VII. Proof of Theorem 1

McDiarmid's Inequality

Let $Z_1, \dots, Z_m$ be independent random variables and let $f : \mathcal{Z}^m \to \mathbb{R}$ be a function satisfying the bounded difference property, i.e., for all i and for all $z_1, \dots, z_m, z_i'$ in $\mathcal{Z}$,

$$|f(z_1, \dots, z_{i-1}, z_i, z_{i+1}, \dots, z_m) - f(z_1, \dots, z_{i-1}, z_i', z_{i+1}, \dots, z_m)| \leq c_i,$$

where c_i are non-negative constants. Then, for any $a > 0$, the following inequality holds:

$$\mathbb{P}\left(|f(Z_1, \dots, Z_m) - \mathbb{E}[f(Z_1, \dots, Z_m)]| \geq a\right) \leq 2 \exp\left(-\frac{2a^2}{\sum_{i=1}^m c_i^2}\right).$$

Hoeffding's Inequality

Let $X_1, \dots, X_n$ be independent random variables with X_i taking values in the interval $[a_i, b_i]$ almost surely. Define $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and let $\mu = \mathbb{E}[\overline{X}]$. Then, for any $t > 0$, the following inequality holds:

$$\mathbb{P}(|\overline{X} - \mu| \geq t) \leq 2 \exp\left(-\frac{2nt^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Assumption 6 (Bounded Loss)**.** *W.L.O.G. for client i, $\forall \xi_i^m, \xi_i^n$, there exists a constant c, such that $|F_i(w, \xi_i^m) - F_i(w, \xi_i^n)| < c$.*

Definition 4 (Rademacher Complexity)**.** *Let $\mathcal{F} = F \circ w \triangleq \{\xi \to F(w, \xi) \mid w \in \mathcal{W}\}$, for any given dataset of FL $S = (\xi^1, ..., \xi^n)$, the set of loss mappings is denoted as $\mathcal{F}_{|S} \triangleq \{(F(\xi^1), ..., F(\xi^n)) \mid F \in \mathcal{F}\}$, the Rademacher Complexity is then defined as follows:*

$$\mathcal{R}(\mathcal{F}_{|S}) = \frac{1}{N}\mathbb{E}_{\tau\sim\{\pm1\}^n} \sup_w \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \tau F_i(\xi_i^j), \tag{15}$$

where τ has an equal probability of 0.5 of being either 1 or -1.

Definition 5 (Representation of Generalization Error)**.** *Given the data distribution D and the training data S, the representation of generalization error is defined as the supremum of the difference between the expected risk and the empirical risk over the hypothesis space $\mathcal{W}$. Formally, it is expressed as:*

$$Rep_D(S; \mathcal{W}, \mathcal{F}) = \sup_{w\in\mathcal{W}} (F_D(w) - F_S(w)), \tag{16}$$

where $F_{\mathcal{D}}(w) = \mathbb{E}_{\tilde{\xi}\sim\mathcal{D}} F(w, \tilde{\xi})$, and $F_S(w) = \frac{1}{|S|}\sum_{\xi\in S} F(w, \xi)$.

Lemma 1. *Consider a data distribution $\mathcal{D}$ and any train set S. $\mathbb{E}_{S\sim D} Rep_{\mathcal{D}(S;\mathcal{F})} \leq 2\mathbb{E}_{S\sim D} R(\mathcal{F}_{|\mathcal{S}})$.*

Proof. Given any dataset $S = (\xi_1, \xi_2, ..., \xi_n) \sim \mathcal{D}$, $S' = (\xi_1^{'}, \xi_2^{'}, ..., \xi_n^{'}) \sim D$.

