since $\eta^{(t)} \leq \gamma_k^{(t)} \leq 2\eta^{(t)}$, we can bound E as

$$
\begin{aligned}
\mathrm{E} &\leq \mathrm{F} + \mathbb{E}\Big\{2\eta^{(t)}\Gamma_{max}\Big\} \\
&= (\eta^{(t)}(1 + \frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L - 1)\mathbb{E}\Big\{\sum_{k=1}^{K}\gamma_k^{(t)}p_k\Big[(H_k(\bar{\boldsymbol{\theta}}^{(t)}) - H^*)\Big]\Big\} \\
&\quad + \mathbb{E}\Big\{\sum_{k=1}^{K}\eta^{(t)}\gamma_k^{(t)}p_k(1 + \frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H^* - H_k^*)\Big\} \\
&\quad + \frac{1}{2\eta^{(t)}}\sum_{k=1}^{K}\gamma_k^{(t)}p_k\Big\{\Big\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\Big\|^2\Big\} + 2\eta^{(t)}\Gamma_{max} \\
&\leq \mathbb{E}\Big\{\sum_{k=1}^{K}\eta^{(t)}\gamma_k^{(t)}p_k(1 + \frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H^* - H_k^*)\Big\} + \frac{1}{2\eta^{(t)}}\sum_{k=1}^{K}\gamma_k^{(t)}p_k\mathbb{E}\Big\{\Big\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\Big\|^2\Big\} + 2\eta^{(t)}\Gamma_{max} \\
&\leq \sum_{k=1}^{K}p_k\mathbb{E}\Big\{\Big\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\Big\|^2\Big\} + \mathbb{E}\Big\{\sum_{k=1}^{K}\eta^{(t)}\gamma_k^{(t)}p_k(1 + \frac{d-1}{p_k|\mathcal{A}_{s_k}|}\lambda)L(H^* - H_k^*)\Big\} + 2\eta^{(t)}\Gamma_{max} \\
&\leq \sum_{k=1}^{K}p_k\mathbb{E}\Big\{\Big\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\Big\|^2\Big\} + 4\eta^{(t)2}L\mathbb{E}\Big\{\sum_{k=1}^{K}p_k(H^* - H_k^*)\Big\} + 2\eta^{(t)}\Gamma_{max} \\
&= \sum_{k=1}^{K}p_k\mathbb{E}\Big\{\Big\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\Big\|^2\Big\} + 4\eta^{(t)2}L\Gamma_K + 2\eta^{(t)}\Gamma_{max}
\end{aligned}
$$

The second inequality holds because $(\eta^{(t)}(1 + \frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L - 1) \leq 0$ and the fourth inequality uses the fact that $1 + \frac{d-1}{p_k|\mathcal{A}_{s_k}|}\lambda \leq 2$ based on the constraint of λ.

Therefore,

$$
\begin{aligned}
A \leq\ & \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\} - \eta^{(t)} \mathbb{E}\left\{ \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta})) \mu \left\| \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\} + \sum_{k=1}^{K} p_k \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|^2 + \mathrm{E} \\
\leq\ & 2 \sum_{k=1}^{K} p_k \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|^2 \right\} + 4\eta^{(t)2} L \Gamma_K + 2\eta^{(t)} \Gamma_{max} + \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\} \\
& - \eta^{(t)} \mathbb{E}\left\{ \sum_{k=1}^{K} p_k (1 - \frac{d-1}{p_k |\mathcal{A}_{s_k}|} \lambda) \mu \left\| \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\} \\
\leq\ & 2 \sum_{k=1}^{K} p_k \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|^2 \right\} + 4\eta^{(t)2} L \Gamma_K + 2\eta^{(t)} \Gamma_{max} + \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\} \\
& - \eta^{(t)} \mathbb{E}\left\{ \sum_{k=1}^{K} p_k^2 (1 - \frac{d-1}{p_k |\mathcal{A}_{s_k}|} \lambda) \mu \left\| \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\} \\
\leq\ & 2 \sum_{k=1}^{K} p_k \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|^2 \right\} + 4\eta^{(t)2} L \Gamma_K + 2\eta^{(t)} \Gamma_{max} + \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\} \\
& - \eta^{(t)} \mathbb{E}\left\{ (1 - \frac{d-1}{\min\{p_k |\mathcal{A}_{s_k}|\}} \lambda) \mu \frac{1}{K} \left\| \sum_{k=1}^{K} p_k \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\} \\
=\ & 2 \sum_{k=1}^{K} p_k \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|^2 \right\} + 4\eta^{(t)2} L \Gamma_K + 2\eta^{(t)} \Gamma_{max} + (1 - \eta^{(t)} (1 - \frac{d-1}{\min\{p_k |\mathcal{A}_{s_k}|\}} \lambda) \frac{\mu}{K}) \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^* \right\|^2 \right\}
\end{aligned}
$$

The third inequality uses the fact that $0 \leq p_k \leq 1$ and $-p_k^2 \geq -p_k$. The last inequality uses the fact that $\left\| \sum_{k=1}^{K} p_k \boldsymbol{\theta}_k \right\|^2 \leq K \sum_{k=1}^{K} \| p_k \boldsymbol{\theta}_k \|^2 = K \sum_{k=1}^{K} p_k^2 \| \boldsymbol{\theta}_k \|^2$ and $1 - \frac{d-1}{p_k |\mathcal{A}_{s_k}|} \lambda \geq 1 - \frac{d-1}{\min\{p_k |\mathcal{A}_{s_k}|\}} \lambda$.

