# Convergence Analysis

## I. Full Notations

We first provide a comprehensive set of additional variables to illustrate the convergence analysis beforehand, without a sudden introduction into the analysis process. Let us denote $\mathcal{F}_{\theta_m}$ and $f_{\theta_m}$ are the joint source-channel auto-encoder and the local representation feature extraction functions for the client $m$, respectively.

Our client's loss function is presented as follows:

$$\mathcal{L}(\theta_i; I, \hat{I}) = \mathcal{L}_{\text{recon}}(\mathcal{F}(\theta_m; I), \hat{I}) + \lambda \| f_m(\theta_m; I) - G\|_2^2, \quad (1)$$

where the global representation is constructed by

$$G = \frac{1}{M}\sum_{m=1}^{M} F_m^S \quad (2)$$

with the local representation feature of client $m$ being calculated as

$$F_m^S = \frac{1}{D_m}\sum_{i=1}^{D_m} f_{\theta_m}(I_i). \quad (3)$$

For the iteration notation for local and global rounds, we denote $r$ as the current global communication round, and $E$ as the total local iteration round. Therefore, we refer to the local iteration $e$ in the communication round $r$ as $rE + e$.

## II. Assumptions

**Assumption 1.** *The learning objective function at local client is $L_1$-Lipschitz smooth, in other words, the gradient of the loss is $L_1$-Lipschitz continuous,*

$$\|\nabla \mathcal{L}_{r_1} - \nabla \mathcal{L}_{r_2}\|_2 \leq L_1 \|\theta_{m,r_1} - \theta_{m,r_2}\|_2, \forall r_1, r_2 > 0, \quad (4)$$

*which indicates the following quadratic bound,*

$$\mathcal{L}_{r_1} - \mathcal{L}_{r_2} \leq \langle \nabla \mathcal{L}_{r_2}, (\theta_{m,r_1} - \theta_{m,r_2})\rangle + \frac{L_1}{2}\|\theta_{m,r_1} - \theta_{m,r_2}\|_2^2 \quad (5)$$

**Assumption 2.** *The stochastic gradient $g_{m,r} = \nabla \mathcal{L}(\theta_{m,r}, \xi_r)$ is an unbiased estimator of the local gradient for training the client. The expectation of its gradient is*

$$\mathbb{E}_{\xi_m \sim D_m} = \nabla \mathcal{L}(\theta_{m,r}) = \nabla \mathcal{L}_r, \quad (6)$$

*and the variance of it is bounded by $\sigma^2$*

$$\mathbb{E}[\|g_{m,r} - \nabla \mathcal{L}(\theta_{m,r})\|_2^2] \leq \sigma^2 \quad (7)$$

**Assumption 3.** *The expectation of the stochastic gradient is bounded by $V$*

$$\mathbb{E}[\|g_{m,r}\|_2] = V \quad (8)$$

**Assumption 4.** *The local feature extraction function is $L_2$-Lipschitz continuous*

$$\|f_m(\theta_{m,r_1}) - f_m(\theta_{m,r_2}))\|_2 \leq L_2 \|\theta_{m,r_1} - \theta_{m,r_2}\|_2 \quad (9)$$

## III. Key Lemmas

**Lemma 1.** *The assumptions 1 and 2 hold. From the beginning of the communication round $r+1$ to the last local update step, the loss function of a client from the network can be bounded as:*

$$\mathbb{E}[\mathcal{L}_{(r+1)E}] \leq \mathcal{L}_{rE+1/2} - (\eta - \frac{L_1\eta^2}{2})\sum_{e=1/2}^{E-1}\|\nabla \mathcal{L}_{rE+e}\|_2^2 + \frac{L_1 E\eta^2}{2}\sigma^2 \quad (10)$$

*Proof.* As we consider this lemma for an arbitrary client, we can freely omit the client notation $m$, then we rewrite *Assumption 1* as follows

$$\mathcal{L}_{rE+1} \leq \mathcal{L}_{rE+1/2} + \langle \nabla \mathcal{L}_{rE+1/2}, (\theta_{rE+1} - \theta_{rE+1/2})\rangle + \frac{L_1}{2}\|\theta_{rE+1} - \theta_{rE+1/2}\|_2^2, \quad (11)$$

and let $\theta_{rE+1} = \theta_{rE+1/2} - \eta g_{rE+1/2}$, thus we have

$$\mathcal{L}_{rE+1} \leq \mathcal{L}_{rE+1/2} - \eta\langle \nabla \mathcal{L}_{rE+1/2}, g_{rE+1/2}\rangle + \frac{L_1}{2}\|\eta g_{rE+1/2}\|_2^2 \quad (12)$$