$$\begin{aligned}
\mathbb{E}_S Rep_{\mathcal{D}}(S; \mathcal{W}, \mathcal{F}) &= \mathbb{E}_S\left[\sup_{w\in\mathcal{W}}\{F_{\mathcal{D}}(w) - F_{\mathcal{S}}(w)\}\right], \\
&= \mathbb{E}_S\left[\sup_{w\in\mathcal{W}}\{\mathbb{E}_{S'}[F_{S'}(w)] - F_S(w)\}\right], \\
&= \mathbb{E}_{S,S'} \sup_{w\in\mathcal{W}}\{\frac{1}{n}\sum_{i=1}^n (F(\xi_i') - F(\xi_i))\}, \\
&= \mathbb{E}_{S,S',\tau\sim\{\pm1\}^n} \sup_{w\in\mathcal{W}}\{\frac{1}{n}\sum_{i=1}^n (\tau_i F(\xi_i') - \tau_i F(\xi_i))\}, \\
&\leq \mathbb{E}_{S',\tau} \sup_F\{\frac{1}{n}\sum_{i=1}^n \tau_i F(\xi_i')\} + \mathbb{E}_{S,\tau}\sup_F\{\frac{1}{n}\sum_{i=1}^m \tau_i F(\xi_i)\}, \\
&= 2R(\mathcal{F}_{|\mathcal{S}})
\end{aligned}$$

□

Theorem 3 (Bounded Generalization Error)**.** *Consider a FL system with N clients, and parameter hypothesis space $\mathcal{W}$. If Assumption 6 holds. Then, for $\forall \delta \in [0,1]$ with probability of at least $1-2\delta$, for $\forall w \in \mathcal{W}$, the generalization error can be upper bounded as:*

$$Rep_D(S;\mathcal{W},\mathcal{F}) \leq 2\mathcal{R}(\mathcal{F}_{|S}) + 3c\sqrt{\frac{2\ln\frac{2}{\delta}}{N}}. \tag{17}$$

Proof. Let's start with McDiarmid's Inequality, let $f(S) = Rep_D(S;\mathcal{F})$, we have

$$\mathbb{P}\left(\mathrm{Rep}_{\mathcal{D}}(S;\mathcal{F}) - \mathbb{E}[\mathrm{Rep}_{\mathcal{D}}(S;\mathcal{F})] \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{mc^2}\right). \tag{18}$$

Set $\delta = \exp\left(-\frac{2\epsilon^2}{Nc^2}\right)$, we then have

$$\epsilon = c\sqrt{\frac{2\ln(2/\delta)}{N}}, \tag{19}$$

Next, we consider empirical Rademacher complexity. According to 1, with a probability of at least $1-\delta$, we can obtain

$$Rep_{\mathcal{D}}(S;\mathcal{F}) \leq 2\mathbb{E}_S R(\mathcal{F}|_S) + c\sqrt{\frac{2\ln(2/\delta)}{N}}. \tag{20}$$

Then consider $\mathbb{E}_S R(\mathcal{F}|_S)$, given $c' = \frac{c}{2}\sqrt{N}$, with a probability of at least $1-\delta$, we have

$$\mathbb{E}_S R(\mathcal{F}|_S) \leq R(\mathcal{F}|_S) + c\sqrt{\frac{2\ln(2/\delta)}{N}}. \tag{21}$$

Combine 20 and 21, and thus with a probability of at least $1-2\delta$, we have

$$Rep_{\mathcal{D}}(S;\mathcal{F}) \leq 2R(\mathcal{F}|_S) + 3c\sqrt{\frac{2\ln(2/\delta)}{N}}. \tag{22}$$

□

Corollary 2 (Bounded Generalization Error)**.** *Under the same conditions as Theorem 3, let S, D be the training set and generalization data distribution, $w_S = ERM_w(S) = argmin_{w\in\mathcal{W}} F(w;S)$, $w^* = argmin_{w\in\mathcal{W}} F(w;D)$. Then, for $\forall\delta \in [0,1]$ with probability of at least $1-3\delta$, for $\forall w \in \mathcal{W}$, the generalization error can be bounded as:*

$$F_{\mathcal{D}}(w_S) - F_{\mathcal{D}}(w^*) \leq 2\sup_{w\in\mathcal{W}} Var(F) + 4c\sqrt{\frac{2\ln(2/\delta)}{N}}, \tag{23}$$

where $Var(F) = \sqrt{\frac{1}{N}\sum_{i=1}^N (F_i(w) - \frac{1}{N}\sum_{i=1}^N F_i(w))^2}$.