Part II: Bounding Term $\sum_{k=1}^{K} p_k \mathbb{E}\left\{ \left\| \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} \right\|^2 \right\}$ in Term A For any iteration $t \geq 0$, denote by $t_0 \leq t$ the index of previous communication iteration before t. Since the FL algorithm requires one communication each E steps, we know $t - t_0 \leq E - 1$ and $\boldsymbol{\theta}_k^{(t_0)} = \bar{\boldsymbol{\theta}}^{(t_0)}$. Assume $\eta^{(t)} \leq 2\eta^{(t+E)}$. Since $\eta^{(t)}$ is decreasing, we have

$$\begin{aligned}
\mathbb{E}\left\{\sum_{k=1}^{K} p_k \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2\right\} &= \mathbb{E}\left\{\sum_{k=1}^{K} p_k \left\|(\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)}) - (\bar{\boldsymbol{\theta}}^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)})\right\|^2\right\} \\
&\leq \mathbb{E}\left\{\sum_{k=1}^{K} p_k \left\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)}\right\|^2\right\} \\
&= \mathbb{E}\left\{\sum_{k=1}^{K} p_k \left\|\sum_{t=0}^{t-1} \eta^{(t)} g_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)})\right\|^2\right\} \\
&\leq \mathbb{E}\left\{\sum_{k=1}^{K} p_k (t - t_0) \sum_{t=0}^{t-1} \eta^{(t)2} \left\|g_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)})\right\|^2\right\} \\
&\leq \sum_{k=1}^{K} p_k \sum_{t=t_0}^{t-1} (E-1)\eta^{(t)2} G^2 \leq \sum_{k=1}^{K} p_k \sum_{t=t_0}^{t-1} (E-1)\eta^{(t_0)2} G^2 \\
&\leq \sum_{k=1}^{K} p_k (E-1)^2 \eta^{(t_0)2} G^2 \leq 4\eta^{(t)2}(E-1)^2 G^2.
\end{aligned}$$

Part III: Bounding Term B By assumption, it is easy to show

$$\mathbb{E}\left\{\eta^{(t)2} \left\|\boldsymbol{g}^{(t)} - \bar{\boldsymbol{g}}^{(t)}\right\|^2\right\} \leq \eta^{(t)2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 \sigma_k^2.$$

Part IV: Proving Convergence So far, we have shown that

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^*\right\|^2\right\} \leq \mathrm{A} + \mathrm{B} \\
&\leq 8\eta^{(t)2}(E-1)^2 G^2 + 4\eta^{(t)2} L\Gamma_K + 2\eta^{(t)}\Gamma_{max} + (1 - \eta^{(t)}(1 - \frac{d-1}{p_k|\mathcal{A}_{s_k}|}\lambda)\mu)\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2\right\} \\
&\quad + \eta^{(t)2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k|\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 \sigma_k^2 \\
&= (1 - \eta^{(t)}(1 - \frac{d-1}{\min\{p_k|\mathcal{A}_{s_k}|\}}\lambda)\frac{\mu}{K})\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2\right\} + {\eta^{(t)}}^2 \xi
\end{aligned}$$

where $\xi = 8(E-1)^2 G^2 + 4L\Gamma_K + 2\frac{\Gamma_{max}}{\eta^{(t)}} + \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k|\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 \sigma_k^2$.

Let $\eta^{(t)} = \frac{\beta}{t+\gamma}$ with $\beta > \frac{1}{(1 - \frac{d-1}{min\{p_k|\mathcal{A}_{s_k}|\}}\lambda)\frac{\mu}{K}}$ and $\gamma > 0$. Define $\epsilon := (1 - \frac{d-1}{min\{p_k|\mathcal{A}_{s_k}|\}}\lambda)$. Let $v = \max\{\frac{\beta^2\xi}{\beta\epsilon\mu - 1}, (\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\right\|^2\}$. We will show that $\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2 \leq \frac{v}{\gamma+t}$ by induction. For $t = 0$, we have

$\left\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right\|^2 \leq (\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right\|^2 \leq \frac{v}{\gamma+1}$. Now assume this is true for some t, then

$$\begin{aligned}
\mathbb{E}\Big\{\left\|\bar{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*\right\|^2\Big\} &\leq (1-\eta^{(t)}\epsilon\mu)\mathbb{E}\Big\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\Big\}+\eta^{(t)^2}\xi \\
&\leq (1-\frac{\beta\epsilon\mu}{t+\gamma})\frac{v}{t+\gamma}+\frac{\beta^2\xi}{(t+\gamma)^2} \\
&= \frac{t+\gamma-1}{(t+\gamma)^2}v+\frac{\beta^2\xi}{(t+\gamma)^2}-\frac{\beta\epsilon\mu-1}{(t+\gamma)^2}v.
\end{aligned}$$

It is easy to show $\frac{t+\gamma-1}{(t+\gamma)^2}v+\frac{\beta^2\xi}{(t+\gamma)^2}-\frac{\beta\epsilon\mu-1}{(t+\gamma)^2}v \leq \frac{v}{t+\gamma+1}$ by definition of v. Therefore, we proved $\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2 \leq \frac{v}{\gamma+t}$.