Then we take the expectation for both sides of the above equation on the random variable $\xi_{rE+1/2}$, and we obtain

$$\mathbb{E}[\mathcal{L}_{rE+1}] \leq \mathcal{L}_{rE+1/2} - \eta\mathbb{E}[\langle \nabla \mathcal{L}_{rE+1/2}, g_{rE+1/2}\rangle] + \frac{L_1\eta^2}{2}\mathbb{E}[\|g_{rE+1/2}\|_2^2], \quad (13)$$

then we apply the *Assumption 2* to replace $\mathbb{E}[g_{rE+1/2}]$ with the $\nabla \mathcal{L}_{rE+1/2}$,

$$\mathbb{E}[\mathcal{L}_{rE+1}] \leq \mathcal{L}_{rE+1/2} - \eta\|\nabla \mathcal{L}_{rE+1/2}\|_2^2 + \frac{L_1\eta^2}{2}\mathbb{E}[\|g_{rE+1/2}\|_2^2], \quad (14)$$

$$\leq \mathcal{L}_{rE+1/2} - \eta\|\nabla \mathcal{L}_{rE+1/2}\|_2^2 + \frac{L_1\eta^2}{2}(\|\mathcal{L}_{rE+1/2}\|_2^2 + Var(g_{rE+1/2})) \quad (15)$$

$$= \mathcal{L}_{rE+1/2} - (\eta - \frac{L_1\eta^2}{2})\|\nabla \mathcal{L}_{rE+1/2}\|_2^2 + \frac{L_1\eta^2}{2}Var(g_{rE+1/2}) \quad (16)$$

$$\leq \mathcal{L}_{rE+1/2} - (\eta - \frac{L_1\eta^2}{2})\|\nabla \mathcal{L}_{rE+1/2}\|_2^2 + \frac{L_1\eta^2}{2}\sigma^2, \quad (17)$$

from (14) to (15), we consider the $Var(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$, while (16) to (17) follow the *Assumption 2* and take the expectation of $\theta$ on both sides. Through telescoping the sequence for all E iterations, we derive

$$\mathbb{E}[\mathcal{L}_{(r+1)E}] \leq \mathcal{L}_{rE+1/2} - (\eta - \frac{L_1\eta^2}{2}) \sum_{e=1/2}^{E-1} \|\nabla \mathcal{L}_{rE+e}\|_2^2 + \frac{L_1 E\eta^2}{2}\sigma^2. \tag{18}$$

**Lemma 2.** *Consider the assumption 3 and 4 hold, the global representation aggregation at the server, the loss function for a random client within the network can be bounded as:*

$$\mathbb{E}[\mathcal{L}_{(r+1)E+1/2}] \leq \mathcal{L}_{(r+1)E} + \lambda L_2 \eta E G. \tag{19}$$

*Proof.* *as below*

$$\begin{aligned}
\mathcal{L}_{(r+1)E+1/2} &= \mathcal{L}_{(r+1)E} + \mathcal{L}_{(r+1)E+1/2} - \mathcal{L}_{(r+1)E} && (20)\\
&= \mathcal{L}_{(r+1)E} + \lambda \|f(\theta_{(r+1)E}) - G_{r+2}\|_2 \\
&- \lambda \|f(\theta_{(r+1)E}) - G_{r+1}\|_2) && (21)\\
&\leq \mathcal{L}_{(r+1)E} + \lambda \|G_{r+2} - G_{r+1}\|_2 && (22)\\
&= \mathcal{L}_{(r+1)E} + \lambda \| \sum_{m=1}^{M} \frac{F_{m,(r+1)E}}{M} - \sum_{m=1}^{M} \frac{F_{m,rE}}{M} \|_2 && (23)\\
&= \mathcal{L}_{(r+1)E} + \lambda \| \sum_{m=1}^{M} \frac{F_{m,(r+1)E} - F_{m,rE}}{M} \|_2 && (24)\\
&= \mathcal{L}_{(r+1)E} + \lambda \| \sum_{m=1}^{M} \frac{1}{MD_m} \sum_{i=1}^{D_m} (f_m(\theta_{m,(r+1)E}, I_i) \\
&- f_m(\theta_{m,rE}, I_i)\|_2 && (25)\\
&\leq \mathcal{L}_{(r+1)E} + \lambda \sum_{m=1}^{M} \frac{1}{MD_m} \sum_{i=1}^{D_m} \|(f_m(\theta_{m,(r+1)E}, I_i) \\
&- f_m(\theta_{m,rE}, I_i)\|_2 && (26)\\
&\leq \mathcal{L}_{(r+1)E} + \lambda L_2 \sum_{m=1}^{M} \frac{1}{M} \|\theta_{m,(r+1)E} - \theta_{m,rE}\|_2 && (27)\\
&= \mathcal{L}_{(r+1)E} + \lambda L_2 \eta \sum_{m=1}^{M} \frac{1}{M} \| \sum_{e=1/2}^{E-1} g_{m,rE+e}\|_2 && (28)\\
&\leq \mathcal{L}_{(r+1)E} + \lambda L_2 \eta \sum_{m=1}^{M} \frac{1}{M} \sum_{e=1/2}^{E-1} \|g_{m,rE+e}\|_2, && (29)
\end{aligned}$$