Proof. Denote w_S, w^* as $w_S = ERM_{\mathcal{W}}(S) = \arg\min_{w\in\mathcal{W}}\{F_S(w)\}, w^* = \arg\min_{w\in\mathcal{W}}\{F_D(w)\}$. Then we can rewrite generalization error as

$$F_{\mathcal{D}}(w_S) - F_{\mathcal{D}}(w^*) = \underbrace{[F_{\mathcal{D}}(w_S) - F_S(w_S)]}_{P_1} + \underbrace{[F_S(w_S) - F_S(w^*)]}_{P_2} + \underbrace{[F_S(w^*) - F_{\mathcal{D}}(w^*)]}_{P_3}. \tag{24}$$

$$\begin{aligned} P_1 &= F_{\mathcal{D}}(w_S) - F_S(w_S) && (25)\\ &\leq Rep_{\mathcal{D}}(S;\mathcal{F}) \leq^{1-2\delta} 2R(\mathcal{F}|_S) + 3c\sqrt{\frac{2\ln(2/\delta)}{N}}. && (26)\end{aligned}$$

$$\begin{aligned} P_2 &= F_S(w_S) - F_S(w^*) && (27)\\ &\leq 0. && (28)\end{aligned}$$

According to the McDiarmid's inequality.

$$F_S(w^*) - F_{\mathcal{D}}(w^*) \leq^{1-\delta} \frac{2c}{\sqrt{m}}\sqrt{\frac{m\ln(2/\delta)}{2}}. \tag{29}$$

Combining 25 to 29, and following [20], with a probability of at least $1-3\delta$, we obtain

$$F_{\mathcal{D}}(w_S) - F_{\mathcal{D}}(w^*) \leq 2 \sup_{w \in \mathcal{W}} Var(F) + 4c\sqrt{\frac{2\ln(2/\delta)}{N}}. \tag{30}$$

□

VIII. Derivations of the Update Rule for the Player

We mainly refer [19] to the iterative derivation of p. We introduce the Legendre function related to p as $\Phi_p(p) = \sum_{i=1}^N p_i \log p_i$. Assume the mirror gradient ascent step $t+1$, without loss of generality, in the dual space is q^{t+1}. The mirror gradient ascent step of the dual space can be defined as follows:

$$\nabla\Phi_p(q^{t+1}) = \nabla\Phi_p(p^t) + \eta_b \mathbf{F}(w^t), \tag{31}$$

where $\mathbf{F}(w^t)$ denotes the loss vector of all participants, η_b is the step size of the dual space, also known as MAB step size in main paper. For any client i, the corresponding i-th element of $\nabla\Phi_p(q^{t+1})$ is:

$$\nabla\Phi_p(q_i^{t+1}) = 1 + \log q_i^{t+1}. \tag{32}$$

Combining 31 and 32, the representational relationship between q and p is deduced as:

$$q_i^{t+1} = e^{(\nabla\Phi_p(p_i^t) + \eta_b F_i(w_i^t) - 1)}. \tag{33}$$

After updating the dual space weights q^t, in order to map them back to the original problem space to get update of p^t, we need to solve the following problem:

$$p^{t+1} = \arg\min_{p \in \mathcal{P}_{\rho,\mathbf{N}}} D_{\Phi_p}(p, q^{t+1}), \tag{34}$$

where $D_{\Phi_p}(p, q^{t+1}) = \Phi_p(p) - \Phi_p(q^{t+1}) - \langle \nabla\Phi_p(q^{t+1}), p - q^{t+1} \rangle$ is the Bregman Divergence.

Incorporating the constraints of p, we establish the Lagrangian function as follows:

$$\begin{aligned} F(p^{t+1}, \beta, \lambda) &= \sum_{i=1}^N p_i^{t+1} \log \frac{p_i^{t+1}}{q_i^{t+1}} \\ &- \beta\left(\sum_{i=1}^N p_i^{t+1} - 1\right) - \lambda\left(\rho - \sum_{i=1}^N p_i^{t+1} \log p_i^{t+1} N\right). \end{aligned} \tag{35}$$

Using the first order conditions, we have

$$p_i^{t+1} = (q_i^{t+1})^{\frac{1}{1+\lambda}} N^{-\frac{\lambda}{1+\lambda}} \exp\left(\frac{\alpha}{1+\lambda} - 1\right). \tag{36}$$

Combining with $\sum_{i=1}^N p_i^{t+1} = 1$, the constant part can be replaced by

$$N^{-\frac{\lambda}{1+\lambda}} \exp\left(\frac{\alpha}{1+\lambda} - 1\right) = \frac{1}{\sum_{i=1}^N (q_i^{t+1})^{\frac{1}{1+\lambda}}}, \tag{37}$$

and then we have

$$p_i^{t+1} = (q_i^{t+1})^{\frac{1}{1+\lambda}} / \left(\sum_{i=1}^N (q_i^{t+1})^{\frac{1}{1+\lambda}}\right). \tag{38}$$