By definition, we know H is $\sum_{k=1}^K p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2}L$-smooth. Therefore,

$$\begin{aligned}
\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\}-H^* &\leq \frac{\sum_{k=1}^K p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2}L}{2}\mathbb{E}\Big\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\Big\} \\
&\leq \frac{\sum_{k=1}^K p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2}L}{2}\frac{v}{\gamma+t}.
\end{aligned}$$

By choosing $\beta=\frac{2}{\epsilon\frac{\mu}{K}}$ We have

$$v=\max\{\frac{\beta^2\xi}{\beta\epsilon\mu-1},(\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right\|^2\} \leq \frac{\beta^2\xi}{\beta\epsilon\mu-1}+(\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right\|^2 \leq \frac{4\xi}{\epsilon^2\mu^2}+(\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right\|^2.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(T)})\Big\}-H^* &\leq \frac{\sum_{k=1}^K p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2}L}{2}\frac{1}{\gamma+T}\Big\{\frac{4\xi}{\epsilon^2\mu^2}+(\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right\|^2\Big\} \\
&\leq \frac{\sum_{k=1}^K p_k \frac{(1+\frac{\lambda(d-1)}{p_k|\mathcal{A}_{s_k}|})}{2}L}{2}\frac{1}{\gamma+T}\Big\{\frac{4\xi}{\epsilon^2\mu^2}+(\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right\|^2\Big\} \\
&\leq \frac{L}{2}\frac{1}{\gamma+T}\Big\{\frac{4\xi}{\epsilon^2\mu^2}+(\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right\|^2\Big\}.
\end{aligned}$$

We thus proved our convergence result. □

Theorem 3. *Assume at each communication round, central server sampled a fraction α of devices and those local devices are sampled according to the sampling probability p_k. Additionally, assume Assumptions in the main paper hold. For $\gamma,\mu,\epsilon>0$, we have*

$$\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(T)})\Big\}-H^* \leq \frac{L}{2}\frac{1}{\gamma+T}\Big\{\frac{4(\xi+\tau)}{\epsilon^2\mu^2}+(\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\right\|^2\Big\},$$

$\tau = \frac{E^2}{\lceil \alpha K \rceil} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 G^2$.

Proof.

$$\begin{aligned}
\mathbb{E}\Big\{ \left\| \bar{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^* \right\|^2 \Big\} &= \mathbb{E}\Big\{ \left\| \bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)} + \bar{\boldsymbol{w}}^{(t+1)} - \boldsymbol{\theta}^* \right\|^2 \Big\} \\
&= \mathbb{E}\Big\{ \left\| \bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)} \right\|^2 + \left\| \bar{\boldsymbol{w}}^{(t+1)} - \boldsymbol{\theta}^* \right\|^2 + 2\langle \bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)}, \bar{\boldsymbol{w}}^{(t+1)} - \boldsymbol{\theta}^* \rangle \Big\} \\
&= \mathbb{E}\Big\{ \left\| \bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)} \right\|^2 + \left\| \bar{\boldsymbol{w}}^{(t+1)} - \boldsymbol{\theta}^* \right\|^2 \Big\}.
\end{aligned}$$

Note that the expectation is taken over subset $\mathcal{S}_c$.

Part I: Bounding Term $\mathbb{E}\Big\{ \left\| \bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)} \right\|^2 \Big\}$ Assume $\lceil \alpha K \rceil$ number of local devices are sampled according to sampling probability p_k. During the communication round, we have $\bar{\boldsymbol{\theta}}^{t+1} = \frac{1}{\lceil \alpha K \rceil} \sum_{l=1}^{\lceil \alpha K \rceil} \boldsymbol{w}_l^{(t+1)}$. Therefore,

$$\begin{aligned}
\mathbb{E}\Big\{ \left\| \bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)} \right\|^2 \Big\} &= \mathbb{E}\Big\{ \frac{1}{\lceil \alpha K \rceil^2} \left\| \sum_{l=1}^{\lceil \alpha K \rceil} \boldsymbol{w}_l^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)} \right\|^2 \Big\} \\
&= \mathbb{E}\Big\{ \frac{1}{\lceil \alpha K \rceil^2} \sum_{l=1}^{\lceil \alpha K \rceil} \left\| \boldsymbol{w}_l^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)} \right\|^2 \Big\} \\
&= \frac{1}{\lceil \alpha K \rceil} \sum_{k=1}^{K} p_k \left\| \boldsymbol{w}_k^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)} \right\|^2.
\end{aligned}$$

We know

$$\sum_{k=1}^{K} p_k \left\| \boldsymbol{w}_k^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)} \right\|^2 = \sum_{k=1}^{K} p_k \left\| (\boldsymbol{w}_k^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t_0)}) - (\bar{\boldsymbol{w}}^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t_0)}) \right\|^2 \leq \sum_{k=1}^{K} p_k \left\| (\boldsymbol{w}_k^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t_0)}) \right\|^2,$$

where $t_0 = t - E + 1$. Similarly,

$$\begin{aligned}
\mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)}\right\|^2 \Big\} &\leq \frac{1}{\lceil \alpha K \rceil}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k \left\|(\boldsymbol{w}_k^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t_0)})\right\|^2 \Big\} \\
&\leq \frac{1}{\lceil \alpha K \rceil}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k \left\|(\boldsymbol{w}_k^{(t+1)} - \boldsymbol{\theta}_k^{(t_0)})\right\|^2 \Big\} \\
&\leq \frac{1}{\lceil \alpha K \rceil}\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k E \sum_{m=t_o}^{t} \left\|\eta^{(m)} \nabla H_k(\boldsymbol{\theta}_k^{(m)}; \zeta_k^{(t)})\right\|^2 \Big\} \\
&\leq \frac{E^2 \eta^{(t_0)2}}{\lceil \alpha K \rceil} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 G^2 \\
&\leq \frac{E^2 \eta^{(t)2}}{\lceil \alpha K \rceil} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 G^2
\end{aligned}$$

using the fact that $\eta^{(t)}$ is non-increasing in t.