Then take the expectations of the random variable $\xi$ on both sides of the equation

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{(r+1)E+1/2}] &\leq \mathcal{L}_{(r+1)E} + \lambda L_2 \eta \sum_{m} \frac{1}{M} \sum_{e=1/2}^{E-1} \mathbb{E}[\|g_{m,rE+e}\|_2] && (30)\\
&\leq \mathcal{L}_{(r+1)E} + \lambda L_2 \eta E V, && (31)
\end{aligned}$$

where from (20) to (21), we replace the definition of local loss with our proposed loss, (21) to (22) is the result of this math $\|a-b\|_2$-$\|a-c\|_2 \leq \|b-c\|_2$, from (22) to (23) is the result of following the definition of global representation in (2). Similarly, (24) to (25) follows the definition of local representation, (25) to (26) and (28) to (29) follow $\|\sum a_i\|_2 \leq \sum\|a_i\|_2$. From (26) to (27) is the result of the *Assumption 4*, $L_2$-*Lipschitz* continuity and finally, (30) to (31) is the result by following *Assumption 3*.

## IV. Theorems and Corollary

**Theorem 1.** *When the Assumptions 1 to 4 hold, the local loss of an arbitrary client within the FL network after every communication round is presented*

$$\mathbb{E}[\mathcal{L}_{(r+1)E+1/2}] \leq \mathcal{L}_{rE+1/2} - (\eta - \frac{L_1\eta^2}{2}) \sum_{e=1/2}^{E-1} \|\nabla \mathcal{L}_{rE+e}\|_2^2 + \frac{L_1 E\eta^2}{2}\sigma^2 + \lambda L_2 E V. \tag{32}$$

**Corollary 1.** *The loss function $\mathcal{L}$ of a random client monotonously decreases in every communication round when*

$$\eta_{e'} < \frac{2\big(\sum_{e=1/2}^{e'} \|\mathcal{L}_{rE+e}\|_2^2 - \lambda L_2 E V\big)}{L_1\big(\sum_{e=1/2}^{e'} \|\nabla \mathcal{L}_{rE+e}\|_2^2 + E\sigma^2\big)}, \tag{33}$$

and

$$\lambda_r = \frac{\|\nabla \mathcal{L}_{rE+1/2}\|_2^2}{L_2 E V}. \tag{34}$$

*Thus, the loss function converges.*

**Theorem 2.** *Let Assumption 1 to 4 hold, while $\Delta = \mathcal{L}_0 - \mathcal{L}^*$, for any client and any $\epsilon > 0$, after*

$$R = \frac{2\Delta}{E\epsilon(2\eta - L_1\eta^2) - E\eta(L_1\eta\sigma^2) - 2\lambda L_2 V} \tag{35}$$

*communication rounds of FedDoM, we have*

$$\frac{1}{RE} \sum_{r=0}^{R-1} \sum_{e=1/2}^{E-1} \mathbb{E}[\|\nabla \mathcal{L}_{rE+e}\|_2^2] < \epsilon, \tag{36}$$

*if*

$$\eta < \frac{2(\epsilon - \lambda L_2 V)}{L_1(\epsilon + \sigma^2)}, \tag{37}$$