By substituting 38 back into the Lagrangian function and taking the derivative with respect to lambda, we have

$$\begin{aligned} \frac{d}{d\lambda}\mathcal{L}(\lambda) &= \log N - \rho - \log \sum_{i=1}^N (q_i^{t+1})^{\frac{1}{1+\lambda}} \\ &- \frac{\sum_{i=1}^N log(q_i^{t+1})(q_i^{t+1})^{\frac{1}{1+\lambda}}}{(1+\lambda)\sum_{i=1}^N (q_i^{t+1})^{\frac{1}{1+\lambda}}}. \end{aligned} \tag{39}$$

Combining 39 and 33, the update formula for p can be expressed as

$$p_i^{t+1} = \frac{e^{\frac{1}{1+\lambda^*}(\log p_i^t + \eta_b F_i(w_i^t))}}{\sum_{i=1}^N e^{\frac{1}{1+\lambda^*}(\log p_i^t + \eta_b F_i(w_i^t))}}, \tag{40}$$

where λ^* is the kernel of $\frac{d}{d\lambda}\mathcal{L}(\lambda)$ which is denoted as $f(\lambda)$ in the main paper.

IX. Convergence of FedMABA

Recall the assumptions we've made to derivate the convergence of FedMABA:

Assumption 7 (Smoothness). *Each objective function of clients is Lipschitz smooth, that is, there exists a constant $L > 0$, such that $\|\nabla F_i(\boldsymbol{x}) - \nabla F_i(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\|, \forall i \in \{1, 2, \ldots, N\}$.*

Assumption 8 (Unbiased Gradient and Bounded clients' Variance). *The stochastic gradient calculated by each client can be an unbiased estimator of the clients' gradient $\mathbb{E}_\xi[g_i(\boldsymbol{p^{avg}}|\xi)] = \nabla F_i(\boldsymbol{p^{avg}})$, and has bounded variance $\mathbb{E}_\xi[\|g_i(\boldsymbol{p^{avg}}|\xi) - \nabla F_i(\boldsymbol{p^{avg}})\|^2] \le \sigma^2, \forall i \in \{1, 2, \ldots, N\}, \sigma^2 \ge 0$.*

Assumption 9 (Bounded Dissimilarity of Clients' Gradient). *For any sets of weights $\{p_i^t \ge 0\}_{i=1}^N, \sum_{i=1}^N p_i^t = 1$,, there exist constants $(\gamma^2 + 1) \ge 1$, $A^2 \ge 1$, such that $\sum_{i=1}^N p_i^t \|\nabla F_i(\boldsymbol{p^{avg}})\|^2 \le \gamma^2 \left\|\sum_{i=1}^N p_i^t \nabla F_i(\boldsymbol{p^{avg}})\right\|^2 + A^2$.*

Assumption 10 (Bounded Weights Divergence). *For any sets of weights p^t derived by FedMABA, $\chi^2_{p^a\|p^t}$, the chi-square divergence of p^t and the average weights $p^a = \left[\frac{1}{N}, ..., \frac{1}{N}\right]$ can be upper bounded by κ, that is, $\forall t = 1, ..., T$, there exists a constant κ, such that $\chi^2_{p^a\|p^t} \le \kappa$, where $\chi^2_{p^a\|p^t} = \sum_{i=1}^N \left(p_i^a - p_i^t\right)^2 / p_i^t$.*

We first explain the gradient differences caused by the aggregation probability bias, following [31].

Lemma 2 (Including bias in the error bound.). *For any model parameter w, the difference between the gradients of $F^{avg}(w) = \sum_{i=1}^N p_i^{avg} F_i(w_i)$ and $F(w) = \sum_{i=1}^N p_i F_i(w_i)$ can be bounded as follows:*

$$\|\nabla f^{avg}(w) - \nabla f(w)\|^2 \le \chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}}\left[(\gamma^2 - 1)\|\nabla f(w)\|^2 + A^2\right], \tag{41}$$

where $\chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}}$ denotes the chi-square distance between $\boldsymbol{p^{avg}}$ and $\boldsymbol{p}$, i.e., $\chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}} = \sum_{i=1}^N \left(p_i^{avg} - p_i\right)^2 / p_i$. $f(x)$ is the global objective with $f(w) = \sum_{i=1}^N p_i f_i(w)$ where $\boldsymbol{p^{avg}}$ is usually the data ratio of clients. $f(w) = \sum_{i=1}^N p_i f_i(w)$ is the objective function of FedMABA with the reweight aggregation probability $\boldsymbol{p}$.