Part II: Convergence Result As aforementioned,

$$\begin{aligned}
\mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^*\right\|^2 \Big\} &= \mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{w}}^{(t+1)}\right\|^2 + \left\|\bar{\boldsymbol{w}}^{(t+1)} - \boldsymbol{\theta}^*\right\|^2 \Big\} \\
&\leq \frac{E^2 \eta^{(t)2}}{\lceil \alpha K \rceil} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 G^2 + (1 - \eta^{(t)} \epsilon \frac{\mu}{K}) \mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2 \Big\} + {\eta^{(t)}}^2 \xi \\
&= (1 - \eta^{(t)} \epsilon \frac{\mu}{K}) \mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2 \Big\} + {\eta^{(t)}}^2 \Big(\xi + \frac{E^2}{\lceil \alpha K \rceil} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 G^2 \Big).
\end{aligned}$$

Let $\tau = \frac{E^2}{\lceil \alpha K \rceil} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 G^2$. Let $\eta^{(t)} = \frac{\beta}{t+\gamma}$ with $\beta > \frac{1}{\epsilon \frac{\mu}{K}}$ and $\gamma > 0$. Let $v = \max\{\frac{\beta^2(\xi+\tau)}{\beta\epsilon\mu - 1}, (\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\right\|^2\}$. Similar to the full device participation scenario, we can show that $\mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2 \Big\} \leq \frac{v}{\gamma+t}$ by induction.

By definition, we know H is $\sum_{k=1}^{K} p_k \frac{(1+\frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))}{2} L$-smooth. Therefore,

$$\begin{aligned}
\mathbb{E}\Big\{ H(\bar{\boldsymbol{\theta}}^{(t)}) \Big\} - H^* &\leq \frac{\sum_{k=1}^{K} p_k \frac{(1+\frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))}{2} L}{2} \mathbb{E}\Big\{ \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2 \Big\} \\
&\leq \frac{\sum_{k=1}^{K} p_k \frac{(1+\frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))}{2} L}{2} \frac{v}{\gamma+t}.
\end{aligned}$$

By choosing $\beta = \frac{2}{\epsilon \frac{\mu}{K}}$ We have

$$v = \max\{\frac{\beta^2 \xi}{\beta\epsilon\mu - 1}, (\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\right\|^2\} \leq \frac{\beta^2 \xi}{\beta\epsilon\mu - 1} + (\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\right\|^2 \leq \frac{4\xi}{\epsilon^2\mu^2} + (\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\right\|^2.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(T)})\Big\} - H^* &\leq \frac{\sum_{k=1}^{K} p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2} L}{2} \frac{1}{\gamma+T}\Big\{\frac{4(\xi+\tau)}{\epsilon^2\mu^2} + (\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\right\|^2\Big\} \\
&\leq \frac{L}{2}\frac{1}{\gamma+T}\Big\{\frac{4(\xi+\tau)}{\epsilon^2\mu^2} + (\gamma+1)\left\|\bar{\boldsymbol{\theta}}^{(0)} - \boldsymbol{\theta}^*\right\|^2\Big\}
\end{aligned}$$

□

3.4 Convergence (Non-convex)

Lemma 2. *If* $\eta^{(t)} \leq \frac{2}{L}$, *then* $\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} \leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)})\Big\}$.

Proof.

$$\begin{aligned}
\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t+1)})\Big\} &= \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)}\sum_{k=1}^{K} p_k \nabla H_k(\boldsymbol{\theta}_k^{(t)};\zeta_k^{(t)}))\Big\} \\
&= \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)}\sum_{k=1}^{K} p_k \nabla H_k(\bar{\boldsymbol{\theta}}^{(t)};\zeta_k^{(t)}))\Big\} \\
&= \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)} g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}))\Big\}
\end{aligned}$$

Here we used the fact that $\bar{\boldsymbol{\theta}}^{(t)} = \boldsymbol{\theta}_k^{(t)}$ since the aggregated model parameter has been distributed to local devices. By Taylor's theorem, there exists a $\boldsymbol{w}^{(t)}$ such that

$$\begin{aligned}
\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t+1)})\Big\} &= \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)}) - \eta^{(t)} g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)})^T g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}) + \frac{1}{2}(\eta^{(t)} g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}))^T g^{(t)}(\boldsymbol{w}^{(t)})(\eta^{(t)} g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}))\Big\} \\
&\leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)}) - \eta^{(t)} g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)})^T g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}) + \eta^{(t)2}\frac{\sum_{k=1}^{K} p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2} L}{2}\left\|g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}\right\|^2\Big\} \\
&\leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} - \eta^{(t)}\left\|g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}\right\|^2 + \eta^{(t)2}\frac{L}{2}\left\|g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}\right\|^2
\end{aligned}$$

since H is $\sum_{k=1}^{K} p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2} L$-smooth. It can be shown that if $\eta^{(t)} \leq \frac{2}{L}$, we have

$$-\eta^{(t)}\left\|g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}\right\|^2 + \eta^{(t)2}\frac{L}{2}\left\|g^{(t)}(\bar{\boldsymbol{\theta}}^{(t)}\right\|^2 \leq 0.$$