*and*

$$\lambda < \frac{\epsilon}{L_2 V}. \tag{38}$$

## V. Completing the Proof of Theorems and Corollary

**Proof for Theorem 1 and Corollary 1.** We take the expectation of $\theta$ on both sides in Lemma 1 and 2, then sum them up, we derive

$$\mathbb{E}[\mathcal{L}_{(r+1)E+1/2}] \leq \mathcal{L}_{rE+1/2} - (\eta - \frac{L_1\eta^2}{2}) \sum_{e=1/2}^{E-1} \|\nabla \mathcal{L}_{rE+e}\|_2^2 + \frac{L_1 E\eta^2}{2}\sigma^2 + \lambda L_2 E V. \tag{39}$$

then to make $-(\eta - \frac{L_1\eta^2}{2})\sum_{e=1/2}^{E-1}\|\nabla\mathcal{L}_{rE+e}\|_2^2 + \frac{L_1E\eta^2}{2}\sigma^2 + \lambda L_2 EV < 0$, we obtain

$$\eta < \frac{2(\sum_{e=1/2}^{E-1}\|\nabla\mathcal{L}_{rE+e}\|_2^2 - \lambda L_2 EV)}{L_1(\sum_{e=1/2}^{E-1}\|\nabla\mathcal{L}_{rE+e}\|_2^2 + E\sigma^2)}, \tag{40}$$

and

$$\lambda < \frac{\sum_{e=1/2^{E-1}}\|\nabla\mathcal{L}_{rE+e}\|_2^2}{L_2EV} \tag{41}$$

Therefore, in practice, we use

$$\eta_{e'} < \frac{2(\sum_{e=1/2}^{E-1}\|\nabla\mathcal{L}_{rE+e}\|_2^2 - \lambda L_2 EV)}{L_1(\sum_{e=1/2}^{E-1}\|\nabla\mathcal{L}_{rE+e}\|_2^2 + E\sigma^2)}, e' = 1/2, 1\ldots, E-1 \tag{42}$$

and

$$\lambda_r < \frac{\|\nabla\mathcal{L}_{rE+1/2}\|_2^2}{L_2EV} \tag{43}$$

So, the convergence of $\mathcal{L}$ holds.

**Proof for Theorem 2**

Take the expectation of $\theta$ on both sides in (32), then telescope through communication round from $r=0$ to $r=R-1$, with timestep from $e=1/2$ to $e=E$ in each communication round, we have

$$\frac{1}{RE}\sum_{r=0}^{R-1}\sum_{e=1/2}^{E-1}\mathbb{E}[\|\nabla\mathcal{L}_{rE+e}\|_2^2] \leq \frac{\frac{1}{RE}\sum_{r=0}^{R-1}(\mathcal{L}_{rE+1/2} - \mathbb{E}[\mathcal{L}_{(r+1)E+1/2}]) + \frac{L_1\eta^2}{2}\sigma^2 + \lambda L_2\eta V}{\eta - \frac{L_1\eta^2}{2}}, \tag{44}$$

given any $\epsilon > 0$, let

$$\frac{\frac{1}{RE}\sum_{r=0}^{R-1}(\mathcal{L}_{rE+1/2} - \mathbb{E}[\mathcal{L}_{(r+1)E+1/2}]) + \frac{L_1\eta^2}{2}\sigma^2 + \lambda L_2\eta V}{\eta - \frac{L_1\eta^2}{2}} < \epsilon, \tag{45}$$

that is

$$\frac{\frac{2}{RE}\sum_{r=0}^{R-1}(\mathcal{L}_{rE+1/2} - \mathbb{E}[\mathcal{L}_{(r+1)E+1/2}]) + L_1\eta^2\sigma^2 + 2\lambda L_2\eta V}{2\eta - L_1\eta^2} < \epsilon, \tag{46}$$

Let $\Delta = \mathcal{L}_0 - \mathcal{L}^*$. Since $\sum_{r=0}^{R-1}(\mathcal{L}_{rE+1/2} - \mathbb{E}[\mathcal{L}_{(r+1)E+1/2}]) \leq \Delta$, the above equation holds when

$$\frac{\frac{2\Delta}{RE} + L_1\eta^2\sigma^2 + 2\lambda L_2\eta V}{2\eta - L_1\eta^2} \leq \epsilon, \tag{47}$$

and that is

$$R > \frac{2\Delta}{E\epsilon(2\eta - L_1\eta^2) - E\eta(L_1\eta\sigma^2 + 2\lambda L_2 V)}. \tag{48}$$

Therefore, we have

$$\frac{1}{RE}\sum_{r=0}^{R-1}\sum_{e=1/2}^{E-1}\mathbb{E}[\|\nabla\mathcal{L}_{rE+e}\|_2^2] \leq \epsilon, \tag{49}$$

when

$$\eta < \frac{2(\epsilon - \lambda L_2 V)}{L_1(\epsilon + \sigma^2)} \tag{50}$$

and

$$\lambda < \frac{\epsilon}{L_2 V}. \tag{51}$$