Proof.

$$\begin{aligned}\nabla f^{avg}(x) - \nabla f(\boldsymbol{x}) &= \sum_{i=1}^N \left(p_i^{avg} - p_i\right) \nabla f_i^{avg}(w) \\ &= \sum_{i=1}^N \left(p_i^{avg} - p_i\right) \left(\nabla f_i^{avg}(w) - \nabla f(w)\right) \\ &= \sum_{i=1}^M \frac{p_i^{avg} - p_i}{\sqrt{p}_i} \cdot \sqrt{p}_i \left(\nabla f_i^{avg}(w) - \nabla f(w)\right).\end{aligned} \tag{42}$$

Applying Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}\|\nabla f^{avg}(w) - \nabla f(w)\|^2 &\le \left[\sum_{i=1}^N \frac{\left(p_i^{avg} - p_i\right)^2}{p_i}\right]\left[\sum_{i=1}^N p_i \left\|\nabla f_i^{avg}(w) - \nabla f(w)\right\|^2\right] \\ &\le \chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}}\left[(\gamma^2 - 1)\|\nabla f(w)\|^2 + A^2\right],\end{aligned} \tag{43}$$

where the last inequality uses Assumption 9. Note that

$$\begin{aligned}\|\nabla f^{avg}(w)\|^2 &\le 2\|\nabla f^{avg}(w) - \nabla f(w)\|^2 + 2\|\nabla f(w)\|^2 \\ &\le 2\left[\chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}}(\gamma^2 - 1) + 1\right]\|\nabla f(w)\|^2 + 2\chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}} A^2.\end{aligned} \tag{44}$$

As a result, we obtain

$$\min_{t \in [T]} \left\|\nabla f^{avg}\left(w^t\right)\right\|^2 \le \frac{1}{T}\sum_{t=0}^{T-1} \left\|\nabla f^{avg}\left(w^t\right)\right\|^2 \tag{45}$$

$$\le 2\left[\chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}}(\gamma^2 - 1) + 1\right]\frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla f\left(w^t\right)\right\|^2 + 2\chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}} A^2 \tag{46}$$

$$\le 2\left[\chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}}(\gamma^2 - 1) + 1\right]\epsilon_{\text{opt}} + 2\chi^2_{\boldsymbol{p^{avg}}\|\boldsymbol{p}} A^2, \tag{47}$$

where $\epsilon_{\text{opt}} = \frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla f\left(w^t\right)\right\|^2$ denotes the optimization error. □

Following [4], we present the following convergence analysis.

Lemma 3 (Local updates bound.)**.** *For a sufficiently small client's step size $\eta_c \leq \frac{1}{8LK}$, the local updates can be bounded as:*

$$\mathbb{E}\|w_{i,k}^t - w^t\|^2 \leq 20K^2(\eta_c^2\sigma^2 + \eta_c^2A^2 + \gamma^2\eta_c^2\|\nabla F(x^t)\|^2). \tag{48}$$

Proof.