Therefore, By choosing $\eta^{(t)} \leq \frac{2}{L}$, we proved $\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} \leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)})\Big\}$. □

Theorem 4. *Assume Assumptions in the main paper hold and* $|\mathcal{S}_c| = K$. *If* $\eta^{(t)} = \mathcal{O}(\frac{1}{\sqrt{t}})$ *and* $\eta^{(t)} \leq \mathcal{O}(\frac{1}{L})$, *then for* > 0

$$\min_{t=1,\dots,T}\mathbb{E}\left\{\left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} \le \frac{1}{\sqrt{T}}\left\{2(1+2KL^2\sum_{t=1}^{T}\eta^{(t)2})\mathbb{E}\left\{H(\bar{\boldsymbol{\theta}}^{(0)})-H^*\right\}+2\sum_{t=1}^{T}\xi^{(t)}\right\},$$

where $\xi^{(t)} = 2KL^2\eta^{(t)2}\Gamma_K + (8\eta^{(t)3}KL^2(E-1) + 8KL\eta^{(t)2} + 4(2+4L)KL\eta^{(t)4}(E-1))G^2 + (2L\eta^{(t)2} + 8KL\eta^{(t)2})\sum_{k=1}^{K}p_k\sigma_k^2$

Proof. Since H is $\sum_{k=1}^{K}p_k\frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2}L$-smooth, we have

$$\mathbb{E}\left\{H(\bar{\boldsymbol{\theta}}^{(t+1)})\right\} \le \mathbb{E}\left\{H(\bar{\boldsymbol{\theta}}^{(t)})\right\} + \underbrace{\mathbb{E}\left\{\langle\nabla H(\bar{\boldsymbol{\theta}}^{(t)}), \bar{\boldsymbol{\theta}}^{(t+1)}-\bar{\boldsymbol{\theta}}^{(t)}\rangle\right\}}_{\text{A}} +$$

$$\frac{\sum_{k=1}^{K}p_k\frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2}L}{2}\underbrace{\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t+1)}-\bar{\boldsymbol{\theta}}^{(t)}\right\|^2\right\}}_{\text{B}}.$$

Part I: Bounding Term A We have

$$\begin{aligned}
\text{A} &= -\eta^{(t)}\mathbb{E}\left\{\langle\nabla H(\bar{\boldsymbol{\theta}}^{(t)}), \sum_{k=1}^{K}p_k\nabla H_k(\boldsymbol{\theta}_k^{(t)};\zeta_k^{(t)})\rangle\right\} = -\eta^{(t)}\mathbb{E}\left\{\langle\nabla H(\bar{\boldsymbol{\theta}}^{(t)}), \sum_{k=1}^{K}p_k\nabla H_k(\boldsymbol{\theta}_k^{(t)})\rangle\right\} \\
&= -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K}p_k\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2\right\} + \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)}) - \sum_{k=1}^{K}p_k\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2\right\} \\
&= -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K}p_k\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2\right\} + \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K}p_k\nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}) - \sum_{k=1}^{K}p_k\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2\right\} \\
&\le -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K}p_k\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2\right\} + \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{K\sum_{k=1}^{K}p_k\left\|\nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}) - \nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2\right\} \\
&\le -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K}p_k\nabla H_k(\boldsymbol{\theta}_k^{(t)})\right\|^2\right\} + \\
&\quad \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{K\sum_{k=1}^{K}p_k((1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L)^2\underbrace{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}_k^{(t)}\right\|^2}_{\text{C}}\right\}.
\end{aligned}$$

In the convex setting, we proved that

$$\text{C} \le 4\eta^{(t)2}(E-1)G^2.$$

This is also true for the non-convex setting since we do not use any property of convex functions.

Part II: Bounding Term B We have

$$
\begin{aligned}
\mathrm{B} = \mathbb{E}\Big\{ \left\| \eta^{(t)} g^{(t)} \right\|^2 \Big\} &= \mathbb{E}\Big\{ \left\| \eta^{(t)} \sum_{k=1}^{K} p_k \nabla H_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) \right\|^2 \Big\} \\
&= \mathbb{E}\Big\{ \left\| \eta^{(t)} \sum_{k=1}^{K} p_k (\nabla H_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) - \nabla H_k(\boldsymbol{\theta}_k^{(t)})) \right\|^2 \Big\} + \mathbb{E}\Big\{ \left\| \eta^{(t)} \sum_{k=1}^{K} p_k \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \right\|^2 \Big\} \\
&= \eta^{(t)2} \sum_{k=1}^{K} p_k^2 \mathbb{E}\Big\{ \left\| \nabla H_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) - \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \right\|^2 \Big\} + \mathbb{E}\Big\{ \left\| \eta^{(t)} \sum_{k=1}^{K} p_k \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \right\|^2 \Big\} \\
&\leq \eta^{(t)2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 \sigma_k^2 + \eta^{(t)2} \mathbb{E}\Big\{ K \sum_{k=1}^{K} p_k^2 \left\| \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \right\|^2 \Big\}.
\end{aligned}
$$