$$\begin{aligned}
\mathbb{E}_t\|w_{i,k}^t - w^t\|^2 &= \mathbb{E}_t\|w_{i,k-1}^t - w^t - \eta_c g_{i,k-1}^t\|^2 &(49)\\
&= \mathbb{E}_t\|w_{i,k-1}^t - w^t - \eta_c(g_{i,k-1}^t - \nabla F_i(w_{i,k-1}^t) + \nabla F_i(w_{t,k-1}^t) - \nabla F_i(w^t) + \nabla F_i(w^t))\|^2 &(50)\\
&\leq (1+\frac{1}{2K-1})\mathbb{E}_t\|w_{i,k-1}^t - w^t\|^2 + 2K\mathbb{E}_t\|\eta_c(g_{i,k-1}^t - \nabla F_i(w_{i,k}^t))\|^2\\
&\quad + 2K\mathbb{E}_t[\|\eta_L(\nabla F_i(w_{i,K-1}^t) - \nabla F_i(w^t))\|^2] + 2K\eta_c^2\mathbb{E}_t\|\nabla F_i(w^t)\|^2 &(51)\\
&\leq (1+\frac{1}{2K-1})\mathbb{E}_t\|w_{i,k-1}^t - w^t\|^2 + 2K\eta_c^2\sigma^2 + 2K\eta_c^2L^2\mathbb{E}_t\|w_{i,k-1}^t - w^t\|^2\\
&\quad + 2K\eta_c^2A^2 + 2K\gamma^2\|\eta_c\nabla f(w^t)\|^2 &(52)\\
&\leq (1+\frac{1}{K-1})\mathbb{E}_t\|w_{i,k-1}^t - w^t\|^2 + 2K\eta_c^2\sigma^2 + 2K\eta_c^2A^2 + 2K\gamma^2\|\eta_c\nabla F(w^t)\|^2. &(53)
\end{aligned}$$

By recursively applying the above formula, we can obtain,

$$\begin{aligned}
\mathbb{E}_t\|w_{i,k}^t - w^t\|^2 &\leq \sum_{p=0}^{k-1}(1+\frac{1}{K-1})^p\left[4K\eta_c^2\sigma^2 + 4K\eta_c^2A^2 + 4K\gamma^2\|\eta_c\nabla F(w^t)\|^2\right] &(54)\\
&\leq (K-1)\left[(1+\frac{1}{K-1})^K - 1\right]\left[4K\eta_c^2\sigma^2 + 4K\eta_c^2A^2 + 4K\gamma^2\|\eta_c\nabla F(w^t)\|^2\right] &(55)\\
&\leq 20K^2(\gamma^2\eta_c^2\|\nabla F(w^t)\|^2 + \eta_c^2\sigma^2 + \eta_c^2A^2)\,. &(56)
\end{aligned}$$

□

We thus can formulate the convergence analysis of FedMABA.

Theorem 4 (Convergence bound)**.** *Under Assumption 7 to 10, let η_s, η_c be the server updating step size, and the clinet's one, respectively. Set η_c small enough as $\eta_c < \min(\frac{1}{8LK}, C)$, where C is a constant, such that $\frac{1}{2} - 10L^2\frac{1}{N}\sum_{i=1}^N K^2\eta_c^2\gamma^2(\chi^2_{\boldsymbol{p}^{avg}\|\boldsymbol{p}}(\gamma^2) - 1)^2 + 1) \leq c \leq 0$, and $\eta_s < \frac{1}{\eta_c L}$, the expectation of the gradient norm can be upper bounded when running FedMABA as:*

$$\underset{t\in[T]}{min}\,\mathbb{E}_t\|\nabla F(w^t)\|^2 \leq \frac{F^0 - F^*}{c\eta_s\eta_c KT} + \Psi, \tag{57}$$

where

$$\Psi = \frac{1}{C}[10\eta_c^2K^2L^2(\sigma^2 + A^2) + \frac{\eta_s\eta_c L}{2}\sigma^2 + 20L^2K^2\gamma^2\eta_c^2\chi^2_{\boldsymbol{p}^{avg}\|\boldsymbol{p}}A^2]\,. \tag{58}$$

where $\chi^2_{\boldsymbol{p}^{avg}\|\boldsymbol{p}} = \sum_{i=1}^N\left(p_i^{avg} - p_i\right)^2/p_i$ represents the chi-square divergence between the average aggragation weights p^{avg} and the MAB aggregation weights p.

Proof.

$$\begin{aligned}
&E_t[F(w^{t+1})] &(59)\\
&\leq F(w^t) + \left\langle\nabla F(w^t), E_t[w^{t+1} - w^t]\right\rangle + \frac{L}{2}E_t[\left\|w^{t+1} - w^t\right\|^2] &(60)\\
&= F(w^t) + \left\langle\nabla F(w^t), E_t[\eta_s\Delta_t + \eta_s\eta_c K\nabla F(w^t) - \eta_s\eta_c K\nabla F(w^t)]\right\rangle + \frac{L}{2}\eta_s^2E_t[\|\Delta_t\|^2] &(61)\\
&= F(w^t) - \eta_s\eta_c K\left\|\nabla F(w^t)\right\|^2 + \eta_s\underbrace{\left\langle\nabla F(w^t), E_t[\Delta_t + \eta_c K\nabla F(w^t)]\right\rangle}_{P_1} + \frac{L}{2}\eta_s^2\underbrace{E_t\|\Delta_t\|^2}_{P_2}, &(62)
\end{aligned}$$

$$
\begin{aligned}
P_1 &= \left\langle \nabla F(w^t), E_t[\Delta_t + \eta_c K \nabla F(w^t)] \right\rangle && (63)\\
&= \left\langle \nabla F(w^t), E_t[-\sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \eta_c g_{i,k}^t + \eta_c K \nabla F(w^t)] \right\rangle && (64)\\
&= \left\langle \nabla F(w^t), E_t[-\sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \eta_c \nabla F_i(w_{i,k}^t) + \eta_c K \nabla F(w^t)] \right\rangle && (65)\\
&= \left\langle \sqrt{\eta_c K} \nabla F(w^t), -\frac{\sqrt{\eta_c}}{\sqrt{K}} E_t[\sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} (\nabla F_i(w_{i,k}^t) - \nabla F_i(w^t))] \right\rangle && (66)\\
&= \frac{\eta_c K}{2} \|\nabla F(w^t)\|^2 + \frac{\eta_c}{2K} E_t \left\| \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} (\nabla F_i(w_{i,k}^t) - \nabla F_i(w^t)) \right\|^2 \\
&\quad - \frac{\eta_c}{2K} E_t \| \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \nabla F_i(w_{i,k}^t) \|^2 && (67)\\
&\leq \frac{\eta_c K}{2} \|\nabla F(w^t)\|^2 + \frac{\eta_c}{2} \sum_{k=0}^{K-1} \sum_{i=1}^{N} p_i^{avg} \mathbb{E}_t \left\| \nabla F_i(w_{i,k}^t) - \nabla F_i(w^t) \right\|^2 \\
&\quad - \frac{\eta_c}{2K} \mathbb{E}_t \| \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \nabla F_i(w_{i,k}^t) \|^2 && (68)\\
&\leq \frac{\eta_c K}{2} \|\nabla F(w^t)\|^2 + \frac{\eta_c L^2}{2N} \sum_{i=1}^{N} \sum_{k=0}^{K-1} \mathbb{E}_t \left\| w_{i,k}^t - w^t \right\|^2 - \frac{\eta_c}{2K} \mathbb{E}_t \| \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \nabla F_i(w_{i,k}^t) \|^2 && (69)\\
&\leq \left(\frac{\eta_c K}{2} + 10 K^3 L^2 \eta_c^3 \gamma^2 \right) \|\nabla F(w^t)\|^2 + 10 L^2 \eta_c^3 K^3 \sigma^2 + 10 \eta_c^3 L^2 K^3 A^2 \\
&\quad - \frac{\eta_c}{2K} \mathbb{E}_t \| \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \nabla F_i(w_{i,k}^t) \|^2 , && (70)
\end{aligned}
$$

$$
\begin{aligned}
P_2 &= \mathbb{E}_t \|\Delta_t\|^2 && (71)\\
&= \mathbb{E}_t \left\| \eta_c \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} g_{i,k}^t \right\|^2 && (72)\\
&= \eta_c^2 \mathbb{E}_t \left\| \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} g_{i,k}^t - \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \nabla F_i(w_{i,k}^t) \right\|^2 + \eta_c^2 \mathbb{E}_t \left\| \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \nabla F_i(w_{i,k}^t) \right\|^2 && (73)\\
&\leq \eta_c^2 \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \mathbb{E} \| g_i(w_{i,k}^t) - \nabla F_i(w_{i,k}^t) \|^2 + \eta_c^2 \mathbb{E}_t \| \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \nabla F_i(w_{i,k}^t) \|^2 && (74)\\
&\leq \eta_c^2 K \sigma^2 + \eta_c^2 \mathbb{E}_t \| \sum_{i=1}^{N} p_i^{avg} \sum_{k=0}^{K-1} \nabla F_i(w_{i,k}^t) \|^2 . && (75)
\end{aligned}
$$