Since H_k is $(1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))L$-smooth, we know

$$
\left\| \nabla H_k(\boldsymbol{\theta}_k^{(t)}) \right\|^2 \leq 2(1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta})) L (H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*).
$$

Therefore,

$$
\begin{aligned}
\mathrm{B} \leq\ & \eta^{(t)2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 \sigma_k^2 + \\
& \quad \eta^{(t)2} \mathbb{E}\Big\{ K \sum_{k=1}^{K} 2 p_k^2 (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta})) L (H_k(\boldsymbol{\theta}_k^{(t)}) - H_k^*) \Big\} \\
=\ & \eta^{(t)2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 \sigma_k^2 + \\
& \quad \eta^{(t)2} \mathbb{E}\Big\{ K \sum_{k=1}^{K} 2 p_k^2 (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta})) L (H_k(\boldsymbol{\theta}_k^{(t)}) - H^* + H^* - H_k^*) \Big\} \\
\leq\ & \eta^{(t)2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 \sigma_k^2 + \\
& \quad \eta^{(t)2} \mathbb{E}\Big\{ K \sum_{k=1}^{K} 2 p_k (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta})) L (H_k(\boldsymbol{\theta}_k^{(t)}) - H^* + H^* - H_k^*) \Big\}
\end{aligned}
$$

since $0 \leq p_k \leq 1$ and $p_k^2 \leq p_k$.

Therefore,

$$
\begin{aligned}
\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t+1)})\Big\} \leq\ & \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{\Big\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\Big\|^2\Big\} \underbrace{-\frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{\Big\|\sum_{k=1}^{K} p_k \nabla H_k(\boldsymbol{\theta}_k^{(t)})\Big\|^2\Big\}}_{\text{D<0}} + \\
& \frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{K\sum_{k=1}^{K} p_k((1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L)^2 4\eta^{(t)2}(E-1)G^2\Big\} \\
& + \frac{\sum_{k=1}^{K} p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2} L}{2}\Big[\eta^{(t)2}\sum_{k=1}^{K} p_k^2(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))^2\sigma_k^2 \\
& + \eta^{(t)2}\mathbb{E}\Big\{K\sum_{k=1}^{K} 2p_k(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L(H_k(\boldsymbol{\theta}_k^{(t)}) - H^* + H^* - H_k^*)\Big\}\Big] \\
\leq\ & \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{\Big\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\Big\|^2\Big\} \\
& \frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{K\sum_{k=1}^{K} p_k((1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))L)^2 4\eta^{(t)2}(E-1)G^2\Big\} \\
& + \frac{\sum_{k=1}^{K} p_k \frac{(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))}{2} L}{2}\Big[\eta^{(t)2}\sum_{k=1}^{K} p_k^2(1+\frac{\lambda}{p_k|\mathcal{A}_{s_k}|}r_k(\boldsymbol{\theta}))^2\sigma_k^2 + \\
& \underbrace{4KL\eta^{(t)2}\mathbb{E}\Big\{\sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H^*) + \sum_{k=1}^{K} p_k(H^* - H_k^*)\Big\}}_{\text{E}}\Big] \\
\leq\ & \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{\Big\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\Big\|^2\Big\} + \frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{K\sum_{k=1}^{K} 4p_k L^2 4\eta^{(t)2}(E-1)G^2\Big\} \\
& + \frac{L}{2}\Big[\eta^{(t)2}\sum_{k=1}^{K} 4p_k^2\sigma_k^2 + \underbrace{4KL\eta^{(t)2}\mathbb{E}\Big\{\sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H^*) + \sum_{k=1}^{K} p_k(H^* - H_k^*)\Big\}}_{\text{E}}\Big]
\end{aligned}
$$

Here

$$
\begin{aligned}
\text{E} &= 4KL\eta^{(t)2}\mathbb{E}\Big\{\sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H^*)\Big\} + 4KL\eta^{(t)2}\mathbb{E}\Big\{\sum_{k=1}^{K} p_k(H^* - H_k^*)\Big\} \\
&= 4KL\eta^{(t)2}\mathbb{E}\Big\{\sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k(\bar{\boldsymbol{\theta}}^{(t)}))\Big\} + 4KL\eta^{(t)2}\mathbb{E}\Big\{\sum_{k=1}^{K} p_k(H_k(\bar{\boldsymbol{\theta}}^{(t)}) - H^*)\Big\} + 4KL\eta^{(t)2}\Gamma_K \\
&= 4KL\eta^{(t)2}\underbrace{\mathbb{E}\Big\{\sum_{k=1}^{K} p_k(H_k(\boldsymbol{\theta}_k^{(t)}) - H_k(\bar{\boldsymbol{\theta}}^{(t)}))\Big\}}_{\text{F}} + 4KL\eta^{(t)2}\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)}) - H^*\Big\} + 4KL\eta^{(t)2}\Gamma_K.
\end{aligned}
$$