Now we take expectation over iteration on both sides of expression:

$$F(w^{t+1}) \tag{76}$$

$$\leq F(w^t) - \eta_s\eta_c K\mathbb{E}_t \left\|\nabla F(w^t)\right\|^2 + \eta_s\mathbb{E}_t \left\langle \nabla F(w^t), \Delta_t + \eta_c K\nabla F(w^t)\right\rangle + \frac{L}{2}\eta_s^2\mathbb{E}_t\|\Delta_t\|^2 \tag{77}$$

$$\begin{aligned}&\leq F(w^t) - \eta_s\eta_c K\left(\frac{1}{2} - 20L^2K^2\eta_c^2\gamma^2(\chi^2_{\boldsymbol{p}^{\boldsymbol{avg}}\|\boldsymbol{p}}(\gamma^2-1)+1)\right)\mathbb{E}_t\left\|\nabla f(w^t)\right\|^2\\ &\quad + 10\eta_s\eta_c^3L^2K^3(\sigma^2+A^2) + \frac{\eta_s^2\eta_c^2KL}{2}\sigma^2 + 20L^2K^3\gamma^2\eta_s\eta_c^3\chi^2_{\boldsymbol{p}^{\boldsymbol{avg}}\|\boldsymbol{p}}A^2\\ &\quad - \left(\frac{\eta_s\eta_c}{2K} - \frac{L\eta_s^2\eta_c^2}{2}\right)\mathbb{E}_t\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(x_{i,k}^t)\right\|^2\end{aligned} \tag{78}$$

$$\begin{aligned}&\leq F(w^t) - C\eta_s\eta_c K\mathbb{E}\left\|\nabla f(w^t)\right\|^2 + 10\eta_s\eta_c^3L^2K^3(\sigma^2+A^2) \end{aligned} \tag{79}$$

$$\begin{aligned}&\quad + \frac{\eta_s^2\eta_c^2KL}{2}\sigma^2 + 20L^2K^3\gamma^2\eta_s\eta_c^3\chi^2_{\boldsymbol{p}^{\boldsymbol{avg}}\|\boldsymbol{p}}A^2\\ &\quad - \left(\frac{\eta_s\eta_c}{2K} - \frac{L\eta_s^2\eta_c^2}{2}\right)\mathbb{E}_t\left\|\frac{1}{N}\sum_{i=1}^{N}\sum_{k=0}^{K-1}\nabla F_i(w_{i,k}^t)\right\|^2\end{aligned} \tag{80}$$

$$\begin{aligned}&\leq F(w^t) - c\eta_s\eta_c K\mathbb{E}_t\|\nabla F(w^t)\|^2 + 10\eta_s\eta_c^3L^2K^3(\sigma^2+A^2)\\ &\quad + \frac{\eta_s^2\eta_c^2KL}{2}\sigma^2 + 20L^2K^3\gamma^2\eta_s\eta_c^3\chi^2_{\boldsymbol{p}^{\boldsymbol{avg}}\|\boldsymbol{p}}A^2\,,\end{aligned} \tag{81}$$

By recursively summing the the error difference above, we can obtain

$$\sum_{t=1}^{T-1} C\eta_s\eta_c K\mathbb{E}\|\nabla F(w^t)\|^2 \leq F(w^0) - F(w^T) + T(\eta_s\eta_c K)\Psi, \tag{82}$$

where

$$\Psi = \frac{1}{C}[10\eta_c^2K^2L^2(\sigma^2+A^2) + \frac{\eta_s\eta_c L}{2}\sigma^2 + 20L^2K^2\gamma^2\eta_c^2\chi^2_{\boldsymbol{p}^{\boldsymbol{avg}}\|\boldsymbol{p}}A^2]\,. \tag{83}$$

Corollary 3. *Suppose η_s and η_c are $\eta_c = \mathcal{O}\left(\frac{1}{\sqrt{T}KL}\right)$ and $\eta_s = \mathcal{O}\left(\sqrt{KN}\right)$ such that the conditions mentioned above are satisfied. Then for sufficiently large T, the iterates of FedMABA satisfy:*

$$\min_{t\in[T]}\mathbb{E}\left\|\nabla f\left(w\right)\right\|^2 \leq \mathcal{O}\left(\frac{1}{\sqrt{NKT}} + \frac{1}{T}\right). \tag{84}$$

□