We can bound term F as

$$
\begin{aligned}
\mathrm{F} &= \mathbb{E}\Big\{ \sum_{k=1}^{K} p_k (H_k(\boldsymbol{\theta}_k^{(t)}) - H_k(\bar{\boldsymbol{\theta}}^{(t)})) \Big\} \\
&\leq \mathbb{E}\Big\{ \sum_{k=1}^{K} p_k (\langle \nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}), \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \rangle + \frac{(1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))L}{2} \underbrace{\left\| \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \right\|^2}_{\leq 4\eta^{(t)2}(E-1)G^2}) \Big\}
\end{aligned}
$$

where we use the fact that H_k is $(1 + \frac{\lambda}{p_k|\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))L$-smooth. To bound the inner product, we again use the inequality of arithmetic and geometric means and Cauchy–Schwarz inequality:

$$
\langle \nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}), \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \rangle \leq \left\| \nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}) \right\| \left\| \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \right\| \leq \frac{\left\| \nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}) \right\|^2 + \left\| \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \right\|^2}{2}.
$$

It can be shown that

$$
\begin{aligned}
\mathbb{E}\Big\{ \left\| \nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}) \right\|^2 \Big\} &= \mathbb{E}\Big\{ \left\| \nabla F_k(\boldsymbol{\theta}_k^{(t)}, D_k^{(t)}) \right\| \Big\}^2 + \mathbb{E}\Big\{ \left\| \nabla F_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) - \nabla F_k(\boldsymbol{\theta}_k^{(t)}) \right\|^2 \Big\} \\
&\leq \mathbb{E}\Big\{ \left\| \nabla F_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) \right\|^2 \Big\} + \mathbb{E}\Big\{ \left\| \nabla F_k(\boldsymbol{\theta}_k^{(t)}; \zeta_k^{(t)}) - \nabla F_k(\boldsymbol{\theta}_k^{(t)}) \right\|^2 \Big\} \\
&\leq (1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))^2 (G^2 + \sigma_k^2) \leq 4(G^2 + \sigma_k^2)
\end{aligned}
$$

Therefore, we can simplify F as

$$
\begin{aligned}
\mathrm{F} &\leq \mathbb{E}\Big\{ \sum_{k=1}^{K} p_k (\frac{\left\| \nabla H_k(\bar{\boldsymbol{\theta}}^{(t)}) \right\|^2 + \left\| \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \right\|^2}{2} + \frac{(1 + \frac{\lambda}{p_k |\mathcal{A}_{s_k}|} r_k(\boldsymbol{\theta}))L}{2} \underbrace{\left\| \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)} \right\|^2}_{\leq 4\eta^{(t)2}(E-1)G^2}) \Big\} \\
&\leq \mathbb{E}\Big\{ \sum_{k=1}^{K} p_k (\frac{4(G^2 + \sigma_k^2) + 4\eta^{(t)2}(E-1)G^2}{2} + 4L\eta^{(t)2}(E-1)G^2) \Big\} \\
&= 2\mathbb{E}\Big\{ \sum_{k=1}^{K} p_k \sigma_k^2 \Big\} + 2G^2 + (2 + 4L)\eta^{(t)2}(E-1)G^2
\end{aligned}
$$

Combining with E, we obtain

$$
\mathrm{E} \leq 4KL\eta^{(t)2} \Big(2\sum_{k=1}^{K} p_k \sigma_k^2 + 2G^2 + (2 + 4L)\eta^{(t)2}(E-1)G^2 \Big) + 4KL\eta^{(t)2}\mathbb{E}\Big\{ H(\bar{\boldsymbol{\theta}}^{(t)}) - H^* \Big\} + 4KL\eta^{(t)2}\Gamma_K
$$

Part III: Proving Convergence Therefore,

$$
\begin{aligned}
&\frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{ \left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2 \Big\} \\
&\leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} - \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t+1)})\Big\} + \frac{1}{2}\eta^{(t)3}\mathbb{E}\Big\{K\sum_{k=1}^{K} 4p_k L^2 4(E-1)G^2\Big\} + \\
&\frac{L}{2}\Big[\eta^{(t)2}\sum_{k=1}^{K} 4p_k^2\sigma_k^2 + 4KL\eta^{(t)2}\Big(2\sum_{k=1}^{K} p_k\sigma_k^2 + 2G^2 + (2+4L)\eta^{(t)2}(E-1)G^2\Big) + 4KL\eta^{(t)2}\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)}) - H^*\Big\} \\
&\qquad + 4KL\eta^{(t)2}\Gamma_K\Big] \\
&= \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} - \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t+1)})\Big\} + 2KL^2\eta^{(t)2}\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)}) - H^*\Big\} + 2KL^2\eta^{(t)2}\Gamma_K \\
&\qquad + (8\eta^{(t)3}KL^2(E-1) + 8KL\eta^{(t)2} + 4(2+4L)KL\eta^{(t)4}(E-1))G^2 + (2L\eta^{(t)2} + 8KL\eta^{(t)2})\sum_{k=1}^{K} p_k\sigma_k^2.
\end{aligned}
$$

Let $\xi^{(t)} = 2KL^2\eta^{(t)2}\Gamma_K + (8\eta^{(t)3}KL^2(E-1) + 8KL\eta^{(t)2} + 4(2+4L)KL\eta^{(t)4}(E-1))G^2 + (2L\eta^{(t)2} + 8KL\eta^{(t)2})\sum_{k=1}^{K} p_k\sigma_k^2$, then

$$
\begin{aligned}
\frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{ \left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2 \Big\} &\leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} - \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t+1)})\Big\} + 2KL^2\eta^{(t)2}\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)}) - H^*\Big\} + \xi^{(t)} \\
&\leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} - \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t+1)})\Big\} + 2KL^2\eta^{(t)2}\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)}) - H^*\Big\} + \xi^{(t)}
\end{aligned}
$$

since $\eta^{(t)} \leq \frac{1}{\sqrt{2KL}}$ and $\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t)})\Big\} \leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)})\Big\}$ by Lemma 2. By taking summation on both side, we obtain

$$
\begin{aligned}
\sum_{t=1}^{T}\frac{1}{2}\eta^{(t)}\mathbb{E}\Big\{ \left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2 \Big\} &\leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)})\Big\} - \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(t+1)})\Big\} + 2KL^2\sum_{t=1}^{T}\eta^{(t)2}\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)}) - H^*\Big\} + \sum_{t=1}^{T}\xi^{(t)} \\
&\leq \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)})\Big\} - \mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{*})\Big\} + 2KL^2\sum_{t=1}^{T}\eta^{(t)2}\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)}) - H^*\Big\} + \sum_{t=1}^{T}\xi^{(t)} \\
&= (1 + 2KL^2\sum_{t=1}^{T}\eta^{(t)2})\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)}) - H^*\Big\} + \sum_{t=1}^{T}\xi^{(t)}.
\end{aligned}
$$

This implies

$$
\min_{t=1,\ldots,T}\mathbb{E}\Big\{ \left\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2 \Big\}\sum_{t=1}^{T}\eta^{(t)} \leq 2(1 + 2KL^2\sum_{t=1}^{T}\eta^{(t)2})\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)}) - H^*\Big\} + 2\sum_{t=1}^{T}\xi^{(t)}
$$

and therefore

$$\min_{t=1,\dots,T} \mathbb{E}\Big\{ \Big\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\Big\|^2 \Big\} \leq \frac{1}{\sum_{t=1}^{T}\eta^{(t)}} \Big\{ 2(1+2KL^2\sum_{t=1}^{T}\eta^{(t)2})\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)}) - H^*\Big\} + 2\sum_{t=1}^{T}\xi^{(t)}\Big\}.$$

Let $\eta^{(t)} = \frac{1}{\sqrt{t}}$, then we have $\sum_{t=1}^{T}\eta^{(t)} = \mathcal{O}(\sqrt{T})$ and $\sum_{t=1}^{T}\eta^{(t)2} = \mathcal{O}(\log(T+1))$. Therefore,

$$\min_{t=1,\dots,T} \mathbb{E}\Big\{ \Big\|\nabla H(\bar{\boldsymbol{\theta}}^{(t)})\Big\|^2 \Big\} \leq \frac{1}{\sqrt{T}} \Big\{ 2(1+2KL^2\sum_{t=1}^{T}\eta^{(t)2})\mathbb{E}\Big\{H(\bar{\boldsymbol{\theta}}^{(0)}) - H^*\Big\} + 2\sum_{t=1}^{T}\xi^{(t)}\Big\}.$$

□

4 Additional Experiments

We conduct a sensitivity analysis using the FEMNIST-3-groups setting. Results are reported in Figure 1. Similar to the observation in the main paper, it can be seen that as λ increases, the discrepancy between two groups decreases accordingly. Here kindly note that we did not plot group 3 for the sake of neatness. The line of group should stay in the middle of two lines.

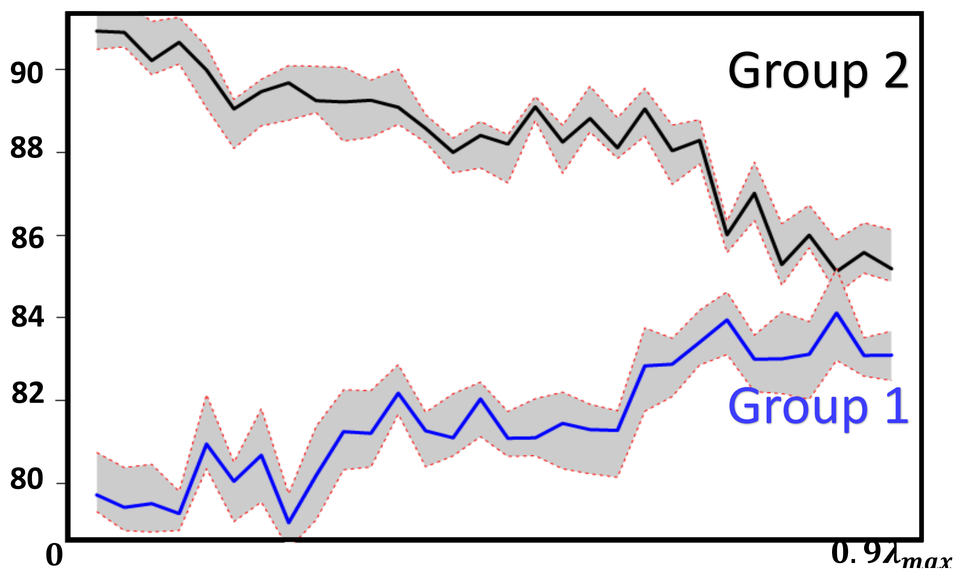

Figure 1: Sensitivity analysis on FEMNIST

References

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. (2019). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

Mohri, M., Sivek, G., and Suresh, A. T. (2019). Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR.